Predicting the legal status of recreational marijuana in U.S. states

Alex Kristoffer Crozier

Master's thesis

Spring 2022

Department of Comparative Politics
University of Bergen

Abstract

Following decades of a war on drugs in the United States and the demonization of marijuana, a wave of recreational legalization began in 2012 with Washington state and Colorado. Recreational marijuana legalization is a novel phenomenon, both within and outside the U.S. borders. This thesis adds to the marijuana legalization research by creating prediction models that classifies observations (state-year) on whether U.S. states have legalized recreational marijuana or not in the timespan 2010 to 2018. It addresses the following research question:

'*To what extent, and how, is it possible to predict if a state has legalized recreational marijuana in the United States?'*

I have drawn from theories and literature that explain why policies change, as well as from theories and literature on why individuals support marijuana legalization. My focus is on public opinion and policy responsiveness. Such theories and literature play an important role in predicting and describing states that have legalized recreational marijuana. A wide range of data sources have been used in order to create the best predictive model possible. For instance, data from such as the General Social Survey and United States Census Bureau have been used to create input variables. Furthermore, the machine learning iteration of multilevel regression with post-stratification is central in terms of this thesis' data. This method, using the R package *autoMrP*, has allowed for the simulated disaggregation of the nationally representative public opinion variable in the General Social Survey.

Tree-based classification algorithms, shrinkage methods and support vector machines have been used to predict the legality of marijuana. A focus on description using machine learning (ML) has been done due to what I see as necessities and to illustrate the use of ML in the social sciences – even when traditional regression models were a viable alternative. Having created multiple models, support vector machines and gradient boosting machines proved to be the best prediction algorithms. In terms of *how* marijuana legality is best predicted, the abstraction of individual-level determinants, and the use of public opinion and medical legalization as input variables are central aspects of creating strong predictive models of recreationally legal marijuana. This thesis concludes with the need for studying medical legalization and its relationship to recreational legalization. This is because the results show that medical legalization is important in predicting recreational legality.

Acknowledgements

Jeg vil selvfølgelig takke familien min. Her er det mye å takke for, så jeg takker rett og slett for alt. Takk mamma, pappa, Tom, og Nico.

Videre retter jeg en stor takk til min veileder Cornelius Cappelen. Takk for gode tilbakemeldinger og samtaler, for at du svarte på spørsmålene mine raskt, og ikke minst for å ha oppmuntret meg de (heldigvis) få gangene jeg møtte veggen det siste året. Jeg vil også takke Reto Wüest for å ha hjulpet meg med det metodiske ved oppgaven min.

En spesiell takk til min trofaste kollokvie-partner og gode venn Ida Marie for all hjelpen, støtten, og latteren de siste to årene. Og ikke minst en stor takk til Nora Elvestad for tilbakemeldinger på masteroppgaven min. Jeg vil også takke alle mine venner, særlig dere jeg har delt de siste to årene med i Sofie Lindstrøms hus. Til slutt vil jeg takke alle i Odin Styrkeløftklubb.

Table of Contents

1. Introduction, context, and concepts **1**

1.1 Introduction **1**

1.2 Marijuana in the United States: from prohibition to legalization **3**

1.3 Models of marijuana regulation and legality **5**

2. Theory **8**

2.1 Structure and aim of the chapter **8**

2.2 Studying policy change and marijuana policy **9**

2.3 Public opinion and responsiveness **10**

2.3.1 What is public opinion? 11

2.3.2 Public opinion on marijuana legalization 11

2.3.3 Public opinion, responsiveness, and congruence 16

2.4 Medical marijuana – a trojan horse? **21**

2.5 Policy diffusion and policy learning **22**

2.6 State level determinants of marijuana legalization **24**

2.7 Expectations **25**

3. Data and Methods **28**

3.1 Why prediction and not explanation/regression? **29**

3.2 Measuring marijuana policy **31**

3.2.1 Marijuana policy in this thesis 31

3.2.2 Marijuana policy data 32

3.3 Simulating data - Multilevel regression with post-stratification **33**

3.3.1 MrP using autoMrP in R– a new approach 33

3.3.2 autoMrP data – General Social Survey, U.S. Census data and more 36

3.3.3 autoMrP models and results 40

3.4 Interpretive Predictive Machine Learning Models **42**

3.4.1 Input Data and Data Splitting 43

3.4.2 Choice of machine learning methods 46

3.4.3 How each machine learning method works 47

3.4.4 Model building 55

4. Results **60**

4.1 Predictions and model accuracy 61

4.2 Variable Importance Plots – opening the black box 66

4.3 Interaction plots.. 72

4.4 Variable coefficients – Shrinkage 77

5. Discussion and Conclusion 79

5.1 Conclusion... 79

5.2 Prediction of Recreational Marijuana Legality... 80

5.2.1 Which model predicts best, and why?..80

5.3 Contributions... 81

5.3.1 Theoretical implications and contributions...81

5.3.2 Methodological contributions..84

5.4 Limitations 85

6. References 88

A. Appendix 97

A.1 Logistic regression models.. 97

A.2 Variable list and dataset overview 98

A.3 Interaction plots.. 99

List of Figures

Figure 3.1 *Support towards marijuana legalization, United States, 2010-2018* 41

Figure 3.2 *autoMrP Ensemble Model Averaging (EBMA) by USCB Divisions, 2010-2018* ... 42

Figure 3.3 *2D and 3D hyperplanes* 49

Figure 3.4 *Kernel trick visualization* 50

Figure 3.5 *Decision tree partitioning* 51

Figure 3.6 *Decision Tree 2* 52

Figure 3.7 *Decision tree 1* 56

Figure 3.8 *Shrinkage penalties – first models* 58

Figure 3.9 *SVM Cost vs Accuracy (CV)* 59

Figure 4.1 *Ridge 1: Variable Importance Plot* 67

Figure 4.2 *Ridge 2: Variable Importance Plot (medyrs excluded)* 67

Figure 4.3 *ElasticNet 2: Variable Importance Plot (medyrs excluded)* 68

Figure 4.4 *Lasso 2: Variable Importance Plot (medyrs excluded)* 68

Figure 4.5 *SVM 1: Variable Importance Plot* 69

Figure 4.6 *SVM 3: Variable Importance Plot (marijuana variables excluded)* 70

Figure 4.7 *DT 2: Variable Importance Plot (medyrs excluded)* 71

Figure 4.8 *RF 1: Variable Importance Plot* 71

Figure 4.9 *GBM 3: Variable Importance Plot* 72

Figure 4.10 *SVM 1 Interaction Plot* 73

Figure 4.11 *SVM 1 Interaction Plot (medyrs)* 74

Figure 4.12 *SVM 2 Interaction Plot* 75

Figure 4.13 *SVM 2 Interaction Plot (hrelig)* 75

Figure 4.14 *GBM 3 Interaction Plot* 76

Figure 4.15 *GBM 3 Interaction Plot (ballot)* 76

Figure A.3.1 *SVM 1 Interaction Plot (diffusion)* 99

Figure A.3.2 *SVM 2 Interaction Plot (ballot)* 100

Figure A.3.3 *SVM 2 Interaction Plot (grass)* 100

Figure A.3.4 *GBM 3 Interaction Plot (diffusion)* 101

List of Tables

Table 4.1 *Confusion Matrix (example)* 60

Table 4.2 *Shrinkage Confusion Matrices* 62

Table 4.3 *SVM, DT and RF Confusion Matrices* 63

Table 4.4 *Boosting Confusion Matrices* 63

Table 4.5 *Shrinkage variable coefficients* 77

Table A.1.1 *Logistic Regression Table* 98

Table A.2.1 *Variable list* 98

Table A.2.2 *Dataset Overview* 99

Abbreviations

RML – Recreational Marijuana Legalization/Legality

MML – Medical Marijuana Legalization/Legality

MrP – Multilevel Regression with Post-stratification

EBMA – Ensemble Bayesian Model Averaging

SVM – Support Vector Machine

GBM – Gradient Boosting Machine

DT – Decision Tree

RF – Random Forest

VIP – Variable Importance Plot

GSS – General Social Survey

BLS – United States Bureau of Labor Statistics

USCB – United States Census Bureau

1. Introduction, context, and concepts

1.1 Introduction

Drug policy has significantly changed the last ten years in the United States. Beginning with Washington and Colorado in 2012, a total of 19 U.S. states have legalized marijuana for recreational use (NORML 2021). This momentum has not stopped: 6 of these 19 states legalized marijuana for recreational use in 2021. Despite this, marijuana is still illegal at the federal level. The incongruence between federal and state legislation motivates me to explore recreational legalization further. Specifically, this thesis examines what separates the states that have legalized marijuana for recreational use from those that have not. As such, the research question for this thesis is the following:

To what extent, and how, is it possible to predict if a state has legalized recreational marijuana in the United States?

The research question focuses on description and exploration, rather than causality. There are multiple motivations for such an approach. Firstly, in order to explain a phenomenon and its causes, a thorough descriptive understanding of it is essential. Additionally, explaining why states legalize marijuana for recreational use has not yet been done in the literature, apart from a few case studies. I therefore consider it as necessary for this thesis to do the groundwork through a descriptive approach. I analyse the differences between states that have, and have not, legalized marijuana for recreational use. Predictive machine learning models, such as ridge regression and gradient tree boosting, are useful tools precisely for this objective: description. Another reason I focus on prediction rather than regression relates to the nature of the available data. The available data faced several challenges, such as small number of observations, multicollinearity and high VIF scores, and sub-optimal operationalizations of certain variables. I therefore apply machine learning methods such as shrinkage regressions, Support Vector Machines, and tree-based methods to predict whether a state has legalized recreational marijuana or not, with biannual data from 2012 to 2018. As such, this is first and foremost a methodological thesis.

My focus is on recreational marijuana rather than medical marijuana. There are many reasons for this. First, few scholars have studied recreational marijuana, particularly political scientists. Secondly, making valid conceptualizations and operationalizations of recreational marijuana is a less encompassing task than doing so for medical marijuana. In light of this being a master's thesis, this is an important consideration. Thirdly, the legalization of recreational marijuana marks a significant shift in U.S. drug policy. The legalization of recreational marijuana and the U.S. war on drugs are in many ways complete opposites.

This thesis applies mainly two categories of theory/previous research to build a strong predictive model. More precisely, I use literature that aims to explain policy change. Public opinion and policy responsiveness is central considering a wide range of literature show that policy responds to the public's preferences (Erikson 1976; Page and Shapiro 1983; Lowery, Gray, and Hager 1989; Monroe 1979; Wlezien 1995; Lascher, Hagen, and Rochlin 1996; Burstein 2003; Lax and Phillips 2009a; Burstein 2020). Furthermore, I use theories and insight from previous research that attempts to explain the determinants of individuals' support towards marijuana legalization. This literature is used by abstracting individual-level determinants to the state-level as input variables for the predictive models. To illustrate, previous literature shows that religious people are less likely to support marijuana legalization. Therefore, one of the input variables is an index measuring the religiosity of each state. Having built machine learning models that predict marijuana legality quite accurately, I conclude this thesis by arguing that it is possible to predict the legality of recreational marijuana to a large extent – particularly if one builds multiple machine learning models that differ in their fundamentals, and if the input variables are theoretically justified. I also conclude with the need for understanding the relationship between medical marijuana and recreational marijuana legalization better.

Structure of this thesis

This thesis has the following structure. The present chapter includes the introduction, as well as a brief history of marijuana prohibition and legality in the United States. It furthermore includes a section discussing the different models of marijuana legality (illegality, decriminalization, medicalization, and recreational legality).

The second chapter – Theory – includes a review of the literature explaining marijuana policy. For instance, case-studies focused on describing and explaining medicalization and recreational legalization (e.g., Hannah and Mallinson (2018)). A substantial part of this chapter is dedicated to describing and explaining public opinion/policy responsiveness, as well as the specifics of marijuana public opinion. This chapter is therefore both a theory chapter and a literature review. Furthermore, a discussion of medicalization as a ‘trojan horse’, policy diffusion, and possible state-level determinants of recreational legalization are included. This chapter serves the purpose of justifying input variables for the prediction models

The third chapter, Data and Methods, describes the data and methods that I employ. How General Social Survey data is disaggregated using multilevel regression with post-stratification (MrP), as well as other sources of data, are discussed. The chapter is divided in two parts: the first explains MrP and how MrP has been conducted, as well as presenting the results of MrP (i.e., the disaggregated public opinion data). The second part explains the predictive machine learning methods; more precisely, the data used for them, why the particular methods have been chosen and how they differ, as well as how the models in this thesis have been built.

The penultimate fourth chapter, Results, includes a presentation of the results of the prediction models. The results are presented mainly in terms of each model’s predictive accuracy. Nevertheless, variable importance and interaction effects are also presented.

The fifth (final) chapter – Discussion and Conclusion – has a fourfold structure. The first part includes a discussion of why the best predictive models predict well, and what this implies for the research question. The second part includes a discussion this thesis’ theoretical and methodological contributions. Then follows a discussion of limitations related to this thesis’ methodology, design and focus is included, before I conclude.

1.2 Marijuana in the United States: from prohibition to legalization

The purpose of this section is to give a brief presentation of the history of marijuana prohibition and legalization in the United States.

Beginning in the late 1960s, the prominent ‘Mexican hypothesis’ was thought to explain the origins of marijuana use in the United States – gaining prominence through the work of David Musto (Campos 2018, 6-9). In short, the idea is that Mexicans imported marijuana to the United States in the early 1900s, both literally and culturally (Bonnie 2018, 577). This hypothesis has been widely accepted since its formation. However, historian Isaac Campos disagrees. Campos points out evidence of marijuana use and experimentation in the United States dating back to the mid 1800s (Campos 2018, 19). He furthermore suggests that there is little evidence of marijuana use being an exceptionally Mexican phenomenon in the United States in the early 1900s (2018, 21), as well as suggesting that marijuana use was uncommon across the southern border (Campos 2018, 7). During the early 1900s, use, sale and cultivation of marijuana, cocaine and heroin was not prohibited in the United States (Bonnie 2018, 577; MacCoun and Reuter 2001, 8). In addition to these findings and reflections, Campos points out that the use of recreational marijuana was not common in Mexico, being confined to mostly prisoners and soldiers barracks (Campos 2018, 10).

The prohibition of marijuana in the United States began at the local and state level during the 1910s and 20s (Bonnie 2018, 577). Following this, the Marihuana tax act of 1937 became law, de-facto prohibiting marijuana at the federal level (Stringer and Maggard 2016, 429; Campos 2018, 9). The tax act criminalized possession of marijuana without a federal tax stamp (Stringer and Maggard 2016, 429). However, tax stamps were difficult to obtain. Stringer and Maggard (2016, 429) have described them as "unattainable", while Campos (2018, 9) referred to the tax-act as a "de facto prohibition of cannabis nationwide".

In 1936, following the Reorganization act, the Federal Bureau of Narcotics (FBN) shifted its attention to marijuana – previously regarded as unworthy of the bureau's time. Marijuana was now equated with violent crime in the eyes of law-enforcement (Stringer and Maggard 2016, 429). During this time period 'educational' propaganda films such as Reefer Madness were created, solidifying marijuana's image of being a hard drug (Stringer and Maggard 2016, 429). In 1970, President Nixon declared a 'War on drugs', and the Controlled Substances Act (CSA) was enacted (Bonnie 2018, 578; USC n.d.). The following year, the National Commission on Marijuana and Drug Abuse published their report recommending the decriminalization of marijuana (Bonnie 2018, 578). This marked a temporary shift in US marijuana policy, ensued by a wave of decriminalization where a total of eleven states decriminalized marijuana (Bonnie 2018, 578). During this brief period of slight liberalization, President Carter endorsed the decriminalization of marijuana (Bonnie 2018, 584). This shift in sentiment took a turn following President Reagan and George H. W. Bush's administrations, who further consolidated the war on drugs (Bonnie 2018, 585).

Following their presidencies, a wave of medicalization occurred in the late 1990s and early 2000s. California was the first, legalizing medical marijuana in the 1996 Compassionate Use

Act (Bonnie 2018, 588). This brings us to today. Since Colorado and Washington state legalized recreational marijuana in 2012, a total of 19 states have legalized marijuana for recreational use in the U.S. (DISA 2022). Despite this, the Controlled Substances Act of 1971's categorization of marijuana as a Schedule I drug remains, sharing classification with heroin and ecstasy (USC n.d.). Marijuana is therefore still federally considered as a drug with "no currently accepted medical use and a high potential for abuse", despite showing potential for treating disorders such as Tourette's, multiple sclerosis and anorexia (DEA 2018; Mouhamed et al. 2018)

1.3 Models of marijuana regulation and legality

Studying marijuana legalization and policy requires a basic understanding of what is meant by terms such as legalization and medicalization. An understanding of how marijuana regulation presents itself empirically, and how one describes these models of regulations, is important. Equally important is knowing what recreational marijuana legalization is not. Distinguishing between different types of reform and laws is also important as this forms the basis of how the dependent variable, and some control variables, are coded and analysed in this thesis. Therefore, this section aims to define and explain the four main models of marijuana regulation and laws.

There are four main categories of marijuana laws. They are: 1) illegal (prohibited), 2) decriminalized, 3) medically legal, and 4) recreationally legal (Pacula and Smart 2017, 400-401). When broken down into these four categories, marijuana laws seem quite simple. Illegality, also known as prohibition, entails the criminal status of marijuana possession, use, cultivation, sale and distribution (Pacula and Smart 2017, 400). Marijuana use is prohibited at the federal level in the U.S., yet legal in some states due to state-level legislation (Johns 2015, 194). Besides varying by levels (state vs federal), marijuana varies in its criminal status – often by degree of offense (Pacula and Smart 2017, 400). The criminal status is often separated into two categories: misdemeanours and felonies – the former being less grave than the latter (Natapoff 2011, 1313). Prohibition of marijuana is naturally not dichotomous – any combination of use, distribution, sale, and cultivation may be prohibited, in theory. In the Netherlands, for instance, 'coffee shops' may sell marijuana to customers, yet the supply side (cultivators and distributors) is still illegally ran (Korf 2008, 151). For conceptual and analytical clarity, prohibition is in this thesis understood as an all-encompassing prohibition, where use, possession, distribution, sale, and cultivation are all prohibited by law, irrespective of the level of crime attributed (misdemeanour or felony).

Marijuana decriminalization is more nuanced and conceptually unclear than prohibition. First defined by the Schafer Commission in 1972, decriminalization of marijuana refers to policies that "do not define possession for personal use or casual (nonmonetary) distribution as a criminal offense" (Pacula and Smart 2017, 400). Decriminalization is, in theory, straightforward: the removal/non-existence of criminalization of an action. However, the term decriminalization is often used to describe policies that de-penalize marijuana use. Depenalization simply entails a reduction of penalties. For instance, reduced fines, reduced/removed mandatory jail sentences and other forms of punishment de-escalation (Pacula et al. 2005, 3-5). Decriminalization is thus a form of depenalization where penalties and punishments are removed entirely. Pacula and colleagues have strongly emphasized the importance of distinguishing between decriminalization and depenalization, and the shortcomings of failing to make this distinction (Pacula et al. 2005). Decriminalization may cover any combination of possession, use, distribution, sale, and cultivation. Nonetheless, it generally refers to the non-criminalization of possession and use.

There are two types of marijuana legalization, medical marijuana legalization (MML) and recreational marijuana legalization (RML). The former refers to the legalization of marijuana use for medical purposes. MML means different things in practice, as states have legislated medical marijuana differently (Bestrashniy and Winters 2015). In essence, MML implies that use and possession of marijuana is legal when used to treat illnesses and medical conditions. A medical marijuana card is needed to prove eligibility, often provided by a medical doctor. In Colorado, for instance, a medical card may be obtained after consulting with a health care provider if one is a resident, 18 or older, and has a qualifying medical condition (CDPHE 2021).

How medicalization is implemented varies significantly across states, especially regarding potency and availability. This heterogeneity of medical marijuana models is important because it justifies my focus on recreational marijuana legalization instead of medicalization. RML's homogeneity makes conceptualizing and operationalizing it in a valid manner more achievable in the context of a master's thesis.

It is important to understand how heterogeneous MML is, especially considering it is used as an input variable. In some states where medical marijuana is legal, the only way to obtain it legally is through cultivating it oneself. In practice, this makes marijuana unobtainable for many medical users as growing marijuana at home is time consuming (Bradford and Bradford 2017, 82). States also vary in the number of conditions that grant eligibility to medical marijuana. In 2014, Illinois permitted medical marijuana for 40 conditions, whilst Washington permitted for only 6 conditions (Bestrashniy and Winters 2015, 641). Medicalization also varies with respect to the extent to which THC (tetra-hydro-cannabinol), the primary psychoactive chemical in marijuana, is legal (Room et al. 2010, 5-6). Florida (in 2019), for instance, allowed for a mere 0.8% THC content in marijuana products (Mosher and Atkins 2019, 160). For illustration, dispensary marijuana in Nevada and Washington varies around the 20% potency-mark (Zoorob 2021). In practice this means that medical users of marijuana in Florida have access to a significantly different (and milder) drug than people in Nevada.

The way in which RML is enacted and implemented is more empirically consistent than MML. For instance, in states where marijuana is legal for recreational use, anyone that is 21 years old or older may obtain it from a dispensary, and consume it completely legally (IIHS 2022). But what does it mean that marijuana is legal for recreational use? Simply put, it refers to the non-criminality (i.e., legality) of marijuana use, production, possession, distribution, and sale for recreational purposes. In other words, one does not need any medical reason to obtain and use marijuana legally. Recreational marijuana often entails taxation of the drug. Colorado, for instance, Colorado has a 2.9% tax on the sale of marijuana, and collected more than $420 million in taxes in 2021 (Gray 2022). Furthermore, marijuana is often regulated in a similar manner as alcohol, and entails a governmental regulatory structure (Barry and Glantz 2018, 914; CGA 2022).

2. Theory

2.1 Structure and aim of the chapter

The aim of this section is three-fold. First, to provide an overview of previous literature. Two types of literature are especially considered: 1) literature concerned with explaining policy, both marijuana policy specifically and policy change more generally, and 2) literature concerned with explaining what makes individuals supportive of legalizing marijuana. Second, to describe and explain theories on how public opinion relates to policy change. Third, to discuss the implications that previous literature and theories have for my objective of predicting marijuana legalization.

Theories of why policy changes shed light on what state-level variables may explain/predict recreational marijuana legalization. The second set of theories are also applied to the state-level, since states are my units of analysis. Determinants of individuals' support is applied to the state-level due to the following logic: if people of x demographic are more likely to be for legalizing recreational marijuana, states with a high share of people of x demographic may be more likely to legalize recreational marijuana. For instance, republicans are less likely to be for legalizing marijuana. An abstraction of this would be that states with a high share of republicans may be less likely to have legalized marijuana. Studies explaining attitudes may therefore shed light on what may explain the implementation of the policy in consideration – recreational marijuana legalization.

To summarize, this section a) presents previous literature and findings; b) explains what public opinion is and its role in studying policy change; c) discusses what policy responsiveness is and its implications for this thesis; d) explores the role of medical marijuana in explaining recreational legalization and how and why marijuana policies may diffuse to other states; e) discusses state-level factors that may affect policy and finally; f) a section on the expected predictive ability of individual variables.

A note on causality is warranted before continuing this chapter. Theories and findings of causality are relevant for creating strong predictive models for the following reason: causes should also be predictors. Following this logic, if a theory indicates that a factor may be causal, it is also implied that this factor is a strong predictor. Therefore, theories of causality and explanation are discussed, but these assumptions will not be tested using hypotheses.

2.2 Studying policy change and marijuana policy

This section gives a brief overview of the existing literature on the subject of policy change and marijuana legalization. It furthermore explains which theories will be focused on and how they will be used in light of this thesis' goal. Since the existing literature exploring why and how marijuana becomes legal is limited, I explore relevant literature related to similar types of "morality" polices, such as such as LGBT rights.

A limited scholarship exists on the nature of marijuana legalization and its causes. For example, von Hoffman has studied the case of marijuana legalization in Uruguay, employing two theoretical perspectives when explaining Uruguay's legalization: top-down and bottom-up factors (von Hoffmann 2020). The top-down perspective contending that Uruguay's President Mujica played the biggest role in legalization. The bottom-up perspective, on the other hand, contends that activism played the bigger part (von Hoffmann 2020).

Hannah studied Ohio's medical marijuana legalization (MML) through the policy diffusion and policy learning theories (Hannah 2018). In a similar vein, Mallinson and Hannah study MML through the policy-diffusion and policy-learning perspective in their more recent study of MML in the United States (Mallinson and Hannah 2020). The fact that states choose to oppose federal law when legalizing marijuana is at the analytical centre of their study, a puzzle in itself to be solved: What causes states to defy federal law? This question is, in part, one of the motivations for writing this thesis – despite prediction being the aim. What diffusion is, how it may affect policy, and how it is modelled is therefore discussed in detail later in this chapter.

This thesis' main theoretical focus is public opinion, differing from the previously mentioned studies attempting to explain marijuana legalization. Public opinion has been chosen due to its prominence in the literature concerned with explaining why public policy changes. An example of such research is Lax and Phillips study of LGBT-laws, which finds that public opinion has a strong effect on policy (Lax and Phillips 2009a). There are a myriad of studies showing that public opinion affects policy, based on the idea that policy makers are responsive to what the public prefer (Erikson 1976; Erikson, Wright, and McIver 1993; Burstein 2003; Haider-Markel and Kaufman 2006; Wlezien and Soroka 2021; Bernardi, Bischof, and Wouters 2021). It is therefore reasonable to expect that public opinion serves as a predictor of RML, even though this has not yet been sufficiently explored empirically. I therefore employ public opinion as an input variable for the models attempting to predict which states have, at a given point in time, recreationally legal marijuana. It should be noted that I do not explicitly focus on predicting policy change. Rather, the focus is on predicting whether or not marijuana is legal in a state. Nonetheless, theories of policy change are relevant for predicting whether or not a policy is in place.

This thesis is more encompassing than the existing literature studying recreational marijuana legalization, which has tended to focus on a explaining and describing marijuana policy change and support towards legalization in a few U.S. states, Uruguay, or a combination of both (Hannah 2018; von Hoffmann 2020). This thesis, in contrast, studies all U.S. states from 2010 to 2018, regardless of whether or not they legalized recreational marijuana in this time period.

2.3 Public opinion and responsiveness

From the mid/late-1970s and onward, research has consistently shown public opinion to be related to policy change (Erikson 1976; Page and Shapiro 1983; Lowery, Gray, and Hager 1989; Monroe 1979; Wlezien 1995; Lascher, Hagen, and Rochlin 1996; Burstein 2003; Lax and Phillips 2009a; Burstein 2020). The idea is policy makers responds to public opinion through representative mechanisms. Public opinion has been shown to affect both dichotomously measured policies (e.g., gay marriage laws) and continuously measured policy changes (e.g., defence spending) (Lax and Phillips 2009a; Wlezien 1995). Wlezien quite succinctly captures this relationship through his thermostatic model of public opinion: the public acts as a thermostat, indicating whether the current policy 'temperature' is in line with their preferred policy 'temperature', thus leading policy makers to respond to these preferences (Wlezien 1995). Exactly how public opinion may affect policy change is detailed later in this section.

Particularly relevant to this thesis, public opinion has been studied in relation to marijuana laws. However, to the best of my knowledge, public opinion has not been used as an independent variable to explain or predict recreational marijuana legalization/the legal status of marijuana. The focus is generally split between describing trends in support towards legalization and explaining the determinants of support towards legalization (Denham 2019; Felson, Adamczyk, and Thomas 2019). For instance, Cruz, Queirolo and Boidi (2016) have examined the determinants of public support towards marijuana legalization in Uruguay, the United States and El Salvador. The findings and implications of this literature is discussed further in the subsequent sections.

2.3.1 What is public opinion?

Public opinion, in a political science perspective, generally refers to political attitudes citizens have towards existing or hypothetical public policy, political parties or ways of governing (Berinsky 2017, 310). There are multiple purposes of measuring and collecting public opinion data. The most obvious purpose is description. Measuring what citizens think about political issues, policies, or political systems has value in and of itself as it allows scholars, politicians, journalists, and the public to learn about how others perceive the world. Knowing what people think may often be the beginning to uncovering and subsequently solving puzzles. It may inspire one to ask, and subsequently answer: 'Why do people feel this way?' This is particularly true for political scientists and sociologists. This seems to be the motivation for previous literature in the field of marijuana legalization – exploring what people think of marijuana legalization and trying to explain individuals' attitudes (Cruz, Queirolo, and Boidi 2016; Denham 2019; Felson, Adamczyk, and Thomas 2019). Moving beyond the descriptive or inductive uses of public opinion, public opinion may furthermore be used to explain/predict phenomena – like policy. This is how public opinion is used in this thesis: to predict the legal status marijuana.

In the United States, there have been conducted numerous public opinion polls on people's attitudes toward marijuana legislation – however only at the national level or with few observations. This thesis uses General Social Survey data disaggregated/simulated to the U.S. census division-level, in an attempt to predict marijuana legalization. Details of disaggregation and data is further explained in the Data and Methods chapter.

2.3.2 Public opinion on marijuana legalization

This section details how support towards marijuana legalization has been measured in previous studies, trends of support towards legalization, and what is thought to explain support towards marijuana legalization at the individual level. How support is operationalized and what data scholars have chosen is discussed prior to their findings. This is because the way in which the studies measure support is critical to keep in mind when analysing their findings, and what their results and choices of data imply for this thesis. The purpose of looking at previous studies is to choose an appropriate dataset and measure of support towards RML.

Looking at marijuana public opinion research gives an indication of the levels of support, and what may cause support. This is particularly important for the disaggregation of public opinion data using multilevel regression with post-stratification.

In short, the existing literature measures and conceptualizes marijuana legalization in a similar manner, shows an increase in support towards marijuana legalization the past decades, and finds similar determinants of support. MacCoun and Kilmer describe the trend differently than most scholars – they argue that we are essentially "seeing more of a shrinking of opposition to legalization than a growth in enthusiasm" (Kilmer and MacCoun 2017, 8). In line with most of the literature, this thesis uses data from the General Social Survey (GSS) due to its operational and conceptual relevance, as well as the data availability. This choice is elaborated in this section as well as in the data and methods chapter.

Operationalizations in previous literature

In their descriptive and explanatory study of public opinion towards marijuana legalization, Felson, Adamczyk and Thomas use the General Social Survey (GSS) item "Do you think the use of cannabis should be made legal or not?" with three possible answer options: yes, no, and don't know (2019, 16). They furthermore employ Gallup polling data to track support for legalizing marijuana use. The Gallup question was quite similarly formulated as the GSS question, the answers also being ordinally measured: support or do not support legalizing marijuana (Newport 2011; Felson, Adamczyk, and Thomas 2019, 13). Felson and colleagues furthermore used a National Survey on Drug Use and Health (NSDUH) variable asking respondents whether they approve of "adults trying cannabis or hashish once or twice" with three possible answers: strongly disapprove, somewhat disapprove, or neither approve nor disapprove (2019, 17).

Stringer and Maggard, in their study of the relationship between media exposure and attitudes towards marijuana, also used GSS' marijuana-legalization variable (Stringer and Maggard 2016, 433). GSS' variable is a quite valid measure of support towards recreational marijuana legalization. However, the variable lacks specificity: The question does not specify in what way cannabis use should be made legal. The question may be interpreted by the respondents (and thus researcher) as referring to legalizing medical or recreational marijuana, or even decriminalization in the minds of the respondents, confusing these marijuana policy models. NSDUH's variable is also problematic, as it mentions the legal status of marijuana and may, at best, be considered a proxy for support of marijuana legalization. It is not unlikely that an individual may disapprove of marijuana use, yet at the same time be inclined to support legalization for libertarian reasons. For this reason, the NSDUH variable is inappropriate for measuring the public's stance towards legalizing marijuana and will not be used in this thesis.

Subbaraman and Kerr used a variable similar to the GSS variable in their study of Washington state, using original public opinion data (2017, 206). They asked: "Do you think marijuana should be legal for adults?", and: "Do you think adults should be able to grow their own marijuana for personal use?". This approach suffers the same validity constraints as Felson and colleagues' approach. However, combining the former (legality) and latter (home-growing) questions may result in a more detailed measure of public opinion towards legalization: the main measure being the question of legality, yet the degree to which the public are liberal in their views on marijuana may be further distinguished through analysing their views towards home-growing. Finally, Cruz, Quierolo and Boidi have written a series of papers on attitudes towards marijuana legalization in the U.S., Uruguay, and El Salvador (Cruz, Queirolo, and Boidi 2016; Cruz, Boidi, and Queirolo 2018a, 2018b). Their articles mainly study what causes support for *existing* legalization schemes and models. For instance, whether marijuana should be sold at drug stores and the amount of marijuana one can legally grow for personal consumption. Empirically they focus on to Uruguay, El Salvador and Washington State (Cruz, Queirolo, and Boidi 2016, 312).

The data that has been used for measuring attitudes towards marijuana in previous studies is not particularly valid for measuring support towards recreational legalization. This is because none of the survey data specifies whether respondents are for or against the legalization of *recreational* marijuana. Rather, the questions used in previous studies reflect general sentiments towards the legality of consuming marijuana or whether respondents approve of adults trying marijuana. Where legality is mentioned, for instance the General Social Survey, recreational marijuana is not explicitly specified. This makes the data difficult to interpret, as respondents may associate legality with either decriminalization, medicalization, or recreational legalization. This lack of specificity may not be as important when considering how legislators view such polls. Policy makers likely consider data showing support towards making marijuana legal, such as GSS', as an indication of support towards the policy of recreational legalization. This is important as that is the basis of the public opinion-policy change relationship: how policy makers perceive attitudes. Despite potential issues, I nonetheless consider GSS' operationalization as adequate for this thesis as it touches upon the legality of marijuana, and not just whether individuals approve of others consuming it. For these reasons, and the fact that the GSS is simply the most valid operationalization of support towards RML that is publicly available and nationally representative, the GSS' support towards legalization variable (*grass*) is used in this thesis.

Findings in previous literature – determinants of support

The survey literature unanimously finds an increase in support for marijuana legalization over time, using the aforementioned measures of public opinion as operationalizations of support. Felson and colleagues, for instance, find a substantial increase in support towards legalization in the US. More precisely, they found that support has increased by a 100%, from roughly 20-25% of the population supporting legalization in the 1980s and 1990s, to more than 50% in the early/mid 2010s (Felson, Adamczyk, and Thomas 2019, 23). They found differences in support by age, education, a lack of religious affiliation and political affiliation (democrats being more supportive). In general, the increases in support were insignificantly different by socio-demographic subgroups. The changes in public opinion were also similar across different regions within the U.S. Attitudes did not change in states that had undergone marijuana policy reform (nor states neighbouring liberalizing states) (Felson, Adamczyk, and Thomas 2019, 23-25). It is of utmost importance to point out that Felson, Adamczyk and Thomas used the question "do you approve of adults trying cannabis or hashish once or twice?", considering the operational problems of using this to measure support for reform.

Subbaraman and Kerr, in contrast to Felson and colleagues, found an increase in support for legal recreational marijuana in Washington state four years following legalization (Subbaraman and Kerr 2017, 207). If public opinion predicts legal status, it may be that it is due to legal status affecting public opinion. Tese authors also found a positive relationship between support for legalization and prevalence of marijuana use, though the effect was small (Felson, Adamczyk, and Thomas 2019, 23-25). Perhaps the most significant finding as a determinant of public opinion was the positive effect of exposure to media on support towards legalization (Felson, Adamczyk, and Thomas 2019, 24-25).

Stringer and Maggard, using GSS data (like Felson and colleagues), also found an increase in support for marijuana throughout the past decades (Stringer and Maggard 2016). They furthermore found a statistically significant and positive relationship between media exposure and support for legalization. However, this relationship was only statistically significant after 1990 (Stringer and Maggard 2016, 438). They, amongst other scholars, explain this as being due to the increasing positive coverage of marijuana in the 1990s compared to the 1980s (Stringer and Maggard 2016, 439). Again, in alignment with Felson and colleagues, they found that higher education, later year of birth and non-religion was associated with more support towards legalization (Stringer and Maggard 2016, 439-440).

Cruz, Boidi and Queirolo found that amongst respondents who have tried marijuana, younger respondents and males are more supportive of marijuana legalization (Cruz, Boidi, and Queirolo 2018b, 431). Furthermore, the same authors, in an earlier article, found a positive relationship between education, being left-leaning politically and support for legalization in Uruguay, but the latter relationship is not as prevalent in the U.S. (Cruz, Queirolo, and Boidi 2016, 318-319). They again found a relationship between use and support in Uruguay in another article from 2018. Interestingly, personal marijuana use history also served to significantly reduce the effect of Catholicism and evangelical affiliations (Cruz, Boidi, and Queirolo 2018a, 72-73).

The literature also shows that religious people are more likely to oppose legalization and consumption of marijuana (Stringer and Maggard 2016, 439-440; Felson, Adamczyk, and Thomas 2019, 23-25). The exact mechanisms for why are unknown. One reason could be that religious people are more culturally conservative. In contrast, higher educated people were more likely to be supportive of legalization and positive to the notion of consuming marijuana (Stringer and Maggard 2016, 439-440; Felson, Adamczyk, and Thomas 2019, 23-25; Cruz, Queirolo, and Boidi 2016, 318-319). It is difficult to say exactly why higher education is associated with support towards legalizing marijuana. Stringer and Maggard suggested that it may be that higher educated people have greater knowledge of the reality of marijuana and its potential for non-abusive use (Stringer and Maggard 2016, 439). Political affiliation, consistently predicting support in the studies mentioned, may be tied to the degree to which respondents are culturally progressive or conservative – left-leaning individuals and democrats being more likely to be in favour of legalization than right-leaning individuals and republicans. Other individual-level determinants of support towards legalization include gender, age, and media exposure.

Explaining why public opinion has shifted, apart from individual-level determinants, is challenging. I suggest it is simply due to the ‘Reefer madness’ generation having less political power, and due to the apparent failure of the ‘war on drugs’. Additionally, it is possible to argue that people may be becoming more exposed to the growing body of evidence that marijuana can be used safely/medically.

These findings serve as useful theoretical and empirical starting points for choosing input variables for the data-disaggregation algorithm, as well as input variables for the predictive models. Demographic variables such as share of republicans, median age, proportion of the population with a high education, and religiosity – all at the state-level – will thus be used for both the predictive and the disaggregation models.

2.3.3 Public opinion, responsiveness, and congruence

Having conceptualized and chosen an operationalization of public support towards marijuana legalization, this section discusses policy responsiveness and opinion-policy congruence. Both concepts are important since they form the basis for why public opinion is thought to predict policy. This section includes, amongst others, sub-headings for factors thought to impact the extent to which policy is responsive to public opinion. For instance, the sub-heading 'morality politics'.

What is responsiveness and policy congruence?

The main mechanism theorized to connect public opinion and policy is political responsiveness (Manza and Cook 2002, 633). This theory rests on the argument that political elites "benefit from pursuing policies that are (or appear to be) in accord with the wishes of citizens" (2002, 633). Responsiveness, as the term suggests, refers to the degree to which political elites act in accordance with the publics and electorate's preferences. High responsiveness means that "governments act in the interests of their citizens so that there is congruence between what citizens want and what their governments do" (Klüver and Pickup 2019, 91). Opinion-policy congruence is therefore understood as the extent to which public policy is in line with the public's policy preferences. It is simply a description of what 'the people' want, and what the government does. Whilst congruence describes the agreement between public opinion and public policy, responsiveness concerns how, and the extent to which, public policy is moved by public opinion. However, there is some disagreement as to how, when and whether public opinion affects policy change. This is explored later in this subchapter.

As previously emphasized, Wlezien, employs an analogy of a thermostat to illustrate how responsiveness works. He attempts to explain the changes in defence spending in the U.S. using public opinion data (Wlezien 1995). He describes the feedback between the public opinion and public policy as thermostatic – policy makers react to the 'thermostat' that is public opinion. In the case of defence spending, he argues that when the public desire more defence spending than is currently the case, the public will respond, signalling that they desire more spending. If government then increases public spending sufficiently, the relationship is in balance. If, for instance, government increases public spending more than the median preference, the ‘thermostat’ will keep blinking, signalling that the favoured policy “temperature” is not reached (Wlezien 1995, 982). The core idea is that policy makers respond to the public, and vice versa. There is a feedback loop amongst voters and politicians that hinges on the availability of information – the public being informed about policy, and policy makers being informed about the public’s preferences (Wlezien 1995, 993).

When is policy responsive?

Erikson, Wright and McIver, in their seminal work ‘Statehouse Democracy’, discuss the possibility of reverse causality, where public policy affects public opinion, as opposed to the inverse (Erikson, Wright, and McIver 1993, 88-89). However, they concluded that public opinion likely affects policy, as opposed to there being an instance reverse causality (Erikson, Wright, and McIver 1993, 88-89). In their article on gay rights and pro-gay legislation, Lax and Phillips similarly dismissed the possibility of reverse causality being present (Lax and Phillips 2009a, 382).

Mayhew argues that the dominance of certain parties and incumbents in some state and national-level elections reduces opinion-policy congruence (Mayhew 1974; Haider-Markel and Kaufman 2006, 165). Considering the mechanism by which public opinion converges with policy is elite incentives, this theory makes sense: the largest incentives for elites to act on public opinion and preferences, re-election, is diminished in states where the incumbent’s states party dominates. Dye also supports the idea of responsiveness, highlighting citizens’ lack of knowledge, interest and consistency in their opinions as a factor decreasing public opinion’s effect on policy (Dye 1984, 326).

Morality politics

Whether a policy is a form of morality policies or not is thought to affect its responsiveness to public opinion. Morality policies are “those which seek to regulate social norms or which evoke strong moral responses from citizens for some other reason” or “instigate debate over first principles, resulting in uncompromising clashes of values” (Mooney and Lee 1995, 600). Morality policies are often contrasted to purely economic policies (Johns 2015, 195). Furthermore, morality policies are often tied to fundamental beliefs, such as religious or ideological affiliations. The concept of morality policies is by some seen as a continuum. On one end of the spectrum are policies that are purely moral, where both sides argue for/against the policy on moral grounds – for instance abortion. On the other end of the continuum, there are policies where one side employs economic arguments and the other side moral arguments (e.g., gambling). Johns places drug-policy in the middle of this continuum, a so-called hybrid, where both sides argue on both moral and economic premises (Johns 2015, 196).

Haider-Markel and Kaufman argue that morality policies/policies that evoke strong feelings of right and wrong are more likely to be reflective of public opinion because of their salience and importance (Haider-Markel and Kaufman 2006, 165-166). This theory is empirically supported for gay-rights policies amongst other policies (Tatalovich and Daynes 1988; Carmines and Stimson 1989; Haider-Markel and Kaufman 2006; Lax and Phillips 2009a).

Others argue that because people have strong and stable opinions on morality issues, they are more likely to become active in interest groups and influence policies through these interest groups (Norrander and Wilcox 1999, 707). Public opinion on morality policies may be thought to impact public policy more due to the importance held by the public of such issues. Morality policies can therefore invoke higher responsiveness due to the impact enacting such a policy may have on whether or not people vote to re-elect the reigning candidate or political party.

This begs the question of whether legalizing recreational marijuana should be thought of as a morality policy. I argue that RML fits into this category, mainly because the prohibition of marijuana regulates social norms by virtue of one being a criminal if one consumes it. It has also been classified as a morality policy due to it involving "conflict over first principles", lacking "amenability to compromise" and having "little technical complexity" in terms of implementation compared to non-morality politics (Ferraiolo 2014, 347). Because RML can be considered a morality issue, the morality aspect of responsiveness is therefore one of the major theoretical backgrounds for why marijuana policy is expected to be impacted by public opinion.

Issue salience

Issue salience is thought to regulate the extent to which government is responsive to public opinion (Monroe 1998; Manza and Cook 2002; Lax and Phillips 2009a; Klüver and Pickup 2019; Burstein 2020). In contrast to morality politics, issue salience is more dynamic and less static. Issue salience generally refers to how important an issue is for people, quite similar to the issue of morality. According to this view, governmental actors are more likely to act in accordance with public opinion if the issue has high salience as they are more likely to, in the first place, know how the electorate feels. In terms of (re)election strategies, it is not wise to go against popular opinion on an issue that is important to the voters (Lax and Phillips 2009a, 370).

Monroe (1998) studies the congruence between policy and public opinion, finding American political institutions to be rigid and conservative – resisting change. He further describes that issue salience decreased this bias (Monroe 1998, 6). Burstein even goes as far as saying that most social scientists agree that the more salient an issue is to the public, the more likely it is that government will be responsive (Burstein 2003, 29). Burstein also argues salience to be important in terms of elections: "Citizens who care about an issue are especially likely to take elected officials' actions on that issue into account on election day" (Burstein 2003, 30). The implication here is that elected officials are aware of this, and the extent to which an issue is salient.

Regarding my thesis, I argue marijuana legalization and drug reform is a salient issue, particularly from the late 1990s when California legalized medical marijuana. Furthermore, I argue that marijuana legalization became increasingly salient after Colorado and Washington legalized recreational marijuana in 2012. It is also considered salient as it is a form of morality policy. Norrander and Wilcox, for instance, argued that morality politics are inherently salient (Norrander and Wilcox 1999, 708). Considering RML's salience, public opinion is expected to have an effect on recreational marijuana legalization. And, therefore, serve as a strong predictor. However, salience will not be measured in this thesis. Rather, RML is considered salient in the period that is studied (2010-2018). Salience is sometimes measured by looking at how often an issue is discussed in the media, or simply by conducting surveys asking to what extent an issue is important to the respondents (Epstein and Segal 2000). No public data operationalizing marijuana salience exist, either surveys or media-analyses. Creating a measure of salience as a mediator of how well public opinion should predict policy, or as a predictor in and of itself, is beyond the scope this thesis.

Interest groups

Interest groups are also thought to condition the effect public opinion has on policy change. Lax and Phillips, for instance, found that the existence of religious conservative interest groups increases conservative LGBT opinion-policy congruence, as well as decreasing liberal LGBT opinion-policy congruence (Lax and Phillips 2009a, 380). In other words, they found powerful interest groups to increase congruence when the majority is of the same opinion, as well as a decrease in congruence when the majority is of a differing opinion. Page and Shapiro make a similar point to what Lax and Phillips made, referring to scholars such as Schattschneider (1960) and McConnell (1966). They point out that powerful interest groups may obstruct opinion-policy congruence as the American democratic system favours well-organized pressure-group politics, as opposed to individual's preferences (Page and Shapiro 1983, 175). There is, of course, the possibility that public opinion affects policy *through* interest groups – making interest groups the mechanism in the opinion-policy relationship (Page and Shapiro 1983, 186). Hansen, for instance, has also made this argument. He argued that interest groups enhance responsiveness through providing legislators with information of the public's preferences (Burstein 2003, 31). Interest groups fighting for legalizing marijuana may, based on this literature, have a positive effect on the legalization of marijuana not only directly due to their influence, but also as an interaction effect with public opinion leading to an increased opinion-policy congruence.

There are many interest groups in the U.S. attempting to impact marijuana policy on both sides of the legalization question. For instance the Drug Free America Foundation (DFAF) on the one side of the issue, and National Organization for the Reform of Marijuana Laws (NORML) on the other (DFAF 2022; NORML 2022). Measuring interest group activities on both sides of the legalization question would be a useful predictor-variable and mediator of the predictive ability of public opinion. However, creating such a measure is too encompassing for this thesis. It would be achievable if this thesis' focus was on a handful of U.S. states. Due to the focus being on all U.S. states over a significant span of time, creating such a measure is impracticable.

Ballot initiatives

A ballot initiative, also known as popular initiative, is defined as "a means by which citizens may propose to create, amend, or repeal a state law or constitutional provision through collecting petition signatures from a certain minimum number of registered voters" (Ballotpedia 2022). Empirically, the presence of ballot initiatives is strongly related to marijuana policy changes – both medical and recreational legalization. In their seminal article *Gun Behind the Door?*', Lascher, Hagen and Rochlin examine this relationship, asking: '*Are states that have ballot initiatives more responsive to public opinion?'*. The availability of ballot initiatives may serve, as the title of their article suggests, as a "gun behind the door": citizens may resort to ballot initiatives if their legislators are not acting on their behalf (Lascher, Hagen, and Rochlin 1996, 760). The more direct way in which ballot initiatives are thought to affect public policy is that it makes citizens participate more in politics, giving them some policy-making power (Lascher, Hagen, and Rochlin 1996, 760). This is the main mechanism linking public opinion and ballot initiatives to policy change. Lascher and colleagues nonetheless concluded that ballot initiatives do not make governments more responsive (Lascher, Hagen, and Rochlin 1996).

There is an empirical link between ballot initiatives and marijuana policy change: seven of the eight states that legalized medical marijuana from 1996 to 2000 did so through ballot initiatives (Kilmer and MacCoun 2017, 194). The pattern is even more clear for recreational marijuana legalization: 13 of the 19 states that have legalized recreational marijuana have done so starting with ballot initiatives, as of 2021 (Ballotpedia 2021). Arcenaux has also focused on referenda, public opinion, and policy change. In contrast to Lascher and colleagues, Arceneaux found an interactive effect between public opinion and allowing ballot initiatives on abortion policy (Arceneaux 2002, 383). The reason why legalization has often started with a ballot initiative may be that marijuana policy generally has had a low congruence with what the public believes. To elaborate, public opinion has long been in favour of legalizing marijuana, yet policy has been lagging (Felson, Adamczyk, and Thomas 2019, 23). Such a democratic institution as ballot initiatives allows the public to correct for such a lack of congruence by mobilizing. For these reasons, a dummy variable indicating whether citizens can propose new legislations through the ballot initiative process is included in the predictive models.

Concluding this section, the presence of interest groups working for legalizing marijuana, the issue salience, and its relation to morality politics indicates that public opinion should have an effect on legalizing recreational marijuana. However, only the availability of ballot initiatives is included in the predictive models.

2.4 Medical marijuana – a trojan horse?

The legalization of medical marijuana is by some deemed as critical for legalizing recreational marijuana. Mosher and Atkins, for instance, argued that medical marijuana “paved the way for the legalization of recreational marijuana” (Mosher and Atkins 2019, 15). They furthermore point to Balko, a Washington Post journalist, who referred to medical marijuana legalization as “gateway legislation” (Mosher and Atkins 2019, 15). Similarly, General Barry McCaffrey, Bill Clinton’s Director of the White House Office of National Drug Control Policy, stated that “medical marijuana is a stalking-horse for legalization […] This is about legalizing dangerous drugs” (Kilmer and MacCoun 2017, 5). This trojan horse argument, often used against legalizing medical marijuana, has featured somewhat prominently in the public debate. The term ‘trojan horse’ comes from a fictional tale in *The Odyssey*, and is used as a saying to describe when “someone is duped into letting an enemy past their defences” (Kilmer and MacCoun 2017, 5). In the case of medical marijuana legalization, opponents using the argument are essentially saying MML is a ploy to legalize marijuana under the guise of compassion. It is essentially a ‘slippery slope’ argument – first one legalizes medical marijuana, then the goal post will move, and recreational marijuana will be legalized. This argument is in many ways one of internal policy diffusion – one policy leads to another. The ‘trojan horse’ also considers medical legalization as a de-facto legalization of recreational marijuana. Meaning that, in practice, MML *is* RML due to doctors over-prescribing medical marijuana to non-needing people (Bonnie 2018, 588).

Due to its presence in the public debate, and Kilmer and MacCoun highlighting a relationship between MML and RML, MML is used as an input variable in this thesis’ analyses. The expectation is that MML serves as a strong predictor of RML, because all states that have legalized marijuana for recreational use have previously legalized medical marijuana. However, I will use different operationalizations of MML: a dichotomous variable indicating whether medical marijuana is legal and a variable measuring years since the given state legalized medical marijuana. Considering that all states that have legalized recreationally have legalized medically, MML seems to be a prerequisite for RML – at least empirically speaking. Thus, it can be expected to be a very strong predictor. If so, it is good in the sense that the models will have a high accuracy if they are included. However, MML overshadowing other variables is not as interesting in terms of theory, as it is in many ways obvious, and does not really help one to explore what is unique about states that have legalized.

It must be noted that the trojan horse theory is not considered in its entirety as this is inappropriate in a quantitative analysis. Trojan horses concern deception and conspiracy. In this case the assumed intentions of moving the goal post to recreational legalization once medical marijuana has been legalized or de-facto legalizing marijuana for recreational purposes by legalizing medical marijuana. This is quite speculative, and a qualitative design is needed for discerning intentions. Whether MML is a trojan horse or not is not fully relevant for this thesis – it may still prove to be a predictor variable.

2.5 Policy diffusion and policy learning

Policy diffusion is the process by which governments emulate each other’s policies (Grossback, Nicholson-Crotty, and Peterson 2004, 521). The mechanism in diffusion is policy learning: governments observe each other, learn from each other and subsequently mimic each other

(Shipan and Volden 2012, 790). States may mimic each other in competing for employers, skilled labour, or other groups of people that find a certain policy to be desirable. This is called 'Tiebout sorting' (Bradford and Bradford 2017, 77). Diffusion is generally two-fold: geographical and ideological. Diffusion is often thought of as geographical: neighbouring municipalities, be it states or cities, look to each other and learn from one another. Ideologically, states may look not only to their neighbours, but ideologically similar states (Hannah and Mallinson 2018, 411; Grossback, Nicholson-Crotty, and Peterson 2004, 526).

The view that states and cities compete against each other economically plays into how diffusion may occur. Particularly, diffusion may occur if economic gains are expected from neighbouring states. Or, states may mimic a neighbouring state's policy if the neighbour's enactment of a policy has led to economic spill over to that state (Shipan and Volden 2008, 842). Both empirically and theoretically relevant, Hannah and Mallinson found that marijuana policy may diffuse, both ideologically and geographically (Hannah and Mallinson 2018). Hannah and Mallinson study the diffusion of medical, rather than recreational, marijuana legalization. Their study still provides a useful framework as their focus is by and large the same: what causes marijuana policy change? Thus, this thesis uses geographical diffusion as an input variable for the predictive models. Geographical diffusion is measured as the proportion of neighbouring states that have legalized recreational marijuana.

Bradford and Bradford have also studied the diffusion of medical marijuana legalization. They built a model capturing the different elements and mechanisms of policy diffusion: ideological and geographical diffusion, motivation effects and resources and obstacles (Bradford and Bradford 2017, 78). Motivation effects include factors that are likely to motivate citizens and legislators to pursue a policy. In the case of medical marijuana legalization, the following variables have been used in previous studies to measure motivation for implementing MML: state-citizen ideology and number of AIDS cases per 100,000 residents (Bradford and Bradford 2017, 78). State-citizen ideology is the median ideological stance of the state's citizenry. Diffusion is thought to be more likely when the motivation for implementing a policy is higher.

What resources and obstacles a given state has for implementing certain policies is also thought to moderate the effect of diffusion. Resources and obstacles are largely economic factors, those determining for instance the opportunity for lobbying, or state fiscal health – whether the state has the resources to enact and implement a policy (Bradford and Bradford 2017, 78). These are all important aspects to consider when studying the diffusion of policy. In short, Bradford and

Bradford (2017, 80.82) find ideological and geographical diffusion to be empirically supported for MML, motivation effects to be minute, and resource effects to be modest. As part of the resource effects, they find unemployment to increase the likelihood of adopting MML, each percentage point in unemployment increasing the likelihood by +1.2% (Bradford and Bradford 2017, 82). These findings serve as useful starting points for building predictive models of recreational marijuana legalization.

2.6 State level determinants of marijuana legalization

This section includes discussions of what variables and factors may increase the probability of a state having legalized recreational marijuana. The purpose of this section is to gain insight into what input-variables may be useful both for the predictive models and the disaggregation methods. It therefore features discussions of variables and concepts that have been highlighted previously in this chapter, as well as variables and factors that have not been discussed.

In their study *Defiant innovation*, Hannah and Mallinson attempted to explain why some U.S. states deviate from federal law by legalizing medical marijuana (Hannah and Mallinson 2018). Being quite similar to this thesis' design, their study offers useful insights on variables that may cause states to defy federal laws and policies. One of which is the states ideology. They argued that states are more likely to adopt defiant policies, in this case MML, if they are ideologically similar to states that have previously adopted said policy (Hannah and Mallinson 2018, 411). This is understood as ideological, as opposed to the more traditional geographical, policy diffusion. The idea is quite simple: ideologically similar states (either citizens or the government) enact similar policies. Empirically, more democrat states have legalized marijuana than republican states, both medically and recreationally.

The share of each state's population's electorate that voted for the republican candidate in the last presidential election is therefore used as an input variable in both the disaggregation and prediction models. This input variable is both empirically and theoretically grounded. It is empirically grounded as, as mentioned, most states that have legalized were majority democrat at the time. The theoretical foundation is tied to the determinants of support towards RML at the individual level. As discussed in an earlier section, democrats and left-wing people are more likely to be in favour of RML, likely due to the cultural progressiveness associated with the left. The share of votes to the republican candidate reflects, in short, both ideological diffusion and the idea that being left leaning may cause one to be in favour of legalizing recreational marijuana.

The availability of ballot initiatives is likely significant as well. As discussed earlier in this chapter, a large majority of states that have legalized marijuana – both recreationally and medically have done so through the ballot initiative. Despite disagreements on whether ballot initiatives increase responsiveness, this variable is likely important in the prediction of marijuana legalization considering that there is both an empirical and theoretical link between legalization and availability of ballot initiatives (Ballotpedia 2021).

Economic factors may determine policy outcomes. For instance, GDP per capita and unemployment rate. This is in part related to the 'resource and obstacles' aspect of diffusion theory. These factors also give the models breadth in terms of what kind of variables are considered – in this case economic variables. Policy may also be impacted by a state's demographics. This is mostly based on the idea that certain policies are more/less favoured by demographics. For instance, as discussed in section 2.3.2, higher educated people are more likely to be in favour of legalization. Therefore, demographic state-level variables such as median age, share of population with a high education, and race demographics may impact policy outcomes.

2.7 Expectations

This section acts as a substitute to a hypothesis-section, as hypotheses are not appropriate for predictive models. This section is dedicated to discussing expectations of which input variables prove to be the most important for predicting RML, as well as the general accuracy of the predictions.

The primary expectation concerns public opinion as a predictor of legalizing recreational marijuana. It is expected that public opinion towards marijuana legalization is a strong predictor of legalizing recreational marijuana. Public opinion shows promise as a causal variable for marijuana legalization, both based on theoretical discussions in this chapter and the empirical findings of other studies. Extending the assumption of public opinion being a causal variable to prediction, it is fair to presume it will fare well as a predictor due to the simple logic that causal variables must be able to predict. Combined with the fact that RML is, as discussed, a kind of morality policy, and a salient issue, there is reason to believe a high degree of responsiveness to public opinion. A few important things regarding this expectation should be noted. Firstly, there is no empirical basis specific to marijuana policy that warrants this expectation. The expectation is, as mentioned, based on theoretical discussions and the effect of public opinion on other types of policy. Despite causality and opinion-policy congruence being aspects of this expectation, I am not directly concerned with the effect of public opinion on marijuana legalization as this thesis' methodological design and data foundation is not suited for such a query. The aim is not to uncover whether public opinion affects RML. The aim is to predict RML as well as possible. This is discussed in detail in chapter 3.

The sub-optimal disaggregation of the public opinion variable likely decreases its potential as a predictor. However, the degree to which the measurement-level of public opinion (census division as opposed to state-level) negatively affects public opinion as a predictor is difficult to foresee. Despite these issues, public opinion is still expected to have a strong predictive ability.

But what is meant by 'strong predictive ability'? This is difficult to operationalize. The strength of a predictor in predictive models is on the one hand comparative. In this regard, public opinion is expected to be one of the best predictors in the models comparatively speaking. This may be measured by running alternative models without public opinion data and comparing how these fare compared to those with the public opinion data. However, this is not necessarily an appropriate way of measuring variable-importance in machine learning models, which is to be detailed in the next chapter.

Furthermore, total number of years a state has had medically legal marijuana is expected to be a stronger predictor of RML than any other variable. This is due to the empirical observation that all states that have legalized recreational marijuana have legalized medical marijuana prior to the enactment of RML. Despite being tied to the MML trojan horse theory, the expectation of years-since-MML being a strong predictor is mostly empirical. The models obviously do not pinpoint causal mechanisms of each variable. As such, MML does not have to have a theoretically stronger basis for expecting causality in order to be a more efficient predictor than public opinion.

The availability of ballot initiatives and share of the electorate that voted for the Republican presidential candidate are also expected to be strong predictors of RML. This expectation is in some ways based on a middle-way between the expectation of public opinion and MML to be strong predictors: empirical observations *and* theory. Empirically, most states that have legalized recreational marijuana have done so through ballot initiatives, in addition to being democrat-majority/left-leaning states. This empirically based expectation is also theoretically grounded, for instance the cultural progressiveness associated with left-leaning ideologies and the 'back-door' aspect of ballot initiatives, both concepts discussed in previous sections.

Concerning the models more generally, I expect them to be able to predict whether a state has legal marijuana or not quite accurately. Quantifying this expectation by giving a percentage of expected accuracy is difficult. However, considering this thesis uses a wide range of prediction models, and includes a breadth of theoretically and empirically substantiated input variables, there is good reason to believe that the best prediction model is quite accurate.

3. Data and Methods

The purpose of this chapter is to explain and give an overview of the data and methods that are used in this thesis. In short, the data includes self-collected data and data from the United States Census Bureau (USCB), United States Bureau of Labor Statistics (BLS), Ballotpedia and more. The methods used in this thesis are split into two categories: Multilevel regression with post-stratification (MrP) and interpretive predictive machine learning models. MrP is used to disaggregate nationally representative public opinion data to the USCB-defined division level. The prediction methods are used to predict whether or not marijuana is legal for recreational purposes in a state. Having two categories of methods complicates the structure of this chapter, warranting a thorough explanation of how and why this chapter is structured as it is. In short, this chapter is structured based on these two categories of methods.

The first section features a discussion of why machine learning prediction models, as opposed to regression models, have been chosen. The second part (3.2) concerns how marijuana laws are measured and is relevant for both categories of methods as marijuana policy variables are used for both the disaggregation/simulation method, and for the prediction methods. Then follows a section dedicated to discussing the details of MrP and autoMrP, the data used for this disaggregation method, and the results of this method. The results from MrP are, as mentioned, used as input in the prediction models. Since the prediction models build on and use the results from MrP, MrP is discussed prior to the prediction models.

The MrP section is followed by the prediction methods section. This part of the chapter is structured in the following manner: a detailed discussion of the input variables and how they are measured, followed by a discussion of each type of machine-learning model that is used in this thesis. The following predictive machine learning algorithms are used: Shrinkage methods (ridge, lasso and ElasticNet), tree-based methods (single decision trees, random forest, and gradient boosting machines) as well as support vector machines (SVM). These models essentially tell us what the differences are between a state that has legal marijuana and a state that does not have legal marijuana through classification. Finally, as part of the prediction-methods section, a discussion of how the models is tuned and built is included.

3.1 Why prediction and not explanation/regression?

Why should one create predictive models as opposed to explanatory regression models? There are multiple reasons why I have chosen prediction models, both reasons specific to this thesis and its data, and more general reasons.

One specific reason relates to data availability. Since public opinion is simulated to the divisional level, rather than to the state-level, attempting to use this variable as an explanatory variable is problematic mainly for two reasons. First, because using it to explain marijuana legality supposes public opinion to be heterogeneous between divisions, and homogenous within divisions. In other words, that public opinion varies across divisions, and is similar within divisions. Using this variable in a regression analysis is inappropriate simply due to its lack of validity as a measure of state-level public opinion. One may argue that this variable therefore should have been removed altogether. However, it has been kept because previous literature suggests a strong link between public opinion and public policy. To understand marijuana legislation, it is imperative to explore public opinion. In order to minimize validity concerns, a prediction-approach has been deemed more appropriate than a regression approach. One of the reasons why prediction has been chosen as opposed to explanation/regression, is thus that it allows the theoretical focus to remain on public opinion's relationship with recreational marijuana legality. A more 'careful' approach (prediction) is more appropriate, however this comes with the cost of needing to be more careful when discussing the relationships between input variables and the output variable. However, prediction may tell us if there is anything there worth studying further with state-level data and regression analyses.

Furthermore, multiple variables[1] that are used for the simulation of disaggregated public-opinion data (MrP) to the division-level are also used as input variables in the prediction models. Using *grass* (public opinion towards marijuana legalization) and the variables that have been used to create *grass* in the same model as explanatory variables poses the risk of *grass* being correlated with the other input variables. The correlation plot A.1 in the appendix shows exactly this: *grass* and *unemp*, for instance, are correlated. One could argue that unemployment could have been removed, and a logistic regression with *grass* could have been ran. However, multiple other variables are correlated with each other to an extent that makes it inappropriate

[1] Such as unemployment, share of population that voted for the republican presidential candidate in the last election, and state religiosity index

to include them in the same regression model. For instance, GDP per capita and the share of the population that have a high education (*heduc*) are highly correlated. Furthermore, unemployment (*unemp*) and the share of population with a low education (*leduc*) also have a correlation coefficient of more than 0.5. Machine learning models, on the other hand, deal with multicollinearity in a better way than regression models. For instance, through variable selection (shrinkage methods) (Chan et al. 2022, 2). The consequence of multicollinearity in regression models is the decreased validity of coefficient estimates (Chan et al. 2022, 2). This is not a problem with prediction models as coefficients are not the main focus. Rather, accuracy of predictions, variable importance/decrease in accuracy when a variable is omitted is focused on, and interaction plots are at the analytical centre.

Furthermore, the logistic regression models that have been ran (see table A.1 in the appendix) show alarmingly high Variable Inflation Factor (VIF) scores, essentially rendering the regression models redundant. Multicollinearity – both in relation to *grass* and how it is constructed, as well as the other variables in the dataset – is thus a reason why prediction has been chosen over regression.

More generally, machine learning models allow for a higher level of flexibility. Particularly, regressions make a functional form assumption: that the relationship between the dependent and independent variables are linear (James et al. 2015, 21). There is nothing inherently wrong about such an assumption. However, machine learning methods that do not make such an assumption have the potential to describe the phenomena better if it is not linearly related to the independent variables.

Another reason why prediction has been chosen is to illustrate the use of machine learning models in the social sciences. I, in other words, use this thesis as an opportunity for demonstrating the use of innovative statistical methods and algorithms that may become even more central in comparative politics than they are today.

Despite having chosen prediction as opposed to regression/explanation, logistic regression models have been run and can be found in appendix A.1. They have been run in order to illustrate the necessity of prediction models (considering VIF/multicollinearity), and because the data had already been collected and processed.

3.2 Measuring marijuana policy

3.2.1 Marijuana policy in this thesis

The legality of marijuana for recreational use is the primary focus of marijuana policy in this thesis. Focusing on drug reform more generally, for instance decriminalization, depenalization, medical legalization, or perhaps law enforcement prioritization would achieve a similar goal as studying RML: surveying the predictive ability of public opinion, amongst other variables, on drug reform. However, due to constraints of time, issues of succinctness, and the novelty and controversial nature of RML, RML is the focal point of this thesis. Studying RML in light of public opinion serves a specific purpose – understanding recreational marijuana legalization – and a more general purpose: assessing existing policy change theories in light of drug reform, by answering the question ‘can the public’s attitudes predict drug reform?’.

In terms of measuring policy change, there are two primary ways of doing this quantitatively: ordinally and continuously (Caughey and Warshaw 2016, 900). Marijuana policy is in this thesis measured ordinally. The dependent variable *rec* is measured dichotomously: marijuana is either recreationally legal (1), or it is not (0). Questions of degree are not considered in the coding of this variable. Breadth has in this aspect been prioritized over depth. Classifying recreational marijuana policies on degrees of liberalism – a continuous measure - is a task in itself warranting a master’s thesis, as it requires a deep understanding of each state’s models of marijuana legality. For instance, the accessibility of marijuana, age restrictions and home-growing restrictions. Due to the complexities of measuring RML continuously, RML is in this thesis operationalized through a dichotomous variable. This way of measuring reflects policy and is distinct from measuring policy change. Measuring policy change would imply that the dependent variable is coded as 1 only for the year marijuana became recreationally legal. Studying the point in time when a policy changes is more theoretically interesting and perhaps more accurate for measuring public opinions predictive power in line with the responsiveness theory. However, the dependent variable is measured so that policy, and not policy change, is predicted. This is mainly due to there not being sufficient observations where the dependent variable is ‘1’ if an intervention (policy change) focus were to be chosen. This is because all observations following the intervention would have to be omitted.

Measuring marijuana policy in the ‘either legal or not’-manner is efficient in terms of time, yet also valid: recreational marijuana use is either legal or not. There are of course nuances that are not captured by a dichotomous measure. The dichotomous route has nonetheless been chosen due to its sufficient validity and pragmatic benefits.

Recreational legality is not the only possible type of marijuana policy used in this thesis. Medical legalization and decriminalization are also used as variables in this thesis. They are used as input variables for both the disaggregation models and predictive models. Their operationalizations are discussed in the next paragraphs.

3.2.2 Marijuana policy data

To the best of my knowledge, there is no publicly available dataset with variables showing marijuana laws by state (as of May 2022). I have therefore created a dataset with the four marijuana policy variables:

rec (marijuana legal for recreational use)

med (marijuana legal for medical use)

medyrs (years since marijuana became medically legal)

dec (marijuana is decriminalized)

This dataset is publicly available on my GitHub profile[2]. For each year in my analysis, the variable *rec* is coded as either '1' or '0'. To illustrate, consider Colorado in 2014. In 2014 the *rec* variable is set to 1. On the other hand, the variable is set to 0 for Colorado in 2010. This is because Colorado legalized marijuana for recreational use in 2012.

In constructing this dataset, an excel spreadsheet with the following variables was created manually: state abbreviation, state name, decriminalization year, medical legalization year and recreational legalization year. This spreadsheet was then used to construct a dataset in R spanning from 1976 to 2022 with each U.S. state. Only years 2010 to 2018 are covered in the analyses due to dependencies on other data, for instance public opinion data. Ballotpedia, Marijuana Policy Project (MPP), DISA and Pacula et al. were the sources used for the manual creation of the marijuana-laws dataset (Ballotpedia 2021, 2022; MPP 2022; DISA 2022; Pacula, Crhiqui, and King 2003)

It should be noted that *rec, dec, med* and *medyrs* are coded by the year in which marijuana legalization was enacted, not implemented. This is most significant for *rec*, as this is the

[2] https://github.com/alexcroz

dependent variable. To illustrate, recreational legalization was enacted in Colorado in 2012. However, legalization was not fully implemented until 2014, when state-licensed marijuana dispensaries opened (Monte, Zane, and Heard 2015). This poses issues of validities for this thesis. Measuring *rec* through either enactment or implementation implies studying two slightly different phenomena. Enactment implies the study of the change in law and criminality, whilst implementation implies a greater focus on the availability and access of marijuana. Despite being similar phenomena, it may impact the results of the analyses. In essence, this encourages one to ask the question of whether the actual legalization of recreational marijuana is more significant in terms of the observable relationships between the input variables and the dependent variable, or if the enactment of legalization (i.e., dispensaries) is more significant. Legal status has been chosen in part due to simplicity, yet also because I consider non-criminality to be a significant aspect of legalization.

3.3 Simulating data - Multilevel regression with post-stratification

3.3.1 MrP using autoMrP in R– a new approach

This section gives an overview of Multilevel regression with post-stratification, and addresses how it is used in this thesis, and why it is used.

As discussed, conceptually valid longitudinal public opinion data at the state-level does not exist. This implies that using public opinion as an input variable for the predictive models is not possible. However, thanks to MrP, accurate disaggregation of nationally representative data is feasible. GSS data is therefore disaggregated to the USCB division-level using the method known as Multilevel regression with post-stratification. The Census Bureau's nine divisions are the following: New England, Middle-Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain and Pacific (USCB 2022).

As mentioned, the reason why data is not simulated to the state-level is due to it not being possible with the publicly available version of the General Social Survey. In theory, it is possible because the data exists – it is just not publicly available. One can pay $750 and go through an application process in order to obtain the variable indicating what state respondents are from (NORC 2018). This would allow for simulating state-level data, which for instance Arceneaux has done. However, he used the disaggregation method, not MrP (Arceneaux 2002, 148).

Due to lack of funds and long waiting times, this thesis uses data that allows for simulation/disaggregation to the census division-level rather than the state-level – despite being sub-optimal. Having public opinion data at the divisional level poses difficulties for the validity and interpretability of the public opinion variable. Having the same public-opinion value for each year for up to nine states significantly decreases the variation in public-opinion across the dataset and assumes homogeneity of public opinion within census-divisions. I nonetheless simulate data for two main reasons: to illustrate the use of the innovative autoMrP package, which uses machine learning algorithms to improve upon traditional MrP, and to give an indication of the predictive ability of public opinion. If this operationalization of public opinion helps the models predict whether a state has legalized marijuana or not, there is reason to believe that a more valid operationalization of public opinion (state-level) would fare very well in predicting RML.

How does MrP work?

Prior to the introduction of MrP by Gelman and Little in 1997, disaggregation of national level public opinion data was the standard for dealing with a lack of representative data at the levels of interest (in this case U.S. states) (Arceneaux 2002; Lax and Phillips 2009b, 107; Buttice and Highton 2013, 450). Disaggregation of data is done by pooling multiple national-level surveys into one dataset and disaggregating the data to create state-level averages of public opinion. The method is simple and relatively accurate if the prerequisite of having a large dataset comprised of multiple national-level surveys (or one large national-level survey) is met. However, the method suffers from poor accuracy compared to MrP when there is no large sample to disaggregate (Buttice and Highton 2013, 451-452; Lax and Phillips 2009b, 109; Broniecki, Leemann, and Wüest 2021, 1).

MrP, on the other hand, uses nationally representative individual-level data to *simulate* averaged values for lower geographical units. These simulations are based on individual-level determinants of public opinion towards marijuana legalization, as well as context-level predictors. Wang et al. describes the steps of MrP as follows:

> "The core idea is to partition the population into cells based on combinations of various demographic and political attributes, use the sample to estimate the response variable within each cell, and finally aggregate the cell-level estimates up to a population-level estimate by weighting each cell by its relative proportion in the population" (Wang et al. 2015, 982).

One thus uses context-level predictors, in this case geographic estimators unique for each division, and individual-level predictors, followed by post-stratification in order to simulate data.

MrP using autoMrP – accuracy

MrP has consistently been shown to be more accurate than Erikson and colleagues' disaggregation approach and is continually improving (Wang et al. 2015; Lax and Phillips 2009b; Buttice and Highton 2013; Broniecki, Leemann, and Wüest 2021). Broniecki, Leemann and Wüest advance MrP's efficacy in their recent R package *autoMrP*. Their version of MrP improves on traditional MrP by using five different machine learning classifiers and combining the results. autoMrP uses the following five different machine learning classifiers:

> "multilevel regression with best-subset selection of context level predictors, multilevel regression with principal components of context-level predictors (PCA), multilevel regression with L1 regularization (Lasso), gradient tree boosting, support vector machine" (Broniecki, Leemann, and Wüest 2021b, 4).

The results from these models are then combined using ensemble Bayesian modelling (EBMA) (Broniecki, Leemann, and Wüest 2021b, 4). EBMA essentially weighs the different machine learning model's predictions and combines them to create a more accurate prediction than any of the models by themselves (Montgomery, Hollenbach, and Ward 2012, 274).

As Broniecki and colleagues have done, testing MrP's accuracy is in theory quite simple. Testing the accuracy can be done if one has actual representative data at the state-level. The authors of autoMrP illustrate and test the accuracy of autoMrP using the 2008 National Annenberg Election Studies (NAES) data set, consisting of over 50,000 respondents, by simulating and subsequently comparing the accuracy of autoMrP and other MrP methods to the complete state-level representative dataset (Broniecki, Leemann, and Wüest 2021b, 4).

Prior to running the MrP models Broniecki and colleagues average responses for each state based on the full NAES dataset and treat these state-level averages as the true public opinion. They then select 1500 respondents from the dataset (minimum 5 respondents per state) and run autoMrP and other MrP methods using the partitioned dataset (Broniecki, Leemann, and Wüest 2021b, 16). They then compare the MrP methods with the 'truth' dataset. autoMrP improves upon standard MrP by reducing Mean Squared Error (MSE) by almost 20%. The MSE is, furthermore, substantively low – at respectively 0.0025 for standard MrP and slightly more than

0.0020 for autoMrP (Broniecki, Leemann, and Wüest 2021b, 17). However, testing MrP's accuracy on marijuana public opinion is not possible as there is no 'truth' dataset to compare the simulated data to. Were there such a dataset, MrP would not be necessary. The previous example illustrates that MrP is indeed a valid approach for dealing with non-representative data in general, implying this to be the case for marijuana public opinion as well.

To reiterate how MrP works, the following formulation from Broniecki, Leeman and Wüest succinctly captures the essentials of MrP:

> "Most MrP models consist of two parts: a set of random effects for individual-level socioeconomic variables and a set of fixed effects for context-level variables" (2021, 597).

Building autoMrP models can be summed up with the following requirements: survey-data to disaggregate, individual-level predictors from the survey-data (in this case demographic ideal types), context-level predictors from other datasets and census-data on the proportion of each ideal-type living in each division.

3.3.2 autoMrP data – General Social Survey, U.S. Census data and more

This section includes a description of the data and variables used for performing autoMrP. The main dataset is GSS. The GSS is a nationally representative survey conducted since 1972 (ARDA 2022). This data is the foundation of this thesis' data-simulation, as it contains the variable theorized to be an important predictor of RML – attitudes towards legalization. The variable, *grass*, is the following survey question: "Do you think marijuana should be made legal or not?". The values, except for refusals and 'don't know' are dichotomous: 'Should be made legal' and 'Should not be made legal'. This variable has been the same for all years used in this thesis, the years being 2010, 2012, 2014, 2016 and 2018. The datasets contain roughly 1500-2000 respondents for each year, similar to the size of datasets that have previously been used to simulate data using MrP (Broniecki, Leemann, and Wüest 2021; Buttice and Highton 2013, 450).

As mentioned, the fact that the GSS is conducted biannually means that MrP can only be applied biannually. Apart from not having data on which state respondents are from, the GSS is an appropriate dataset for this thesis, allowing insight onto marijuana legalization with never-before used data.

Other individual-level variables from the GSS are also used to simulate the disaggregated data. These variables are the respondents age, race, and gender. The respondents' ages have been coded into five categories. The variables have then combined to create one ideal-type variable for each respondent: age-sex-race (*asr*). This variable is crucial in the post-stratification step of MrP, particularly in unison with the proportion variable created from Census data (Broniecki, Leemann, and Wüest 2021b, 5).

United States Census Bureau data

This brings us to the next dataset: The United States Census Bureau (USCB). The USCB estimate the population and demographics of each year between the decennial censuses – the American Community Survey (ACS). The ACS 1-year estimates, as opposed to the 5-year estimates, have been used for the MrP models. Using the 5-year estimates for one point in time is inappropriate as it should be considered an average of these (USCB 2018, 13).

A 'proportion' variable has been created in order to indicate the amount of people living in each division that correspond to the individual level variable *asr*. For instance, the proportion of the population in New England in 2012 that were white males in the first age group (18-32) was 15%. This proportion variable is then used with GSS data to weigh every GSS respondent's answer in the post-stratification step. Apart from the proportion of people living in each region, the following variables have been taken from, or created using, census data: total population, share of Hispanics (*hprop*), share of whites (*wprop*), median age (*age*), proportion of population that are highly education (*heduc*) and the proportion of people who have low education (*leduc*).

Each of these variables are at the division-level, as this is the level of disaggregation. The proportion of people 25 years or older with a bachelor's degree (or higher) are considered highly educated, whilst people 25 years or older without a high school degree are considered to have a low level of education in the dataset. Both variables have been added as both low education and high education have shown, separately to be strong determinants of public opinion – as discussed in the theory chapter.

The inclusion of race demographics is based in part on theory and previous studies, and simple analyses of GSS data. Stringer and Maggard, for instance, found non-white people to be more likely to be in favour of legalization of marijuana (Stringer and Maggard 2016, 437). The biggest reason why race demographics has been included using census data is, however, as it demographically describes the census divisions, allowing the autoMrP models to more accurately – and discriminately – make predictions for support towards legalization in each division.

Age as a predictor of attitudes towards marijuana legalization – both as an individual level variable and a context-level (median age) variable – is more theoretically and empirically grounded. Younger people are more likely to support marijuana legalization and older people are less likely to support marijuana legalization. Particularly, Stringer and Maggard's analyses showed that birth year was statistically significant and had a positive coefficient in all models on attitudes towards legalization (Stringer and Maggard 2016, 437). Cruz, Boidi and Quirelo only found a relationship between age and support in one of their 2018 articles. They found younger people to be more liberal in their support towards different models of access to legal marijuana, for instance self-cultivation and cannabis clubs (Cruz, Boidi, and Queirolo 2018b, 432). They did not find a relationship in their second 2018 article. However, the age categories were large – the first one including people 26-40 years old (Cruz, Boidi, and Queirolo 2018a, 71). At the individual-level, GSS' age-variable has been used to construct the ideal type, as mentioned. Median age in each division from the U.S. Census Bureau has been used as context-level variables.

The education-variables gathered from the ACS 1-year estimates indicate, as mentioned, the proportion of the population in each division that have a low level of education and high level of education, respectively. These variables have been chosen based on the studies of the determinants of support towards marijuana legalization. Cruz, Boidi and Queirolo found a strong relationship between higher education and support towards legalization in both their 2016 and 2018 articles (2016, 315; 2018a, 71). In addition, Stringer and Maggard also found education to be related to support for legalization (2016, 437).

Pew Research, U.S. Bureau of Labor Statistics and MIT data

Besides USCB data, data from Pew Research, U.S. Bureau of Labor Statistics (BLS) and Massachusetts Institute of Technology (MIT) have been used for the context-level autoMrP variables. An index variable indicating the share of each state's population that are 'highly religious' has been used to create a context-level variable. The variable is part of the Pew Research Religious Landscape Study, a survey of more than 35,000 respondents (PRC 2014). This variable is included because religion (and non-religion) has, as discussed in the theory chapter, consistently shown to have an effect on support towards legalization on the individual-level (Cruz, Queirolo, and Boidi 2016; Stringer and Maggard 2016; Cruz, Boidi, and Queirolo

2018a; Felson, Adamczyk, and Thomas 2019). The index variable only has values for 2014 but has been applied to all years. Conveniently, this is the middle of the timespan of the simulated data (2010-2018). This minimizes the obvious reliability issues of using a static/time-specific variable for different points in time. The variable has been altered to represent religiosity at the divisional-level by using census data indicating the total population in each state/division.

BLS data has been used to construct a context-level variable measuring the share of the population in each division that are unemployed. Employment status has been shown to increase the likelihood of a state legalizing marijuana medically (Bradford and Bradford 2017). Despite MML differing theoretically from predicting opinions, unemployment has nonetheless been included in order to give the autoMrP models more data by which it can simulate public opinion data, as it contributes to contextualize each division – and in order to reflect diffusion theory.

Finally, MIT data has been used to create a variable showing how many that, in percent, voted for the republican nominee in the last presidential election. Political affiliation, at the individual-level, correlates with RML preferences. Felson and colleagues found democrats to be more supportive of RML and republicans to be less supportive, whilst Cruz and colleagues found left-leaning people to be more supportive than right-leaning respondents (Cruz, Queirolo, and Boidi 2016, 315; Cruz, Boidi, and Queirolo 2018a, 71; Felson, Adamczyk, and Thomas 2019, 22).

Overview of autoMrP input variables

To summarize, the autoMrP models have one L1 (individual-level) and eight L2 (context-level) variables. All variables are empirically and/or theoretically substantiated, be it in previous literature studying the determinants of public opinion, or the likelihood of a state legalizing medical marijuana. The L1 variable is an ideal-type consisting of three variables. Having more L1 variables would be beneficial, particularly considering that variables such as ideological self-placement and religiosity have proven to be strong predictors of support towards RML at the individual level. Due to the nature of MrP, the post-stratification step requires the model to know the proportion of people with each combination of all L1 variables that reside in each division. For instance, say self-placement on the left-right scale (political affiliation) was included as an L1 variable, and furthermore that observation 1 - a white male in the first age group from New England - placed himself on the middle of the scale. This would require knowledge of how many white men in the first age group self-identify as centrists in New England. This demands detailed census data, which is not available. For this reason, the ideal type (age-sex-race) is the only level 1 variable included. Variables reflecting politics (share of the electorate who voted for the republican candidate), and religion (religiosity index) have been included at the context-level in order to capture some of the variations at the division-level support towards marijuana legalization, as individual-level operationalizations of these variables are not possible to include in the models.

3.3.3 autoMrP models and results

As previously emphasized, it is only possible to simulate/disaggregate data at the divisional level with the available data. Disaggregating to the divisional level and using that data to predict state-level marijuana legalization essentially has two assumptions: public opinion is heterogeneous across divisions, and homogenous within divisions. For the data to meaningful, these assumptions must be met to a certain extent. If the assumptions are not met, the public opinion variable is not a meaningful operationalization of the public's attitudes towards marijuana legalization. The consequence of this is that public opinion will have little to no predictive effect on RML in the prediction models when it otherwise, with for instance state-level data, could predict well. These are quite strong assumptions when taking into consideration that the divisions consist of anywhere from three to nine states. For instance, the assumption of homogeneity of public support in the South Atlantic division, consisting of nine states, is quite demanding.

As discussed in 3.2.1 the autoMrP package in R allows for combining five machine learning models into one MrP prediction. All five algorithms have been run for all time points (2010, 2012, 2014, 2016 and 2018). A total of five datasets, one for each model, were created for each time-point, which were then been combined into one using EBMA.

Figure 3.1 shows the averaged support towards marijuana legalization in the U.S. from 2010 to 2018 according to GSS data. 1 being 100% of the population supporting marijuana legalization, 0 being 0% of the population supporting marijuana legalization. 2012 saw an overall dip in support compared to 2010, subsequently increasing by roughly 15-25% each year. The reason why support dropped in 2012 is not known. One can expect most divisions to display a similar pattern as the national average..

Figure 3.1 *Support towards marijuana legalization, United States, 2010-2018*

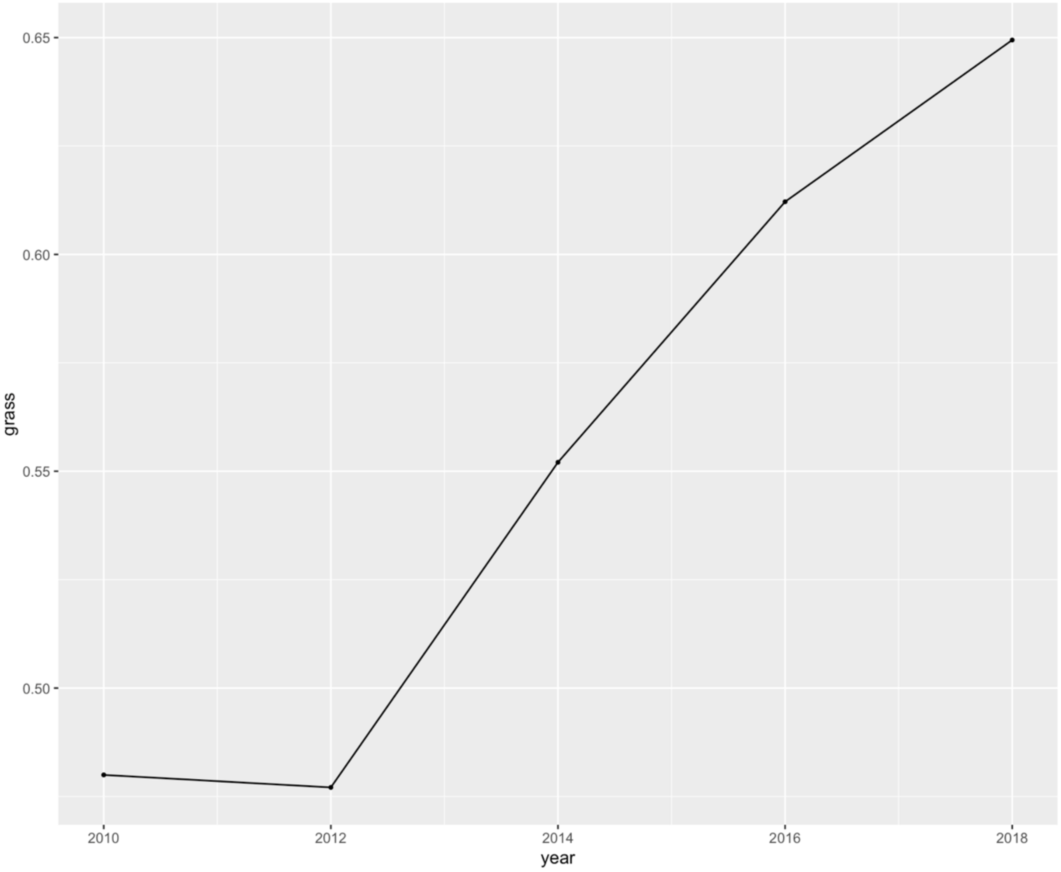
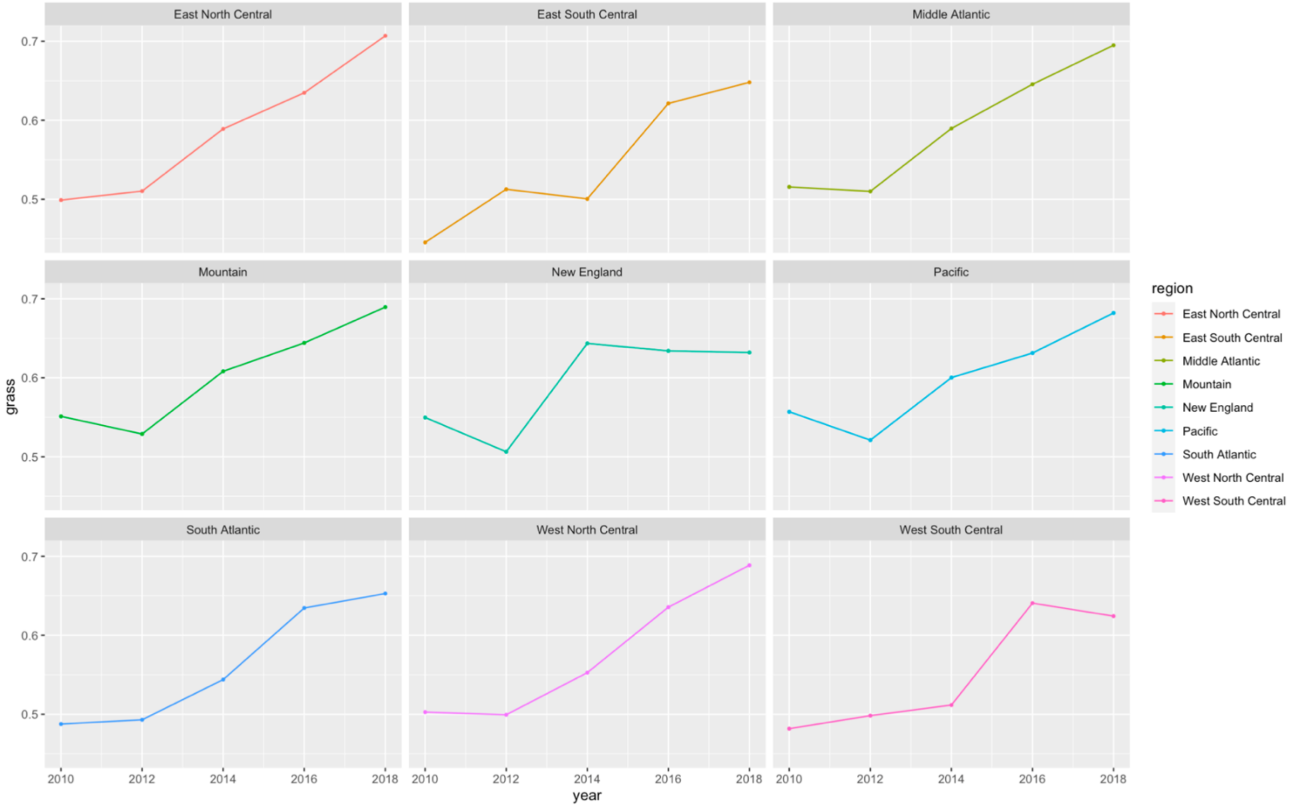


Figure 3.2 shows the results from the autoMrP models. That is, the simulated disaggregated data at the USCB division-level. Most divisions show a similar development of support towards legalization as the national average: a slight dip in 2012, followed by an increase each subsequent year. The assumption that public opinion is heterogeneous across divisions is in many ways fulfilled, despite the divisions showing similar patterns. The simulated data shows support to be higher in every division in 2018 compared to 2010, but the divisions have different peaks in support: New England for instance peaked in support in 2014, whilst others peaked in 2016 or 2018. The data displayed in figure 3.2 will be used in its raw form for the predictive models.

Substantively speaking, most divisions have a quite high proportion of the population supporting the legalization of marijuana. From 2014 and onwards, the majority of every division's population is in favour of marijuana legalization. That is, if this data can be considered as valid.

Figure 3.2 *autoMrP Ensemble Model Averaging (EBMA) by USCB Divisions, 2010-2018*



3.4 Interpretive Predictive Machine Learning Models

The purpose of this sub-section is to describe how the legality of recreational marijuana is predicted in this thesis. In particular, what machine learning algorithms are used, how the models are built, why these different algorithms have been chosen, and how they differ from each other. Furthermore, I discuss a concept known as "interpretive predictive machine learning", how interpretability is applied, and how it differs from so-called 'black box' models.

There are two main aims of the predictive machine learning models: 1) creating strong predictive models and identifying the model/algorithm that predicts best and 2) explaining how the models predict. Explaining how models predict may give insight into future areas of focus on both causal and predictive endeavours. The first aim may be fulfilled with so-called 'black box' models – models that predict well but are not interpreted (Boehmke and Greenwell 2019, 305). They are called black boxes because of their complex inner workings.

Considering the second aim, methods for interpreting machine learning models are applied in order to give insight into *how* the models predict. One of which is variable importance. Variable importance illustrates how important a variable is for making accurate predictions in each model (Boehmke and Greenwell 2019, 312). This is done to get insight into which variables are the most important, both for this thesis and for future research. A method for uncovering how variables interact with each other in each model – so called interaction plots – is the second aspect of what makes this thesis' models interpretive.

Interaction plots measures interactions between variables. Interactions between variables are known as "the change in the prediction that occurs by varying the features after considering the individual feature effects" (Molnar 2019, 124). In other words, interaction values indicate the extent to which a variable's ability to predict depends on another variable's value. For instance, if variable x has a high interaction value, it means that it depends on other variables (for instance variable z) in order to predict Y. In this thesis, holistic variable interaction scores are presented, as well as scores for combinations of variables (specific interaction score). Figure 4.10 (SVM 1 interaction plot) is an example of the prior, whilst figure 4.11 (*medyrs* in SVM 1 interaction plot) is an example of the latter.

Because the dependent variable recreational marijuana legalization is binary, the machine learning algorithms used in this thesis are classificatory. The models categorize each state-year observation as either recreationally legal ('yes' or '1'), or not legal ('no' or '0').

3.4.1 Input Data and Data Splitting

Input data

As discussed in the previous sections, simulated public opinion data at the division-level is one of the main input variables (*grass*). In addition to this, a series of marijuana-policy specific variables are included in the predictive models. One of these is a variable operationalizing geographical policy diffusion (*diffusion*). This variable simply measures the percent of neighbouring states that have legalized recreational marijuana and is created by using the marijuana policy dataset I have created. This variable has the advantage of being simple to understand, and time-efficient to create. However, the models do not fully cover diffusion theory. Apart from including unemployment rate partly as a variable reflecting motivation for enacting RML, the models do not directly consider the ideological aspect of policy diffusion. An example of a variable measuring ideological diffusion would be one that measures the ideological differences between each state and the states that have previously legalized marijuana. However, the share of each state's population that voted for the republican presidential candidate in the previous presidential election (*repshare*) has been included. This, in some ways, reflects ideological diffusion, considering that most states that legalized are democrat-majority states. Perhaps more importantly, it captures the ideological differences in people who are pro-legalization and those that are against.

Furthermore, medical marijuana and decriminalization are included. A variable indicating whether the state has legalized medical marijuana (*med*), a variable indicating whether the state has decriminalized marijuana (*dec*), and a variable measuring the years since MML was enacted (*medyrs*) are included in the models. *Medyrs* has a negative value if the year in question is *prior* to the year the state legalized medically, a positive value if the year in question is after the state legalized medically, and zero if it is the year of legalization *or* if the state does not have medically legal marijuana. The latter aspect of the operationalization is problematic, as states that legalized medically that year are put in the same category as states that have not (prior to 2018) legalized medical marijuana. This unfortunate operationalization is, however, necessary as the algorithms do not deal well with missing values. However, this issue is to some extent offset by the variable *med*. If there is a pattern between MML and RML, particularly the time it takes for MML to develop into RML, the latter variable will capture this relationship.

As discussed in the theory section, the availability of ballot initiatives is important both empirically and theoretically. Therefore, a dichotomous variable indicating whether a state allows for ballot initiatives (*ballot*) is included as an input variable

In addition to these variables, a range of demographic variables have been included. These are included both on a theoretical basis, and in order for the models to distinguish between the states (i.e., description). GDP per capita (*gdppc*), for instance, is a variable that has been included to capture the economic context of each state. Furthermore, the proportion of black people (*bprop*), Hispanic people (*hprop*) and white people (*wprop*) in all states have been included as input variables, gathered from the USCB. The proportion of men in each state (*male*), the median age (*age*), as well as the proportion of people with a high education (*heduc*) and low education (*leduc*) have all been included as input variables. These are mainly based on the literature describing and analysing the determinants of individual support towards marijuana legalization.

Data splitting – training data vs test data

Before building and testing the models, one needs to split the dataset into two parts: a training dataset and a test dataset. The machine learning models are built on the training dataset, and predictions are tested on a test dataset. The training set is usually 50-80% of the total dataset, whilst the remaining observations fall into the test-dataset. Based on the model built on the training dataset, the observations in the test dataset are categorized as either ‘legal’ or ‘not legal (‘yes’ and ‘no’). There is no universally recognized optimal split between training- and test set. One must therefore make a judgement of how much of the original dataset should be put in the training dataset (Boehmke and Greenwell 2019, 16). This judgement is based on the size of the full dataset. More specifically, the training set should be large enough for the models not to be biased by too few observations. At the same time, one should have a test dataset that is large enough for the results to be generalizable, and that one has sufficient variation in each variable, particularly the dependent one. When working with very large datasets (multiple thousands of observations), the training and test sets will be large enough independent of whether one chooses a 50-50 split or a 75-25 split.

The results are more sensitive to how the data is split when one has a small dataset – such as the one used in this thesis (roughly 250 observations. This is because each observation in the training set has a higher impact on the model than in a large dataset. The impact of each individual observation (bias) is one of the biggest limitations of this thesis. Any dataset under 1000 observations is generally considered quite small for machine learning models (Vabalas et al. 2019, 17). However, due to this thesis focusing on a relatively new phenomenon, and observations being at the state-level, a small dataset is to be expected.

Due to few observations of the ‘legal’ class, a 70-30 split has been chosen. A 50-50 split would render the training dataset too small to create unbiased models, and an 80-20 split would have too few observations of the ‘legal’ class in the test dataset. With too few observations of this class, the overall prediction accuracy of each model is greatly impacted by the prediction of very few ‘legal’ observations. There are a total of 247 observations in the full dataset (roughly 49 observations per year). Because 70% of the observations are in the training data, and 30% are in the test data, the models are trained on 174 observations and tested on 73 observations.

It is, as mentioned, generally recognized that machine learning models thrive with large datasets (Vabalas et al. 2019, 1). As such, scholars have considered the issue of having a small dataset, and how to deal with this when creating machine learning models. One issue with a small dataset is bias. Considering the fact that each observation is a larger percentage of the total dataset, the models are more likely to be skewed and essentially be statistically insignificant because the patterns the models find are specific to the limited observations of the training dataset. Particularly, Vabalas et al. find that K-fold cross validation (CV), a validation approach used in this thesis (and explained thoroughly later in this chapter) is significantly negatively impacted by having small datasets (N<1000) (Vabalas et al. 2019, 17). Bradley and Greenwell make a similar point, comparing K-fold CV to bootstrapping (2019, 26).

3.4.2 Choice of machine learning methods

The supervised machine learning algorithms used in this thesis includes: 1) shrinkage methods (lasso, ridge and ElasticNet), 2) support vector machines (SVM), 3) single decision trees, 4) random forests (RF) and 5) gradient boosting machines (GBM). The point of running a wide range of machine learning methods is to gain insight on which method is best suited for predicting recreational marijuana legality, and *how* it is best predicted. The no-free-lunch (NFL) theorem is also central in the choice of performing multiple models/algorithms. NFL states that *"averaged over all optimization problems, without re-sampling all optimization algorithms perform equally well"* (Adam et al. 2019, 58). In essence, the no-free-lunch theorem implies that there is no one best prediction algorithm, and that a range of different algorithms, tuned differently, should be ran in order to get a strong predictive model for the particular prediction problem.

The above-mentioned methods have also been chosen because they differ on multiple key areas: whether they are parametric or non-parametric and their flexibility, interpretability, and functional-form assumptions. Most statistical learning models, including those in this thesis, can by and large be categorized as being either parametric or non-parametric (James et al. 2015, 21). Parametric methods differ from non-parametric methods in that they assume the functional form of *f*, the function for estimating the dependent variable. For instance, a form assumption can be linear. Such an assumption implies that Y – the dependent variable – is linearly related to the independent variables. Linear regression models are an example of parametric methods that assume the functional form of f to be linear (James et al. 2015, 21). James et al. use the following example to illustrate:

$$income \approx \beta_0 + \beta_1 \times education + \beta_2 \times seniority$$

In this example, income is a function of the constant (β_0), education and seniority. In this linear function, the prediction problem boils down to estimating β_0, β_1 and β_2 (James et al. 2015, 22). Parametric models are less flexible than non-parametric models because they make a functional form assumption. However, they benefit from making such an assumption by simplifying the problem to estimating a set of coefficients (James et al. 2015, 22). Therefore, they have a high degree of interpretability, and can have faster run times. However, this comes with a trade-off. If the assumed functional form is far from the *actual* functional form of *f*, the model will not predict well because it distorts the reality of the relationship by forcing it into a functional form that is not its true form (James et al. 2015, 104). The true form being the actual 'function' that models the way in which the output variable is related to the input variables.

Non-parametric methods, on the other hand, do not assume a functional form and force the problem into this assumed form – be it linear, quadratic, or any other functional form. Non-parametric models are therefore more flexible and can provide better predictions due to their flexibility (James et al. 2015, 104). There is, nonetheless, one major disadvantage with non-parametric models: they require a very large number of observations because they do not estimate *f* by a few parameters, as parametric methods do (James et al. 2015, 23). They essentially 'create' *f* using the input data. Non-parametric models also run the risk of overfitting the training data – a phenomenon that occurs when the models picks up on patterns in the training data that are purely random as opposed to a function of the prediction problem's 'true' functional form (James et al. 2015, 32).

Due to their different characteristics and the no-free-lunch theorem, parametric methods (ridge, lasso and ElasticNet) and non-parametric methods (support vector machines, random forests (RF), single decision trees and gradient boosting machines (GBM)) are used to predict the legality of recreational marijuana. However, the single decision trees are primarily used as a pedagogical tool for understanding GBM and RF better, as they are not particularly strong predictors.

3.4.3 How each machine learning method works

Having explained the basis of choosing the different machine learning methods for predicting marijuana legality, a more detailed discussion of each algorithm is warranted. Therefore, this section explains how the different methods used in this thesis work. Understanding how the different methods work is important for the interpretation of the results, and why some methods prove to be better at predicting marijuana legality than others.

Shrinkage methods

Shrinkage methods, also known as regularization methods, are a type of linear parametric machine learning methods that regularize – or shrink – the estimated coefficients towards zero. This can help reduce the variance and out of sample prediction error (Boehmke and Greenwell 2019, 121). Variance is the degree to which the model changes when built on different data – an important metric for assessing prediction models (James et al. 2015, 34). The point of shrinking coefficients is to eliminate predictors that are model-specific (i.e., reduce variance) as well as highlighting which predictors are most important.

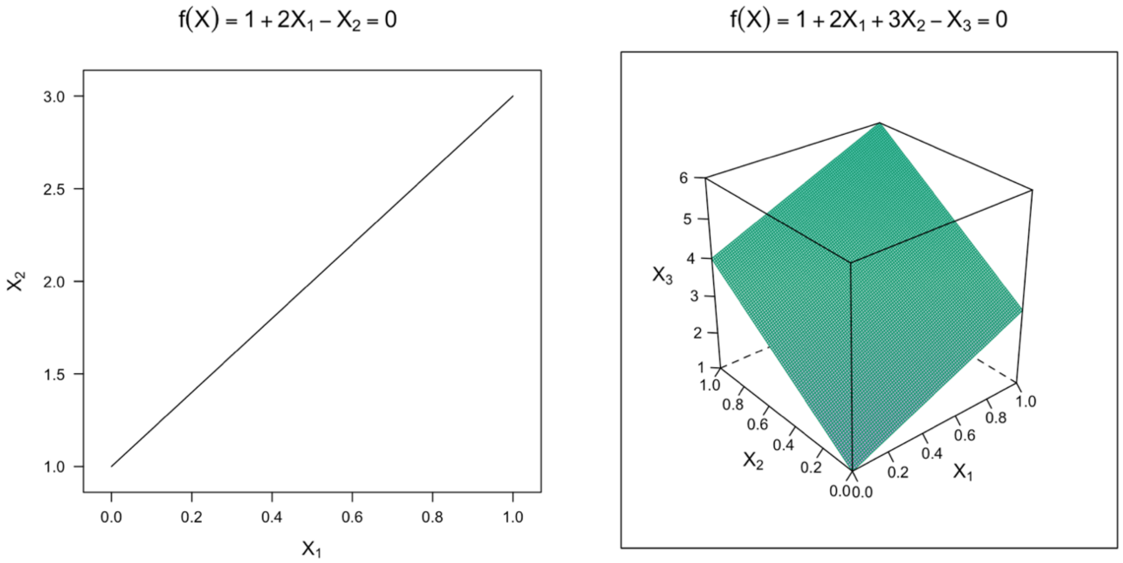
There are two main types of shrinkage methods: ridge and lasso. The main difference between ridge and lasso is that ridge shrinks the coefficients towards zero, but never to zero. Lasso, on the other hand, can shrink coefficients to zero (James et al. 2015, 219). Strict penalties have the advantage of eliminating variables that are not particularly useful to the model. This is useful from a theoretical perspective as it highlights which variables are important whilst eliminating redundant variables. This advantage is more relevant when one has many input variables, which is not the case in this thesis. A third type of shrinkage method is called ElasticNet (EN). ElasticNet is a middle way between ridge and lasso (Zou and Hastie 2005, 303). The shrinkage penalty applied to predictors that are not useful is regulated by λ (lambda), a tuning parameter (James et al. 2015, 215). Tuning parameters, also known as hyperparameters, are configurations that must be specified by the person creating the models. Different types of machine learning models have different number and kinds of hyperparameter. As mentioned, shrinkage methods only have one: shrinkage penalty.

The optimal λ value is identified using what is known as k-fold cross-validation. K-fold cross-validation (CV) is "a resampling method that randomly divides the training data into *k* groups (aka folds) of approximately equal size" (Boehmke and Greenwell 2019, 23). When building predictive models, it is very important that the test dataset remains unseen, as this is how one gauges its performance. However, when attempting to find the optimal tuning parameter for the models, one needs to test their performance. This is *not* done with the test-data. This is where the validation approach comes into play. K-folds CV is a form of validation method used to compare models with different tuning parameters built on the same training dataset. K-folds CV works by splitting the training data into K-folds, usually 10. K number of models are then built on $K - 1$, and tested on the one fold that was left out. For instance, the first model is built using folds 1 through 9. The prediction model is then tested on the 10th fold. K-folds CV is used to test different tuning-parameters without exposing the models to the test-dataset, as this must be kept ‘secret’ until the final model is built. In this case, CV is used in order to compare ridge models with different lambda values to then find the value for lambda resulting in the highest prediction accuracy. The “one standard error rule” (1se rule) has not been used when choosing lambda values. The idea of the 1se rule is to choose the largest lambda that has a cross validated error rate within one standard error of the error rate produced using the optimal lambda value. The reason for this is to have a parsimonious model. The result is that one tunes a model that is simpler than, yet almost as accurate as the most accurate model (Chen and Yang 2021, 868). The reason why this rule of thumb has not been used is because it produced significantly lower cross validated accuracy than using the minimum lambda value.

Support vector machines

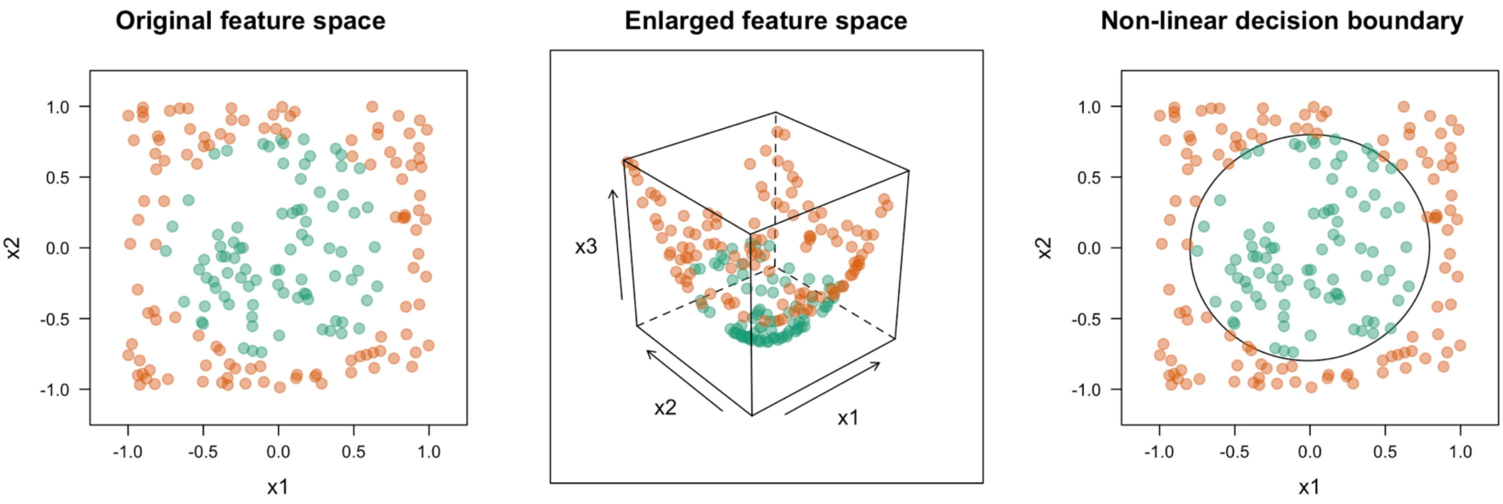
Support vector machines, the second prediction method, is a non-parametric machine learning method. Support vector machines are commonly used for binary classifications, as is the nature of this thesis’ problem (classifying the two categories ‘legal’ and ‘not legal’) (Boehmke and Greenwell 2019, 271). SVM essentially attempts to separate the two classes by finding a hyperplane in the feature space that does so. This hyperplane acts as a decision boundary, separating the classes (Boehmke and Greenwell 2019, 273). In a two-dimensional space, a hyperplane is simply a line. In a three-dimensional, a hyperplane is a plane, as visualized in figure 3.3. If the prediction problem only included two predictors (p = 2), SVM would classify based on which side of the line the observations fall (see left graph, figure 3.3).

Figure 3.3 *2D and 3D hyperplanes*



Left: 2D hyperplane. Right: 3D hyperplane (Boehmke and Greenwell 2019, 273)

Figure 3.4 *Kernel trick visualization*



Left: Original feature space. Middle: Enlarged feature space. Right: Decision boundary from the enlarged feature space applied to the original feature space (Boehmke and Greenwell 2019, 278)

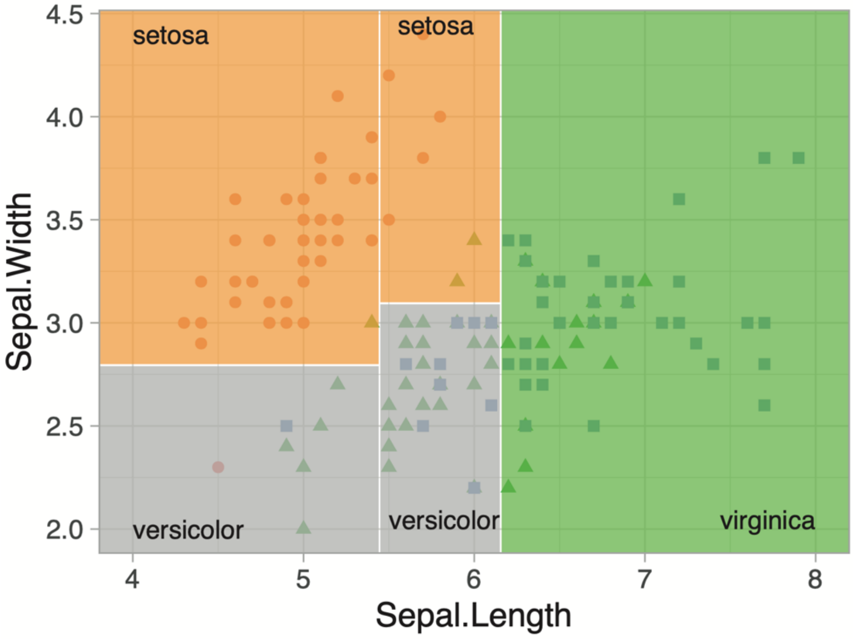
In addition to the concept of hyperplanes, SVMs build on the mathematical concept known as the 'kernel trick'. This 'trick' involves the enlarging of the feature space, making it more likely that a hyperplane is able to separate the data into two separate classes (Boehmke and Greenwell 2019, 277). This is because separating classes by drawing a line, as visualized in figure 3.3, is often too simple to be accurate. Drawing a line in figure 3.3 (left) would not separate the classes well. One therefore uses the kernel trick to enlarge the feature space. This enlargement is visualized in the progression of the graphs from the left of figure 3.4 to the middle of figure 3.4. Following the enlargement of the feature space, the decision boundary is drawn on the enlarged feature space. Finally, the decision boundary is applied to the original feature space, resulting in a non-linear decision boundary that separates the classes with a higher degree of accuracy than a 2D hyperplane (Boehmke and Greenwell 2019, 278). The kernel trick uses different kernel functions. SVMs generally use one of three kernel functions: *d*-th degree polynomial function, radial basis function and hyperbolic tangent (Boehmke and Greenwell 2019, 278-279).

SVMs have a higher number of tuning parameters than shrinkage methods, which only have one (lambda). The different kernel functions are one example of hyperparameters that the model-builder must specify. Furthermore, each type of kernel function has different hyperparameters. Cost is the common denominator of hyperparameters for all of the three mentioned kernel functions (Boehmke and Greenwell 2019, 279). In short, the cost is the extent to which the SVM model allows observations to be on the 'wrong' side of the decision boundary. With a low cost, the SVMs have a soft margin classifier. With a high cost, the model has a so-called 'hard margin' (Boehmke and Greenwell 2019, 276). The cost is essentially the degree to which the models are penalized for allowing outliers. Further hyperparameters include degree and scale for the polynomial kernel function, and sigma for the radial basis kernel function. These hyperparameters regulate aspects of SVM like the flexibility of the decision boundary, and the degree to which outliers/observations that are close to the decision boundary impact the fit of the decision boundary (Boehmke and Greenwell 2019, 276). Due to these hyperparameters, SVMs being non-parametric, and the nature of SVM's kernel trick, SVMs are very flexible.

Decision trees

Decision trees also belong to the non-parametric family of machine learning algorithms (Boehmke and Greenwell 2019, 175). They work by searching for patterns in each class in terms of their values on the input variables. Decision trees partition the feature space into non-overlapping regions (see figure 3.5). These regions are assigned to one of the dependent variables' classes. In this particular example, different types of flowers (setosa, versicolor and virginica) are classified based on the sepal length and width. The figure is taken from Boehmke and Greenwell (2019, 179).

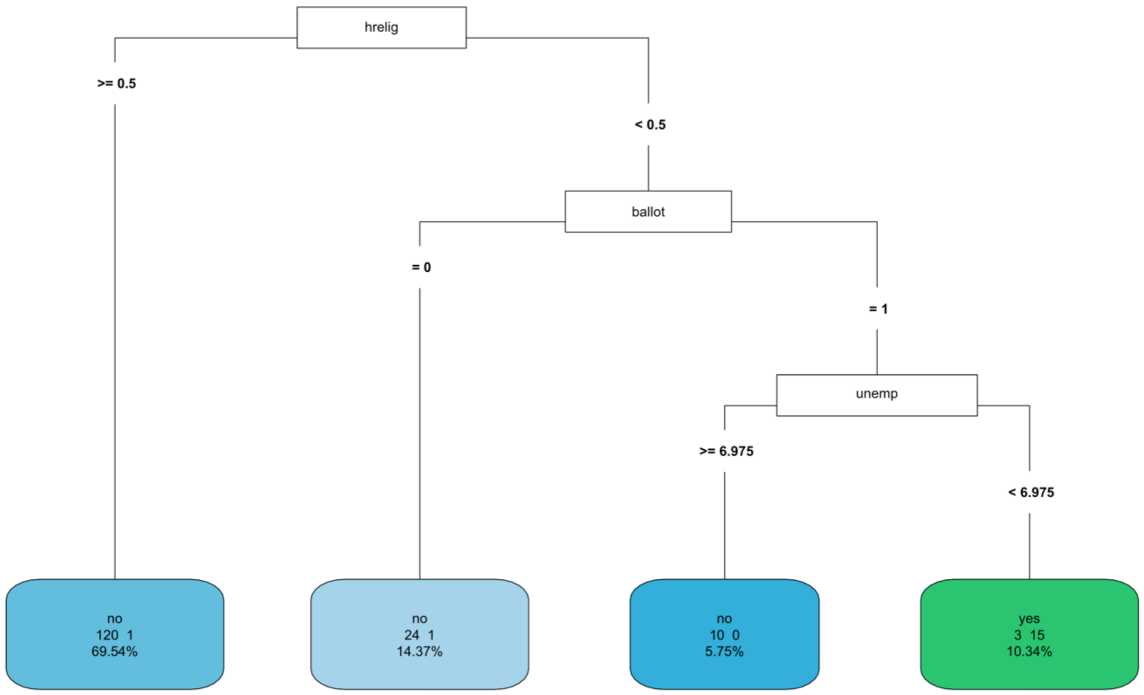
Figure 3.5 *Decision tree partitioning*



(Boehmke and Greenwell 2019, 179)

However, visualizing them as a tree (as seen in figure 3.6) is more common. They are therefore easy to interpret, although they are very likely to overfit the data and produce a low degree of prediction accuracy (Boehmke and Greenwell 2019, 175). They have been included mainly due to their visual interpretability, as well as to serve as a stepping-stone to understanding more complex and accurate tree-based methods such as random forest and gradient boosting machines. There are multiple algorithms for fitting decision trees. This thesis' decision trees are fit using CART – classification and regression tree algorithm. Figure 3.6 shows one of the decision trees created in this thesis. Trees are built up of nodes. The top of a decision tree is called the root node. In figure 3.6, the root node is *hrelig*. There are two decision nodes in this tree: *ballot* and *unemp*. *ballot*, for instance, splits into two categories: 0 and 1. Splits are followed by branches, connecting to the next node. At the bottom of the tree is what is called terminal/leaf nodes. Leaf nodes do not split into further nodes. Finally, depth is also used to describe decision trees. The decision tree in figure 3.6 has a depth of three, as one does not count the root node (James et al. 2015, 304-324).

Figure 3.6 *Decision Tree 2*



Decision trees, despite being quite simple, have a range of hyperparameters that can be tuned. When tuned optimally, these parameters can to some extent decrease its inherently high risk of overfitting. The most important parameter is how deep the decision tree is. A deeper tree is more detailed, however this comes at a cost: overfitting the model on the training data (Boehmke and Greenwell 2019, 180). Early stopping (shallow trees) as well as a concept known as pruning are important for decreasing the risk of overfitting. Pruning, however, is not a hyperparameter. Rather, pruning is a process where one grows a large and complex tree and then pruning it back to select a subtree of the original tree (James et al. 2015, 307-308). It should be emphasized that these single decision trees are built more for the purpose of visualizing how tree-based methods work, rather than as a purely predictive means. This is, as mentioned, because single decision trees are prone to overfitting on the training data. Due to the fact that decision trees are in this thesis used as pedagogical tools, rather than for predictive accuracy, they have not been tuned optimally.

Random forest, bootstrapping, and bagging

Random forests is a type of non-parametric tree-based algorithm, and builds on a concept known as bagging (Boehmke and Greenwell 2019, 204). As a tree-based model, random forests build on the same principles as decision trees. In order to understand random forests, an understanding of bagging is required. Bagging gets its name from the terms ‘bootstrapping’ and ‘aggregating’ (Boehmke and Greenwell 2019, 204). Bootstrapping is a form of random sampling, where the random sample is the same size as the original dataset (Boehmke and Greenwell 2019, 26). In the case of this thesis, the dataset that undergoes bootstrapping is the training dataset. At each step of creating the new dataset, a random row of observations is selected. What makes bootstrapping unique is that it does not exclude a row from being picked once it has been picked (Boehmke and Greenwell 2019, 26). In practice, this means that the new bootstrapped dataset includes duplicates of some of the rows, as it always creates a dataset as large as the original dataset. Observations that are not included in the bootstrapped dataset are called *out-of-bag* (*OOB*). The OOB samples are used for validating the models (Boehmke and Greenwell 2019, 26). Aggregation is applied to bootstrapping by taking the ‘base learner’ – the prediction model built on the original training data – and applying this algorithm to multiple bootstrapped samples. Each observation is then classified by taking the plurality ‘vote’ of each bootstrapped prediction, hence aggregation (Boehmke and Greenwell 2019, 192). The main reason for using bootstrapped samples is to reduce the variance of the model (Boehmke and Greenwell 2019, 204).

Random forests build upon bagging by introducing another layer of randomness. This is done through what is known as *split variable randomization* (Boehmke and Greenwell 2019, 204). This is a process where only a random set of variables are considered at each split of the tree-building process. This essentially forces some variables to be excluded at each split, decreasing the correlation between trees (Boehmke and Greenwell 2019, 204). Random forest has become popular due to predicting out-of-bag samples well (Boehmke and Greenwell 2019, 205).

Random forest has a wide array of hyperparameters that may be tuned. The forest needs a certain number of trees in order for the results to be robust and to stabilize the error rate. However, number of trees is not technically a hyperparameter. The optimal number of trees will vary based on the values of the other hyperparameters, but more trees is generally better in terms of error rates and the stability of variable importance (Boehmke and Greenwell 2019, 206). This does, however, come at the cost of high computational times – time increasing linearly for each tree added to the forest (Boehmke and Greenwell 2019, 206). The variables considered at each split is, as discussed, random and limited. One can, however, tune the number of variables that are considered at each split (*mtry*). A high number of variables considered at each split would increase the correlation between trees compared to a low value for *mtry* because the trees are less 'random' when they have more variables to choose from at each split (Boehmke and Greenwell 2019, 207).

A third hyperparameter is complexity, which may be an inclusion of multiple hyperparameters such as tree depth, minimum number of observations in each node or max number of terminal nodes (Boehmke and Greenwell 2019, 207-208). Finally one can tune the hyperparameters sampling scheme and splitting rule (Boehmke and Greenwell 2019, 206). The default sampling scheme is bootstrapping, where the bootstrapped sample is as large as the training set. However, one can adjust sample size. A smaller size of each bootstrapped dataset will lead to a higher diversity of trees, reducing correlation between the trees (Boehmke and Greenwell 2019, 207). The default splitting rule is to choose the split that minimizes Gini impurity, a measure of classification inaccuracy. This default rule favours variables with a high number of values (continuous variables or variables with a high number of categories). One can minimize this bias with a method called conditional inference trees, but Boehmke and Greenwell argue this has yet to be proven superior to the traditional splitting rule (2019, 209-210).

Gradient boosting machines

Gradient boosting machine (GBM) is also a non-parametric machine learning algorithm that is usually applied to decision trees (Boehmke and Greenwell 2019, 221-222). Again in similarity to random forests, GBMs combine multiple models (Boehmke and Greenwell 2019, 222). GBMs build trees based on previous trees, improving on the original tree at each step. This contrasts with random forest, which builds each tree independent of the other trees. The main idea is that GBMs sequentially try to improve upon the previous model by addressing its weaknesses (Boehmke and Greenwell 2019, 222-223).

The way in which GBMs build on the previous tree is through attempting to explain the residuals. This is done by fitting a single decision tree (F_1) to the training data. F_1 is the base learner upon which the next trees/algorithms are built on. Subsequently, a new decision tree (h_1) is fit to the residuals of F_1. This new tree is then added to the model, such that $F_2$is a combination of F_1 and h_1. Following this, a third tree is created, attempting to explain the residuals of F_2, and finally added to the function. The process of fitting trees on the residuals of the previous function continues until a mechanism, for instance cross-validation, stops it (Boehmke and Greenwell 2019, 223).

GBMs can also be tuned by an array of hyperparameters. As with random forest, one can tune the number of trees. However, having a large number of trees does not give the same benefit as when doing random forest. The sequential aspect of GBMs means that fitting more trees increases the risk of overfitting, simply because each tree attempts to explain the residuals of the previous. (Boehmke and Greenwell 2019, 227). Therefore, optimally tuned GBMs generally have fewer trees than random forest. Another hyperparameter to tune is the learning rate. In short, the learning rate decides the extent to which each tree contributes to the final model (Boehmke and Greenwell 2019, 227). One should also tune the tree depth and minimum observations in each terminal node, as one does in random forest and decision trees (Boehmke and Greenwell 2019, 227).

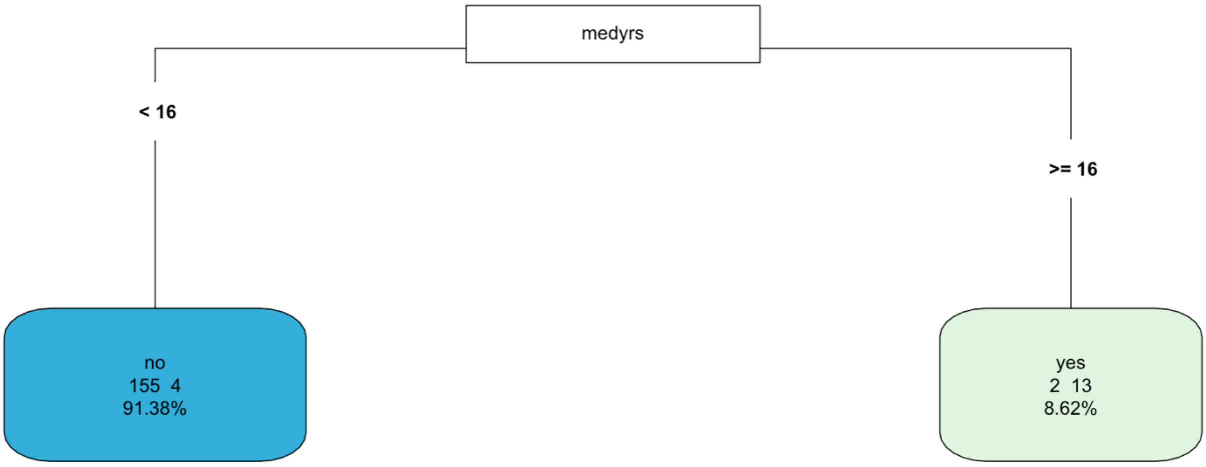
3.4.4 Model building

In this section I describe how the different machine learning models in this thesis have been built. Particularly, how the models have been tuned in order to create the best model for each method, as well as a description of the three different sets of input-variables used for each method. This section serves as a background for the results chapter, to not mix methodological choices with the presentation of each model’s predictive ability.

Different datasets and omission of variables

One of the most significant choices in model-building in this thesis is the omission of certain variables. The models are built on three datasets, each smaller than the previous. The first dataset includes all input variables described in section 3.3.1. For an overview of variables and datasets see appendix A.2. The second dataset is the same as the full dataset, except that the variable *medyrs* has been excluded. There are multiple reasons for its exclusion in the second set of models. One reason is the fact that it mainly builds on empirical observations that are quite obvious: every state that has legalized recreational marijuana has legalized marijuana for medical use prior. This is not to say that this is uninteresting or not useful, rather, it is obvious. Furthermore, this variable dominated the models in which it was included in terms of its variable importance score. As illustrated in figure 3.7, the first decision tree is based completely on each state's value for the variable *medyrs*. Another reason is how it is measured. As discussed in 3.3.1, observations that legalized medically that year and observations that have yet to legalize marijuana recreationally are coded as 0. Giving two very different types of observations the same value is problematic. *medyrs* and its contribution to the model's predictive accuracy is discussed further in the discussion section of this thesis.

Figure 3.7 *Decision tree 1*



The third, and last, dataset, is one where all variables directly related to marijuana legalization, decriminalization and medicalization have been excluded – except for the public opinion variable *grass*. There are several reasons why I run the models on three datasets. One of them is that one can argue that medicalization is a prerequisite for recreational legalization, at least empirically speaking. Removing such variables will also shine a light on the extent to which these are prerequisites for predicting marijuana legalization accurately. Marijuana variables were shown to be important, but to know exactly how important one needs to perform interpretive methods (such as variable importance) *and* remove these variables entirely from the models. Exploring the extent to which the models can predict without any knowledge of marijuana laws such as decriminalization truly points to the importance of the models knowing whether a state has decriminalized, medicalized or if its neighbouring states have legalized recreationally. *grass* has not been omitted as it is the main independent variable both in terms of theory and in light of autoMrP. In hindsight, a model without *grass* would be useful to gauge its importance. In short, the marijuana variables have been removed as it can be argued that they give too much information to the models, producing strong yet theoretically problematic models.

Shrinkage – tuning the most optimal models

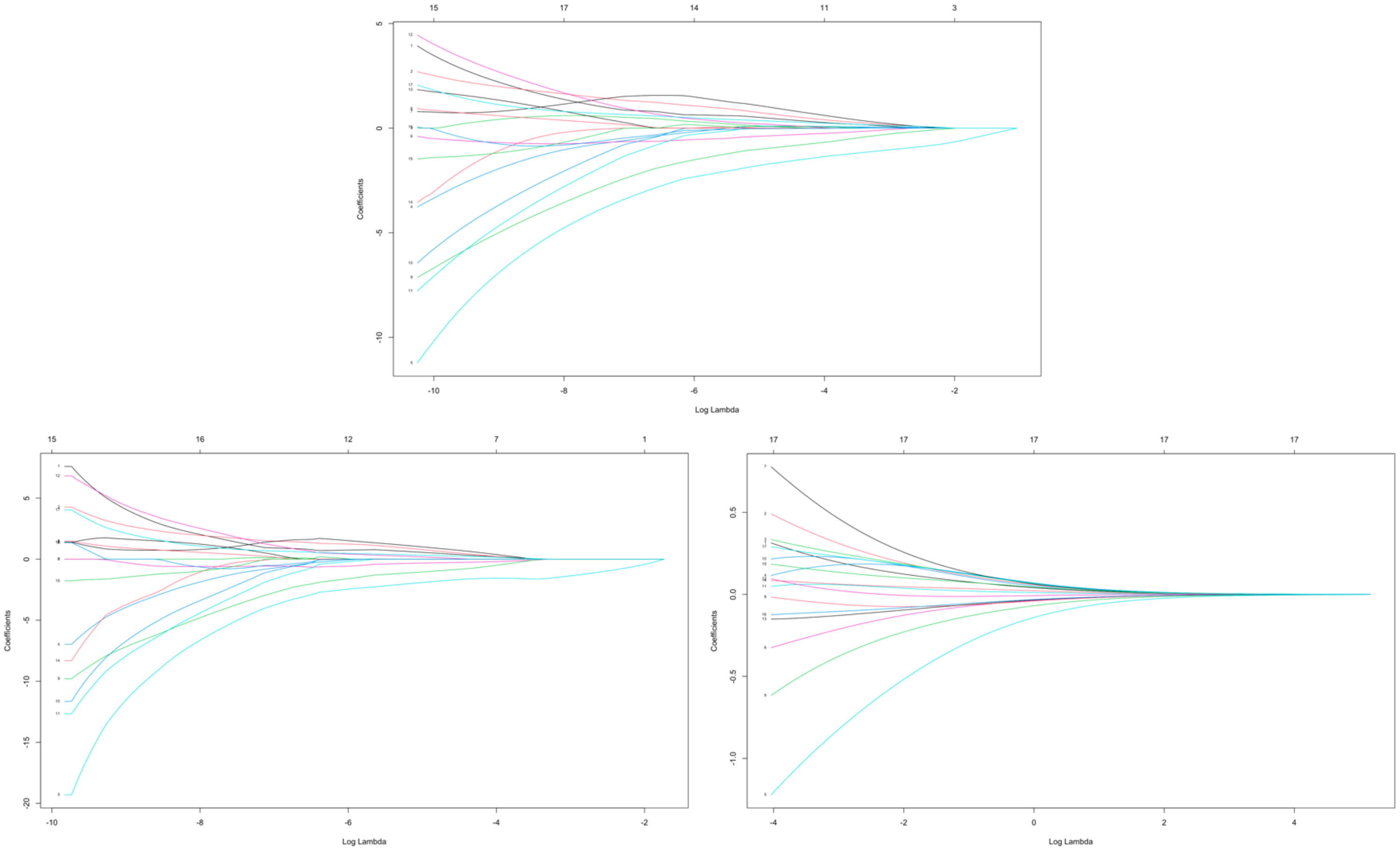
As discussed earlier in this chapter, K-fold cross-validation is used to tune the most optimal model, despite its issues of bias with smaller datasets (Vabalas et al. 2019, 17; Boehmke and Greenwell 2019, 26). Optimal shrinkage parameter is the one that leads to the lowest misclassification rate. One hundred different lambda values were tested using K-fold CV. According to K-fold CV, the optimal ridge model with the first dataset (*medyrs* excluded) had a λ (lambda, shrinkage parameter) of 0.030. ElasticNet and lasso had the lowest k-fold CV error rate with a λ of respectively 0.029 and 0.033. The most optimal ridge model had an error rate of roughly 7.5%, whilst EN and lasso had the same error rate of roughly 6%.

Figures 3.8 visualize the shrinkage penalty for ridge, ElasticNet and lasso. These graphs show how the methods differ in the extent to which they shrink the coefficients towards zero.

The second ridge model (*medyrs* excluded) had an optimal λ of 0.04 and a classification error of 9%. The second ElasticNet and lasso models had the optimal lambda value of 0.022 and 0.009, respectively, and a CV classification error rate of 9% and 6%.

The third ridge model (all marijuana variables excluded, except *grass*) had an optimal shrinkage parameter of 0.03, and an error rate of slightly less than 9%, whilst the optimal lambda for EN was 0.005, producing an error rate of 7%. Finally, the last lasso model had an optimal lambda of 0.01 and a misclassification rate of less than 6%. Thus, the lasso algorithm produced the optimal CV results on all three datasets. Visualizations of the second and third models' shrinkage penalty is not included in this text, as the visualization of the first model's shrinkage penalty is sufficient for comparing each of the three shrinkage methods' strictness.

Figure 3.8 *Shrinkage penalties – first models*

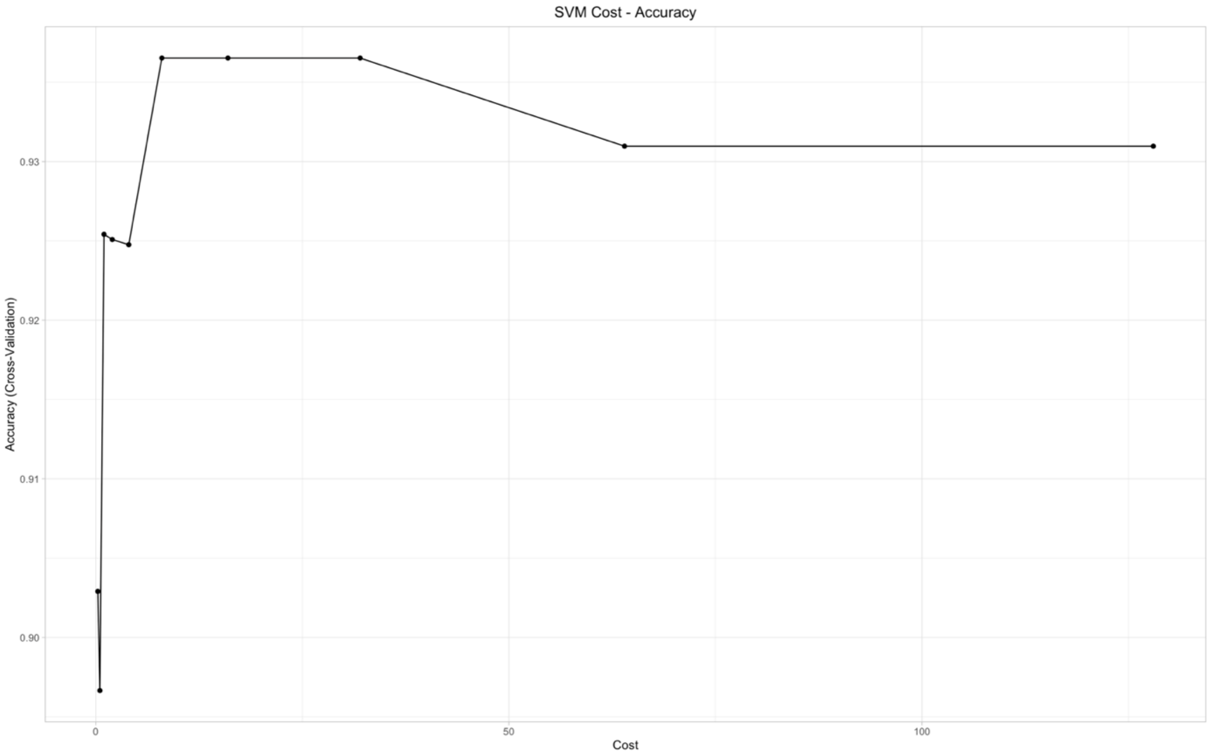


Top: Ridge shrinkage penalty. Left: ElasticNet shrinkage penalty. Right: Lasso shrinkage penalty

Support Vector Machines tuning

The main hyperparameter one can tune in SVMs are the kernel function, as discussed earlier in this chapter. For all three SVM models, the radial basis function was found to be the optimal using K-fold CV. Furthermore, K-fold CV found 8 to be the optimal cost for the first SVM model, and 0.046 as the optimal sigma value. This produced a cross-validated accuracy of 93.7%, as seen in figure 3.9. The second model had an optimal cost-parameter of 64, and sigma of 0.046. This also produced a CV accuracy of 93.7%. The third model had the same optimal cost-parameter as the second model, but the optimal sigma-value was 0.060. These parameters produced a CV accuracy of 93.1%, quite similar to the first and second models. When tuning SVMs, one not only gets an output of CV-accuracy, but one also gets a metric of true/false positives and true/false negatives. Considering the aim of this thesis is to predict which states have legalized marijuana, the degree to which the models predict positives correctly has been prioritized. It should be noted that the cross-validated accuracy is *not* the same as the prediction accuracy of the test data, which is to be discussed in the results chapter.

Figure 3.9 *SVM Cost vs Accuracy (CV)*



Random Forest tuning

As with all other methods, there are three different random forest models. The main tuning-parameter for random forests is, as mentioned, the number of variables tried at each split (*mtry*). This is the only variable that has been tuned using cross-validation. The *mtry* that produced the lowest out of bag (OOB) error for the first model was 4. CV showed 2 to be the optimal *mtry* for models two and three. All three random forests were built using 500 trees. After reaching a few hundred trees, the number of trees did not impact OOB error.

Gradient Boosting Machines tuning

More hyperparameters for the GBM models have been tuned than for any other algorithm in this thesis. The following values for the hyperparameters produced the lowest CV error rate for the first GBM model: 85 number of trees, an interaction depth of 1, a minimum number of 5 observations in each node and a shrinkage coefficient of 0.2. For the second model, the corresponding values 140, 5, 15 and 0.2. For the third model, the corresponding values were 84, 3, 5 and 0.05.

4. Results

This chapter presents the results from all the prediction models. This is largely done by presenting the accuracy of each model in predicting negatives (recreational marijuana not legal) and positives (marijuana legal for recreational use). How important each variable is for the different prediction models is presented as well. The interaction between variables is also visualized and discussed using interaction plots. I furthermore present the variable coefficients in the shrinkage models are also presented. The chapter is split into four parts: prediction accuracy, variable importance, interaction plots, and variable coefficients. As already mentioned, logistic regression models have also been run, and the result of these regression models are in appendix A.1. The logistic regression models are, however, not discussed in this chapter.

Since the prediction problem is classificatory, the accuracy/results of each model is presented in what is known as a confusion matrix. Table 4.1 shows a confusion matrix. This confusion matrix does not show the actual results. Rather, table 4.1 is included for illustrative purposes. The middle of the confusion matrix shows how many of the observations fall into each of the following categories: true positives (TP, 95 observations), false positives (FP, 5 observations), true negatives (TN, 45 observations) and false negatives (FN, 5 observations). Classifying negatives incorrectly (false positive) is known as a type I error, whilst classifying positives incorrectly (false negative) is known as a type II error (James et al. 2015, 148). Observations that have legalized marijuana are referred to as positives, whilst those that have not are referred to as negatives.

Table 4.1 *Confusion Matrix (example)*

	True				
Predicted	No	Yes	Total	Accuracy	
No	45	5	50	No	90%
Yes	5	95	100	Yes	95%
Total	50	100	150	Total	93%

Variable importance plots (VIPs) are presented for most models, as this is applicable for all methods and essentially what makes my models interpretive, and not ‘black-box’ models. The variable importance plots presented are mostly of the best predictive models. However, at least one VIP per dataset is included to get a more nuanced picture of which variables are important.

Interaction plots are only presented for the three best models. The final chapter discusses why some models/variables predict better than others.

4.1 Predictions and model accuracy

Tables 4.2, 4.3, and 4.4 show the prediction results from all prediction models. There are three models for each algorithm, one for each dataset. Models with '1' at the end use all the variables, models with a '2' at the end use all variables except *medyrs*, and models with a '3' at the end use all the variables except *medyrs*, *med*, *dec* and *diffusion*.

Table 4.2 *Shrinkage Confusion Matrices*

Ridge 1		**True**					**Ridge 2**		**True**					**Ridge 3**		**True**				
	Predicted	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy	
	No	66	3	69	No	100%		No	66	4	70	No	100%		No	66	5	71	No	100%
	Yes	0	4	4	Yes	57.2%		Yes	0	3	3	Yes	42.8%		Yes	0	2	2	Yes	28.6%
	Total	66	7	73	Total	95.8%		Total	66	7	73	Total	94.5%		Total	66	7	73	Total	93.2%
ElasticNet 1		True					**ElasticNet 2**		**True**					**ElasticNet 3**		**True**				
	Predicted	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy	
	No	66	4	70	No	100%		No	66	4	70	No	100%		No	66	5	71	No	100%
	Yes	0	3	3	Yes	42.8%		Yes	0	3	3	Yes	42.8%		Yes	0	2	2	Yes	28.6%
	Total	66	7	73	Total	94.5%		Total	66	7	73	Total	94.5%		Total	66	7	73	Total	93.2%
Lasso 1		True					**Lasso 2**		**True**					**Lasso 3**		**True**				
	Predicted	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy	
	No	66	4	70	No	100%		No	66	4	70	No	100%		No	66	5	71	No	100%
	Yes	0	3	3	Yes	42.8%		Yes	0	3	3	Yes	42.8%		Yes	0	2	2	Yes	28.6%
	Total	66	7	73	Total	94.5%		Total	66	7	73	Total	94.5%		Total	66	7	73	Total	93.2%

Table 4.3 *SVM, DT and RF Confusion Matrices*

SVM 1		**True**					**DT 1**		**True**					**RF 1**		**True**				
	Predicted	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy	
	No	66	2	68	No	100%		No	64	3	67	No	96.9%		No	66	3	69	No	100%
	Yes	0	5	5	Yes	71.4%		Yes	2	4	6	Yes	57.1%		Yes	0	4	4	Yes	57.1%
	Total	66	7	73	Total	97.3%		Total	66	7	73	Total	93.2%		Total	66	7	73	Total	95.9%
SVM 2		**True**					**DT 2**		**True**					**RF 2**		**True**				
	Predicted	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy	
	No	66	2	68	No	100%		No	62	2	64	No	93.9%		No	66	4	70	No	100%
	Yes	0	5	5	Yes	71.4%		Yes	4	5	9	Yes	71.4%		Yes	0	3	3	Yes	42.9%
	Total	66	7	73	Total	97.3%		Total	66	7	73	Total	91.8%		Total	66	7	73	Total	94.5%
SVM 3		**True**					**DT 3**		**True**					**RF 3**		**True**				
	Predicted	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy	
	No	64	2	66	No	96.9%		No	62	2	64	No	93.9%		No	66	4	70	No	100%
	Yes	2	5	7	Yes	71.4%		Yes	4	5	9	Yes	71.4%		Yes	0	3	3	Yes	42.9%
	Total	66	7	73	Total	94.5%		Total	66	7	73	Total	91.8%		Total	66	7	73	Total	94.5%

Table 4.4 *Boosting Confusion Matrices*

GBM 1		**True**					**GBM 2**		**True**					**GBM 3**		**True**				
	Predicted	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy			**Predicted**	No	Yes	Total	Accuracy	
	No	66	3	69	No	100%		No	66	3	69	No	100%		No	66	2	68	No	100%
	Yes	0	4	4	Yes	57.1%		Yes	0	4	4	Yes	57.1%		Yes	0	5	5	Yes	71.2%
	Total	66	7	73	Total	95.9%		Total	66	7	73	Total	95.9%		Total	66	7	73	Total	97.3%

Most prediction algorithms predict at least as good with *medyrs* as an input variable compared to when *medyrs* is excluded. For most methods, the models with *medyrs* predict better than models without *medyrs*. This indicates that the total number of years a state has implemented medically legal marijuana is an important and strong predictor of whether marijuana is legal for recreational purposes in the given state at a given point in time. More generally, there is a quite wide span of predictive accuracy between methods and models, particularly regarding accuracy in predicting true positives. The final shrinkage models (table 4.2), for instance, only predict two of the seven true positives in the test-dataset correctly. The SVM models (table 4.3), however, predict a total of five true positives correctly. The models range from 28-71% accuracy for predicting positives, 43-100% for predicting negatives, and 91-97% overall predictive accuracy.

Accuracy of shrinkage models

Looking closer at the shrinkage models in table 4.2, the methods largely predict similarly in terms of accuracy when given the same input variables. The only exception is Ridge 1, which predicts positives slightly better than Lasso 1 and ElasticNet 1. A slightly less strict shrinkage function is therefore better for predicting legal marijuana with the given input variables. When *medyrs* is included, it is thus important to not shrink the other input variables' coefficients too harshly, as they seem to help with predictive accuracy. It should be noted that the difference between Ridge 1 and Lasso/ElasticNet 1 is only the prediction of one observation. The variable coefficients in the first shrinkage models are illustrated in table 4.4, and further visualized in the variable importance plots, figure 4.1 to 4.4. These tables and graphs can be found on pages 75 and 69-70, respectively.

In terms of their actual accuracy, all shrinkage models have a 100% accuracy for the prediction of observations that have not legalized recreational marijuana. However, the prediction accuracy for positives is roughly 40-60% for the first models, 40% for the second models, and 30% for the third models. These models, in other words, do not predict the 'recreationally legal' class well. However, these numbers are significantly impacted by the correct prediction of single observations, as there are only seven positive observations in the whole test dataset. This is discussed further in the limitations sub-section of the discussion and conclusion chapter; however, it deserves mentioning here as well: the empirical reliability and validity of all models in this thesis is quite weak due to having few observations.

Accuracy of SVM models

Looking at the SVM models in table 4.3, the SVMs predict better than the shrinkage models. This may be due to their non-linear functional form assumption. All SVM models predict five out of the seven positive observations correctly (71.4%), but model 3 is not 100% accurate in classifying negatives. This indicates that the SVMs are slightly more sensitive and have a somewhat lower specificity than the shrinkage models – however, this is only true for the last SVM model. Sensitivity refers to how accurately the classifiers classify positive observations, whilst specificity refers to a models ability to classify negative observations correctly (Boehmke and Greenwell 2019, 35). Considering the aim of this thesis is first and foremost to assess the degree to which different methods can predict a state that has recreationally legal marijuana, SVMs fare better than the shrinkage models.

Accuracy of tree-based models

Looking at the tree-based models in table 4.4, the gradient boosting machine models have the highest overall prediction accuracy. The third GBM model is the strongest of all the tree-based models, with an overall prediction accuracy of 97.1% (100% for negatives, 71.4% for positives). This model is, in addition, the best model overall, sharing first place with SVM 1 and SVM 2. The fact that the three best models use three different datasets is noteworthy. What is striking about the tree-based models is how well the single decision trees predict. The single decision trees are first and foremost used as pedagogical tools to build a foundation for understanding how GBMs and Random Forests work. The single decision trees nonetheless predict the positives better than the random forest models – but their specificity is not as high. Why the single decision trees predict so well, despite not being tuned optimally, is difficult to pinpoint. It may be due to the low sample size of both the training data and particularly the test data.

Overall, the confusion matrices in tables 4.2, 4.3 and 4.4 show that all prediction models by and large predict quite well. Most models predict negatives quite well – a majority of the models have 100% prediction accuracy for negative observations. The sensitivity, however, is what distinguishes the different models and algorithms from one another in terms of prediction power. The accuracy in predicting positives correctly, as mentioned, ranges from 28% to 71%. The shrinkage methods have the highest type II error (false negatives), particularly the third shrinkage models. SVMs have the lowest overall type II error, and the best overall models (SVM 1 and 2), alongside GBM 3. To conclude this section, SVMs and GBMs produced the best models, with the single decision trees at a close third due to their surprisingly high true positive rate.

4.2 Variable Importance Plots – opening the black box

This section builds on the previous section by opening the 'black box' using variable importance plots. Importance-scores measure how important variables are for making accurate predictions. Variable importance plots are only presented for the best models within each method, in order for this section to not be repetitive. A total of 9 VIPs are included in the text (out of a total of 21 models). I consider variables that are influential in the best predictive models as the most important variables, in contrast to the variables that have a high degree of influence on the models that do not predict well.

Variable importance plots for four of the nine shrinkage models are presented in this thesis: Ridge 1, Ridge 2, ElasticNet 2 and Lasso 2. Ridge 1 was the best model. However, it included the somewhat problematic variable *medyrs*. The second models predict equally well, sharing second place amongst all the shrinkage models.

Figure 4.1 shows the variable importance for the first ridge model. The y-axis shows the names of the variables, and the x-axis measures the variable importance. One thing is quite evident from looking at figure 4.1: the total number of years a state has hade medically legal marijuana is an important predictor of whether or not a state has legalized marijuana for recreational use. Considering ElasticNet and lasso shrink coefficients closer to zero than ridge, the effect of *medyrs* becomes even more evident when looking at figure 4.2 and 4.3. However, one should not confuse variable importance with coefficients. Variable importance shows the degree to which variables are influential in predicting the output variable. The variable importance is not shrunk. The coefficients, on the other hand, are subject to penalization. If a variable has been shrunk to zero, its variable importance score will also be zero.

One can see in figure 4.1 that whether a state has the opportunity for ballot initiatives, the extent to which a state is considered 'highly religious' and the public opinion variable *grass* are also important predictors in this ridge model. The variable importance plot does not say anything about direction of effects. This is discussed in more detail when the coefficients of some of the shrinkage models are presented (table 4.5).

Figure 4.1 *Ridge 1: Variable Importance Plot*

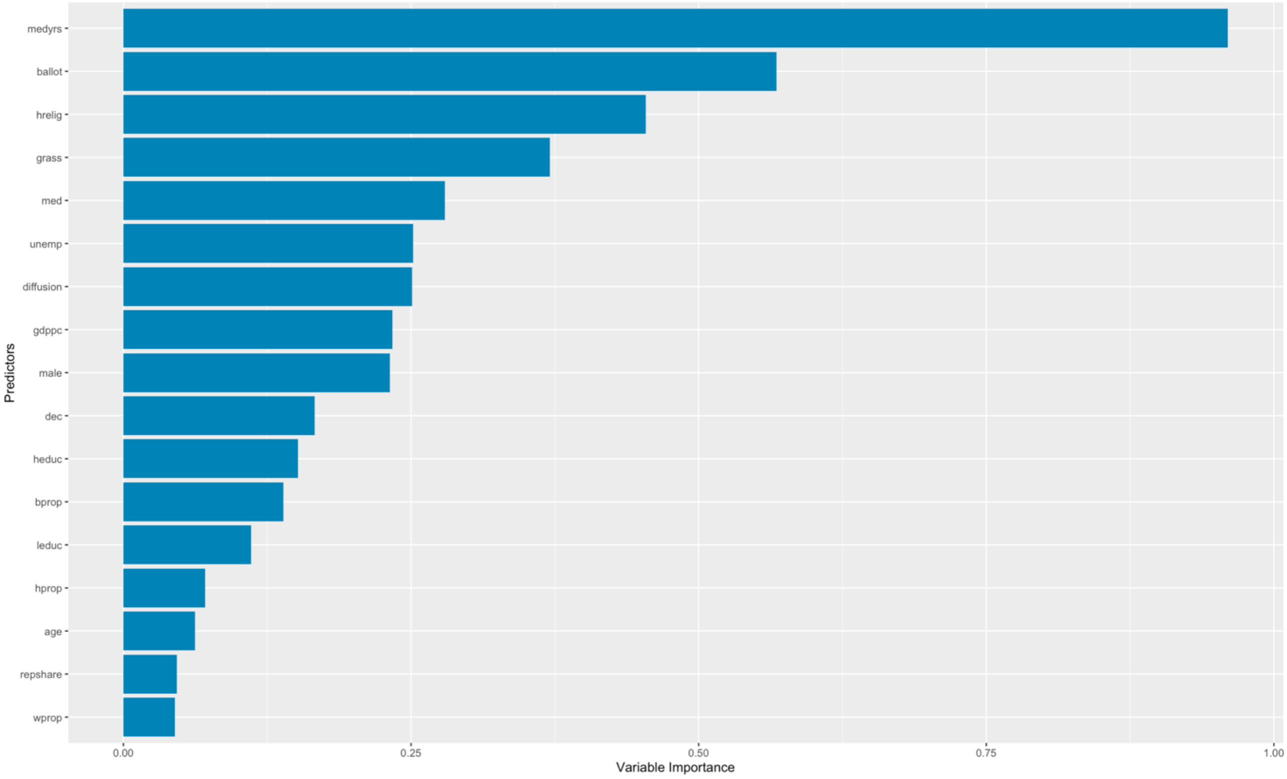


Figure 4.2 *Ridge 2: Variable Importance Plot (medyrs excluded)*

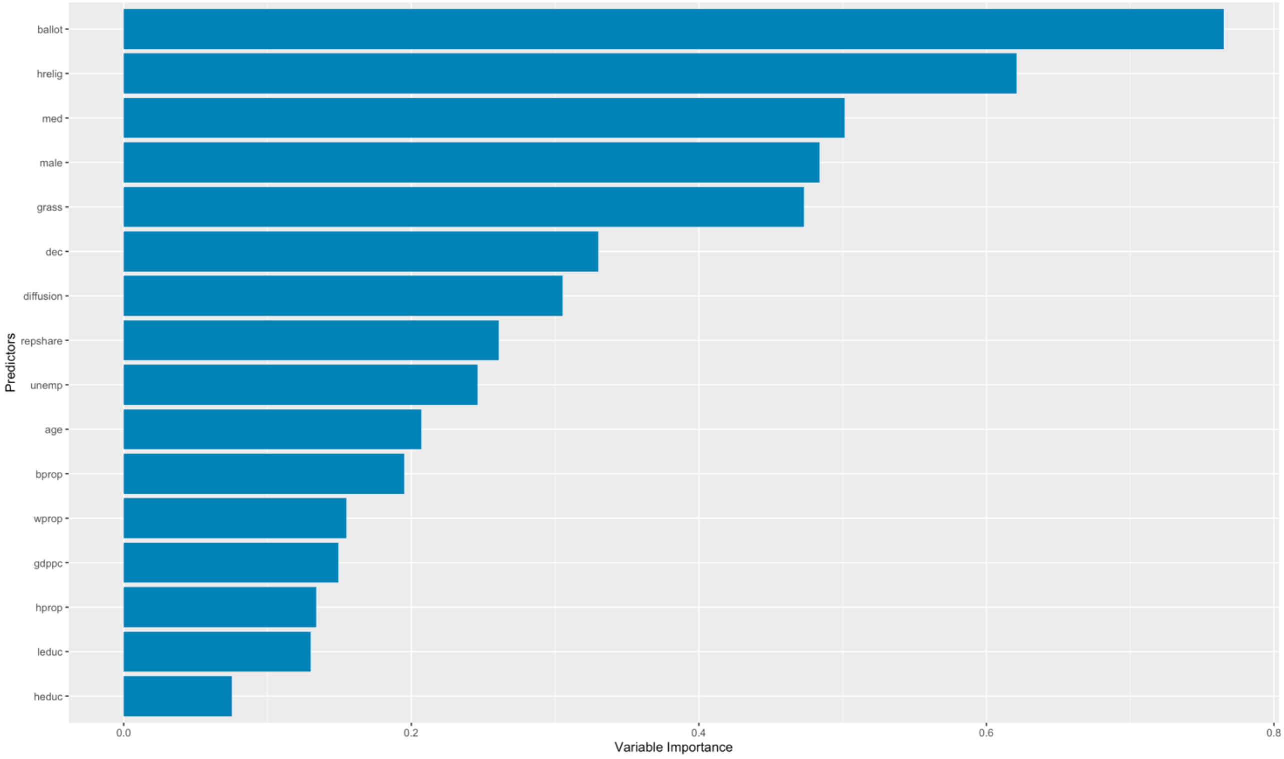


Figure 4.3 *ElasticNet 2: Variable Importance Plot (medyrs excluded)*

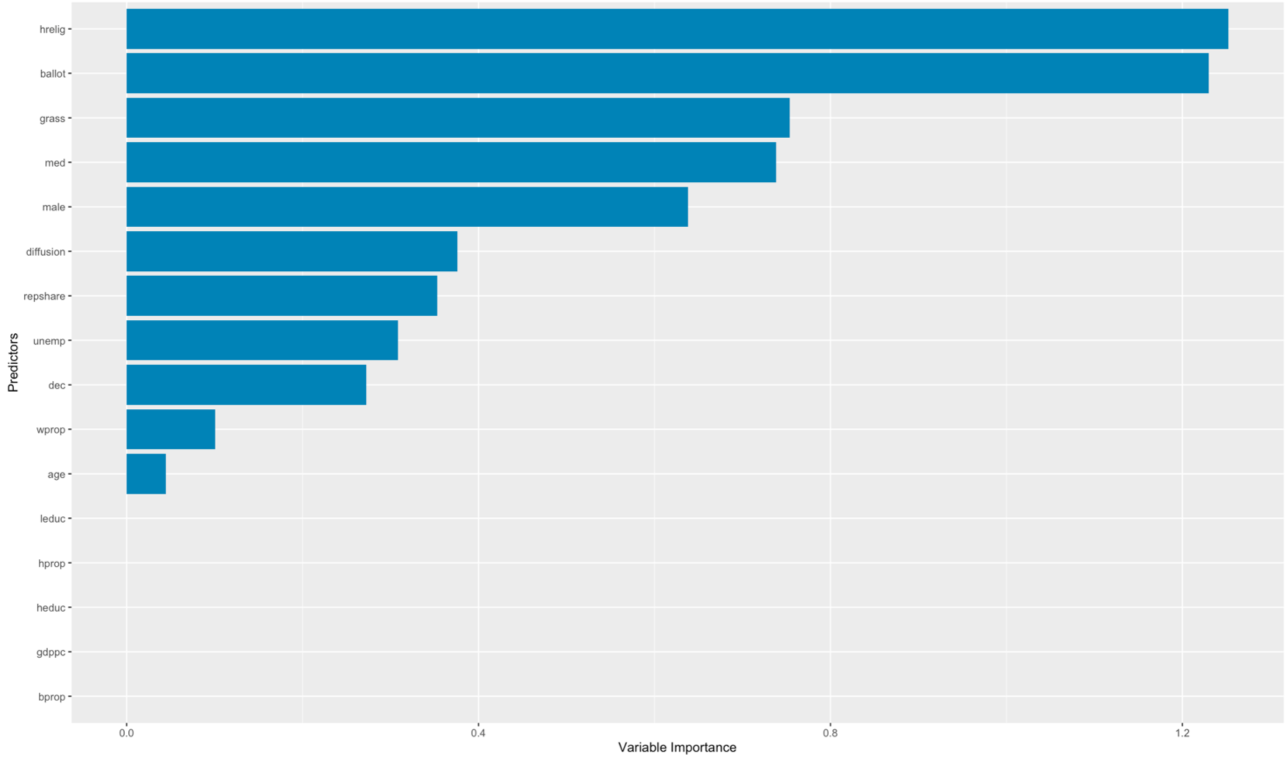
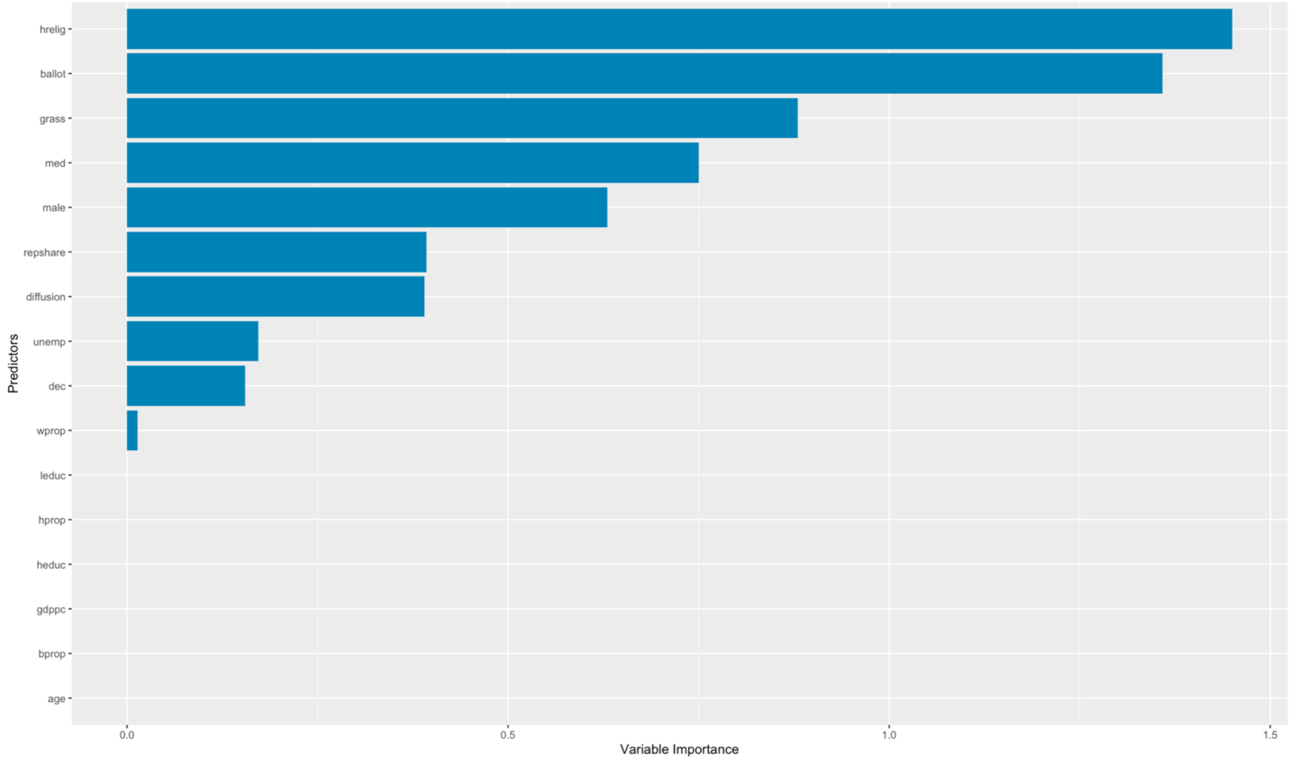
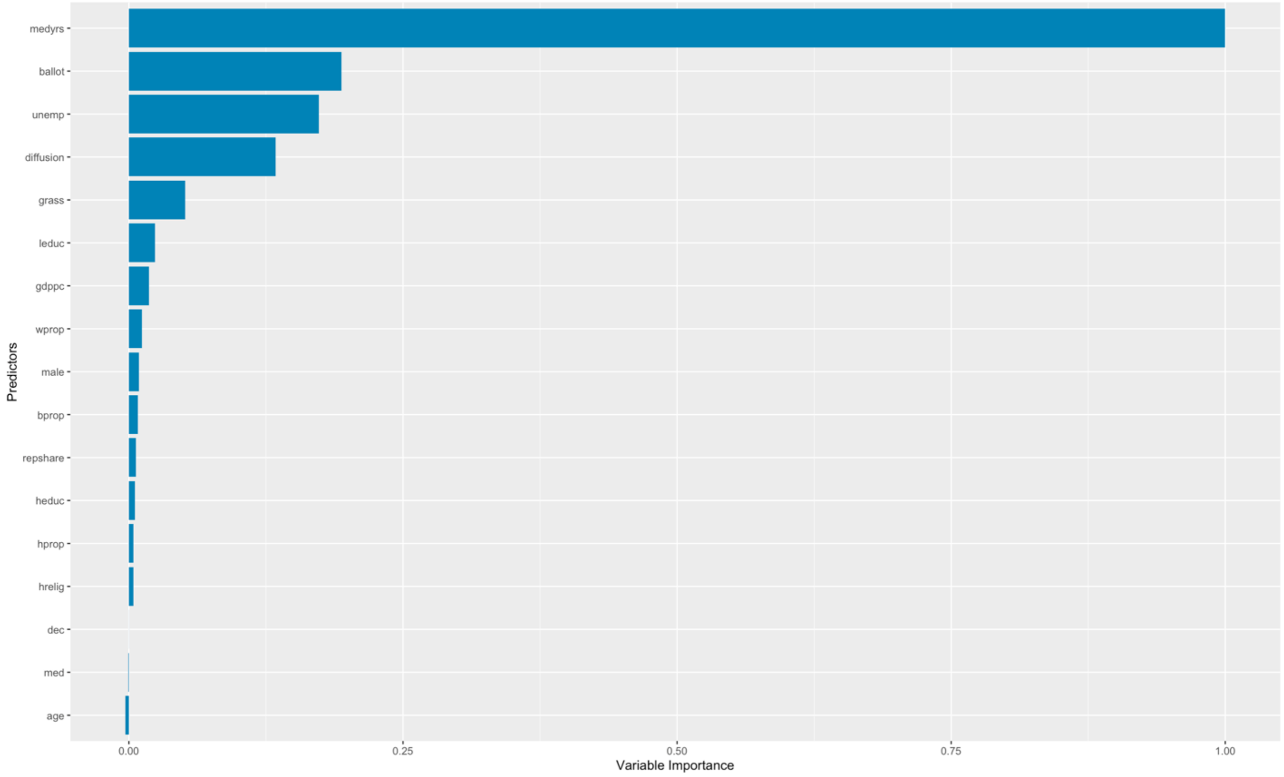


Figure 4.4 *Lasso 2: Variable Importance Plot (medyrs excluded)*



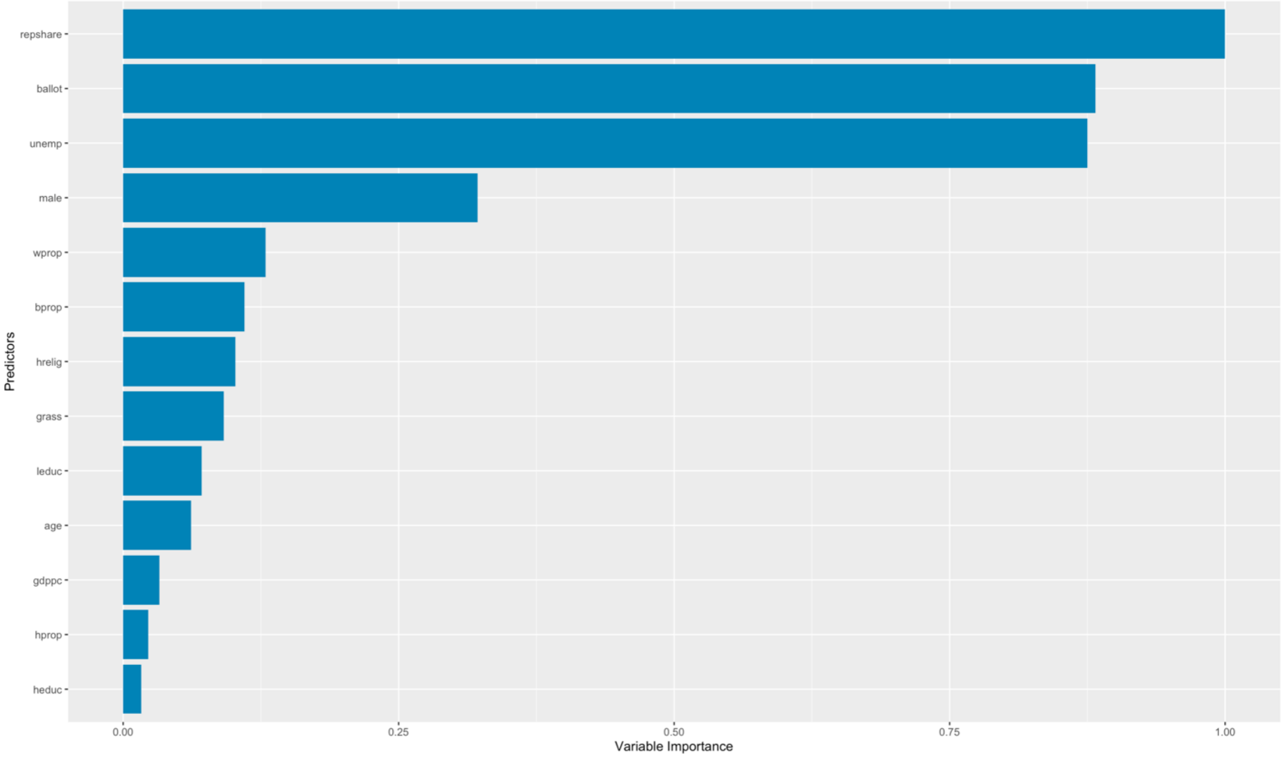
Figures 4.2, 4.3 and 4.4 illustrate the progression of shrinkage penalties from Ridge to ElasticNet and Lasso. Considering that Ridge never shrinks coefficients to zero, all variables have an importance-value of more than 0. In figure 4.3 and 4.4, on the other hand, one can see that some variables have an importance of 0. This is because their coefficients have been shrunk to 0. In terms of the variables' importance, *hrelig*, *ballot* and *grass* score high in these models as well. *med* has gained a higher importance after *medyrs* was removed. Overall, the variable importance plots indicate that medicalization, religiosity, ballot initiatives and public opinion are important predictors of RML. The worst/least important predictors in the shrinkage models are proportion of black/white/Hispanic people, proportion of population with a high/low education, median age and GDPPC. However, it should be noted that the shrinkage models are far from the best predictive models, as seen in the confusion matrices.

Figure 4.5 *SVM 1: Variable Importance Plot*



The SVMs variable importance plots differ somewhat from the shrinkage models' variable importance plots. However, *medyrs* also dominates the VIP of SVM 1 (figure 4.5), as is the case for the first shrinkage models. *medyrs* has a significantly higher importance-score than all other variables combined in SVM 1. Furthermore, *ballot* is also an important predictor in the SVM models, Unemployment rate, diffusion, share of males and share of people who voted for the republican candidate in the last presidential election are also important predictors of the dependent variable, in quite stark contrast to the shrinkage models. However, *repshare* is insignificant when *medyrs* is included. The biggest difference between the shrinkage models and SVMs is the fact that *hrelig* is not an important predictor in the SVM models. What is particularly noteworthy for these variable importance plots (figures 4.5 and 4.6) is that *med* and *age* have negative coefficients. These variables, in other words, decrease the model's predictive performance. This is the only VIP where variables are considered harmful for the prediction of the output variable

Figure 4.6 *SVM 3: Variable Importance Plot (marijuana variables excluded)*



Decision tree 2's variable importance plot shares some similarities with the shrinkage and SVM models' variable importance plots. *grass* scores quite high, as does *hrelig*, *ballot*, and *repshare*. In contrast to the other models, *unemp* is at the top as the most important variable predicting accurately. In the VIPs presented previously in this sub-section, *unemp* has been placed roughly in the middle, if not towards the bottom. In terms of its importance, the other method's evaluation of *unemp* as a predictor is more valid considering they are tuned and considering the weaknesses/bias of building a single tree.

As with most first models, RF 1 shows *medyrs* to be the most important variable by a large margin. This is followed by *hrelig, ballot* and *diffusion*. Nonetheless, these variables are not significantly more important than the variables at the middle of the graph (variables such as *bprop*, *leduc, gdppc* and *unemp*). The low importance of *grass* is quite noteworthy, considering it has been deemed as quite an important variable by most of the other models. GBM 3, one of the strongest predictive models, shows similar patterns as the other variable importance plots in this thesis. *hrelig*, *ballot* and *grass* are the three most important variables.

Figure 4.7 *DT 2: Variable Importance Plot (medyrs excluded)*

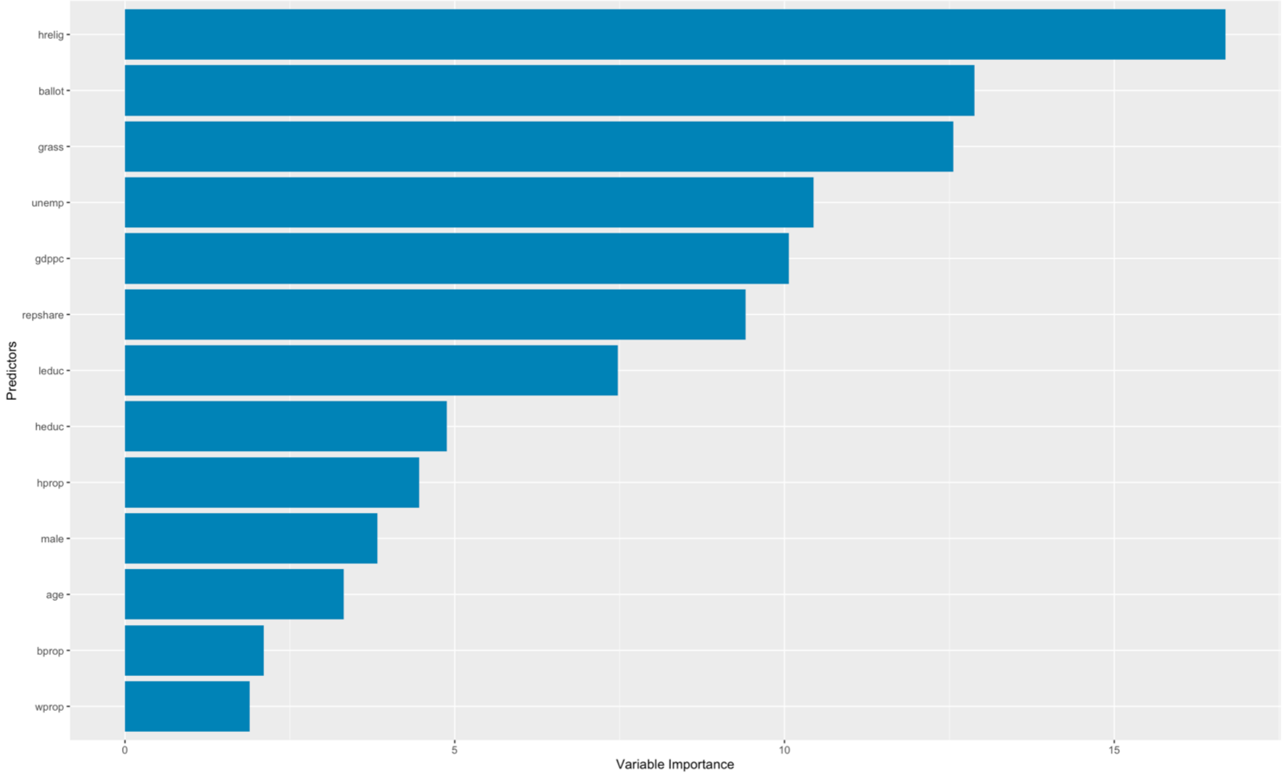


Figure 4.8 *RF 1: Variable Importance Plot*

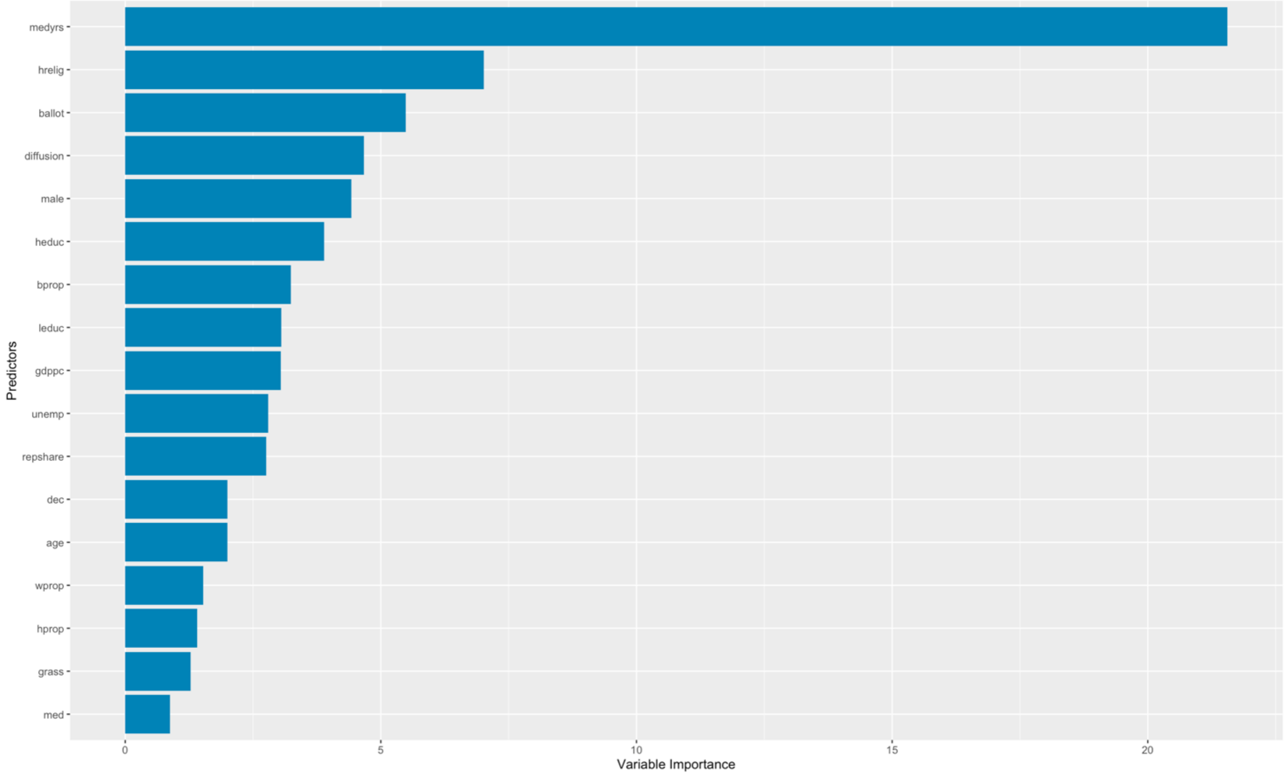
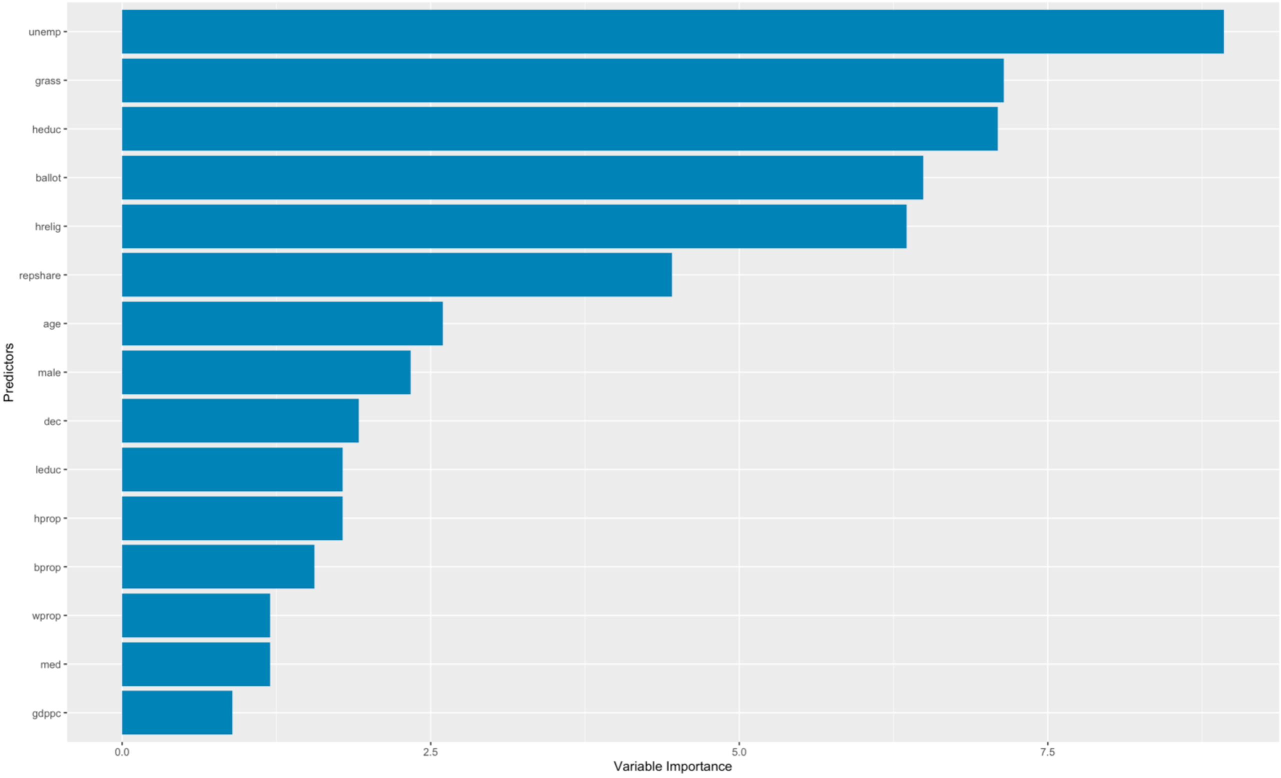


Figure 4.9 *GBM 3: Variable Importance Plot*



When considering the models' predictive abilities, the four most important variables (except for *medyrs*) are: *hrelig, ballot, grass* and *repshare*. Despite *repshare* not scoring high in terms of importance in most models it is the most important variable in one of the strongest models: SVM 3. I therefore consider it an important variable. Amongst the least important variables are the race-demographic variables (*bprop*, *wprop* and *hprop*), as well as median age and share of males. It is, however, difficult to determine which variables are important and which are not, considering the results from each model differ quite widely. *repshare*, for instance, is almost ranked last in Ridge 1, yet ranked first in SVM 3. One reason why there is such a difference between variable importance from model to model may be because variable may be correlated with each other. For instance, the inclusion of *medyrs* may decrease the importance of *grass* and *repshare* as they correlate with each other, as seen in figure A.1 in the appendix.

4.3 Interaction plots

Interaction plots are only presented for the three best models (SVM 1, SVM 2, and GBM 3). A total of 12 interaction plots have been made: one for each model, and one for each of the three most important variables (according to the VIPs) in each model. The interaction plots for each model, as well as the interaction plot for the most important variable in each model, are discussed and presented in this sub-section. The other interaction plots can be found in appendix A.3.

Figure 4.10 *SVM 1 Interaction Plot*

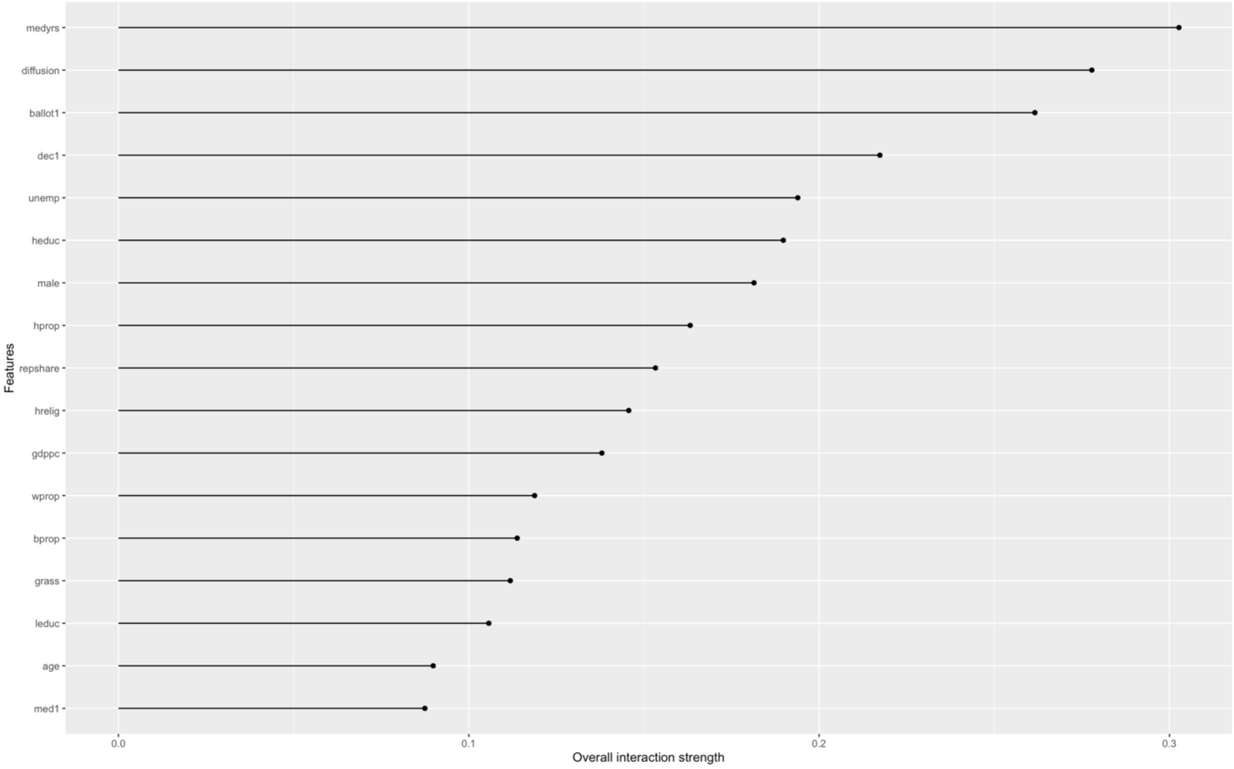


Figure 4.10 shows the interaction plot for the first SVM model. That is, the degree to which variables 'use' each other in making predictions. The y-axis shows each variable in the model and the x-axis shows the overall interaction strength. Figure 4.10 shows that most variables range between an interaction strength of 0.1 and 0.2. Of all the variables in the model, three stand out as depending on interactions the most: *medyrs*, *diffusion*, and *ballot*. *medyrs* is also the most important variable in the model, as seen in SVM 1's variable importance plot (figure 4.5). Overall, figure 4.10 does not really tell us much other than the general level of interactions in the model. Figure 4.11, however, illustrates the interaction strength of *medyrs* specifically. This figure is included due to *medyrs* being the variable with the highest interaction score, and due to it being the most important variable.

Figure 4.11 highlights two quite interesting mechanisms in the models: *medyrs* reliance on *diffusion* and *grass*. These are amongst the most important variables in SVM 1 according to figure 4.5, but far from as important as *medyrs*. However, *medyrs* relies on *diffusion* and *grass*, making these variables more significant than the variable importance plots suggest. In short, figure 4.11 tells us that in order for *medyrs* to predict the output variable it needs to 'know' about the share of neighboring states that have legalized recreational marijuana, and the public support towards legalization. However, directions are not specified in this figure, I am inclined to suggest that recreational marijuana is common in states that have medically legal marijuana and a high proportion of neighboring states that have legalized recreational marijuana. A high *medyrs* combined with a high public support is also typical for states that have recreationally legal marijuana.

Figure 4.11 *SVM 1 Interaction Plot (medyrs)*

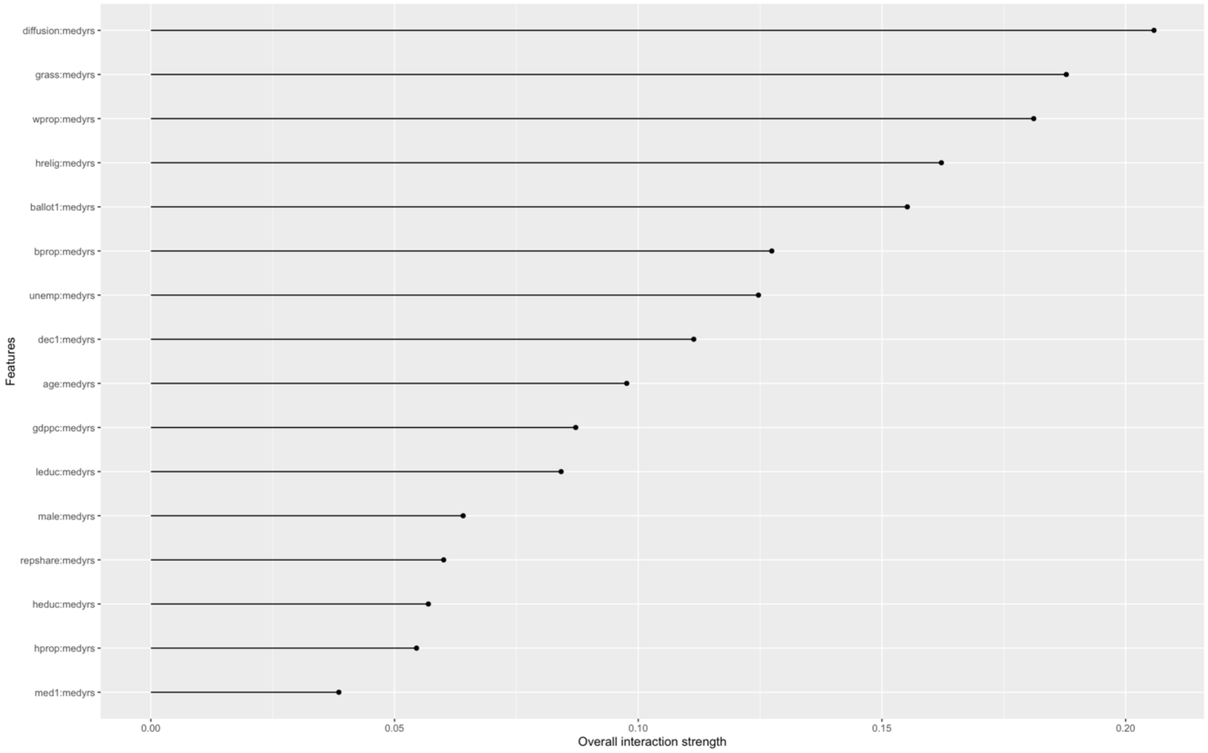


Figure 4.12 is quite different from figure 4.10. The overall interaction scores are less centered, and the average seems to be lower. Values range from practically 0 to almost 0.4. The three variables most reliant on interactions are *hrelig*, *ballot* and *grass.* Looking at figure 4.13, *hrelig* seems to be quite reliant on *ballot*, *unemp* and *grass.* The coefficients of the shrinkage models (table 4.5) and logistic regression model (table A.1) suggest that the models predict that a state has legal marijuana if it has a low *hrelig* score, and the opportunity for ballot initiatives.

Figure 4.14 shows the interaction strength in GBM 3 – the highest of the three best models. Values range from 0.05 to more than 0.6. *ballot* and *diffusion* show quite high interaction scores, the prior reaching more than 0.6. Turning to figure 4.15, one can see that *ballot* relies on a multitude of variables, including *hrelig*, *unemp*, *repshare*, *leduc* and *grass*. *hrelig* stands out as the variable *ballot* is the most dependent on, with a score of almost 0.4. This is quite similar to what figure 4.13 shows – *hrelig*'s reliance on *ballot*.

Figure 4.12 *SVM 2 Interaction Plot*

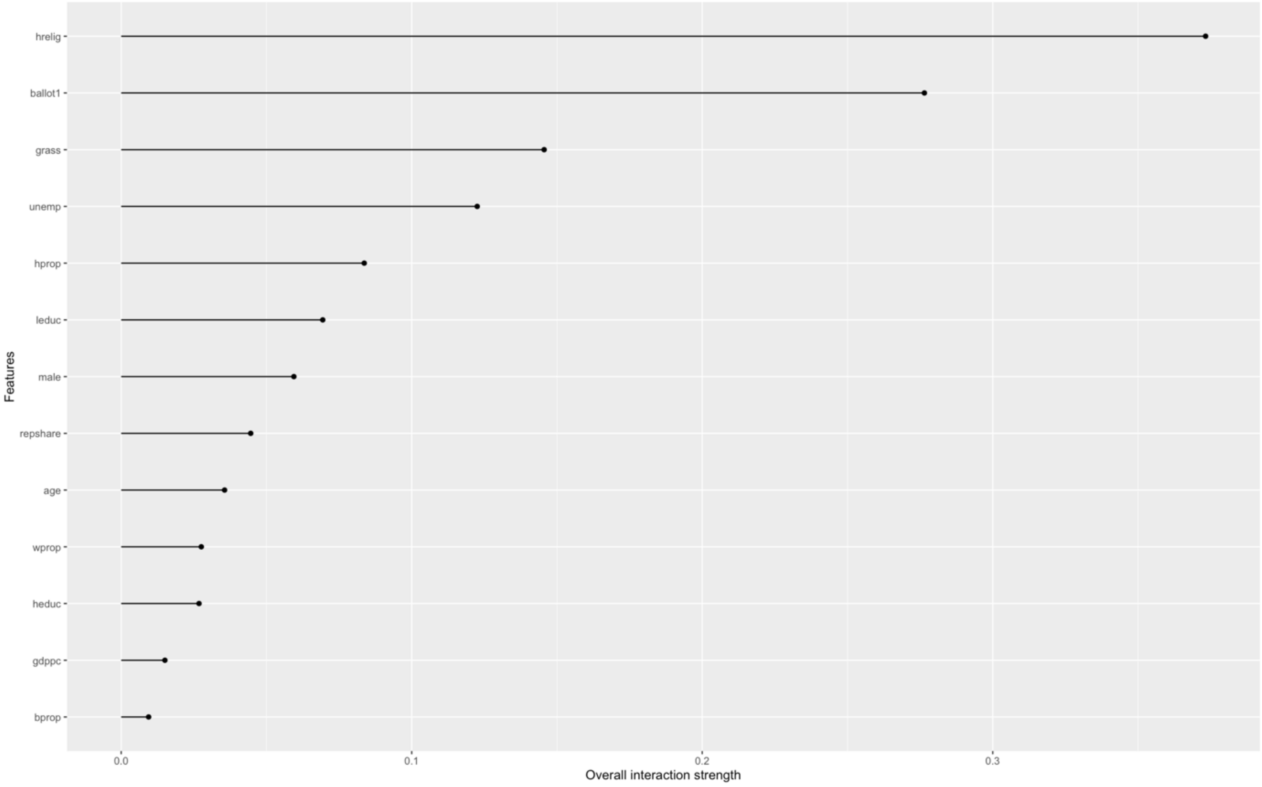


Figure 4.13 *SVM 2 Interaction Plot (hrelig)*

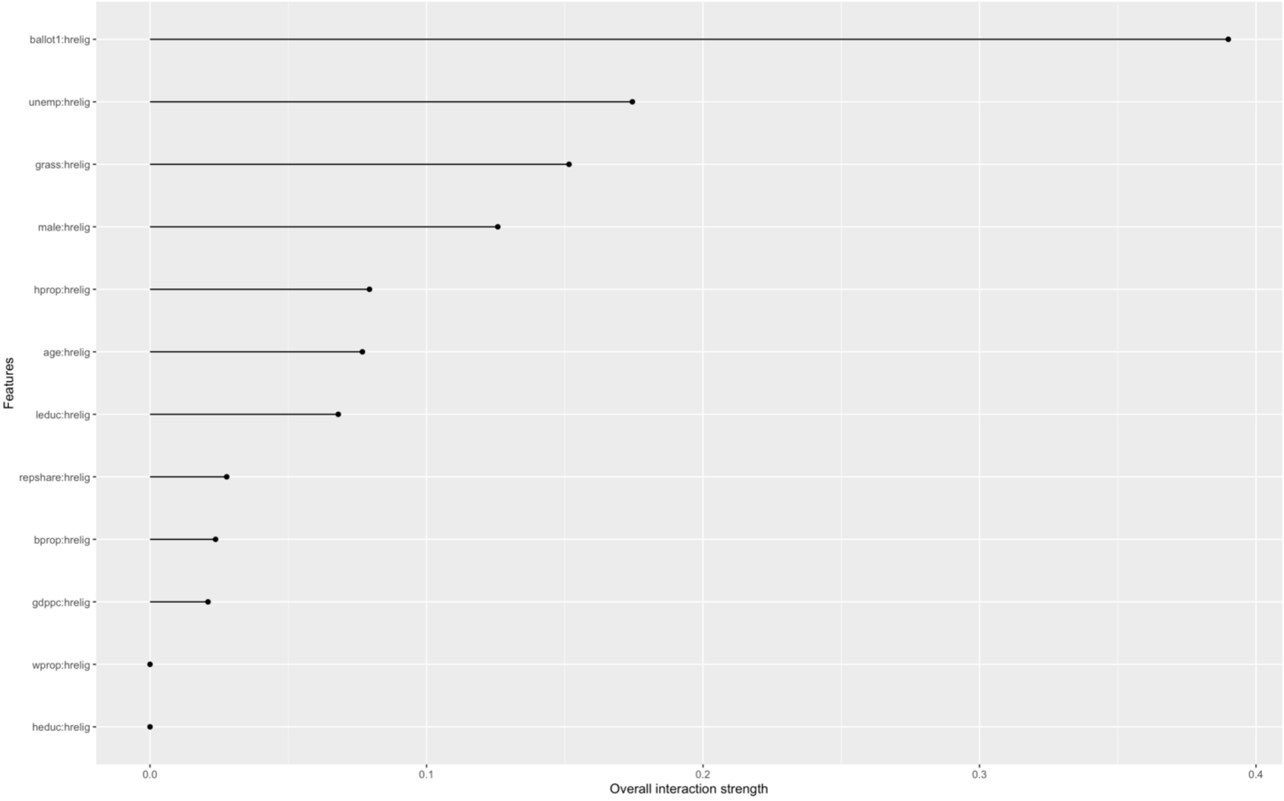


Figure 4.14 *GBM 3 Interaction Plot*

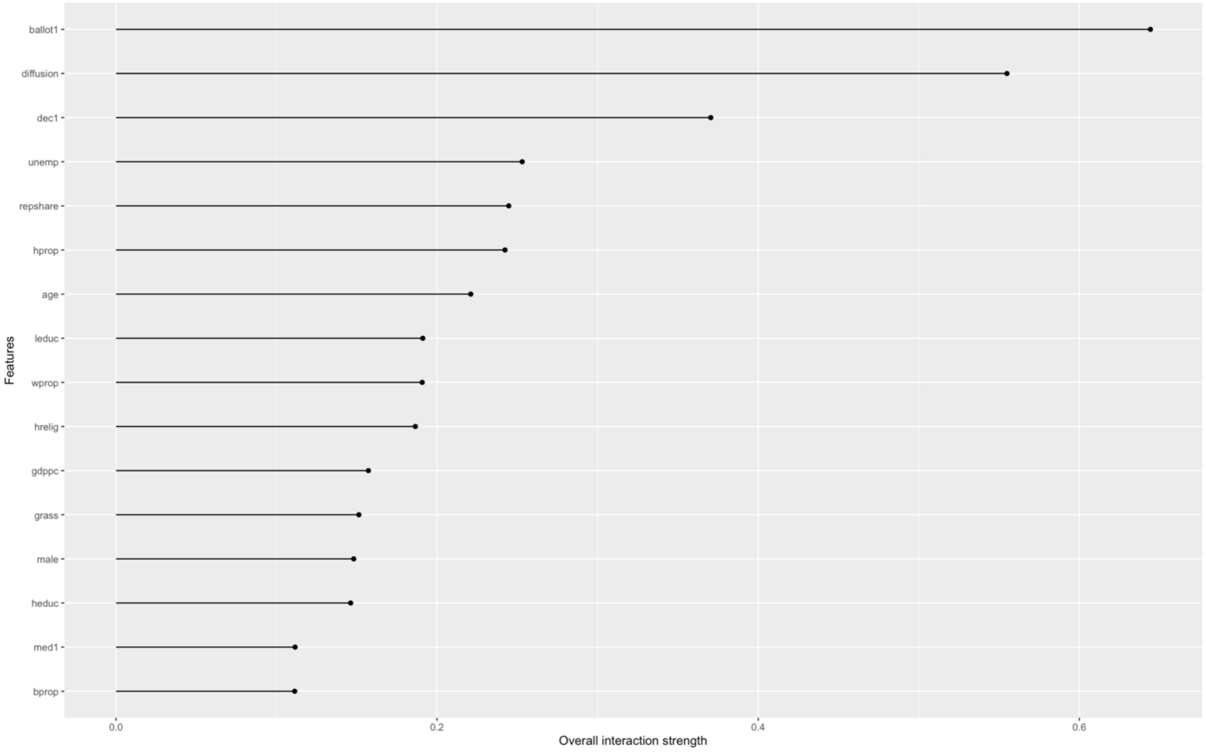
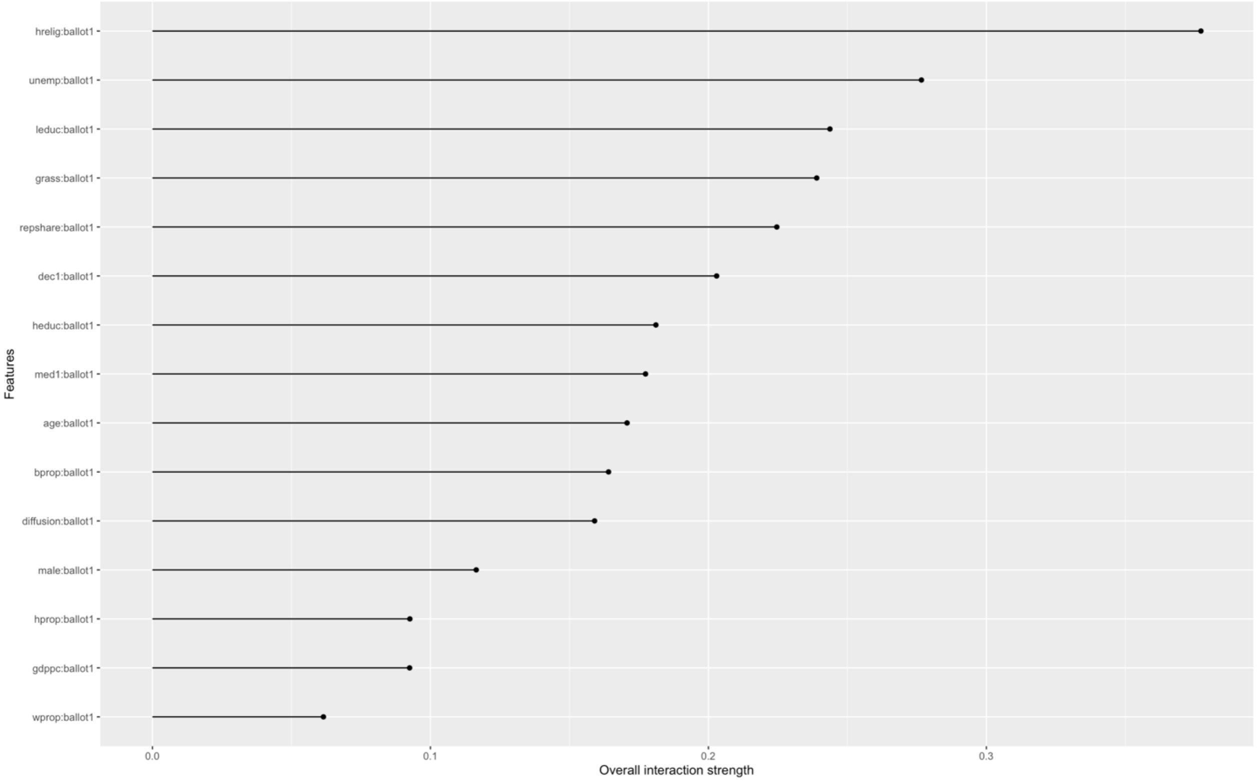


Figure 4.15 *GBM 3 Interaction Plot (ballot)*



4.4 Variable coefficients – Shrinkage

In this section I present the variable coefficients of the shrinkage models. This is done to see the how the variables affect/predict the output variable. I concentrate the shrinkage models since SVMs- and tree-based methods (DT, RF and GBM) do not have variable coefficients. However, since all models use the same datasets, I find it appropriate to generalize the coefficient directions form the shrinkage models to the other models.

Table 4.5 *Shrinkage variable coefficients*

Model 1				Model 2				Model 3			
Variable	Ridge	ElasticNet	Lasso	**Variable**	Ridge	ElasticNet	Lasso	**Variable**	Ridge	ElasticNet	Lasso
gdppc	0.23	0.19	0.00	*gdppc*	0.15	0.05	0.00	*gdppc*	0.13	0.00	0.00
grass	0.37	0.28	0.00	*grass*	0.42	0.51	0.85	*grass*	0.57	0.94	0.97
unemp	-0.25	-0.18	-0.05	*unemp*	-0.22	-0.15	-0.15	*unemp*	-0.26	-0.45	-0.18
ballot	0.57	0.41	0.00	*ballot*	0.66	0.75	1.29	*ballot*	0.81	1.49	1.12
repshare	-0.05	0.00	0.00	*repshare*	-0.23	-0.21	-0.37	*repshare*	-0.41	-0.74	-0.54
hrelig	-0.45	-0.49	-0.03	*hrelig*	-0.54	-0.81	-1.39	*hrelig*	-0.80	-1.70	-1.60
male	0.23	0.07	0.00	*male*	0.43	0.41	0.60	*male*	0.56	1.02	0.66
age	0.06	0.00	0.00	*age*	0.17	0.00	0.00	*age*	0.28	0.31	0.00
wprop	0.04	0.00	0.00	*wprop*	-0.13	0.00	0.00	*wprop*	-0.16	-0.11	0.00
bprop	-0.14	0.00	0.00	*bprop*	-0.18	0.00	0.00	*bprop*	-0.26	-0.17	0.00
hprop	0.07	0.00	0.00	*hprop*	0.12	0.00	0.00	*hprop*	0.28	0.44	0.24
heduc	0.15	0.02	0.00	*heduc*	0.08	0.00	0.00	*heduc*	0.19	0.08	0.00
leduc	-0.11	0.00	0.00	*leduc*	-0.12	0.00	0.00	*leduc*	-0.05	0.00	0.00
diffusion	0.25	0.19	0.03	*diffusion*	0.29	0.30	0.38				
med	0.28	0.00	0.00	*med*	0.45	0.52	0.71				
dec	0.17	0.03	0.00	*dec*	0.32	0.27	0.16				
medyrs	0.96	1.21	1.61								

However, considering the shrinkage models overall are amongst the poorest predictive models, one should be careful when generalizing the coefficients to the other models/methods. This is especially true as the other methods do not make a linear functional form assumption. GDPPC has positive coefficients in the models where it has not been shrunk to zero, indicating that states that have legal marijuana also have higher GDPPC. However, this variable was shown to be quite insignificant in most variable importance plots. Ballot initiative, one of the most important variables, has a positive coefficient in all models, and is quite large – indicating that having ballot initiative is typical for state that have legalized marijuana. Unemployment, *repshare*, *hrelig*, *bprop* and *leduc* have negative coefficients in all models. Most of these relationships are as theorized: religiosity, low education, and being republican are all negatively related to support towards marijuana legalization at the individual-level. As such, they are thought to be negatively related at the state-level.

In terms of other demographic variables, a higher proportion of men increases the probability of a state having legalized recreational marijuana, and, interestingly, a higher median age also increases this probability. The direction of *age* is not as expected based on the literature exploring the determinants of support at the individual level. All marijuana-specific variables have the expected direction of coefficients: *medyrs*, *med*, *dec* and *diffusion* all increase the probability of *rec* as they themselves increase. In terms of magnitude, *medyrs*, *ballot*, *hrelig*, *repshare*, *grass* and *med* have the largest coefficients – as observed in the variable importance plots for most models.

5. Discussion and Conclusion

The purpose of this chapter is to discuss the results in light of the research question “To what extent is it possible to predict the legal status of recreational marijuana in the US?”. In addressing this question, I also discuss why SVM 1, SVM 2, and GBM 3 ended up being the best models. Furthermore, this chapter includes a discussion of this thesis’ theoretical and methodological contributions to the literature. Finally, I briefly discuss the limitations of this thesis.

5.1 Conclusion

I have in this thesis explored what separates states that have legalized marijuana for recreational use from state that have not. This has been done using machine learning prediction models in order to answer the research question “To what extent is it possible to predict if a state has legalized recreational marijuana in the United States?”. To answer the research question quite bluntly, it is to a large extent possible to predict whether a state has legalized recreational marijuana or not. Despite having a low number of observations, the most powerful models built in this thesis predict both positives and negatives well. But how can one predict the legality of recreational marijuana?

Methodologically, the largest implication of this thesis’ analysis is the no-free-lunch theorem: the need for building multiple machine learning models that differ in their fundamentals and data-inputs in order to end up with powerful predictive models. In terms of theory, the application of public opinion and marijuana policy-specific variables are important. The application of public opinion is not just the use of public opinion as a variable. Rather, it concerns the abstraction of determinants of individual-level support towards marijuana legality to the state-level. For instance, using a religiosity index as a state-level input variable due to it being a determinant of individuals’ support towards legalization. In terms of policy-specific variables, medicalization is central in prediction (and perhaps explaining) the legality of recreational marijuana. If there is one theoretical consideration to draw from this thesis it is the following: in order to understand, explain and predict the legality of recreational marijuana, a thorough understanding of why states legalize medical marijuana, and its relationship to recreational legalization, is important.

5.2 Prediction of Recreational Marijuana Legality

5.2.1 Which model predicts best, and why?

The first and second Support Vector Machine models, as well as the third Gradient Boosting Machine model, are the best predictive models of recreational marijuana legality in this thesis. But why do these models predict better than the other models? In many ways it is not possible to answer this question to the fullest extent. This is due to the no-free-lunch theorem that states that "averaged over all optimization problems, without re-sampling all optimization algorithms perform equally well" (Adam et al. 2019, 58). However, there are some conclusions that can be drawn from the fact that SVM and GBM predict better than the shrinkage methods, RF and DT.

Firstly, both SVMs and GBMs are non-parametric, in contrast to the shrinkage methods. In other words, they do not assume the function describing the relationship between dependent and independent variables. This is not necessarily a benefit. However, in the case of this thesis and its data, it seems to be. Considering that the SVMs and GBMs predict better than the shrinkage models, one can assume that the relationship between the legal status of recreational marijuana and the input variables is not linear. However, due to the complex inner-workings of these models, determining how the output variable is related to the input variables is difficult, and even the assumption that the input variables and output variable are not linearly related is still not certain.

Explaining why the best models' performance in terms of their input-data is challenging as they all use different datasets. The three best models all use different sets of data-inputs. SVM 1, as shown in figure 4.5, relies heavily on the variable *medyrs* to make predictions. So much so that the other variables are practically insignificant for the prediction of marijuana legalization.

However, considering that GBM 3 predicts well without *medyrs* refutes the idea that *medyrs* is necessary for high predictive accuracy. But what can the different inputs say about *why* SVM and GBM predict so well? There are two common denominators amongst the four most important variables in SVM 1, SVM 2 and GBM 3: *ballot* and *unemp*. These variables are clearly important, since the three best models consider them important (as seen in the variable importance plots). *ballot* is also one of the variables that uses interactions with other variables in the models the most. However, the inclusion of these variables is not unique to the two aforementioned methods – so it can only partially explain why SVM 1 and 2, and GBM 3 predict so well. If anything, the fact that SVM 1 and GBM 3 predict well with different datasets illustrates that there is no "free lunch" in machine learning prediction. The fact that they reach the same prediction solidifies the idea that one needs to test different machine learning methods and data-inputs to build a strong model.

Interaction effects may be why GBM 3 predicts so well despite using fewer variables than the SVM models. GBM 3 had a high overall interaction score, and a very high score for certain variables (such as *ballot* and *diffusion*). A reason why GBM 3 predicts so well may therefore be because it uses variables together – not just by themselves.

Answering the research question

Having discussed the three most powerful models, trying to discern why they predict so well, brings us to this thesis' research question:

To what extent is it possible to predict the legal status of recreational marijuana in the United States?

Since the best models had an overall accuracy of 97.3% (100% for negatives, 71.4% for positives), I am inclined to conclude that it is possible to a large degree. Even with such a small training dataset, the models predict the unseen data well. However, it is difficult to conclude with certainty that this high level of accuracy would remain if there were more observations in the test dataset.

5.3 Contributions

This section gives insight into the theoretical, as well as methodological, contributions of this thesis – beginning with the theoretical.

5.3.1 Theoretical implications and contributions

As emphasized multiple times already, one must be wary when using prediction models for causal inferences and implications. Despite this, prediction models – particularly interpretive models – may be used to explore causality, or at the very least suggest areas for future research to explore in terms of causality. This thesis is, therefore, not without its theoretical implications and contributions – despite being a methodologically centered thesis.

Public opinion and responsiveness

The main input variable is the variable *grass*, measuring public opinion at the USCB division level. The discussions in the theory chapter suggest that this variable should be able to contribute well to the prediction of recreational marijuana legality. The idea is quite simple:

government is responsive to the citizens' policy preferences. Empirically, this relationship has been found by multiple scholars, and is quite established in the field of policy research (Tatalovich and Daynes 1988; Carmines and Stimson 1989; Haider-Markel and Kaufman 2006; Lax and Phillips 2009a).

The results suggests that public opinion is an important input variable for the models to determine whether a state has legalized recreational marijuana or not. The models suggest that there is a relationship between public opinion and the legality of recreational marijuana. However, it may simply be that states that have legalized see an increase in public support to the legalization of marijuana.

The public opinion variable is amongst the four or five most important input variables in the models as a whole. This is not exceptional. If one takes its measurement at the USCB division-level into consideration (as opposed to the state-level), the fact that the models benefit from including *grass* is significant. The theoretical implication is not necessarily that public opinion affects marijuana policy, as prediction models are inappropriate for such an inference. However, the implication is that high support towards legalizing marijuana is, at the very least, an indication of whether or not a state has legalized recreational marijuana. This is an important contribution considering that public opinion has not been used as an input-variable for prediction models of marijuana legality, nor as an independent variable in regression models. What this suggests for future research is that attaining state-level public opinion data (through autoMrP using the geographical identification variable) for the purpose of studying the effect of public opinion on marijuana policy and/or marijuana policy change is a worthwhile endeavour.

Medical marijuana – trojan horse?

As is clear, *medyrs* and *med* contributed significantly to the prediction of recreational marijuana's legal status in many of the models. But what does this imply for the trojan horse theory? It is not necessarily confirmed – nor disproven – as this is not the purpose of this thesis. This is not only because of the prediction vs. regression dimension. It is because intent is an important aspect of trojan horse, as discussed. This is not only inappropriate for prediction models to investigate, but for regression analyses as well. However, one thing is clear: predicting whether or not a state has legalized marijuana for recreational use is easier if one uses these variables, particularly *medyrs*. Empirically, it seems like a prerequisite (as all states that have legalized marijuana recreationally have done so medically as well), and the prediction model suggests the same as well. I am not saying that medicalization is a necessary condition, as recreational legalization can in theory happen without medicalization. For future research, *medyrs* dominance may suggest that in order to gain insight and understanding of why states legalize marijuana for recreational use, or what separates these states from others (as I have done in this thesis), a thorough analysis of why states legalize marijuana for medical use in the first place is useful.

In light of the interaction plots, *medyrs* seems reliant on other variables in order to make predictions. This somewhat weakens the notion that *medyrs* in and of itself is important. The implication is that *medyrs* needs the other variables to reach its full potential as a predictor.

Diffusion

The variable *diffusion* did not help greatly in distinguishing states by the legality of recreational marijuana. However, diffusion is not unimportant. Rather, simply having neighboring states with legal marijuana did not help with predictions significantly. The variable *unemp*, on the other hand (signifying 'motivation' for diffusion), was quite important in some of the prediction models. Individually, they help predictions a little. However, these variables should be used together as interaction variables in a regression analysis to fully understand diffusion and explore this theory.

Unemployment and diffusion were nevertheless shown to be important variable in terms of interactions – at least in the three best models. In SVM 1, for instance, diffusion was the most important companion for *medyrs* (see figure 4.11).

State level determinants

Some of the most important predictive variables were state level variables discussed under the 'state level determinants' section of the theory chapter. Particularly, these variables have strong ties to determinants of individual level support towards marijuana legalization – such as religiosity and whether one votes republican or democrat. These two variables, in particular, are amongst the most important variables in this thesis' prediction models. Without discussing causality and the effects these variables might have on marijuana legalization, the inclusion of these variables (and their predictive power) particularly highlights one thing: the usefulness of abstracting individual-level variables to the state-level. This is especially relevant for predicting/explaining policy. That is, using state-level demographic variables that correspond to the determinants of support for a certain policy. For instance, *repshare*[3] was used because whether an individual is a democrat or republican is correlated with whether they support legalizing marijuana or not. This is in many ways also a methodological consideration/contribution.

5.3.2 Methodological contributions

Machine learning methods – insights for future endeavors

The main methodological contribution of this thesis is demonstrating the use of prediction models in the social sciences. On the one hand, this contribution is pedagogical: explaining how each of the methods used work, and how one should approach such a prediction task. On the other hand, I have also demonstrated when prediction models can be appropriate. Particular to this thesis when there is multicollinearity in the data and operationalizations are sub-optimal. Furthermore, I have demonstrated the importance of utilizing theories when choosing input variables for machine learning models. Despite the results being somewhat atheoretical, theory has played an important part in creating such powerful prediction models. However, this thesis is not a tutorial for how to apply machine learning models, particularly considering the R-packages used and the scripts created have not been shared nor explained thoroughly. This somewhat diminishes the methodological contributions.

Data and autoMrP

Perhaps an even greater contribution than the demonstration of machine learning prediction models is the demonstration and use of the R-package autoMrP. As discussed, autoMrP is a machine-learning package in R that improves upon traditional Multilevel regression and post-stratification. Apart from Broniecki, Leeman and Wüest's articles and R-vignette explaining and demonstrating the use of this package, it has not been used by scholars. The disaggregated public opinion values created using autoMrP were useful for the prediction of recreational marijuana legality – despite only being disaggregated to the USCB division-level. The fact that this data proved useful despite its sub-optimal operationalization is an important demonstration of this package's potential for future research where representative data at lower levels of analysis is lacking.

[3] share of population that voted for the republican candidate in the last presidential election

In addition, using this ‘new’ data and new R-package, I have uploaded a dataset with an overview of marijuana laws in each U.S. state from 1976 to 2021 on my GitHub profile[4].

5.4 Limitations

Prediction vs Regression – a limitation?

One limitation of this thesis is that it is not designed to explore the effects that variables have on the legalization of recreational marijuana in detail. This is mainly due to its use of prediction models rather than regression models. However, due to high VIF-scores, multicollinearity, and sub-optimal operationalization, studying the effect of variables on the legalization of marijuana would be difficult even with regressions. One can even go as far as saying that in order to see the true effects of x on Y, a quantitative *and* qualitative design is necessary: regressions for determining statistical significance and coefficients, and case-studies for process-tracing and mechanism-identification.

Regardless, choosing to create prediction models as opposed to regression models has its downsides. One of the downsides is parsimony and accessibility. Machine learning methods are not necessarily easy to understand, particularly for readers who have little experience with them. This makes it difficult for readers to both understand the models and the conclusions drawn from them, as well as critique the methodology and identify methodological mistakes. An extreme consequence of this would be that reader’s do not see the necessity of machine learning methods in the social sciences, and why it has been applied to studying the legal status of recreational marijuana.

Another limitation, which is more methodological, is that Ensemble Bayesian Model Averaging (EBMA) has not been applied to the final prediction models. This would, in all likelihood, make the predictions even better – as is done in autoMrP. Furthermore, in terms of predictions, I have not identified which observations have been accurately/falsely predicted. For instance, the two false negatives in GBM 3 have not been identified. Identifying these would indicate which states are anomalies, or least likely cases that turned out to have legal recreational marijuana. This would be useful for future endeavors, particularly qualitative research.

[4] https://github.com/alexcroz

Data

A second set of limitations of this thesis concerns its data. The primary limitation is the small number of observations. The dataset used only has around 250 observations, with the years 2010, 2012, 2014, 2016 and 2018. This is, as discussed previously, due to dependencies on the public opinion variable from GSS – this variable is only available biannually. Only having 250 observations increases the risk of overfitting the training data, making the models biased by these observations. Furthermore, having such few observations makes it difficult to assess the models, as the model accuracy is greatly altered by the classification of a few variables. It is therefore difficult to conclude anything with certainty – be it the importance of individual variables or the predictive power of each model. The way I see it, there are two solutions to this problem. The first being to simply add more observations pre-2010. For instance, by adding 2006 and 2008 to the dataset. However, this would only provide the models with more observations of states that have not legalized, because no states had legal marijuana in this timespan. This would not give the models more positive observations to be trained on. The positive observations are, after all, the ones the models struggled the most with classifying. Therefore, observations from 2006 and 2008 were not added to the dataset due to assumed diminishing returns.

The second option, which would greatly increase the number of observations and would allow the models to be built on annual data, would be to not use the public opinion variable. This variable is the reason why the dataset only has biannual data. However, this comes with great costs. Removing these could negatively affect prediction accuracy. It would also decrease this thesis' methodological and theoretical contributions: *grass* would no longer be an input variable, and autoMrP would not have been used in this thesis. The use of autoMrP is, the way I see it, one of the biggest contributions of this thesis, as it is arguably more important and has greater areas of use than prediction models. The final data-related limitation also concerns *grass*. The fact that grass is measured at the division-level, as opposed to the state-level, is unfortunate both in terms of operationalization, and in terms of demonstrating how useful autoMrP can be for future research.

Focus and design

There are two main limitations regarding this thesis' focus and design. The first is the type of marijuana policy that is focused on: recreationally legal marijuana. On the one hand, this subject is contemporary and controversial, and has not been studied much before. This is why I chose to focus on it. On the other hand, the results indicate that in order to understand recreational marijuana legality, it is perhaps more worthwhile to put an effort into understanding medical legalization since *medyrs* dominated the models. However, this was obviously not known to me when choosing to study recreational marijuana. One can argue that the field should 'finish' studying medicalization in order to get a better understanding of how to study recreational legalization. Alternatively, the theoretical focus could have been more on the relationship between medicalization and recreational legalization, and the question of whether the prior is a prerequisite for the former.

Another limitation relates to my focus on policy rather than policy change. The dependent variable is measured as legality of recreational marijuana in the US, rather than as an intervention variable. The results therefore boil down to 'what is the difference between states that have legalized recreational marijuana and states that have not?', as opposed to predicting the year of legalization (intervention/event). The implication of this is that the policy responsiveness/policy change theory discussed in the theory section is not as applicable. Nonetheless, the policy (as opposed to policy change) was chosen due to increased number of observations, amongst other considerations.

6. References

Adam, Stavros P., Stamatios-Aggelos N. Alexandropoulos, Panos M. Pardalos, and Michael N. Vrahatis. 2019. "No Free Lunch Theorem: A Review." In *Approximation and Optimization : Algorithms, Complexity and Applications*, edited by Ioannis C. Demetriou and Panos M. Pardalos, 57-82. Cham: Springer International Publishing.

Arceneaux, Kevin. 2002. "Direct Democracy and the Link between Public Opinion and State Abortion Policy." *State Politics & Policy Quarterly* 2 (4): 372-387. https://doi.org/10.1177/153244000200200403. https://journals.sagepub.com/doi/abs/10.1177/153244000200200403.

ARDA. 2022. "General Social Surveys." Association of Religion Data Archives. https://www.thearda.com/Archive/GSS.asp.

Ballotpedia. 2021. "History of marijuana on the ballot." Accessed 11.11.2021. https://ballotpedia.org/History_of_marijuana_on_the_ballot.

---. 2022. "Ballot initiatives." https://ballotpedia.org/Ballot_initiative.

Barry, Rachel A., and Stanton A. Glantz. 2018. "Marijuana Regulatory Frameworks in Four US States: An Analysis Against a Public Health Standard." *American Journal of Public Health* 108 (7): 914-923. https://doi.org/10.2105/ajph.2018.304401. https://ajph.aphapublications.org/doi/abs/10.2105/AJPH.2018.304401.

Berinsky, Adam J. 2017. "Measuring Public Opinion with Surveys." *Annual Review of Political Science* 20 (1): 309-329. https://doi.org/10.1146/annurev-polisci-101513-113724. https://www.annualreviews.org/doi/abs/10.1146/annurev-polisci-101513-113724.

Bernardi, Luca, Daniel Bischof, and Ruud Wouters. 2021. "The public, the protester, and the bill: do legislative agendas respond to public opinion signals?" *Journal of European Public Policy* 28 (2): 289-310. https://doi.org/10.1080/13501763.2020.1729226. https://doi.org/10.1080/13501763.2020.1729226.

Bestrashniy, Jessica, and Ken C. Winters. 2015. "Variability in medical marijuana laws in the United States." *Psychology of addictive behaviors : journal of the Society of Psychologists in Addictive Behaviors* 29 (3): 639-642. https://doi.org/10.1037/adb0000111. https://pubmed.ncbi.nlm.nih.gov/26415061

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4588056/.

Boehmke, Bradley, and Brandon Greenwell. 2019. In *Hands-On Machine Learning with R*, 488. CRC Press.

Bonnie, Richard J. 2018. "The Surprising Collaps of Marijuana Prohibition: What Now?" *UC Davis Law Review* 50 (2): 573-593.

Bradford, Ashley C., and David W. Bradford. 2017. "Factors driving the diffusion of medical marijuana legalisation in the United States." *Drugs: Education, Prevention and Policy* 24 (1): 75-84. https://doi.org/10.3109/09687637.2016.1158239. https://doi.org/10.3109/09687637.2016.1158239.

Broniecki, Philipp, Lucas Leemann, and Reto Wüest. 2021. "Improved Multilevel Regression with Post-stratification through Machine Learning (autoMrP)." *The Journal of Politics* 84 (1): 597-601. https://doi.org/10.1086/714777. https://doi.org/10.1086/714777.

---. 2021b. autoMrP: Multilevel Models and Post-Stratificaion (MrP) Combined with Machine Learning in R. 1-22.

Burstein, Paul. 2003. "The Impact of Public Opinion on Public Policy: A Review and an Agenda." *Political Research Quarterly* 56 (1): 29-40. https://doi.org/10.2307/3219881. http://www.jstor.org/stable/3219881.

---. 2020. "The Determinants of Public Policy: What Matters and How Much." *Policy Studies Journal* 48 (1): 87-110. https://doi.org/https://doi.org/10.1111/psj.12243. https://onlinelibrary.wiley.com/doi/abs/10.1111/psj.12243.

Buttice, Matthew K., and Benjamin Highton. 2013. "How Does Multilevel Regression and Post-stratification Perform with Conventional National Surveys?" *Political Analysis* 21 (4): 449-467. http://www.jstor.org/stable/24572674.

Campos, Isaac. 2018. "Mexicans and the Origins of Marijuana Prohibition in the United States: A Reassessment." *The Social History of Alcohol and Drugs* 32: 6-37. https://doi.org/10.1086/shad3201006. https://www.journals.uchicago.edu/doi/abs/10.1086/SHAD3201006.

Carmines, Edward G.;, and James A. Stimson. 1989. *Issue evolution : race and the transformation of American politics*. Princeton, N.J.: Princeton University Press.

Caughey, Devin, and Christopher Warshaw. 2016. "The Dynamics of State Policy Liberalism, 1936–2014." *American Journal of Political Science* 60 (4): 899-913. https://doi.org/https://doi.org/10.1111/ajps.12219. https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12219.

CDPHE. 2021. "How to apply for a Colorado medical marijuana card." Accessed 25.10.2021. https://cdphe.colorado.gov/apply-colorado-medical-marijuana-card.

CGA. 2022. "Marijuana Taxes." Colorado General Assembly. Accessed 03.06.22. https://leg.colorado.gov/agencies/legislative-council-staff/marijuana-taxes%C2%A0.

Chan, Jireh Yi-Le, Steven Mun Hong Leow, Khean Thye Bea, Wai Khuen Cheng, Seuk Wai Phoong, Zeng-Wei Hong, and Yen-Lin Chen. 2022. "Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review." *Mathematics* 10 (8): 1283. https://www.mdpi.com/2227-7390/10/8/1283.

Chen, Yuchen, and Yuhong Yang. 2021. "The One Standard Error Rule for Model Selection: Does It Work?" *Stats* 4 (4): 868-892. https://www.mdpi.com/2571-905X/4/4/51.

Cruz, José Miguel, Maria Fernanda Boidi, and Rosario Queirolo. 2018a. "Saying no to weed: Public opinion towards cannabis legalisation in Uruguay." *Drugs: Education, Prevention and Policy* 25 (1): 67-76. https://doi.org/10.1080/09687637.2016.1237475. https://doi.org/10.1080/09687637.2016.1237475.

---. 2018b. "The status of support for cannabis regulation in Uruguay 4 years after reform: Evidence from public opinion surveys." *Drug and Alcohol Review* 37 (S1): S429-S434. https://doi.org/https://doi.org/10.1111/dar.12642. https://onlinelibrary.wiley.com/doi/abs/10.1111/dar.12642.

Cruz, José Miguel, Rosario Queirolo, and María Fernanda Boidi. 2016. "Determinants of Public Support for Marijuana Legalization in Uruguay, the United States, and El Salvador." *Journal of Drug Issues* 46 (4): 308-325. https://doi.org/10.1177/0022042616649005. https://journals.sagepub.com/doi/abs/10.1177/0022042616649005.

DEA. 2018. "Drug Scheduling." United States Drug Enforcement Administration. Accessed 02.06.22. https://www.dea.gov/drug-information/drug-scheduling.

Denham, B. E. 2019. "Attitudes toward legalization of marijuana in the United States, 1986-2016: Changes in determinants of public opinion." *Int J Drug Policy* 71: 78-90. https://doi.org/10.1016/j.drugpo.2019.06.007.

DFAF. 2022. "About Us." Drug Free America Foundation. https://www.dfaf.org/about-us/.

DISA. 2022. "Map of Marijuana Legality by State." Accessed 14.01.2022. https://disa.com/map-of-marijuana-legality-by-state.

Dye, Thomas R. 1984. *Understanding public policy*. Fifth edition. Englewood Cliffs, N.J. : Prentice-Hall, [1984] ©1984.

Epstein, Lee, and Jeffrey A. Segal. 2000. "Measuring Issue Salience." *American Journal of Political Science* 44 (1): 66-83. https://doi.org/10.2307/2669293. http://www.jstor.org/stable/2669293.

Erikson, Robert S. 1976. "The Relationship between Public Opinion and State Policy: A New Look Based on Some Forgotten Data." *American Journal of Political Science* 20 (1): 25-36. https://doi.org/10.2307/2110507. http://www.jstor.org/stable/2110507.

Erikson, Robert S., Gerald C. Wright, and John P. McIver. 1993. *Statehouse Democracy: Public Opinion and Policy in the American States*. Cambridge: Cambridge University Press.

Felson, Jacob, Amy Adamczyk, and Christopher Thomas. 2019. "How and why have attitudes about cannabis legalization changed so much?" *Social Science Research* 78: 12-27. https://doi.org/https://doi.org/10.1016/j.ssresearch.2018.12.011. https://www.sciencedirect.com/science/article/pii/S0049089X17310232.

Ferraiolo, Kathleen. 2014. "Morality Framing in U.S. Drug Control Policy: An Example From Marijuana Decriminalization." *World Medical & Health Policy* 6 (4): 347-374. https://doi.org/https://doi.org/10.1002/wmh3.114. https://onlinelibrary.wiley.com/doi/abs/10.1002/wmh3.114.

Gray, Shannon, 2022, "Colorado sets new record for marijuana tax and fee revenue in a single year," https://sbg.colorado.gov/sites/sbg/files/220111_December_and_November_2021_Marijuana_Sales_and_Tax_Revenue_Press_Release.pdf.

Grossback, Lawrence J., Sean Nicholson-Crotty, and David A. M. Peterson. 2004. "Ideology and Learning in Policy Diffusion." *American Politics Research* 32 (5): 521-545. https://doi.org/10.1177/1532673X04263801. https://doi.org/10.1177/1532673X04263801.

Haider-Markel, Donald, and Matthew Kaufman. 2006. "Public Opinion and Policy Making in the Culture Wars: Is There a Connection Between Opinion and State Policy on Gay and Lesbian Issues?", 163-82.

Hannah, A Lee. 2018. "The Politics of Passing and Implementing Medical Marijuana in ohio." *The Journal of Economics and Politics* 24 (1). https://collected.jcu.edu/jep/vol24/iss1/1.

Hannah, A Lee, and Daniel Mallinson. 2018. "Defiant Innovation: The Adoption of Medical Marijuana Laws in the American States." *Policy Studies Journal* 46. https://doi.org/10.1111/psj.12211.

IIHS. 2022. "Marijuana laws by state." Insurance Institute for Highway Safety. Accessed 03.06.22. https://www.iihs.org/topics/alcohol-and-drugs/marijuana-laws-table.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2015. *An introduction to statistical learning: with applications in R*. 2 ed.*Springer Texts in Statistics*. New York: Springer.

Johns, Tracy L. 2015. "Managing a Policy Experiment:Adopting and Implementing Recreational Marijuana Policies in Colorado." *State and Local Government Review* 47

(3): 193-204. https://doi.org/10.1177/0160323x15612149. https://journals.sagepub.com/doi/abs/10.1177/0160323X15612149.

Kilmer, Beau, and Robert J. MacCoun. 2017. "How Medical Marijuana Smoothed the Transition to Marijuana Legalization in the United States." *Annual Review of Law and Social Science* 13 (1): 181-202. https://doi.org/10.1146/annurev-lawsocsci-110615-084851. https://www.annualreviews.org/doi/abs/10.1146/annurev-lawsocsci-110615-084851.

Klüver, Heike, and Mark Pickup. 2019. "Are they listening? Public opinion, interest groups and government responsiveness." *West European Politics* 42 (1): 91-112. https://doi.org/10.1080/01402382.2018.1483662. https://doi.org/10.1080/01402382.2018.1483662.

Korf, Dirk. 2008. "An open front door: the coffee shop phenomenon in the Netherlands." In *A cannabis reader: global issues and local experiences: Volume 1*, edited by EMCDDA, 352. The European Monitoring Centre for Drugs and Drug Addiction.

Lascher, Edward L., Michael G. Hagen, and Steven A. Rochlin. 1996. "Gun Behind the Door? Ballot Initiatives, State Policies and Public Opinion." *The Journal of Politics* 58 (3): 760-775. https://doi.org/10.2307/2960443. http://www.jstor.org/stable/2960443.

Lax, Jeffrey, and Justin Phillips. 2009a. "Gay Rights in the States: Public Opinion and Policy Responsiveness." *American Political Science Review* 103 (3): 367-386. https://www.cambridge.org/core/journals/american-political-science-review/article/gay-rights-in-the-states-public-opinion-and-policy-responsiveness/3B905084E7544CFB9035B79F127FEBD3.

---. 2009b. "How Should We Estimate Public Opinion in The States?" *American Journal of Political Science* 53: 107-121. https://doi.org/10.1111/j.1540-5907.2008.00360.x.

Lowery, David, Virginia Gray, and Gregory Hager. 1989. "Public Opinion and Policy Change in the American States." *American Politics Quarterly* 17 (1): 3-31. https://doi.org/10.1177/1532673x8901700101. https://journals.sagepub.com/doi/abs/10.1177/1532673X8901700101.

MacCoun, Robert, and Peter Reuter. 2001. *Drug war heresies: Learning from other vices, times, and places*. doi:10.1017/CBO9780511754272.*Drug war heresies: Learning from other vices, times, and places.* New York, NY, US: Cambridge University Press.

Mallinson, Daniel J, and A Lee Hannah. 2020. "Policy and Political Learning: The Development of Medical Marijuana Policies in the States." *Publius: The Journal of*

Federalism 50 (3): 344-369. https://doi.org/10.1093/publius/pjaa006. https://doi.org/10.1093/publius/pjaa006.

Manza, Jeff, and Fay Lomax Cook. 2002. "A Democratic Polity?:Three Views of Policy Responsiveness to Public Opinion in the United States." *American Politics Research* 30 (6): 630-667. https://doi.org/10.1177/153267302237231. https://journals.sagepub.com/doi/abs/10.1177/153267302237231.

Mayhew, David R. 1974. "Congressional elections: The case of the vanishing marginals." *Polity* 6 (3): 295-317.

Molnar, Christoph. 2019. *Interpretable Machine Learning*. Lulu.com.

Monroe, Alan D. 1979. "Consistency between Public Preferences and National Policy Decisions." *American Politics Quarterly* 7 (1): 3-19. https://doi.org/10.1177/1532673X7900700101. https://doi.org/10.1177/1532673X7900700101.

---. 1998. "Public Opinion and Public Policy, 1980-1993." *The Public Opinion Quarterly* 62 (1): 6-28. http://www.jstor.org/stable/2749715.

Monte, Andrew A., Richard D. Zane, and Kennon J. Heard. 2015. "The Implications of Marijuana Legalization in Colorado." *JAMA* 313 (3): 241-242. https://doi.org/10.1001/jama.2014.17057. https://doi.org/10.1001/jama.2014.17057.

Montgomery, Jacob M., Florian M. Hollenbach, and Michael D. Ward. 2012. "Improving Predictions using Ensemble Bayesian Model Averaging." *Political Analysis* 20 (3): 271-291. https://doi.org/10.1093/pan/mps002. https://www.cambridge.org/core/article/improving-predictions-using-ensemble-bayesian-model-averaging/11866974EE2888D4A2988309FC6B602F.

Mooney, Christopher Z., and Mei-Hsien Lee. 1995. "Legislative Morality in the American States: The Case of Pre-Roe Abortion Regulation Reform." *American Journal of Political Science* 39 (3): 599-627. https://doi.org/10.2307/2111646. http://www.jstor.org/stable/2111646.

Mosher, Clayton J, and Scott Atkins. 2019. "In the Weeds: Demonization, Legalization, and the Evolution of US Marijuana Policy." *Social Forces* 98 (4): 1-3. https://doi.org/10.1093/sf/soaa003. https://doi.org/10.1093/sf/soaa003.

Mouhamed, Yara, Andrey Vishnyakov, Bessi Qorri, Manpreet Sambi, Sarah Frank, Catherine Nowierski, Anmol Lamba, Umrao Bhatti, and Myron Szewczuk. 2018. "Therapeutic potential of medicinal marijuana: an educational primer for health care professionals."

Drug, Healthcare and Patient Safety 10: 45-66. https://doi.org/https://doi.org/10.2147/DHPS.S158592.

MPP. 2022. "State Policy." Marijuana Policy Project. Accessed 04.06.22. https://www.mpp.org/states/.

Natapoff, Alexandra. 2011. "Misdemeanors." *S. Cal. L. Rev.* 85: 1313.

Newport, Frank. 2011. "Record-High 50% of Americans Favor Legalizing Marijuana Use." Gallup Politics. Accessed 28.10.2021. https://news.gallup.com/poll/150149/record-high-americans-favor-legalizing-marijuana.aspx.

NORC. 2018. "Obtaining GSS Sensitive Data Files." National Opinion Reserach Center. Accessed 15 March. https://gss.norc.org/documents/other/ObtainingGSSSensitiveDataFiles.pdf.

NORML. 2021. "Legalization." Accessed 14.09.21. https://norml.org/laws/legalization/.

---. 2022. "About Marijuana." https://norml.org/marijuana/.

Norrander, Barbara, and Clyde Wilcox. 1999. "Public Opinion and Policymaking in the States: The Case of Post-Roe Abortion Policy." *Policy Studies Journal* 27 (4): 707-722. https://doi.org/https://doi.org/10.1111/j.1541-0072.1999.tb01998.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0072.1999.tb01998.x.

Pacula, Rosalie, Jamie Crhiqui, and Joanna King. 2003. "Marijuana Decriminalization: What does it mean in the United States?" *Working Paper Series*. https://doi.org/10.3386/w9690. http://www.nber.org/papers/w9690.

Pacula, Rosalie, Robert MacCoun, Peter Reuter, Jamie Chriqui, Beau Kilmer, Katherine Harris, Letizia Paoli, and Carsten Schäfer. 2005. "What Does It Mean to Decriminalize Marijuana? A Cross-National Empirical Examination." *Advances in health economics and health services research* 16: 347-69. https://doi.org/10.1016/S0731-2199(05)16017-8.

Pacula, Rosalie, and Rosanna Smart. 2017. "Medical Marijuana and Marijuana Legalization." *Annual Review of Clinical Psychology* 13 (1): 397-419. https://doi.org/10.1146/annurev-clinpsy-032816-045128. https://www.annualreviews.org/doi/abs/10.1146/annurev-clinpsy-032816-045128.

Page, Benjamin I., and Robert Y. Shapiro. 1983. "Effects of Public Opinion on Policy." *The American Political Science Review* 77 (1): 175-190. https://doi.org/10.2307/1956018. http://www.jstor.org/stable/1956018.

PRC. 2014. "Religious Landscape Study." Pew Research Center. Accessed 22.03.22. https://www.pewforum.org/religious-landscape-study/.

Room, Robin, Benedikt Fischer, Wayne Hall, Simon Lenton, and Peter Reuter. 2010. "Cannabis Policy: Moving beyond stalemate."

Shipan, Charles R., and Craig Volden. 2008. "The Mechanisms of Policy Diffusion." *American Journal of Political Science* 52 (4): 840-857. https://doi.org/https://doi.org/10.1111/j.1540-5907.2008.00346.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5907.2008.00346.x.

---. 2012. "Policy Diffusion: Seven Lessons for Scholars and Practitioners." *Public Administration Review* 72 (6): 788-796. https://doi.org/https://doi.org/10.1111/j.1540-6210.2012.02610.x. https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6210.2012.02610.x.

Stringer, Richard J., and Scott R. Maggard. 2016. "Reefer Madness to Marijuana Legalization: Media Exposure and American Attitudes Toward Marijuana (1975-2012)." *Journal of Drug Issues* 46 (4): 428-445. https://doi.org/10.1177/0022042616659762. https://doi.org/10.1177/0022042616659762.

Subbaraman, Meenakshi Sabina, and William C. Kerr. 2017. "Support for marijuana legalization in the US state of Washington has continued to increase through 2016." *Drug and Alcohol Dependence* 175: 205-209. https://doi.org/https://doi.org/10.1016/j.drugalcdep.2017.02.015. https://www.sciencedirect.com/science/article/pii/S0376871617301710.

Tatalovich, Raymond, and Byron W. Daynes. 1988. "Conclusion: Social Regulatory Policymaking." In *Social Regulatory Policy*, edited by Raymond Tatalovich, Byron W. Daynes and Theodore J Lowi. New York: Routledge.

USC. n.d. "Overview of Controlled Substances and Precursor Cheimcals." University of South Carolina. Accessed 02.06.22. https://ehs.usc.edu/research/cspc/chemicals/#:~:text=The%20Controlled%20Substances%20Act%20(CSA,hallucinogens%2C%20anabolic%20steroids%2C%20and%20other.

USCB. 2018. *Understanding and Using*

American Community Survey Data. (United States Census Bureau). https://www.census.gov/content/dam/Census/library/publications/2018/acs/acs_general_handbook_2018.pdf.

---. 2022. "Geographic Levels." United States Census Bureau. https://www.census.gov/programs-surveys/economic-census/guidance-geographies/levels.html.

Vabalas, Andrius, Emma Gowen, Ellen Poliakoff, and Alexander J. Casson. 2019. "Machine learning algorithm validation with a limited sample size." *PLOS ONE* 14 (11): e0224365. https://doi.org/10.1371/journal.pone.0224365. https://doi.org/10.1371/journal.pone.0224365.

von Hoffmann, Jonas. 2020. "“Someone has to be the First”: Tracing Uruguay’s Marijuana Legalisation Through Counterfactuals." *Journal of Politics in Latin America* 12 (2): 177-199. https://doi.org/10.1177/1866802x20937415.

Wang, Wei, David Rothschild, Sharad Goel, and Andrew Gelman. 2015. "Forecasting elections with non-representative polls." *International Journal of Forecasting* 31 (3): 980-991. https://EconPapers.repec.org/RePEc:eee:intfor:v:31:y:2015:i:3:p:980-991.

Wlezien, Christopher. 1995. "The Public as Thermostat: Dynamics of Preferences for Spending." *American Journal of Political Science* 39: 981.

Wlezien, Christopher, and Stuart N. Soroka. 2021. Public Opinion and Public Policy. Oxford University Press.

Zoorob, Michael J. 2021. "The frequency distribution of reported THC concentrations of legal cannabis flower products increases discontinuously around the 20% THC threshold in Nevada and Washington state." *Journal of Cannabis Research* 3: 1-6. https://doi.org/http://dx.doi.org/10.1186/s42238-021-00064-2. https://www.proquest.com/scholarly-journals/frequency-distribution-reported-thc/docview/2546375219/se-2

Zou, Hui, and Trevor Hastie. 2005. "Regularization and variable selection via the elastic net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301-320. https://doi.org/https://doi.org/10.1111/j.1467-9868.2005.00503.x. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00503.x.

Appendix

A.1 Logistic regression models

Figure A.1.1 *Correlation Matrix*

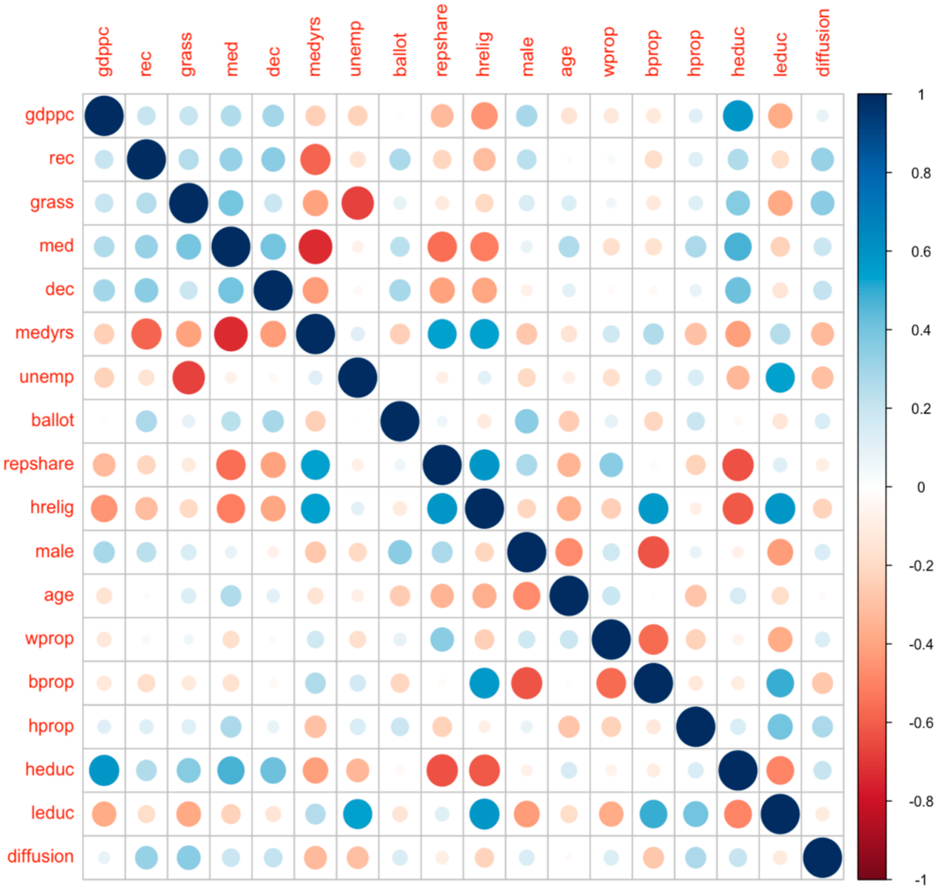


Table A.1.1 *Logistic Regression Table*

Model 1			
Variable	**Coefficient**	**Std. Error**	**VIF**
(Intercept)	-2.61	0.295	NA
grass	1.01***	0.275	NA

Model 2			
Variable	**Coefficient**	**Std. Error**	**VIF**
(Intercept)	-26.01	1491.68	NA
grass	8.39	5.18	1.06
med	17.79	1491.67	1.00
dec	1.85***	0.67	1.02
diffusion	1.99*	1.05	1.07

Model 3			
Variable	**Coefficient**	**Std. Error**	**VIF**
(Intercept)	-26.42	1448.02	NA
grass	10.09	6.76	1.14
med	13.50	1448.01	1.00
dec	1.57*	0.84	1.09
diffusion	1.20	1.48	1.01
medyrs	0.34***	0.08	1.06

Model 4			
Variable	**Coefficient**	**Std. Error**	**VIF**
(Intercept)	-33.55	2524.29	NA
gdppc	-1.49	1.13	11.76
grass	0.11	0.91	3.73
med	18.21	2524.29	1
dec	0.61	1.24	2.19
unemp	-1.37	1.21	6.67
ballot	8.01***	2.54	4.35
repshare	-1.42	0.93	8.40
hrelig	-4.11**	1.91	16.26
male	2.19*	1.16	15.22
age	-0.56	1.14	17.37
wprop	-0.19	1.14	9.02
bprop	0.18	1.33	6.84
hprop	1.06	1.02	12.09
heduc	0.73	0.99	6.85
leduc	-0.97	1.65	16.73
diffusion	0.74*	0.39	2.91

Model 5			
Variable	**Coefficient**	**Std. Error**	**VIF**
(Intercept)	-29.20	2686.83	NA
gdppc	2.09	2.31	26.8
grass	0.42	1.59	6.8
med	14.24	2686.83	1.00
dec	-2.26	2.05	3.35
unemp	-1.49	1.76	9.77
ballot	2.56	3.46	4.09
repshare	0.67	1.08	3.76
hrelig	-6.70**	2.96	20.48
male	-3.35	2.23	41.49
age	-5.93**	2.63	55.85
wprop	6.62	4.80	91.01
bprop	4.44	3.56	37.74
hprop	2.44	1.77	16.15
heduc	-1.88	1.80	12.37
leduc	-2.53	2.77	29.08
diffusion	0.71	0.82	4.35
medyrs	7.62**	3.01	16.68

Notes:	***: p<0.01, **: p<0.05, *p<0.10

A.2 Variable list and dataset overview

Table A.2.1 *Variable list*

Variable name	Meaning	Measurement
medyrs	years since observation legalized medical marijuana	Integer. Negative value for years prior to medicalization, positive for years following.
med	medical marijuana legal or not	Dichotomous. 0 and 1.
dec	marijuana decriminalized or not	Dichotomous. 0 and 1.
rec	recreational marijuana legal or not	Dichotomous. 0 and 1.
diffusion	share of neighbouring states that have legalized marijuana for recreational purposes	Numeric. 0 to 1.
grass	support towards marijuana legalization	Numeric. 0 to 1.
hrelig	religiosity index	Numeric. 0 to 1.
repshare	share of population that voted for the republican candidate in the last presidential election	Numeric. 0 to 1.
ballot	ballot initiative possible or not	Dichotomous. 0 and 1.
hprop	proportion of hispanic people in state	Numeric. 0 to 1.
bprop	proportion of black people in state	Numeric. 0 to 1.
wprop	proportion of white people in state	Numeric. 0 to 1.
gdppc	gross domestic product per capita	Integer.
age	median age	Integer.
male	proportion of males in state	Integer. 0 to 1
unemp	unemployment rate	Integer.

Table A.2.2 *Dataset overview*

Dataset 1	**Dataset 2**	**Dataset 3**
medyrs	*med*	*rec*
med	*dec*	*grass*
dec	*rec*	*hrelig*
rec	*diffusion*	*repshare*
diffusion	*grass*	*ballot*
grass	*hrelig*	*hprop*
hrelig	*repshare*	*bprop*
repshare	*ballot*	*wprop*
ballot	*hprop*	*gdppc*
hprop	*bprop*	*age*
bprop	*wprop*	*male*
wprop	*gdppc*	*unemp*
gdppc	*age*	
age	*male*	
male	*unemp*	
unemp		

A.3 Interaction plots

Figure A.3.1 *SVM 1 Interaction Plot (diffusion)*

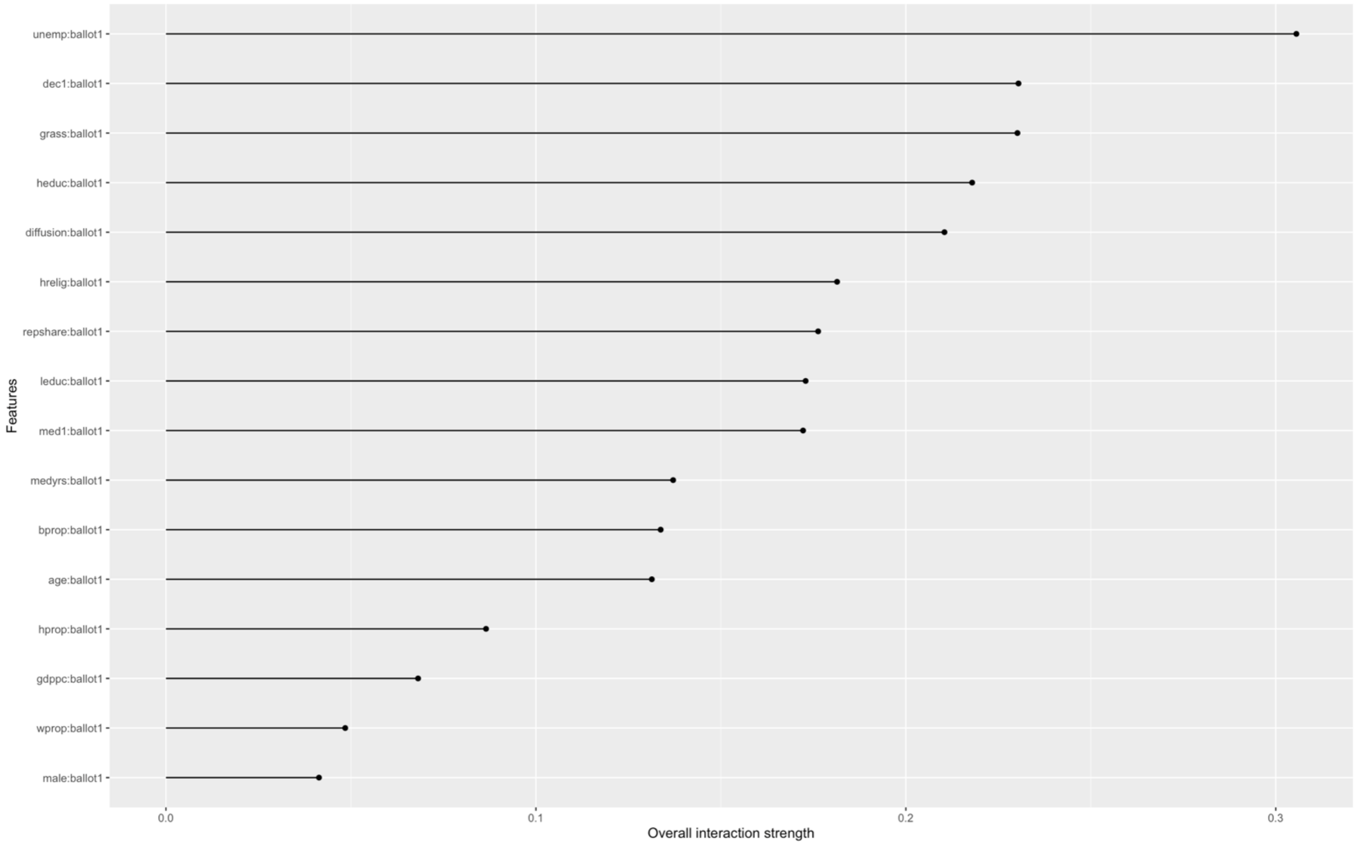


Figure A.3.2 *SVM 2 Interaction Plot (ballot)*

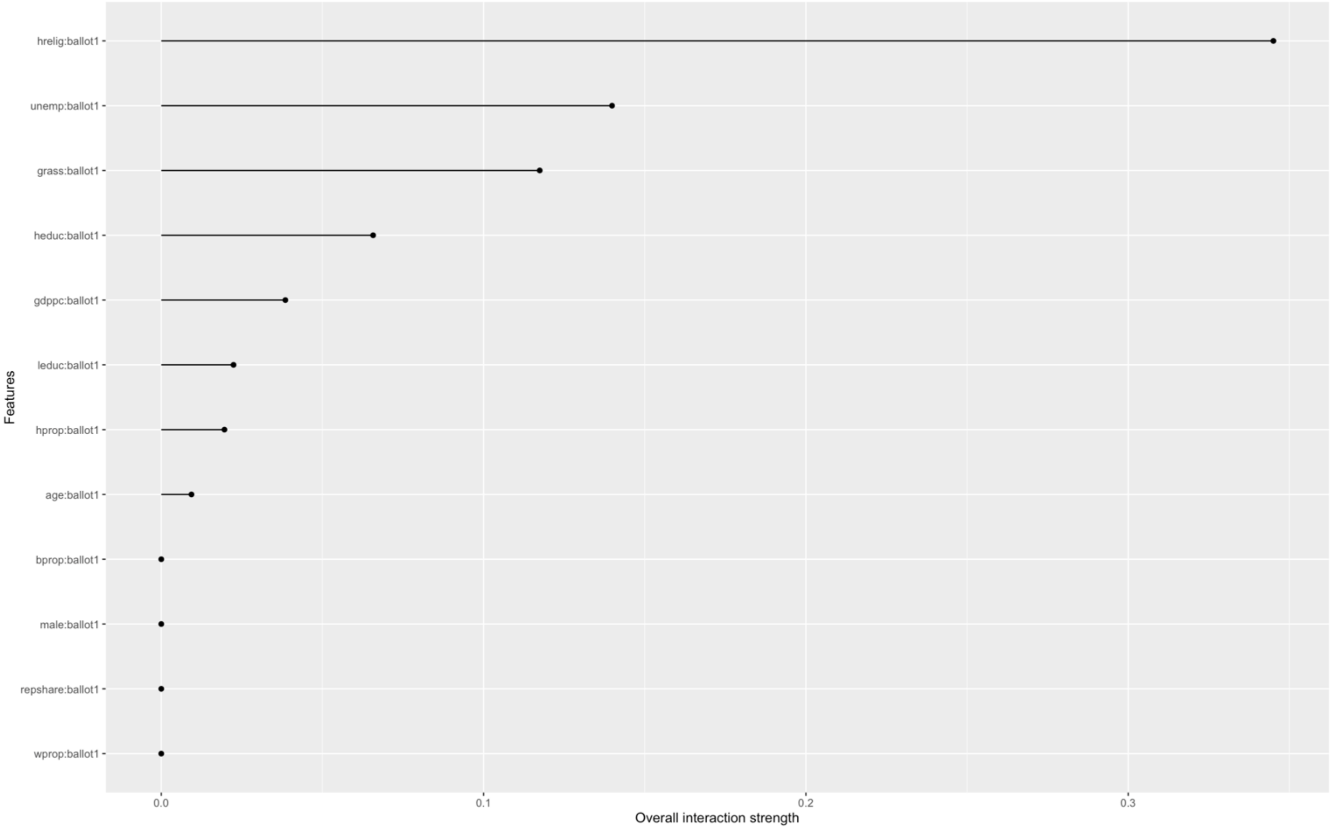


Figure A.3.3 *SVM 2 Interaction Plot (grass)*

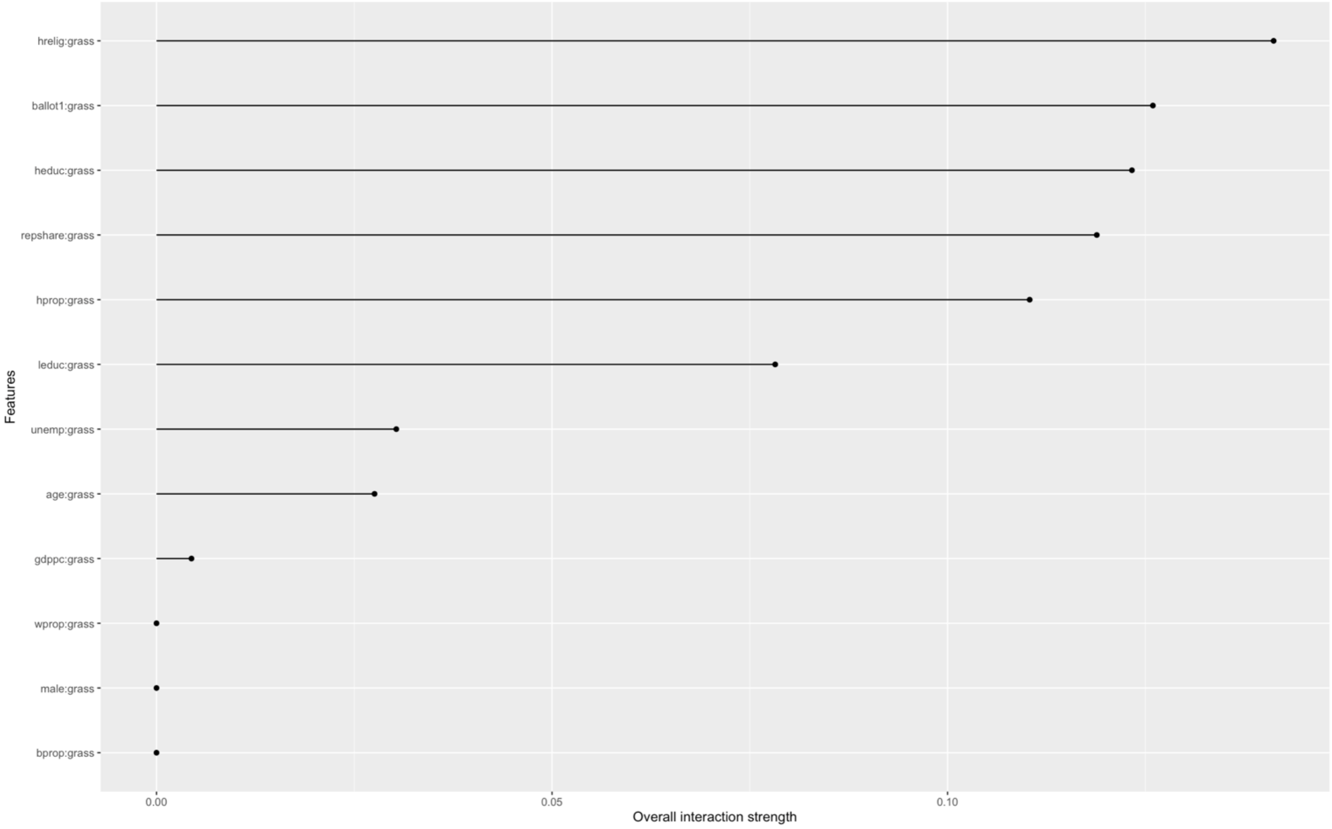
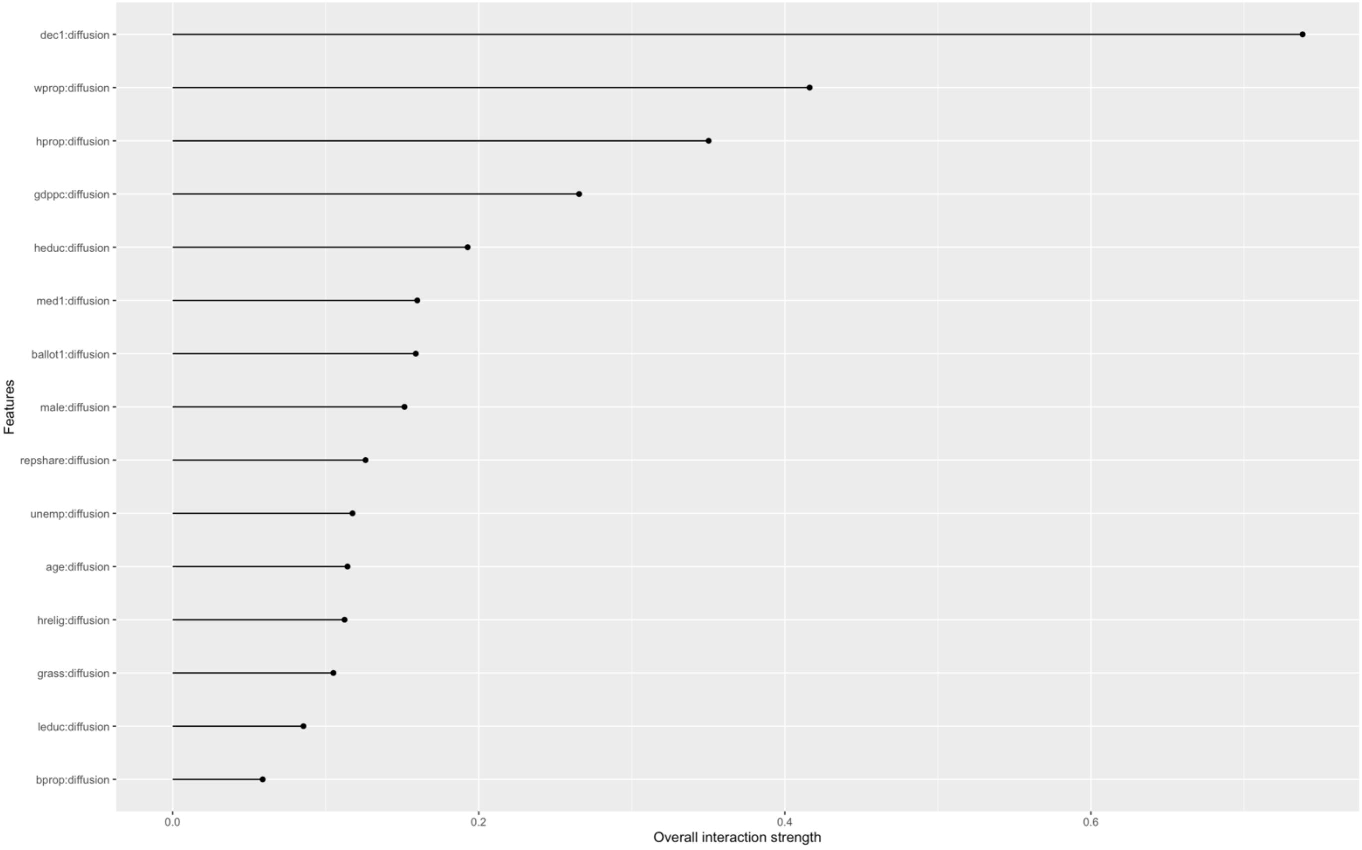


Figure A.3.4 *GBM 3 Interaction Plot (diffusion)*



′