12-2023

Statistical Analysis of Health Habits for Incoming College Students

Wendy Isamara Lizarraga Noriega

Statistical Analysis of Health Habits for Incoming College students

---

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

in

Mathematics

---

by

Wendy Lizarraga Noriega

December 2023

Statistical Analysis of Health Habits for Incoming College students

---

A Thesis

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Wendy Lizarraga Noriega

December 2023

Approved by:

Dr. Hani Aldirawi, Committee Chair

Dr. Suthakaran Ratnasingam, Committee Member

Dr. Dalton Marsh, Committee Member

Dr. Madeleine Jetter, Chair, Department of Mathematics

Dr. Corey Dunn, Graduate Coordinator

Abstract

Health habits among college students are commonly overseen, especially for students transitioning from high school right into college. These students are becoming independent young adults, and learning how to adapt to a different scenery when it comes to their learning environment. As these young adults transition into college, this is the perfect time for the students to become more vulnerable and comfortable with their independence, and their weight begins to fluctuate.

Many variables come into consideration when increasing weight as an incoming first-year student. Students are more likely to live alone, get a job, and rely on fast food and quick snacks, all while juggling school, which may also lead to stress and binge eating. Not only may their stress lead to binge eating, but it may also include alcohol consumption as well as Marijuana consumption.

This topic is essential because if unmeasured, weight gain may lead to health problems, which may also interfere with the student's learning. It is also crucial because making students aware of their surroundings and possible outcomes may prevent them from gaining substantial unhealthy weight.

Our study consists of about 900 students from a large Hispanic-serving institution in the Western United States. We applied three different statistical models to predict the student's weight status given 34 covariates such as gender, ethnicity, stress, and alcohol assumptions. The analysis based on our dataset shows that the neural network model has the highest performance. Our findings have significant implications for students' health.

Acknowledgements

First and foremost, I am most thankful for my parents, Marcos Lizarraga and Isabel Noriega. I would not be where I am if it were not for their hard work, dedication, advice, and sacrifices, but most importantly, for their pure, unconditional love and support. My family has been my backbone through this journey, and I could not have done it without any of them. I am also thankful for my siblings' support, Sandy Lizarraga, Casandra Lizarraga, and Marcos Lizarraga, and I wish to one day inspire their kids: Carlitos Lomeli Lizarraga, Damian Lomeli Lizarraga, Christopher Herrera Lizarraga, Sol-Luna Lizarraga, Catalina Herrera Lizarraga, Amias Lizarraga, and Ilisapeta Yuseili Lizarraga-Nand.

In the same manner, words cannot express my gratitude to my advisor, Dr. Hani Aldirawi. I thank him for his feedback, support, late-night back-and-forth emails, and the immense patience he has demonstrated throughout my research journey. He has inspired and helped me through the process and has continued to help me grow as a student and educator. I could not have come this far without him, and I am significantly thankful for that.

I am also grateful to my committee, Dr. Dalton Marsh and Dr. Suthakaran Ratnasingam, for taking the time to support me throughout the writing of my thesis. I thank them for their valuable comments and suggestions.

I would additionally like to thank my former professors From San Bernardino Valley College and California State University, San Bernardino, for motivating me and sharing their knowledge and love of Mathematics. I hope to become soon a professor who inspires students just as much as they have all inspired me.

Table of Contents

Abstract **iii**

Acknowledgements **iv**

List of Figures **vii**

List of Tables **viii**

1 Introduction **1**
1.1 Background 1
1.2 Research Objectives 3
1.3 Organization of Thesis 3

2 Literature Review **5**
2.1 Freshman 15 Factors: An Overview 5
2.2 Statistical Analysis and Machine Learning Methods for Studying Freshman 15: An Overview 9

3 Methods and Statistical Analysis **12**
3.1 Introduction 12
3.2 Data Source and Description 12
3.3 Statistical Analysis and Model Comparison 13
3.4 Logistic Regression 16
3.5 Model Selection 17
3.6 Neural Networks 18
3.7 Decision Tree Method 20
3.8 Cross-Validation 21
3.9 Evaluation Metrics 23

4 Results and Discussion **26**
4.1 Introduction 26
4.2 Results and Discussions 26

4.2.1 Descriptive Analysis 26
4.2.2 Statistics and Machine Learning Models 30
4.3 Logistic Regression 30
4.3.1 Significant Variables and Interpretations 31
4.4 Neural Network Analysis 34
4.5 Decision Trees Analysis 35
4.6 Model Training and Evaluation 36

5 Conclusion **39**

Appendix A Survey **41**

List of Figures

3.1 Survey Email 13
3.2 Neural network example. Circles present nodes (Neurons). 19
3.3 Decision tree model Example. The tree is made up of nodes that represent the various features in the dataset, and it starts at the root node, which represents the entire dataset. 20
3.4 Cross-validation Principle: Overview of the machine learning models used to predict the weight status. The architecture includes data splitting, feature selection, model training, and model evaluation. Image designed by the Shanthababu Pandian, 2022. 22
3.5 A diagram of 10-fold cross-validation. The dataset is randomly divided into five folds of equal size. One of the folds is kept as the validation set in each iteration, while the remaining four folds are used to train the model. 23
3.6 An example of a confusion matrix. The matrix displays the model's true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions. 24

4.1 Histogram:"Weight increased" Variable 27
4.2 Boxplot:"Weight increased" Variable 28
4.3 Weight Status During Freshman Year 29
4.4 Confusion matrices for the evaluated model on the original dataset and using the cross-validation training dataset. The diagonal elements represent the number of correctly classified weight status cases, while the off-diagonal elements represent weight status cases that were misclassified. A 0.5 cut-off value was used. 38

A.1 IRB Approval Letter 48

List of Tables

4.1 Description Statistic of "weight increased" Variable 27
4.2 Weight Status 28
4.3 Weight and Ethnicity 30
4.4 Logistic Regression Result 31
4.5 Precision, Sensitivity, and Accuracy for Comparison ANNs Models With Different Number of Neurons 35
4.6 Evaluation of Model Performance for Weight Status Prediction . . . 37

Chapter 1

Introduction

1.1 Background

"The best is yet to come!" This statement is what high school seniors hear before they graduate. They believe they will soon be free because they turn 18 years of age shortly after. In reality, once they attend college, their lives truly begin. They become independent and work for what they want on their own. What the students need to know is that their freedom comes with responsibilities. Not only do they begin college, but for some, they do something they have never done before: they move out, get a job, and begin to pay bills. The life-changing events for incoming college students create a sense of freedom, which leads to young adults gaining a substantial amount of weight as they face the variables that come into play as first-year college students. It is found that most of the increase in weight happens when transitioning from adolescence to adulthood [Nel+08]. Transitioning from adolescence to adulthood is also when life-changing events occur, such as educational transitions, financial situations, independent decisions, and others. Since first-year students are adapting to life after high school, they are known to gain ten to fifteen or more pounds of weight. Weight gain can be caused by many factors such as eating habits [Roy+19; VE10], physical activity [YH20], stress [Hai+18], alcohol consumption [SMB11], smoking [GTM10], the environment they are in [Cli13], ethnicity [DSo+15], adaptation to work, and living status. Due to these changing life

events, freshmen have been known to have the highest weight gain in their first years of college."Freshman 15" is a term for first-year college students averaging a weight increase of 15 pounds.

This topic is important because many students experience this as they are transitioning from high school to college. Many young adults do not realize a pattern until it has escalated [YH20]. It is important to make the students aware because increasing unhealthy weight leads to health issues as well as affecting their performance in school. Gaining unhealthy weight as an incoming college student is not addressed enough as it should be because not only does one gain weight, but one also adopts unhealthy habits that may affect one's health in the fullness of time. Unhealthy weight increase in young adults can lead to obesity, and obesity is known to be associated with significant diseases such as diabetes and heart attack [Cen+19]. Not only is being overweight bad for the health, but it may also affect how students learn. It may affect them because it can cause trouble with sleeping, which will cause students to be fatigued and not fully paying attention in their courses.

To further understand why young adults who attend college right after high school gain significant weight in their first two years, research must be done to identify the factors that contribute to the problem. By identifying which factors have the more significant impact, solutions may be found to help students know why and what they can do to prevent weight gain. Even though the consequences of being overweight might not affect one's health in the present, it unfortunately affects one's health in the future. Collecting data on college students will help identify which predictors lead to obesity and, by doing so, which factors can be prevented by choice.

The university under study is located in a low-income community and serves a largely Hispanic population. Obesity is associated with a lack of access to healthy foods which in turn puts individuals from low-income and minority backgrounds at higher risk of developing diabetes. Hispanics are 1.5 times more likely to have diabetes than non-Hispanic white Americans [Sel+14]. Therefore, the results of

this study have important implications for understanding weight gain in college students from high-risk populations.

1.2 Research Objectives

The overall research goal is to analyze the health habits of incoming college students. The main goal is to determine which factors are more significant in resulting unhealthy weight gain as an incoming college freshman. More specifically, the objective is to declare possible predictors that lead to freshman 15. Knowing those factors helps prevent weight gain during their first year of college.

Our study has 34 independent variables (predictors) and one dependent variable (weight status). Please see the attached appendix A for more details. We analyzed a survey that was given to the university students and performed some statistical analysis methods.

1.3 Organization of Thesis

This thesis consists of five parts. Chapter 1 introduces the topic of the study, providing background information and identifying the specific problem being investigated. This chapter also outlines the study's objectives and the methods that will be used to achieve them.

Chapter 2 presents literature reviews related to the topic of the study. This chapter includes a discussion of relevant studies, previous research studies, and any other information that is pertinent to the research question.

Chapter 3 provides an in-depth discussion of the methods used to build the thesis. Here, we detail the methods and procedures used to gather and analyze data and any tools or techniques utilized throughout the analysis.

Chapter 4 presents some discussions and results of the study, using visual representations such as tables and graphs to illustrate key findings. This chapter also includes a detailed discussion of the results.

Finally, Chapter 5 provides the study's conclusions, highlighting the essential findings and their implications. This chapter also includes some recommendations for future research work. Overall, this thesis provides a comprehensive and detailed investigation of the chosen topic, using various methods and techniques to explore and analyze some major research questions.

Chapter 2

Literature Review

2.1 Freshman 15 Factors: An Overview

In this section, many of the research articles discuss the results of the transition as an incoming college student and how their eating habits, physical activity, alcohol, marijuana, and electronic-cigarettes (e-cigarettes) consumption, and gender may have affected the students to increase their weight as college freshmen. Freshman 15 is a significant public health problem worldwide, and its prevalence is higher in students with a lack of exercise, poor sleep and eating habits, and immoderate consumption of alcohol. According to Mihalopoulos, Freshman 15 refers to "the belief that college students frequently gain 15 lbs during their first year" [MAK08].

Vadeboncoeur et al. (2015) research study was conducted to determine if gaining weight as a first-year college student is a myth or a fact [VTF15]. This article cites 32 studies from 1980 to 2014. Five thousand five hundred twenty-nine studies were part of the research to average the weight gained during the first year. The mean for the students was 3 pounds for every 5 months, which averaged an amount of about 8 pounds. Although some of the studies did not show that students gain an average of 15 pounds in their first year of college, the studies do, however, show that college students gain more weight than non-college students.

One of the factors that contributes to freshman 15 is student's eating habits. According to Roy et al. (2019), The variables that play a role in eating habits are

all-you-can-eat buffets, vending machines, and food trucks on campuses, which give students access to "unhealthy" food and as a result, weight gain occurs [Roy+19]. It was also found that the highly promoted food on campus via email, posters, or social media posts influences students to obtain it more frequently.

Vending machines placed on campus may be convenient to the students due to their lack of time [Byr+12]. As many students are running late to class, with a short time in between breaks, hurriedly and hungry, they tend to catch themselves buying candy and chips from the school's vending machine. It is easy, quick, and convenient but damaging to the body. An assessment by Byrd et al. (2012) identified that the snacks sold the most were sweet and salty snacks; however, all the options sold in the vending machines were considered unhealthy, high in carbohydrates, poor nutrients and high in sugar or salt. In conclusion, there was a high correlation between sugar/salted snacks and weight gain [Byr+12].

Sometimes, the problem is more than just eating junk food or overeating. It can also be not eating enough, which can also be a factor in weight gain. Pendergast et al. (2016) study focuses on young adults aged 18 to 30 who tend to skip meals [Pen+16]. Even though eating habits have been known to be an issue, meal skipping needs to be talked about more. Based on this article, meal skipping is known as having worse diet quality. Skipping on necessary nutrients causes less energy, fewer vitamins consumed, and an increase in weight. The participants in this research were required to be 18 to 30, and they needed to be recommended by a college. Their research found a correlation between students skipping meals and students with a self-perception of being overweight. On top of the choices students make to eat unhealthy food and snacks, stress also affects how students eat and how often.

Goldschmidt et al. (2008) found that Binge Eating Disorder (known as BED) has been highly correlated and negatively affected by stress [Gol+08]. According to the American Psychiatric Association, BED is the most common disorder that causes people to eat significant portions in a short interval of time [AA+13]. In fact, individuals who binge eat are more likely to become obese than those who do not [Hud+07].

Alongside poor eating habits, physical activity is also crucial to one's health and weight. As students transition from high school to college, less physical activity is performed due to the lack of time. In high school, students are required to have Physical Education, which is a by-choice in college. In high school, various sports are offered for all levels or any activities that require them to be active. High school students mainly live with their legal guardians, where most of the time, they are not required to have a job, which helps with staying active.

In addition to eating habits and physical activity, alcohol and nicotine use disorders have been highly correlated with college students [Hef+19]. Both alcohol and e-cigarettes are common substances young college adults utilize as they are getting through their first years of college, and they are both a public health concern for these young adults [WWZ22]. Wechsler et al. (2000) state that America's colleges are the most known for heavy episodic drinking, and it also states that college students binge drink more than students who do not go to college [Wec+00]. One of the reasons why college students binge drink is because of easy accessibility. There are bars either located on campus or within a mile from campus [Wec+00]. Another reason why college students are more prone than non-college students is that, based on Wechsler et al. (1995) and their findings, being involved with athletics, living on campus, attending social school events, and having high interactions with friends had a higher chance of being related to binge drinking while in college[Wec+95]. Alcohol is a high-calorie beverage that can also become addicting to those who consume it. Even though it is not proven that this is always the case, there is a high correlation between alcohol and weight gain for those who turn into heavy drinkers [SMB11]. Alcohol is also calorically dense, just like fast food, and alcohol consumption may induce alcohol-related eating [CKG13]. After smoking, alcohol consumption is the next highest leading cause of premature death in the United States. [OKe+14]. Similarly, when stressed, Students also tend to lean towards drinking alcohol and utilizing Electronic cigarettes, which are also known as E-cigarettes because they consider them stress-relieving [GBG14]. E-cigarettes are products that contain nicotine product. This product is created by heating

a solution made up of propylene glycol or glycerol. There are different types of e-cigarettes and the ways they can be utilized. E-cigarettes were mainly created to experience smoking conventional cigarettes but in the form of a pen instead. Since young adults smoke, whether it is for fun or to de-stress, e-cigarettes have been known to also cause weight gain [Sut+13]. They mainly cause the weight increase after one stops using after having used it for a while. During the time these young adults are smoking e-cigarettes, their appetite decreases; they either maintain weight or thin out slightly, but after they stop consuming, they jump to an average of about 10 pounds because of the appetite change[BP18]. Young adults from the age 18-21 who vape have been found to have long-term health problems as well as it leading to eating disorders [Hen].

Moreover, gender can also be a factor in Freshman 15. Males and females digest differently and the effect on their health and weight can vary [Llo+09]. A study was conducted where 904 students, 18 years or older, full-time students, and US citizens at a college in Indiana were examined in their first and second years of college. The second one was to examine the weight gain for a sample of 386 students at a private university in Rhode Island. Both studies had a rough average of 74 percent weight gain in the first year, especially in the first semester. In year 1, study one, both males and females had the same weight fluctuation of an estimated 8-10 pounds. In the second study, however, males and females did not have the same weight gain, and their weight gain was also less than in study 1. Male and female weight gain difference was about 3 pounds from one another, where females gained about 5 pounds, and males gained about 8 pounds. The second year was a higher percentage weight gain, where the difference from the beginning of their first year was an average of 12 pounds. Body Mass Index difference for study one had a 10 percent increase over the two years, whereas the difference for study 2 was a 3 percent increase.

Bodenlos et al. (2015) study focuses on gender differences in first-year students' weight gain. The difference between gender weight gains their first year [BGS15]. During September 2010, 2011, and 2012, 304 freshmen participated in this study.

The number of females and males was not stated. However, the difference between each gender's weight gain was not too significantly different. The average gain weight for males was almost 7 pounds, and the average for females was almost 5 pounds. Males were involved with alcohol. The conclusion was that most weight was gained the first semester, and different factors played into each gender.

2.2 Statistical Analysis and Machine Learning Methods for Studying Freshman 15: An Overview

In recent years, various statistical analysis methods and machine learning algorithms have been increasingly used to study freshman 15 factors. This chapter provides an overview of two parts. The first part is studying the current literature discussing freshman 15 factors. The second part focuses on the use of statistical models and machine learning algorithms for studying Freshman 15.

Gillen et al. (2011) applied the logistic regression model for predicting the freshman 15. The independent variables in the study were gender, African American ethnicity, mother's education, SAT score, and campus group activity. SAT scores were a significant predictor of freshman 15 (B=.04, Odds Ratios=0.79). There was a pattern for students who did better than others on the SAT which came into effect on identifying who would most likely fall under the category of gaining an average of 15 pounds. Students who performed slightly better on the SAT by a difference of 10 points resulted in 3% less chance of falling under the Freshman 15 category. Another factor that showed to be a pattern was students who were more involved with campus activities. The odds of increasing weight and falling under the Freshman 15 category was 28% less likely [GL11]. Ethnicity, gender, and mother's education were insignificant variables for predicting freshman 15. Although men and African Americans weighed more than women and some other racial groups, their weight change is not significant. In addition, Gillen et al. (2011) found that students with lower SAT scores had a greater likelihood of experiencing increasing weight during their first year. Given that those were the patterns related to not

increasing weight, it is connected to the fact that students who are more involved as well as more academically active have their time occupied and well-strategized. When students have a set agenda as far as doing well in their classes, attending clubs and meetings, and getting their work done without procrastination, it does not leave room for them to procrastinate and stress which leads to binge eating [GL11].

Baum et al. (2017) applied a multivariate regression model to study the relationship between college attendance and gaining weight. For one of the observations, the average annual weight since the age of sixteen is being examined. Another observation was the annual weight change from a previous survey. The statistics for the 17-23 year-olds in the survey gained 3-4 pounds per year, about 47 percent of the students who participated were attending college, and most of the respondents were freshmen [Bau17].

Deforche et al. (2015) discussed the factors that affect freshman 15 [Def+15]. A sample of 291 students was collected. The student's weight was measured and tracked at the beginning of their high school senior year and until their second year of college. The independent variables were physical activity, dietary intake, and sedentary behavior. ANOVA analyses were conducted repeatedly to keep track of their measures. Another method used for the study was stepwise multiple regression analysis to understand the difference between gaining weight and health behavior changes and to see the weight changes between genders. The results over the years were that the males increased an average of 10 pounds, and the females increased an average of 5 pounds. The variables that affected the students the most to a higher BMI were higher alcohol consumption, less physical activity, and less study time. The primary variable for the male group was alcohol consumption, whereas for the girls, it was the eating habits.

Singh et al. (2019) study is based on the logistic regression model and artificial neural networks (ANN) approach for estimating the weight gain for young adults [ST19]. The results indicated that the ANN algorithm using 20 neurons in the hidden layer performs better compared to the logistic regression and other ANN

approaches using different numbers of neurons. The best prediction accuracy of 93.4% is based on ANN.

Babajide et al. (2020) study was based on the connection between dietary behavior and overweight [Bab+20]. They investigated some machine learning algorithms to improve the prediction of body weight. The results indicate that Artificial Neural Networks and Random Forests perform better than Support vector machines and Linear regression in predicting body weight.

Following the review of relevant literature on the suggested topic, this work seeks to utilize statistical analysis techniques to evaluate the factors of freshman 15.

Chapter 3

Methods and Statistical Analysis

3.1 Introduction

In this chapter, we discuss the research method, data collection, variables included in the study, estimation procedures, methods in building the model, and some relevant statistical computations.

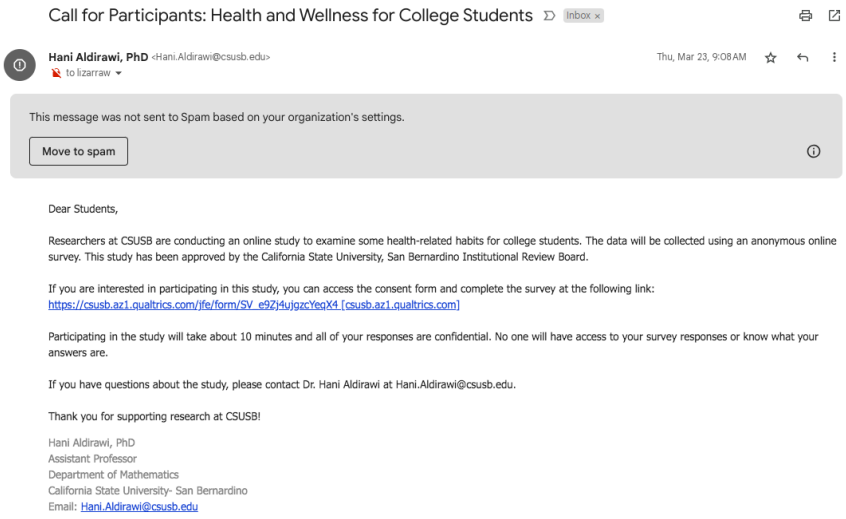
3.2 Data Source and Description

Dr. Hani Aldirawi created an online survey and sent it to the Institutional Review Board (IRB) for review. When it was approved (IRB Number: FY2023-245), a survey was sent out to students in a Hispanic-serving institution in the Western United States via their school email. In the email the students received, it was stated there was research being done to examine health-related habits for college students. It was also reassured the data would be anonymous. Therefore, students felt safe to respond to the survey freely. This information was also shared with Dr. Aldirawi's colleagues. They were given the survey link, and they were asked to share it with their students/classes if there was an opportunity. The survey given to the students asked questions about some health-related issues. Below is the image of the email that was sent out.

We sent out the survey to the students from the university under study in

March 2023. The students who participated in the survey were students of all levels currently attending the university, but their answers were based on their first year of college. The response (dependent) variable is whether the student's weight increased or not during their freshman year.

There are 34 independent variables (predictors) in our study. The predictors are college, status, credits, residence, area, sex, sex orientation, race, relationship, first generation, GPA, employment status, income, parental income, parental degree, high school sport, high school job, high school activities, parent cooking, activities, exercise, fast food, vending machine, school restaurant, depression, stress, TV, video games, sleeping, E-cigarette, alcohol, and marijuana. For more details about the survey, please see the attached appendix A. We Analyzed a survey that was given to the university and performed some statistical analysis methods.


Call for Participants: Health and Wellness for College Students Inbox ×

Hani Aldirawi, PhD <Hani.Aldirawi@csusb.edu> Thu, Mar 23, 9:08 AM
to lizarraw

This message was not sent to Spam based on your organization's settings.

Move to spam

Dear Students,

Researchers at CSUSB are conducting an online study to examine some health-related habits for college students. The data will be collected using an anonymous online survey. This study has been approved by the California State University, San Bernardino Institutional Review Board.

If you are interested in participating in this study, you can access the consent form and complete the survey at the following link:
https://csusb.az1.qualtrics.com/jfe/form/SV_e9ZJ4ujgzcYeqX4 [csusb.az1.qualtrics.com]

Participating in the study will take about 10 minutes and all of your responses are confidential. No one will have access to your survey responses or know what your answers are.

If you have questions about the study, please contact Dr. Hani Aldirawi at Hani.Aldirawi@csusb.edu.

Thank you for supporting research at CSUSB!

Hani Aldirawi, PhD
Assistant Professor
Department of Mathematics
California State University- San Bernardino
Email: Hani.Aldirawi@csusb.edu

Figure 3.1: Survey Email

3.3 Statistical Analysis and Model Comparison

There are two main methods for modeling data: "supervised learning" and "unsupervised learning". Supervised learning consists of both independent variables

and a dependent variable whose value is to be estimated. Unsupervised learning consists of a set of independent variables but no dependent variable. The idea of unsupervised learning is to group information based on similarities, patterns, and differences instead of estimation.

In this thesis, we are going to predict whether the students will gain weight during their first year of college or not. Our dependent variable is a binary response (Yes or No). Logistic regression, Neural networks, and decision trees are some choices to predict whether a student gains weight or not. The following is a short comparison between the three models. However, a complete description of these three methods is described later in this chapter.

1. **Logistic Regression**:

    - Advantages: Logistic regression is an easy way to interpret and understand. Logistic regression is a relatively simple algorithm, and its predictions are easy to understand. In addition, it's less computationally expensive compared to other algorithms, such as neural networks. This can be important for problems with a large amount of data or where the model needs to be deployed in real-time. Logistic regression is more robust to overfitting and less prone to overfitting than neural networks. This can be important for problems where the data is not perfectly clean or where there is a small amount of data.
    - Disadvantages: It requires some conditions, such as independence of errors and linearity in the logit for continuous variables. Those conditions make less flexible other models, such as neural networks. Logistic regression may not be accurate if the sample size is too small.

2. **Decision Trees**:

    - Advantages: Decision trees are non-parametric. This means that Decision trees do not have any assumption on the data distribution. It is effective in capturing non-linear relationships, which can be difficult

to achieve with some other models such as linear regression. Decision trees are displayed graphically so they are an easy way to explain and understand without requiring statistical knowledge or complex concepts.

- Disadvantages: Sometimes the prediction accuracy level is less than some other regression and classification approaches. Trees can be non-robust. For example, a small change in the data can cause a large change in the prediction. It's computationally expensive on large datasets. It requires a large number of observations to create a stable and accurate model.

3. **Neural Networks**:

   - Advantages: It requires less formal statistical training, and it can model non-linear relationships. This can be important for problems where the relationship between the dependent and independent variables is not linear. It's able to detect all different interactions between the independent variables. It can often achieve higher accuracy than some other models. This is because neural networks can learn more complex relationships between the features and the target variable.
   - Disadvantages: Neural Networks are computationally expensive so they may take a long time to develop. It Requires a large dataset.

Here are some additional things to consider when choosing between various different models:

- The size of the dataset: Logistic regression is a good choice for problems with a small dataset, while neural networks and decision trees are a good choice for problems with a large dataset.
- The complexity of the problem: Logistic regression is a good choice for problems with a simple relationship between the independent variables and the dependent variable, while neural networks are a good choice for problems with a complex relationship.

- The availability of resources: Logistic regression is a computationally efficient algorithm, while neural networks can be computationally expensive.

3.4 Logistic Regression

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable [EM17]. The most common outcome is a binary outcome. Suppose Y is a binary response, say $y \in \{0, 1\}$, then logistic regression is to model the distribution of the random variable Y given X. The value of $\mathrm{P}(Y = 1 \mid X)$, which is called $p(X)$ will be in the interval [0,1]. To satisfy this condition,

$$e^{\beta_0+\beta_1 X} = \frac{p(X)}{1-p(X)} \tag{3.1}$$

By taking logarithm of both sides, we can write the above equation as

$$\text{logit}(P) = \log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{3.2}$$

The ratio $p(X)/[1-p(X)]$ is called the odds.

We can generalize the simple logistic regression as follows:

$$\log\left(\frac{p(\mathbf{X})}{1-p(\mathbf{X})}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{3.3}$$

where $X = (X_1, X_2, \ldots, X_p)$ are p predictors. The above equation can be rewritten as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}} \tag{3.4}$$

Logistic regression will estimate the odds that an existing student's weight increased or not. To fit the logistic regression model, we minimize the Residual Sum of Squares (RSS)

$$RSS = \sum_{i=1}^{n} \left(y_i - f\left(x_i\right)\right)^2 \tag{3.5}$$

Where y_i is the actual response value, and $f(x_i)$ is the estimated predictor value. In this thesis, x_i are the independent variables such as ethnicity, gender, alcohol consumption, fast food consumption, etc.

3.5 Model Selection

Model selection is the process of choosing the best model from among several candidate models depending on specific criteria.

Given candidate models of similar predictive power, the simplest model is most likely to be the best choice [RG00]. Too few variables in the model might cause under-fitting. However, so many variables in a model might lead to overfitting. Therefore, it is essential to choose the most optimal variables for a high-performance model.

There are some model selection criteria for selecting a few models from a large class of candidate models. The Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are the most common model selection criteria.

AIC and BIC have been widely used in the literature for model selection (see, for example, [Has+09], for a comprehensive review). It is well known that AIC and BIC are both penalized-likelihood information criteria. The formula for both are as follows:

$$AIC = -2 \cdot loglik + 2k \tag{3.6}$$

The Bayesian information criterion (BIC) is given by:

$$BIC = -2 \cdot loglik + (logn) \cdot k \tag{3.7}$$

Where k is the number of model parameters. In statistics, step-wise model selection is a method of fitting regression models based on an iterative process of removing or adding independent variables. Usually, this takes the form of a forward or backward selection. The main approaches for step-wise selection are:

- **Forward step-wise selection** (or forward selection), begins with a model that contains no variables (Null Model), then tests the addition of each vari-

able to the model using some criterion, Adding the variable that results in the largest statistically significant improvement in the fit, and repeating this step until none of the variables improves the model to a statistically significant level.

- **Backward step-wise selection** (or backward elimination), it begins with a model that contains all variables (Full Model), testing the deletion of each variable using some criterion, deleting the variable that causes the least statistically significant deterioration in model fit, and repeating this process until no more variables can be deleted without causing a statistically significant loss of fit.

3.6 Neural Networks

Neural networks, also called artificial neural networks (ANN), are models for prediction and classification. ANN have been demonstrated to be a highly effective method of solving nonlinear problems in a variety of disciplines, including engineering and biology [Wu+23; Lu+23; DRG+23].

The concept of neural networks is based on a very flexible combination of independent variables that captures relationships between these variables and between them and the dependent variable. The ANN algorithm is predicated on the assumption that information processing occurs at several neurons, which are linked to one another via connection links, each with a weight that multiplies the signal transmitted. To determine their output signal, neurons apply an activation function to their net input, which is the sum of the weighted input signals.

Other models, such as regression analysis, are less adaptable than neural networks. In linear regression models, for example, there should be a relationship between the dependent and independent variables. Many times, the precise nature of the relationship is unknown. In comparison, in neural networks, we don't required to specify the correct form. The network tries to learn about such relationships from the data. In fact, logistic regression and linear regression models

are special cases of neural networks that have only input and output layers and no hidden layers.

Neural networks consist of three types of nodes (input, hidden, and output) separately for three types of layers. The below Figure 3.2 (Arden Dertat, 2017) is an example of ANN.

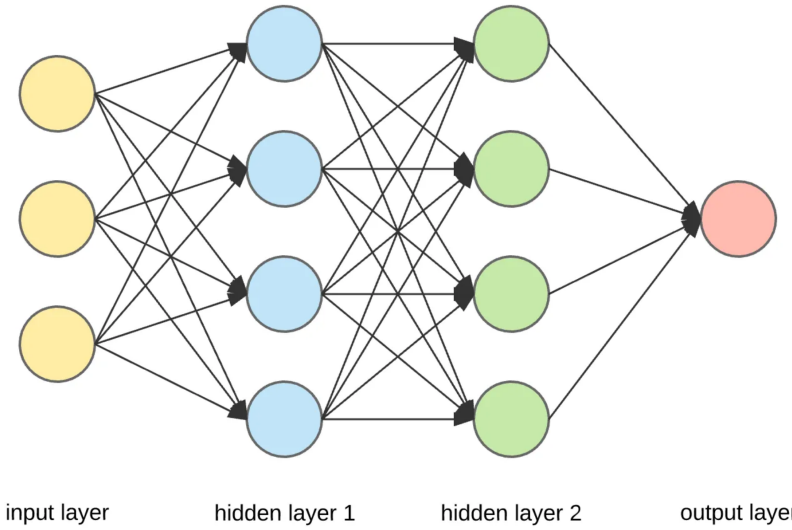


Figure 3.2: Neural network example. Circles present nodes (Neurons).

In the neural networks shown in Figure 3.2, the input layer consists of three nodes, each of which feeds into a node in the hidden layer. Two hidden layers, each with four nodes, receive input from all of the input nodes.

For a set of input values $x_1, x_2, \ldots, x_p$, we compute the output of node j by taking the weighted sum $\theta_j + \sum_{i=1}^{p} w_{ij}x_i$, where $\theta_j, w_{1,j}, \ldots, w_{p,j}$ are weights that are initially set randomly, then adjusted as the network "learns." The constant θ_j controls the level of contribution of node j.

Next, we take a monotone function g of this sum. The function g, is called the activation function. There are several options for the activation function g such as a linear function, an exponential function, and a logistic function. The logistic (It's called sigmoidal) function is the most commonly used activation function in neural

networks. It produces output in the range of 0–1 and introduces a certain amount of nonlinearity into the output.

For example, If g is a logistic activation function, the output of node j can be written as:

$$\text{Output}_j = g\left(\theta_j + \sum_{i=1}^{p} w_{ij}x_i\right) = \frac{1}{1 + e^{-\left(\theta_j + \sum_{i=1}^{p} w_{ij}x_i\right)}} \tag{3.8}$$

3.7 Decision Tree Method

The decision tree method is a popular method for approximating discrete-valued functions that uses a flowchart-like tree structure to represent the learned function. Each internal node in this structure represents a test on an attribute, each branch represents a test result, and each leaf node holds a class label. Figure 3.3 is an example of a decision tree model.

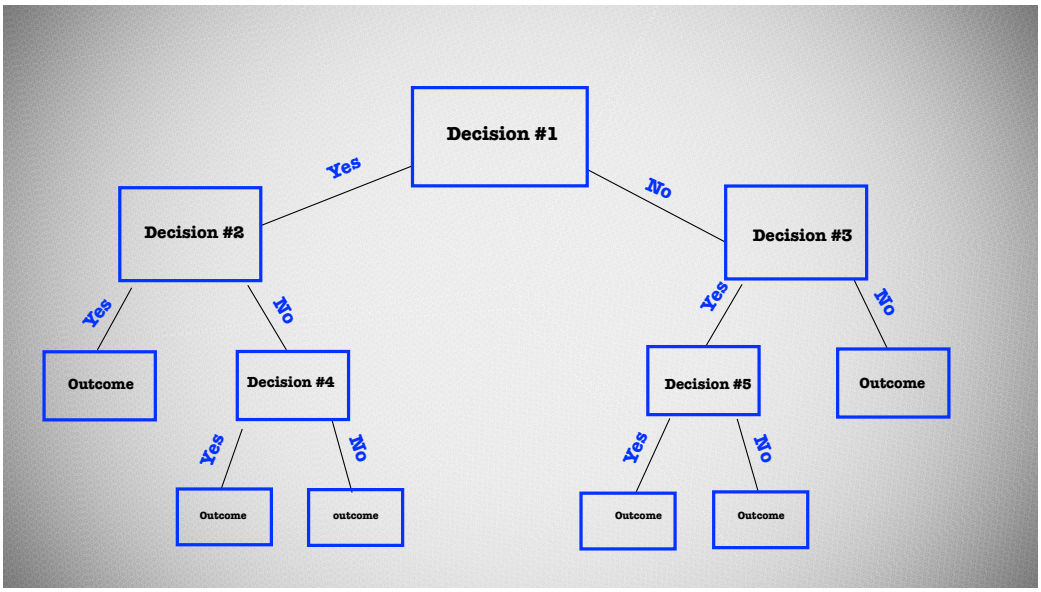


Figure 3.3: Decision tree model Example. The tree is made up of nodes that represent the various features in the dataset, and it starts at the root node, which represents the entire dataset.

Classification trees and regression trees are two types of decision trees. Classification trees determine whether an event occurred or not, similar to a "yes" or "no" outcome. Regression trees forecast continuous values using prior data or information sources.

In this thesis, our goal is to predict whether the student's weight increases or not, so it's a decision tree classifier.

The leaves of the tree in Figure 3.3 represent class labels, while the branches represent feature conjunctions that lead to these class labels. At each step, the method divides the data into subsets based on the most informative feature until a stopping criterion is met. As a result, the algorithm is capable of capturing complex relationships between features and class labels, making it suitable for classification tasks.

Following the construction of the decision tree model, some branches may reflect noise in the training data. By reducing over-fitting, tree pruning attempts to identify and remove unwanted branches from the tree in order to optimize its performance and improve classification accuracy.

3.8 Cross-Validation

Cross-validation is a significant topic when it comes to testing learning models. When using cross-validation to train a model, the data set is split into a training and a testing set.

Cross-validation is essential because it is capable of making an estimate of unseen data on the model being used that is not able to be captured during the training set.

The purpose of testing unseen data while using cross-validation is to understand the dataset in the Machine Learning model. There are different machine learning methods such as Artificial Neural Work, deep learning, logistic Regression, Random Forest, and many more. While using these machine learning methods, not all are the best options for the prediction of the model being used. If the wrong machine

learning method is chosen, it can overstate the prediction of the model.

This method allows to use of cross-validation on any machine learning method and identifies which would be the best model to use that will give the most accurate outcome.

When splitting data into testing and training, it can be split into different increments. The ratios could be random such as 80-20 training-testing, 75-25 training-testing, etc. Below Figure 3.4 ( )illustrates the general idea of Cross-validation of 75-25 percent cases.

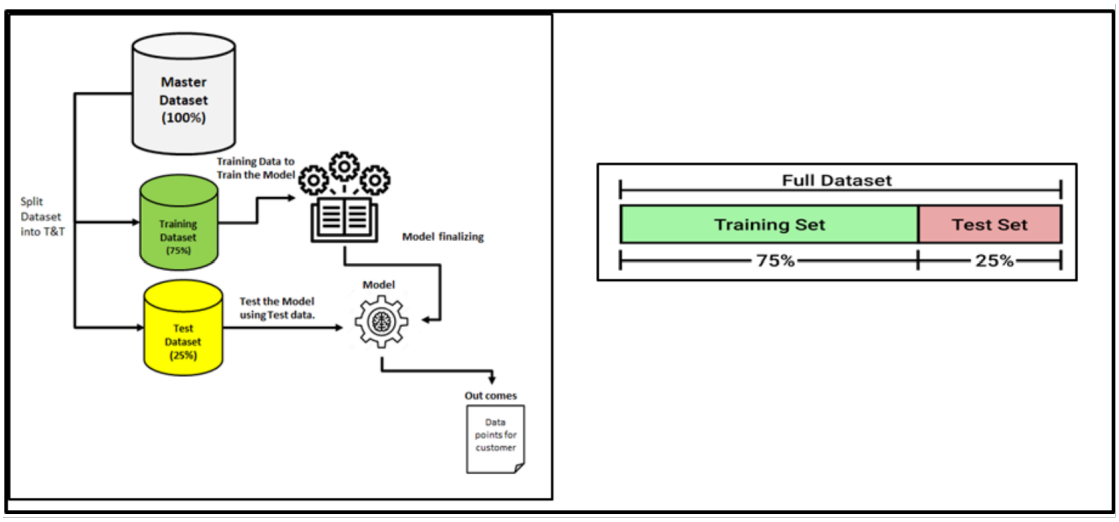


Figure 3.4: Cross-validation Principle: Overview of the machine learning models used to predict the weight status. The architecture includes data splitting, feature selection, model training, and model evaluation. Image designed by the Shanthababu Pandian, 2022.

One technique for using cross-validation is splitting the data into subsets known as folds. k-fold describes the number of subsets that are being split in the data set for training/testing, where k is the number of folds the data set is being split by.

For example, if we use 10-fold for cross-validation (k-10), the data set will be split into 10 folds, where 90 percent is going to training, and 10 percent is going to the validation. Figure 3.5 (Leslie Myint, 2021) provides an illustration of the 10-fold cross-validation mechanism.

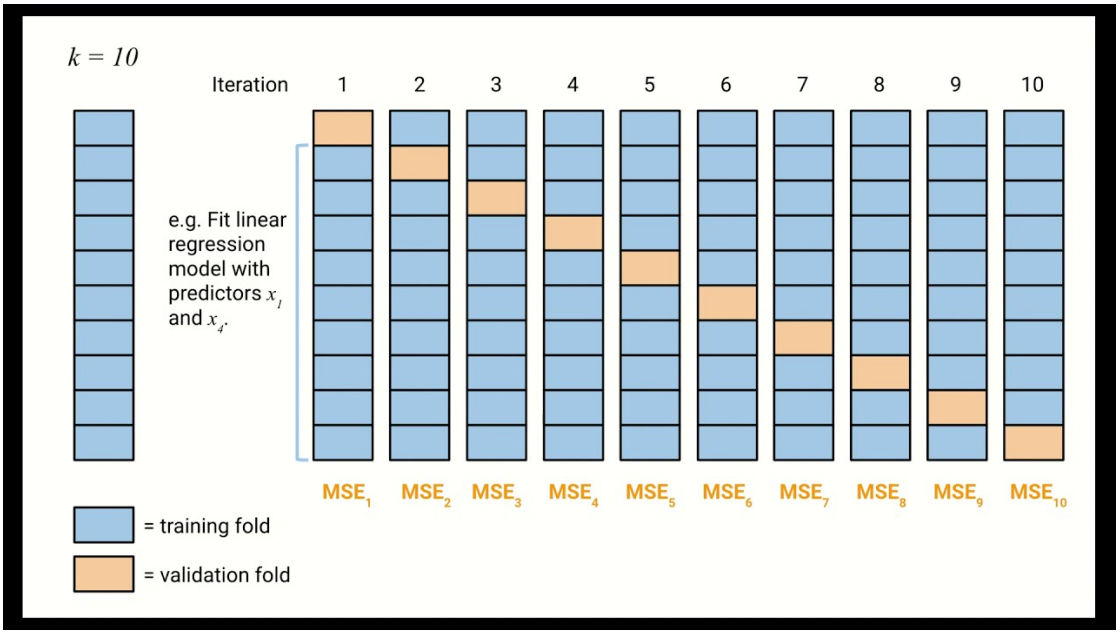


Figure 3.5: A diagram of 10-fold cross-validation. The dataset is randomly divided into five folds of equal size. One of the folds is kept as the validation set in each iteration, while the remaining four folds are used to train the model.

Another technique for using cross-validation is Leave-one-out cross-validation (LOOCV). This technique is only appropriate to use when the dataset size is small or when an accurate estimate of model performance is more important than the computational cost of the model. This technique also k-fold strengthens and it is known to be an expensive computationally expensive procedure to perform.

Lastly, Leave-pair-out cross-validation (LPOCV) is a technique to perform cross-validation. This technique identifies the repetitions of how many examples could be used for testing the model and how many should be left for training the model. When p=1, LOOCV is considered to be a technique for LPOCV.

3.9 Evaluation Metrics

In machine learning algorithms, after applying cross-validation and determining the best machine learning model, it is also crucial to determine if the model cho-

sen will be dependable enough to proceed with analyzing the data. This is when evaluation metrics come in. Evaluation metrics will evaluate the learning machine model chosen and will help identify if there are any modifications needed as well as improve the model as much as possible to further conclude accurate results.

The confusion matrix will help identify the predicted versus the actual outcome of the model. The layout of a confusion matrix is shown in Figure 3.6 (Yunus Yalman, 2022) [Yal+22].

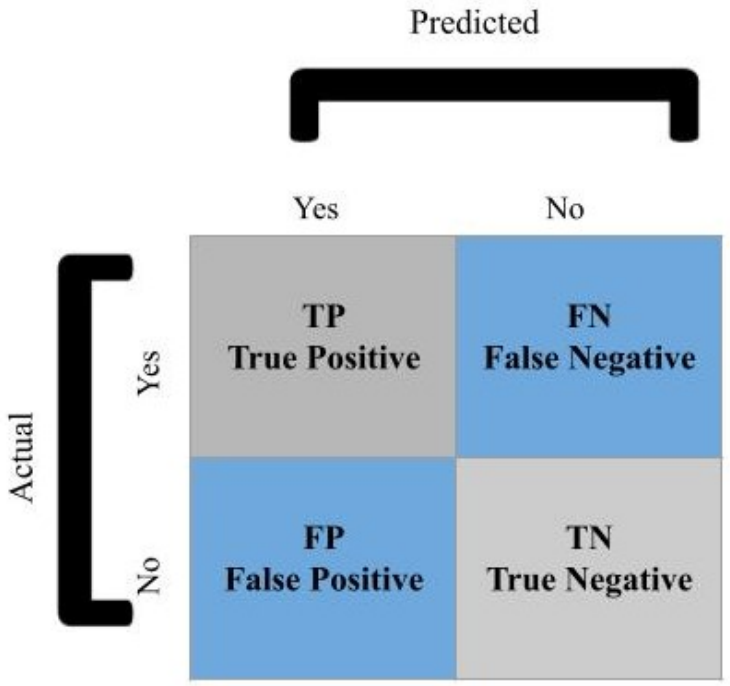


Figure 3.6: An example of a confusion matrix. The matrix displays the model's true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions.

The confusion matrix has four possible outcomes: True Positive (TP) indicates that the model predicted true and the observation was true. True Negative (TN) indicates that the model predicted a false result and that the actual result was also false. False Positive (FP) indicates that the model predicted a true outcome but the

observed result was false. False Negative (FN) denotes that the model predicted a false outcome while the actual observation was true.

Confusion matrices can be used to compute classification model performance metrics. Accuracy, precision, and sensitivity (recall) are the most commonly used metrics. Below is a brief description of each:

1. Accuracy: It's the measure of all true positive and true negative cases divided by the number of all cases. It's given by

$$\frac{TP + TN}{TP + FP + FN + TN} \tag{3.9}$$

2. Precision: Precision is the measure of true positives over the number of total positives predicted by your model. In other words, it allows you to calculate the rate at which your positive predictions are actually positive. The formula is given by:

$$\frac{TP}{TP + FP} \tag{3.10}$$

3. Sensitivity (Recall): is the measure of the true positive over the count of actual positive outcomes. In other words, sensitivity is used to identify the actual true result. The formula is given by:

$$\frac{TP}{TP + FN} \tag{3.11}$$

The confusion matrix can be used to build three metrics: accuracy, precision, and recall. When we want to predict both 0 and 1 correctly and our dataset is balanced enough, we can use accuracy. We use precision when we want our model to predict 1 as accurately as possible, and recall when we want our model to detect as many true 1s as possible. Choosing the right metric for our model can increase its predictive power and provide us with a significant competitive advantage.

Chapter 4

Results and Discussion

4.1 Introduction

In this chapter, we present the results obtained from the data analysis. The analysis consists of various techniques, including descriptive analysis, logistic regression, machine learning analysis, and results comparison. The analysis was conducted using the R program.

The analysis aimed to predict whether the students' weight increased or not, and identify the variables that are associated with freshman 15.

The results obtained will provide a better understanding of the factors that are associated with freshman-15, and will be useful in guiding healthcare practices and policies towards the prevention and management of freshman-15.

4.2 Results and Discussions

4.2.1 Descriptive Analysis

This study below was done to verify if Freshman 15 was true. Freshman 15 is a term known for incoming college students who are known to increase an average of 15 pounds. In this study, 889 students were surveyed, and their outcome showed that the average weight increase is 35.4 pounds which is greater than 15. Please

see below Table 4.1 for brief descriptive statistics.

n	889
Mean	35.4
Median	30
Standard Deviation	26.1

Table 4.1: Description Statistic of "weight increased" Variable

The following Figure 4.1 is a "Weight increased" histogram.

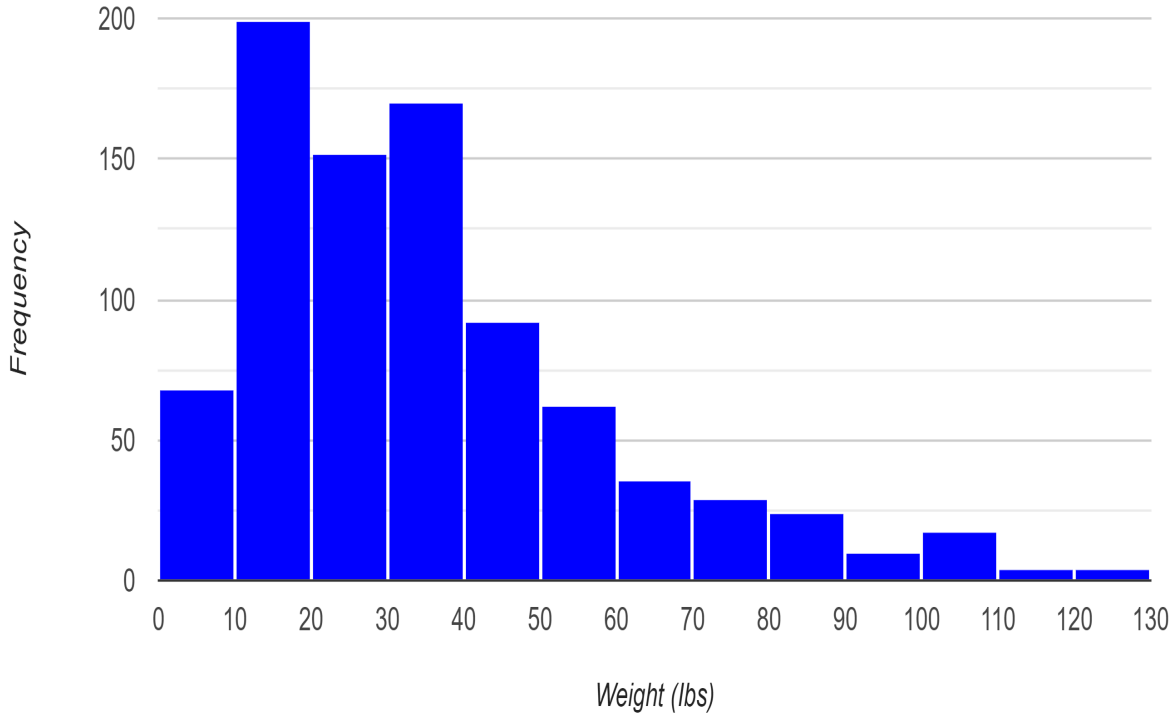


Figure 4.1: Histogram:"Weight increased" Variable

The histogram in Figure 4.1 shows the amount of weight increased. As shown, about 40 students have gained anywhere from 0-10 pounds. Similarly, about 200 students gained anywhere between 10-20 pounds, about 150 students gained anywhere between 20-30 pounds, and so on. Couples of students gained more than 100

pounds. After closely reviewing the graphs, the average still shows to be close to 35.4 pounds which was also found computationally.

Due to the skewness of the histogram, we created a boxplot. The median (Q2) is 30, Q1 is 17, and Q3 is 52. Our dataset contains some outliers.

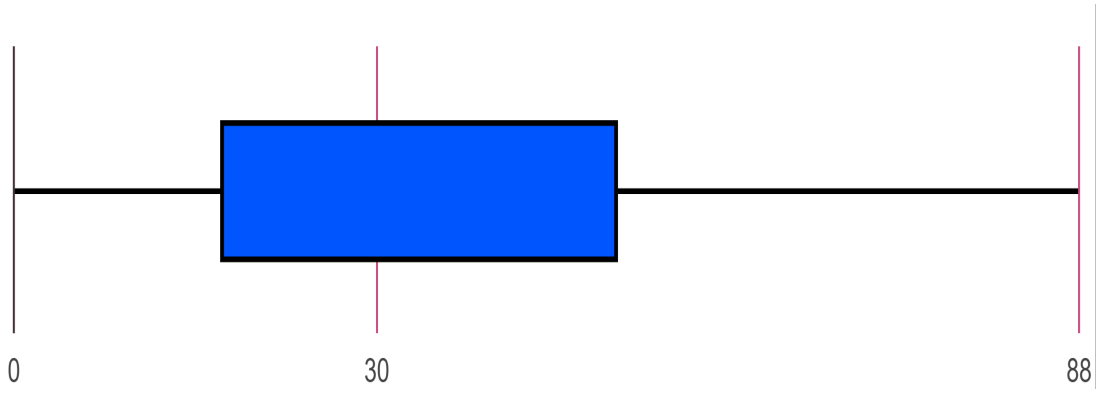


Figure 4.2: Boxplot:"Weight increased" Variable

We asked the university students the following questions: "Compared to high school, does your current weight increase, decrease, or stay the same?"

Table 4.2 shows that out of the 889 students who answered the survey, a significant majority 63.55% reported that their weight increased during their freshman year, 12.37% reported that their weight decreased, and 24.07% of the students reported that their weight didn't change.

Weight Fluctuation		
Weight	Percentage	Number of Students
Increase	63.55%	565
Decrease	12.37%	110
Stay the same	24.07%	214

Table 4.2: Weight Status

The below Figure 4.2 is a bar graph illustrating the data above. It shows a visual of the students whose weight increased during their freshmen year as they

transitioned into college. It also shows the students whose weight decreased and whose weight stayed the same. As the figure shows, The student's weight which increased as incoming freshmen is the highest. Students whose weight stayed the same followed, and the least change for the students who answered the survey were those who decreased in weight.

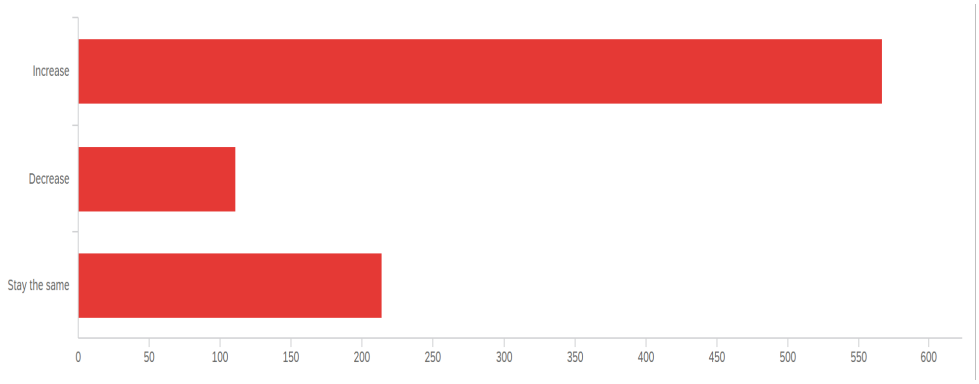


Figure 4.3: Weight Status During Freshman Year

According to various studies such as Hoerr et al (2002) and Sira et al (2010) [Hoe+02; SP10], certain racial and ethnic groups have a significant impact on eating disorders while in college that it is impacting their academic performance. This is the reason for the importance of studying ethnicity as a factor for incoming college student's weight increase.

Table 4.3 below summarizes the weight status by ethnicity. Given that the university has a high number of Hispanics (about 66%), our hypothesis was that ethnicity would have been a significant factor. African Americans have the highest percentage for increasing weight as an incoming college student, but it is still relatively close to the other ethnicities.

As shown in Table 4.3, the percentage for increase, decrease, and stay the same was roughly close to one another. It seems that there is no statistical difference between weight status and ethnicity.

Table 4.3: Weight and Ethnicity

Percentage Weight Fluctuation by Ethnicity			
	Increase	Decrease	Stay the same
Hispanic (595)	62.4% (371)	12.7% (75)	24.9%(149)
African American (45)	77.8%(35)	12.5%(6)	8.3% (4)
White (143)	67.1%(96)	11.9%(17)	21% (30)
Asian (65)	53.9%(35)	12.31%(8)	33.8% (22)
Other (41)	68.3%(28)	9.8%(4)	21.9% (9)
Total: 889	**63.6%**	**12.4%**	**24%**

4.2.2 Statistics and Machine Learning Models

As we described before, our sample size is 889 students. There are 34 independent variables and one dependent variable. The dependent variable in our study is whether or not students gain weight during their first year of college, which is a binary response (yes or no).

In this thesis, we applied three different approaches to model the weight status. Artificial Neural Networks (ANNs), Logistic Regression, and Regression tress methods have been applied. We applied cross-validation and calculated the confusion matrix for each model, then we have a direct comparison of the results.

4.3 Logistic Regression

With the help of logistic regression, we can determine which of the independent variables influenced the most college freshmen students to increase their weight. Given that there is an influence from any of these independent variables, then we can predict how likely it is for an incoming freshman to increase weight in their first year of college.

First, we cleaned our dataset. After excluding two independent variables (parental background and High School activities) from the model, the variance inflation factors (VIF) ranged from 1.31 to 3.90, indicating that they were all less than 5, which is a frequently cited critical value for that diagnostic. [Obr07].

The logistic regression model was then used. The model included all of the independent variables. Then, using stepwise model selection, we identified the most significant (important) predictors.

Coefficients				
Predictor	Estimate	Std. Error	z-value	Pr ($>\|z\|$)
(Intercept)	2.66549	0.55621	4.792	$1.65e$-0.6 $***$
Residence: Off Campus	-0.39810	0.26137	1.523	0.12772
Gender: Female	0.21865	0.17067	-1.281	0.20013
Stress: Yes	0.78455	0.15969	-4.913	$8.97e$-07 $***$
E-cigarette: Yes	0.32079	0.18257	-1.757	0.07890
Marijuana: Yes	0.38771	0.18192	-2.131	0.03307 $*$
Employment: Part-time	-0.48097	0.17590	2.734	0.00625 $**$
Employment: Full-time	-0.95110	0.22121	4.299	$1.71e$-05 $***$
Fast Food: Yes	0.20797	0.09516	-2.184	0.02893 $*$
Alcohol: Yes	0.66295	0.16825	-3.940	$8.14e$-05 $***$

Table 4.4: Logistic Regression Result

Table 4.4 displays the results of the stepwise model selection, along with their p-values. The variables' β coefficients are reported, along with the standard errors.

4.3.1 Significant Variables and Interpretations

Our estimate column in the regression table is the estimate on the intercept, and estimated partial slope on the rest of the Predictors. Given β_0=2.665, β_1=-0.398, β_2=0.219, β_3=0.785 and so on, Raising e to the power of our β 's such as e^{β_0}, e^{β_1}, e^{β_2}, e^{β_3}, we get the chance of the college students gaining weight based on each predictor.

$X = (X_1, X_2, \ldots, X_p)$ are p predictors. It can be rewritten as

$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

The standard error column is the uncertainty associated with the variables. The standard error explains how far off it is expected the estimate to be. Z-value,

also known as Z statistics is the ratio of the estimate column to the Standard error column. The last column, Pr $(>|z|)$ shows the statistical significance of the coefficient's estimates for each predictor. The smaller the number is, the more significant the predictors are in affecting the college students to increase in weight.

The significant variables (p-value < 0.05) are alcohol, fast food, employment, marijuana, and stress. The E-cigarette p-value is 0.079 which is low but not significant. Although the resident and Gender variables are non-significant, it seems that it's necessary to include them in the logistic regression model. A positive sign of the coefficient indicates that the student is more likely to gain weight.

One way to interpret the logistic regression model is by looking at the odds ratio (OR). The odds ratio is the ratio of a specific predictor's outcome. The larger the odd ratios, the more chances it is found with exposure. When the odds ratio exceeds one, it denotes a positive relationship. If the odds ratio is less than one, this indicates a negative relationship.

Stress is one of the most significant variables. We asked the students "Do you catch yourself binge eating when stressed?" and the response was either no or yes. Before we interpret the stress coefficient, note that it's a categorical variable. The reference value is 0 or no, which means no stress. 1 or yes means there is stress. Stress' OR is $e^{0.785} = 2.19$. This indicates that there is a positive relationship. The positive relationship means that as stress "increases", which means changes from No to Yes, the odds of gaining weight increase. The OR 2.19. This indicates that the odds of gaining weight for students who feel stress is 2.19 times that of students who don't feel stress.

Our result is consistent with some other studies. For example, Torres et al. (2007) state that stress can impact students by affecting their eating behaviors [TN07]. Regarding stress eating, weight gain seemed to be dependent on the nutrients taken in which are usually the ones high in sugar. Stress-related eating was found to be significantly associated with obesity, but only in women, not men. Slochower et al (1981) study on female students discussed how stress affected their eating habits during and after their college exams [SKM81]. Obese women had

anxiety and were stress eating more than non-obese females, but the correlation between their anxiety was similar during exams. They indicate that the reason might be because women lean more towards food and men may lean towards alcohol or smoking due to stress.

Similarly, alcohol consumption is one of the most significant variables. We asked the students "Do you consume alcohol?" and the response was either no or yes. Alcohol OR is $e^{0.663} = 1.94$. This indicates that the odds of gaining weight for students who drink alcohol are almost as twice as the of students who don't drink alcohol.

Baum et al. (2017) study found a high correlation between alcohol consumption and weight gain among college freshmen [Bau17]. It states that students are dealing with the pressure of transitioning from high school to college and given that they might also have their parent's pressure on their shoulders, they are more likely to rely on drinking to forget about their worries. Colditz et al. (1991) Also explained the science behind alcohol affecting fluctuation in weight [Col+91]. It is assumed that alcohol consumption may result in weight gain because it is an energy-dense macronutrient that promotes fat storage and increases appetite. Lloyd et al. (2008) study showed that College students who drink more alcohol leads them to have a higher appetite which increases their weight [Llo+08]. About 66% of the students in this study who reported drinking alcohol were unaware of the calorie content of the beverages they were consuming. Overall, correlation analysis revealed that eating habits after drinking were related to changes in their first semester and freshman year.

Employment is one of the most significant variables as well. We asked the students "Are you employed?", and the responses were not employed, part-time, and full-time. There appears to be a negative relationship between employment and weight status. Employment (part-time) OR is $e^{-0.481} = 0.618$. This indicates that the odds of gaining weight for students who work part-time is 0.618 times that of students who are not employed. Employment (full-time) OR is $e^{-0.951} = 0.386$. This indicates that the odds of gaining weight for students who work full-time

is 0.386 times that of students who are not employed. Our result is consistent with some other studies. For example, Ross et al. (1990) state that for most American women, employment improves physical activity which improves physical and psychological well-being and that leads to losing weight [RMG+90]. Based on Marcus et al. (2014), a person who loses their job is more likely to begin smoking, be more stressed, and as mentioned before, both of these factors correlate to weight increase and may lead to obesity [Mar14].

Marijuana is a significant variable. We asked the students " Have you ever used marijuana or cannabis even just one time in your entire lifetime?", and the responses were no or yes. It seems that there is a positive relationship between marijuana usage and weight status. Marijuana OR is $e^{0.388} = 1.474$. This indicates that the odds of gaining weight for students who use marijuana are 1.474 times that of students who are not using marijuana. This is an expected result, as several studies have found a positive relationship between marijuana, alcohol, and the freshman 15 [WLP05; Jon+01].

4.4 Neural Network Analysis

Before we apply the neural network (ANN) algorithm, we normalized our data. Normalization can aid in neural network training because the different features are on a similar scale, which helps to stabilize the gradient descent step, allowing us to use higher learning rates or help models converge faster for a given learning rate.

To fit ANN to our data, we randomly divided the dataset into two parts: 80% for training and 20% for testing. The standard feed-forward neural network algorithm with three layers is then applied: an input layer, a hidden layer, and an output layer. The input layer contained 34 neurons (the independent variables); the hidden layer contained a different number of neurons: 5, 10, 15, 20, 25, 30, 35, and 40; and the output layer contained two neurons (Yes or No). The sigmoid activation function was used. The models were run until the average squared error was less than 0.05.

Table 4.5 lists the eight models that were used to choose the best ANN. Indica-

tors such as precision, sensitivity, and accuracy were compared. When compared to other neural networks, the ANN with 20 neurons in the hidden layer performed better. As a result, the ideal neural network to compare with the logistic regression model was one with 20 neurons in the hidden layer.

Number of Neurons	Precision	Sensitivity	Accuracy
ANN(5)	0.814	0.807	0.810
ANN(10)	0.887	0.876	0.871
ANN(15)	0.901	0.894	0.886
ANN(20)	**0.912**	**0.890**	**0.886**
ANN(25)	0.877	0.869	0.858
ANN(30)	0.839	0.846	0.828
ANN(35)	0.793	0.758	0.763
ANN(40)	0.459	0.467	0.432

Table 4.5: Precision, Sensitivity, and Accuracy for Comparison ANNs Models With Different Number of Neurons

4.5 Decision Trees Analysis

As an explanatory model, the Decision Tree is the most preferred machine learning model [Jeo+23]. A decision tree is a powerful machine learning algorithm that is widely used in different disciplines such as public health, finance, and marketing.

In this study, we used a measure of Gini impurity that is used for the categorical target variable (weight status). The Gini index is defined by:

$$G = \sum_{k=1}^{K} \hat{p}_{mk} \left(1 - \hat{p}_{mk}\right) \tag{4.1}$$

where, $\hat{p}_{mk}$ represents the proportion of training observations in the m th region that are from the k^{th} class.

The Gini index calculates the total variance across the K classes; it is low if all of the $\hat{p}_{mk}$'s are close to zero or one.

4.6 Model Training and Evaluation

In our predictive analysis, we divided our dataset into a training set (80%) and a testing set (20%) using the cross-validation method. This method yielded the best results for our analysis. We employed three algorithms to perform the predictive analysis; logistic regression, neural networks, and decision trees. The algorithm that performed the best on the test set based on overall performance metrics such as test accuracy, sensitivity (recall), and precision would be selected.

We first ran the algorithms on the original dataset and recorded their performance. We then applied the cross-validation approach, allowing the models to learn more from the training data, and recorded their performance again. Overall, our approach aimed to optimize the performance of our predictive models by selecting the best algorithm and incorporating techniques to improve the quality of the training data. By doing so, we were able to make more positive predictions for the dataset.

The confusion matrices in Figure 4.3 provided a detailed breakdown of the performance of the three models using the original dataset and the cross-validation data as well. For each model, they show the number of true positives, true negatives, false positives, and false negatives. These metrics are critical in determining the model's accuracy, precision, and sensitivity (recall). A visual representation of the confusion matrix can help in identifying which models are better at identifying the weight status, and which models may require further tuning to improve their performance. In addition, Table 4.6 provides a detailed overview of the performance of the various models on the original and cross-validation datasets.

The model using the neural network model with 20 neurons (ANN 20) reported an accuracy of 88.6%. The recall is 89%, and the precision is 91.2%. All are higher than those of the other two models. The results of the k-fold cross-validation model were nearly identical to those of the confusion matrix report model.

	ML Algorithm	Precision	Sensitivity (Recall)	Accuracy
Original Dataset	Logistic Regression	.904	.872	.833
	Decision tree	.879	.852	.800
	Neural Network	.912	.89	.886
Cross Validation	Logistic Regression	.908	.870	.835
	Decision tree	.884	.847	.799
	Neural Network	.913	.888	.85

Table 4.6: Evaluation of Model Performance for Weight Status Prediction

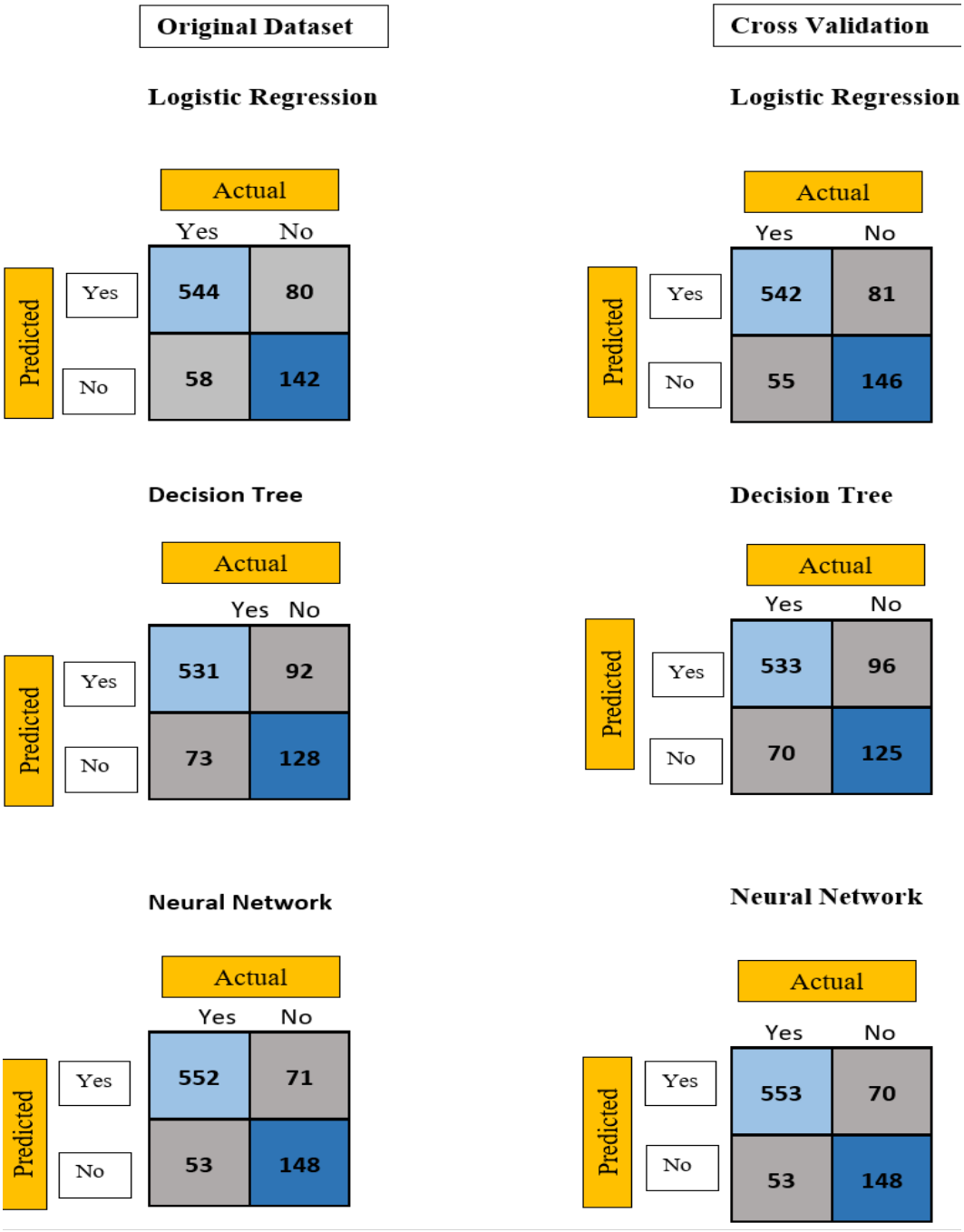


Figure 4.4: Confusion matrices for the evaluated model on the original dataset and using the cross-validation training dataset. The diagonal elements represent the number of correctly classified weight status cases, while the off-diagonal elements represent weight status cases that were misclassified. A 0.5 cut-off value was used.

Chapter 5

Conclusion

This survey was conducted to verify if the term "Freshman 15" is a myth or an occurring event. As we have stated, freshman 15 is a term used to identify incoming college students increasing an average of fifteen pounds in their first year in college.

Despite the fact that the study's goal is to see if students gain an average of 15 pounds in their first years of college, the data and graphs show that students gain an average of 36 pounds in this study. Because the data was collected in 2023, it is assumed that some of the students who participated in the survey were in their first year of college during the COVID-19 pandemic in 2020. Based on the survey, it is unclear how many students from the study were freshmen during the pandemic, but the chances of it affecting the current data are more likely. This could be the reason why the average was double what was expected to verify the term "Freshman 15"

In this study, we proposed three methodologies for studying freshman 15 factors and classifying the weight-gaining status. We used cross-validation to divide our dataset into a training set (80%) and a testing set (20%). Then, for each of the three algorithms (logistic regression, neural networks, and decision trees), we generated two confusion matrices. A confusion matrix for the original dataset and a second matrix for the cross-validation data. We calculate the accuracy, sensitivity, and precision of each model using the confusion matrix.

According to the findings of this study, an ANN with 20 neurons in the hidden

layers outperformed the other eight ANNs. The ANN model in the study was more accurate in predicting weight status and had higher overall performance than the logistic regression and decision tree models. As a result of the findings, artificial neural network (ANN) methods are appropriate for the classification of a weight status-dependent variable. This is not a surprising result. Actually, logistic regression can be thought of as a one-layer neural network. The logistic regression model follows the ANN. The least accurate algorithm is the decision trees model. It has the lowest precision, sensitivity, and accuracy.

Our study concludes that the significant variables (p-value < 0.05) are stress, alcohol, employment, and marijuana, which all lead to bad eating habits and weight increase for first-year students. For example, based on the logistic regression model, the OR for alcohol identifies that the chances of the student increasing weight are 1.94 greater if they drink alcohol than the ones who do not. Consuming marijuana is also 1.74 more likely to lead to weight increase than for those who do not consume marijuana. All these factors were caused by stress, working more hours, and adapting to a new independent environment.

Future studies could consider using a more extensive and diverse dataset to improve the generalizability of the findings. For example, even though this university's population serves a higher number of Hispanic students, based on our logistic regression model, ethnicity was not a significant variable for freshman 15. However, in some other studies, it is a significant factor [GL11; Hof+06]. Finally, future research could investigate using machine learning models to improve the model's accuracy.

Appendix A

Survey

1. Which college is your primary major in? **(College)**
   (a) College of Arts and Letters (CAL)
   (b) College of Education (CE)
   (c) College of Natural Sciences(CNS)
   (d) College of Social and Behavioral Sciences (CSBS)
   (e) College of Business and Public Administration (JHB)
   (f) College of Extended and Global Education (CEGE)
2. Are you a full-time or part-time student? **(Status)**
   (a) Full-time
   (b) Part-time
3. How many school credits are you taking this semester? (Please enter a number) **(Credits)**
4. Do you live on-campus or off-campus? **(Residence)**
   (a) On-Campus
   (b) Off-Campus

5. Do you live in a rural or urban area? **(Area)**
   (a) Rural
   (b) Urban
6. What sex were you assigned at birth? **(Sex)**
   (a) Male
   (b) Female
7. What is your Sexual orientation (Select all that apply)? **(Sex.Orientation)**
   (a) Heterosexuality (Heterosexuality is romantic attraction, sexual attraction or sexual behavior between people of the opposite sex)
   (b) Lesbian/gay
   (c) Bisexual
   (d) Others
8. What is your race/ethnicity? **(Race)**
   (a) Asian/ Pacific Islander
   (b) Black/ African American
   (c) Hispanic/ Latino/ Latina
   (d) Middle Eastern/ Indian
   (e) White/ Caucasian
   (f) Not Listed
9. What is your weight (in pounds)? **(Weight)**
10. Are you in any relationship? **(Relationship)**
    (a) Single
    (b) Married

(c) Living with a partner

11. Are you a first-generation college student? **(First.Generation)**

    (a) Yes

    (b) No

12. What is your current overall GPA? **(GPA)**

13. Are you employed? **(Employment)**

    (a) Not employed

    (b) Part-time

    (c) Employed full-time

14. What is your annual self income? **(Income)**

    (a) Less than 5,000

    (b) 5,001 - 15,000

    (c) 15,001 - 30,000

    (d) 30,001 - 50,000

    (e) 50,001 - 70,000

    (f) More than 70,000

15. What is your parental income? **(Parental.Income)**

    (a) Less than 25,000

    (b) 25,001 - 50,000

    (c) 50,001 - 70,000

    (d) 70,001 - 100,000

    (e) 100,001 - 200,000

    (f) More than 201,000

(g) I don't know

16. What is the highest degree or level of education completed by any of your parents/guardians? **(Parent.Degree)**

    (a) Less than high school

    (b) Some high school

    (c) High school diploma / GED

    (d) Associate's degree

    (e) Bachelor's degree

    (f) Master's degree

    (g) Ph.D. or other doctoral degrees

17. Did you play sports in High school? **(Sport.HS)**

    (a) No

    (b) Yes

18. Did you have a job in High School? **(Job.HS)**

    (a) No

    (b) Yes

19. What was your weight in High School? **(Weight.HS)**

20. Compared to high school, does your current weight increase, decrease, or stay the same? **(Weight.Status)**

    (a) Increase

    (b) Decrease

    (c) Stay the same

21. If your answer to the previous question (increases). Approximately, how many pounds?

22. How many hours out of the day were you active in HS? (Active meaning playing a sport, working, walking/running) **(Active.HS)**

    (a) Less than one hour

    (b) 1-2 hours

    (c) 3-5 hours

    (d) 6+ hours

23. Did a parent/ guardian cook food for you while in HS? **(Parent.Cook)**

    (a) No

    (b) Yes, a few days per week

    (c) Yes, at least once a day

    (d) Yes, 2 meals a day

24. How often are you active daily (Active meaning playing a sport, working, walking/running)? **(Active)**

    (a) Less than one hour

    (b) 1-2 hours

    (c) 3-5 hours

    (d) 6+ hours

25. How often do you exercise a week? **(Exercise)**

    (a) I do not exercise

    (b) 1-3 days

    (c) 3-5 days

    (d) 5+ days

26. How often do you eat fast food? **(Fast.Food)**

(a) Once a day

(b) 2-3 times a day

(c) Average of 3 times a week

(d) I don't eat fast food

27. Do you use the vending machines regularly to purchase snacks? **(Vending.Machine)**

(a) Yes

(b) No

28. How often do you eat at school restaurants? **(School.Restaurant)**

(a) 2-3 times a week

(b) Once a month

(c) Daily

29. Do you consider yourself depressed? **(Depression)**

(a) No

(b) Yes

30. Do you catch yourself binge eating when stressed? **(Stress)**

(a) No

(b) Yes

31. How often do you watch television? **(TV)**

(a) 1 hour a day

(b) 2-3 hours a day

(c) 4+ hours a day

(d) I hardly ever watch Tv

32. How often do you play video games? **(Video.Games)**

    (a) 1 hour a day

    (b) 2-3 hours a day

    (c) 4+ hours a day

    (d) I do not play video games

33. How many hours do you sleep each night? **(Sleeping)**

    (a) less than 6

    (b) 6-8 hours

    (c) 8 hours or more

34. Do you consume alcohol? **(Alcohol)**

    (a) No

    (b) Yes

35. Have you ever used marijuana or cannabis even just one time in your entire lifetime? **(Marijuana)**

    (a) No

    (b) Yes

IRB-FY2023-245 - Initial: IRB Admin./Exempt Review Determination Letter

do-not-reply@cayuse.com <do-not-reply@cayuse.com>
Mon 2023-03-06 12:12 PM
To:Hani Aldirawi <Hani.Aldirawi@csusb.edu>

CALIFORNIA STATE UNIVERSITY
SAN BERNARDINO

March 6, 2023

CSUSB INSTITUTIONAL REVIEW BOARD
Administrative/Exempt Review Determination
Status: Determined Exempt
IRB-FY2023-245

Prof. Hani Aldirawi
CNS - Mathematics
California State University, San Bernardino
5500 University Parkway
San Bernardino, California 92407

Dear Prof. Hani Aldirawi :

Your application to use human subjects, titled "Health and Wellness for College Students" has been reviewed and determined exempt by the Chair of the Institutional Review Board (IRB) of CSU, San Bernardino. An exempt determination means your study had met the federal requirements for exempt status under 45 CFR 46.104. The CSUSB IRB has weighed the risks and benefits of the study to ensure the protection of human participants.

This approval notice does not replace any departmental or additional campus approvals which may be required including access to CSUSB campus facilities and affiliate campuses. Investigators should consider the changing COVID-19 circumstances based on current CDC, California Department of Public Health, and campus guidance and submit appropriate protocol modifications to the IRB as needed. CSUSB campus and affiliate health screenings should be completed for all campus human research related activities. Human research activities conducted at off-campus sites should follow CDC, California Department of Public Health, and local guidance. See CSUSB's COVID-19 Prevention Plan for more information regarding campus requirements.

You are required to notify the IRB of the following as mandated by the Office of Human Research Protections (OHRP) federal regulations 45 CFR 46 and CSUSB IRB policy. The forms (modification, renewal, unanticipated/adverse event, study closure) are located in the Cayuse IRB System with instructions provided on the IRB Applications, Forms, and Submission webpage. Failure to notify the

Figure A.1: IRB Approval Letter

Bibliography

[SKM81] Joyce Slochower, Sharon P Kaplan, and Lisa Mann. "The effects of life stress and weight on mood and eating". In: *Appetite* 2.2 (1981), pp. 115–125.

[RMG+90] Catherine E Ross, John Mirowsky, Karen Goldsteen, et al. "The impact of the family on health: The decade in review". In: *Journal of Marriage and the Family* 52.4 (1990), pp. 1059–1078.

[Col+91] Graham A Colditz et al. "Alcohol intake in relation to diet and obesity in women and men". In: *The American journal of clinical nutrition* 54.1 (1991), pp. 49–55.

[Wec+95] Henry Wechsler et al. "Correlates of college student binge drinking." In: *American journal of public health* 85.7 (1995), pp. 921–926.

[RG00] Carl Rasmussen and Zoubin Ghahramani. "Occam's razor". In: *Advances in neural information processing systems* 13 (2000).

[Wec+00] Henry Wechsler et al. "Environmental correlates of underage alcohol use and related problems of college students". In: *American journal of preventive medicine* 19.1 (2000), pp. 24–29.

[Jon+01] Sherry Everett Jones et al. "Binge drinking among undergraduate college students in the United States: Implications for other substance use". In: *Journal of American College Health* 50.1 (2001), pp. 33–38.

[Hoe+02] Sharon L Hoerr et al. “Risk for disordered eating relates to both gender and ethnicity for college students”. In: *Journal of the American College of Nutrition* 21.4 (2002), pp. 307–314.

[WLP05] Helene R White, Erich W Labouvie, and Vasiliki Papadaratsakis. “Changes in substance use during the transition to adulthood: A comparison of college students and their noncollege age peers”. In: *Journal of Drug Issues* 35.2 (2005), pp. 281–306.

[Hof+06] Daniel J Hoffman et al. “Changes in body weight and fat mass of men and women in the first year of college: A study of the” freshman 15””. In: *Journal of American College Health* 55.1 (2006), pp. 41–46.

[Hud+07] James I Hudson et al. “The prevalence and correlates of eating disorders in the National Comorbidity Survey Replication”. In: *Biological psychiatry* 61.3 (2007), pp. 348–358.

[Obr07] Robert M O’brien. “A caution regarding rules of thumb for variance inflation factors”. In: *Quality & quantity* 41 (2007), pp. 673–690.

[TN07] Susan J Torres and Caryl A Nowson. “Relationship between stress, eating behavior, and obesity”. In: *Nutrition* 23.11-12 (2007), pp. 887–894.

[Gol+08] Andrea B Goldschmidt et al. “Subtyping children and adolescents with loss of control eating by negative affect and dietary restraint”. In: *Behaviour research and therapy* 46.7 (2008), pp. 777–787.

[Llo+08] Elizabeth E Lloyd-Richardson et al. “The relationship between alcohol use, eating habits and weight change in college freshmen”. In: *Eating behaviors* 9.4 (2008), pp. 504–508.

[MAK08] Nicole L Mihalopoulos, Peggy Auinger, and Jonathan D Klein. “The Freshman 15: is it real?” In: *Journal of American College Health* 56.5 (2008), pp. 531–534.

[Nel+08] Melissa C Nelson et al. “Emerging adulthood and college-aged youth: an overlooked age for weight-related behavior change”. In: *Obesity* 16.10 (2008), p. 2205.

[Has+09] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Vol. 2. Springer, 2009.

[Llo+09] Elizabeth E Lloyd-Richardson et al. “A prospective study of weight gain during the college freshman and sophomore years”. In: *Preventive medicine* 48.3 (2009), pp. 256–261.

[GTM10] Rachel W Gow, Sara E Trace, and Suzanne E Mazzeo. “Preventing weight gain in first year college students: an online intervention to prevent the “freshman fifteen””. In: *Eating behaviors* 11.1 (2010), pp. 33–39.

[SP10] Natalia Sira and Roman Pawlak. “Prevalence of overweight and obesity, and dieting attitudes among Caucasian and African American college students in Eastern North Carolina: a cross-sectional survey”. In: *Nutrition research and practice* 4.1 (2010), pp. 36–42.

[VE10] Rachel A Vella-Zarb and Frank J Elgar. “Predicting the ‘freshman 15’: Environmental and psychological predictors of weight gain in first-year university students”. In: *Health Education Journal* 69.3 (2010), pp. 321–332.

[GL11] Meghan M Gillen and Eva S Lefkowitz. “The ‘freshman 15’: trends and predictors in a sample of multiethnic men and women”. In: *Eating behaviors* 12.4 (2011), pp. 261–266.

[SMB11] Carmen Sayon-Orea, Miguel A Martinez-Gonzalez, and Maira Bes-Rastrollo. “Alcohol consumption and body weight: a systematic review”. In: *Nutrition reviews* 69.8 (2011), pp. 419–431.

[Byr+12] Carol Byrd-Bredbenner et al. “Sweet and salty. An assessment of the snacks and beverages sold in vending machines on US post-secondary institution campuses”. In: *Appetite* 58.3 (2012), pp. 1143–1151.

[AA+13] DSMTF American Psychiatric Association, American Psychiatric Association, et al. *Diagnostic and statistical manual of mental disorders: DSM-5*. Vol. 5. 5. American psychiatric association Washington, DC, 2013.

[Cli13] Aurel Ion Clinciu. "Adaptation and stress for the first year university students". In: *Procedia-social and behavioral sciences* 78 (2013), pp. 718–722.

[CKG13] Elizabeth Culnan, Jacqueline D Kloss, and Michael Grandner. "A prospective study of weight gain associated with chronotype among college freshmen". In: *Chronobiology international* 30.5 (2013), pp. 682–690.

[Sut+13] Erin L Sutfin et al. "Electronic cigarette use by college students". In: *Drug and alcohol dependence* 131.3 (2013), pp. 214–221.

[GBG14] Rachel Grana, Neal Benowitz, and Stanton A Glantz. "E-cigarettes: a scientific review". In: *Circulation* 129.19 (2014), pp. 1972–1986.

[Mar14] Jan Marcus. "Does job loss make you smoke and gain weight?" In: *Economica* 81.324 (2014), pp. 626–648.

[OKe+14] James H O'Keefe et al. "Alcohol and cardiovascular health: the dose makes the poison... or the remedy". In: *Mayo Clinic Proceedings*. Vol. 89. 3. Elsevier. 2014, pp. 382–393.

[Sel+14] Elizabeth Selvin et al. "Trends in prevalence and control of diabetes in the United States, 1988–1994 and 1999–2010". In: *Annals of internal medicine* 160.8 (2014), pp. 517–525.

[BGS15] Jamie S Bodenlos, Kara Gengarelly, and Rachael Smith. "Gender differences in freshmen weight gain". In: *Eating behaviors* 19 (2015), pp. 1–4.

[DSo+15] Malcolm J D'Souza et al. "Effect of gender and lifestyle behaviors on BMI trends in a sample of the first state's undergraduate population". In: *American Journal of Health Sciences* 6.1 (2015), p. 59.

[Def+15] Benedicte Deforche et al. “Changes in weight, physical activity, sedentary behaviour and dietary intake during the transition to higher education: a prospective study”. In: *International Journal of Behavioral Nutrition and Physical Activity* 12 (2015), pp. 1–10.

[VTF15] Claudia Vadeboncoeur, Nicholas Townsend, and Charlie Foster. “A meta-analysis of weight gain in first year university students: is freshman 15 a myth?” In: *BMC obesity* 2.1 (2015), pp. 1–9.

[Pen+16] Felicity J Pendergast et al. “Correlates of meal skipping in young adults: a systematic review”. In: *International Journal of Behavioral Nutrition and Physical Activity* 13.1 (2016), pp. 1–15.

[Bau17] Charles L Baum. “The effects of college on weight: examining the “freshman 15” myth and other effects of college over the life cycle”. In: *Demography* 54.1 (2017), pp. 311–336.

[EM17] Thomas W Edgar and David O Manz. *Research methods for cyber security*. Syngress, 2017.

[BP18] Brooke L Bennett and Pallav Pokhrel. “Weight concerns and use of cigarettes and e-cigarettes among young adults”. In: *International Journal of Environmental Research and Public Health* 15.6 (2018), p. 1084.

[Hai+18] Suzan A Haidar et al. “Stress, anxiety, and weight gain among university and college students: a systematic review”. In: *Journal of the Academy of Nutrition and Dietetics* 118.2 (2018), pp. 261–274.

[Cen+19] Jenny C Censin et al. “Causal relationships between obesity and the leading causes of death in women and men”. In: *PLoS genetics* 15.10 (2019), e1008405.

[Hef+19] Kathryn R Hefner et al. “E-cigarettes, alcohol use, and mental health: Use and perceptions of e-cigarettes among college students, by alcohol use and mental health status”. In: *Addictive behaviors* 91 (2019), pp. 12–20.

[Roy+19] Rajshri Roy et al. “Exploring university food environment and on-campus food purchasing behaviors, preferences, and opinions”. In: *Journal of nutrition education and behavior* 51.7 (2019), pp. 865–875.

[ST19] Balbir Singh and Hissam Tawfik. “A machine learning approach for predicting weight gain risks in young adults”. In: *2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT)*. IEEE. 2019, pp. 231–234.

[Bab+20] Oladapo Babajide et al. “A machine learning approach to short-term body weight prediction in a dietary intervention program”. In: *Computational Science–ICCS 2020: 20th International Conference, Amsterdam, The Netherlands, June 3–5, 2020, Proceedings, Part IV 20*. Springer. 2020, pp. 441–455.

[YH20] Zi Yan and Alexandra Harrington. “Factors that predict weight gain among first-year college students”. In: *Health Education Journal* 79.1 (2020), pp. 94–103.

[WWZ22] Brendan E Walsh, Callon M Williams, and Emily L Zale. “Expectancies for and Pleasure from Simultaneous Alcohol and E-Cigarette Use among Young Adults”. In: *Substance Use & Misuse* 57.14 (2022), pp. 2101–2109.

[Yal+22] Yunus Yalman et al. “Prediction of Voltage Sag Relative Location with Data-Driven Algorithms in Distribution Grid”. In: *Energies* 15.18 (2022), p. 6641.

[DRG+23] Pankaj Dagur, Gourav Rakshit, Manik Ghosh, et al. “Virtual screening of phytochemicals for drug discovery”. In: *Phytochemistry, Computational Tools and Databases in Drug Discovery*. Elsevier, 2023, pp. 149–179.

[Jeo+23] Seungjin Jeon et al. “Machine learning-based obesity classification considering 3D body scanner measurements”. In: *Scientific Reports* 13.1 (2023), p. 3299.

[Lu+23] Mingkun Lu et al. “Artificial intelligence in pharmaceutical sciences”. In: *Engineering* (2023).

[Wu+23] Yanling Wu et al. “A consensual machine-learning-assisted QSAR model for effective bioactivity prediction of xanthine oxidase inhibitors using molecular fingerprints”. In: *Molecular Diversity* (2023), pp. 1–16.

[Hen] Kathrin Hennigan. “E-Cigarette and Eating Disorders: Current Research”. In: ().