




Analysis of Health Status of Alcoholic Students Based on the Gradient Boosting Decision Tree

Yijie Tang¹^a, Yixiang Yang²^{b*} and Baoshu Zhao³^c

¹Information Engineering College, Changsha Medical University, No. 1501, Leifeng Avenue, Wangcheng District, Changsha, China

²Maynooth International Engineering College, Fuzhou University, No.2 Wulongjiang North Avenue, Fuzhou, China

³Tianjin Foreign Language School, Tianjin, 30000, China

Keywords: Health Status, Alcoholic, Gradient Boosting Decision Tree.

Abstract: Underage alcohol consumption has become an increasingly serious issue in contemporary society. Not only does it violate legal regulations, but it may also lead to a series of severe social problems, such as poor academic performance and mental health issues. Therefore, conducting comprehensive and in-depth research on underage alcohol consumption can help better understand its influencing factors and potential consequences, thus taking corresponding measures to prevent and address them. To address this issue, this study employs the Gradient Boosting Decision Tree (GBDT) model for deep learning and prediction using a dataset of students' academic performance. The analysis includes insights into the features and attributes based on GBDT, as well as an examination of the relationship between alcohol consumption and other characteristics. The results reveal that among students engaged in alcohol consumption, 25% ultimately achieve unsatisfactory academic performance. Simultaneously, 50% of students with the lowest alcohol consumption enter the next semester with the lowest grades. However, students consuming moderate amounts of alcohol on weekdays experience consistent failure in both aspects.


1 INTRODUCTION


In modern society, the issue of alcohol abuse is prevalent among student populations, causing serious negative impacts on individual health and societal stability (Yue, 2019). Therefore, conducting a scientific analysis of the lifestyle of students engaged in alcohol abuse is crucial for developing effective strategies to prevent and intervene in alcohol-related problems.


Alcohol abuse poses significant challenges to both the physical and mental well-being of individuals and societal stability. Behaviors of alcohol abuse during student years often coincide with academic issues, family conflicts, and mental health disturbances, reflecting unfavorable life conditions. Over time, alcohol abuse has been identified as a high-risk behavior associated with adverse consequences such as psychological dependence, substance abuse, traffic

accidents, and criminal activities (Peng, 2023). Therefore, gaining a deep understanding of the lifestyle of students engaged in alcohol abuse helps uncover the relationships between alcohol abuse and individual psychology, family environment, and societal influences.

Currently, research has predominantly focused on the influencing factors and severity of alcohol abuse, lacking a comprehensive and in-depth understanding of the lifestyle of students engaged in alcohol abuse and its relationship with their alcohol-related behaviors. Hence, this study, in conjunction with gradient boosting trees, aims to explore the lifestyle characteristics of students engaged in alcohol abuse, identify potential influencing factors, and construct a predictive model to provide long-term scientific support for intervening in student alcohol abuse (Wang, 2021).

^a <https://orcid.org/0009-0008-1392-0611>

^b <https://orcid.org/0009-0004-0546-8793>

^c <https://orcid.org/0009-0033-1022-0329>

Through the adoption of the gradient boosting algorithm, this research efficiently processes and analyzes large-scale data to accurately predict the lifestyle of students engaged in alcohol abuse. In the research process, extensive data on alcohol-abusing students, including personal characteristics, family environment, and mental health information, will be collected. Subsequently, gradient boosting trees will be utilized to analyze and model this data, uncovering the most relevant features and factors associated with the lifestyle of students engaged in alcohol abuse. Through model validation, we aim to precisely predict the lifestyle of students engaged in alcohol abuse and delve into the correlations with alcohol-related behaviors.

2 PREPARATION PREDICTION METHODS BASED ON DEEP LEARNING

In this study, the Python compiler was employed to analyze and process the dataset. The dataset was input into the model for learning, and various visualizations such as heat maps were generated. By examining the degree of correlation between factors, predictions could be made. In this process, the following models were utilized:

In the field of machine learning, logistic regression is a simple and interpretable classification algorithm. It is particularly suitable for handling linearly separable data and can provide probability estimates for each class.

Decision trees are another intuitive and easily interpretable model that requires minimal data preprocessing and can handle both numerical and categorical data.

Random forests improve model robustness by ensemble learning, combining multiple decision trees. They perform well, especially in handling high-dimensional data and providing feature importance rankings.

GBDT reduces bias and variance by iteratively adding trees, achieving excellent predictive performance and being applicable to different types of predictors. In simple terms, it boasts superior predictive performance, can handle different types of predictors (numerical and categorical), reduce bias and variance, and to some extent, address overfitting issues (Chen,2023; Pang,2020).

XGBoost is an upgraded version of the gradient boosting model with regularization features to prevent overfitting. It also offers fast and scalable

advantages due to support for parallel and distributed computing (Yu, 2023).

Adaboost combines weak learners into a strong learner, exhibiting anti-overfitting characteristics and compatibility with various base learners (Zeng, 2023).

DNN are models capable of learning complex nonlinear relationships, suitable for solving highly complex problems. Their powerful representation learning ability enables them to automatically extract and learn features (Li,2023).

GBDT, through iterative addition of trees, reduces bias and variance, achieving outstanding predictive performance and applicability to different types of predictors. In simple terms, it boasts superior predictive performance, can handle different types of predictors (numerical and categorical), reduce bias and variance, and to some extent, address overfitting issues.

In the field of deep learning, the aforementioned seven models each have their own characteristics. The first six belong to traditional machine learning algorithms; however, compared to traditional machine learning methods, deep learning models exhibit greater potential in handling large-scale and complex tasks (Liu,2019).

3 RESEARCH RESULTS AND ANALYSIS

3.1 Data Set Introduction

The dataset focuses on the relationship between alcohol consumption behavior and academic performance among high school students. It comprises 649 valid records of high school students engaged in alcohol consumption, with 31 features covering various aspects of students' lives, including school, gender, age, living arrangement, family size, parental relationship status, mother and father's education level, occupation, reasons for school selection, family support, study time, additional educational support, participation in extracurricular activities, internet usage, relationship status, family relationship quality, extracurricular time utilization, time spent with friends, weekday and weekend drinking habits, health status, school absenteeism, and grades for two semesters. By analyzing these features, we can gain a comprehensive understanding of various factors influencing high school students in both academic and life aspects, as well as the potential connections between these factors and

alcohol consumption behavior. This information is crucial for devising intervention measures and enhancing the overall quality of students' well-being.

3.2 Experimental Setup and Evaluation Metrics

Among the seven models, namely logistic regression, decision tree, random forest, GBDT, AdaBoost, XGBoost, and DNN, for the sake of convenience in research, we integrated the label column "Health Status" into binary form, focusing on whether the high school students engaged in alcohol consumption are healthy. In addressing this issue, accuracy is a crucial metric, typically represented as the percentage of correctly classified samples out of all samples diagnosed. Additionally, for disease diagnosis, recall often carries greater significance than false positives, making it an essential performance evaluation metric for such classification models. It is defined as the probability that the samples of these students are correctly detected as being in a healthy state. Furthermore, AUC (Area Under the Curve), as a model performance metric, is commonly used to measure the model's generalization ability, calculated from the area under the ROC curve (Hu,2022). The comparative results of these three metrics are shown in Figure 1, Figure 2, and Figure 3, with an overall comparison presented in Table 1.

In this experiment, we utilized the Scikit-learn platform for logistic regression, decision tree, random forest, GBDT, AdaBoost, and XGBoost, while employing TensorFlow for deep learning with the DNN model. Simultaneously, default hyperparameters for each model method were used on both the Scikit-learn and TensorFlow platforms. Through these platforms, we conducted a comparative analysis of accuracy, recall, and AUC, ultimately selecting the best-performing model.

3.3 Performance Comparison Analysis

We observed that the accuracy of GBDT consistently surpasses 70%, reaching above 72%, compared to other models, excluding Random Forest and DNN, which generally fall below this threshold. The recall rate also exceeds 90%. Additionally, GBDT's AUC performance significantly outperforms other models, reaching 61%.

In terms of nonlinear relationship modeling capability, GBDT effectively captures nonlinear relationships between features. Given the potential complex nonlinear relationships between students' academic performance and alcohol consumption, excessive alcohol consumption may lead to a decline in academic performance, but other interacting factors, such as family environment and study habits, may also play a role. Handling mixed data types is crucial in this dataset, which includes both numerical features (e.g., age, grades) and categorical features (e.g., gender, family environment). GBDT is capable of simultaneously handling both types of features and automatically selecting optimal split points and features during the training process. Regarding feature importance analysis, GBDT allows for the ranking of the importance of each feature in predicting students' academic performance and alcohol consumption. This aids in understanding which factors have a greater impact on students' academic and behavioral outcomes, providing targeted intervention measures and recommendations. In summary, GBDT demonstrates superior accuracy, recall, and AUC, showcasing its effectiveness in capturing complex relationships and handling mixed data types. The feature importance analysis further enhances its utility in identifying influential factors for students' academic and behavioral outcomes, facilitating tailored interventions and recommendations.

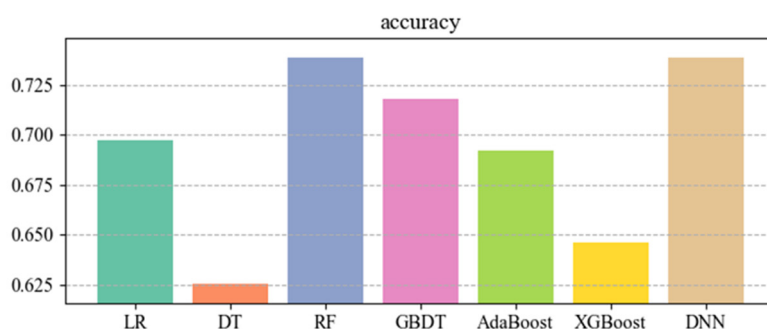


Figure 1: Comparison of Accuracy of various deep learning models (Photo/Picture credit: Original).

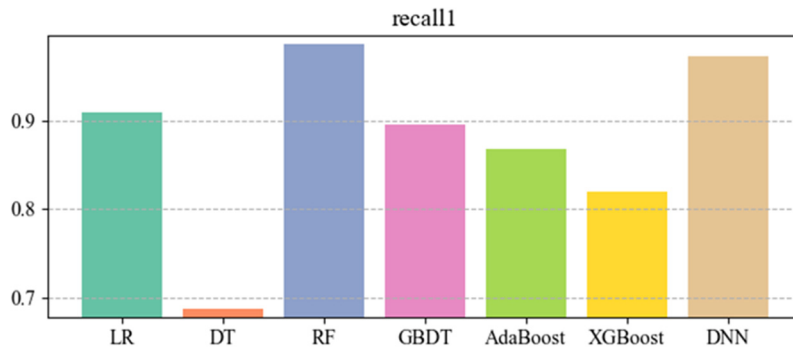


Figure 2: Comparison of recall rates of various deep learning models (Photo/Picture credit: Original).

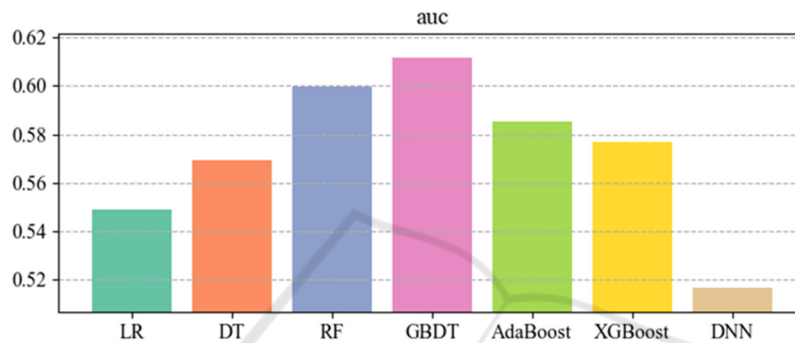


Figure 3: Comparison of AUC of various deep learning models (Photo/Picture credit: Original).

Table 1: Comparison of various deep learning models.

	LR	DT	RF	GBDT	AdaBoost	XGBoost	DNN
Accuracy	0.697436	0.625641	0.738462	0.717949	0.692308	0.646154	0.738462
Recall1	0.909722	0.687500	0.986111	0.895833	0.868056	0.819444	0.972222
AUC	0.548883	0.569240	0.599537	0.611520	0.585376	0.576797	0.516748

3.4 Feature Attribute Analysis Based on GBDT

Through the GBDT, we generated heatmaps depicting the relationships between various features and other features, as shown in Figure 4. The visualizations demonstrate the degree of relationship between health status and other features. Using a threshold of 0.05, we identified significant correlations between health status and various features, such as age (as shown in Figure 5), weekday alcohol consumption (as shown in Figure 6), weekend alcohol consumption (as shown in Figure 7), living alone (as shown in Figure 8), and

family size (as shown in Figure 9). Notably, we found that lower alcohol consumption is associated with better health status. In terms of age, younger individuals tend to have better health status. Living with parents is correlated with better health status. Larger family size is also associated with better health status. Additionally, spending less time with friends is correlated with better health status. Next, we will delve into the specific features of these correlation relationships to gain a more comprehensive understanding of the factors influencing health status.

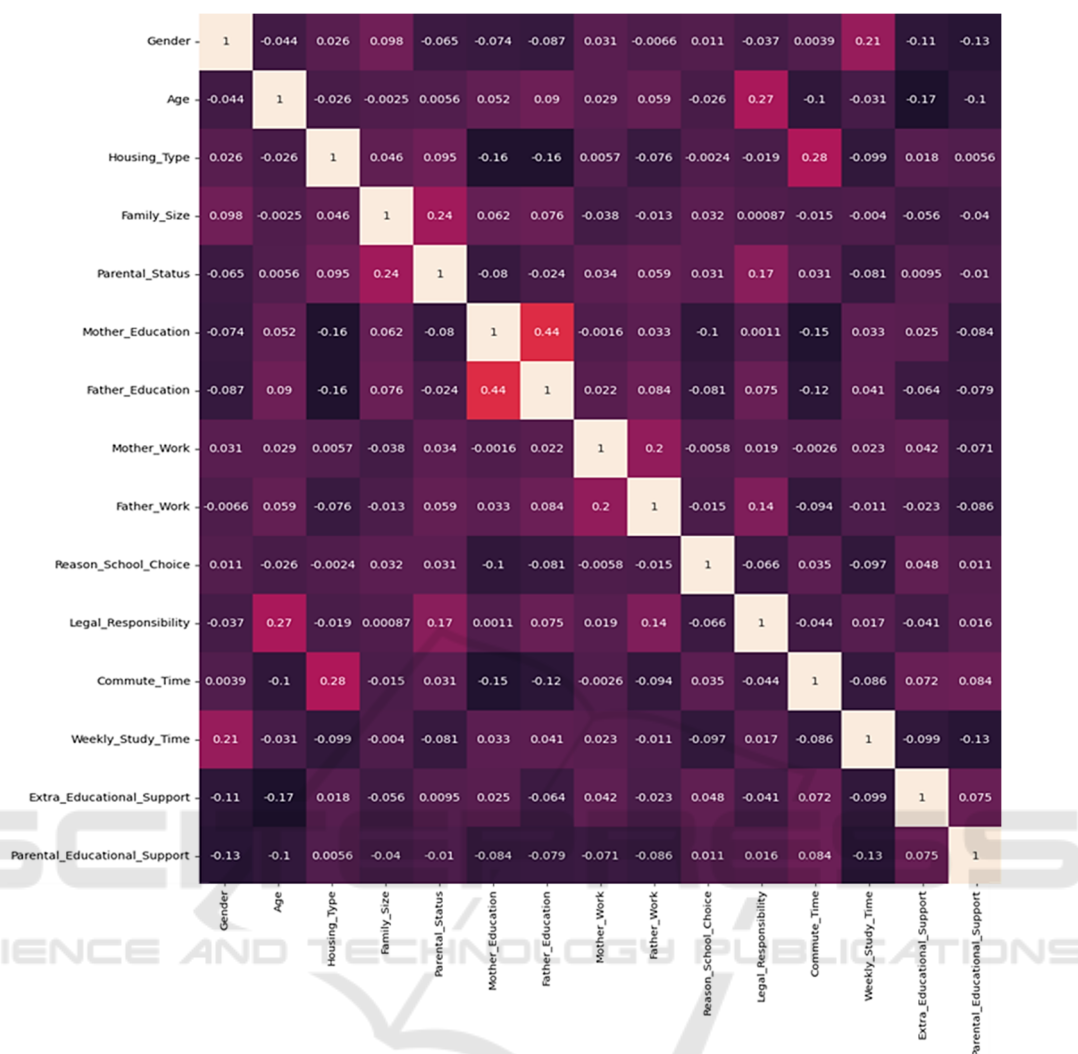


Figure 4: Heatmap of Each Feature with Other Features (Photo/Picture credit : Original).

Firstly, we further analyzed the distribution of health status across different age groups to determine if there is a trend indicating certain age groups are more susceptible to health issues. We observed that among adolescents aged 15-17, there is a high proportion with good health status, but as age increases, influenced by academic pressure and lifestyle habits, the proportion of individuals with poor health status gradually rises. Secondly, the correlation between weekday and weekend alcohol consumption and health status suggests that drinking behavior may significantly impact health. Ultimately, we found that higher levels of alcohol consumption, whether on weekdays or weekends, have a negative impact on health status. Additionally, the correlation between living alone and health status may imply the importance of a social support network. Upon a deeper exploration of the lifestyle and social

interactions of individuals living alone, we found that they exhibit noticeably lower social engagement compared to those living with others. Individuals living alone may lack close individuals for encouragement and assistance in times of difficulty (Roberts,2021). Finally, the correlation between family size and health status is also worthy of in-depth investigation. We examined the differences in health status among different family sizes and found that only children, compared to those with siblings, show distinct negative emotions such as food preferences and rejection, stubbornness, immediate demands, disrespect for elders, timidity, anger, and other negative emotion.

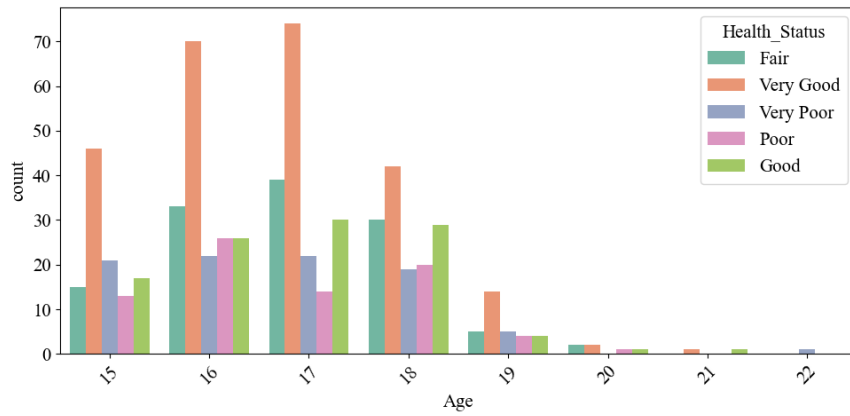


Figure 5: Relationship between Health Status and Age (Photo/Picture credit: Original).

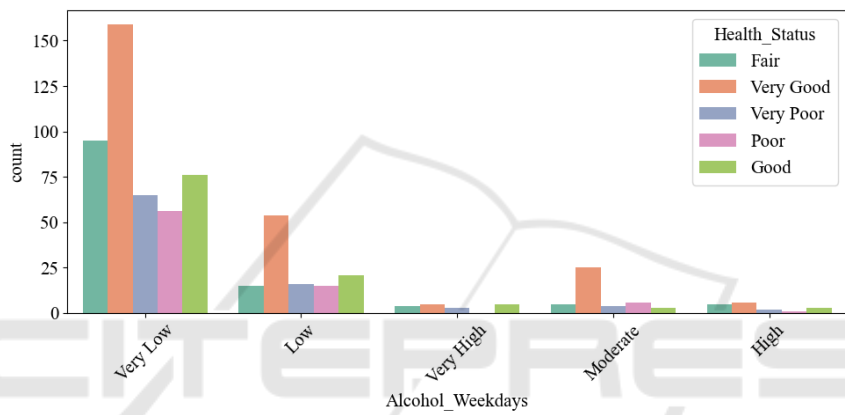


Figure 6: Relationship between Health Status and Weekday Alcohol (Photo/Picture credit: Original).

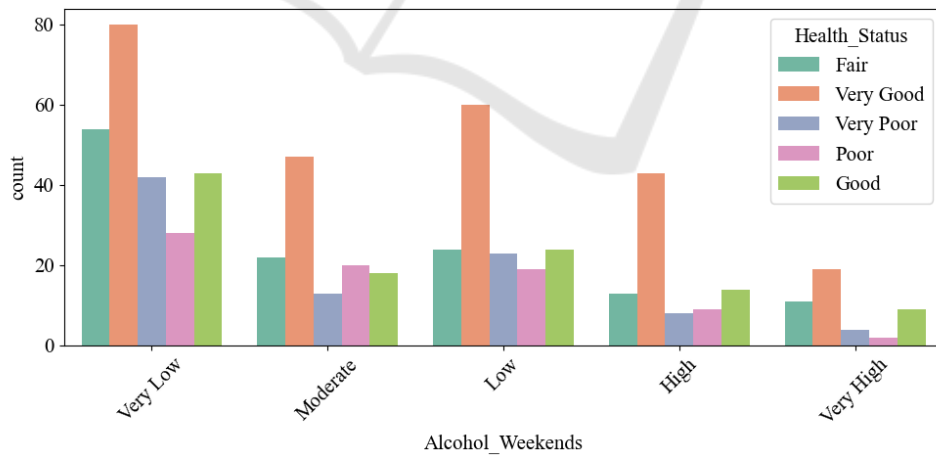


Figure 7: Relationship between Health Status and Weekend Alcohol (Photo/Picture credit: Original).

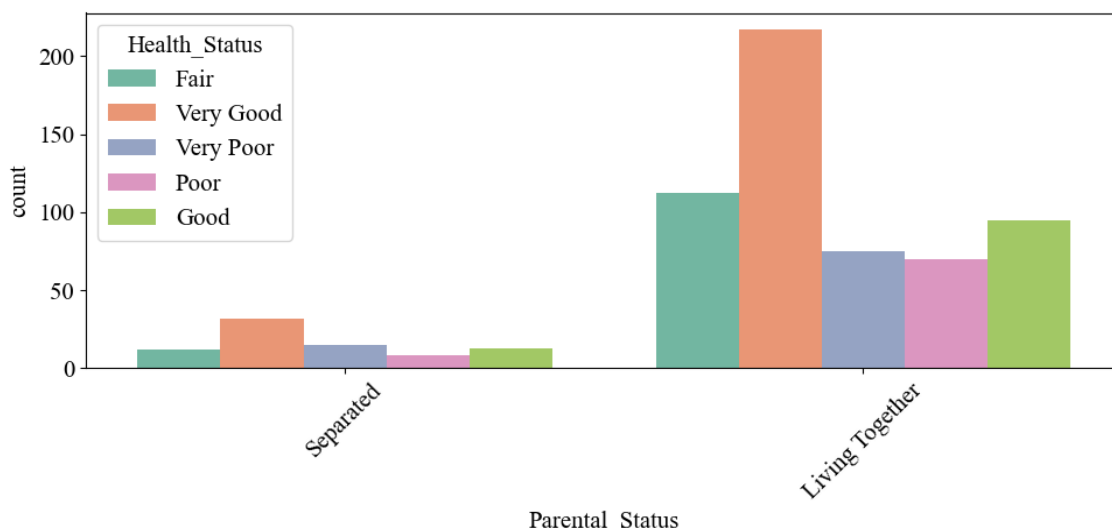


Figure 8: Relationship between Health Status and Parental Status (Photo/Picture credit: Original).

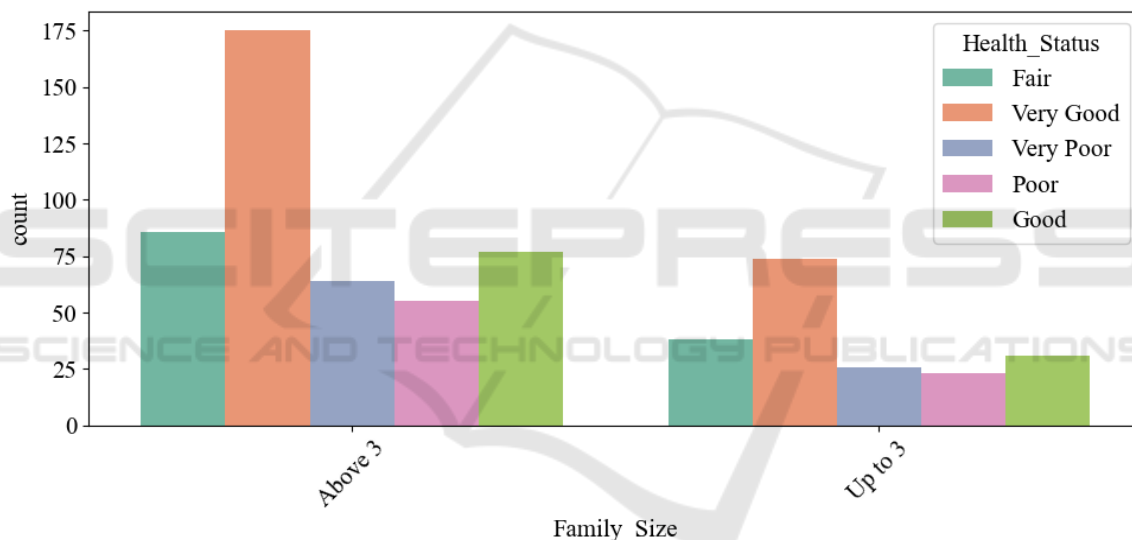


Figure 9: Relationship between Health Status and Family Size (Photo/Picture credit: Original).

3.5 Analysis of Alcohol Consumption and Other Attributes

While conducting a horizontal analysis of the factors influencing alcohol consumption, alongside the analysis of the health status feature, we observed strong correlations between alcohol consumption and study time (as shown in Figure 10), age (as shown in Figure 11), academic performance in two semesters (as shown in Figure 12), and family relationships (as shown in Figure 13).

We observed that among students who spend 2 hours studying per week, there is a significantly higher level of alcohol consumption. Additionally, as study time increases, especially among students with

study times exceeding 10 hours, the alcohol consumption is higher compared to students with study times between 5 to 10 hours. Interestingly, students who spend 2 to 5 hours studying per week exhibit the lowest levels of alcohol intoxication.

Age shows a slight correlation with alcohol consumption on weekdays or weekends. The primary drinking population consists of students aged 16-18. Most of them consume minimal alcohol on weekdays, with 17-18-year-old students having higher alcohol consumption, especially high or very high. However, on weekends, the alcohol consumption is nearly equal across all age groups. Additionally, students aged 20-22 tend to have high alcohol consumption on weekdays.

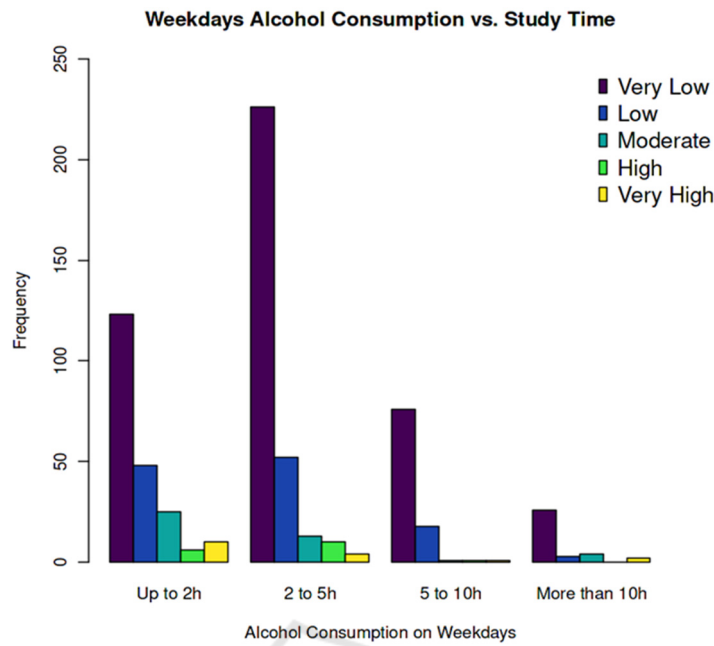


Figure 10: Relationship between weekday alcohol consumption and study time (Photo/Picture credit: Original).

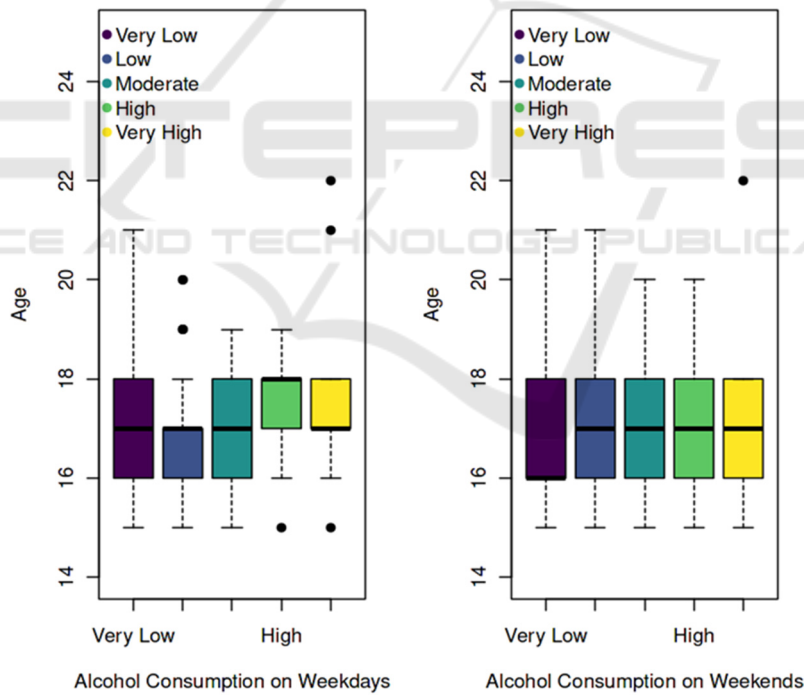


Figure 11: Relationship between alcohol consumption and age (Photo/Picture credit: Original).

According to the dataset, students can pass the semester if they achieve 60% of the total score, meaning they need to score 12 out of 20 points. In the bar chart, we can observe the relationship between grades in two semesters and alcohol consumption on weekdays and weekends. Notably, among students who consume alcohol, 25% of them fail in both

semesters, even with low alcohol content. Meanwhile, among students with the lowest alcohol consumption, 50% enter the next semester with the lowest grades. However, students who consume alcohol moderately, highly, or very highly on weekdays experience consistent failure in both aspects.

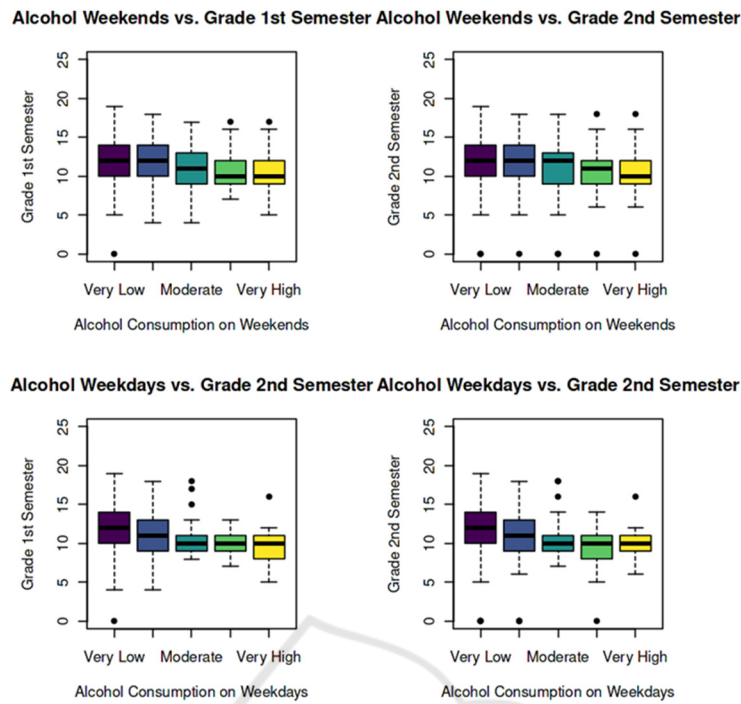


Figure 12: Relationship between alcohol consumption and academic performance in two semesters (Photo/Picture credit: Original).

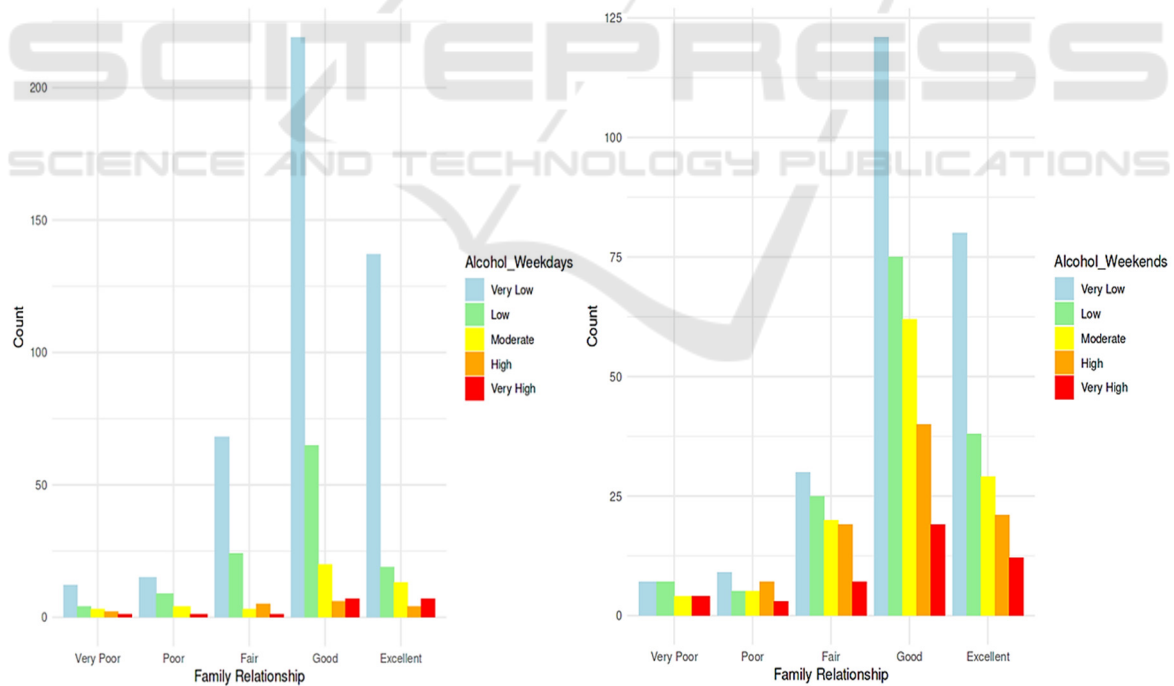


Figure 13: Relationship between alcohol consumption and family relationships(Photo/Picture credit: Original).

Family relationships may contribute to increased alcohol consumption. However, in the above charts, many students with good family relationships consume more alcohol than those with poor family

relationships. Moreover, on weekends, alcohol consumption significantly increases among students with good or even better family relationships. This unexpected observation leads us to the conclusion that

there is a correlation between family background and student alcohol consumption. This connection becomes more apparent on weekends, with most students staying at home during this time.

4 CONCLUSION

Based on the above study, in the predictive model for the health status of high school students engaged in alcohol consumption, there is an inseparable relationship between health status and alcohol consumption—the more alcohol consumed, the greater the negative impact on health status. Simultaneously, the health status of high school students engaged in alcohol consumption is highly correlated with four feature variables: age, living alone, family size, and time spent with friends. This study designed a Gradient Boosting (GBDT) model to predict the health status of high school students engaged in alcohol consumption and conducted a comparative analysis with five traditional machine learning algorithms (logistic regression, random forest, decision tree, XGBoost, Adaboost) and a DNN model used for detecting the health status of high school students engaged in alcohol consumption. Comparing the performance of the seven models using accuracy, AUC, and recall, it was found that the Gradient model has certain advantages, with an accuracy of 72% and an AUC of 66.3%.

While studying the model for the health status of high school students engaged in alcohol consumption, we also conducted a horizontal comparison of weekday and weekend alcohol consumption with other features. We found a strong correlation between the level of alcohol consumption and academic performance, study time, and family relationships. Therefore, parents, in understanding their child's alcohol consumption behavior, should not rely solely on health status for judgment. Instead, it is crucial to assess the child's daily interactions and gain a comprehensive understanding of their life situation to more accurately grasp their alcohol consumption behavior.

AUTHORS CONTRIBUTION

All the authors contributed equally and their names were listed in alphabetical order.

REFERENCES

- Chen, X. L., Cheng, S., Chen, K., Xiao, Z. Y. 2023, Research on Influencing Factors of Housing Prices in First-tier Cities Based on Machine Learning Methods. *Nankai Journal (Philosophy, Literature and Social Science Edition)*, (06): 146-163.
- Hu, C. Y., Hu, L. P. 2022, Reasonable Multiple Logistic Regression Analysis - Combined with ROC Curve Analysis. *Sichuan Mental Health*, 35(06): 493-499.
- Li, R. P., Zhu, J. J. 2023, Coronary Heart Disease Prediction Based on Improved Borderline-SMOTE-GBDT. *Chinese Journal of Medical Physics*, 40(10): 1278-1284.
- Liu, H. C. 2019, Comparative Analysis of Image Classification Algorithms Based on Traditional Machine Learning and Deep Learning. *Computer and Information Technology*, 2019, 27(05): 12-15.
- Pang, C., Jiang, Y., Liao, C. W., Wu, T., Yu, W., Wang, L. 2020, Research on Anti-Interference Technology for Strong Vibration Observation Based on AdaBoost Ensemble Learning. *Sichuan Earthquake*, (04): 14-18.
- Peng, S. T. 2023, Causes, Process, and Coping Strategies of Alcohol Addiction in Youth: In-Depth Interviews with Members of a Sobriety Association. *Youth Research*, (02), 82-93+96.
- Roberts, T., Krueger, J. 2021, Loneliness and the Emotional Experience of Absence. *South. J. Philos.*, 59: 185-204.
- Wang, H. B., Wu, J. J., Wu, X., Chen, C. Q., Chen, P. Y., Zhang, T. A. 2021, Research on Monthly Electricity Consumption Prediction of Office Buildings Based on Gradient Boosting Trees. *Electric Power Science and Engineering*, 37(04): 30-36.
- Yu, J. L. 2023, Evolutionary Algorithm-Based Multi-Objective Deep Neural Network Architecture Search. *Shandong University*, 58.
- Yue, J., & Zheng, X. Q. 2019, Reflections on Legal Issues Related to Guardianship of Alcohol Abusers in China. *China Health Law Review*, 27(06), 30-33.
- Zeng, S. R., Kong, M. 2023, Measurement Model of Phase Distribution in Gas-Liquid Two-Phase Flow Based on GBDT. *Chemical Industry and Engineering Progress*, 10, 17: 1-11.