Utilizing Machine Learning for Effects of Alcohol on the Study

by

Class Roll:
Bushra Akter (1965)
Zannat Hossain Tamim (1970)
Sabina Yeasmim (1984)

Submitted To :
Md. Mahmudur Rahman
Lecturer
IIT,JU

Institute of Information Technology
Jahangirnagar University
Savar, Dhaka-1342

October 2023

ACKNOWLEDGEMENTS

We feel pleased to have the opportunity to express our heartfelt thanks and gratitude to those who rendered their cooperation in making this report. This thesis is performed under the supervision of Md. Mahmudur Rahman, Lecturer, Institute of Information Technology (IIT), Jahangirnagar University, Savar, Dhaka. During the work, he supplied us with several books, journals, and materials related to the present investigation. Without his help, kind support, and the generous period he has given, we could not have performed the project work successfully in due time. First and foremost, we wish to acknowledge our profound and sincere gratitude to him for his guidance, valuable suggestions, encouragement, and cordial cooperation. We express our gratitude to all other sources that we have found helpful.

ABSTRACT

Excessive alcohol consumption is a major public health problem, profoundly affecting academic performance. Addressing this problem is crucial for promoting healthier lifestyles and supporting educational success. College students are particularly vulnerable to the harmful effects of alcohol, as they are still developing their brains and bodies. These problems can make it difficult for students to succeed in their studies. This project explores the effect of alcohol on academic outcomes. By examining existing literature and collecting and analyzing relevant data, including alcohol consumption patterns, study habits, and grades, this study intends to employ correlational analysis and regression modeling techniques to uncover potential links between these variables. This project aims to uncover potential correlations between alcohol intake and study efficacy. The findings could provide insights into the complex relationship between alcohol consumption and study performance. contribute to informed discussions on student well-being, and suggest strategies for promoting healthier study habits. The results of the project will be used to develop interventions to help students who are struggling with alcohol-related academic problems.

Keywords: Machine learning, Data analysis, Accuracy

LIST OF ABBREVIATIONS

AC	Accuracy
DT	Decision Tree
FP	False Positive
FPR	False Positive Rate
FN	False Negative
FNR	False Negative Rate
KNN	k-Nearest Neighbour
SVM	Support Vector Machine
ML	Machine Learning
N	Negative
NIAAA	National Institute on Alcohol Abuse and Alcoholism
NumPy	Numerical Python
P	Positive
RF	Random Forest
TN	True Negative
TP	True Positive

LIST OF FIGURES

Figure

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

In today's fiercely competitive academic landscape, educational institutions must remain attuned to the multifaceted factors influencing student achievement. One such factor with implications for academic outcomes is alcohol consumption.

In contemporary society, student alcohol consumption remains a prevalent and intricate concern. Students engage in drinking for various motives, ranging from social interactions and gatherings to seeking relaxation. This behavior is intricately intertwined with factors like peer pressure, evolving cultural norms, and the delicate balance between academic stress and leisure pursuits. Yet, this practice is not without challenges, including potential harm to academic performance, health-related risks, and the potential to foster unhealthy coping mechanisms, and these complexities are graphically represented in figure 1.1. In these complexities, comprehending the current landscape of student alcohol consumption assumes paramount importance.

To address this concern, the project proposes an investigation into the "Effects of Alcohol on Study." This project will harness the power of machine learning to delve into the intricate interplay between alcohol consumption and academic achievement.

In this project, we use machine learning to understand alcohol's impact on studying for several distinct benefits. Machine learning excels at detecting intricate patterns in these datasets, revealing subtle correlations missed by traditional methods. It efficiently incorporates multiple variables, capturing the issue's complexity. Machine learning predicts outcomes, aiding scenario assessment and projecting how alcohol habits influence academics, a challenge for other approaches. It ensures unbiased analysis, another reason for selecting a machine learning approach.

The project aims to comprehensively explore the relationship between alcohol consumption and academic performance using machine learning. The primary goal is to leverage advanced data analysis to uncover insights into how alcohol habits influence

Figure 1.1: Effect of alcohol consumption

studying. Objectives include data collection, model development, and prediction of alcohol's impact on academic outcomes. By analyzing patterns and trends in the data, we aim to gain a better understanding of how alcohol consumption affects students' academic performance. This information could be used to develop interventions to help students who are struggling with alcohol abuse and to prevent academic problems from occurring in the first place.

CHAPTER II

LITERATURE REVIEW

Understanding the effects of alcohol consumption on academic performance is a subject of growing importance. Prior research has highlighted the negative impact of alcohol on cognitive functions, memory retention, and overall academic achievement. Several studies have employed traditional statistical analyses to examine the relationship between alcohol consumption and educational outcomes. These studies often rely on self-reported surveys and questionnaires, which might suffer from response bias and inaccuracies in reporting. Moreover, while these studies offer valuable insights, they often need to capture the complex and non-linear interactions between various factors that contribute to the effects of alcohol on studying.

There is a growing body of research on the use of machine learning to predict the effects of alcohol. A study by the National Institute on Alcohol Abuse and Alcoholism (NIAAA) found that alcohol use can impair cognitive function, including memory, attention, and learning. The study also found that alcohol use can lead to academic problems, such as lower grades and increased dropout rates [1] . The studies did not always measure cognitive function in the same way. The studies did not always adjust for other factors that could have influenced the results, such as underlying mental health conditions.

Another study, published in the journal "Alcoholism: Clinical and Experimental Research," found that alcohol use can disrupt sleep patterns, which can also impair cognitive function. The study also found that alcohol use can lead to mood problems, such as anxiety and depression, which can also affect academic performance t[2] . The study was limited to a predominantly white college student sample in the Midwest. The findings may not generalize to other populations due to the small sample size and the missing data, which was not completely random.

The project aims to address the limitations of current approaches by developing a machine learning model that can be trained on a large, high-quality dataset of data

on alcohol use and its effects on learning. The model will be developed using a variety of machine-learning techniques, including supervised learning.

CHAPTER III

BACKGROUND TOOLS AND TECHNOLOGIES

3.1 Overview

This chapter provides a comprehensive overview of the libraries employed to investigate the relationship between alcohol consumption and academic performance. It also serves as an introduction to various well-established models of machine learning, along with an exploration of the terminologies associated with them. The chapter concludes by summarizing key findings and insights derived from the analysis.

3.2 Machine Intelligence Libraries

The next section lists a few of the libraries that were utilized to implement the models.

3.2.1 Seaborn

Seaborn is a powerful and flexible Python library for data visualization. It is built on top of Matplotlib and integrates well with the PyData stack, including Pandas and NumPy. Seaborn makes it easy to create beautiful and informative statistical graphics, even for beginners. One of the key benefits of Seaborn is its built-in themes and color palettes[3]. .These themes and color palettes are designed to produce visually appealing and informative graphics with minimal effort from the user. However, Seaborn also provides a high degree of customization, allowing users to tailor the appearance of their graphics to meet their specific needs. Seaborn also provides high-level functions for creating many common types of plots, such as scatter plots, line plots, bar plots, and heat maps. These functions take care of many of the details of

setting up the plot, such as setting axis labels and titles, and can be customized with many different options.

3.2.2 Pandas

Pandas is an open-source Python library for data analysis and manipulation. It is built on top of the NumPy library and provides a high-level interface for working with structured data. Pandas is widely used by data scientists, analysts, and researchers to perform a variety of tasks, including data cleaning, filtering, aggregation, and visualization. [4] One of the key features of Pandas is its data structures, which are designed to be efficient and easy to use. The most important data structures in Pandas are the Series and DataFrame objects. A series is a one-dimensional data structure that can store any type of data, such as numbers, strings, or dates. A data frame is a two-dimensional data structure that can store different types of data in columns.

3.2.3 Numerical Python

NumPy stands as a formidable Python library designed for numerical computing, offering robust data structures and algorithms tailored for the manipulation of multi-dimensional arrays and matrices[5] . Leveraging the power of the C programming language, NumPy ensures swift and efficient operations on these arrays and matrices. At the core of NumPy's offerings is the n-array, an abbreviation for "n-dimensional array," serving as a versatile container for homogeneous data types like integers or floats. The library encompasses an extensive array of functions facilitating the creation, manipulation, and execution of calculations on arrays.
Some of the key features of NumPy include:
- Multi-dimensional array operations
- Broadcasting
- Linear algebra operations
- Random number generation

3.2.4 Matplotlib

Matplotlib is a Python library for creating static, animated, and interactive visualizations in Python. It is widely used in scientific computing, data analysis, and machine learning for generating high-quality plots, histograms, bar charts, scatter-

plots, and other types of visualizations. Matplotlib provides a wide range of plotting functions and customization options for creating and customizing plots[6] .
Some of the key features of NumPy include:
• Multi-dimensional array operations
• Broadcasting
• Linear algebra operations
• Random number generation

3.2.5 Scikit Learn

Scikit-learn, also known as sklearn, is a free, open-source machine learning library for Python. It is one of the most popular machine learning libraries in the world and is used by researchers and practitioners alike. Scikit-learn provides a wide range of tools for supervised and unsupervised learning, including classification, regression, clustering, dimensionality reduction, and model selection[7] . It also includes various tools for pre-processing and feature selection, cross-validation, and model evaluation. Scikit-learn is known for its simplicity, flexibility, and efficiency. It provides a consistent and easy-to-use interface for building and evaluating machine learning models. It also provides a wide range of algorithms for different types of problems, making it a versatile tool for machine learning.
Some of the key features of NumPy include:
• Multi-dimensional array operations
• Broadcasting
• Linear algebra operations
• Random number generation

3.3 Preprocessing

3.3.1 Data Cleaning

The act of locating and eliminating flaws, inconsistencies, and inaccuracies from a dataset is known as data cleaning. Data cleaning aims to enhance the quality of the data and make it appropriate for modeling or analysis. Data cleaning often involves the following steps.
Handling missing data: Missing data can be imputed (replaced with estimated

values) using various techniques such as mean imputation, median imputation, or regression imputation[8] .

Removing duplicates: Duplicates can occur in datasets due to errors or multiple sources of data. Removing duplicates can improve the accuracy of analysis and modeling.

Data cleaning is an iterative process that involves multiple steps and techniques. The choice of techniques and methods depends on the nature of the problem, the characteristics of the data, and the goals of analysis or modeling.

3.3.2 Feature Selection

A machine learning approach called feature selection is used to pick a subset of important features (or variables) from a wider pool of potential features. By removing unnecessary or redundant information, feature selection aims to increase the precision, interpretability, and effectiveness of machine learning models. Filter methods, wrapper methods, and embedding methods are just a few of the techniques that can be used to pick features[9] . The following are a few of the most popular feature selection methods:

Wrapper methods: Wrapper approaches employ a machine learning algorithm to assess how well a subset of features perform. They evaluate different subsets of features by training and testing the machine learning model iteratively. Wrapper methods can be computationally expensive but can lead to better performance than filter methods.

Feature selection can lead to various benefits like improving model performance, reducing computational cost, and increasing interpretability.

3.3.3 Encoding Categorical Variables

Encoding categorical variables is an important step in preparing data for machine learning models. Categorical variables are variables that represent categories or groups, such as gender, color, or product type. Machine learning algorithms require numerical data as input, so categorical variables need to be encoded into numerical values[10] . There are several ways to encode categorical variables

One-hot encoding: One-hot encoding adds a new binary column for each category, giving it a value of 1, and giving all other columns a value of 0. For example, if the categorical variable is "color" with the categories "red", "green", and "blue", one-hot encoding will create three binary columns with values of 1 or 0 to indicate whether

the color is red, green, or blue.

Ordinal encoding: Ordinal encoding assigns a unique integer value to each category based on its rank or order. For example, if the categorical variable is "size" with the categories "small", "medium", and "large", ordinal encoding will assign the values 1, 2, and 3, respectively, to indicate the relative size of the categories.

Label encoding: Label encoding assigns a unique integer value to each category without any order or ranking. For example, if the categorical variable is "gender" with the categories "male" and "female", label encoding will assign the values 0 and 1 to represent the categories.

The choice of encoding method depends on the nature of the problem and the characteristics of the categorical variables. One-Hot Encoding is commonly used when there is no natural order or ranking among the categories, while Ordinal Encoding is used when there is a natural order or ranking. Label encoding is used when the categories are binary or when the order of the categories does not matter.

3.3.4 Class Imbalance Sampling

Class imbalance is a common issue in machine learning where the distribution of classes in the training dataset is not equal. In many real-world scenarios, one class (the minority class) is significantly less frequent than the other class or classes (the majority class or classes). This can cause machine learning models to be biased towards the majority class, leading to poor performance for the minority class. To address this problem, various techniques, including sampling methods, can be employed. One such technique is class imbalance sampling[11] .

3.3.5 Scaling

Scaling in machine learning refers to the process of normalizing or standardizing the features of a dataset. It is a crucial step in many machine learning algorithms that are sensitive to the scale of input features. When features in a dataset have different scales, some algorithms might give more weight to features with larger magnitudes, leading to biased or incorrect results. Scaling helps in bringing all features to a similar scale, ensuring that the algorithm treats all features equally.

3.4 Machine Learning Models Related to Our Work

This section includes detailed information about machine learning algorithms (Linear Regression, Logistic Regression, SVM, KNN, Decision Tree, Random Forest, etc.)

related to our work.

3.4.1 Linear Regression

Linear regression is a supervised machine learning algorithm that is used to model the relationship between a dependent variable and one or more independent variables. The dependent variable is the variable that we are trying to predict, and the independent variables are the variables that we are using to predict it.

To build a linear regression model, we first need to collect a dataset of input and output variables. Once we have our dataset, we can use a linear regression algorithm to train the model. The training process will find the best-fit line through the data points. This best-fit line will be used to predict the dependent variable for new values of the independent variables[12] .

Some of the key features of NumPy include:

- Multi-dimensional array operations
- Broadcasting
- Linear algebra operations
- Random number generation

3.4.2 Logistic Regression

Logistic regression is a type of generalized linear model that is used to predict the probability of a binary outcome. It is similar to linear regression, but it uses a logistic function to transform the linear output into a probability. This logistic function ensures that the predicted probability is always between 0 and 1.

Logistic regression is a widely used algorithm in machine learning, and it is used for a variety of tasks, such as predicting whether a customer will churn, predicting whether a patient has a disease, and predicting whether a loan will be repaid[13] .

The working procedure of linear regression :

1. Collect a dataset of input and output variables.
2. Prepare the data.
3. Split the data into training and test sets.
4. Choose a logistic regression algorithm.
5. Train the model.
6. Evaluate the model.
7. Deploy the model.

3.4.3 Support Vector Machines

Support vector machines (SVM) are a powerful machine learning algorithm that can be used for both classification and regression tasks. SVMs are particularly well-suited for classification tasks with data that can be separated into linear or non-linear categories. In a binary classification task, SVMs seek to find the hyperplane with the largest margin of separation between the two classes. The margin is the distance between each class's nearest data points to the hyperplane. The wider the margin, the better the hyperplane can generalize to new data points[14] .

The working procedure of linear regression :

1. Collect a dataset of input and output variables.
2. Prepare the data.
3. Split the data into training and test sets.
4. Choose a kernel function.
5. Train the model.
6. Evaluate the model.
7. Deploy the model.

3.4.4 KNN

A straightforward yet efficient non-parametric classification and regression approach is K-Nearest Neighbors (KNN). Using KNN, a new data point is classified by comparing it to its K nearest neighbors in the training data and selecting the majority class label or the KNN mean value as the forecast[15] .

KNN algorithm for classification:

1. Given a new data point x, calculate the distance between x and all the training data points using a distance metric such as Euclidean distance or Manhattan distance.

2. Select the KNN to x based on the calculated distances.

3. Assign the majority class label among the KNN as the predicted class label for x.

KNN algorithm for regression:

1. Given a new data point x, calculate the distance between x and all the training data points using a distance metric.

2. Select the KNN to x based on the calculated distances.

3. Assign the mean value of the target variable among the KNN as the predicted value for x.

The choice of K is a hyperparameter that can be tuned based on the performance on a validation set. A smaller K tends to result in a more flexible model with higher variance but lower bias, while a larger K tends to result in a more rigid model with lower variance but higher bias.

3.4.5 Decision Tree

A decision tree is a supervised machine learning algorithm that can be used for both classification and regression tasks. It is a tree-like structure where each internal node represents a feature in the data, each branch represents a decision rule based on the feature value, and each leaf node represents a prediction[16] .

The working procedure of decision tree :

1. Collect a dataset of input and output variables.
2. Prepare the data.
3. Split the data into training and test sets.
4. Choose a splitting criterion.
5. Build the decision tree.
6. Evaluate the decision tree.
7. Deploy the decision tree.

3.4.6 Random Forest

Random forest is a popular machine learning algorithm that combines multiple decision trees to make predictions. It is a type of ensemble learning algorithm, which means that it combines the predictions of multiple models to produce a more accurate prediction.Random forests work by creating a set of decision trees on randomly selected subsets of the training data. Each decision tree is trained independently of the others, and the predictions of all the trees are then averaged to produce the final prediction[17] .

Random forest algorithm working procedure:

1. Given a training dataset, randomly select a subset of the data and features for each decision tree in the forest.

2. Grow a decision tree on the selected subset using a criterion such as information gain or Gini impurity.

3. Repeat step 1 and 2 to create a forest of decision trees. 4. To make a prediction on a new data point, pass the data point through each decision tree in the forest and obtain the majority vote of their predictions for classification or the mean value of their predictions for regression.

3.5 Other Related Terms to Machine Learning

This section includes information for other words related to machine learning.

3.5.1 Cross-validation

Cross-validation is a technique used in machine learning to evaluate the performance of a model on an independent dataset. The goal of cross-validation is to estimate the generalization performance of a model[18] .

Cross-validation algorithm working procedure:

1. Create training and testing sets from the dataset. The model is trained using the training set, and its effectiveness is assessed using the testing set.

2. Partition the training set into k equal-sized subsets or folds.

3. For each fold, train the model on the remaining k-1 folds and evaluate its performance on the hold-out fold.

4. Repeat the process k times, using a different fold as the hold-out set each time.

5. Calculate the average performance of the model across all k folds to obtain an estimate of its generalization performance.

K-fold cross-validation is the most popular kind of cross-validation, with k often set to 5 or 10. Depending on the dataset and the issue at hand, different cross-validation techniques, such leave-one-out cross-validation and stratified cross-validation, may be utilized.

3.6 Conclusion

Several machine learning libraries (Seaborn, Pandas, NumPy, etc) are described in this chapter in detail. Then different preprocessing techniques that are followed before feature extraction and feed data to model are described . Description of different machine learning models related to our work including KNN, GBC, Random Forest etc is given. This chapter ends by detailing other terminology related to machine learning.

CHAPTER IV

METHODOLOGY

In our project, we recognize the complexity of the relationship between student grade and alcohol consumption, acknowledging factors such as family stability, support, internet usage, and previous academic failures etc to the overall dynamics. Our data collection strategy encompasses these multifaceted variables to gain a comprehensive understanding. The subsequent phases of data preprocessing and model evaluation will consider this holistic dataset, aiming to reveal brief insights beyond just student grades and alcohol consumption. The evaluation process will examine model accuracy to ascertain the predictive prowess in revealing insights into the intricate relationship between student results and alcohol consumption. Figure 4.1 visually outlines the sequential steps in our methodology. It begins with comprehensive data collection, encompassing student records, family stability, internet usage, and past academic performance etc. The subsequent phases include meticulous data preprocessing and model evaluation, ensuring a systematic and thorough exploration of the complex relationship between student academic performance and alcohol consumption.

4.1 Data Collection

In the current situation in Bangladesh, alcohol consumption is a very common and growing threat [19] . Our main objective is to organize the machine learning concepts and try to make predictions about alcohol's impact on academic outcomes. So, primarily, we have collected data (Student Alcohol Consumption) from Kaggle [20]

We take a data set (Student Alcohol Consumption).Our data set includes data of two schools(Gabriel Pereira,Mousinho da Silveira).Data set include 65.2

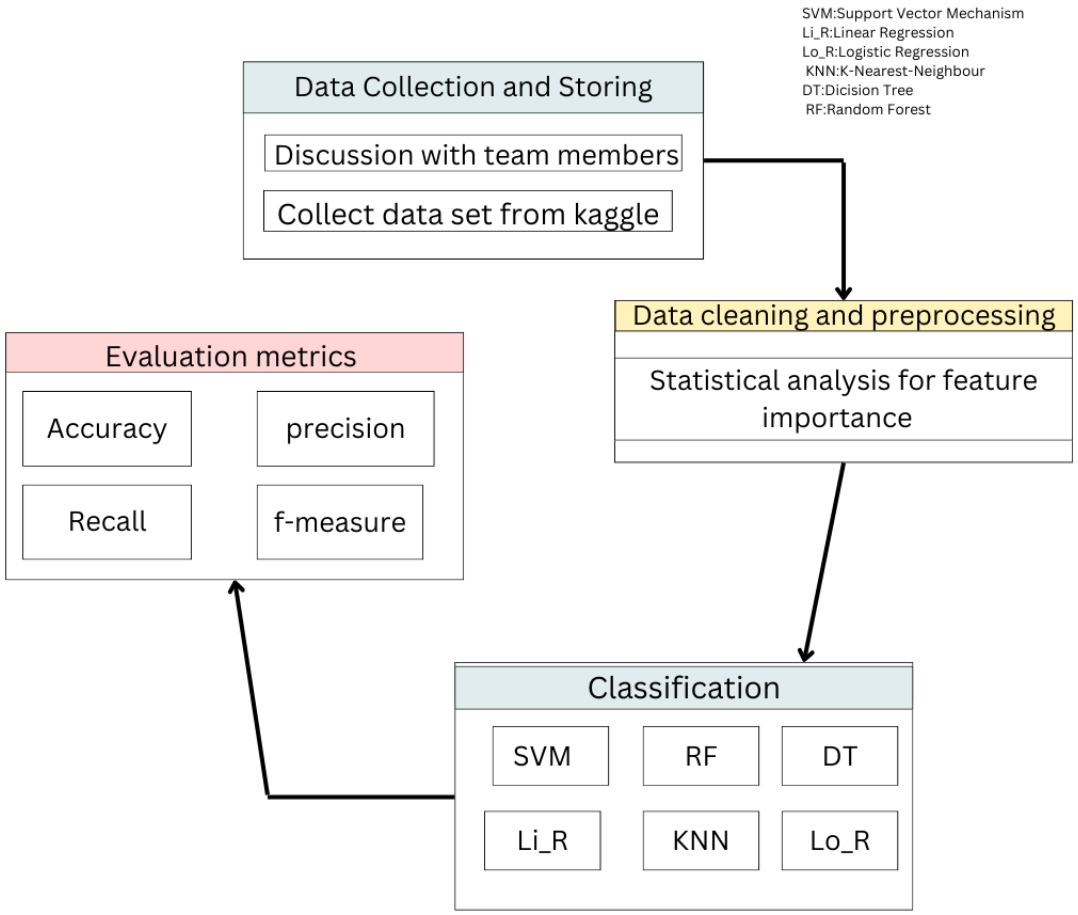

Figure 4.1: Methodology .

```
data and features : (649, 33)

features are: Index(['school', 'sex', 'age', 'address', 'famsize', 'Pstatus', 'Medu', 'Fedu',
       'Mjob', 'Fjob', 'reason', 'guardian', 'traveltime', 'studytime',
       'failures', 'schoolsup', 'famsup', 'paid', 'activities', 'nursery',
       'higher', 'internet', 'romantic', 'famrel', 'freetime', 'goout', 'Dalc',
       'Walc', 'health', 'absences', 'G1', 'G2', 'G3'],
      dtype='object')
```

Figure 4.2: Features of our dataset

4.2 Data Pre-processing

In the data preprocessing phase, our methodology adheres to a structured flowchart depicted in Figure 4.4. Prior to this, we conduct an exploratory analysis, revealing insightful relationships through figures such as a scree plot and graphical represen-

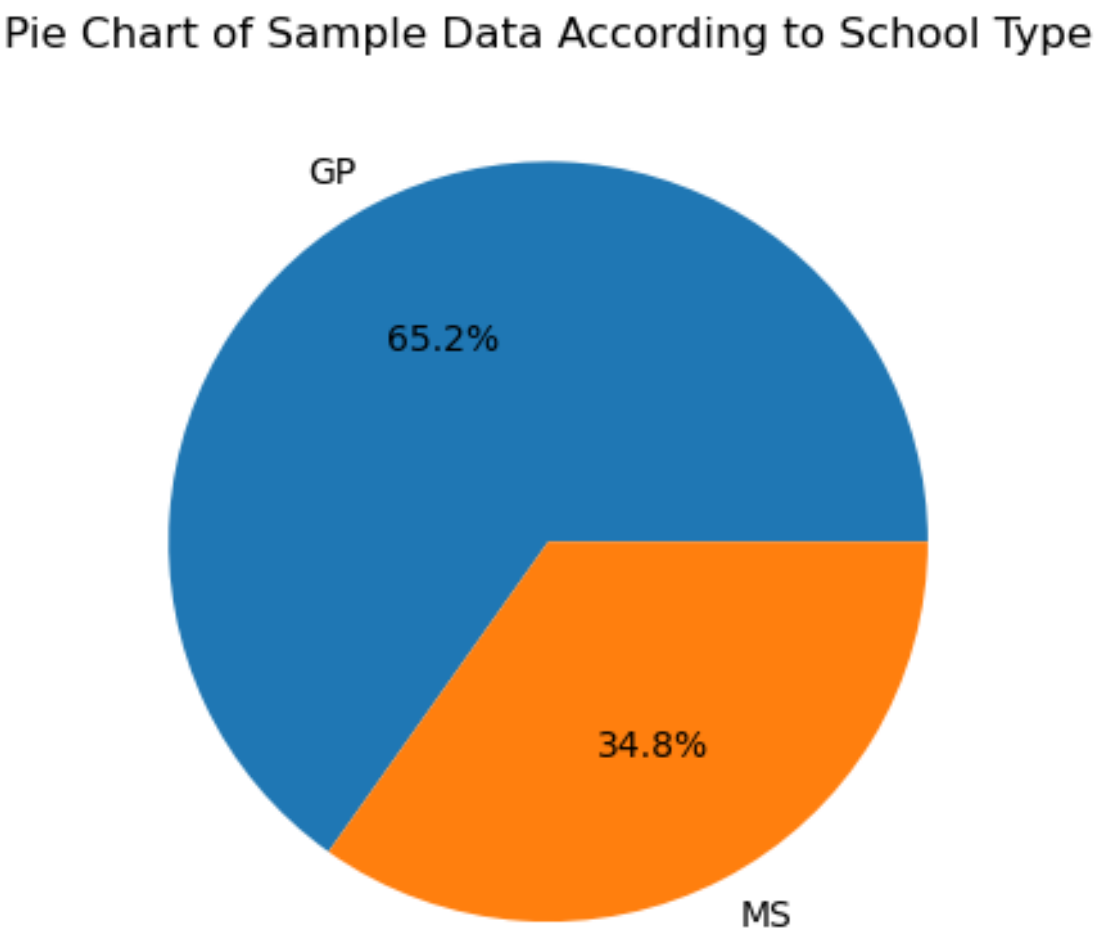

Figure 4.3: Features of our dataset

tations of key features. Noteworthy visualizations include the correlation between weekly alcohol consumption and support from school or home.

Figure 4.5 represents the reciprocal relationship between alcohol consumption (We selected two features, Dalc (weekday consumed alcohol) and Walc (weekend consumed alcohol), and concatenated them into a new feature called weekly alcohol consumption.)and school support, underscoring a notable decrease in alcohol consumption when robust support structures exist. Similarly, Figure 4.5 sheds light on the intricate dynamics between alcohol consumption and support from home, revealing a parallel decline in alcohol consumption when support is present in the familial context.

Expanding on these insights, Figure 4.6 aptly captures the interplay between alcohol consumption and age, highlighting a pronounced impact of teenage years on drinking patterns. Figure 4.7shows the brief relationship between alcohol consumption and family stability, demonstrating that increased family stability correlates with

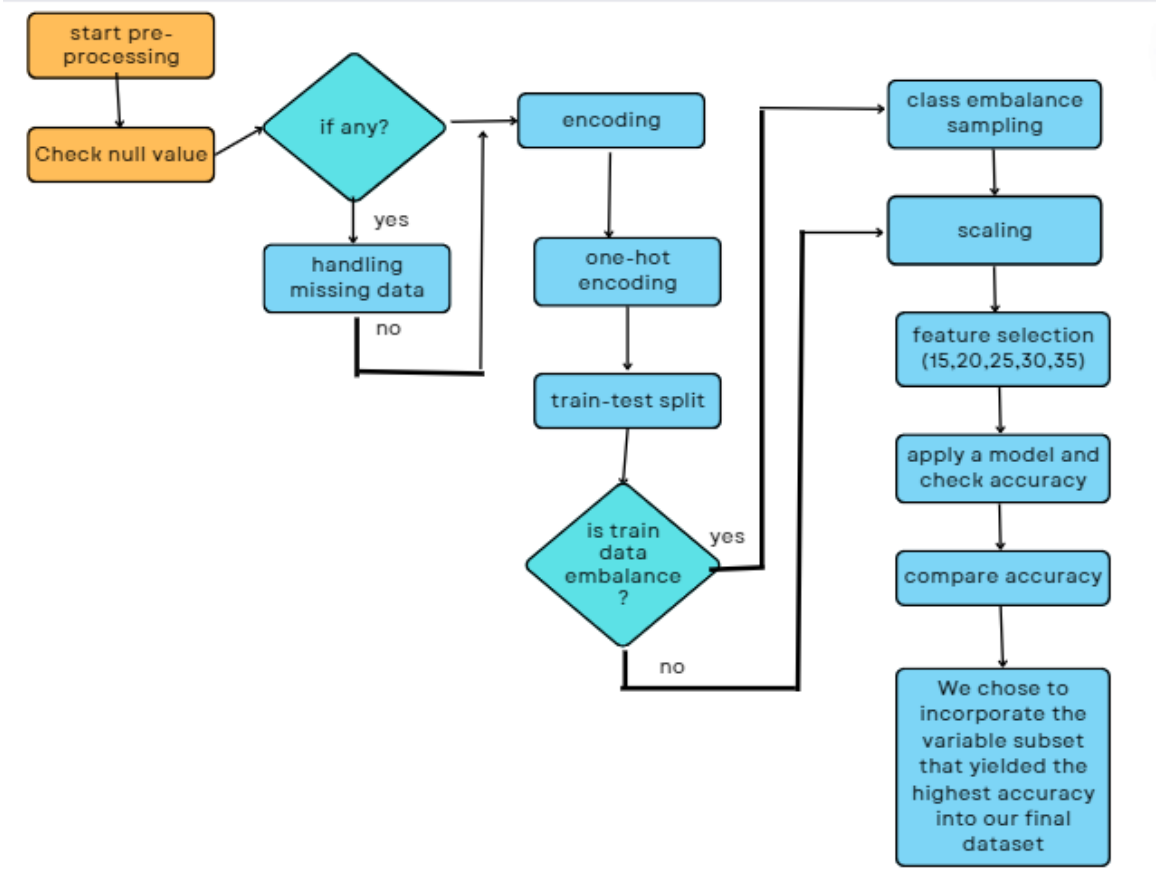

Figure 4.4: Features of our dataset

lower alcohol consumption. In a parallel fashion, Figure 4.8 explores the influence of academic support, revealing a similar trend whereas heightened academic support is associated with reduced alcohol consumption. These visual narratives not only underscore the central role of support structures but also emphasize the nuanced impact of age, family stability, and academic support on alcohol consumption patterns.

Moreover, our analytical approach includes a pivotal categorical classification. Specifically, we categorize weekly alcohol rates within the range of 0-5 as indicative of light drinkers, assigned a corresponding feature value of 0. Besides, rates falling between 5-10 are identified as characteristic of heavy drinkers, assigned a feature value of 1. This classification schema serves as a foundational element in our exploration, enabling a more nuanced understanding of alcohol consumption patterns among study participants.

In Figure 4.9, the revelation that heavy alcohol consumers number 120 students, contrasted with 529 identified as light drinkers, emphasizes the imbalance in class sampling. To address this, Figure 4.10 introduces a scree plot, aiding in determining

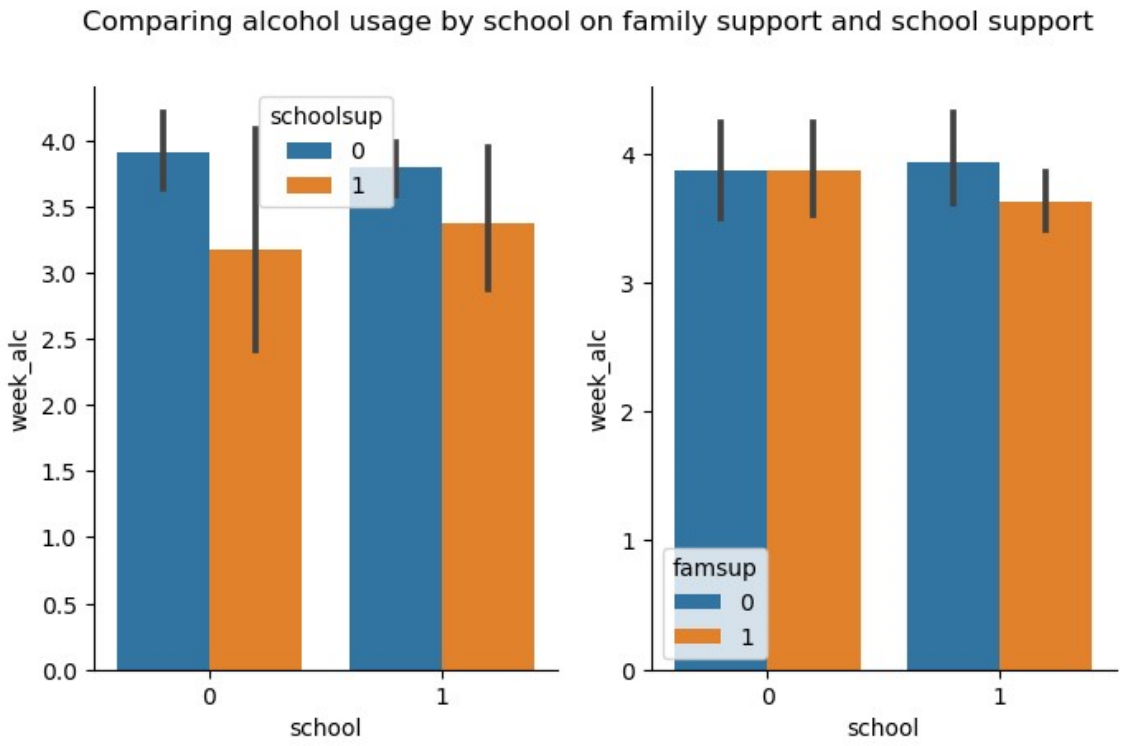

Figure 4.5: Comparing alcohol usage by school on family support and school support

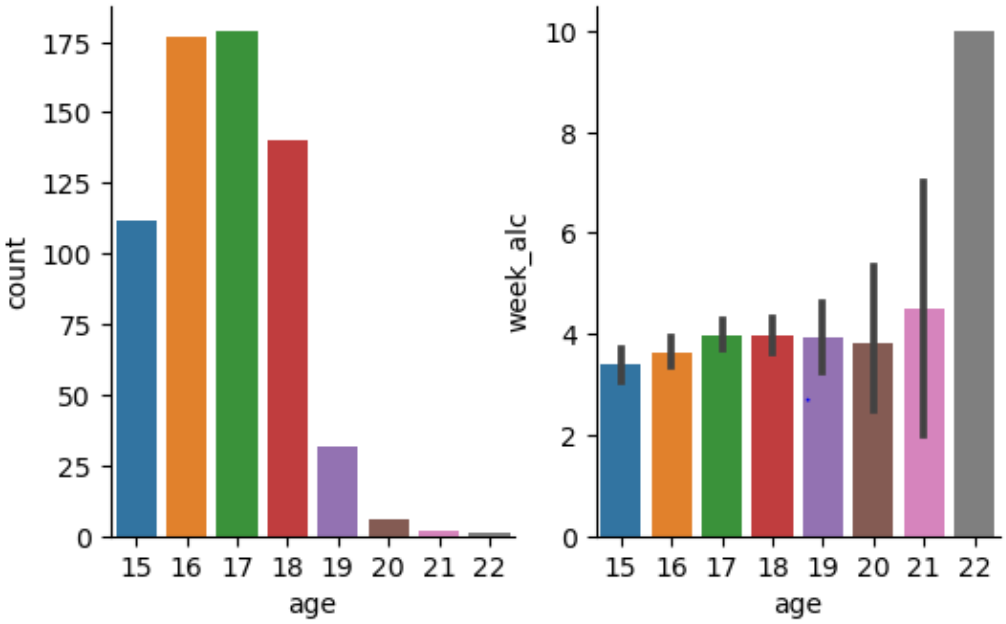

Figure 4.6: Comparing student

optimal class balance adjustments. Meanwhile, Figure 4.11 provides a compelling visual representation linking academic grades (g1, g2, g3) with alcohol classifications (0 for light drinkers, 1 for heavy drinkers). Notably, the data in Figure 4.12 further

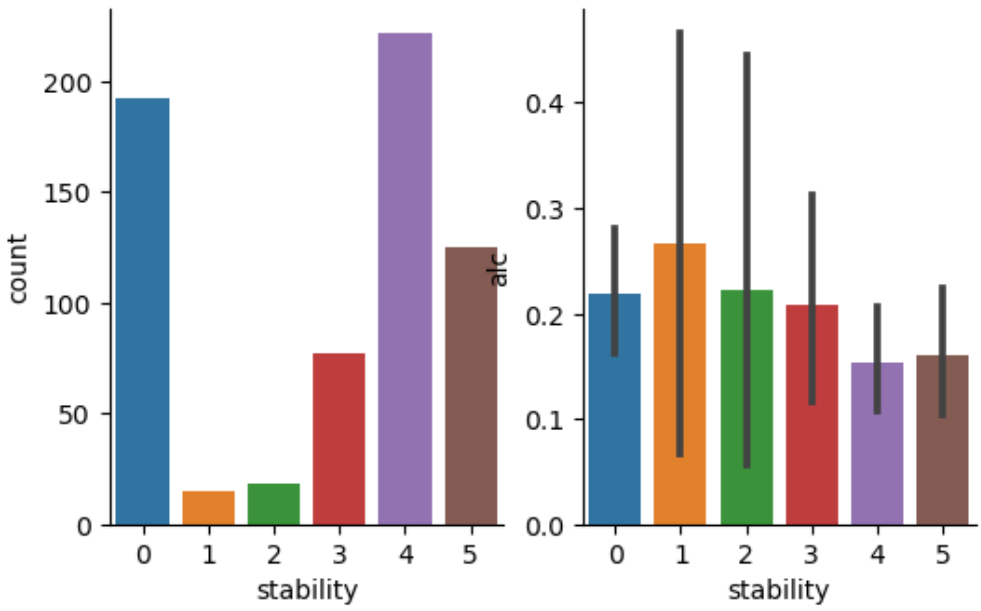

Figure 4.7: Feature quantifying family stability

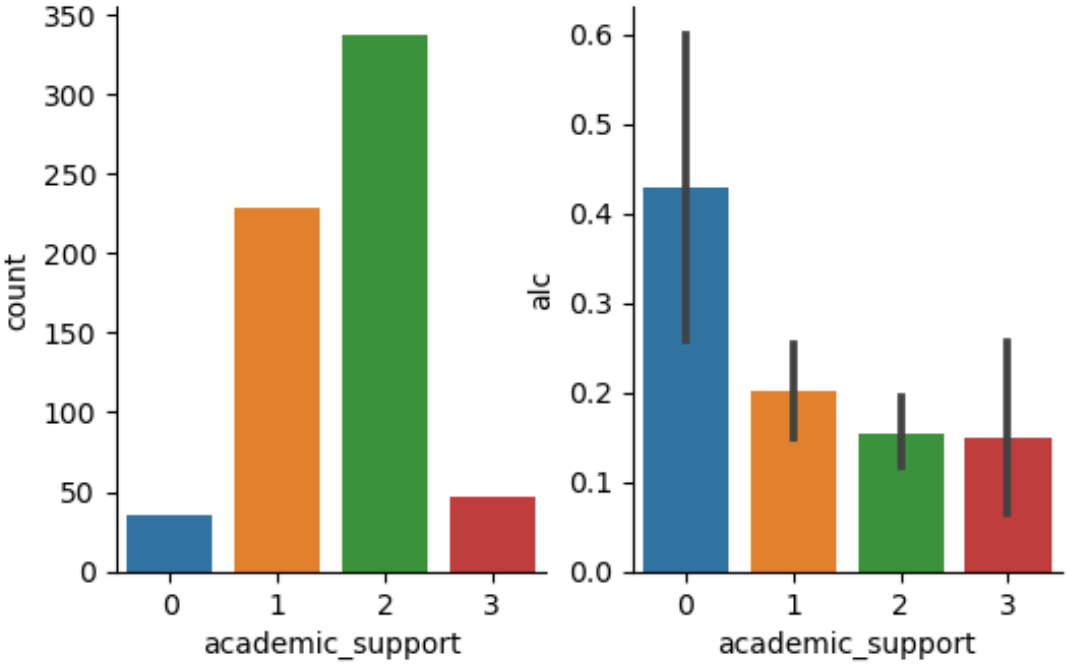

Figure 4.8: Feature quantifying academic support network

elaborate this relationship, depicting the frequency distribution of students based on average grade (x-axis), with orange bars denoting light drinkers and blue representing heavy drinkers. The pattern emerges that higher alcohol consumption (coded as 1) correlates with lower academic grades.

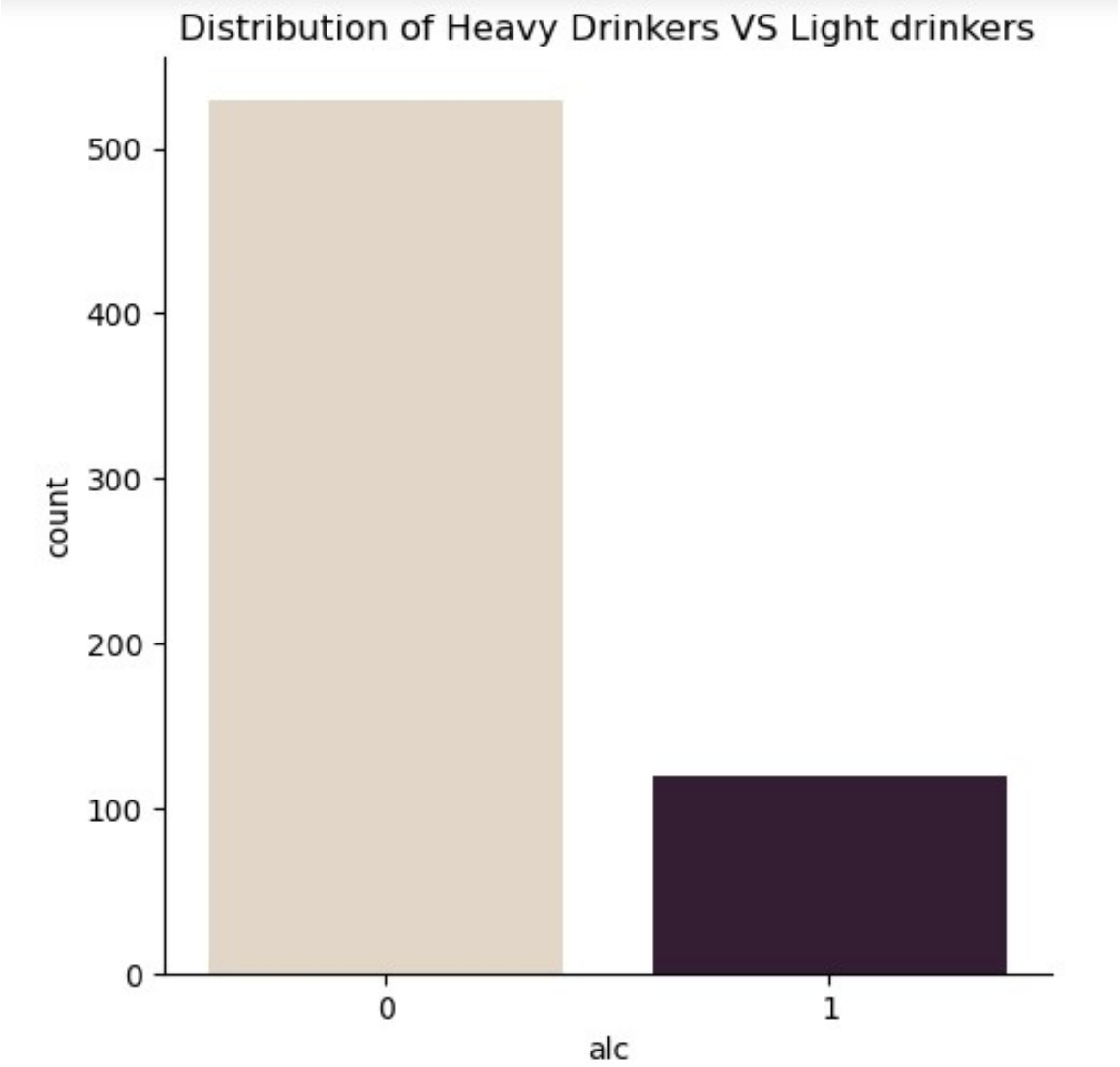

Figure 4.9: Distribution of Heavy Drinkers vs. light drinkers

4.2.1 Handling Missing Data

In our dataset, there are 649 observations and 33 attributes in total. Since we have found no missing values in our dataset using the isnull() command, we do not need to handle missing data.

4.2.2 Encoding Categorical Variables

By utilizing a label encoder, we transformed categorical features such as school, student's sex, and address into numeric values, simplifying the dataset for analysis. However, it's worth noting that if we had opted for one-hot encoding instead, the number of columns would have expanded from 33 to 51 due to the creation of binary

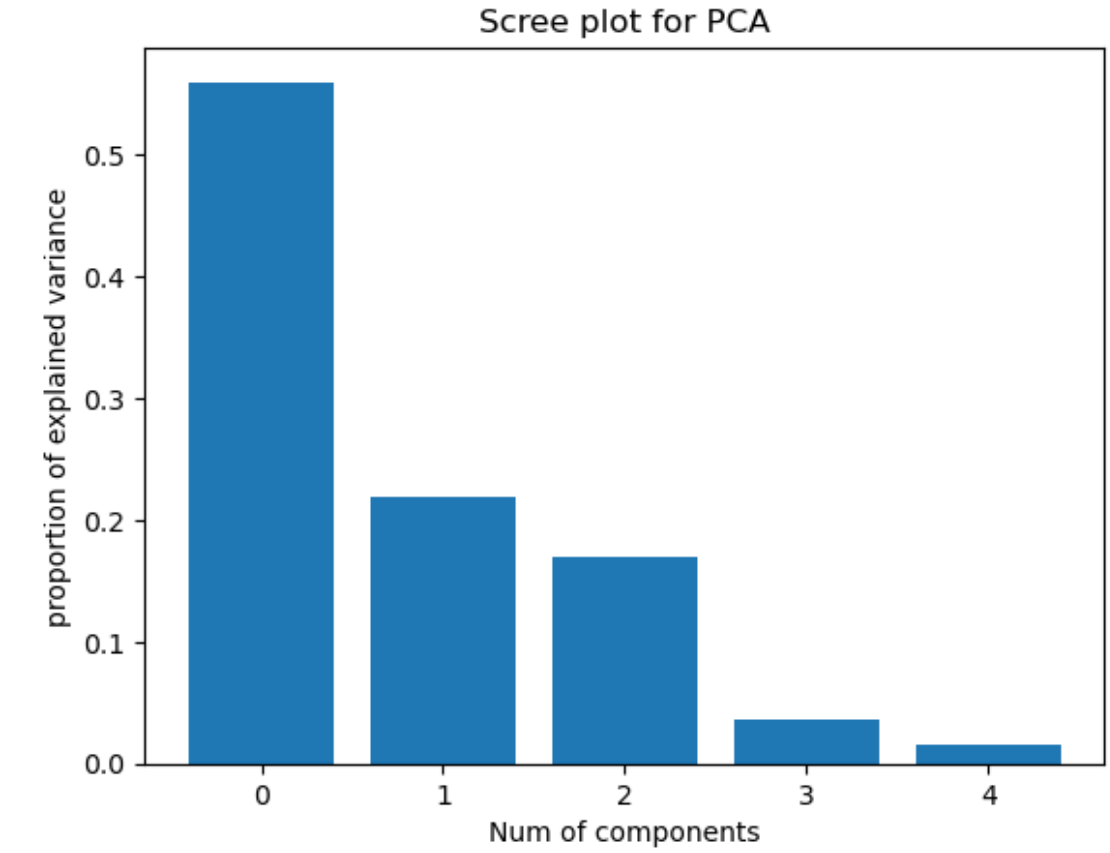

Figure 4.10: Distribution of Heavy Drinkers vs. light drinkers

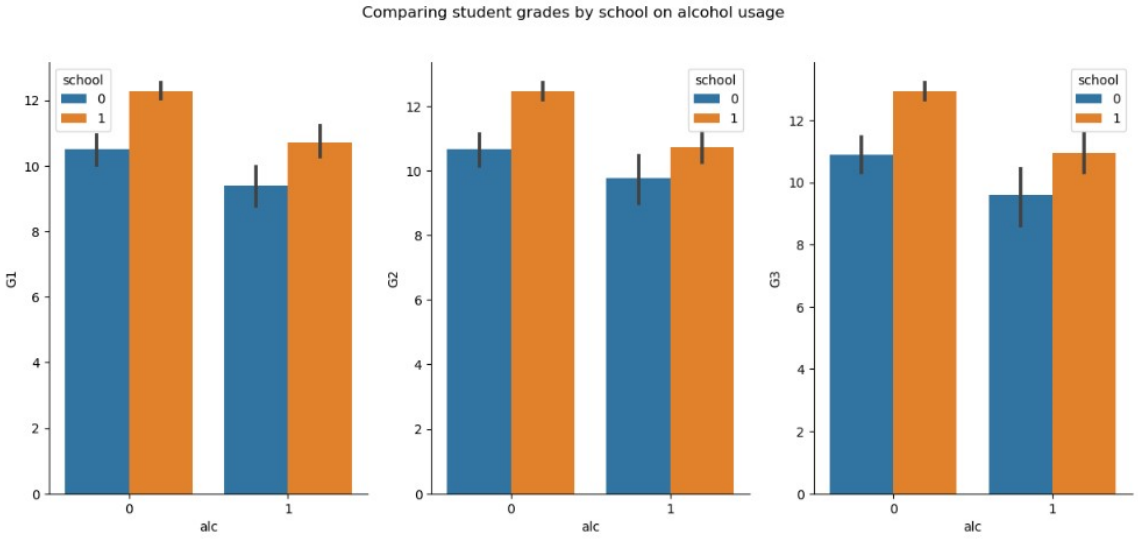

Figure 4.11: Distribution of Heavy Drinkers vs. light drinkers

columns for each category. Our choice of label encoding maintained dataset efficiency by condensing the information into fewer columns, ensuring it remained manageable while retaining its integrity and relevance for subsequent analyses.Fig 4.2.2.1 shows

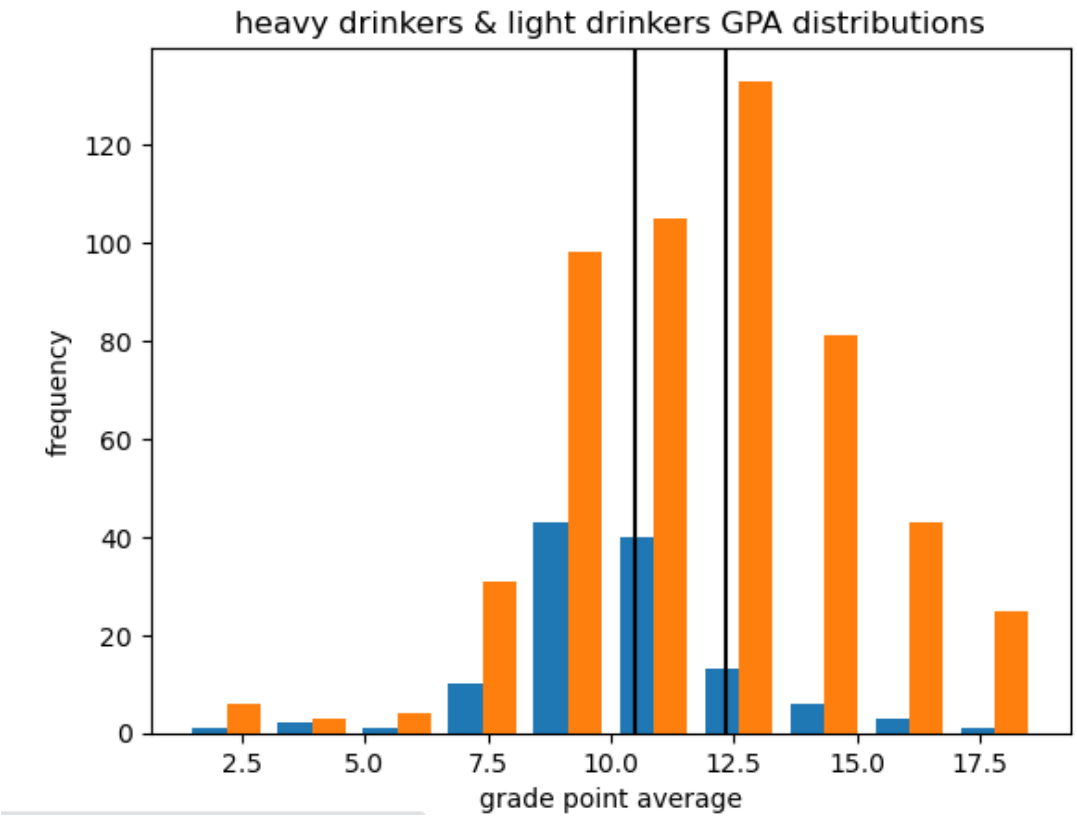

Figure 4.12: Distribution of Heavy Drinkers vs. light drinkers

the features after encoding.

4.2.3 Train-test split

In this project, our primary goal is to predict the average grade (grade avg) using alcohol consumption as well as additional features. Here, the target variable is the average grade calculated from individual grades (g1, g2, g3). Additionally, we aim to predict whether a student is a heavy drinker or a light drinker based on their grade and other features. In this case, the target variable is the alcohol consumption category, with values indicating whether a student falls into the heavy drinker (1) or light drinker (0) category. These dual prediction tasks provide a comprehensive understanding of the interplay between academic performance and alcohol consumption patterns among students.

To evaluate the performance of our machine learning model, we randomly split the dataset into two subsets: a training set and a test set. The training set (80The training set has 519 rows and 50 columns, including the 51th column, which is the alcohol target variable. The test set has 130 rows and 50 columns, including the alcohol target variable.

4.2.4 Class Imbalance Sampling

In our training dataset, we initially observed an imbalance between light drinkers (count: 426) and heavy drinkers (count: 93). This inconsistency posed a significant challenge as machine learning models tend to be biased towards the majority class, potentially resulting in poor performance for the minority class. To rectify this imbalance, we employed class imbalance sampling techniques. Through careful sampling methods, we balanced the dataset, achieving an equal count of both light drinkers and heavy drinkers (each with a count of 426). As a result, our dataset now consists of a total of 852 instances, each described by 50 features. This balanced and comprehensive dataset ensures that our machine learning models are trained on fair and representative data, paving the way for more accurate and unbiased predictions.

4.2.5 Scaling

In our dataset, scaling was applied to address the significant variations in the magnitude of features. This variation in scales could potentially skew the behavior of certain machine learning algorithms, leading to biased results. Scaling ensures that all features are on a similar scale, preventing any particular feature from dominating the learning process due to its larger magnitude.

After applying scaling techniques such as Min-Max scaling or standardization, the dataset underwent a transformative change. Features that previously spanned different ranges and units were harmonized to a standard scale. Min-Max scaling, for example, transformed features to a range between 0 and 1, preserving the relationships between the data points. On the other hand, standardization scaled features to have a mean of 0 and a standard deviation of 1, making the data follow a standard normal distribution.

This meticulous preprocessing step paved the way for more accurate and reliable predictions, ensuring that the model's performance was optimized across all features in the dataset.

4.2.6 Feature Selection

In our dataset, comprising 51 features where 50 features are variable and one is target variable, a strategic decision was made to enhance the model's efficiency. Subsequently, we embarked on a meticulous analysis to identify the optimal number of features for our model. We systematically experimented with varying sets of variables, including 10, 15, 20, 25, 30, and 35 features. For each set, we trained the model

and evaluated its accuracy. Through this iterative process, we discerned that as the number of features increased, so did the accuracy of the model. Crucially, the highest accuracy was achieved with a specific number of features, guiding our selection process. We chose to incorporate the variable subset that yielded the highest accuracy into our final dataset. This strategic approach not only streamlined the model but also ensured that it was enriched with the most influential features, enhancing its predictive performance. By leveraging these carefully chosen variables, our dataset was optimized, empowering the model to make accurate predictions while maintaining simplicity and interpretability.

4.3 Classification

Classification is a supervisory technique that categorizes the data into the desired number of classes [21] . The goal of this work is to find out the factors behind alcohol consumption and predict a person's probability of being an alcoholic. In that manner, we can try to keep people, especially teenagers, away from this deadly situation.
In our comprehensive approach to this project, we employ two distinct prediction tasks to unfold the intricate dynamics between academic performance and alcohol consumption among students.
For the first task, our objective is to predict the average grade (grade avg) using linear regression. We closely split the dataset into training and testing sets, training a linear regression model on a suite of features, including alcohol consumption and other relevant factors. The model's efficiency is then assessed on the testing set using regression metrics, such as mean squared error, providing insights into the nuanced relationship between academic achievement and various influencing factors.
In parallel, our second prediction task involves classifying students into heavy drinkers or light drinkers. To accomplish this, we convert the problem into a binary classification challenge, with 0 denoting light drinkers and 1 denoting heavy drinkers. The dataset is once again divided into training and testing sets, and we deploy a range of classification algorithms.This ensemble includes Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), and k-Nearest Neighbors (k-NN). Each model is individually trained and evaluated on the testing set, allowing for a meticulous comparison of their predictive capabilities.
In addition to these individual models, we may explore ensemble methods to synergistically combine predictions from multiple classifiers, aiming to enhance overall accuracy. Throughout the process, emphasis is placed on model interpretation, partic-

ularly with logistic regression, where coefficients and feature significance are examined to discern the impact of each factor on the likelihood of being a heavy drinker.
The methodology includes rigorous validation through techniques such as cross-validation, ensuring the robustness of our models. Fine-tuning hyperparameters further optimizes the performance of each classification algorithm. Through this multifaceted approach, our project endeavors to provide nuanced insights into the dual dynamics of academic performance prediction and alcohol consumption classification among students.

CHAPTER V

RESULT ANALYSIS AND DISCUSSION

5.1 Overview

In this section, the first data analysis and then the output evaluation are described. For our result analysis, Logistic Regression, k-Nearest Neighbors (kNN), Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM) are evaluated using accuracy, precision, recall, and the F1 score. Additionally, Linear Regression's predictive performance is assessed through mean absolute error and mean squared error, providing a comprehensive understanding of both classification and regression model outcomes.

5.2 Dataset Analysis

Correlation heatmap: A correlation heatmap is a machine learning technique that visually displays the correlation between variables in a dataset. It is a graphical representation of a correlation matrix, which shows the correlation coefficients between every pair of variables in the dataset. Correlation heatmaps can be used to find patterns and connections between variables in a dataset and are especially helpful in locating strongly correlated variables that may result in multicollinearity and influence the effectiveness of machine learning models.

Figure 5.1 shows the correlation heatmap for our dataset. A higher value represents a strong correlation. We can see the correlation between a number of variables related to student performance.

Pairplot: The pairwise correlations between variables in a dataset are displayed using the visualization approach known as a data pair plot in machine learning. It is a kind of scatterplot matrix that compares every variable in the dataset to every other variable. A scatterplot of each variable's combined distribution and a diagonal plot

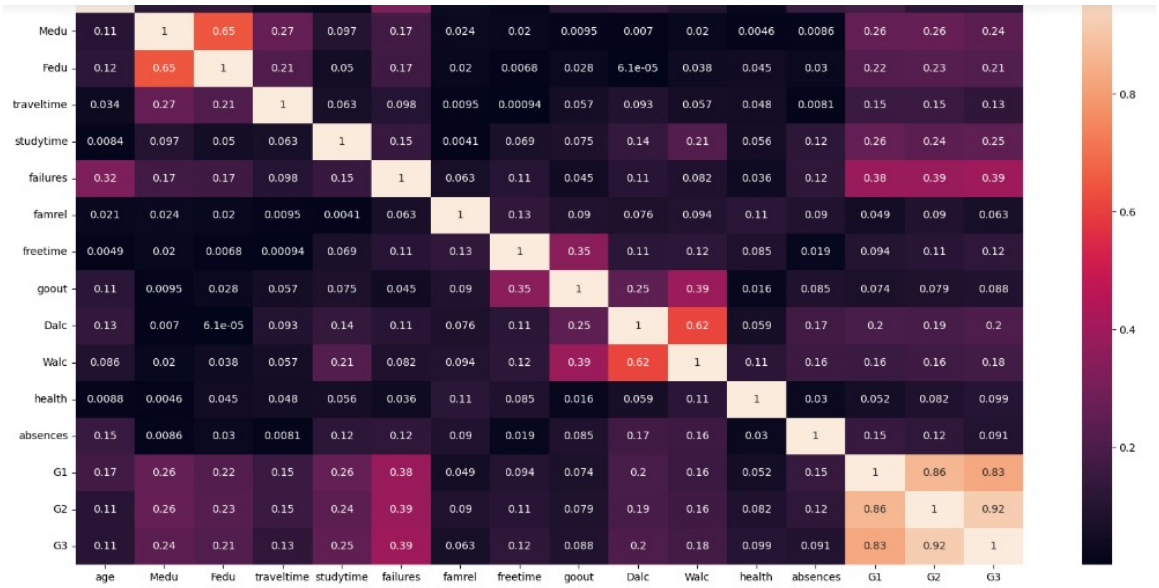

Figure 5.1: Correlation Heatmap

of each variable's distribution serves as the representations of each variable in a pair plot. The scatterplots can display any potential outliers or patterns in the data, as well as the strength and direction of the link between the two variables.

Our dataset contains 33 predictor variables and plotting the pair plot for our dataset would result in a 33 by 33 matrix that is too large. That's why we have selected 4 important features from our dataset based on the correlation coefficient value and have tried to show the pair plot for the selected features. Figure 5.2 shows the pair plot.

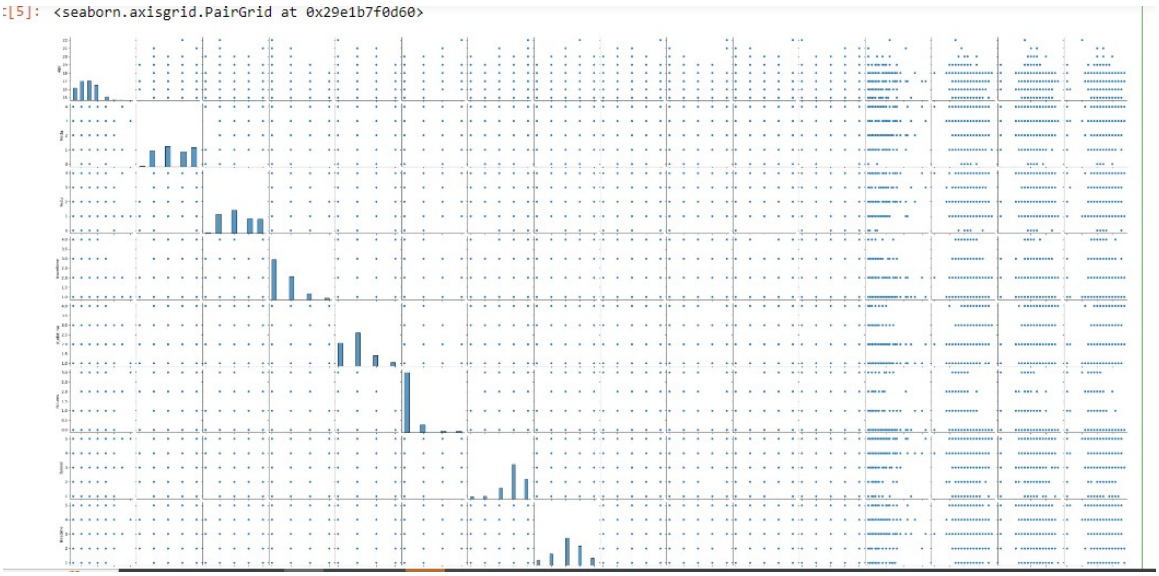

Figure 5.2: Pairplot of Important Features

5.3 Output Evaluation

5.3.1 Accuracy

Accuracy is a highly efficient and useful indicator to assess machine learning prediction accuracy in both of its forms. It is one of the metrics that are most frequently employed in research, where it is typical to have clean and balanced datasets to allow for attention to improvements in the algorithmic approach. AI accuracy is the ratio of correctly classified data that a trained machine learning model produces, or the proportion of correct predictions to all other forecasts combined. Accuracy is frequently referred to as AC. Depending on the scale selected, AC is expressed as a value between [0,1] and [0,100]. If the classifier's accuracy is 0, it consistently guesses the incorrect label; if it's 1, or 100, it consistently predicts the right label. True positives (TP) and true negatives (TN) make accurate predictions (TN). The whole set of positive (P) and negative (N) instances make up each forecast. False positives (FP) and false negatives (TN) make up P, while FP and TN make up N. (FN).[24].
Whole dataset accuracy for traditional ML models: To evaluate the performance of traditional machine learning models on our dataset, we employed six different classifiers: Linear Regression, Logistic Regression, Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and K-Nearest Neighbors (KNN).
First, we used Linear Regression to predict the grades of students. Then, we used the other five models to predict the effect of alcohol consumption on study habits, based on grade point average and other features that indicate whether the student is a heavy drinker or a light drinker.

5.3.1.1 Precision, Recall and F1 score

Precision, recall, and F1 score are common evaluation metrics used in machine learning to measure the performance of a classification model. In general, a model with high precision, recall, and F1 score is considered to be a good model. By analyzing the whole dataset, we got the highest precision, recall, and F1 score for DT and RF for 10 features and the lowest precision, recall, and F1 score for 10 features.

5.3.2 Confusion Matrix

A confusion matrix is a useful tool for understanding the performance of a classification model. It helps to identify cases where the model is making errors and can

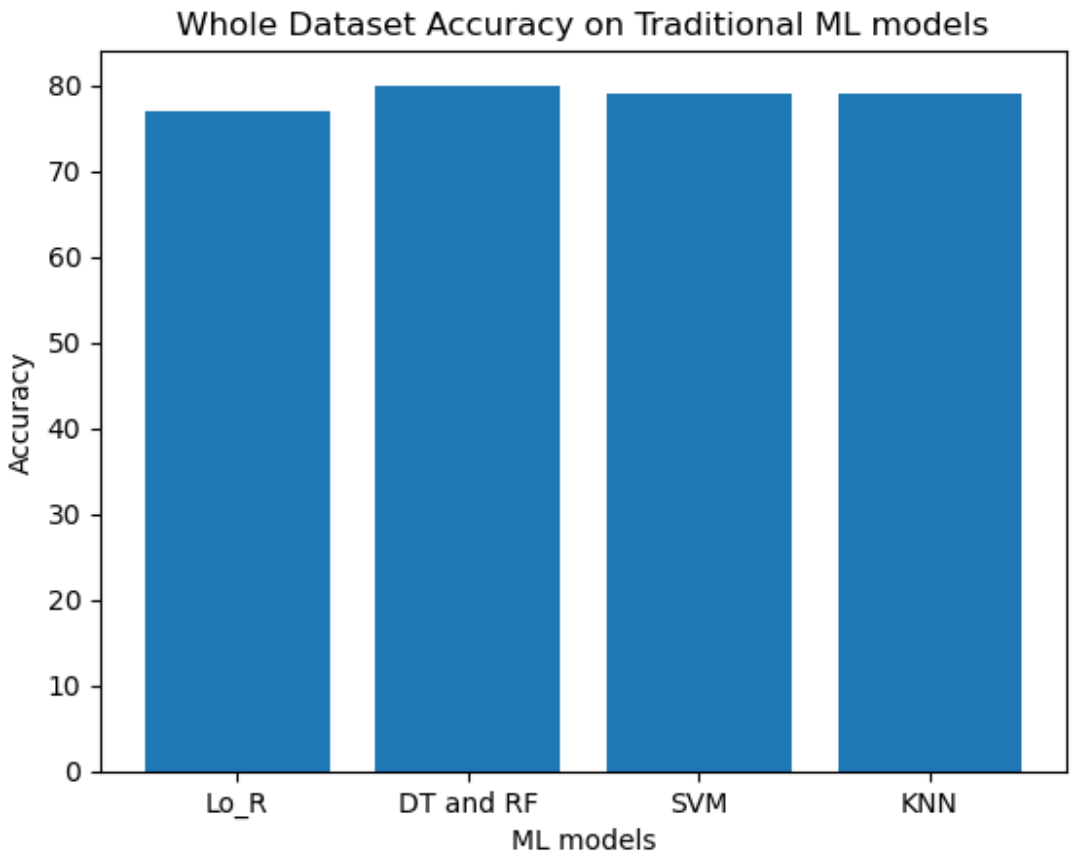

Figure 5.3: Whole Dataset Accuracy on Traditional ML models

be used to guide improvements to the model. Additionally, a confusion matrix can be used to compare the performance of different classification models and to choose the best model for a given problem.[22]

5.4 Discussion

We have developed a model that predicts the effects of alcohol on the study with an accuracy of 80 percentage. The model was trained with 80 percentage of the data, and testing was performed with the remaining 20 percentage of the data. This model's performance was significantly increased by using feature selection and SMOTE.

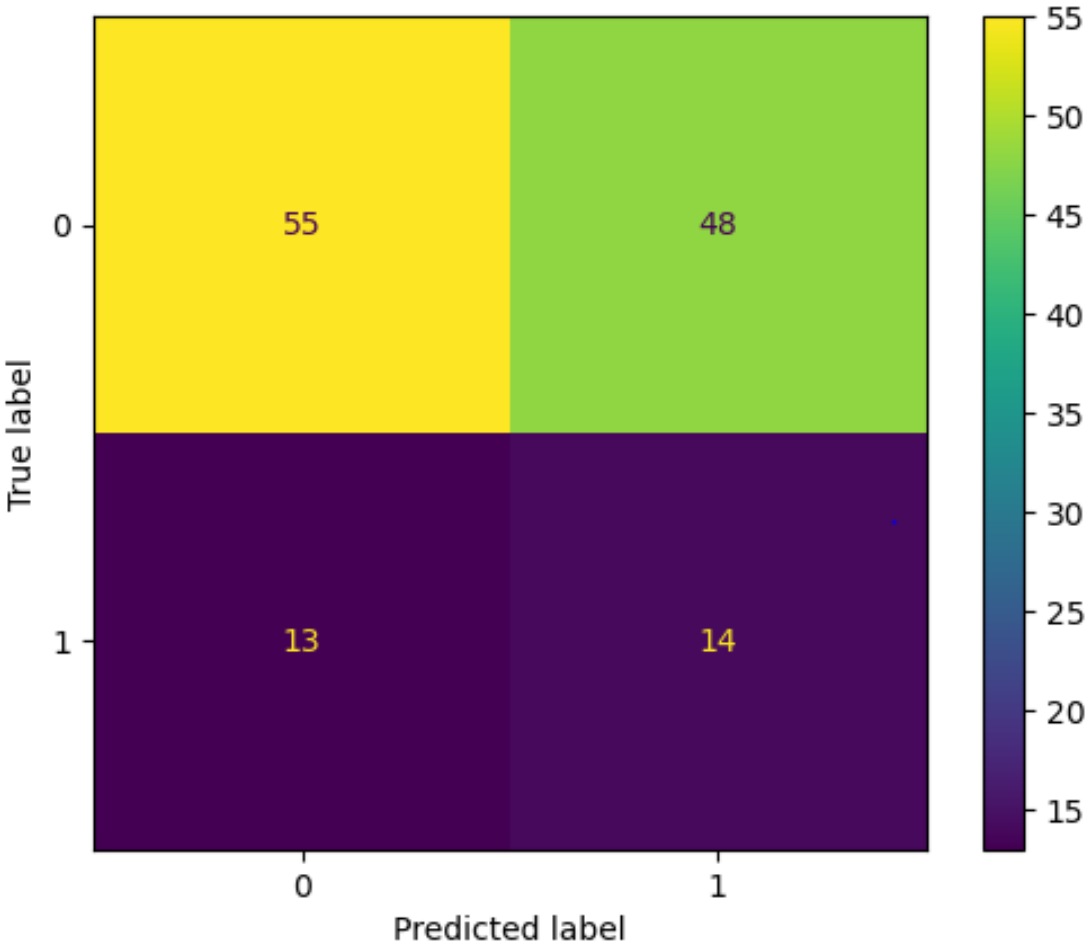

Figure 5.4: Precision, Recall, and F1 Score For RF with 10 featured variable

CHAPTER VI

CONCLUSION AND FUTURE WORK

6.1 Limitations and Future Work

Our study finds two significant limitations. Firstly, the relatively small size of our dataset, sourced from records of just two schools, raises concerns about the generalizability of our findings. The unique characteristics of these schools may not fully encapsulate the diverse dynamics found in broader student populations. Acknowledging this limitation is crucial for interpreting and applying our conclusions in a more extensive educational context.

Secondly, the absence of real-time integration or a dedicated website limits the immediate applicability of our research. A real-time system or an interactive online platform could empower students with timely insights into the relationship between alcohol consumption and academic performance. To address these limitations, future work should focus on expanding the dataset to include a more diverse range of schools. Simultaneously, the development of a real-time integrated platform or website could enhance the practical utility of our findings, providing students with immediate and personalized guidance. This forward-looking approach aligns with the overarching goal of translating research outcomes into actionable tools that directly benefit students in navigating the complexities of alcohol-related decisions and academic success.

6.2 Conclusion

In summary, our project employs linear regression to predict average grades, considering features such as alcohol consumption. The classification task, utilizing Logistic Regression, Decision Tree, Random Forest, SVM, and k-NN, targets the prediction of heavy or light drinking, with 120 heavy drinkers and 529 light drinkers identified

in our dataset. Ensemble methods may further refine predictions. Cross-validation and hyperparameter tuning enhance model robustness, offering insights into the complex dynamics of academic performance and alcohol consumption among students. This multifaceted approach aims to provide nuanced perspectives for educators, policymakers, and researchers based on a dataset comprising various influencing factors and 649 student records. Precision, recall, and F1 score were used to measure the performance of the classification model and represented as graphs. We also used Confusion Matrix to find where the model is making errors.

References

[1] C. M. Walid El Ansari, Christiane Stock1, "Is alcohol consumption associated with poor academic achievement in university students?," *National Library of Medicine*, vol. 4, no. 10, p. 1175, 2013. Publisher: PMC.

[2] A. K. L. Angela M. Haeny1 and K. J. Sher1, "Limitations of lifetime alcohol use disorder assessments: A criterion-validation study," *National Library of Medicine*, vol. 59, no. 10, p. 95, 2016. Publisher: sciencedirect.

[3] "Seaborn." Available at: `https://www.w3schools.com/python/numpy/numpy_random_seaborn.asp` [Last Access on August 12, 2023,7.00 pm].

[4] "Introduction to panda." Available at: `https://www.geeksforgeeks.org/introduction-to-pandas-in-python/`[Last Access on August 15, 2023,09.00 pm].

[5] "python numpy tutorial." Available at: `https://www.mygreatlearning.com/blog/python-numpy-tutorial/`[Last Access on August 18, 2023,08.00 pm].

[6] "An introduction to matplotlib." Available at: `https://www.simplilearn.com/tutorials/python-tutorial/matplotlib`[Last Access on August 19, 2023,07.00 pm].

[7] "Scikit learn tutorial." Available at: `https://www.tutorialspoint.com/scikit_learn/index.htm`[Last Access on August 23, 2023,7.00 pm].

[8] "working with missing data." Available at: `https://www.geeksforgeeks.org/working-with-missing-data-in-pandas/`[Last Access on August 30, 2023,7.00 pm].

[9] "Feature selection techniques in machine learning." Available at: `https://www.javatpoint.com/`

feature-selection-techniques-in-machine-learning[Last Access on August 31, 2023,7.00 pm].

[10] "Categorical data encoding techniques." [Last Access on September 05, 2023,7.00 pm].

[11] "Imbalanced data : How to handle imbalanced classification problems." Available at: https://www.analyticsvidhya.com/blog/2017/03/imbalanced-data-classification/#:~:text=Introduction,belonging%20to%20the%20other%20classes[Last Access on September 08, 2023,7.00 pm].

[12] "Linear regression in machine learning." Available at: https://www.geeksforgeeks.org/ml-linear-regression/[Last Access on September 13, 2023,7.00 pm].

[13] "An introduction to logistic regression." Available at: https://www.simplilearn.com/tutorials/machine-learning-tutorial/logistic-regression-in-python#:~:text=BootcampEXPLORE%20PROGRAM-,What%20is%20Logistic%20Regression%3F,one%20or%20more%20independent%20variables[Last Access on September 20, 2023,7.00 pm].

[14] "Support vector machine algorithm." Available at: https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm[Last Access on September 26, 2023,7.00 pm].

[15] "k-nearest neighbors algorithm." Available at: https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm[Last Access on September 30, 2023,7.00 pm].

[16] "Decision tree." Available at: https://www.geeksforgeeks.org/decision-tree/[Last Access on October 20, 2023,7.00 pm].

[17] "Random forest algorithm." Available at: https://www.javatpoint.com/machine-learning-random-forest-algorithm[Last Access on October 22, 2023,07.00 pm].

[18] "Cross-validation techniques." Available at: https://www.analyticsvidhya.com/blog/2021/05/4-ways-to-evaluate-your-machine-learning-model-cross-validation-techniques-wi#:~:text=Cross%2Dvalidation%20is%20a%20technique%20used%20in%

20machine%20learning%20and,to%20obtain%20reliable%20performance%20metrics[Last Access on October 23, 2023,09.00 pm].

[19] I. Z. Kotsiantis, Sotiris B. and P. Pintelas., "Supervised machine learning: A review of classification techniques," *Informatica*, vol. 31, no. 10, p. 245, 2007. Publisher: researchgate.

[20] "Student alcohol consumption." Available at: https://www.kaggle.com/datasets/uciml/student-alcohol-consumption [Last Access on October 24, 2023,09.00 pm].

[21] e. a. Rahman, Mohammad Mizanur, "Psycho-social factors associated with relapse to drug addiction in bangladesh," *Journal of Substance Use*, vol. 21, no. 6, p. 627, 2015. Publisher: researchgate.

[22] A. C. Müller and S. Guido, *Introduction to Machine Learning with Python*. United States of America, 2016.