2021

Statistical and Machine Learning Approaches to Depressive Disorders Among Adults in the United States: From Factor Discovery to Prediction Evaluation

Minhwa Lee
The College of Wooster, mlee21@wooster.edu

Statistical and Machine Learning Approaches to Depressive Disorders Among Adults in the United States: From Factor Discovery to Prediction Evaluation

Independent Study Thesis

Presented in Partial Fulfillment of the Requirements for
the Degree Bachelors of Arts in the
Computer Science and Mathematics in the Department of
Mathematical & Computational Sciences at The College of
Wooster

by
Minhwa Lee

The College of Wooster
2021

Advised by:

Dr. Sofia Visa (Computer Science)

Dr. Marian Frazier (Mathematics)

THE COLLEGE OF WOOSTER

Abstract

According to the National Institutes of Mental Health (NIMH), depressive disorders (or major depression) are considered one of the most common and serious health risks in the United States. Our study focuses on extracting non-medical factors of depressive disorders diagnosis, such as overall health states, health risk behaviors, demography, and healthcare access, using the Behavioral Risk Factor Surveillance System (BRFSS) data set collected by the Center of Disease Control and Prevention (CDC) in 2018.

We set the two objectives of our study about depressive disorders in the United States as follows. First, we aim to utilize machine learning algorithms and statistical methods to build models that will discover the factors of depressive disorders for young, middle, and old adulthood in the United States. Second, based on the mined attributes from each adult group, we predict depressive disorders for each group and evaluate the performances of those prediction tasks. Throughout the study, we obtain an in-depth understanding of what impacts the depressive disorders diagnosis for each adult group in the United States, as well as how machine learning and statistical approaches are useful in mining information about the factors and predicting the depressive disorders.

Acknowledgments

First of all, I would like to thank God for giving me this great opportunity to learn what he destined me to do in this new environment. Without God's will and help, I could not have started my new journey in this land.

Also, I would like to acknowledge my IS advisors, Dr. Marian Frazier and Dr. Sofia Visa, for their excellent guidance and warm-hearted words of encouragement for my long thesis throughout the entire year. I would also like to extend my gratitude for their assistance in preparing for graduate school applications and helping me to continue my academic passion toward Computer Science in the United States. I will remember their endless support. Thank you so much.

Lastly, I would like to express my deepest gratitude to my parents, grandparents, and two adorable sisters, Beenhwa "Grace" and Seunghwa "Victoria." Looking back on the past four years of college, I realize that my academic journey in the United States has been continued entirely due to the support of all my family members. I will always think of their prayers and sacrifice whenever I have hard time throughout my entire life.

Vita

Fields of Study Major field: Computer Science and Mathematics

Minor field: Statistical & Data Sciences

Contents

APPENDIX PAGE

List of Figures

List of Tables

Table Page

CHAPTER 1

Introduction

1.1 Topic of Interest: Depressive disorders in the U.S.

Mental health is a significant portion of humans' overall health, and sometimes we face problems in maintaining good mental health conditions. In particular, depressive disorders (or major depression) are one of the most common and serious mental health risks for all age groups in the United States. They may cause other major mental health disorders such as suicide attempts, bipolar disorders, and even schizophrenia.

Current research indicates that depressive disorders are caused by genetic, biological, environmental, and psychological factors [18]. Also, previous research has found that family history of depression, major stress, and certain medications due to physical illnesses can be major risk factors of depressive disorders [18]. Furthermore, individuals can be diagnosed with depressive disorders at any age, but they usually realize their suffering from the related symptoms in their adulthood. In particular, people in midlife or old age groups can have complex physical illnesses that can co-occur with depressive disorders due to side effects of medications [18].

Therefore, we set **the two objectives of our study** as follows. First, we discover factors with regard to overall health states, demographic features, health risk behaviors, and healthcare access that would impact depressive disorder diagnosis for three adult groups (young, middle, and old adulthood) by using machine learning techniques with some statistical approaches. Then, based on the extracted features of depressive disorders from each adult group, we predict the depressive disorders by constructing models with the same methods that we used for the discovery process, and we evaluate the performance of the prediction tasks.

1.2 Research trends about Depressive Disorders

One research paper from Tara W. Strine claimed that chronic health issues, sex, and marital status are associated with mental health issues [21]. Strine also states that healthcare access for both physical and mental health should be improved. Therefore, the article provides a clue that demography and healthcare access can affect one's depressive disorders. Different research from Strine also found a significant relationship between depressive disorders and health-related risk behaviors such as smoking, physical inactivity, and alcohol consumption [22]. The main result of the article is that adults with depressive disorders tend to smoke, have obesity, be physically inactive, and drink heavily. Thus, we can grasp an understanding of the impact of health risk behaviors on an individual's depression.

Also, the application of machine learning (ML) to the research of depressive disorders is substantial. For example, one recent study discusses the performance of several machine learning algorithms in the detection of depressive disorders from a clinical data set [8]. The article presents some insights that several ML models, such as logistic regression, CART trees, and support vector machines, have overall good performance in prediction tasks [8]. Also, it points out that the ML models can bring a broader view of depressive disorders by tracking the demographic factors, such as household income and educational background of individuals [8]. Following the current research trend over the topic of our interest, we therefore decide to use ML algorithms to not only find risk factors of depressive disorders for each age group of the U.S. adults, but also to predict the onset of depressive disorders for each group.

1.3 Machine Learning Approaches

We live in a world of data, where the analysis of the data is now considered to be essential in every academic field. Thus, we employ an advanced technique called **machine learning** to extract knowledge from, and find patterns in, the data of interest. In this section, we present an overview of machine learning, as in later chapters we will apply this technique to the data of our interest. This enables us to uncover valuable, organized information about depressive disorders among the U.S. adult residents in 2018.

1.3.1 Machine Learning: Supervised learning

Machine learning enables computers to learn the patterns of data by themselves. Also, it is academically denoted as a *learning process* from experience E with respect to some task T and some performance measure P, as its performance on T improves with experience E [16]. In simpler words, machine learning gives computers the capacity to automatically learn the data with low time-complexity so as to find complex patterns and perform better decisions [12]. There are major three types of machine learning techniques: *supervised, unsupervised,* and *reinforcement learning.* Since our study will use supervised learning methods, we explore the concept of supervised learning, as our main focus of machine learning fields.

In **supervised learning**, training data is provided with the correct class labels, and based on discovered patterns from the training set, a machine learning method responds to the testing data with unknown class labels [15]. A typical example is *classification*, where the trained model classifies the incoming inputs into each class label. If a target variable is numerical, *regression* can predict the numeric value of the target based on results from the trained model. Notice that logistic regression is an exception of regression methods, as it is widely used as a classification method and calculates the probability of belonging to the given class labels [11].

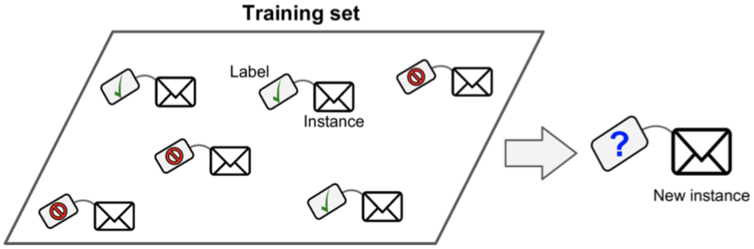

Figure 1.1: An example of supervised learning as a classification method of spam emails [12]

For example, Figure 1.1 shows how supervised machine learning method can classify spam emails. The learning method obtains characteristics of spam email from the examples of the training set. The method also predicts the label of the incoming email as either spam or non-spam, based on the findings from the training process [12].

1.4 Outline of Study

In this chapter, we have introduced the motivation of our study regarding depressive disorders in the U.S., reviewed some related literature, and explored the overview of machine learning. These contents provide us the essentials to investigate advanced supervised machine learning algorithms: decision trees, logistic regression with several statistical hypothesis tests, and support vectors machines along with prediction metrics, which are to be examined in Chapter 2, 3, and 4, respectively.

After exploring the three methodologies, we put all theories of methods into practice. Chapter 5 discusses the data set of our interest, called the *BRFSS*, and performs data transformation procedures, followed by exploratory data analysis of the BRFSS. Then, Chapter 6 presents all the results produced by the three methods. To be specific, this chapter shows the factors of depressive disorder diagnosis that have been produced by decision trees and logistic regression models, as well as the interpretations and insights into those discovered factors. Then, the chapter displays all prediction metrics that all three methods - decision trees, logistic regression, and support vector machines - have performed. In Chapter 7, we have an in-depth discussion about the results presented in Chapter 6 by first comparing decision trees and logistic regression in terms of factor discovery and then providing insights into the prediction metrics of the three methods. Finally, Chapter 8 concludes our study regarding depressive disorders among the U.S. adults, as well as provides potential topics of future research as an extension of this study.

CHAPTER 2

Decision Trees

In this chapter, we present the theory for building a decision tree using the CART algorithm, and we illustrate an example of how the CART algorithm builds a decision tree on a simple data set.

2.1 Introduction

Decision trees are one of the supervised machine learning methods that are most frequently utilized for drawing inference on a given data set. There are several reasons for the popularity of decision trees in the current field of machine learning, but the following two properties of the trees have attracted most people to utilize this method. First, decision trees are computationally low-cost, as they mainly apply if-then rules (or binary decision rules) that simplify the process of querying the trained algorithms. Second, unlike most machine learning techniques which operate in a black-box setting, decision trees are transparent in their decision-making. Therefore, these two benefits of decision trees consolidate the credibility of the algorithm as well as its results. In particular, decision trees have been applied to a wide range of classification tasks such as medical diagnosis, credit card fraud detection, and even sports analysis. Since we aim to investigate what affects the diagnosis of depressive disorder, we confirm that decision tree algorithm is a suitable method for our study.

2.2 Foundation of Decision Trees

Decision trees create classification rules by tracing down the tree from the *root node* – the node in the highest level - to a set of *leaf nodes* in the bottom line. Each node in the tree contains an explanatory variable that contributes to the response, and each branch connecting two nodes corresponds to a logical "AND." Therefore, the main evaluation of any decision trees always starts at the root node, answers the branch from this node, and moves downward the tree. Figure 2.1 shows an example of a decision tree that classifies four animals based on the specific features that each possesses. From the root node, we apply if-then rules to move downward the trees. For example, if an animal has feathers and can fly, then the animal is classified as *dove*. If an animal does not have feathers and fins as well, then it is classified as *lion*. Therefore, decision trees are suitable for presenting clues to problems that necessitate discrete output values, such as boolean responses (e.g. Yes or No).

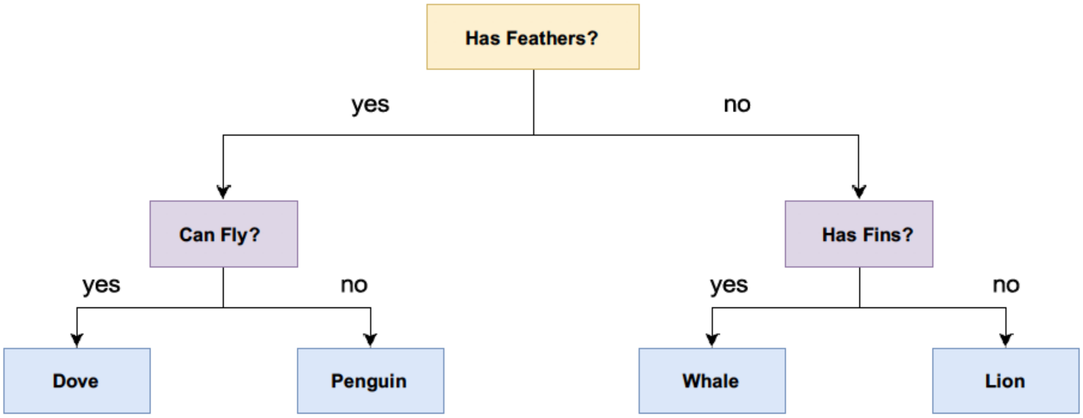

Figure 2.1: An example of decision tree

2.3 CART (Classification and Regression Tree Algorithm)

Next, we investigate how to build a decision tree: to be specific, how to choose the features to be put on the nodes of the tree. One fundamental algorithm for building decision trees is called **classification and regression trees** (**CART**), and it was developed by Leo Breiman alongside many other colleagues in the 1980s [10]. One advantage of the CART algorithm is that it only builds a binary tree: that is, a tree that has only two separated branches to the following nodes. Thus, a CART decision tree can be readily interpreted by using humans' intuition that makes decisions based on the dichotomous responses from each node in the tree. In addition, this binary nature of a CART decision tree creates a visualization of the decision-making process in a highly appealing manner [10, 16, 15]. In addition, the CART constructs a classification tree for the categorical response variable and a regression tree for the numeric one. Since the objective of our study is to classify depressive disorders in the U.S., we would solely concentrate on how the CART builds a classification tree.

2.3.1 Tree-Structured Classifiers

How does the CART build a decision tree for classification? The CART algorithm defines its classification trees in mathematical context as follows [16].

First, tree-structured classifiers are built by beginning with the entire sample D itself at the root and repeatedly splitting D into two descendant subsets X_i and X_j. Figure 2.2 is a hypothetical binary classification tree. X_1 and X_2 are *disjoint*, as $D = X_1 \cup X_2$. Likewise, X_3 and X_4 are disjoint as $X_1 = X_3 \cup X_4$. Also, the following subsets - X_3, X_5, X_7, X_8, X_9, and X_{10} - are denoted as **terminal subsets**. Those terminal subsets, also known as *leaves* or *terminal nodes*, are assigned a class label. For instance, the terminal sets X_8 and X_9 are classified as class B, as the common characteristics of training examples in these two subsets belong to class B.

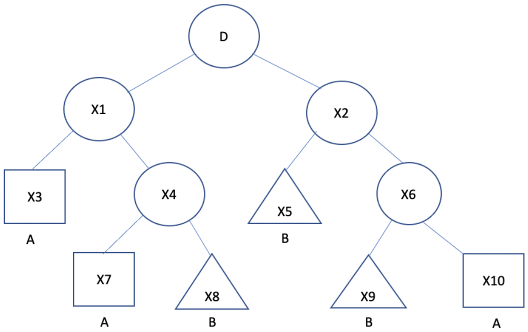

Figure 2.2: A hypothetical two-class decision tree

Suppose that the vector $\vec{x}$ is a multidimensional data point of the examples used in building a decision tree. The splitting rules assign one of the labels $j \in \{1, ..., J\}$ to every vector $\vec{x} \in D$. For example, the first splitting rule assigns to every vector $\vec{x} \in D$ either X_1 or X_2, by applying its defined condition to the $\vec{x}$ and determining whether $\vec{x}$ belongs to X_1 or X_2. Repeating the process and finally reaching a terminal subset in the bottom level of the tree, the tree classifier labels $\vec{x}$ with the class of that terminal subset. Thus, each terminal subset is a partition of D, and it is assigned to one of the two class labels A or B, as shown in Figure 2.2. Then, the partition corresponding to each class is obtained through compiling all terminal subsets within the same class. For example, we write A as a partition of D where all vectors $\vec{x}$ in this partition are assigned to class A. Similarly, set B is the other partition of D where all remaining $\vec{x}$ belong to the class B. Hence, Figure 2.2 presents the following partitions made through the tree: $\boldsymbol{A = X_3 \cup X_7 \cup X_{10}}$ and $\boldsymbol{B = X_5 \cup X_8 \cup X_9}$. Definition 2.1 summarizes all procedures of the classification tree.

Definition 2.1 *A classification tree is a partition of D into j disjoint subsets $X_1, ..., X_j$, where $D = X_1 \cup X_2 \cup ... \cup X_j$ for every $x \in X_j$ is assigned to the class label j. Each terminal subsets of D becomes **terminal nodes** and D is then a **root node** of the tree.* [16]

2.3.2 Splitting Criterion

The main focus of building classification trees is to understand when and where to split the data. In the thesis we will examine one splitting criterion that has been most popularly utilized - **Gini impurity**.

2.3.2.1 Impurity of a Data Set

What does the term 'impurity' represent? First, consider a random data set. If the data set is **pure**, there exists only one class label of the data. Also, if the data set X is said to be pure and contains two classes A and B, then all data points in A have one characteristic a, whereas the remaining data points in B have the other characteristic b which is opposite to a. Suppose a data set M contains 50 red and 0 blue points, then M is a *pure* data. If the data set does not show whether every A has a and every B has b, then it is considered as **impure**. For example, a data set N that contains 20 red and 30 blue points is said to be *impure*. Therefore, the data should be impure in order to build a decision tree based on it.

In formal words, we define a splitting tuple σ as:

$$\sigma =< i, t_k >, \tag{2.1}$$

where i indicates the feature that we choose for splitting the data inputs at a given node of the decision tree, and t_k is the threshold value that determines the left and right branch to proceed. Then, the total impurity $I(D, \sigma)$ of the splitting tuple σ is defined in Equation 2.2:

$$I(D, \sigma) = \frac{N_{left}}{N_D} I(D_{left}) + \frac{N_{right}}{N_D} I(D_{right}), \tag{2.2}$$

where N_D is the number of the entire data inputs at the selected node N, N_{left} and N_{right} are the number of the resulting subsets from the splitting tuple σ [4]. For example, suppose that we have a data set of 100 cases with $N_{left} = 50$ and $N_{right} = 50$. Then, the impurity of this split can be expressed as below:

$$\therefore I = \frac{50}{100} \times I(D_{left}) + \frac{50}{100} \times I(D_{right})$$

2.3.2.2 Gini Impurity

How do we compute the impurities $I(D_{left})$ and $I(D_{right})$? **Gini impurity** is a widely-used measure of impurity [4]. It is the sum of a product of the probability p_i that a data point is correctly classified as the label i and the probability p_j that the data point is misclassified at a certain node N. Notice that p_j is equivalent to $1 - p_i$.

Definition 2.1. *As defined in Equation 2.3, the **Gini impurity** $I_G(N)$ is a measure of how an input from the data set is not correctly classified as the class i at a node N under the assumption that the label is randomly chosen using the probability distribution of the branch [16].*

$$I_G(N) = \sum_{i=1}^{J} p_i p_j = \sum_{i=1}^{J} p_i(1 - p_i) = \sum_{i=1}^{J} (p_i - p_i^2) = 1 - \sum_{i=1}^{J} p_i^2 \tag{2.3}$$

The Gini impurity reaches its minimum 0 when all inputs are correctly classified as only one class. On the other hand, the impurity is at maximum when the number of those inputs for each class is evenly distributed. Thus, the graph of the Gini impurity is concave down, with its minimum of 0, as observed in Figure 2.3.

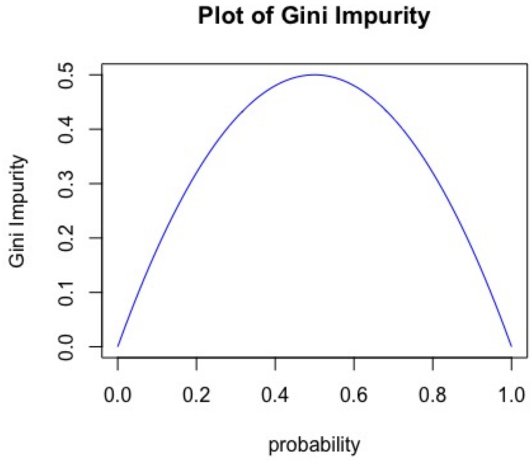

Figure 2.3: The Gini Impurity index as a function of the probability p_i

2.3.3 Splitting Process

Equation 2.3 shows the formula of Gini impurity of each subset of the data set. The subset with the lowest Gini impurity is considered as the feature that splits the data best. Thus, we use this selected feature as splitting criteria for any current node N. Repeating the process of computing the Gini impurity of the remaining subset, the one with the lowest Gini impurity is then placed at the node of the next level of the decision tree.

In mathematical context, the splitting process of decision trees can be explained as the maximization of the change in Gini impurity after each split s at node N.

$$\Delta I(s, N) = I(N) - I(N_L) - I(N_R) \tag{2.4}$$

In Equation 2.4, $I(N)$ is the overall tree impurity, and $I(N_L)$ and $I(N_R)$ denote the impurity of the left and right child nodes, respectively. Therefore, the objective of splitting procedures of a CART decision tree are either to minimize the overall Gini impurity $I(N)$ of the tree, or to maximize the change in Gini impurity $\Delta I(s, N)$.

2.4 Example of Decision Tree Classification

This section investigates a real-world example of building a CART decision tree that uses Gini impurity as a splitting criterion. Sourced from Kaggle, the data set in this example includes information about weather conditions of 14 days, such as Outlook, Temperature, Humidity, and Wind. Based on those given weather conditions, the data set also includes a response column about whether a participant would play tennis or not in that specific day. Thus, the CART tree for this example investigates which features impact their decisions on playing tennis outside [13].

2.4.1 Root Node

First, we explore which variable can be placed at the root node of the decision tree. Given the entire data set as an input for the root node, we compute the Gini impurity of each explanatory variable.

Day#	Outlook	Temperature	Humidity	Wind	Play Tennis
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rainfall	mild	high	weak	yes
5	rainfall	cool	normal	weak	yes
6	rainfall	cool	normal	strong	no
7	overcast	cool	normal	wrong	yes
8	sunny	mild	high	weak	no
9	sunny	cool	normal	weak	yes
10	rainfall	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rainfall	mild	high	strong	no

Table 2.1: The example data set of playing tennis [13]

Gini Impurity (GI) of Outlook

We will compute the Gini impurity when we split the data by using Outlook. Table 2.2 is a two-way table that counts decisions within each category of Outlook. Suppose $I_G(t)$ is a Gini impurity for the category t within Outlook variable.

Outlook	Play Yes	Play No	Total
sunny	2	3	5
overcast	4	0	4
rainfall	3	2	5

Table 2.2: The decision table for the Outlook variable

Referring to Equation 2.3 and Table 2.2, we then compute Gini impurity for each category of Outlook, as described in Equation 2.5, 2.6, and 2.7.

$$I_G(outlook = sunny) = 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = \frac{12}{25} = 0.48 \tag{2.5}$$

$$I_G(outlook = overcast) = 1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 = 0 \tag{2.6}$$

$$I_G(outlook = rainfall) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = \frac{12}{25} = 0.48 \tag{2.7}$$

Then, the weighted sum of Gini impurity for Outlook can be calculated using Equation 2.2, as shown in Equation 2.8:

$$\begin{aligned} \therefore I_G(outlook) &= \frac{N_{sunny}}{N_{total}} \times I_G(sunny) + \frac{N_{overcast}}{N_{total}} \times I_G(overcast) + \frac{N_{rainfall}}{N_{total}} \times I_G(rainfall) \\ &= \left(\frac{5}{14} \times 0.48\right) + \left(\frac{4}{14} \times 0\right) + \left(\frac{5}{14} \times 0.48\right) = 0.342 \end{aligned} \tag{2.8}$$

GI of Temperature

We take the same procedures to calculate the Gini impurity of Temperature, using Table 2.3 which displays the two-way table for the response within each Temperature category.

Temperature	Play Yes	Play No	Total
hot	2	2	4
cool	3	1	4
mild	4	2	6

Table 2.3: The decision table for the Temperature variable

Equation 2.9, 2.10 and 2.11 describe the process of obtaining the Gini impurity for each category of Temperature variable, respectively.

$$I_G(temperature = hot) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = \frac{8}{16} = 0.5 \tag{2.9}$$

$$I_G(temperature = cool) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = \frac{6}{16} = 0.375 \tag{2.10}$$

$$I_G(temperature = mild) = 1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = \frac{16}{36} = 0.445 \tag{2.11}$$

Then, the weighted sum of Gini impurity for Temperature can be calculated as:

$$\begin{aligned} \therefore I_G(temperature) &= \frac{N_{hot}}{N_{total}} \times I_G(hot) + \frac{N_{cool}}{N_{total}} \times I_G(cool) + \frac{N_{mild}}{N_{total}} \times I_G(mild) \\ &= \left(\frac{4}{14} \times 0.5\right) + \left(\frac{4}{14} \times 0.375\right) + \left(\frac{6}{14} \times 0.445\right) = 0.439 \end{aligned} \tag{2.12}$$

GI of Humidity

Table 2.4 displays the two-way table for the binary response from each Humidity category. Similarly, Equation 2.13 and 2.14 describe the process of obtaining the Gini impurity for each category of Humidity variable, respectively. Additionally, Equation 2.15 shows the weighted sum of the total Gini impurity of the Humidity variable.

Humidity	Play Yes	Play No	Total
high	3	4	7
normal	6	1	7

Table 2.4: The decision table for the Humidity variable

$$I_G(humidity = high) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = \frac{24}{49} = 0.489 \tag{2.13}$$

$$I_G(humidity = normal) = 1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 = \frac{12}{49} = 0.244 \tag{2.14}$$

$$\begin{aligned} \therefore I_G(humidity) &= \frac{N_{high}}{N_{total}} \times I_G(high) + \frac{N_{normal}}{N_{total}} \times I_G(normal) \\ &= \left(\frac{7}{14} \times 0.489\right) + \left(\frac{7}{14} \times 0.244\right) = 0.367 \end{aligned} \tag{2.15}$$

GI of Wind

Table 2.5 displays the two-way table for the binary response from each Wind category. In the same manner, Equation 2.16 and 2.17 describe the process of obtaining the Gini impurity for each category of Wind variable, respectively. Additionally, Equation 2.18 indicates the total Gini impurity of the Wind variable.

Wind	Play Yes	Play No	Total
weak	6	2	8
strong	3	3	6

Table 2.5: The decision table for the Wind variable

$$I_G(wind = weak) = 1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2 = \frac{24}{64} = 0.375 \tag{2.16}$$

$$I_G(wind = strong) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = \frac{18}{36} = 0.5 \tag{2.17}$$

$$\begin{aligned} \therefore I_G(wind) &= \frac{N_{weak}}{N_{total}} \times I_G(weak) + \frac{N_{strong}}{N_{total}} \times I_G(strong) \\ &= \left(\frac{8}{14} \times 0.375\right) + \left(\frac{6}{14} \times 0.5\right) = 0.428 \end{aligned} \tag{2.18}$$

Table 2.6 shows the Gini impurity values for all variables of the data set. Since the variable with the lowest Gini impurity is to be selected as a splitting node according to CART, it is clearly identified that Outlook is the root node of the decision tree for this example.

Variables	Gini Impurity
outlook	**0.342**
temperature	0.439
humidity	0.367
wind	0.428

Table 2.6: The Gini impurity for all variables

Therefore, we have our root node of Outlook with three different branches, one for each category within Outlook. Next, we will set three internal nodes for each branch as described in Figure 2.4.

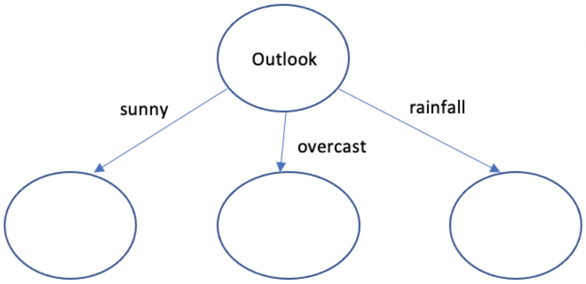

Figure 2.4: The root node of the play tennis tree

2.4.2 Internal Nodes

For setting the internal nodes of the tree, we first formulate three different subsets on the Outlook variable. Then, we apply the same process to figure out the next splitting variable.

Sunny Outlook

Table 2.7 shows all inputs within the sunny category of Outlook variable. Then, for the remaining variables - Temperature, Humidity, and Wind - we will compute the Gini impurity for each one within the same category of the Outlook group.

outlook	**temperature**	**humidity**	**wind**	**play tennis**
sunny	hot	high	weak	no
sunny	hot	high	strong	no
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
sunny	mild	normal	strong	yes

Table 2.7: The subset table for sunny response of Outlook variable

GI of Temperature

Table 2.8 shows the decision of playing tennis based on Temperature categories within sunny outlook.

Temperature	**Play Yes**	**Play No**	**Total**
hot	0	2	2
cool	1	0	1
mild	1	1	2

Table 2.8: The decision table for Temperature on sunny outlook

Then, the Gini impurity for Temperature on sunny outlook will be:

$$I_G(outlook = sunny \,\&\, temperature = hot) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0 \tag{2.19}$$

$$I_G(outlook = sunny \;\&\; temperature = cool) = 1 - \left(\frac{1}{1}\right)^2 - \left(\frac{0}{1}\right)^2 = 0 \tag{2.20}$$

$$I_G(outlook = sunny \;\&\; temperature = mild) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5 \tag{2.21}$$

Therefore, the total Gini impurity for Temperature on sunny Outlook is calculated as:

$$\begin{aligned} \therefore I_G(outlook = sunny \;\&\; temperature) &= \frac{N_{hot}}{N_{total}} \times I_G(hot) + \frac{N_{cool}}{N_{total}} \times I_G(cool) + \frac{N_{mild}}{N_{total}} \times I_G(mild) \\ &= \left(\frac{2}{5} \times 0\right) + \left(\frac{1}{5} \times 0\right) + \left(\frac{2}{5} \times 0.5\right) = 0.2 \end{aligned} \tag{2.22}$$

GI of Humidity

Humidity	Play Yes	Play No	Total
high	0	3	3
normal	2	0	2

Table 2.9: The decision table for Humidity on sunny outlook

$$I_G(outlook = sunny \;\&\; humidity = high) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0 \tag{2.23}$$

$$I_G(outlook = sunny \;\&\; humidity = normal) = 1 - \left(\frac{2}{2}\right)^2 - \left(\frac{0}{2}\right)^2 = 0 \tag{2.24}$$

Then, the weighted sum of Gini impurity for Humidity on sunny outlook is 0.

GI of Wind

Wind	Play Yes	Play No	Total
weak	1	2	3
strong	1	1	2

Table 2.10: The decision table for Wind on sunny outlook

$$I_G(outlook = sunny \;\&\; wind = weak) = 1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2 = 0.44 \tag{2.25}$$

$$I_G(outlook = sunny \;\&\; wind = strong) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5 \tag{2.26}$$

Then, the weighted sum of Gini impurity for Humidity on sunny outlook is 0.466, as explained in Equation 2.27.

$$\therefore I_G(outlook = sunny\&wind) = \left(\frac{3}{5} \times 0.44\right) + \left(\frac{2}{5} \times 0.5\right) = 0.466 \tag{2.27}$$

Since Humidity variable has the lowest Gini impurity value of 0, it will be placed on the next node connected with sunny outlook branch. In addition, we can identify in Table 2.7 that if outlook is sunny and humidity is high, then the person does not play tennis. Since our response variable is binary, the left terminal node of the humidity is *yes* and the right one is *no*, as described in Figure 2.5.

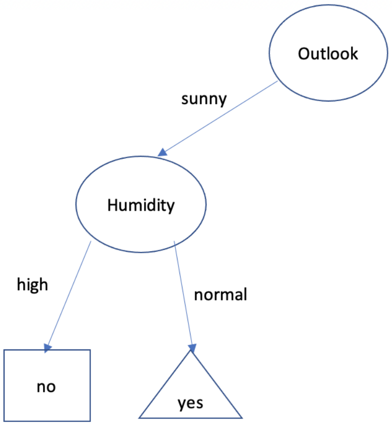

Figure 2.5: The first internal node is Humidity

Overcast Outlook

In Table 2.11, we observe that all the responses for Overcast outlook is 'yes.' Since all those corresponding inputs are classified into a single class 'yes', the Gini impurity of the Overcast variable is 0. The updated decision tree is shown in Figure 2.6.

outlook	**temperature**	**humidity**	**wind**	**play tennis**
overcast	hot	high	weak	yes
overcast	cool	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes

Table 2.11: The subset table for Overcast response of Outlook variable

Rainfall Outlook

Now, we focus on the subsets for rainfall outlook variable, as described in Table 2.12. Then, we compute the Gini impurity for Temperature, Humidity, and Wind variable respectively in order to select the next internal node with the lowest Gini impurity.

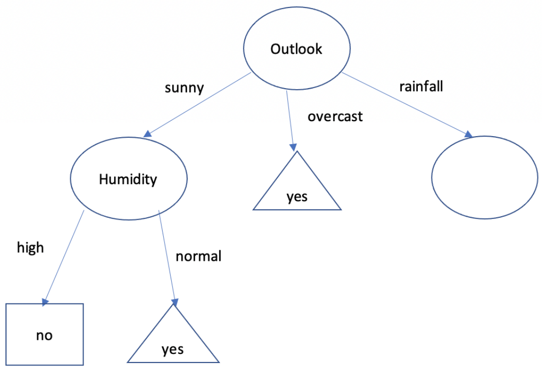

Figure 2.6: The decision tree, added with Overcast outlook

outlook	temperature	humidity	wind	play tennis
rainfall	mild	high	weak	yes
rainfall	cool	normal	weak	yes
rainfall	cool	normal	strong	no
rainfall	mild	normal	weak	yes
rainfall	mild	high	strong	no

Table 2.12: The subset table for Rainfall response of Outlook variable

Temperature	Play Yes	Play No	Total
cool	1	1	2
mild	2	1	3

Table 2.13: The decision table for Temperature on Rainfall outlook

GI of Temperature

Based on Table 2.13, the Gini impurity for Temperature on Rainfall outlook will be:

$$I_G(outlook = rainfall \,\&\, temperature = cool) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5 \tag{2.28}$$

$$I_G(outlook = rainfall \,\&\, temperature = mild) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.444 \tag{2.29}$$

Therefore, the weighted sum of Gini impurity for Temperature on Rainfall outlook is:

$$\begin{aligned} \therefore I_G(outlook = rainfall\& \, temperature) &= \frac{N_{cool}}{N_{total}} \times I_G(cool) + \frac{N_{mild}}{N_{total}} \times I_G(mild) \\ &= \left(\frac{2}{5} \times 0.5\right) + \left(\frac{3}{5} \times 0.444\right) = 0.466 \end{aligned} \tag{2.30}$$

GI of Humidity

$$I_G(outlook = rainfall \,\&\, humidity = high) = 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 = 0.5 \tag{2.31}$$

$$I_G(outlook = rainfall \,\&\, humidity = normal) = 1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2 = 0.444 \tag{2.32}$$

Humidity	Play Yes	Play No	Total
high	1	1	2
normal	2	1	3

Table 2.14: The decision table for Humidity on Rainfall outlook

Therefore, the total Gini impurity for Humidity on rainfall outlook is:

$$\begin{aligned}\therefore I_G(outlook = rainfall \,\&humidity) &= \frac{N_{high}}{N_{total}} \times I_G(high) + \frac{N_{normal}}{N_{total}} \times I_G(normal) \\ &= \left(\frac{2}{5} \times 0.5\right) + \left(\frac{3}{5} \times 0.444\right) = 0.466\end{aligned} \tag{2.33}$$

GI of Wind

Lastly, we compute the Gini impurity for Wind on Rainfall outlook, as described in Equation 2.35, 2.34, and 2.36 based on Table 2.15.

Wind	Play Yes	Play No	Total
weak	3	0	3
strong	0	2	2

Table 2.15: The decision table for wind on rainfall outlook

$$I_G(outlook = rainfall \;\&\; wind = weak) = 1 - \left(\frac{3}{3}\right)^2 - \left(\frac{0}{3}\right)^2 = 0 \tag{2.34}$$

$$I_G(outlook = rainfall \;\&\; wind = strong) = 1 - \left(\frac{0}{2}\right)^2 - \left(\frac{2}{2}\right)^2 = 0 \tag{2.35}$$

$$\therefore I_G(outlook = rainfall \;\&\; wind) = 0 \tag{2.36}$$

Therefore, we have calculated the Gini impurity of all categories of the remaining variables when the Outlook is rainfall. Since the Wind variable has the lowest Gini impurity, we confirm that the next node following rainfall outlook is Wind. In addition, the decisions of playing tennis are always no when the Wind is strong, and they are always yes when the Wind is weak. Therefore, the left terminal node from the Wind is no, and the right one is yes. Figure 2.7 displays the complete decision tree for the example data set.

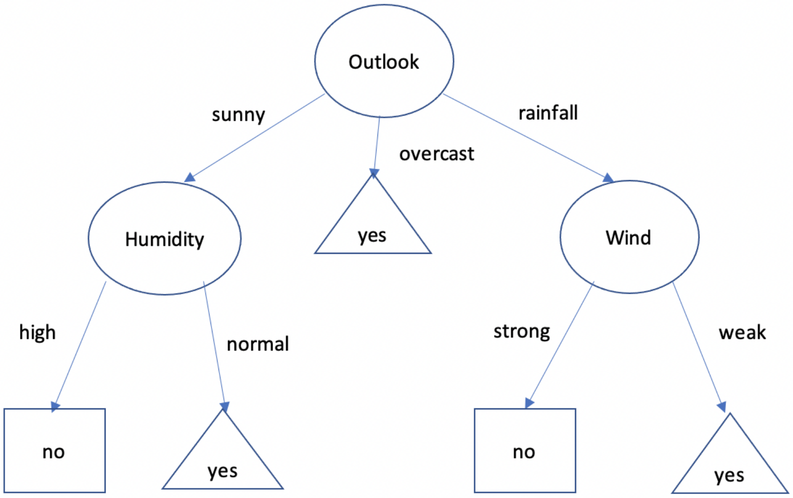

Figure 2.7: The complete decision tree for playing tennis

CHAPTER 3

Logistic regression

In this chapter, we explore our second supervised machine learning method with statistical approach, called *logistic regression*. We discuss essential backgrounds of logistic regression and its application to a real-world example. All explanations of logistic regression through this chapter are referenced from the textbook *Stat2: Building a Model for a World of Data* by Ann Cannon, et al [9].

3.1 Introduction

Regression is a statistical method that summarizes the quantitative relationship between a response variable `Y` and explanatory variables `X`s. This approach lets us predict a new value of `Y` with the `X`s, based on the quantitative information learned from results of the regression. There are various types of regression models, and what determines the types of regression depends primarily on the type of a response variable `Y`.

First, we use *linear regression* for a numeric response variable `Y` with a set of explanatory variables `X`s. Since Y is a continuous variable, fitting a linear regression model to the Y is appropriate for quantitatively extracting relationship between each X and Y and predicting Y by the multiple predictors (explanatory variables). However, when the Y changes its type to a binary variable, which has only two discrete values, then we may consider using another type of regression, called *logistic regression*. This uses a transformed version of `Y` as the response variable. The necessity of transforming an ordinary linear regression is described in Section 3.2.

3.2 Foundations of Logistic Regression

3.2.1 Logistic transformation

As mentioned earlier, logistic regression is utilized when a response variable of a certain model is binary, and it is originated from the transformation of the equation of ordinary linear regression.

First, suppose that we have a data from a random sample of 346 teenagers aged 14 to 18, who respond to the question "On average how many hours of sleep do you get?" The summary of the data is presented in Table 3.1. In Figure 3.1, we observe what relationship exists between age and the proportion of saying "Yes" - it is the proportion of the randomly selected teenagers who sleep at least 7 hours on average weekdays.

Age		14	15	16	17	18
Sleeping At least 7 hours?	Yes	34	79	77	65	41
	No	12	35	37	39	27
Total		46	114	114	104	68
Proportion of Yes		0.74	0.69	0.68	0.63	0.60

Table 3.1: The table of sleeping hours (Yes/No) and age [9]

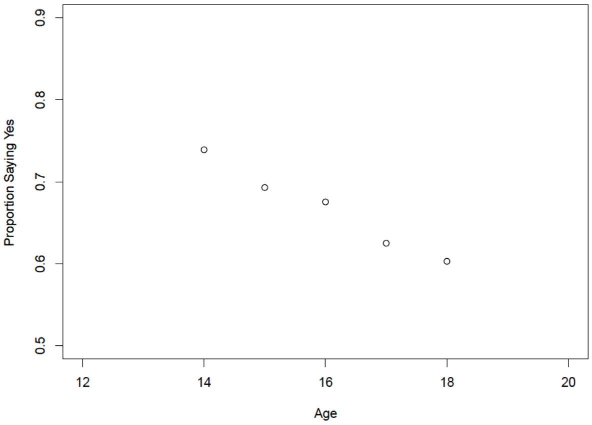

Figure 3.1: The scatterplot of age and proportion of the corresponding teenagers sleeping at least 7 hours (saying "Yes") [9]

We observe a linear relationship between ages of teenagers and the proportion of those sleeping at least 7 hours in weekdays. Hence, the five data points shown in Table 3.1 are fitted well into the linear regression line. However, as we predict other age values not in the table - before 14 and after 18 - with using the fitted line, there will be numerous points whose predicted y-values will be either greater than 1 or less than 0. Considering that definition of the response variable in this example is proportions which must lie between 0 and 1, we confirm that fitting a straight line to those five points is not appropriate. Therefore, we see the necessity to transform the format of linear regression to logistic regression, particularly when using a regression for any response variable that has only two discrete outputs - 0 (No) or 1 (Yes).

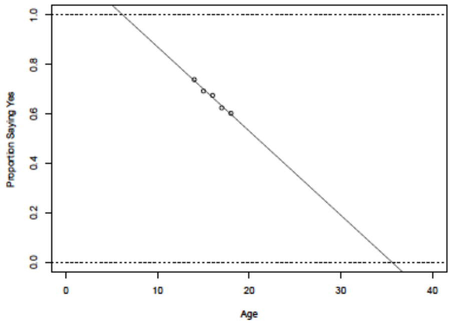

Figure 3.2: The fitted linear regression

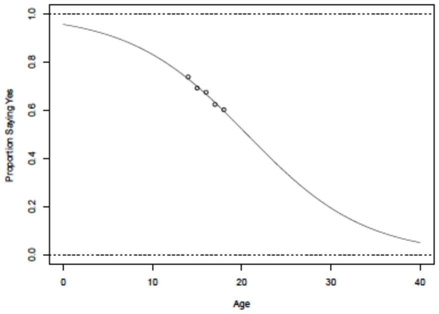

Figure 3.3: The fitted logistic regression

The transformation on the binary response variable Y is called *logit* or *log(odds)*. Also, the logit transformation have two essential features. First, this transformation is reversible due to a nature of logarithm: one π value has only one logit (= log(odds)) form. Also, the logit form of the response variable can take on any real number, which allows a fitted linear equation of a set of predictors. Therefore, we predict the logit (Y) with linear predictors (explanatory variables) of the equation form $\beta_0 + \beta_1 X_1 + ... + \beta_k X_k$, as described in Equation 3.1.

$$\therefore logit(Y) = log(\text{odds of Y}) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \tag{3.1}$$

Definition 3.1. *Let $\pi = Pr(Y = 1)$ = probability that the response variable Y is a "success." The* **odds** *of $Y = 1$ is $\frac{\pi}{1-\pi}$, which is the ratio of the probability of "success" to the probability of "failure." Then, the* **log(odds)** *or* **logit(Y)** *equals* $\log(\frac{\pi}{1-\pi})$, *where the log is indeed natural log.*

Then, the Equation 3.1 can be written as Equation 3.2. Also, Figure 3.4 shows a fitted line of logit form of the proportion of sleeping at least 7 hours a day predicted by age.

$$\therefore log(odds) = log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k \tag{3.2}$$

Based on Equation 3.2, we compute the log (odds) of sleeping at least 7 hours for each age, as listed in Table 3.2. Finally, we draw a plot of log (odds) versus Age , with a fitted line as described in

Figure 3.4. Therefore, we confirm that logit transformation of Y succeeds in fitting given data with a line and predicting other future values at the end.

Age		14	15	16	17	18
Sleeping At least 7 hours?	Yes	34	79	77	65	41
	No	12	35	37	39	27
Total		46	114	114	104	68
Proportion of Yes ($=\pi$)		0.74	0.69	0.68	0.63	0.60
$\log(\frac{\pi}{1-\pi})$		$\frac{0.74}{1-0.74} = 1.05$	$\frac{0.69}{1-0.69} = 0.80$	$\frac{0.68}{1-0.68} = 0.75$	$\frac{0.63}{1-0.63} = 0.53$	$\frac{0.60}{1-0.60} = 0.41$

Table 3.2: Observed values of $\log(\frac{\pi}{1-\pi})$

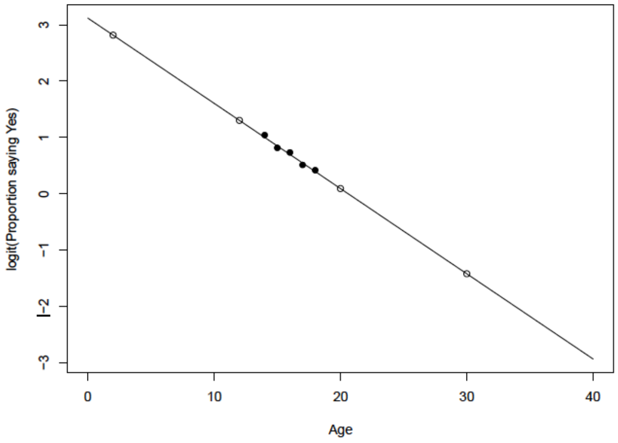

Figure 3.4: The fitted logit plot, with four white points that are not given in the data and five black points that are given [9]

In order to compute the probability of success π of a binary response variable, Equation 3.3 can be used under the assumption that we use k explanatory variables to predict the response, as a modification of Equation 3.2.

$$\therefore P(Y=1) = \pi = \frac{\text{odds}}{1+\text{odds}} = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}} \tag{3.3}$$

3.2.2 Odds Ratio and Fitted Slope

Another essential concept of logistic regression is called *odds ratio*, which is defined as the ratio of two odds. When fitting a linear model of predictors to the logit form of a response variable, the slope of the fitted line is equal to natural log of odds ratio within a single explanatory variable. Thus, the odds ratio equals to $e^{\text{fitted slope}}$. For example, suppose that a categorical explanatory variable X has two discrete values - A and B, and we want to compute the odds ratio of A to B, defined in Equation 3.4, by using the slope of the fitted log-linear model $logit(Y) = aX + b$.

$$\text{odds ratio} = \frac{\text{odds of "success" when X = A}}{\text{odds of "success" when X = B}} \tag{3.4}$$

Assuming the category B is the reference (baseline) group, we will arbitrarily decide that $X = 0$ for $X = B$ and $X = 1$ for $X = A$. Then, we substitute X with 0 and 1 for B and A, respectively. The entire process are presented from Equation 3.5 to 3.8.

$$log(Odds(A)) = a * 1 + b = a + b \tag{3.5}$$

$$log(Odds(B)) = a * 0 + b = b \tag{3.6}$$

$$log(Odds(A)) - log(Odds(B)) = a \tag{3.7}$$

$$\therefore \text{Odds ratio} = \frac{Odds(A)}{Odds(B)} = e^a \tag{3.8}$$

If the X is numeric, we use the same ways for obtaining odds ratio for one-unit increase in X as for the categorical X shown in Equation 3.5 to 3.8. In general, by using a computational method we fit a logistic regression model to the data. Then, from the model summary we obtain estimated slopes of the model to compute the odds ratio of an explanatory variable.

3.3 Assessment of Logistic Regression Models

When using a logistic regression model, we must consider the three requirements to confirm validity of the model for drawing formal inference: **linearity**, **randomness**, and **independence**.

3.3.1 Linearity

When checking the linearity of a logistic regression mode, we examine the following two cases, which primarily depend on the type of given explanatory variables and their corresponding y values. First, if an explanatory variable in the model is **binary** or **categorical**, linearity is automatically satisfied. Otherwise, we draw a empirical logit plot of $log(odds)$ versus the explanatory variable. If the plot shows a linear form as Figure 3.4, then linearity is satisfied. If not linear, we may transform the explanatory variable to have more linear relationship. Lastly, if any transformation does not make the variable have a linear relationship with the $log(odds)$, then using the variable is inappropriate for the logistic regression model.

3.3.2 Randomness and Independence

Since logistic regression derives from a probability model, considering randomness is important for formal inference. We determine the randomness condition based on either how the data has been created or certain situations that we may judge with our discretion. For example, suppose that an experiment design includes random assignments of its participants to either group A or group B. This experimental design satisfies randomness condition.

However, we may take into consideration various situations for checking randomness. For example, if we throw the ball into a box and record the success of each event, the result can be considered as fairly random, because we assume that certain physical forces are engaged in the result of the ball going into the box so that we can apply a probability model.

With regard to independence, we may check whether there are no links to each other. First, if certain events happen with time differences, then it is reasonable to assume that one result does not impact the others in the sequence. Also, if the events include some spatial relationship such as people in the same units, we investigate whether the outcome of one unit is independent of other units. If the events do not explicitly imply independence in the process, then we may use our subjective judgements to determine independence.

3.4 Formal Inference by Statistical Tests

When we utilize formal statistical tests that depend on a probability interpretation of p-values and confidence levels, at least conditions of randomness, independence, and linearity must be satisfied. Once we check the three conditions of a logistic regression model, we can conduct several statistical tests to confirm that the logistic regression model predicts the binary response variable well. The tests can be readily conducted with a computational software such as R.

3.4.1 Wald Test

Wald test examines whether the slope (or coefficient) of an individual explanatory variable is significantly different from 0. If the slope is 0, the odds ratio of the corresponding predictor will be 1 as it equals e^{slope}. This indicates no change in odds between different factor levels within the predictor, and thus the predictor has no impact on the response variable. Hence, Wald tests the following hypotheses, $H_0 : \beta_i = 0$ and $H_a : \beta_i \neq 0$. The corresponding test statistic is defined as Equation 3.9:

$$z = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}}, \tag{3.9}$$

where the z-statistic (or called the Wald statistic) follows the standard normal distribution under the null hypothesis H_0. If we find the small p-values for the slope of the individual predictor in the summary of Wald tests, then we reject the null hypothesis and ultimately conclude that there is a significant log-linear relationship between the predictor and the binary response variable.

Furthermore, we can compute a confidence interval for the slope using Equation 3.10:

$$(\hat{\beta}_i \pm z^* \cdot SE_{\hat{\beta}_i}), \tag{3.10}$$

where z^* is obtained using the normal distribution and the given confidence level. Also, we exponentiate the confidence interval for the slope β_i to obtain an interval for the odds ratio. Therefore, the estimated odds ratio for an one-unit change in the individual predictor x_i is $e^{\hat{\beta}_i}$ with confidence interval $e^{\hat{\beta}_i \pm z^* \cdot SE_{\hat{\beta}_i}}$.

3.4.2 Drop-in-deviance Test

Drop-in-deviance test looks for the overall usefulness of a logistic regression model. In other words, it examines how much improvement was gained by using this logistic model to predict a binary response variable with linear predictors, instead of a constant (or null) model.

To conduct the drop-in-deviance test, we first set up the following hypotheses as below.

$$H_0 : \text{log-linear model is useless}$$

$$H_a : \text{log-linear model is useful}$$

Then, this test compares the null deviance of the constant model to residual deviance of the logistic model. We use the **G-statistic** as defined in Equation 3.11:

$$G = \text{null deviance} - \text{residual deviance}, \tag{3.11}$$

where under the null hypothesis H_0, the G-statistic follows an approximate chi-square distribution with the degrees of freedom equal to the difference in the number of predictors estimated between the constant and the logistic model. Like the Wald test, we observe the associated p-value of the G-statistic. If the p-value is small, we reject the null hypothesis and confirm that there is a compelling evidence that the logistic regression model is useful to predict the response variable by the linear form of explanatory variables.

3.4.3 Nested Drop-in-deviance Test

Nested drop-in-deviance test examines whether adding a predictor to the logistic regression model increases the overall effectiveness of the model. Then, we build a hypothesis as follows:

$$H_0 : \textbf{\textit{Reduced model}}\ \textit{with fewer predictors is sufficient}$$

$$H_a : \textbf{\textit{Full model}}\ \text{is better}$$

We use the G-statistic with a similar formula to Equation 3.11.

$$G = (\text{resid. deviance of the reduced model}) - (\text{resid. deviance of the full model}), \tag{3.12}$$

where the G-statistic follows a chi-square distribution with degrees of freedom equal to difference in the number of predictors between the reduced and the full model. If the associated p-value of the nested G-statistic is significantly close to 0, then the full model is a significant improvement over the reduced model.

3.4.4 Misclassification rate

The last assessment for the logistic regression model is to compute the **misclassification rate** of the model. The misclassifiation rate can be expressed as the proportion of values that the model incorrectly predicts - *Type I (false success)* and *Type II (false failure)* error. The logistic regression model with the lowest misclassification rate is the best model that predicts the binary response variable with the linear form of explanatory variables. For the sake of simplicity, we may compare misclassification rates to determine if an additional explanatory variable can improve the overall performance of the logistic regression model.

3.5 Example Study: Medical School Admissions Results

We now explore how logistic regression can be applied to a real-world problem. The data set *MedGPA* is retrieved from a R-package *'Stat2Data,'* which contains the medical school admission results, GPA, and other standardized test scores for 55 selected students from a liberal arts college in the Midwest region of the United States [7]. We want to use this data to determine how well academic scores and biological sex of individuals predict their medical school admission outcomes. Table 3.3 displays the explanations of the selected 5 variables from the data set *MedGPA*.

Type	**Variable Name**	**Explanation**
Response	Acceptance	Accept: 1 or Denied: 0 (Binary)
Explanatory	Sex	F = female or M = male
	GPA	College GPA
	MCAT	MCAT exam score

Table 3.3: The variables of interest in the MedGPA data set

3.5.1 Single Logistic Regression: One Predictor

First, we explore the characteristics of each selected variable in Table 3.3. We observe that 54.5% of the applicants in the data received an admission from the medical schools. Also, the applicants earned their average college GPA of 3.55, average MCAT score of 36.3, and on average they applied to 8 medical schools.

3.5.1.1 Sex Predictor

First, we examine whether there is a relationship between sex and the medical school admission results. Table 3.4 is a two-way table of Acceptance and Sex, where we obtain information about the percentage of each sex type that received admission.

	Sex	
Acceptance	Female	Male
Accepted	18	12
Denied	10	15

Table 3.4: The two-way table of sex and acceptance results

Of the female applicants of medical schools, $\frac{18}{(18+10)} * 100 = 70.3\%$ were accepted. Of the male applicants of medical schools, $\frac{12}{(12+15)} * 100 = 44.4\%$ received admissions. Then, we fit the logistic regression model LR_{sex} to predict the admission ('Acceptance') from the biological sex of the applicants ('Sex'). Table 3.5 shows the summary of LR_{sex} that predicts the acceptance to medical schools by the applicants' sex. Then, Equation 3.13 shows the linear equation of LR_{sex}:

$$\therefore log(odds) = 0.5878 - 0.811 * Sex(Male) \tag{3.13}$$

Variable	Factor	Coef. Estimate	$e^{Estimate}$	SE	z-statistic	p-value	Null Dev.	Residual Dev.	d.f.
Sex (Female)	Male	-0.811	0.444	0.553	-1.467	0.142	75.79	73.59	1

Table 3.5: The summary table of LR_{sex} (Dev. represents deviance and d.f. indicates degrees of freedom for Chi-square distribution.)

Also, we use Equation 3.8 to obtain the corresponding odds ratio of acceptance by sex: $e^{-0.811} = 0.444$. Hence, the odds of acceptance for male applicants is 44% times the odds of acceptance for female applicants. Based on Equation 3.3, we also obtain that for the male applicants, the probability of acceptance to medical schools is $\frac{e^{(0.5878-0.811)}}{1+e^{(0.5878-0.811)}} = 0.444$. This result matches the proportion of male applicants who received admission, as described above.

In order to draw formal inference, we first check the conditions of the model LR_{sex}. The linearity condition is automatically satisfied because the variable 'Sex' in the model is binary. In terms of randomness and independence, we use our subjective judgements from the speculations of the data collection process. According to the 'Stat2' package manual, the data 'MedGPA' is collected from some students from a randomly selected liberal arts college in the Midwest, which implies reasonable randomness in our model LR_{sex}. Regarding independence, an individual's acceptance does not affect the others' admission results under the assumption that there is no quota in the number of accepted class year for the medical schools. However, since the data 'MedGPA' consists of only 55 students, there is a weakness in drawing generalizations from the results of LR_{sex}.

By using Equation 3.10, we compute a 95% confidence interval for the odds ratio: $(0.146, 1.296)$. Hence, we are 95% confident that the odds of acceptance for a male applicant is between 14.6% and 129.6% the odds of acceptance for a female applicant. However, since the confidence interval includes 1, we conclude that male applicants are not significantly more or less likely to receive admission than female applicants.

Finally, we use two statistical tests to the effectiveness of the model LR_{sex}. According to the Wald test, the associated p-value of the predictor 'Sex' is 0.142, as presented in Table 3.5, which implies that the variable 'Sex' is not statistically significant as a predictor of acceptance to the medical schools. Also, the drop-in-deviance test shows its G-statistic of 2.2 and the associated p-value of 0.14. Hence, we confirm that there is not strong evidence that the odds of acceptance to the medical schools depends on the biological sex of the applicants.

3.5.1.2 GPA Score Predictor

Since the applicants' biological sex does not significantly affect the acceptance to their medical schools admission, we want to alternatively examine the relationship between applicants' GPA score and the acceptance to medical schools. Figure 3.5 shows that the admitted students have higher median of GPA scores than the rejected students.

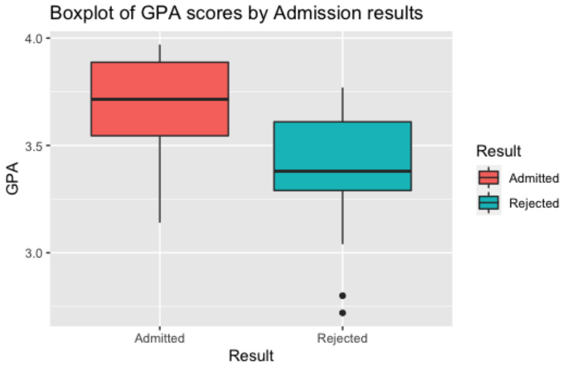

Figure 3.5: The boxplot of GPA scores by admission results

The fitted logistic regression model LR_{GPA} and the summary of it are described in Equation 3.14 and Table 3.6, respectively.

$$\therefore log(odds) = -19.21 + 0.545 * GPA(0.1unit) \tag{3.14}$$

Variable	Coef. Estimate	e^{Coef}	SE	z-statistic	p-value	Null Dev.	Residual Dev.	d.f.
GPA	0.545	1.724	0.158	3.45	0.00055	75.79	56.84	1

Table 3.6: The summary table of LR_{GPA} (Dev. represents deviance, and d.f. indicates degrees of freedom for Chi-square distribution)

We compute the odds ratio of acceptance regarding GPA scores by using the slope coefficient of LR_{GPA}: $e^{0.545} = 1.724$. This means that a 0.1-unit increase in GPA is associated with an 72.4% increase in the odds of acceptance to the medical schools. Also, the fitted probabilities of acceptance for two applicants with each 3.5 and 3.8 GPA are,

$$\hat{\pi}(GPA = 3.5) = \frac{e^{(-19.21+10*0.545*3.5)}}{1 + e^{(-19.21+10*0.545*3.5)}} = 0.466 \tag{3.15}$$

$$\hat{\pi}(GPA = 3.8) = \frac{e^{(-19.21+10*0.545*3.8)}}{1 + e^{(-19.21+10*0.545*3.8)}} = 0.817 \tag{3.16}$$

Since we confirm the randomness and independence of the MedGPA data in Section 3.5.1.1, we only examine the validity of linearity in the model LR_{GPA}. Figure 3.6 does show the reasonable extent of linearity between log (odds) of acceptance and GPA scores.

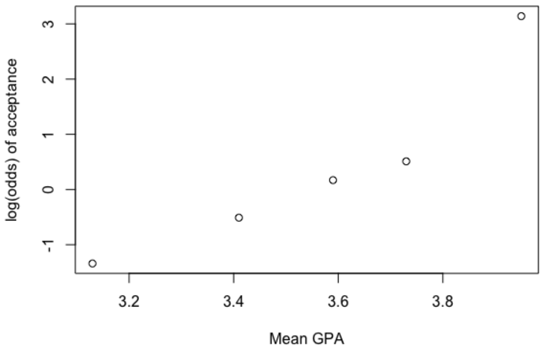

Figure 3.6: Linearity between log odds and GPA

The 95% confidence interval for the odds ratio is $(1.31, 2.45)$, and we interpret it as "We are 95% confident that an 0.1-unit increase in GPA is associated with between 31% and 145% higher odds of acceptance." Since the interval does not include 1, we conclude that higher GPA score is a significant predictor of medical school admission.

Lastly, we use two statistical tests for the overall performance of the model LR_{GPA}. According to the Wald test, the associated p-value with GPA is 0.0005, which implies that GPA is a statistically significant predictor of the acceptance to medical schools. The drop-in deviance test shows its G-statistic of 18.95 and the p-value of 0.0000134, so there is strong evidence that the odds of acceptance relies on the GPA scores. To compare the two logistic models LR_{sex} and LR_{GPA}, we can use misclassification rate. LR_{sex} has the misclassification rate of 0.4, whereas LR_{GPA} has 0.27. Therefore, we conclude that GPA performs better as a predictor of acceptance to medical schools.

3.5.2 Multiple Logistic Regression: Two or More Predictors

Next, we examine the relationship between multiple predictors and the response variable. For instance, we select MCAT to discuss whether an addition of MCAT to the model LR_{GPA} is worthwhile. In Figure 3.7, the admitted students are likely to attain higher MCAT scores. We fit the logistic regression model $LR_{GPA,MCAT}$ with using the two explanatory variables 'GPA' and 'MCAT' simultaneously. The fitted line equation is presented in Equation 3.17 and the summary in Table 3.7.

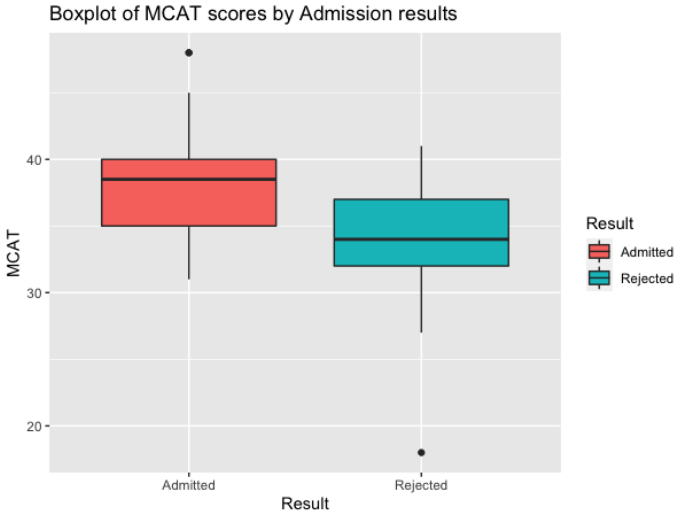

Figure 3.7: The boxplot of MCAT scores by admission results

$$\therefore log(odds) = -22.37 + 0.468 * GPA + 0.164 * MCAT \tag{3.17}$$

Variable	Coef. Estimate	$e^{Estimate}$	SE	z-statistic	p-value	Null Dev.	Residual Dev.	d.f.
GPA	0.468	1.597	0.164	2.85	0.00439	75.79	54.01	2
MCAT	0.164	1.178	0.103	1.56	0.111	75.79	54.01	2

Table 3.7: Summary table of $LR_{GPA,MCAT}$ (Dev. represents deviance, and d.f. indicates the degrees of freedom of Chi-square distribution)

We state that assuming the same MCAT score, a student with a 0.1-point higher GPA has 59.5% higher odds of the acceptance to the medical schools. Assuming the same GPA, a student with a 1-point higher MCAT scores has 17.9% higher odds of the acceptance to the medical schools. Then, we calculate the probabilities of two applicants who have different GPA scores of 3.5 and 3.8 but the same MCAT score of 37.

$$\hat{\pi}(GPA = 3.5 \text{ and MCAT} = 37) = \frac{e^{(-22.37+10*0.468*3.5+0.164*37)}}{1 + e^{(-22.37+10*0.468*3.5+0.164*37)}} = 0.519 \tag{3.18}$$

$$\hat{\pi}(GPA = 3.8 \text{ and MCAT} = 37\ ) = \frac{e^{(-22.37+10*0.468*3.8+0.164*37))}}{1 + e^{(-22.37+10*0.468*3.8+0.164*37)}} = 0.815 \tag{3.19}$$

The 95% confidence intervals for the odds ratio of GPA and MCAT in $LR_{GPA,MCAT}$ are $(1.19, 2.29)$ and $(0.974, 1.468)$, respectively. Assuming constant MCAT score, we are 95% confident that an 0.1-unit increase in GPA is associated with between 19% and 129% higher odds of acceptance. Also, under a condition of same GPA, we are 95% confident that an one-point increase in MCAT scores has between 3% lower and 47% higher odds of the acceptance to medical schools. Since the interval of GPA does not include 1, a higher GPA does impact the odds of acceptance to medical schools. However, the interval of MCAT scores includes 1, so higher MCAT scores does not significantly impact the odds of acceptance to medical schools, as long as GPA is included in the model.

Furthermore, we use two statistical tests for the overall performance of the model $LR_{GPA,MCAT}$. From the Wald test, we see that the associated p-value with GPA and MCAT is 0.00439 and 0.111, which implies that GPA is a statistically significant predictor of the acceptance to medical schools but MCAT is not. Also, the drop-in-deviance test of $LR_{GPA,MCAT}$ shows its G-statistic of 21.78 and the p-value of 0.0000187, thus confirming the overall usefulness of the model $LR_{GPA,MCAT}$.

Lastly, we compare the two logistic models LR_{GPA} and $LR_{GPA,MCAT}$ by using the nested drop-in deviance test and the misclassification rate. First, the nested drop-in deviance test checks the validity of our hypothesis that $LR_{GPA,MCAT}$ is better than LR_{GPA}. We observe that the residual deviance of LR_{GPA} is 56.84 and of $LR_{GPA,MCAT}$ is 54.01, thus the corresponding G-statistic is 2.83 and the associated p-value is 0.093. It means that there is not strong evidence to conclude that $LR_{GPA,MCAT}$ possesses more effectiveness. Also, LR_{GPA} has a misclassification rate of 0.27 and $LR_{GPA,MCAT}$ has the comparable misclassification rate of 0.25. Therefore, for simplicity concerns we may choose LR_{GPA} as the best model for predicting the odds of acceptance to medical schools, based on the Wald-test, the nested drop-in deviance test, and the comparable misclassification rate with $LR_{GPA,MCAT}$.

3.6 Advanced Topics in Logistic Regression

In light of in-depth discussion about logistic regression, the following questions have stimulated the essence of our understanding about the method. *How does a computational software compute the fitted coefficient estimates of each predictor in the logistic regression? What are the fundamentals and mathematical derivations of the calculations of the estimates? Based on what mathematical backgrounds do we confirm that the logistic regression model has good overall effectiveness of predictions?* In this section, we investigate advanced mathematical backgrounds behind the three questions with regards to logistic regression.

3.6.1 Likelihood

Likelihood is defined as the probability of the observed data, expressed as a function of parameters whose values are unfixed [2]. Suppose that a coin is flipped 50 times and 20 heads are observed. Given a probability p of getting a head on a single flip, we can compute the probability of getting 20 heads out of 50 trials, as Equation 3.20:

$$P(20\,H) = \binom{50}{20} \times p^{20} \times (1-p)^{30} \tag{3.20}$$

Since the probability $P(20H)$ depends on the value of the unknown probability p, we can define the likelihood of 20 heads out of 50 trials given that the probability of a head out on a single flip is p, as shown in Equation 3.21.

$$L(20\,H|\,p) = \binom{50}{20} \times p^{20} \times (1-p)^{30} \tag{3.21}$$

For any statistical model that contains parameters whose values are not fixed, we typically assume a probability distribution for the response variable of the model. In light of logistic regression models, we assume that a binary response variable follows Bernoulli distribution with a parameter p_i: the probability of success for an random, independent trial i [1]. Given that a probability of success for a trial i is p_i and the probability of failure is $1 - p_i$, we compute the **likelihood function** of success of the binary response variable for the trial i, as Equation 3.22:

$$L(Y = y_i|\,p_i) = p_i^{y_i} \times (1-p_i)^{1-y_i}, \text{ where } y_i = \{1 \text{ (success)}, \quad 0 \text{ (failure)}\} \tag{3.22}$$

Since the data under Bernoulli distribution are identically independent, the likelihood function of the binary response for all Bernoulli trials is defined as Equation 3.23, and the formula also applies to logistic regression models [23]:

$$L(Y = y_i | \, p_i) = \prod_{i=1} p_i^{y_i} \times (1 - p_i)^{1-y_i}, \text{ where } y_i = \{1 \text{ (success)}, \quad 0 \text{ (failure)}\} \tag{3.23}$$

3.6.2 Deviance

To measure how well a logistic regression model fits to the data of interest, we may conduct the drop-in-deviance test, called *a goodness-of-fit measurement*. This computes the statistic $G = \Delta Deviance$, the difference between the deviance under the null hypothesis and the deviance of the fitted logistic model, as discussed in Equation 3.11. Also, if the difference between these two deviance becomes larger due to the fitted logistic model, then we can state that the model fits to the data well.

Now, we rewrite Equation 3.11 as the difference in the log-likelihoods between the fitted model and the null model, as shown in Equation 3.24.

$$\therefore G = \Delta Deviance = 2[log(L_{fitted}) - log(L_{null})], \tag{3.24}$$

where log is a natural logarithm, L_{null} the likelihood of the null model M_{null} with only an intercept, and L_{fitted} the likelihood of the fitted logistic model M_{fitted} [1]. Since M_{null} does not include any explanatory variable available in the data, L_{null} is constant. Therefore, we confirm that the G-statistic is largely impacted by $log(L_{fitted})$.

Furthermore, we want the G-statistic to be as large as possible, for the larger difference between the null and residual deviance indicates a smaller p-value of Chi-square distribution under the null hypothesis. This also means that the model fits to the data well and thus is useful. Therefore, we conclude that $\mathbf{2log(L_{fitted})}$ should be *maximized* for the overall effectiveness of the fitted logistic model M_{fitted}. This also indicates that the value of L_{fitted} should reach its *maximum* value.

Therefore, a concept of likelihood is essential to compute the deviance for testing the effectiveness of a logistic regression model. The large likelihood of the model indicates a good fit between the explanatory variables used in the model and our data of interest. Thus, a computational software obtains each of the estimated coefficients of explanatory variables in the model, by maximizing the likelihood of a response variable of any given data. One methodology to find the estimated coefficients of predictors in a logistic regression model is called **Maximum Likelihood Estimation (MLE)**. This chooses parameter values of a model to maximize L_{fitted}.

3.6.3 Maximum Likelihood Estimation (MLE)

3.6.3.1 Introduction

As in Equation 3.25, the maximum likelihood estimate (MLE) of a parameter β is the fitted value $\hat{\beta}$ at which the associated likelihood function reaches its maximum [2]. Namely, this is the parameter value under which the observed data have the highest probability of occurrence [1].

$$\therefore \hat{\beta}_{MLE} = \underset{\beta}{\operatorname{argmax}}\, L(Y|X, \beta) \tag{3.25}$$

A simple example of the MLE is a game of rolling a die. Suppose that we have three types of dice, such as a 4-sided, a 6-sided, and a 10-sided die. We roll one of the three dice 5 times and obtain two 3s. Then, the probability p of obtaining one 3 varies by the type of a die that we rolled: that is, $\frac{1}{4}$ for a four-sided, $\frac{1}{6}$ for a six-sided , and $\frac{1}{10}$ for a ten-sided die. Based on the findings, we calculate the likelihood of obtaining two 3s from five trials, as in the following equations 3.26 to 3.28, where Y is the number of 3s obtained in five trials.

$$\text{4-sided die: } L(Y = 2|\, p = \frac{1}{4}) = \binom{5}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^3 = 0.264 \tag{3.26}$$

$$\text{6-sided die: } L(Y = 2|\, p = \frac{1}{6}) = \binom{5}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^3 = 0.161 \tag{3.27}$$

$$\text{10-sided die: } L(Y = 2|\, p = \frac{1}{10}) = \binom{5}{2}\left(\frac{1}{10}\right)^2\left(\frac{9}{10}\right)^3 = 0.073 \tag{3.28}$$

Of those calculations, the largest likelihood occurs at $p = \dfrac{1}{4}$, i.e. rolling a four-sided die. This means that the MLE $\hat{p}_{MLE}$ of the parameter p is $\dfrac{1}{4}$. In more contextual interpretation, the event of obtaining two 3s out of the total five rolls is most likely to occur when we roll a four-sided die.

As an extension of the above example regarding the MLE, suppose that we are not given information about how many sides the dice include. Then, the likelihood functions are modified. First, the likelihood function L_2 with a probability p of getting two 3s out of the total five rolls is,

$$L_2 \equiv L(Y = 2|\, p) = \binom{5}{2} p^2(1-p)^3 = 10\, p^2(1-p)^3 \quad (0 \le p \le 1) \tag{3.29}$$

Also, the likelihood function L_4 with a probability p of getting four 3s out of the total five rolls is,

$$L_4 \equiv L(Y = 4|\, p) = \binom{5}{4} p^4(1-p)^1 = 5\, p^4(1-p) \quad (0 \le p \le 1) \tag{3.30}$$

In Figure 3.8, we observe that the two likelihood functions reach maximum values at different p: L_2 appears to be maximized near $p = 0.4$ whereas L_4 near $p = 0.8$. Therefore, the MLE of L_2 is $\hat{p} = 0.4$ and the MLE of L_4 is $\hat{p} = 0.8$. Hence, the result of two 3s out of the total five rolls is more likely to occur when $p = 0.4$ than any other p value between 0 and 1 [2]. Also, the result of four 3s out of the total five rolls occurs at most when $p = 0.8$, compared to any possible p in the interval $[0, 1]$.

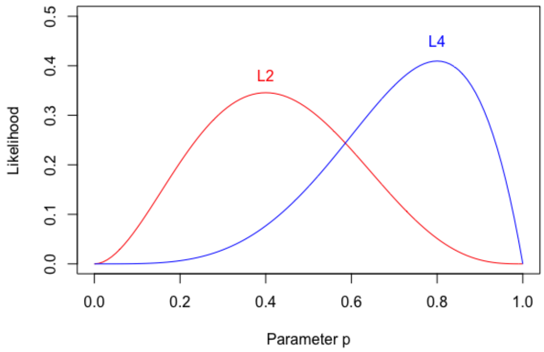

Figure 3.8: The plot of L_2 and L_4, the likelihood functions for two 3s and four 3s, respectively

3.6.3.2 Calculations of MLE

Rather than approximating the MLE of a parameter by either looking at the plot or comparing the possible parameter values, we now want to calculate the MLE of the parameter directly from the likelihood function. Notice that the MLE $\hat{\beta}_{MLE}$ of $L(Y|\beta)$ also maximizes $\log(L(Y|\beta))$, the natural log-transformation of the original likelihood function $L(Y|\beta)$ [1]. In particular, the log-transformation can simplify the computation of MLE, as it will turn the powers and products into the products and sums. Hence, we take the following two steps to calculate the MLE of a parameter with regard to a certain likelihood function.

1. Take the natural logarithm of the original likelihood function of interest.
2. Optimize the log-transformed likelihood by differentiating it in terms of the parameter of interest and setting the differentiated equation to 0. The obtained parameter value is the MLE of the parameter.

The MLEs of p for the likelihood function L_2 in Equation 3.29 and L_4 in Equation 3.30 are computed as below, and the obtained MLEs for L_2 and L_4 also match the findings from Figure 3.8.

$$\begin{gathered}\log(L_2) = \log(10p^2(1-p)^3) = \log 10 + 2\log p + 3\log(1-p) \quad (0 < p < 1)\\ \frac{d\log(L_2)}{dp} = \frac{2}{p} + 3 \cdot \frac{1}{1-p} \cdot (-1) = 0\\ \therefore \frac{2(1-p)-3p}{p(1-p)} = \frac{2-5p}{p(1-p)} = 0\end{gathered}$$

$$\therefore \hat{p}_{MLE_{L_2}} = \frac{2}{5} = 0.4 \tag{3.31}$$

$$\begin{gathered}\log(L_4) = \log(5p^4(1-p)) = \log 5 + 4\log p + \log(1-p) \quad (0 < p < 1)\\ \frac{d\log(L_4)}{dp} = \frac{4}{p} - \frac{1}{1-p} = 0\\ \therefore \frac{4(1-p)-p}{p(1-p)} = \frac{4-5p}{p(1-p)} = 0\end{gathered}$$

$$\therefore \hat{p}_{MLE_{L_4}} = \frac{4}{5} = 0.8 \tag{3.32}$$

3.6.3.3 MLE for Logistic Regression

Lastly, we examine the MLE of each parameter of a logistic regression model. We use the maximum likelihood estimation to find the coefficient estimate of each predictor that maximizes the likelihood function of the fitted logistic model.

For any binary logistic regression model, each response value y follows Bernoulli distribution such that any y will take either 1 or 0, with the probability $p(x)$ and $1 - p(x)$, respectively. For each data point, suppose there is a vector $< x_i, y_i >$ of explanatory features x_i and its response y_i. Since $p(x_i)$ is defined as the probability of success (or $y_i = 1$) for x_i, we can rewrite Equation 3.23, the likelihood function for the logistic regression, as shown in Equation 3.33.

$$\mathbf{L}(\beta) = L(Y = y_i|\, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i}\,(1-p(x_i))^{1-y_i} \quad (y_i = 1\,(success),\;\; 0\,(failure)), \tag{3.33}$$

where $p(x_i)$ is defined by the logistic model, as in 3.34 and 3.35.

$$logit(p(x_i)) = log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n \tag{3.34}$$

$$\therefore \mathbf{p(x_i)} = \frac{e^{\beta_0+\beta_1 x_1+...+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+...+\beta_n x_n}} \tag{3.35}$$

Following the two steps of obtaining the MLE, we first take a natural log of the likelihood function $L(\beta)$ defined from Equation 3.33.

$$\begin{aligned} logL(\beta) &= log\left(\prod_{i=1}^{n} p(x_i)^{y_i}\,(1-p(x_i))^{1-y_i}\right) \\ &= \sum_{i=1}^{n} (y_i\, logp(x_i) + (1-y_i)\, log(1-p(x_i))) \\ &= \sum_{i=1}^{n} (y_i\, logp(x_i) + log(1-p(x_i)) - y_i\, log(1-p(x_i))) \end{aligned}$$

$$\therefore logL(\beta) = \sum_{i=1}^{n} \left(log(1-p(x_i)) + y_i\, log\left(\frac{p(x_i)}{1-p(x_i)}\right)\right) \tag{3.36}$$

We know from Equation 3.34 that $log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i$, so the log-likelihood function $log(L(\beta))$ is

$$log(L(\beta)) = \sum_{i=1}^{n} (log(1-p(x_i)) + y_i(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i)) \tag{3.37}$$

Also, we replace $p(x_i)$ with Equation 3.35:

$$\begin{aligned} log(1-p(x_i)) &= log\left(1 - \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_i x_i}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_i x_i}}\right) \\ &= log\left(\frac{1}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_i x_i}}\right) \\ &= log1 - log(1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_i x_i}) \end{aligned}$$

$$\therefore log(1-p(x_i)) = -log(1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_i x_i}) \tag{3.38}$$

Therefore, the log-likelihood of $L(\beta)$ is,

$$\therefore logL(\beta) = \sum_{i=1}^{n} \left(-log(1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_i x_i})\right) + \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i) \tag{3.39}$$

The second step of obtaining the MLE is to differentiate the log likelihood with respect to each parameters β_j, set the derivatives equal to 0, and find the $\hat{\beta}_j$.

$$\frac{d\ logL(\beta)}{d\beta_j} = -\sum_{i=1}^{n} \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_i x_i}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_i x_i}} * x_j + \sum_{i=1}^{n} y_i x_j \tag{3.40}$$

$$= \sum_{i=1}^{n} x_j(y_i - p(x_i)) = 0 \tag{3.41}$$

Notice that Equation 3.40 is non-linear and do not have a closed form of solutions, so we use iterative optimization to numerically approximate the MLE $\hat{\beta}_{j_{MLE}}$ for each parameter β_j [1, 20]. Therefore, we can define the MLE $\hat{\beta}_{j_{MLE}}$ for each parameter β_j, as Equation 3.42.

$$\therefore \hat{\beta}_{j_{MLE}} = argmax \sum_{i=1}^{n} x_j(y_i - \hat{p}(x_i)), \tag{3.42}$$

where $\hat{p}(x_i) = \dfrac{e^{\hat{\beta}_0+\cdots+\hat{\beta}_i x_i}}{1+e^{\hat{\beta}_0+\cdots+\hat{\beta}_i x_i}}$ is the ML estimate of the probability of success p for x_i.

So far, we have examined the derivation of how to obtain the estimated coefficient of each predictor in the logistic regression model. Based on Equation 3.42, the computational software iteratively calculates the MLE of each parameter β_j using the previous β_{j-1}. Possible numerical optimization methods can be the Newton-Raphson method or the stochastic gradient ascent, as these are the most popular iterative optimization algorithms when performing parameter estimation for logistic regression models [1, 20].

CHAPTER 4

Support Vector Machines & Metrics of Prediction

The second objective of our research is to predict depressive disorders diagnosis among the U.S. adults by using several supervised machine learning methods. In this chapter, we first investigate *support vector machines*, which is widely renowned for its superb classification tasks. Note that the depth of our investigation toward support vector machines is comparatively shallow, as our study focuses more on mining attributes of depressive disorders for each group by using decision trees and logistic regression. Lastly, we explore various metrics of model performance when testing the prediction tasks: *confusion matrix*, *accuracy*, *precision*, *F1 score*, *recall*, and *ROC curve*.

4.1 Support Vector Machines

4.1.1 Historical Background

Support vector machines (SVMs) are a widely-used supervised machine learning algorithm, first developed by a Russian statistician Vladimir Vapnik and his coworkers in the field of computer science [14]. Then, the concepts of SVMs were introduced in the two eminent books of statistics: *Statistical Learning Theory* and *The Nature of Statistical Learning Theory* in early 1970s [14].

Since then, the algorithm of SVMs has been spotlighted because of its distinctness from the traditional learning theories, where SVMs do not require dimension reduction in the high-dimensional data but find the optimal hyperplane that separates the data. In addition, the comparatively simple algorithm of the SVMs vastly expedites the application of SVMs to wider variety of fields, particularly two-class classification tasks.

4.1.2 Linear Support Vector Machine Classifiers

Support vector machines (SVMs) have been utilized in many fields for classification and regression. Since a goal of our study is to predict the diagnosis labels of depressive disorders among the U.S. adults, we examine *support vector machine classifiers (SVMCs)* for predicting data with one of the two responses (i.e. yes or no) for depressive disorders.

The main idea behind SVMCs is to find a line (or **hyperplane**) that separates the data of interest into two or multiple areas for each category of a response variable in the best manner. Then, we predict the labels of incoming data by looking at the area that each data point is placed on. For example, in Figure 4.1 there are two bands of blue and green points, and using the linear SVMC we separate the bands into two areas. We observe that future data points above the line are classified as blue, and those below green. The objective of SVMC is to find a line or hyperplane that has the widest margin, with each of the two decision boundaries going through the closest point in each band of points [4].

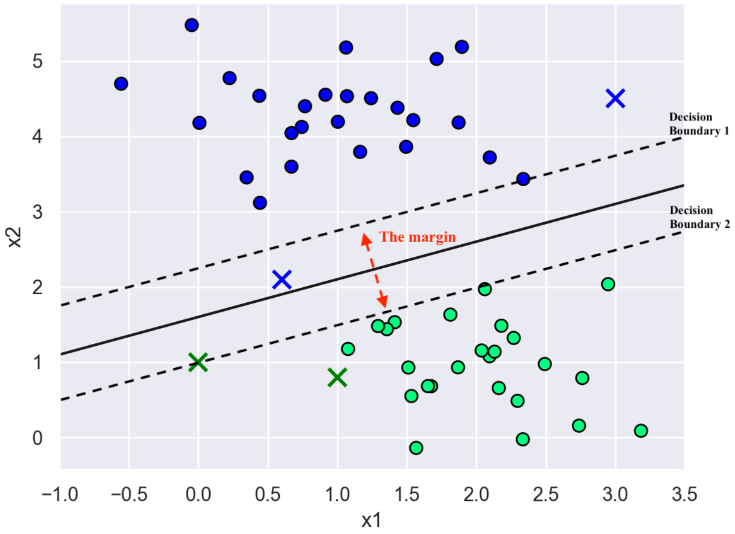

Figure 4.1: The data separated by a linear SVMC (The solid line is the optimal hyperplane that has the widest margin between the two dashed boundary hyperplanes.

Then, how do we find the optimal hyperplane that has the largest margin? Suppose that we have a data set of explanatory vectors $X = \{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n\}$ where $\overline{x}_i \in R^n$ and a two-class vector $Y = \{y_1, y_2, \ldots, y_n\}$ where $y_i \in \{-1, 1\}$. Then, we find the best hyperplane with the maximized margin, as described in Equation 4.1 [4]:

$$\overline{w}^T \overline{x} + b = 0, \tag{4.1}$$

where $\overline{w} =< w_1, \ldots, w_n >$ is the weight vectors, $\overline{x} =< x_1, \ldots, x_n >$ the inputs, and b is bias. Then, for any linearly separable data, the two classes can be split by a margin with two possible boundaries, where some data points from each class, called **support vectors**, lie on the boundaries. In this example, these two boundaries are given by Equation 4.2. Also, we observe in Figure 4.2 that there are four support vectors, two from each class. If we substitute any input vector $\overline{x}_i$ into the equation $\overline{w}^T\overline{x} + b$ and the corresponding output value is above 1, then the vector is placed in the area of blue points. Otherwise, the vector is placed in the area of green points if the output value is below -1.

$$\begin{aligned} \overline{w}^T\overline{x} + b &= 1 \\ \overline{w}^T\overline{x} + b &= -1 \end{aligned} \tag{4.2}$$

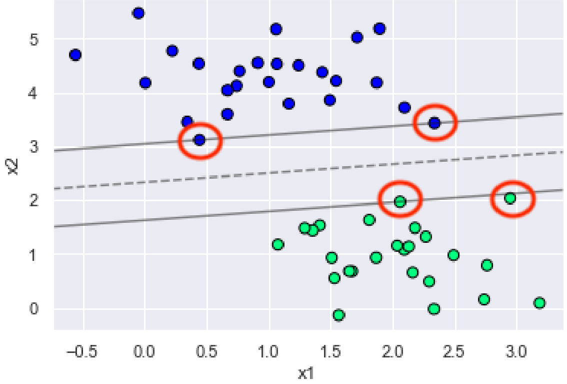

Figure 4.2: The four support vectors marked with red circles: the above solid line is given by the equation $\overline{w}^T\overline{x} + b = 1$ and the below solid line is $\overline{w}^T\overline{x} + b = -1$. The dashed line indicates the optimal hyperplane that has the widest margin from these two boundaries.

Since the main objective of SVMC is to maximize the margin between these two parallels (or boundary hyperplanes) that divide the classes, we select some two vectors that are perpendicular to and ended on each decision boundary, $\overline{x}_1$ and $\overline{x}_2$. Figure 4.3 shows an arbitrary example of the $\overline{x}_1$ and $\overline{x}_2$ on a random data. Then, we compute the distance of these two vectors, which is defined by the length of the margin between the two boundaries, as expressed in Equation 4.3 [4].

$$\overline{x}_2 - \overline{x}_1 = t\overline{w}, \tag{4.3}$$

where the t is the margin. Then, Equation 4.2 can be modified as Equation 4.4:

$$\begin{gathered} \overline{w}^T\overline{x}_2 + b = \overline{w}^T(\overline{x}_1 + t\overline{w}) + b = (\overline{w}^T\overline{x}_1 + b) + t||\overline{w}||^2 = 1 \\ \therefore t = \frac{2}{||\overline{w}||^2} \quad (\because \overline{w}^T\overline{x}_1 + b = -1) \end{gathered} \tag{4.4}$$

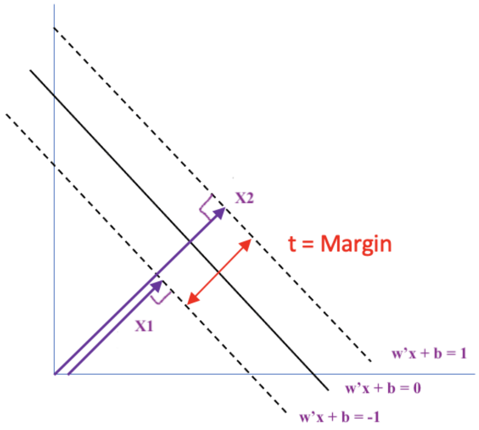

Figure 4.3: The two selected vectors $\vec{x_1}$ and $\vec{x_2}$ for each decision boundary. w' is the transposed weight vectors w.

Since the distance $d(\overline{x}_1, \overline{x}_2)$ is the length of the margin t, we obtain the value of $d(\overline{x}_1, \overline{x}_2)$, as expressed in Equation 4.5.

$$d(\overline{x}_1, \overline{x}_2) = ||t\overline{w}|| = t||\overline{w}|| = \frac{2}{||\overline{w}||^2} \times ||\overline{w}|| = \frac{2}{||\overline{w}||}, \tag{4.5}$$

where $||\overline{w}||$ is the normalized weight vectors. Using the training set, we minimize the value of $||\overline{w}||$ to maximize the distance of $\overline{x}_1$ and $\overline{x}_2$. In addition, since every data point will be classified into one of the two classes, we can impose the constraint into the outputs of every points in each class, as described in the bottom of Equation 4.6 [4].

$$y_i(\overline{w}^T \overline{x}_i + b) \geq 1, \ \forall(\overline{x}_i, y_i), \tag{4.6}$$

where y_i is the class value of $\overline{w}^T \overline{x}_i + b$ - that is, either -1 or 1. Finding the minimum value of $||\overline{w}||$ can be solved by using the Lagrange Multipliers technique, which is beyond the scope of our exploration of support vector machines. Instead, we will use the following computational functions *linearSVC* or *SVC (kernel = 'linear')* in Python machine learning library *Scikit-learn*, which automatically compute the minimized normal weight vector $||\overline{w}||$ and the two decision boundaries as a result.

4.1.3 Non-linear Support Vector Machine Classifiers

We face many data sets that are not linearly separable, such as Figure 4.4. Then, we are not able to use linear SVMCs to separate the points into one of the two classes. A possible solution of this problem is to manipulate the data to become linearly separable, known as **kernel tricks**. The kernel trick is a machine learning technique that projects the original vectors into higher dimensional spaces by using several *kernel functions*, where the projected data thereby become linearly separated [4]. Then, we can apply SVMCs to this modified data and classify the points into the two classes.

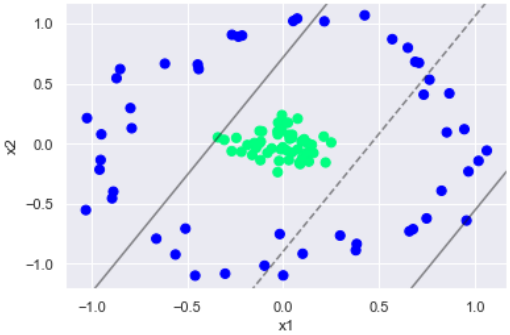

Figure 4.4: An example of non-linearly separable data: the circular data set

The following are three kernel functions K that have been widely used for the kernel trick, given the two-dimensional vectors $\overline{x_1}, \overline{x_2}, \ldots, \overline{x_n}$ with $\overline{x_i} =< x_1, x_2 >$ for all $i = 1, 2, \ldots, n$. Note that γ is a constant term that determines the amplitude of the function, influenced only by the distance [4].

1. The Radial Basis Function (RBF): $K(\overline{x_i}, \overline{x_j}) = e^{-\gamma \|\overline{x}_i - \overline{x}_j\|^2}$

2. The polynomial kernel function: $K(\overline{x_i}, \overline{x_j}) = (\gamma \overline{x}_i^T \overline{x}_j + r)^c$, where c is the parameter degree.

3. The sigmoid kernel function: $K(\overline{x}_i, \overline{x_j}) = \dfrac{1 - e^{-2(\gamma \overline{x}_i^T \overline{x}_j + r)}}{1 + e^{-2(\gamma \overline{x}_i^T \overline{x}_j + r)}}$.

Let us use the RBF kernel functions to transform the original data in Figure 4.4 from non-linear spaces to linear ones. As visualized as a 3D plot in Figure 4.5, we first add a third dimension defined by the RBF kernel function of the original vectors $\overline{x}_i$ and $\overline{x}_j$. Figure 4.5 displays that the two bands of either blue and green data points are now trivially linearly separable. Then, fitting a RBF-kernel SVMC to the three-dimensional space of this modified data can separate the two bands of data points by the red plane $r = 0.7$, as shown in Figure 4.6.

Indeed, if we make a two-dimensional plot of the original data in Figure 4.4 and apply the RBF-kernel SVMC to every data point, we observe two solid circular decision boundaries and support vectors lying on each circular boundary hyperplane, as descried in Figure 4.7. Likewise, the dashed circular hyperplane drawn by the RBF-kernel SVMC has the maximized margin.

Furthermore, we can apply other kernel functions to non-linearly separable data with an increased dimension. Then, we may compare and confirm which kernel function produces the maximized decision boundaries that separate the data into each given class.

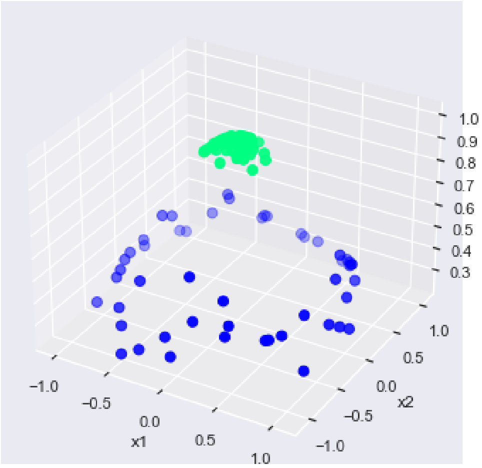

Figure 4.5: 3D-transformed linear data expanded by the RBF kernel in the z-axis

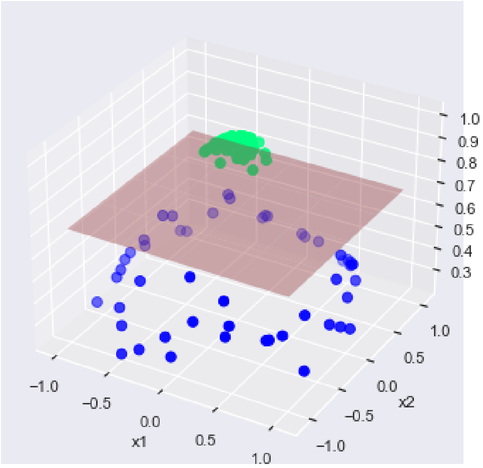

Figure 4.6: 3D-transformed linear data separated by with the SVMC hyperplane

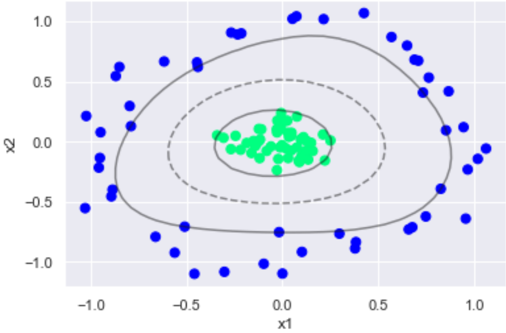

Figure 4.7: The RBF kernelized SVMC to the non-linearly separable data: the dashed circle corresponds to the optimal (3D) hyperplane shown in Figure 4.6.

4.1.4 Soft-margin Classification

All the examples above have perfect and clean decision boundaries: that is, every data point of each class is clearly located in the area of the class. They are called as **hard-margin classification**. However, there are many data set that are not clearly separable, such as the data points in Figure 4.8 where two groups of data points are overlapped in some areas.

In this case, neither a linear SVMC nor kernel tricks would separate the data clearly into one of the two classes. Thus, we use another technique of SVMC, called **soft-margin classification**. This creates a flexible model that allows some of the points to "creep into" the margin as long as this process results in a better performance in classification [4, 24]. Therefore, soft-margin classification not only maximizes a margin but also limits the number of misclassified points in margin violations. Then, the extent of allowing outliers is determined by a tuning parameter C, which controls the

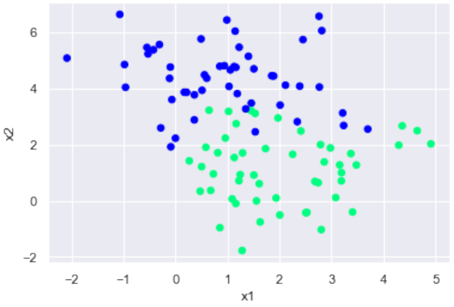

Figure 4.8: The data with overlapped areas of each class

trade-off between the smooth margin and the correct classification of the points [24]. For example, if the value of C is high, then the SVMC prioritizes classifying all data points correctly, thus resulting in a narrower (or "hard") margin. On the other hand, smaller C intends to find the widest margin, thus producing some incorrect classification scores.

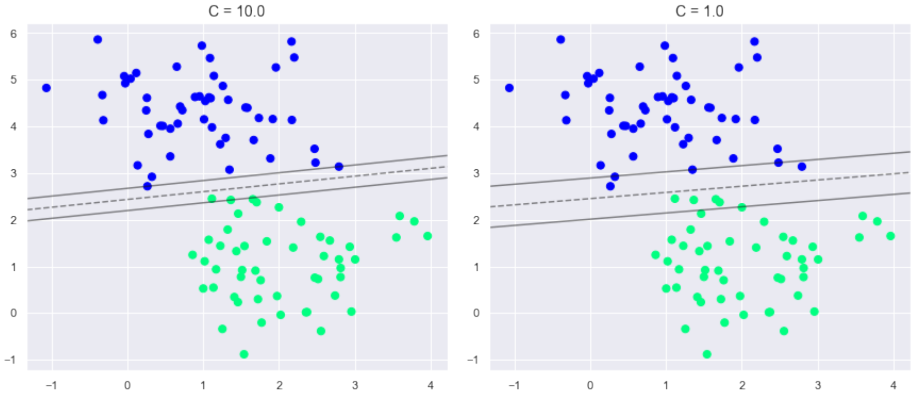

Figure 4.9: An example of soft-margin classification with different values of C. The dotted line is the optimal hyperplane that has the widest margin from the two solid decision lines.

In Figure 4.9, we observe that the SVMC with $C = 10.0$ has a narrower margin but higher rates of correct classification, whereas the SVMC with $C = 1.0$ has a wider margin but encompasses some data points within the margin area. Finding the optimal value of C can be performed by using either cross-validation or a function *GridSearchCV* in Python *Scikit-learn* library, which seeks the best choice of C and other parameters automatically [4, 24].

4.1.5 Assessment of SVMC algorithm

SVMCs are renowned as powerful classifiers. First, the SVMC depends their training process only on a few support vectors with spending a low cost on time and space, thereby producing fast classification. Hence, it can handle high-dimensional data with high precision. In addition, the kernel tricks can be adapted to many types of data [11]. However, the classification results of SVMCs rely significantly on the tuning parameter C, which may increase the computational cost. Furthermore, the SVMC is a black-box model, which does not show the internal procedures that the results have been made by.

4.2 Metrics of Prediction

In this section, we explore some widely-used metrics of prediction. The first important metric is the **accuracy**, which is the ratio of correct classifications to the total size of data. However, the accuracy is not always an appropriate metric for classifiers, particularly when classifying skewed data in which a class is more frequent than the others [11].

The second metric is the **confusion matrix**, also known as the misclassification table from Chapter 3. The confusion matrix counts the number of instances of predicted class ’Yes’ as either actual class ’Yes’ or ’No.’ In other words, we compute the following four possibilities: true positive, true negative, false positive, and false negative cases [11].

	Predicted class		
Actual class		No	Yes
	No	230	12
	Yes	15	210

Table 4.1: An example of confusion matrix

Table 4.1 shows that 230 cases of No are correctly classified (true negatives), while the remaining 12 are wrongly classified as Yes (false positives). The second row also displays that 15 cases of Yes are incorrectly classified as No (false negatives) and the remaining 210 are correctly classified as Yes (true positives). If a confusion matrix have non-zero values only on its diagonal, then the corresponding classifier is said to be perfect [11].

The third metric is the **precision**, which is the accuracy of the positive predictions. It is also defined as the ratio of the number of true positives to the number of all predicted positives.

$$precision = \frac{TP}{TP + FP}, \tag{4.7}$$

where TP is the number of true positives and FP the number of false positives. It tests whether the classifier is able to detect the characteristics that determine the positiveness of sample and to avoid misclassification as negative [4]. Precision is typically computed along with another metric called **recall**, which calculates the ratio of true positive cases to all the actual positives.

$$recall = \frac{TP}{TP + FN}, \tag{4.8}$$

where TP is the number of true positives and FN the number of false negatives. For many cases, we observe that despite the same objectives of the two metrics, recall is usually lower than precision because the number of false negatives is proportionally higher than the number of false positives. Therefore, we use another metric, called **F1 score**, which harmonizes the weighted mean between precision and recall. To be specific, this harmonized mean places more weights to recall.

$$F_1 = \frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{TP}{TP + \frac{FN + FP}{2}}, \tag{4.9}$$

where TP is the number of true positive cases, FN the number of false negatives, and FP the number of false positives [11]. Higher precision outputs the highest F_1 score, while higher recall gives the least F_1 score. Hence, F_1 score performs as a trade-off between high precision and a limited number of false negatives [4]. All the procedures of computing those four classification metrics can be conducted by using *sklearn.metrics* from a scikit-learn library.

The last metric is the **ROC curve** (receiver operating characteristics curve) that enables us to compare several classifiers by assigning a prediction score to each classifier [4]. It plots the true positive rate (TPR) against the false positive rate (FPR) at different thresholds on the curve, where the TPR is equal to *recall* and the FPR is the ratio of negative cases that are incorrectly classified as positive. An example of ROC curve is shown in Figure 4.10, where the yellow curve is the ROC curve of a given classifier and the dashed line represents the ROC curve of a random classifier [25]. Hence, any classifier with a ROC curve above the dashed threshold line performs better than the random classifier. Therefore, the best classifier has a ROC curve that consists of two line segments: one from $(0, 0)$ to $(0, 1)$ and the other from $(0, 1)$ to $(1, 1)$. Then, we aim to find the classifier whose ROC curve should be as close as possible to these line segments [4].

A measure of prediction in light of ROC curves is to compute the **area under curve (AUC)** of the curves. The value of AUC lies between 0 and 1, where the perfect classifier has AUC = 1, the worst one has AUC = 0, and the random classifier has AUC = 0.5. Hence, higher AUC means higher accuracy of classification. Therefore, by plotting the ROC curves and computing the corresponding

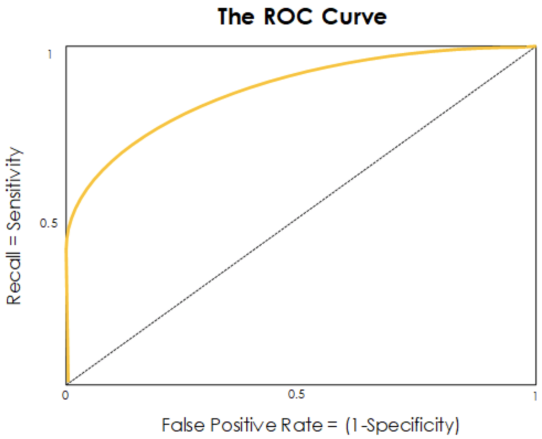

Figure 4.10: An example of a ROC curve [25]

AUC values of multiple classifiers simultaneously, we are able to visualize and compare the quality of classification of each classifier.

4.3 Real-world Example: Breast Cancer Classification

In this section, we apply support vector machine classifiers (SVMCs) to a simple real-world data set, known as *Breast Cancer Wisconsin Diagnostics.* The original data can be retrieved from UCI Machine Learning Repository, but instead we load and fetch the same data from a module `sklearn.datasets.load_breast_cancer` in the Scikit-learn library. The data set of breast cancer consists of 569 rows and 31 columns, which also includes a class column of two tumor states: *malignant (cancer)* and *benign (non-cancer)*. Note that there are no missing values in the data.

In order to classify the states of tumors into either malignant or benign, we first split the entire data into a training set and a testing set. For example, we have 70% of the randomly shuffled data as a training set and the remaining 30% as a testing set. Then, we normalize both the training and testing sets to keep all the values of each attribute within the range $[0, 1]$. Next, we train the RBF-kernel SVMC and $C = 1.0$ as a default by using the training set, and then we predict the labels of each example in the testing set with our trained SVMC. Table 4.2, 4.3, and Figure 4.11 present the confusion matrix, classification report, and the ROC curve for the trained SVMC model.

		Predicted	
		Benign	Malignant
Actual	Benign	104	5
	Malignant	2	60

Table 4.2: The confusion matrix of breast cancer diagnostics with the SVMC

	Precision	Recall	F1 score
"Benign"	0.98	0.95	0.97
"Malignant"	0.92	0.97	0.94
macro average accuracy	0.95	0.96	0.96
weighted average accuracy	0.96	0.96	0.96

Table 4.3: Classification report of breast cancer diagnostics with the SVMC

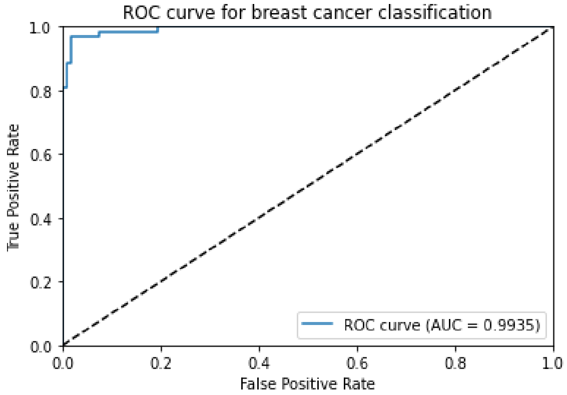

Figure 4.11: The ROC curve of breast cancer diagnostics with the default RBF-kernel SVMC

Based on Table 4.2, we compute the accuracy of the SVMC performance, which is $\frac{104+60}{104+5+2+60} = 0.959$. Also, only two people with actual malignant tumors were predicted as benign ones. Furthermore, the values of precision, recall, and F1 scores are on average from 0.95 or 0.96, as shown in Table 4.3. All those metrics indicate that the default SVMC model performed very well in predicting tumor states of the examples in the given data. Lastly, the ROC curve in Figure 4.11 is almost close to the perfect one, and its AUC value is 0.9935. Thus, we identify the good performance of this SVMC on the breast cancer prediction.

However, we may consider the possibility of overfitting, regardless of the good prediction results of the SVMC. Therefore, we optimize the tuning parameters, such as C, γ, and the kernel function, in order to increase the extent of generalization. All the optimization process can be done by using a function `GridSearchCV` in the Scikit-learn module `sklearn.model_selection`, which grid-searches the optimal parameters to the SVMC. When we consider only two non-linear kernel functions - RBF and polynomial, we find that the polynomial kernel with $C = 0.1$ and $\gamma = 1$ is the best parameters for the SVMC in this data.

According to Table 4.4, the accuracy of the prediction made by the optimized RBF SVMC is $\frac{105+60}{105+4+2+60} = 0.965$. Also, according to Table 4.5, the optimized SVMC has improved values

	Predicted		
Actual		Benign	Malignant
	Benign	105	4
	Malignant	2	60

Table 4.4: The confusion matrix of the optimized SVMC

	Precision	Recall	F1 score
"Benign"	0.98	0.96	0.97
"Malignant"	0.94	0.97	0.95
macro average accuracy	0.96	0.97	0.96
weighted average accuracy	0.97	0.96	0.97

Table 4.5: The classification report of breast cancer diagnostics with the optimized SVMC

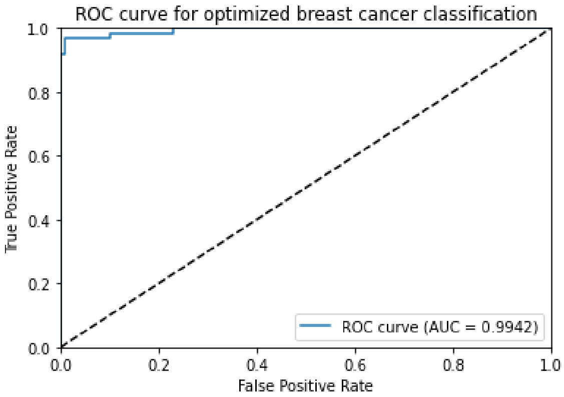

Figure 4.12: The ROC curve of breast cancer diagnostics with the optimized SVMC

of precision, recall, and F1 scores within a range of 0.96 and 0.97, compared to the default SVMC. Furthermore, the ROC curve in Figure 4.12 seems almost perfect, and its AUC value is 0.9942, thereby indicating excellent performance of the optimized SVMC. Therefore, we conclude that the optimized polynomial-kernel SVMC is said to be a better model than a default RBF-kernel SVMC.

CHAPTER 5

Data Description

In this chapter, we examine the data set of our interest, ***2018 BRFSS***, and perform data transformation to produce the finalized data set. Lastly, we conduct exploratory data analysis of three adult groups in our data set, thus enhancing our understanding of the characteristics of each age group.

5.1 Data Source Investigation

The data set of interest is **the Behavioral Risk Factor Surveillance System (BRFSS)** [17], collected in 2018 and released by the Centers of Disease Control and Prevention (CDC) in July 2019. The BRFSS is composed of self-reported responses to a health-related questionnaire distributed through telephone surveys. Also, the surveys collect uniform and state-specific data on health risk behaviors, chronic illnesses conditions, healthcare access, and use of preventive health services, which influence the main factors of death and disability of people in the United States [17]. Its target respondents are the non-institutionalized adult population over 18 years old who reside in the 50 states of the United States, the District of Columbia, and participating U.S. territories (i.e., Guam and Puerto Rico).

The data collection of the BRFSS is operated through a collaboration among health departments of all 53 participating regions and the Centers for Disease Control and Prevention (CDC). To be specific, state health departments cooperate with the CDC in designing the process, conduct their telephone surveys on their randomly selected residents each month, and transmitting the collected responses to the CDC for data editing, processing and analysis [17]. Also, due to the diminishing population of landline telephone users, the BRFSS also includes the responses from cellular telephone surveys on randomly selected people by using a weighting methodology called *raking*. This method adds indicators of demographic characteristics of the participants. Therefore, the dual-frame telephone survey format increases the validity, data values, and the representativeness of the BRFSS [17].

The BRFSS questionnaire has been influenced by established national health surveys, such as the National Health Interview Survey (NHIS) and the National Health and Nutrition Examination Survey (NHANES) [17]. Thus, the reflection of the these eminent surveys in the BRFSS guarantees verification of the responses to questions in the BRFSS.

The questionnaire consists of three main parts: core component, optional BRFSS modules, and state-added questions. First, the core component is a standard set of questions that all 53 participating regions ask to participants, such as current health conditions, potential risk behaviors, and demography. However, optional BRFSS modules and state-added questions are not required to be asked for all regions; they depend on the discretion of the corresponding state health departments. For the sake of eliciting uniform and general inferences, we therefore decide to focus only on **the core component questionnaire** for our study about depressive disorders among the U.S. adults in 2018.

5.2 Data Transformation

Rather than applying machine learning methods to the original data set, we aim to first produce a cleaned, organized version of the BRFSS that contains valid responses only from the core component questions. All the transformation process have been conducted by the statistical software *R*.

5.2.1 Data Cleaning & Wrangling

The core component questionnaire of the original BRFSS consists of 437,436 rows and 84 columns. First, we remove some unnecessary columns of the original BRFSS that are not needed in our study, such as the identifiers of participants. Examples of the respondent identifiers are listed in Table 5.1.

Code	Name	Definitions
STATE	State	State identification
IYEAR	Interview year	The year of interview conducted
PVTRESD1	Private Residence	Is this a private residence?
CELLFON5	Cellular phone	Is this your cell phone?
CADULT	Adult identifier	Are you 18 years of age or older?

Table 5.1: Removed participant identifiers from the BRFSS

Second, we convert all the remaining columns of the BRFSS to be either numerical or categorical. Then, we set the following variable as the target variable of our study: **ADDEPEV2** - ***(Ever told) Do you have a depressive disorder, including depression, major or depression, and dysthymia?*** This question consists of four possible answers from the participants who are asked: *Yes, No, Don't*

Know/Not Sure, and *Refused*. However, since the responses of 'Not Sure' and 'Refused' take up only 0.49% of the total responses, we delete those two unnecessary responses from the column 'ADDEPEV2' and select only the two binary responses ('Yes' and 'No').

Third, we remove irrelevant responses from the explanatory variables in the BRFSS. Each explanatory variable has responses of either '7' or '9', which are not to be included in the data analysis process [17]. Therefore, we examine all explanatory variables in the core component questionnaires and delete these two responses in the columns of explanatory variables. In general, the answers 7, 77, or 7 represent the response of 'Don't know/Not sure', while 9 or 99 indicates 'Refused.'

Up to this point, the BRFSS consists of only the valid responses from the 84 columns of all the 17 core sections questionnaires. Then, we add the column of age from the calculated variables section. Noticing that there are no missing values in the column of age in the BRFSS, we divide the entire BRFSS responses by three categories of age groups for adults: **young adults (18-39), middle-aged adults (40 - 60), and older adults (61 - 85).**

Lastly, we manipulate missing values in the BRFSS. Omitting all missing values in the data set would result in reducing a vast size of the available responses, thus lacking generalization of certain conclusions generated by this study. In order to cope with this problem, we employ an imputation strategy: the missing values are replaced with the mean if the corresponding variable is numeric, or with the mode if the variable is categorical. As the BRFSS data is randomly collected, we can assume that those missing values are likely close to the mean or mode of the distribution of the columns [3]. Therefore, we utilize the mean/mode imputation to handle with the missing values in the data. Hence, we generate a finalized data set of use that consists of 238,219 rows and 85 columns.

5.2.2 Feature Selection

In the previous section, we produced a finalized data set of the BRFSS, which contains 84 explanatory columns and 1 response column as a result. However, it is reasonable to assume that not all of 84 explanatory columns will be good predictors of the column ADDEPEV2. Using irrelevant or partially relevant columns can negatively impact the performance of any predictive models. Therefore, we investigate which explanatory variables out of the total 84 contribute most to the ADDEPEV2.

Feature selection algorithms automatically select explanatory variables that are most relevant to a target variable. Hence, it improves the accuracy of prediction with less time complexity and lower chances of overfitting [5]. One methodology is to calculate the **feature importance** of each explanatory variable available in the data, which represents the relative importance of each variable when making a prediction [6].

To compute feature importance, we examine **CART importance**, which computes the Gini importance scores of each variable available in the data and compares one with the others. We decide on the largest score as it contributes most to predicting a response variable. The Gini importance of an explanatory variable is defined as the total decrease in Gini impurities from splitting on the variable, averaged over all trees [19]. Namely, if the total decrease in Gini impurities due to placing the variable on a node increases, the importance of the variable increases. To find the importance of each variable, we will use Scikit-learn which automatically implements its embedded function `feature_importances_` when implementing a decision tree classifier.

5.2.2.1 Decision Trees

Feature selection for a decision tree can be readily performed by limiting the maximum depth of the tree. In Scikit-learn, we use a parameter named `max_depth` of the function `DecisionTreeClassifier` for limiting the depth of the tree `clf`, as described in the code line below:

```
clf = DecisionTreeClassifier(max_depth = 4, criterion = 'gini')
```

After computing the Gini impurity and importance scores of each explanatory variable in the data, this code will build a binary decision tree that has a depth of 4. The selected variables, descriptions, and importance values are presented in Tables 5.2 (young adult group), 5.3 (middle-aged adult group), and 5.4 (older adult group).

5.2.2.2 Logistic Regression

Unlike decision trees, a logistic regression does not automatically perform feature selection when it constructs a predictive model. Thus, we first determine to utilize the CART importance scores of each explanatory variable selected by decision trees.

Then, we also perform a **Chi-square test** to observe an association between categorical variables and a response variable ADDEPEV2 by examining the associated p-value of Chi-square statistic of each categorical variable. We report that for the young adult group, all categorical variables from Table 5.2 are relevant to the response variable *ADDEPEV2*, as their associated p-values are close to 0. For the middle-age group, all from Table 5.3 but IMFVPLAC are related to the response variable. In terms of the older adult group, all categorical variables in Table 5.4 turn out to have an association with the ADDEPEV2.

5.3 Explanatory Data Analysis (EDA)

Now, we perform **exploratory data analysis** about the relationships between some of the selected explanatory variables and ADDEPEV2 for each adult group. All related tables and plots for EDA procedures are presented in the Appendix A.

5.3.1 Young Adulthood (18 - 39)

First, 20% of the participants in young adulthood have been diagnosed with depressive disorders. According to Table A.1, the participants who have depressive disorders appear to have, on average, 14 *uncomfortable mental days* during the past 30 days, while those who do not have depressive disorders have nearly 10 uncomfortable mental days.

In terms of health status risks, the young adult participants diagnosed with depressive disorders appear to have a higher proportion of the following variables: *difficulties in making decision by themselves*, *doing errands alone*, and *arthritis*. Regarding demography, the young participants who have been diagnosed with depressive disorders are more likely to be either unemployed or a homemaker, and the higher proportion identify themselves as White. Lastly, the young adult participants who have depressive disorders are much more likely to smoke more than 100 cigarettes in their life and take HPV tests at least once.

5.3.2 Middle Adulthood (40 - 60)

According to the summary tables in the Appendix A, 19.9% of the middle-aged participants have been diagnosed with depressive disorders. Those who suffer from depressive disorders have an average of 14 days of *uncomfortable mental status*, while those who do not have an average of 10 days. However, both groups of the middle-age participants have a comparable number of the alcohol drinks in a month on average.

Also, the middle-aged participants with depressive disorders have higher proportion of having difficulties in making decisions by themselves. In terms of demography, these adults have a higher proportion of being unemployed, and they are less likely to be currently married. In addition, those who have depressive disorders have a higher proportion of having arthritis and receiving HIV tests at least once in their lifetime. However, the type of places for receiving a flu shot appears to have a weak relationship with depressive disorders for this middle-age group.

5.3.3 Old-age Adulthood (61 - 85)

15.2% of the old-aged participants answered that they have been diagnosed with depressive disorders. The participants with depressive disorders have a monthly average of 13 days with *uncomfortable mental status*, whereas those without feel an average of 11 days with uncomfortable mental status during the past 30 days.

Like the young and middle adult groups, the old-aged participants who suffer from depressive disorders are more likely to find difficulties in making decisions; suffer from pneumonia and arthritis; receive sigmoidoscopy or colonoscopy exams; and have less access to healthcare services due to high medical costs. For demographic attributes, the old-aged participants who have depressive disorders tend to be divorced and/or separated. Also, they are less likely to be veterans, compared to those who do not have depressive disorders.

Variable	Definition	Responses (Encoded)	Gini Importance
MENTHLTH	# of bad mental days during the past 30 days	Any positive integer (1 - 30)	0.611
DECIDE	Difficulty of making decision alone	Yes (1) / No (2)	0.177
DIFFALON	Difficulty of doing errands alone		0.083
HAVARTH3	Arthritis, lupus, or related illnesses		0.035
HPVTEST	HPV test records		0.024
SMOKE100	Smoke more than 100		0.015
EMPLOY1	Current employment	Employed (1) Unemployed (2) Student (5) Homemaker (6) Retired (7)	0.013
RACE	Race	White (1) Black (2) American Indian (3) Asian (4) Native Hawaiian (5) Multiracial (6) Hispanic (7) Others (8)	0.012
USENOW3	Current use of tobacco or snuff	Everyday (1) Somedays (2) Not at all (3)	0.0108
CHILDREN	# of children	Any positive integer	0.0107
WEIGHT2	Weight in pounds	Any positive integer	0.008

Table 5.2: Selected variables by the CART feature importance, with the definition, responses, and Gini importance value of each variable (young adult group)

Variable	Definition	Responses (Encoded)	Gini Importance
MENTHLTH	# of bad mental days during the past 30 days	Any positive integer (1 - 30)	0.700
DECIDE	Difficulty of making decision alone	Yes (1) / No (2)	0.181
HAVARTH3	Arthritis, lupus, or related illnesses		0.039
EMPLOY1	Current employment	Employed (1) Unemployed (2) Student (5) Homemaker (6) Retired (7)	0.029
AVEDRNK2	Average # of alcohol drinks per month	Any positive integer	0.015
IMFVPLAC	Type of place for flu or vaccine shots	Doctor/Hospital (1) Health department (2) Community health center (3) Non-medical spaces (4)	0.010
HIVTSTD6	HIV test records	Yes (1)/No (2)	0.009
MARITAL	Current marital status	Married (1) Divorced (2) Widowed (3) Separated (4) Never (5) Unmarried couple (6)	0.008

Table 5.3: Selected variables by the CART feature importance, with the definition, responses, and Gini importance value of each variable (middle-aged adult group)

Variable	Definition	Responses (Encoded)	Gini Importance
MENTHLTH	# of bad mental days during the past 30 days	Any positive integer (1 - 30)	0.762
DECIDE	Difficulty of making decision alone	Yes (1) / No (2)	0.141
HAVARTH3	Arthritis, lupus, or related illnesses		0.020
MARITAL	Current marital status	Married (1) Divorced (2) Widowed (3) Separated (4) Never (5) Unmarried couple (6)	0.020
HADSIGM3	Sigmoidoscopy and colonoscopy exam record (intestine exams)	Yes (1)/ No (2)	0.017
SLEPTIM1	Average sleeping hours per day	Any positive integer	0.010
CHCCOPD1	Pulmonary or lung-related disease	Yes (1)/No (2)	0.0094
DIABAGE2	Age when diabetes started	Any positive integer	0.0093
VETERAN3	Veteran status	Yes (1) / No (2)	0.006
MEDCOST	Healthcare inaccessibility due to high medical costs	Yes (1) / No (2)	0.004

Table 5.4: Selected variables by the CART feature importance, with the definition, responses, and Gini importance value of each variable (old-aged adult group)

CHAPTER 6

Results

In this chapter, we examine the results of what decision trees and logistic regression select as the determinants of depressive disorders for each adult group in the BRFSS sample. Also, we present the results of prediction metrics of decision trees, logistic regressions, and support vector machines for each group.

6.1 Decision Trees

Before constructing decision tree models, we first take 10% of the original data of each adult group as a random sample. Then, we set the maximum tree depth as 4, thereby avoiding overfitting.

In the results of the decision trees, the blue leaf nodes represent people not diagnosed with depressive disorders (No), while the brownish leaf nodes are people diagnosed with depressive disorders (Yes). Also, the saturation of colors indicates the extent of Gini impurities of the responses in each leaf node: the more saturated the color, the smaller Gini impurity the response is. We notice that these inequalities '≤ 1.5' on certain nodes indicates Yes, and '≥ 1.5' means No, since all categorical variables have values of either 1 (Yes) and 2 (No).

6.1.1 Young adults (18 - 39)

Figure 6.1 displays the decision tree of whether a young adult in the sample has been ever diagnosed with depressive disorders, based on the features selected by the tree algorithm.

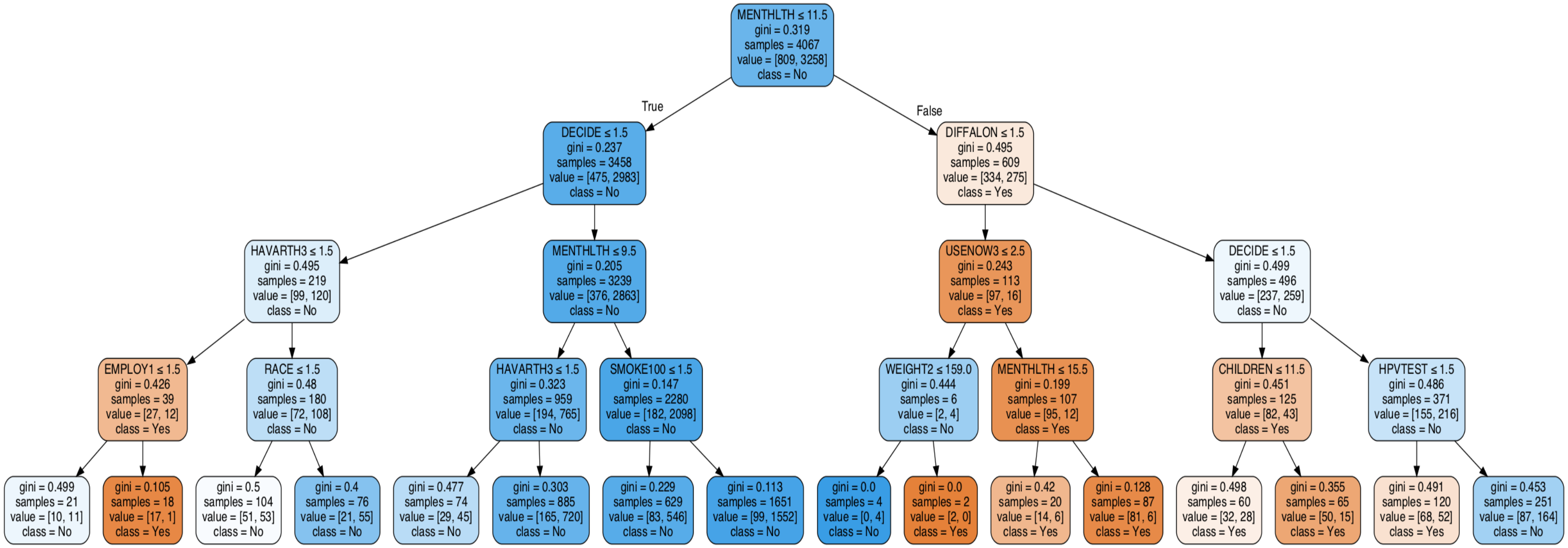

Figure 6.1: The decision tree for the young adult group

Table 6.1 presents all 16 decision rules of the decision tree, as shown in Figure 6.1. The rule numbers of important decision rules are colored in red, as their Gini impurity scores are comparatively low. Now, we examine the important rules for the young adult participants who are diagnosed with depressive disorders in the sample.

According to rule 6, among the young participants who have 11 or fewer bad mental days, feel difficulty in making decision, and suffer from arthritis or related illnesses, those who are not employed (i.e., students, homemaker, retired, unemployed) are more likely to be diagnosed with depressive disorders.

Also, we obtain an insight from rule 12 that for those who have more than 11 bad mental days, decision-making difficulty, and errands-doing difficulty, those who have more than 11 children in their household are more likely to be diagnosed with depressive disorders. In addition, as rules 15 and 16 show, depressive disorders can more likely to occur among those who have 12 or more bad mental days, never snuffed, and have difficulty doing errands.

Therefore, we can conclude that the number of bad mental days, decision-making difficulty, arthritis or related illnesses, employment status, tobacco and snuff records, errands-doing difficulty, and the number of children in the household are most important factors of depressive disorders among the young participants of the sample.

Rule #	Bad mental days	Decision difficulty	Arthritis (muscle)	Employment	Race	Weight (lbs)	Smoke >100 cigarettes?	Snuff	Errands difficulty	HPV test	Children	Diagnosis	# Yes	# No
1	< 10	No	No									No	165	720
2			Yes									No	29	45
3	10 - 11	No					Yes					No	83	546
4							No					No	99	1552
5	≤ 11	Yes	Yes	Employed								No	10	11
6				Non-employed								Yes	17	1
7			No		White							No	51	53
8					Non-white							No	21	55
9	> 11	Yes				≤ 159		Every Some	Yes			No	0	4
10						> 159						Yes	2	0
11									No		≤ 11	Yes	32	28
12											> 11	Yes	50	15
13		No							No	Yes		Yes	68	52
14										No		No	87	164
15	12 – 15							Never	Yes			Yes	14	6
16	> 15							Never	Yes			Yes	81	6

Table 6.1: The decision table for the young adult group (The most important rules are colored in red.)

6.1.2 Middle-aged adults (40 - 60)

Figure 6.2 displays the decision tree of whether an individual in the middle-age adult group has been ever diagnosed with depressive disorders. Table 6.2 presents some selected decision rules of the decision tree, as displayed in Figure 6.2. Also, the rule numbers of important rules have been colored in red.

As rule 8 illustrates, the participants in the middle-age group are more likely to be diagnosed with depressive disorders if they have the following characteristics: have fewer than 11 bad mental days; suffer from decision-making difficulty; currently retired or never employed; and have received flu shot at non-doctor spaces (i.e., health departments, community health centers, schools, and workplaces).

Furthermore, within the group of middle-age participants who have more than 11 bad mental days in the past month, we find two meaningful results regarding depressive disorders diagnosis. First, we focus on rule 9, in which the corresponding participants have difficulties making decisions alone and are students, retired, or unable to work. If the respondents in this group have, on average, 9 or fewer alcohol drinks per month, then they are more likely to be diagnosed with depressive disorders. Second, as rule 13 shows, we look at the group that does not have decision-making difficulty. If the respondents in this group are unable to work and have 3.5 or fewer alcohol drinks, then they are more likely to have depressive disorders.

Therefore, we can conclude that the number of bad mental days per month, decision-making difficulty, current employment status, the number of alcohol drinks per month, and the places of receiving flu shots are the most relevant factors of depressive disorders in most middle-age participants in the sample.

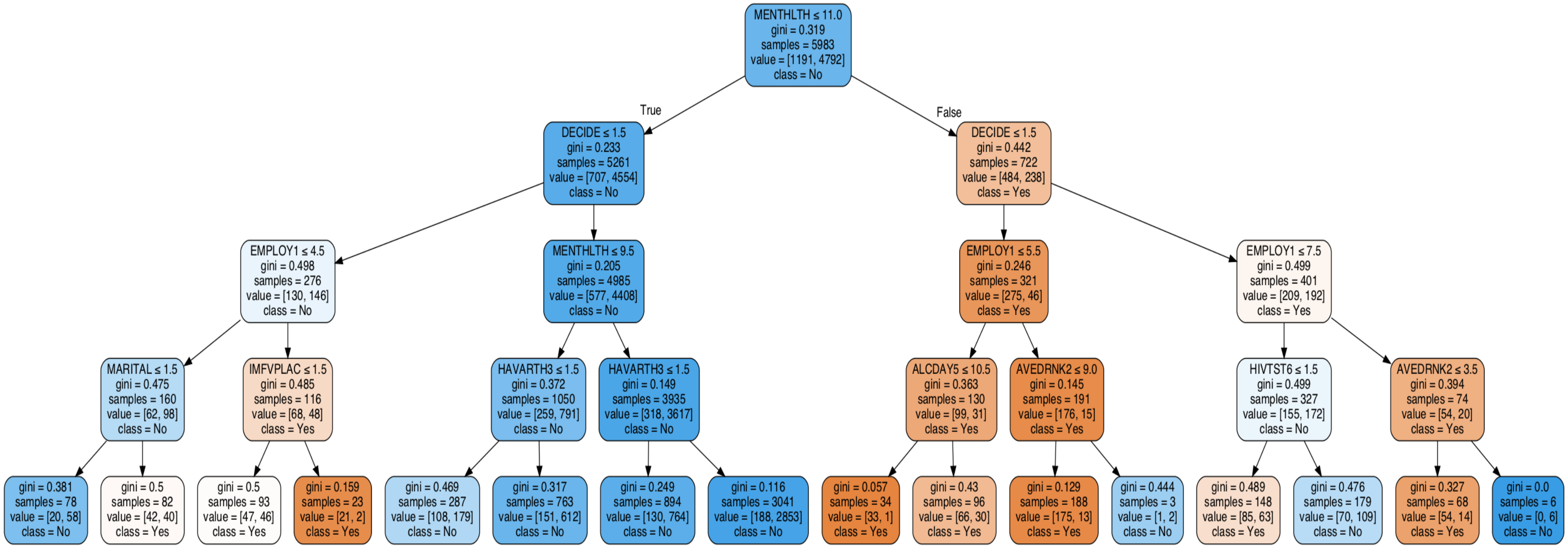

Figure 6.2: A decision tree for the middle-age adult group

Rule #	Bad mental days	Decision difficulty	Arthritis (muscle)	Employment	Marital status	Flu shot place	# of alcohol drinks in a month	HIV test	Diagnosis	# Yes	# No
1	≤ 9	No	No						No	151	612
2			Yes						No	108	179
3	10 - 11	No	No						No	188	2853
4			Yes						No	130	764
5	< 11	Yes		Once/currently work	Married				No	20	58
6					Not-married				Yes	42	40
7				Never, Retired		Medical space			Yes	47	46
8						Health departments Community health center Non-medical spaces			Yes	21	2
9	> 11	Yes		Students, retired, unable			≤ 9		Yes	175	13
10							> 9		No	1	2
11		No		Able to work				Yes	Yes	85	63
12								No	No	70	109
13				Unable to work			≤ 3.5		Yes	54	14
14							> 3.5		No	0	6

Table 6.2: The decision table for the middle-aged adult group (The most important rules are colored in red.)

6.1.3 Older adults (61-85)

Figure 6.3 displays the decision tree for the old-aged adult group of the sample. Also, Table 6.3 presents all decision rules shown in Figure 6.3.

Within the group of old-aged adults in the sample who have more than 11 bad mental days, we can obtain two important results that indicate determinants of depressive disorders among the group. First, rule 9 focuses on the respondents who have difficulties in making decisions alone and suffer from diabetes before 72 years old. If the respondents in this group sleep on average 17 or fewer hours per day, then they are more likely to be diagnosed with depressive disorders. In addition, as rule 15 shows, we look at the group of old-aged participants who do not have decision-making issues and who are separated, never married, or in an unmarried couple. If the respondents in this group receive insufficient healthcare services due to high medical costs at least once in their lifetime, then they are more likely to be diagnosed with depressive disorders. **Therefore, we can conclude that the number of bad mental days per month, decision-making issues, sleeping hours, the age when diabetes has started, marital status, and healthcare inaccessibility due to high costs are the most relevant factors of depressive disorders in most old-aged participants of the sample data of the U.S. residents in 2018.**

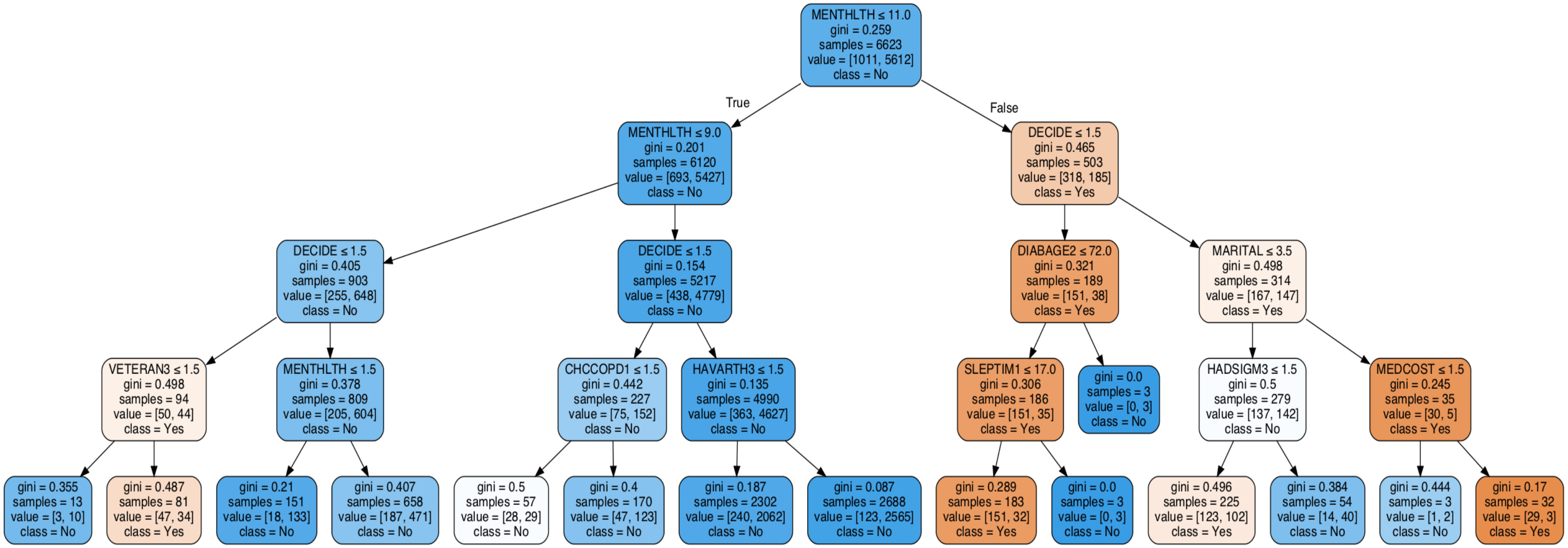

Figure 6.3: A decision tree for the old-aged adult group

Rule #	Bad mental days	Decision difficulty	Arthritis (muscle)	Marital status	Sleeping hours	Veteran status	Lung illnesses	Diabetes age	Intestine exams	Healthcare inaccess (high medical costs)	Diagnosis	# Yes	# No
1	≤ 1	No									No	18	133
2	2 – 9	No									No	187	471
3	≤ 9	Yes				Yes					No	3	10
4						No					Yes	47	34
5	10 - 11	Yes					Yes				No	28	29
6							No				No	47	123
7		No	Yes								No	240	2062
8			No								No	123	2565
9	> 11	Yes			≤ 17			≤ 72			Yes	151	32
10					> 17						No	0	3
11								> 72			No	0	3
12		No		Married, Divorced, Widowed					Yes		Yes	123	102
13									No		No	14	40
14				Separated, Never, Unmarried couple						Yes	No	1	2
15										No	Yes	29	3

Table 6.3: The decision table for the old-aged adult group (The most important rules are colored in red.)

6.2 Logistic Regression

6.2.1 Young adults (18 - 39)

Table 6.4 presents the summary of a logistic model for the young adults. This model was chosen through removing variables with the largest p-value and comparing the one with a smaller number of variables to the model with all variables from Table 5.2. Then, considering simplicity, misclassification rate, and linearity, we decided that the logistic model in Table 6.4 is considered to be the best model that predicts the depressive disorders among the young adults. This model includes all variables from Table 5.2 except WEIGHT2 and SMOKE100. We will call this selected model as M_{young}.

From the Wald tests for M_{young}, the following variables are most statistically significant predictors for depressive disorders among the young adults: the number of bad mental days, difficulty of making decisions, employment status, race, arthritis, difficulty of doing errands alone, the number of children, and HPV test records. All those variables are explained in Table 5.2.

In addition, from Table 6.4, we can interpret the results of some statistically significant variables in M_{young}. First, assuming that all other variables are constant for the young-aged group, we interpret the results of the odds ratio for each significant predictor, as listed below. **According to this interpretation, young-age adult residents of the United States are most likely to have been diagnosed with depressive disorders when they are a white homemaker; have arthritis; have difficulty doing errands and making decisions alone; and have HPV test records.** More children and more bad mental health days slightly increase the likelihood of depressive disorders diagnosis.

- One day increase in the number of bad mental days during the past 30 days is associated with 5% higher odds of depressive disorders diagnosis.
- The odds of the onset of depressive disorders for people who have difficulty in making decisions are 402% higher than the odds of the disorders for people who do not have.
- The odds of depressive disorders diagnosis for people who are homemaker and students are 57% and 24% higher than the odds of the disorders for those who are currently employed.
- Whites have the highest odds of depressive disorders. Specifically, Blacks, American Indians, Asians, Native Hawaiians, and other racial groups have 46%, 70%, 41%, 48%, and 55% times the odds of depressive disorders for Whites, respectively.
- The odds of the depressive disorders for people who have arthritis, lupus, or other related illnesses are 153% higher than the odds of the disorders for those who do not.

Variables	Levels	Coef. Estimate	SE	OR	95% CI for OR	z-test	p-value
Bad mental days	.	0.054	0.002	1.05	(1.052, 1.059)	32.1	**2e-16**
Decision Difficulty (No)	Yes	1.61	0.034	5.02	(4.7, 5.3)	47.1	**2e-16**
Employment Status (Employed)	Homemaker	0.45	0.04	1.57	(1.45, 1.70)	11.2	**2e-16**
	Retired	0.16	0.05	1.17	(1.06, 1.29)	3.1	0.0019
	Student	0.22	0.04	1.24	(1.15, 1.34)	5.4	**7.02e-8**
	Unemployed	0.19	0.25	1.21	(0.73,0.95)	0.75	0.45
Race (White)	Black	-0.76	0.045	0.46	(0.43, 0.51)	-17.1	**2e-16**
	American Indian Alaska Native	-0.36	0.081	0.7	(0.59, 0.82)	-4.4	**9.76e-6**
	Asian	-0.9	0.077	0.41	(0.35, 0.47)	-11.7	**2e-16**
	Native Hawaiian Pacific Islanders	-0.74	0.14	0.48	(0.36, 0.62)	-5.3	**1.15e-7**
	Multiracial	-0.51	0.15	0.60	(0.44, 0.81)	-3.3	0.0009
	Hispanic	-0.075	0.065	0.93	(0.81, 1.05)	-1.1	0.25
	Other	-0.60	0.03	0.55	(0.51, 0.59)	-17.1	**2e-16**
Arthritis (No)	Yes	0.93	0.04	2.53	(2.35, 2.72)	25.3	**2e-16**
Doing errands alone (No)	Yes	1.15	0.06	3.15	(2.82, 3.52)	20.3	**2e-16**
Snuff Use (Everyday)	Some days	0.25	0.10	1.29	(1.05, 1.58)	2.4	0.015
	Not at all	0.3	0.078	1.35	(1.16, 1.57)	3.85	0.0001
Children	.	0.015	0.002	1.01	(1.01, 1.02)	9.1	**2e-16**
HPV Test (No)	Yes	0.78	0.025	2.18	(2.08, 2.29)	30.9	**2e-16**

Table 6.4: Coefficient estimate, standard error (SE), the odds ratio and its 95% CI, and results from Wald-Test statistics in M_{young} (Most statistically significant variables are marked as bold, as presented in Table 5.2)

- The odds of the depressive disorders for people who have difficulties in doing errands alone are 215% higher than the odds of the disorders for those who do not.
- One increase in the number of children is associated with 1% higher odds of the depressive disorders.
- The odds of the depressive disorders for people who have HPV test records are 118% higher than the odds for those who do not.

For simplicity, we interpret 95% confidence intervals for odds ratios of only two statistically significant predictors. Holding all other variables constant, we are 95% confident that young adult participants with decision-making difficulties will have between 370% and 430% higher odds of the depressive disorders diagnosis than those without. Also, we are 95% confident that the young participants who have HPV test records will have between 108% and 129% higher odds of the depressive disorder than those who do not.

We now assess M_{young}. First, the linearity condition is satisfied because the empirical odds ratio of both the number of bad mental days and children does not vary a lot, and all other predictors are categorical. According to the data collection process in Chapter 5, randomness and representativeness are also automatically met, as well as independence, since one's depressive disorders are not generally affected by the others. From the drop-in-deviance test, the corresponding G-statistic is 10,318 and its associated p-value is 0, which means that M_{young} predicts the depressive disorders diagnosis of the young adults in the sample well. Finally, the misclassification rate of M_{young} is 16.45%. To be specific, the percentage of young adults who are predicted to have, but do not have, depressive disorders is 2.4%, and the percentage of young adults who are predicted to not have, but actually have, depressive disorders is 14.1%.

6.2.2 Middle-aged adults (40 - 60)

Table 6.5 presents the summary of the selected logistic model for the middle-aged adult participants. This model was also chosen through repetitive procedures of removing variables with the largest p-value and comparing the one with a smaller number of variables to the model with all variables from Table 5.3. Then, considering simplicity, misclassification rate, and linearity, we decided that the logistic model in Table 6.5 is the best that predicts the depressive disorders among the middle-aged participants. This model includes all variables from Table 5.3 except AVEDRNK2. We now call the model M_{middle}.

From the Wald tests for M_{middle}, the following variables are most statistically significant predictors of depressive disorders of the middle-age group: the number of bad mental days, difficulty of making decisions, employment status, arthritis, HIV test records, and marital status. All those variables are explained in Table 5.3 . In addition, from Table 6.5 we can interpret the results of the most statistically significant variables in M_{middle}. First, assuming that all other variables are constant for the middle-age group, we interpret the results of the odds ratio for each significant predictor as follows. **According to the results, the middle-aged participants are most likely to have been diagnosed with depressive disorders when they are unemployed, divorced individuals who have arthritis, HIV test records, and difficulty making decisions alone.** More days of bad mental health per month slightly increase the likelihood of depressive disorders for the middle-aged adults.

Variables	Levels	Coef. Efficient	SE	OR	95% CI for OR	z-test	p-value
Bad mental days	.	0.048	0.001	1.05	(1.046, 1.052)	32.88	**2e-16**
Decision Difficulty (No)	Yes	1.71	0.028	5.54	(5,25 5.86)	60.33	**2e-16**
Employment Status (Employed)	Homemaker	0.525	0.046	1.69	(1.54, 1.85)	11.49	**2e-16**
	Retired	0.200	0.033	1.22	(1.12, 1.33)	4.49	**6.95e-6**
	Student	0.512	0.13	1.67	(1.29, 2.14)	3.96	**7.48e-5**
	Unemployed	0.762	0.026	2.14	(2.04,2.25)	29.53	**2e-16**
Marital Status (Divorced)	Married	-0.411	0.026	0.66	(0.63, 0.70)	-16.03	**2e-16**
	Never	-0.242	0.034	0.78	(0.73, 0.84)	-7.13	**1e-12**
	Separated	0.077	0.053	1.08	(0.97, 1.20)	1.45	0.15
	Unmarried couple	-0.296	0.058	0.74	(0.66, 0.83)	-4.09	**3.48e-7**
	Widowed	0.030	0.052	1.03	(0.93, 1.14)	0.575	0.56
Arthritis (No)	Yes	0.825	0.020	2.28	(2.19, 2.38)	40.41	**2e-16**
HIV test (No)	Yes	0.437	0.020	1.55	(1.49, 1.61)	22.11	**2e-16**
Flu shot place (Community center)	Doctor /Hospital	-0.107	0.057	0.90	(0.80, 1.00)	-1.87	0.06
	Health department	-0.176	0.14	0.84	(0.63, 1.10)	-1.25	0.21
	Non-medical	0.144	0.06	1.15	(1.02, 1.30)	2.34	0.02

Table 6.5: Coefficient estimate, standard error (SE), the odds ratio and its 95% CI, and results from Wald-Test statistics in M_{middle} (Most statistically significant variables are marked as bold, as explained in Table 5.3)

- One day increase in the number of bad mental days during the past 30 days is associated with 5% higher odds of the depressive disorder diagnosis.

- The odds of depressive disorder diagnosis for the middle-aged participants who have difficulty in making decisions are 454% higher than the odds of the disorders for people who do not have.

- The unemployed people in the middle-age group have the highest odds of depressive disorders, who are 114% higher than those who are employed. Specifically, homemaker, students, and retired people have 69%, 67%, and 22% times higher odds of the depressive disorders than the employed people, respectively.

- The divorced participants who are in their middle adulthood have the highest odds of depressive disorder diagnosis. In particular, those who are never married, widowed, and married have 78%, 74%, and 66%, respectively, the odds of depressive disorders of the divorced.

- The odds of the depressive disorders for the middle-aged participants who have arthritis, lupus, or other related illnesses are 128% higher odds of the disorders than those who do not.
- The odds of the depressive disorders for the middle-aged participants who have HIV test records are 55% higher odds than those who do not.

Likewise, we interpret 95% confidence intervals for the odds ratios of only two significant predictors. Holding all other variables constant for the middle-aged participants, we are 95% confident that one day increase in the number of bad mental days is associated with an increase in the odds of a depressive disorder diagnosis between 4.6% and 5.2% higher. Also, we are 95% confident that the middle-aged participants who are unemployed will have between 104% and 125% higher odds of the depressive disorder diagnosis than those who are employed.

We have identified that linearity condition is satisfied because the empirical odds ratio of the number of bad mental days is fairly constant, and all other predictors are categorical. From the drop-in-deviance test, the corresponding G-statistic is nearly 16,984 and its associated p-value is 0, which means that M_{middle} predicts depressive disorder diagnosis of the middle-age participants in the sample data well. Finally, the misclassification rate of M_{middle} is 15.5%. To be specific, the percentage of middle-aged adults from the sample who are predicted to have, but actually do not have, depressive disorders is 2.3%, and the percentage of middle-aged adults who are predicted to not have, but actually have, depressive disorders is 13.2%.

6.2.3 Older adults (61-85)

Table 6.6 presents the summary of a logistic model for the old-aged participants. This model was chosen with the same procedures as M_{young} and M_{middle}. Considering simplicity and linearity, we decided that the model that includes all from Table 5.4 except SLEPTIM1 is the best model that predicts depressive disorders diagnosis for the old-age group. We will call this model M_{old}.

From the Wald tests for M_{old}, the following variables are most statistically significant predictors for the depressive disorder diagnosis among the older adults: the number of bad mental days, difficulty of making decisions, the age when diabetes started, arthritis, marital status, intestine exams experiences for colorectal cancer, healthcare inaccessibility due to high medical costs, lung illnesses, and veteran status. All those variables are explained in Table 5.4. In addition, from Table 6.6, we can interpret the results of statistically significant variables in M_{old}. Assuming that all other variables are constant, we interpret the results of the odds ratio for some meaningful predictors as follows. **According to the interpretations, the old-aged participants are most likely to have been diagnosed**

with depressive disorders if they are married non-veterans who have decision-making difficulties, suffer from arthritis and lung illnesses, and who have intestine-related exams but lack healthcare accessibility due to medical costs. More bad mental days and lower ages at which diabetes started slightly increase the likelihood of the depressive disorders diagnosis for this old-age group.

Variables	Levels	Coef. Estimate	SE	OR	95% CI for OR	z-test	p-value
Bad mental days	.	0.054	0.002	1.055	(1.051, 1.058)	32.1	**2e-16**
Decision Difficulty (No)	Yes	1.59	0.028	4.93	(4.66, 5.21)	56.99	**2e-16**
Diabetes	.	-0.0075	0.001	0.99	(0.989, 0.995)	-4.88	**1.1e-6**
Marital Status (Divorced)	Married	0.532	0.026	1.70	(1.62, 1.79)	20.49	**2e-16**
	Never	0.057	0.026	1.06	(1.01, 1.11)	2.23	0.025
	Separated	0.485	0.0748	1.62	(1.40, 1.88)	6.48	**9.2e-11**
	Unmarried couple	0.371	0.039	1.45	(1.34, 1.56)	9.35	**2e-16**
	Widowed	0.500	0.079	1.65	(1.41, 1.92)	6.29	**3.3e-10**
Arthritis (No)	Yes	0.660	0.020	1.93	(1.86, 2.01)	32.57	**2e-16**
Intestine Exams (No)	Yes	0.413	0.027	1.61	(1.43, 1.59)	15.23	**2e-16**
High Medical Cost Difficulty (No)	Yes	0.492	0.037	1.63	(1.52, 1.76)	13.26	**2e-16**
Lung Illnesses (No)	Yes	0.535	0.026	1.71	(1.62, 1.80)	20.33	**2e-16**
Veteran Status (No)	Yes	-0.293	0.025	0.75	(0.71, 0.78)	-11.63	**2e-16**

Table 6.6: Coefficient estimate, standard error (SE), the odds ratio and its 95% CI, and results from Wald-Test statistics in M_{old} (Most statistically significant variables are marked as bold, as explained in Table 5.4.)

- One day increase in the number of bad mental days during the past 30 days is associated with 5.5% higher odds of the depressive disorder diagnosis for the old-aged participants.
- The odds of the depressive disorders diagnosis for the old-aged participants who have difficulty in making decisions are 393% higher odds of the disorders than people who do not have.
- One year increase in the age that the diabetes started is associated with 1% lower odds of the depressive disorder diagnosis for the old-aged participants.
- The married people in the old-age group have the highest odds of depressive disorders. Specifically, the widowed, separated, and unmarried couple have 65%, 62%, and 45% higher odds of depressive disorders than the divorced, respectively.

- The odds of the depressive disorders for the old-aged participants who have arthritis, lupus, or other related illnesses are 93% higher odds of the disorders than those who do not.
- The odds of the depressive disorders for the old-aged participants who have experienced intestine exams for colorectal cancer are 61% higher odds than those who do not.
- The odds of the depressive disorders for the old-aged participants who have healthcare inaccessibility due to high medical costs are 63% higher odds than those who do not.
- The odds of the depressive disorders for the old-aged participants who have lung-related illnesses are 71% higher odds than those who do not.
- The odds of the depressive disorders for the old-aged veterans are 75% the odds for non-veterans.

Likewise, we interpret 95% confidence intervals for odds ratios of two selected predictors. Holding all other variables constant for the old-aged participants, we are 95% confident that people who have healthcare inaccessibility due to economic hardships will have between 52% and 76% higher odds of a depressive disorder diagnosis than those who do not have. Also, we are 95% confident that one day increase in the number of bad mental days is associated with an increase in the odds of the depressive disorder diagnosis between 5.1% and 5.8% higher.

To assess M_{old}, we have identified that linearity condition is satisfied because the empirical odds ratio of the following two numeric variables - the number of bad mental days and the age of diabetes - turns out to be fairly constant, and all other predictors are categorical. The G-statistic is nearly 9,450 and its associated p-value is 0, so M_{old} predicts the depressive disorders diagnosis of the old-aged participants in a good manner. Finally, the misclassification of M_{old} is 13.92%. To be specific, the percentage of old-aged participants who are predicted to have, but do not have, depressive disorders is 1.2%, and the percentage of old-aged participants who are predicted to not have, but actually have, depressive disorders is 12.7%.

6.3 Prediction Metrics

Before making and evaluating the prediction tasks that the three methods perform for each adult group, we randomly split the data of each group into 70% of the training set and the remaining 30% into the testing set.

6.3.1 Decision Trees

We present the prediction metrics of decision tree models for each of the three adult groups, by using the following six metrics: confusion matrix, accuracy, precision, recall, F1 scores, and ROC curve.

6.3.1.1 Young adults

	Predicted No	Predicted Yes
Actual No	1354	58
Actual Yes	228	104

Table 6.7: The confusion matrix of the decision tree model (young adult group)

Table 6.7 is the confusion matrix of the decision tree that predicts the young adult group of the BRFSS data sample. According to the table, among the young adult participants in the sample, 77.6% of people who have not been diagnosed with depressive disorders are correctly predicted (true negatives), while the remaining 3.3% are wrongly predicted as those who have depressive disorders (false positives). In contrary, 13.1% of people who have been suffering from depressive disorders are predicted as those who have not (false negatives), whereas the other 5.9% are correctly predicted (true positives). Considering the medical setting where false diagnosis of actual patients would result in serious issues, we focus on the number of false negatives in Table 6.7 that would affect the value of recall. Furthermore, the accuracy of this decision tree model is then $\left(\frac{1354 + 104}{1354 + 58 + 228 + 104}\right) = 0.8355$, which indicates good performance of the tree model, given that the tree predicts the depressive disorders among the young adult group of the U.S. adult respondents.

	Precision	Recall	F1 score
"No"	0.86	0.96	0.90
"Yes"	0.64	0.31	0.42
macro average accuracy	0.75	0.64	0.66
weighted average accuracy	0.82	0.84	0.81

Table 6.8: Classification report of the decision tree model (young adult group)

Second, we examine the classification report as presented in Table 6.8, which shows the three metrics for whether a young adult has been diagnosed with depressive disorders. Notice that the proportion of people who are diagnosed with depressive disorders in this sample is lower than those who are not. Hence, we look at the weighted average accuracy of each metrics score at any situation that encompasses skewness of one category in a response variable. Then, we confirm that weighted precision, recall, and F1 scores are all above 0.80, which indicates the acceptable quality of prediction that the tree model made for the disorders among the U.S. young adults.

Lastly, as presented in Figure 6.4, the ROC curve has its AUC score of 0.78, which maintains our confirmation that the decision tree model, which is pre-pruned by limiting the number of tree depth to avoid overfitting issues, predicts depressive disorder diagnosis among the young age group of the U.S. adult residents in 2018 in an effective manner.

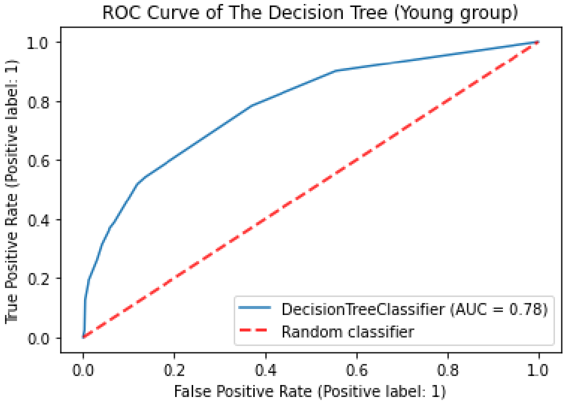

Figure 6.4: ROC curve of the decision tree model (young adult group)

6.3.1.2 Middle-aged adults

Table 6.9 is the confusion matrix of the decision tree for the middle-age group of the BRFSS data sample. Likewise, we interpret the table as follows: among the middle-aged adults in the sample, 75.9% of people who have not been diagnosed with depressive disorders are correctly predicted (true negatives), while the remaining 4.2% are wrongly predicted as those who have been diagnosed with depressive disorders (false positives). Also, 11.2% of people who have been suffering from depressive disorders are predicted as those who have not (false negatives), whereas the other 8.7% are correctly predicted (true positives). Furthermore, the accuracy of the decision tree model is then $\left(\frac{1947+222}{1947+108+288+222}\right) = 0.8456$, which indicates good performance of the tree model in terms of accuracy.

	Predicted No	Predicted Yes
Actual No	1947	108
Actual Yes	288	222

Table 6.9: The confusion matrix of the decision tree model (middle-aged adult group)

Second, Table 6.10 shows the three prediction metrics within the middle-age group. As the young adults shows, the proportion of middle-aged adults who are diagnosed with depressive disorders is much lower than those who are not. Thus, we focus on the weighted average accuracy of each metrics score. Then, we confirm that the weighted precision, recall, and F1 scores are all above 0.83, which indicates the good quality of predictions that the tree model made for the disorders among the U.S. middle-aged adults.

	Precision	Recall	F1 score
"No"	0.87	0.95	0.91
"Yes"	0.67	0.44	0.53
macro average accuracy	0.77	0.69	0.72
weighted average accuracy	0.83	0.85	0.83

Table 6.10: Classification report of the decision tree model (middle-aged adult group)

Lastly, Figure 6.5 shows that the ROC curve has its AUC score of 0.80, which confirms that our selected, pruned decision tree model predicts depressive disorder diagnosis among the middle-age group of the U.S. adults in an acceptable manner.

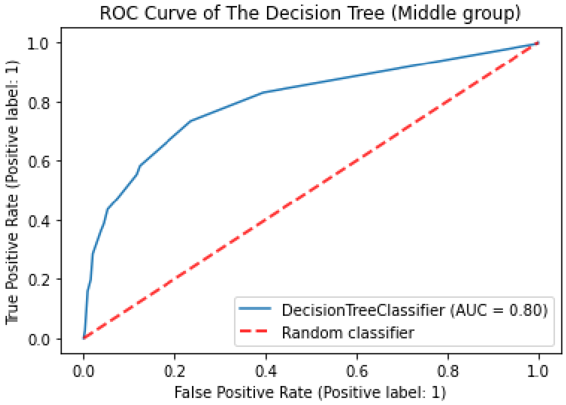

Figure 6.5: ROC curve of the decision tree model (middle-aged adult group)

6.3.1.3 Older adults

Finally, Table 6.11 is the confusion matrix of the decision tree for the older adults. We observe that 82.5% of people who have not been diagnosed with depressive disorders are correctly predicted (true negatives), while the remaining 2.9% are wrongly predicted as those who have been diagnosed with depressive disorders (false positives). Also, 10.1% of people who have been suffering from depressive disorders are predicted as those who have not (false negatives), whereas the other 4.4%

are correctly classified and predicted (true positives). Furthermore, the accuracy of the decision tree model is then $\left(\dfrac{2343 + 126}{2343 + 82 + 288 + 126}\right) = 0.8697$, which indicates good prediction of the tree model for depressive disorder diagnosis among the old-age group.

	Predicted No	Predicted Yes
Actual No	2343	82
Actual Yes	288	126

Table 6.11: The confusion matrix of the decision tree model (old-aged adult group)

	Precision	Recall	F1 score
"No"	0.89	0.97	0.93
"Yes"	0.61	0.30	0.41
macro average accuracy	0.75	0.64	0.67
weighted average accuracy	0.85	0.87	0.85

Table 6.12: Classification report of the decision tree model (old-aged adult group)

Furthermore, Table 6.12 shows the three prediction metrics for the tree of the old adults. Again, the proportion of the group of old-aged people who are diagnosed with depressive disorders is lower than those who are not. Therefore, we focus on the weighted average accuracy of each metrics score. Then, we confirm that the weighted precision, recall, and F1 scores are all above or equal to 0.85, which indicates good prediction of the tree model regarding the disorder diagnosis among the U.S. older adults.

Lastly, Figure 6.6 shows that the ROC curve has its AUC score of 0.77, which proves that our chosen decision tree model with a limitation on the number of tree depth, predicts depressive disorders diagnosis of the U.S. older adults in an acceptable manner.

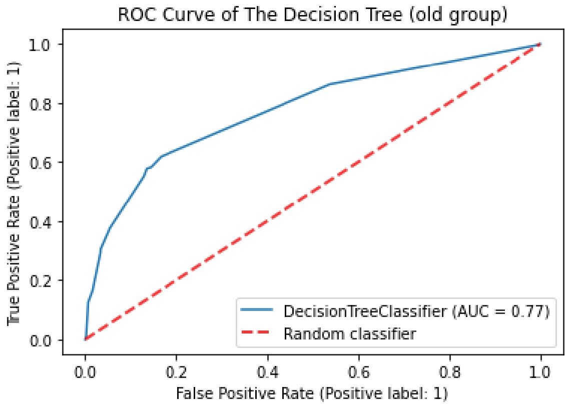

Figure 6.6: The ROC curve of the decision tree model (old-aged adult group)

6.3.2 Logistic Regression

We present prediction metrics of three logistic regression models for the three adult groups. In order to do so, we first take 10% of the entire data for each adult group, thus creating three random samples. Then, we randomly split each of the three samples into 70% of the training set and the remaining 30% into the testing set. Notice that the results from misclassification rates for each group that have been presented in the former section do not match the results from the confusion matrices in this section. We use the testing set of the sample for computing the confusion matrix in this section, while the misclassification rates for each adult group in the former section were calculated by using the entire data set for each group.

6.3.2.1 Young adults

	Predicted No	Predicted Yes
Actual No	1370	42
Actual Yes	235	97

Table 6.13: The confusion matrix of the logistic model M_{young} (young adult group)

Table 6.13 is the confusion matrix of the logistic model M_{young} that predicts the depressive disorders diagnosis among the young adult group sample. We see that 78.5% of people who have not been diagnosed with depressive disorders are correctly predicted (true negatives), while the remaining 2.4% are wrongly predicted as those who have depressive disorders (false positives). In contrary, 13.5% of people who have been suffering from depressive disorders are predicted as those who have not (false negatives), whereas the other 5.6% are correctly predicted (true positives). Furthermore, the accuracy of this logistic model M_{young} is then $\left(\dfrac{1370+97}{1370+42+235+97}\right) = 0.8412$, which indicates good prediction, given that M_{young} predicts depressive disorder diagnosis among the young group of the U.S. adult respondents.

	Precision	Recall	F1 score
"No"	0.85	0.97	0.91
"Yes"	0.70	0.29	0.41
macro average accuracy	0.78	0.63	0.66
weighted average accuracy	0.82	0.84	0.81

Table 6.14: Classification report of the logistic model M_{young} (young adult group)

Second, we examine the classification report scores as presented in Table 6.14. Notice that the proportion of young adult participants who are diagnosed with depressive disorders in the sample is lower than those who are not. Hence, we look at the weighted average accuracy of each metrics score, for our data sample is imbalanced with one category in the response variable. Considering the medical setting where false diagnosis of actual patients could result in serious issues, we similarly focus on the number of false negatives in Table 6.13 that would effect in the value of recall. Then, we confirm that weighted precision, recall, and F1 scores are all above 0.80, which indicate the good quality of prediction that the M_{young} performed.

As presented in Figure 6.7, the ROC curve has its AUC score of 0.74, which maintains our confirmation that M_{young} predicts depressive disorder diagnosis among the young adult group of the U.S. residents from the 2018 BRFSS sample reasonably well.

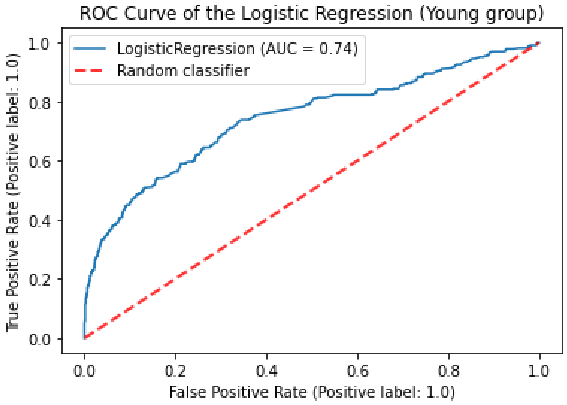

Figure 6.7: ROC curve of the logistic model (young adult group)

6.3.2.2 Middle-aged adults

	Predicted No	Predicted Yes
Actual No	1995	60
Actual Yes	335	175

Table 6.15: The confusion matrix of the logistic model M_{middle} (middle-aged adult group)

Table 6.15 is the confusion matrix of the logistic model M_{middle} that predicts depressive disorders among the middle-aged adults in the sample. In this setting, 77.8% of people who have not been diagnosed with depressive disorders are correctly predicted (true negatives), while the remaining 2.3% are wrongly predicted as those who have depressive disorders (false positives). In contrary, 13.1% of people who have been suffering from depressive disorders are predicted as those who have not (false negatives), whereas the other 6.8% are correctly predicted (true positives). Furthermore,

	Precision	Recall	F1 score
"No"	0.86	0.97	0.91
"Yes"	0.74	0.34	0.47
macro average accuracy	0.80	0.66	0.69
weighted average accuracy	0.83	0.85	0.82

Table 6.16: Classification report of the logistic model M_{middle} (middle-aged adult group)

the accuracy of this logistic model M_{middle} is then $\left(\dfrac{1995+175}{1995+60+335+175}\right) = 0.8460$, which indicates good prediction of the model.

Next, we look at the classification report of M_{middle}, as shown in Table 6.16. Likewise, we look at the weighted average accuracy of each metrics score, for our middle-age group data is also skewed to one category in the response variable. Then, since the weighted precision, recall, and F1 scores are around or above 0.82, M_{middle} shows good prediction for the disorders among the U.S. middle-aged adults in the sample. As presented in Figure 6.8, the ROC curve has its AUC score of 0.73, thus M_{middle} predicts depressive disorder diagnosis among the middle-age group of the U.S. adult residents from the 2018 BRFSS in a fair manner.

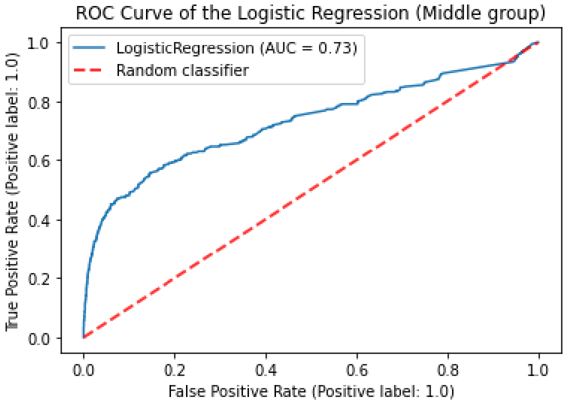

Figure 6.8: ROC curve of the logistic model M_{middle} (middle-aged adult group)

6.3.2.3 Older adults

	Predicted No	Predicted Yes
Actual No	2387	38
Actual Yes	324	90

Table 6.17: The confusion matrix of the logistic model M_{old} (old-aged adult group)

Table 6.17 is the confusion matrix of the logistic model M_{old} that predicts depressive disorders of the older adults in the sample. Interpreting the table, we claim that 84.1% of people who have not been diagnosed with depressive disorders are correctly predicted (true negatives), while the remaining 13.4% are wrongly predicted as those who have depressive disorders (false positives). In contrary, 11.4% of people who have been suffering from depressive disorders are predicted as those who have not (false negatives), whereas the other 3.2% are correctly predicted (true positives). Also, the accuracy of this logistic model M_{old} is then $\left(\frac{2387+90}{2387+38+324+90}\right) = 0.8725$, which indicates its good performance in predicting the depressive disorders among the older adults.

	Precision	Recall	F1 score
"No"	0.88	0.98	0.93
"Yes"	0.70	0.22	0.33
macro average accuracy	0.79	0.60	0.63
weighted average accuracy	0.85	0.87	0.84

Table 6.18: Classification report of the logistic model M_{old} (old-aged adult group)

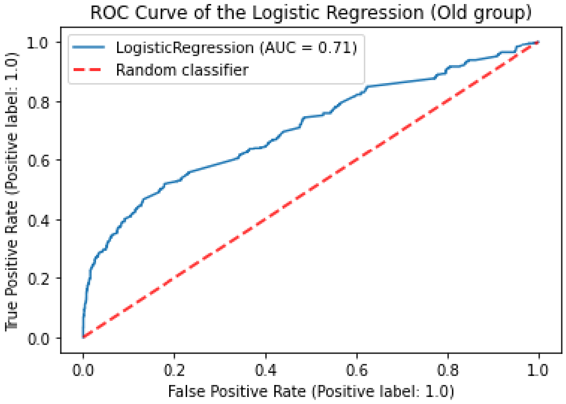

Figure 6.9: ROC curve of the logistic model M_{old} (old-aged adult group)

In Table 6.18, we look at the weighted average accuracy of each metrics score, for our old-age group data is also imbalanced. Since the weighted precision, recall, and F1 scores are around or above 0.84, M_{old} shows a good prediction for the disorders among the U.S. older adults in the sample. As presented in Figure 6.9, the ROC curve has its AUC score of 0.71, thus M_{old} predicts the depressive disorders among the old-age group of the U.S. adult residents from the 2018 BRFSS in an effective manner.

6.3.3 Support Vector Machine Classifiers (SVMC)

First, we take 10% of the entire data for each adult group to make three samples for the three adult groups and normalize the samples. Then, we use a grid-search algorithm to find the optimized parameters of the SVMCs that provide the best balanced accuracy score for each adult group.

6.3.3.1 Young adults

The SVMC that predicts the depressive disorders diagnosis among the U.S. young adults in the sample have the following optimized tuning parameters, as presented in Table 6.19. Hence, the RBF-kernel SVMC with the optimized tuning hyperparameters $C = 10$ and $\gamma = 0.01$ produces the best balanced accuracy scores for the SVMC.

C	γ	kernel function
10	0.01	Radial Basis Function (RBF)

Table 6.19: The optimized parameters of the SVMC for young adult group)

	Predicted No	Predicted Yes
Actual No	1323	89
Actual Yes	217	115

Table 6.20: The confusion matrix of the SVMC model (young adult group)

Table 6.20 is the confusion matrix of the optimized SVMC for the young adult sample. First, 75.9% of people who have not been diagnosed with depressive disorders are correctly predicted (true negatives), while the remaining 5.1% are wrongly predicted as those who have been diagnosed with depressive disorders (false positives). Also, 12.4% of people who have been suffering from depressive disorders are predicted as those who have not (false negatives), whereas the remaining 6.6% are correctly predicted (true positives). Then, the accuracy of prediction is $\left(\frac{1323 + 115}{1323 + 89 + 217 + 115}\right) =$ 0.8245.

	Precision	Recall	F1 score
"No"	0.86	0.94	0.90
"Yes"	0.56	0.35	0.43
macro average accuracy	0.71	0.64	0.66
weighted average accuracy	0.80	0.82	0.81

Table 6.21: Classification report of the optimized SVMC model (young adult group)

In addition, Table 6.21 provides the prediction metric scores of this optimized SVMC. The weighted average scores of all precision, recall, and f1 score are around or above 0.80, which shows a good performance of the SVMC in this prediction task. Lastly, Figure 6.10 shows that the ROC curve of this SVMC has its AUC score of 0.74, which confirms that our optimized SVMC predicts depressive disorders diagnosis among the U.S. young adults of the sample in an acceptable manner.

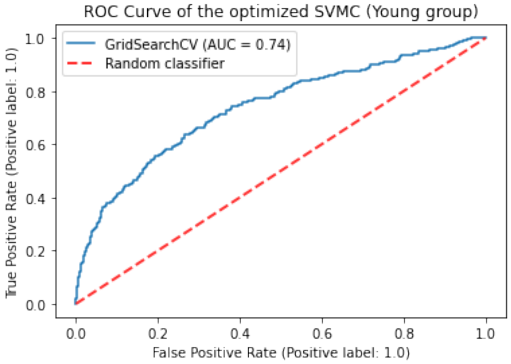

Figure 6.10: ROC curve of the optimized SVMC model (young adult group)

6.3.3.2 Middle-aged adults

The optimized parameters of the SVMC that predicts the depressive disorders among the U.S. middle-aged adults are presented in Table 6.22. The RBF-kernel function with the tuning hyperparameters $C = 10$ and $\gamma = 0.01$ produces the best balanced accuracy scores for the SVMC.

C	γ	kernel function
10	0.01	Radial Basis Function (RBF)

Table 6.22: The optimized parameters of the SVMC for middle-aged adult group)

	Predicted No	Predicted Yes
Actual No	1900	155
Actual Yes	288	222

Table 6.23: The confusion matrix of the optimized SVMC model (middle-aged adult group)

Table 6.23 is the confusion matrix of the optimized SVMC for the random sample of the middle-aged adults. First, 74.1% of people who have not been diagnosed with depressive disorders are correctly predicted (true negatives), while the remaining 6% are wrongly predicted as those who have been diagnosed with depressive disorders (false positives). Also, 11.2% of people who have been suffering from depressive disorders are predicted as those who have not (false negatives), whereas the remaining 8.7% are correctly predicted (true positives). The accuracy score of the prediction made by this SVMC is then $\left(\dfrac{1900 + 222}{1900 + 155 + 288 + 222}\right) = 0.8273$.

	Precision	Recall	F1 score
"No"	0.87	0.92	0.90
"Yes"	0.59	0.44	0.50
macro average accuracy	0.73	0.68	0.70
weighted average accuracy	0.81	0.83	0.82

Table 6.24: Classification report of the optimized SVMC model (middle-aged adult group)

Table 6.24 provides the prediction metrics of this optimized SVMC for the middle-aged adults. The weighted average scores of all precision, recall, and f1 score are around or above 0.81, which demonstrates a good quality of prediction that this optimized SVMC performed. Lastly, Figure 6.11 shows that the ROC curve of the SVMC has its AUC score of 0.78. Therefore, our optimized SVMC predicts depressive disorders diagnosis among the U.S. middle-aged adult residents in an effective manner.

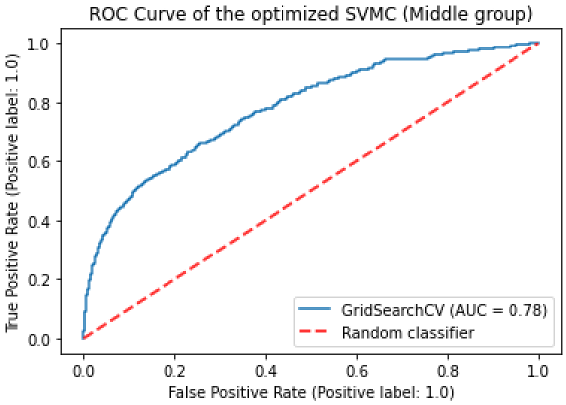

Figure 6.11: The ROC curve of the optimized SVMC model (middle-aged adult group)

6.3.3.3 Older adults

The parameters of this optimized SVMC that predicts depressive disorders among the U.S. older adults from the sample are presented in Table 6.25. The RBF function with the tuning hyperparameters $C = 100$ and $\gamma = 0.001$ produces the best balanced accuracy scores for the SVMC.

Table 6.26 is the confusion matrix of the optimized SVMC for the old-aged adults in the sample. First, 82.7% of people who have not been diagnosed with depressive disorders are correctly predicted (true negatives), while the remaining 2.7% are wrongly predicted as those who have been diagnosed

C	γ	kernel function
100	0.001	Radial Basis Function (RBF)

Table 6.25: The optimized parameters of the SVMC for old-aged adult group)

with depressive disorders (false positives). Also, 11.1% of people who have been suffering from depressive disorders are predicted as those who have not (false negatives), whereas the other 3.5% are correctly predicted (true positives). Then, the accuracy score of predictions made by this SVMC for the older adults is $\left(\frac{2347 + 99}{2347 + 78 + 315 + 99}\right) = 0.8616$.

	Predicted No	Predicted Yes
Actual No	2347	78
Actual Yes	315	99

Table 6.26: The confusion matrix of the SVMC model (old-aged adult group)

	Precision	Recall	F1 score
"No"	0.88	0.97	0.92
"Yes"	0.56	0.24	0.34
macro average accuracy	0.72	0.60	0.63
weighted average accuracy	0.83	0.86	0.84

Table 6.27: Classification report of the optimized SVMC model (old-aged adult group)

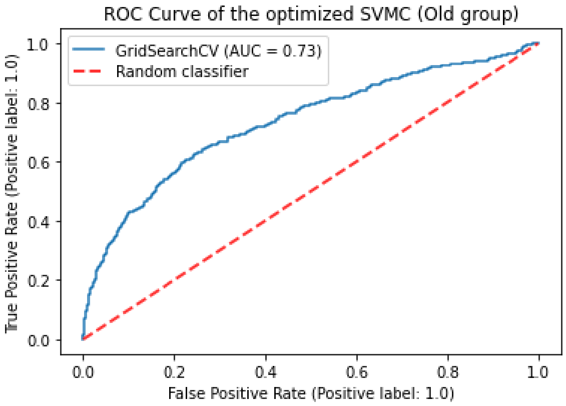

Figure 6.12: ROC curve of the optimized SVMC model (old-aged adult group)

Table 6.27 provides the prediction metrics of this optimized SVMC. Likewise, the weighted average scores of all precision, recall, and f1 score are around or above 0.83, which proves a good performance of this SVMC in this prediction task. Lastly, Figure 6.12 shows that the ROC curve has its AUC score of 0.73. Hence, the results confirm that our optimized SVMC predicts depressive disorder diagnosis among the U.S. old-aged residents in a good manner.

CHAPTER 7

Summary & Discussion

In this chapter, we summarize and discuss the important findings from the previous chapter. First, we compare and discuss our interpretations of the characteristics of U.S. adults diagnosed with depressive disorders, based on both decision trees and logistic regression models for each adult group. Then, we compare the performance of the three methods as predictors of depressive disorders diagnosis for each of the three adult groups.

7.1 Factor Discovery

7.1.1 Young adults

<table>
<tr><th>Decision tree</th><th>Logistic regression</th></tr>
<tr><td colspan="2">- # of bad mental days per month (↑)
- Decision-making issues (Yes)
- Arthritis or related illnesses (Yes)
- Difficulty of doing errands alone (Yes)
- # of children in household (↑)</td></tr>
<tr><td>- Employment status (student, homemaker, unemployed)
- Tobacco & Snuff history (Yes)</td><td>- Employment status (student, homemaker)
- Race (White)
- HPV test records (Yes)</td></tr>
</table>

Table 7.1: The most important variables for the **young adult group**, selected by decision trees and logistic regression: when holding all other conditions constant, categories in parentheses for each variable in the table are relevant factors that increase the likelihood of a depressive disorders diagnosis for the young adults.

Both the decision tree model and the logistic regression model ('M_{young}') present that the following five variables are relevant factors of the depressive disorders diagnosis among the U.S. young adult group in the 2018 BRFSS sample: the number of bad mental days, decision-making difficulty, arthritis, difficulty of doing errands alone, and the number of children in the household. Also, both methods select the current employment status of student and homemaker as another factor, but the only difference is that the decision tree model additionally picks an unemployed status. This means that stability in employment status can impact a young adult's mental state and have a relationship with their likelihood of depressive disorders.

In addition, the decision tree selects tobacco and snuff history as another potential deciding variable, while the M_{young} selects race and HPV test records. One interesting finding can be made in terms of HPV test records. It is widely known that the HPV test is the health exam for women's cervical cancer and that HPV is a serious sexually transmitted disease (STD) among the young-adult women. We observe how stressful it is to take the HPV test for women due to the social stigma around HPV, which may have an association with women's depressive disorders diagnosis.

Hence, we confirm that an individual's overall health status, including physical and mental conditions, demography, use of preventive health services and risk behaviors, impact the diagnosis of depressive disorders among the young adults.

7.1.2 Middle-aged adults

<table>
<tr><th>Decision tree</th><th>Logistic regression</th></tr>
<tr><td colspan="2">- # of bad mental days per month (↑)
- Decision-making issues (Yes)
- Employment status
(homemaker, students, retired, unemployed)</td></tr>
<tr><td>- Flu shot places
(health departments,
community health centers,
schools and workplaces)
- # of alcohol drinks per month (↓)</td><td>- Arthritis (Yes)
- Marital status (divorced)
- HIV test records (Yes)</td></tr>
</table>

Table 7.2: The most important variables for the **middle-aged adult group**, selected by decision tree and logistic regression: when holding all other conditions constant, categories in parentheses for each variable in the table are relevant factors that increase the likelihood of a depressive disorders diagnosis for the middle-aged adults.

The decision tree model and the logistic regression model M_{middle} select the number of bad mental days per month and decision-making difficulty as the important variables of depressive disorders diagnosis for the middle-aged adults in the 2018 BRFSS sample. We observe that an individual's mental states impact their depressive disorders diagnosis. Also, the two methods commonly select the current employment status of all categories but employed as another determinant. Since middle-aged adults are usually in the range of working ages, having jobs with monetary compensation would impact the depressive disorders diagnosis for this adult group.

On top of that, the decision tree selects the average alcohol consumption per month and the type of flu shot place as important determinants. On the other hand, M_{middle} considers arthritis, marital status, and HIV test records as important. We can provide additional insights into some of these results. First, a middle-aged adult's divorced status of marriage turns out to have the strongest relationship with their depressive disorders diagnosis. Hence, staying married is considered as the important feature of lives for the middle-aged adults from this sample data.

Second, HIV test records can determine the depressive disorders diagnosis for the middle-aged adult group. There exists harsh stigma around the HIV and AIDS all over the world, and no perfect treatments have been developed for the diseases. Therefore, the incurable nature of and negative perspective toward HIV and AIDS can impact the depressive disorders diagnosis for those in this adult group.

Lastly, non-doctor spaces such as health departments, community centers, workplaces and schools typically provide flu shots at lower costs. Hence, we reasonably assume that the impact of economic inequality on healthcare services has a certain relationship with depressive disorders diagnosis among the middle-aged adults in the sample.

7.1.3 Older adults

<table>
<tr><th>Decision tree</th><th>Logistic regression</th></tr>
<tr><td colspan="2">- # of bad mental days per month (↑)
- Decision-making issues (Yes)
- Healthcare inaccessibility due to high costs (Yes)
- The age when diabetes started (<72, ↓)</td></tr>
<tr><td>- Marital status (separated, never married, unmarried couple)
- Sleeping hours (<17)</td><td>- Marital status (non-divorced)
- Arthritis (Yes)
- Veteran status (No)
- Intestine-related exams (Yes)
- Lung-related illnesses (Yes)</td></tr>
</table>

Table 7.3: The most important variables for the **old-aged adult group**, selected by decision tree and logistic regression: when holding all other conditions constant, categories in parentheses for each variable in the table are relevant factors that increase the likelihood of a depressive disorders diagnosis for the older adults.

Both the decision tree and the logistic regression model M_{old} select the following variables as determinants of the depressive disorders diagnosis for the older adults in the 2018 BRFSS sample:

the number of bad mental days per month, decision-making difficulty, healthcare inaccessibility due to high costs, and the age when diabetes started. Similar to the young and middle-age groups, we confirm that older adults' mental health states are highly associated with the depressive disorders diagnosis. Also, we can assume that there is healthcare inequity among those in the older group because of high medical costs. Thus, we gain an insight into the impact of individuals' economic status on the quality of medical services among the U.S. older adults in 2018.

Furthermore, both methods state that when an older adult's diabetes starts at lower ages, particularly before 72, they have a higher likelihood of being diagnosed with depressive disorders. In addition, the two methods select the current marital status as another common factor, but the difference is that the decision tree picks a status of separated, never, or unmarried couple as the determinant, while the M_{old} selects all except divorced.

Besides those attributes stated above, the tree model chooses the average sleeping hours per day, while M_{old} picks arthritis, veteran status, intestine exams, and lung-related illnesses. Then, we examine some of those variables that were selected from these two methods. First, an older adult's non-veteran status has a relationship with their depressive disorders diagnosis, and we may guess that the U.S. governments provides some special care and priorities for the old-aged veterans, which highly impacts their mental states.

In addition, we imply that people's having an experience of intestine screening to check any tumor or colorectal cancer represents their usage of preventive services. Then, we can reasonably state that the older adults who had screening exams for preventing colorectal cancers are slightly more likely to be diagnosed with depressive disorders. More importantly, people's active usage of preventive care services increases their access to medical services, including mental healthcare. The older generation, in general, tends to rarely access mental healthcare services, or even related medical services, due to social stigma around a mental disorder diagnosis. If the old-aged adults have more frequent access to medical services, they are more likely to receive mental healthcare services, thereby increasing their likelihood of depressive disorder diagnosis.

7.1.4 Comparison to Literature Review

Our findings indicate that for all adult groups, people's physical and mental well-being, demography, healthcare access, and some health-related risk behaviors, such as high frequency of tobacco use, have significant relationships with their depressive disorder diagnosis. Those results show similar viewpoints with the two research papers presented in Chapter 1. However, our study presents the lower alcohol drinks per month as the relevant factor, which is opposite to the result from the research paper [22]. Also, our findings suggest the importance of one additional factor - the use of preventive services such as intestine-related exams for tumor screening.

7.2 Prediction Evaluation

7.2.1 Young adults

In Table 7.4, we observe that logistic regression performs better in terms of accuracy than other two methods, but it has comparable scores with the decision tree in precision, recall, and F1 score. But, the tree has the highest AUC values. Considering recall and F_1 scores, we conclude that both the decision tree model and logistic regression predict the depressive disorders diagnosis among the young adults of the BRFSS sample in the best manner.

	Accuracy	Precision	Recall	F1 score	AUC
Decision tree	0.836	**0.82**	**0.84**	0.81	**0.78**
Logistic regression	**0.841**	**0.82**	**0.84**	0.81	0.74
Support vector machine	0.825	0.80	0.82	0.81	0.74

Table 7.4: Comparison of classification metrics among the three methods (young adult group)

7.2.2 Middle-aged adults

In Table 7.5, we observe that the logistic regression has slightly better accuracy than others, but it shows comparable performance with the decision tree model in terms of precision and recall. Also, the decision tree has the highest AUC value among the three methods.

	Accuracy	Precision	Recall	F1 score	AUC
Decision tree	0.845	**0.83**	**0.85**	**0.83**	**0.80**
Logistic regression	**0.846**	**0.83**	**0.85**	0.82	0.73
Support vector machine	0.827	0.81	0.83	0.82	0.78

Table 7.5: Comparison of classification metrics among the three methods (middle-aged adult group)

7.2.3 Older adults

In Table 7.6, we observe that except accuracy, the decision tree outperforms or is on parallel with the logistic regression in terms of precision, recall, F1 score, and AUC values. However, the tree shows comparable accuracy scores with the other two methods.

	Accuracy	Precision	Recall	F1 score	AUC
Decision tree	0.870	**0.85**	**0.87**	**0.85**	**0.77**
Logistic regression	**0.872**	**0.85**	**0.87**	0.84	0.71
Support vector machine	0.862	0.83	0.86	0.84	0.73

Table 7.6: Comparison of classification metrics among the three methods (old-aged adult group)

Lastly, we observe that support vector machine classifiers show slightly lower scores for each metric in all adult groups, compared to the decision trees and logistic regression models. Considering that there was no feature selection process before building support vector machines models, we can recognize how powerful this classifier is, which uses all of the 84 explanatory variables without any pre-processing steps and generates comparable performance in prediction tasks.

CHAPTER 8

Conclusion & Future Works

We have created three different models for each adult group using supervised machine learning algorithms. Hence, we have achieved the two objectives of our research. Using the statistical and machine learning methods, we first discovered factors about health risk behaviors and societal attributes of depressive disorders for each group, and then we constructed and evaluated predictive models by comparing several metric scores.

First, we have comprehensively examined that the number of bad mental days and decision-making difficulty are the two most important variables for the depressive disorders diagnosis across all adult groups. Second, one's demographic attributes may accelerate the risk of depressive disorders diagnosis among the U.S. adult residents in 2018: the number of children in household and race for young adults; employment status for young and middle-aged adults; marital status for middle-aged and older adults; and veteran status for older adults.

In terms of health states and risk behaviors, the following factors turns out to be important determinants of depressive disorders in U.S. adults: arthritis for all age groups; difficulty of doing errands alone and tobacco use for young adults; the monthly average number of alcohol drinks for middle-aged adults; and the average sleeping hours, the age when diabetes started, and lung illnesses for older adults. Furthermore, individuals' healthcare access and use of preventive services can be a determinant of depressive disorders among the adults in the U.S., specifically, HPV test records for young adults; the type of flu shot places and HIV test records for middle-aged adults; and intestine-related exams for tumor screening and healthcare inaccessibility due to medical costs for older adults.

Finally, we have found that both decision trees and logistic regression produce similar quality of prediction performance, particularly accuracy scores, and agree on selecting important factors in common that can impact risks of depressive disorders diagnosis. Also, support vector machine classifiers have shown comparable performance with the other two methods, but they lack transparency as to what factors influence the prediction of depressive disorders among all age groups.

8.1 Limitations

First, our research has been conducted only on the BRFSS data collected in 2018. This indicates that the findings and conclusions from this research can be valid only for the survey respondents. Second, the BRFSS data takes the form of a telephone survey, which implies the possibility that the respondents of the BRFSS in 2018 may not provide true answers to the surveyors due to social stigma associated with the depressive disorders. This could generate biases or errors in collecting responses on the survey and ultimately distort the conclusion of our study.

As mentioned earlier, the support vector machine classifiers show good performance in predictions according to each metric score for all adult groups. However, since it uses every explanatory variable in the sample to train itself, this can result in a situation in which irrelevant explanatory variables may disturb the performance of predicting the depressive disorders diagnosis. To overcome this weaknesses in support vector machines, one may use principal component analysis for reducing the dimension of the data and extracting only relevant factors for the training process.

8.2 Future scope of the study

Although our study has examined the important factors of the depressive disorders diagnosis among U.S. adults, it is an indirect approach to the nature of any medical problem: we only analyzed the environmental factors of depressive disorders, not the direct causes in medical settings. Therefore, one may investigate the problem of our study with more clinical approaches: for instance, analyzing the behaviors and speech of the patients who suffer from depressive disorders, and thus detecting biological or behavioral markers of depressive disorders. In order to do so, researchers may use deep learning and computer vision techniques to examine the body gestures and the voice of patients. Also, they can use electronic health records (EHRs) or social media texts to conduct information extraction in the field of natural language processing. This approach will capture the medical indicators of depressive disorders in any linguistic format. Hence, those research processes within more clinical settings will provide us direct causes of depressive disorders among not only U.S. adult residents but also all people over the world.

APPENDIX *A*

Exploratory Data Analysis (EDA) Outputs

In Appendix A, we present the summary tables and several plots that were created during the exploratory data analysis (EDA) procedures. Interpretations about the data exploration for each adult group of our study are fully explained in Chapter 5.

A.1 Young Adults(18-39)

Tables A.1 to A.3 present the summary statistics of the response variable *ADDEPEV2* for young adult group between 18 and 39 years old, grouped by the following numeric variables *MENTHLTH*, *WEIGHT2*, and *CHILDREN*.

ADDEPEV2	Minimum	25% Quartile	Median	Mean	75% Quartile	Maximum
No	1	7	10.71	9.68	10.71	30
Yes	1	6	10.71	13.74	20	30

Table A.1: Summary table for young adult group: MENTHLTH

ADDEPEV2	Minimum	25% Quartile	Median	Mean	75% Quartile	Maximum
No	1	97	127	138.11	155	567
Yes	1	94	126	139.45	162	567

Table A.2: Summary table for young adult group: WEIGHT2

ADDEPEV2	Minimum	25% Quartile	Median	Mean	75% Quartile	Maximum
No	1	2	4	9.06	17	18
Yes	1	2	5	9.31	17	18

Table A.3: Summary table for young adult group: CHILDREN

Figures A.1 to A.7 shows the bar plots of the proportion of each factor levels of the following categorical variables: *DECIDE*, *EMPLOY1*, *DIFFALON*, *SMOKE100*, *HAVARTH3*, *HPVTEST*, *USENOW3* and *RACE*, which impact the response variable *ADDEPEV2*.

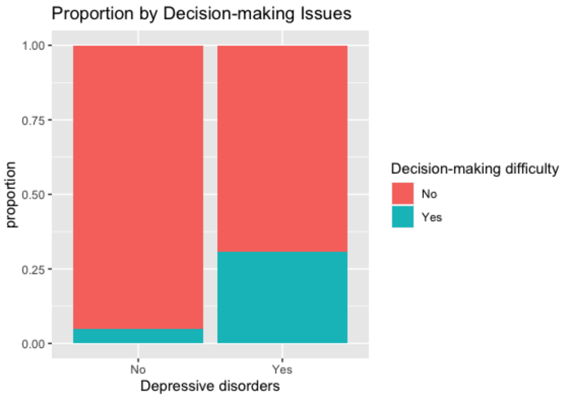

Figure A.1: Difficulties in making decision by oneself (Young adult group)

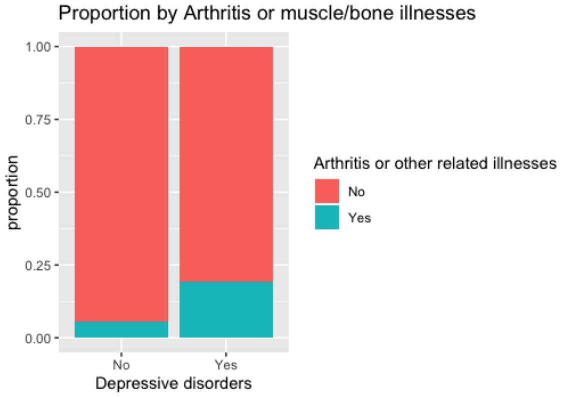

Figure A.2: Arthritis or bone-related illnesses (Young adult group)

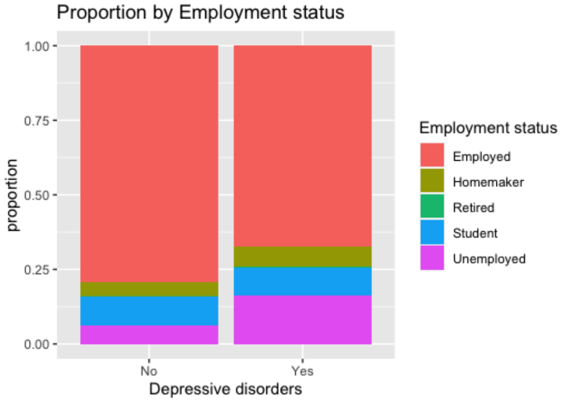

Figure A.3: Current employment status (Young adult group)

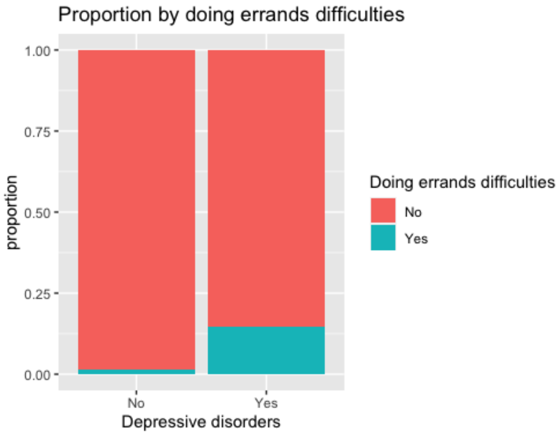

Figure A.4: Difficulties in doing errands alone (Young adult group)

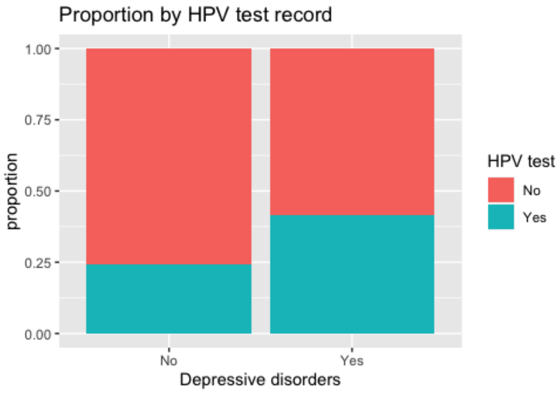

Figure A.5: HPV test records (Young adult group)

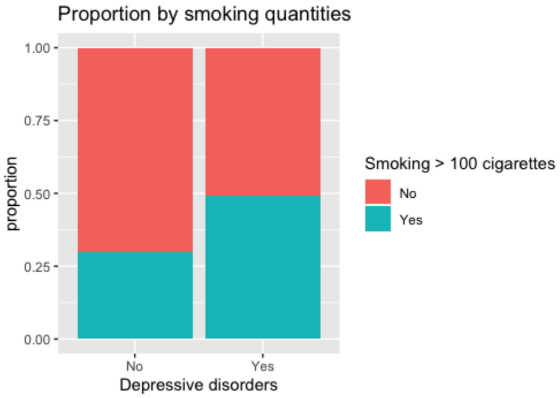

Figure A.6: Smoking more than 100 times (Young adult group)

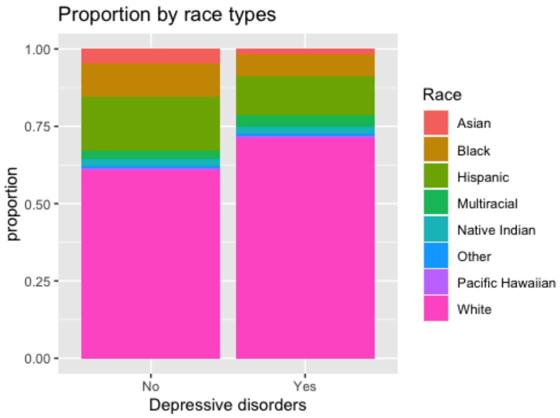

Figure A.7: Race types (Young adult group)

A.2 Middle-aged Adults (40-60)

Tables A.4 to A.5 present the summary statistics of the response variable *ADDEPEV2* for middle-aged adult group between 40 and 60 years old, grouped by the following numeric variables *MENTHLTH* and *AVEDRNK2*.

ADDEPEV2	Minimum	25% Quartile	Median	Mean	75% Quartile	Maximum
No	1	10.71	10.71	10.07	10.71	30
Yes	1	7	10.71	14.26	20	30

Table A.4: Summary table for middle-aged adult group: MENTHLTH

ADDEPEV2	Minimum	25% Quartile	Median	Mean	75% Quartile	Maximum
No	1	2	3.012	3.016	3.012	54
Yes	1	2	3.012	3.146	3.012	54

Table A.5: Summary table for middle-aged adult group: AVEDRNK2

Figures A.8 to A.13 shows the bar plots of the proportion of each factor levels of the following categorical variables: *DECIDE, EMPLOY1, MARITAL, HAVARTH3, IMFVPLAC, HIVTST6*, which impact the response variable *ADDEPEV2*.

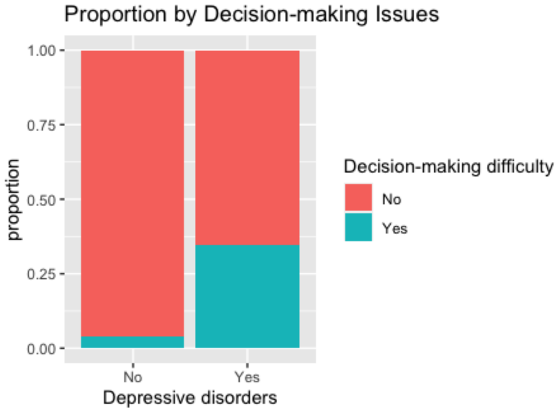

Figure A.8: Difficulties in making decision by oneself (Middle-aged adult group)

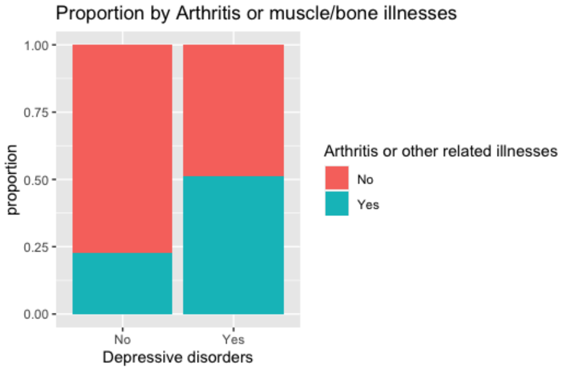

Figure A.9: Arthritis or bone-related illnesses (Middle-aged adult group)

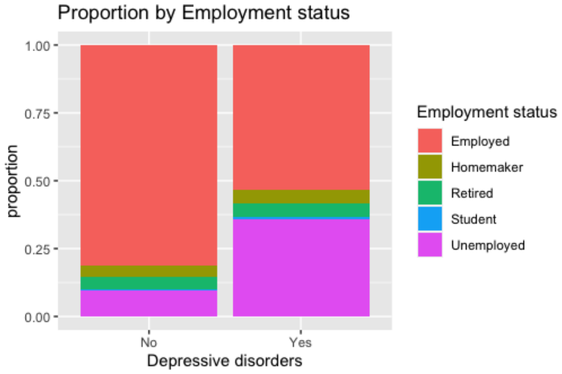

Figure A.10: Current employment status (Middle-aged adult group)

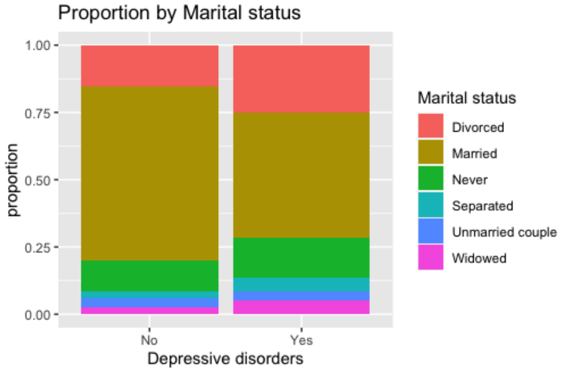

Figure A.11: Current marital status (Middle-aged adult group)

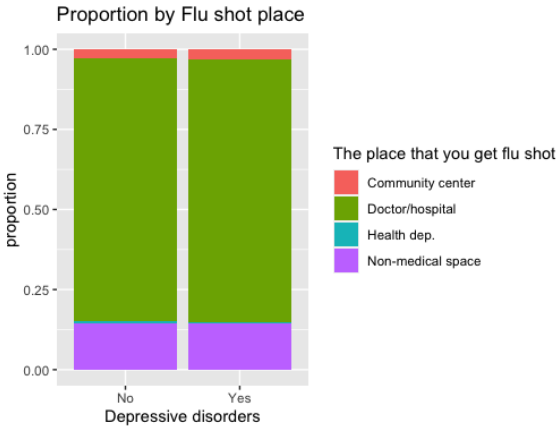

Figure A.12: The type of flu shot places (Middle-aged adult group)

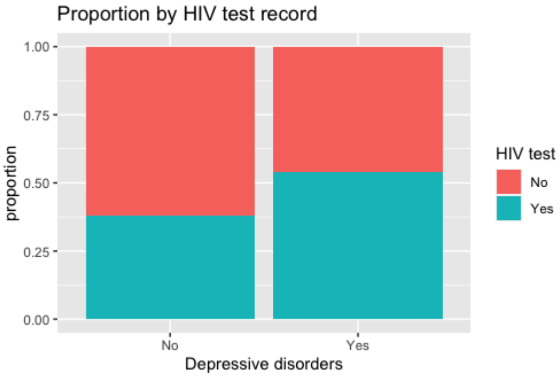

Figure A.13: HIV test records (Middle-aged adult group)

A.3 Older Adults(61-85)

Tables A.6 to A.8 present the summary statistics of the response variable *ADDEPEV2* for older adult group between 61 and 85 years old, grouped by the following numeric variables *MENTHLTH*, *DIABAGE2* and *SLEPTIM1*.

ADDEPEV2	Minimum	25% Quartile	Median	Mean	75% Quartile	Maximum
No	1	10.71	10.71	10.3	10.71	30
Yes	1	7	10.71	12.7	15	30

Table A.6: Summary table for older adult group: MENTHLTH

ADDEPEV2	Minimum	25% Quartile	Median	Mean	75% Quartile	Maximum
No	1	49.65	49.65	50.92	49.65	94
Yes	1	49.65	49.65	50.62	49.65	90

Table A.7: Summary table for older adult group: DIABAGE2

ADDEPEV2	Minimum	25% Quartile	Median	Mean	75% Quartile	Maximum
No	1	6	7	7.25	8	24
Yes	1	6	7	7.20	8	24

Table A.8: Summary table for older adult group: SLEPTIM1

Figures A.14 to A.20 shows the bar plots of the proportion of each factor levels of the following categorical variables: *DECIDE, VETERAN3, MARITAL, HAVARTH3, HADSIGM3, MEDCOST, CHCCOPD1*, which impact the response variable *ADDEPEV2*.

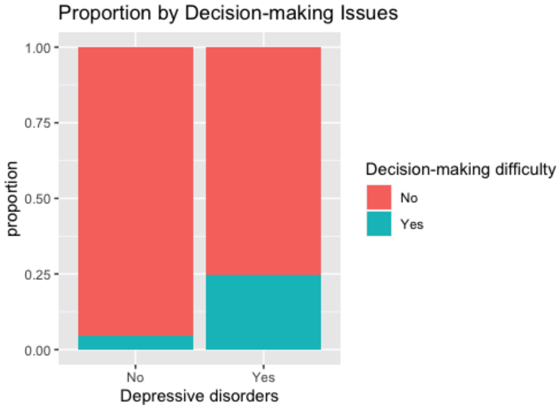

Figure A.14: Difficulties in making decisions by oneself (Older adult group)

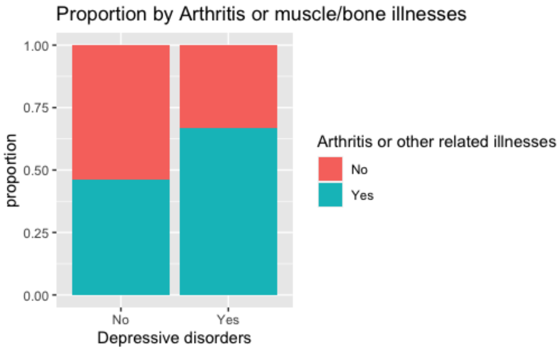

Figure A.15: Arthritis and bone-related illnesses (Older adult group)

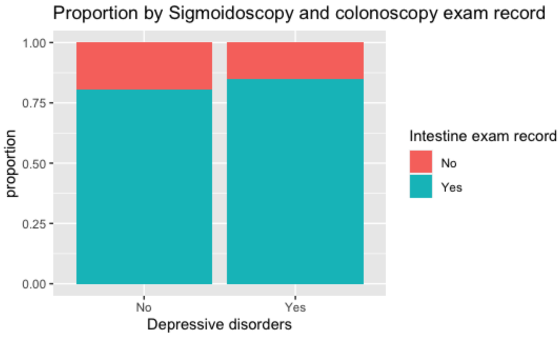

Figure A.16: Sigmoidoscopy and colonoscopy exam records (Older adult group)

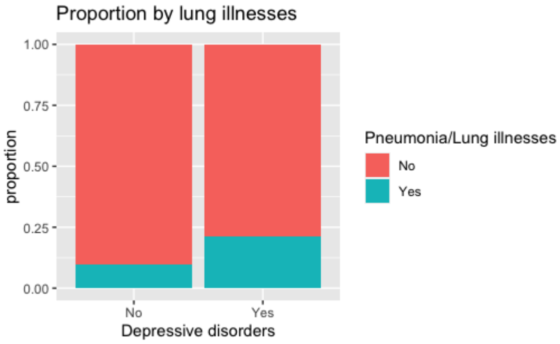

Figure A.17: Pulmonary illnesses (Older adult group)

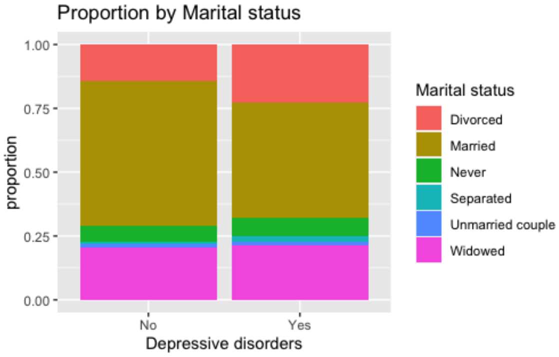

Figure A.18: Current marital status (Older adult group)

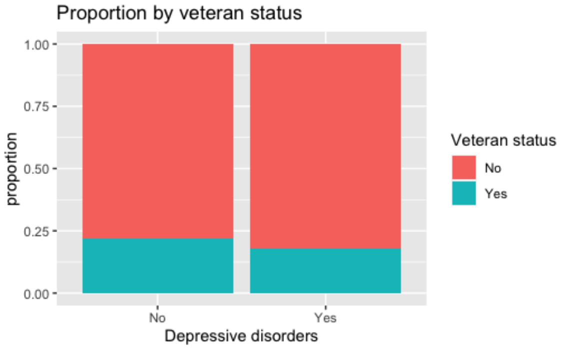

Figure A.19: Current veteran status (Older adult group)

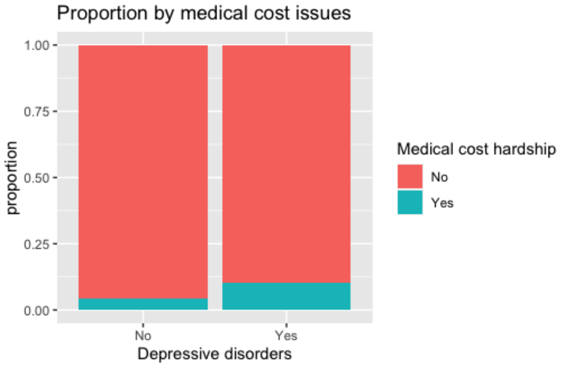

Figure A.20: Medical cost hardships (Older adult group)

References

1. Alan Agresti. *Categorical Data Analysis*. Wiley, USA, 3 edition, 2013. ISBN 978-0-470-46363-5. 34, 35, 36, 37, 40

2. Alan Agresti. *Introduction to Categorical Data Analysis*. Wiley, USA, 3 edition, 2019. ISBN 978-1-119-40526-9. 34, 36, 37

3. Arun Amballa. Feature engineering part 1: Mean/median imputation, 2020. URL https://medium.com/analytics-vidhya/feature-engineering-part-1-mean-median-imputation-761043b95379. 55

4. Giuseppe Bonaccorso. *Machine Learning Algorithms: A Reference Guide to Popular Algorithms for Data Science and Machine Learning*. Packt Publishing, 2017. ISBN 1785889621. 8, 9, 42, 43, 44, 45, 46, 47, 49

5. Jason Brownlee. Feature selection for machine learning in python, 2016. URL https://machinelearningmastery.com/feature-selection-machine-learning-python/. 55

6. Jason Brownlee. How to calculate feature importance with python, 2020. URL https://machinelearningmastery.com/calculate-feature-importance-with-python/. 56

7. Ann Cannon, George Cobb, Bradley Hartlaub, Julie Legler, Robin Lock, Thomas Moore, Allan Rossman, and Jeffrey Witmer. *Stat2Data: Datasets for Stat2*, 2019. URL https://CRAN.R-project.org/package=Stat2Data. R package version 2.0.0. 28

8. Erito Marques de Souza Filho, Helena Cramer Veiga Rey, Rose Mary Frajtag, Daniela Matos Arrowsmith Cook, Lucas Nunes Dalbonio de Carvalho, Antonio Luiz Pinho Ribeiro, and Jorge Amaral. Can machine learning be useful as a screening tool for depression in primary care? *Journal of Psychiatric Research*, 132:1–6, 2021. ISSN 0022-3956. doi: https://doi.org/10.1016/j.jpsychires.2020.09.025. URL https://www.sciencedirect.com/science/article/pii/S0022395620309912. 2

9. Ann R. Cannon et al. *STAT2: Building Models for a World of Data*. W.H. Freeman, 2011. xv, xvii, 20, 21, 23

10. Leo Breiman et al. *Classification and Regression Trees*. Chapman and Hall CRC, 1984. 6

11. Aurélien Géron. *Hands-On Machine Learning with Scikit-learn, Keras, and Tensorflow*. O'Reilly Media, Inc, 2 edition, 2019. 3, 48, 49

12. Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 3 edition, 2011. xv, 3

13. Kaggle. Play tennis - simple dataset with decisions about playing tennis, 2018. URL https://www.kaggle.com/fredericobreno/play-tennis. xvii, 10, 11

14. Yi-Zeng Liang, Qisong Xu, Hong-Dong Li, and Dong-Sheng Cao. *Support Vector Machines and Their Application in Chemistry and Biotechnology*. CRC Press, 03 2013. 41

15. Stephen Marsland. *Machine Learning: An Algorithmic Perspective, Second Edition*. Chapman and Hall CRC, 2nd edition, 2014. ISBN 1466583282. 3, 6

16. Thomas M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., USA, 1 edition, 1997. ISBN 0070428077. 3, 6, 7, 8, 9

17. Center of Disease Control and Prevention. The behavioral risk factor surveillance system (brfss): Overview., 2018. URL https://www.cdc.gov/brfss/annual_data/2018/pdf/overview-2018-508.pdf. 53, 54, 55

18. National Institute of Mental Health. Depression, 2021. URL https://www.nimh.nih.gov/health/topics/depression/index.shtml. 1

19. Alexis Perrier. Feature importance in random forests, 2015. URL https://alexisperrier.com/datascience/2015/08/27/feature-importance-random-forests-gini-accuracy.html. 56

20. Cosma Rohilla Shalizi. *Advanced Data Analysis from an Elementary Point of View*. Department of Statistics, Carnegie Mellon University, 2019. 40

21. Tara W. Strine, Ali H. Mokdad, Lina S. Balluz, Olinda Gonzalez, Raquel Crider, Joyce T. Berry, and Kurt Kroenke. Depression and anxiety in the united states: Findings from the 2006 behavioral risk factor surveillance system. *Psychiatric Services*, 59(12):1383–1390, 2008. doi: 10.1176/ps.2008.59.12.1383. URL https://ps.psychiatryonline.org/doi/abs/10.1176/ps.2008.59.12.1383. PMID: 19033164. 2

22. Tara W. Strine, Ali H. Mokdad, Shanta R. Dube, Lina S. Balluz, Olinda Gonzalez, Joyce T. Berry, Ron Manderscheid, and Kurt Kroenke. The association of depression and anxiety with obesity and unhealthy behaviors among community-dwelling us adults. *General Hospital Psychiatry*, 30 (2):127–137, 2008. ISSN 0163-8343. doi: https://doi.org/10.1016/j.genhosppsych.2007.12.008. URL https://www.sciencedirect.com/science/article/pii/S0163834307002629. 2, 94

23. Kenneth Tay. Deviance for logistic regression, 2019. URL https://statisticaloddsandends.wordpress.com/2019/04/22/deviance-for-logistic-regression/. 35

24. Jake VanderPlas. *Python Data Science Handbook*. O'Reilly Media, Inc, 2016. ISBN 9781491912058. 46, 47

25. Jaime Zornoza. The roc curve: Unveiled, 2019. URL https://towardsdatascience.com/the-roc-curve-unveiled-81296e1577b. xv, 49, 50