PROJECT REPORT

Health Data-Based Predictive Modelling for Identifying Individuals with Alcohol Consumption Patterns

DATA7703_SEM2_2023 Group 3

Xijia Wei - s4718338
Ming Zhun Ku - s4772452
Gaojie Du - s4766696
Xuanyu Qin - s4765132
Lok Yee Joey Cheung - s4763354

We give consent for this to be used as a teaching resource.

Abstract

Alcohol consumption is a global public health concern that not only poses serious health risks to individuals but also gives rise to various potential medical issues. However, there are instances when individuals unconsciously deny their drinking habits. Failing to accurately identify patients with alcohol consumption habits can result in significant consequences, including misdiagnosis, adverse drug reactions, or treatment inefficiency. This impacts the treatment decisions from doctors and the overall treatment outcomes for patients. As such, the aim of this project is to develop a model that utilises medical data to identify patients with alcohol consumption habits.

The data used in this project is sourced from Korean National Health Insurance Service data from Kaggle. It contains 991,346 rows and 24 columns, encompassing a range of health parameters, including attributes such as age, weight, blood pressure, cholesterol levels, and more. Additionally, the dataset includes a labelled variable denoted as "DRK_YN". This dataset forms a comprehensive basis for the development of our identification model.

In this project, we harness the power of machine learning with various algorithms, including Logistic Regression, Decision Tree, Random Forest, LightGBM, AdaBoost, and CatBoost, to analyse the data and evaluate the performance of these models. Then, we utilise GridSearchCV imported from scikit-learn library to adjust the hyperparameters, ensuring each model operates at its best. Additionally, Cross-Validation is utilised to evaluate model performances and identify models with higher accuracies. To further boost predictive power, we create an ensemble test by combining AdaBoost, LGBM, and CatBoost using Majority Voting, resulting in enhanced predictive performance.

Table of Contents

1 Introduction

Intense alcohol consumption has been a detrimental problem with the global average consumption of 6.18 litres of pure alcohol per person per year, which is equivalent to 53 bottles of wine per person [1]. It is a global public health problem that leads to a wide range of negative health influences. According to key facts from the World Health Organization (WHO), the harmful use of alcohol leads to more than 200 disease and injury conditions, resulting in approximately 3 million deaths annually, accounting for 5.3 percent of all global deaths [2].

Not only alcohol consumption poses serious harm to an individual's health, but also causes a lot of underlying healthcare issues. In medical cases, knowing whether or not a patient is an alcoholic is critical to a doctor's diagnosis. For instance, if a doctor is unaware that a person is an alcoholic, they might underestimate the amount of anaesthesia used. In addition, both anaesthesia and alcohol can cause nausea and vomiting [3]. This may increase the risk of aspiration, known as inhaling vomit, which can be potentially fatal.

When a patient has already suffered from underlying medical conditions and alcoholism, the implications can be significant. They will be at greater risk for heart disease, liver diseases, gastrointestinal disease and other illnesses. Under such complex circumstances, doctors must invest more time and effort in selecting specific treatment methods and considerations to avoid medication conflicts.

In certain cases, to effectively address particular illnesses, doctors might recommend alcohol cessation for the patient. However, for alcoholics, when they are admitted to the hospital, they may not be allowed to drink for a period of time based on their treatments, which may result in the emergence of alcohol withdrawal symptoms and potentially severe, life-threatening complications [4]. As a result, physicians are required to develop appropriate alcohol cessation plans based on the patient's physical condition to minimise any adverse effects on treatment due to possible withdrawal syndromes.

However, It is not uncommon that people with alcohol use disorders tend to deny their drinking habits due to unconscious defence mechanisms [5]. Failure to identify potential alcohol addiction and patient refusal to acknowledge alcohol consumption can lead to severe consequences, such as misdiagnosis, adverse drug interactions, or reduced treatment efficacy. This concealment not only affects individual health but also places trouble on healthcare professionals. Ultimately impacts the doctor's treatment decisions, which in turn affects the patient's health status and treatment outcomes. Hence, identifying patients with alcohol consumption habits accurately becomes a great challenge in the healthcare field. If we can predict whether a patient belongs to the category of long-term alcohol users based on their physical indicators, this could greatly assist doctors in formulating personalised treatment plans and offering more effective medical support.

The aim of this project is to develop a model that utilises medical data to accurately identify patients with alcohol consumption habits. This could provide healthcare professionals with a valuable tool for identifying patients who engage in alcohol consumption, thereby preventing patient concealment. In addition, this will enable doctors to formulate treatment strategies and intervention measures more effectively, which minimises medical errors.

2 Data Preprocessing

Data preprocessing is one of the important processes in the data science framework. The purpose of this step is to clean and transform raw data into a specified format that can be used to build models that generate efficient and accurate output. In this project, six data preprocessing techniques were explored, but only five of them were implemented. The following paragraph will provide a thorough justification to clarify this disparity.

2.1 Remove Unnecessary Label

The dataset contains two labels, whether or not they are smokers and whether or not they are drinkers. Since the aim of this project is to evaluate whether a patient is a drinker, the smoker label is removed to lower the dimensionality of the data. As a note, checking for missing and null values had been performed. Fortunately, there were no null or missing values in the dataset, obviating the necessity for removal or the use of imputation methods.

2.2 Label Encoding

Moving on, the categorical features in the dataset such as "sex" and "DRK_YN" were converted into numerical format using the Label Encoder function from scikit-learn library. For example, "Not Drinker" is encoded as 0 and "Drinker" as 1. The purpose of implementing this method is to ensure that the format is capable of fitting into the machine learning algorithms that are intended to be used to build the best model. Most machine learning algorithms are based on mathematical equations or statistics. Hence, it would be a good practice to convert categorical features into numerical features [6].

2.3 Outlier Detection

Next, the definition of an outlier is an observation or data point that is significantly different compared to the whole dataset. Outliers will have an impact on certain statistical analysis and machine learning models. Therefore, the way to deal with outliers needs to be seriously considered. In this project, the decision made to remove the outliers will be the best solution to increase the accuracy and efficiency of model performance. The statistical and visualisation technique used is boxplot. By plotting out the graph, the data points which are out of the upper and lower boundaries are identified as outliers, and they were removed accordingly [7]. The formula is as follows:

$$[Q1 - 1.5IQR, Q3 + 1.5IQR]$$

In this project, the first quartile is 0.1 and the third quartile is 0.9. Eliminating these outliers aims to enhance the robustness and generalisation of our models.

2.4 Data Standardization

Data standardisation is important to ensure that the data points are being evaluated and used in a standard range and measured in the same units [8]. Without this step, there will be several challenges in creating an effective model. For example, when the data points consist of a broad range of values, the result will be highly skewed toward the features with higher values, and hence generating a false conclusion. Z-score standardisation has been implemented as the dataset used in this project is continuous data. The formula to calculate the z-score is subtracting the mean and dividing by the standard deviation for each value of each feature as shown below:

$$Z\text{-}score = (value - mean)/(standard\ deviation)$$

2.5 Feature Correlation Analysis

Feature Correlation Analysis is a statistical technique to explore the relationship between the features. This technique is implemented and it is found that features (as shown in the graph below) such as age, hearing capability, and cholesterol level have a negative relationship with whether the person is a drinker or not. In spite of that, these negatively correlated features were not removed. This is because the experimentation revealed that removing the negatively correlated features negatively impacted model performance. To prioritise overall model performance, these features were retained in our analysis.

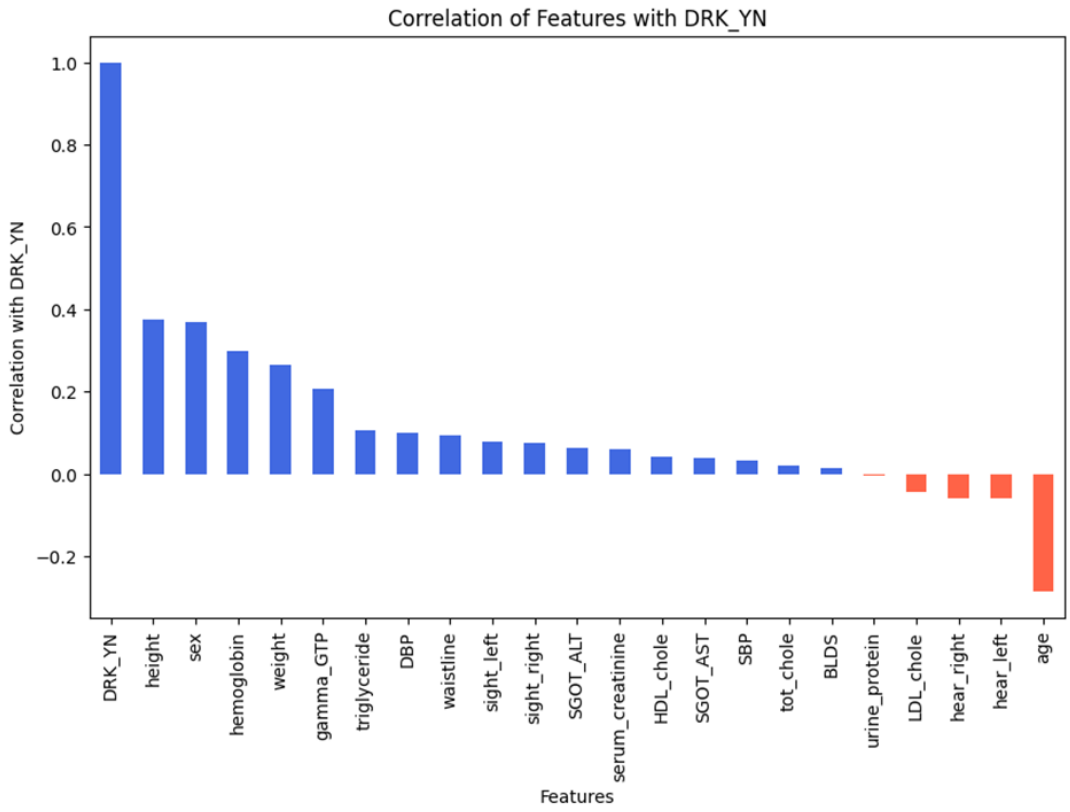

Fig1. Correlation of Features with Label "DRK_YN"

2.6 Train-Test Split

Lastly, the dataset is split into 70 percent training dataset and 30 percent training dataset. The purpose of this procedure is to assess the performance of models and eventually choose the best performance models to be distributed. The reason that 70 percent of the dataset is used as the training dataset is after careful consideration to ensure that it can represent all the data while not overfitting the model. This step emulates real-world scenarios where new and unseen data are being generated continuously.

2.7 Summary

In short, the preparation procedures for the data outlined in this paper have prepared the groundwork for identifying drinking patterns. The quality and reliability of the models created depend on these preprocessing activities, which also set the foundation for the later phases of model development, evaluation, and deployment.

3 Classification Models

3.1 Logistic Regression Model

3.1.1 Modelling and Hyperparameter Tuning

Logistic Regression is a widely used classification method, and we have selected it as our modelling approach due to its simplicity and interpretability.

To improve its performance, we conducted hyperparameter tuning using GridSearchCV from the scikit-learn library. We focused on two crucial hyperparameters: "C" and "max_iter," which significantly influence the Logistic Regression model's behaviour.

Here is how we performed hyperparameter tuning for our Logistic Regression model:

- "C": This hyperparameter controls the regularisation strength. We considered candidate values of 0.1, 1.0, and 10.0 to evaluate their impact on model performance.
- "max_iter": This parameter defines the maximum number of iterations for the solver to converge. We tested candidate values of 100, 200, and 300 to optimise the model.

We used accuracy as the scoring metric to assess and compare the Logistic Regression model's performance across different hyperparameter combinations.

3.1.2 Logistic Regression Model Assessment

After the hyperparameter tuning process, we identified the optimal configuration for our Logistic Regression model. By setting "C" to 10.0 and "max_iter" to 100, we achieved an improvement in model performance. This configuration resulted in an accuracy score of 0.7203 and an F1 score of 0.7115.

3.2 Decision Tree

3.2.1 Modelling and Hyperparameter Tuning

Decision tree is a powerful machine learning algorithm which is commonly used to solve classification problems, fitting the requirements of this project.

To enhance its performance, we employed the GridSearchCV to conduct hyperparameter tuning. We concentrated on two crucial hyperparameters, namely "criterion" and "max_depth," which significantly influence the behaviour of our Decision Tree model. We assessed various candidate values to determine the optimal configuration:

- "criterion": It determines the function used to measure the quality of a split. We

evaluated the 'gini' and "entropy" criteria.

- "max_depth": This hyperparameter controls the maximum depth of the decision tree, affecting the complexity of the model. We explored three candidate values: 10, 20, and 30.

In addition, we utilised accuracy as the scoring metric for comparing the Decision Tree model's performance while varying the "n_estimators" values.

3.2.2 Decision Tree Model Assessment

Following the hyperparameter tuning process, we pinpointed the best configuration for our Decision Tree model. By setting "criterion" to "gini" and "max_depth" to 10, we observed a significant improvement in model performance. This configuration resulted in an accuracy score of 0.7296 on the test data.

3.3 Random Forest

3.3.1 Modelling and Hyperparameter Tuning

Random forest is an integrated learning method and suitable for analysing classification problems. Moreover, random forest usually performs well with high dimensional data as it can handle a large number of features without the need for feature selection.Thus, random forest is suitable for modelling the dataset studied in our project.

To improve the performance of our Random Forest model, we conducted hyperparameter tuning using GridSearchCV and focused on the crucial hyperparameter "n_estimators," which controls the number of decision trees in the ensemble.

- "n_estimators": It corresponds to the number of decision trees in the Random Forest. Candidate values of 50, 100, 150, and 200 were set for comparison.

Furthermore, we selected accuracy as the evaluation metric to assess the performance of the Random Forest model across various "n_estimators" values.

3.3.2 Random Forest Model Assessment

After the hyperparameter tuning process, we identified the optimal configuration for our Random Forest model. By setting "n_estimators" to 200, we achieved an improvement in model performance. This configuration resulted in an accuracy score of 0.7245 and an F1 score of 0.7205.

3.4 LightGBM

3.4.1 Modelling and Hyperparameter Tuning

LightGBM (Light Gradient Boosting Machine) is an ensemble method that is well-suited for binary classification problems. It combines multiple weak classifiers to create a strong classifier. Moreover, it uses a gradient boosting algorithm and performs well with large-scale datasets because its efficiency and memory-friendliness make it possible to train models quickly. Considering the characteristics of our dataset, LightGBM proves to be an apt selection for our modelling.

In order to improve the performance of our LightGBM model, we used GridSearchCV to conduct hyperparameter tuning. We focused on two crucial hyperparameters, namely "n_estimators" and "learning_rate," which play significant roles in shaping the behaviour of our LightGBM model, and we conducted comparisons of various candidate values:

- "n_estimators": It corresponds to the number of boosting rounds performed by LightGBM. Candidate values of 175, 200, and 225 were set for comparison.
- "learning_rate": It is used to regulate the contribution of each weak classifier in the ensemble. Candidate values of 0.075, 0.1, and 0.125 were set for comparison.

We utilised accuracy as the scoring metric to assess the LightGBM model's performance across various parameter configurations.

3.4.2 LightGBM Model Assessment

After the hyperparameter tuning process, we identified the optimal configuration for our LightGBM model. By setting "n_estimators" to 225 and "learning_rate" to 0.1, we achieved an improvement in model performance. This configuration resulted in a test accuracy score of 0.7297.

By refining the hyperparameters, we optimised the LightGBM model, allowing it to reach its maximum potential on our dataset.

3.5 AdaBoost Model

3.5.1 Modelling and Hyperparameter Tuning

AdaBoost (Adaptive Boosting) is an ensemble method that can be very effective in dealing with binary classification problems by combining multiple weak classifiers to create a strong classifier [9]. Given the nature of our dataset, AdaBoost is a suitable choice for our modelling.

In order to make the performance better, we used the GridSearchCV imported from scikit-

learn library to conduct hyperparameter tuning. We focused on three crucial hyperparameters, namely "max_depth", "n_estimators" and "learning_rate", which play significant roles in shaping the behaviour of our AdaBoost model, and we conducted comparisons of various candidate values:

- "max_depth": It is used to control the maximum depth of the base estimator and thus the complexity of a single weak classifier. Candidate values of 5, 10, and 15 were set for comparison.
- "n_estimators": It corresponds to the number of boosting rounds performed by AdaBoost. Candidate values of 25, 50 and 75 were set for comparison.
- "learning_rate": It is used to regulate the contribution of each weak classifier in the ensemble. Candidate values of 0.1, 0.5 and 1.0 were set for comparison.

Additionally, accuracy was set as the scoring metric to compare the performance of the AdaBoost model within different values.

3.5.2 AdaBoost Model Assessment

After the hyperparameter tuning process, we identified the optimal configuration for our AdaBoost model. By setting "max_depth" to 5, "n_estimators" to 75, and "learning_rate" to 0.1, we achieved an improvement in model performance. This configuration resulted in an accuracy score of 0.7281 and an F1 score of 0.7265.

Through hyperparameter adjustments, we fine-tuned the AdaBoost model, ensuring it could perform its full potential on our dataset.

3.6 CatBoost Model

3.6.1 Modelling and Hyperparameter Tuning

CatBoost (Categorical Boosting) is a gradient boosting algorithm on decision trees that can achieve high performance on classification problems. By building a balanced tree structure, CatBoost not only runs efficiently on CPUs, but also reduces prediction time and the risk of overfitting. This balance enables CatBoost to effectively handle large-scale classification problems, ensuring high performance while maintaining model stability [10].

We conducted hyperparameters on "depth" which is to control the maximum depth of a decision tree, "iterations" which is used to determine the number of iterations or boosting rounds, and "learning_rate" which controls the weighting of each new model during each iteration. We compared different candidate values within three hyperparameters:

- "depth": Candidate values of 4, 6 and 8 were set for comparison.

- "iterations": Candidate values of 100 and 500 were set for comparison.
- "learning_rate": Candidate values of 0.01, 0.03, 0.05 and 0.07 were set for comparison.

To determine the optimal hyperparameter values, we employed accuracy as the scoring criterion within CatBoost.

3.6.2 CatBoost Model Assessment

After tuning hyperparameters, we found that setting "depth" to 8, "iterations" to 500, and "learning_rate" to 0.07 resulted in the best F1 score so far, and the best accuracy which was over 0.73.

To gain insights into feature importance within our dataset, we utilised CatBoost's "get_feature_importance" functionality. We visualised the results, as shown in Figure 2. According to the plot, it can be seen that sex and age had the highest importance scores among the 22 features.

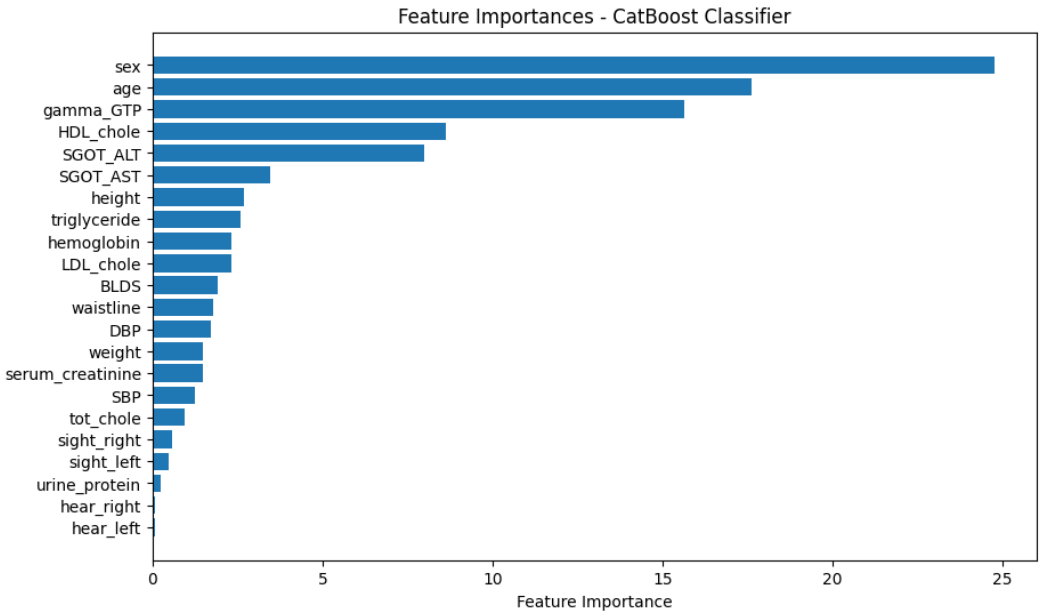

Fig2. Feature Importances within CatBoost Classifier

3.7 Model Performance Evaluation

3.7.1 Cross Validation

Cross-Validation is a significant technique for assessing the robustness and reliability of our models [11]. Therefore, to evaluate the performance of each model, we used Cross-Validation to assess by comparing accuracy scores. We gained some valuable insights into how well our models generalise to unseen data. This process helps us to determine the consistency of our models' performance across different subsets of datasets.

The results gained from Cross-Validation showed that the CatBoost model achieved the highest accuracy, approximately 0.7304, followed by the LGBM model and the Decision Tree model, both at around 0.7294. The AdaBoost model demonstrated an accuracy of 0.7276, while the Random Forest and Logistic Regression models showed accuracy values of 0.7249 and 0.7205, respectively.

3.7.2 ROC Curve

The Receiver Operating Characteristic curve, short for ROC curve, is also an essential tool for model evaluation, particularly in binary classification tasks [12]. The area above the curve represents the True Positive Rate or Recall Rate, which indicates the percentage of positive examples that model successfully identifies, while the area below the curve indicates the percentage of negative examples that the model incorrectly classifies as positive. To assess the effectiveness of our model in distinguishing between positive and negative examples, we also drew a ROC curve for all our models. The results can be seen in Figure 3 below.

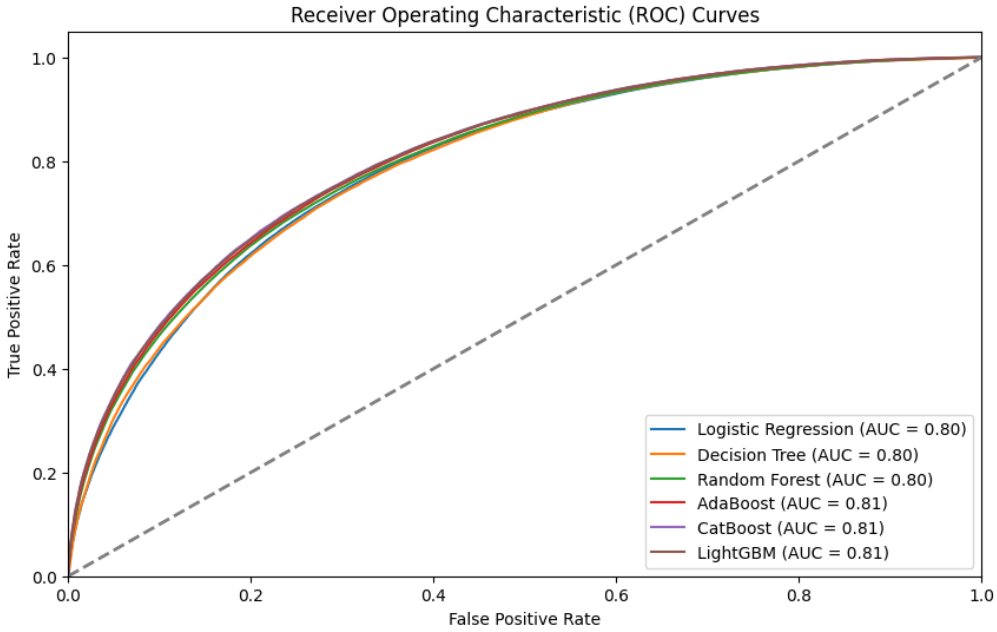

Fig3. ROC Curve Plot of All Models

In this plot, it can be seen that the AUC scores of the Logistic Regression, Decision Tree and Random Forest models' AUC scores are all at 0.80, while the other three models achieve AUC scores of 0.81. And all models perform as well as

3.7.3 Conclusion

Table 1 summarises the performance metrics of different models in our binary classification task.

Table 1. Model Performance Metrics

Model	Accuracy	F1 Score	ROC
Logistic Regression	0.7205	0.7115	0.80
Decision Tree	0.7294	0.7175	0.80
Random Forest	0.7249	0.7205	0.80
LightGBM	0.7294	0.7264	0.81
AdaBoost	0.7276	0.7265	0.81
CatBoost	0.7304	0.7269	0.81

Based on the results and analysis, we concluded that the CatBoost, LightGBM, and Adaboost models stand out as the top performers for our binary classification task, since they consistently demonstrated high accuracy and strong discriminative abilities.

3.8 Majority Voting

In order to improve the performance and robustness of our models for our dataset, we also considered the concept of Majority Voting. It can put multiple better performing models together and combine their prediction results to make more informed decisions.

Therefore, based on the results of model evaluation, we made a combination of AdaBoost, CatBoost and LGBM. After that, we gained a well accuracy score of 0.7303, which is similar to the score obtained using the CatBoost model alone.

5 Limitation and Further Discussion

While the classification model demonstrates strong predictive capabilities for individuals with alcohol consumption, this project does have limitations. The presence of outliers in our dataset is apparent. Notably, when examining the 'waistline' column through a boxplot, with an outlier which particularly stands out was 999. This value is often attributed to input errors. In contrast, when scrutinising the 'gamma_GTP' column, an abundance of outliers, exceeding ten thousand, are observed outside the interquartile range. Although these outliers account for only approximately 1% of our entire dataset, their removal may result in the loss of significant information. To avoid the loss of too much valuable data, we customised Q1 and Q3 as the 10th and 90th quartile respectively during the outlier removal outside [Q1-1.5IQR, Q3+1.5IQR]. Including a wide range of data points keeps some of the extreme values that may still be valid data points.

Moreover, potential bias presents another limitation. Since a single dataset originating from National Health Insurance Service in Korea was used, the clinical data might introduce biases. The models implemented, therefore, might become less applicable and generalizable to individuals from other countries with varying racial backgrounds, dietary habits and genetic profiles. For example, healthy blood pressure levels may differ slightly between countries and recommended cholesterol levels could be impacted by dietary and genetic factors. To further enhance the generalizability of the models on new unseen data, clinical data from diverse international sources could be incorporated.

The curse of dimensionality has also been a challenge. The dataset contains 23 features in total which posed several obstacles in computation and memory usage. After applying PCA to determine the optimal number of features needed to minimise explained variance, we retained 22 features. Our data analysis and processing then occurred in a 22-dimensional space and hence the escalation of computational complexity and memory intensiveness of the algorithms. Fortunately, despite these challenges, other potential problems like data sparsity and overfitting were not exhibited in our model implementations. Our approaches effectively maintained the quality of data and the performance of the models while mitigating some problems associated with high dimensionality.

6 Conclusion

Intense alcohol consumption has been a global public health problem that leads to a wide range of negative health influences. This project aims at developing a model that leverages medical data for the accurate identification of patients with alcohol consumption habits. Our goal was successfully achieved through data analysis and machine learning techniques. The high quality of the dataset played a pivotal role in this achievement, as no missing values were observed. This absence of missing data not only streamlined the initial data preprocessing, but also laid a strong foundation for the subsequent development of our models. It created significant potential and offered numerous benefits for the success and reliability of our model development.

As for model development, it was noticeable that ensemble methods and boosting such as Catboost and Majority Voting have more robust performance on our dataset. They effectively handled complex relationships of data points and adapted well to high dimensional space, which in turn fostered the accuracy of the model. By using permutation importance in Catboost, we reckoned that age, sex, and clinical indicators like HDL cholesterol, gamma-glutamyl transferase level, and alanine transaminase level are features that contributed relatively heavier to the final prediction of drinkers. These clinical indicators play a crucial role in assessing the overall health of the liver and heart, serving as valuable markers for identifying individuals who may be at risk of excessive alcohol consumption. In conclusion, the classification models serve the purpose to effectively predict whether a patient belongs to the category of long-term alcohol abusers based on their physical indicators. Not only could this offer healthcare professionals with more effective medical support, thereby reducing the occurrence of medical errors, but it could also greatly assist doctors in formulating personalised treatment plans.

References

[1] Hannah Ritchie and Max Roser (2018). *"Alcohol Consumption".* Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/alcohol-consumption'

[2] World Health Organization (2022). *Alcohol.* [online] World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/alcohol.

[3] Pietrangelo, A. (2020). *Why You Absolutely Shouldn't Drink Alcohol Before Surgery*. [online] Healthline. Available at: https://www.healthline.com/health/alcohol-before-surgery

[4] Jonathan M. Dolman, Neil D. Hawkes (2005). *Combining the audit questionnaire and biochemical markers to assess alcohol use and risk of alcohol withdrawal in medical inpatients,* Alcohol and Alcoholism, Volume 40, Issue 6, Pages 515–519, https://doi.org/10.1093/alcalc/agh189

[5] N. Saya Des Marais. (2022). *How to help an alcoholic in denial*. GoodRx. https://www.goodrx.com/conditions/alcohol-use-disorder/how-to-help-alcoholic-in-denial

[6] Verma, Y. (2021). *A Complete Guide to Categorical Data Encoding.* [online] Analytics India Magazine. Available at: https://analyticsindiamag.com/a-complete-guide-to-categorical-data-encoding/.

[7] Dhadse, A. (2021). *Removing Outliers. Understanding How and What behind the Magic.* [online] Analytics Vidhya. Available at: https://medium.com/analytics-vidhya/removing-outliers-understanding-how-and-what-behind-the-magic-18a78ab480ff.

[8] Jaadi, Z. (2019). *When and Why to Standardize Your Data?* [online] Built In. Available at: https://builtin.com/data-science/when-and-why-standardize-your-data.

[9] Devin S. (2022). *Adaboost. SERP AI.* [online] Medium. Available at: https://medium.com/serpdotai/adaboost-eea9ad18dbf3

[10] Artem O. (2023). *What is CatBoost?* [online] Builtin. Available at: https://builtin.com/machine-learning/catboost

[11] Chirag G. (2021). *Importance of Cross Validation: Are Evaluation Metrics enough?* [online] Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/05/importance-of-cross-validation-are-evaluation-metrics-enough/

[12] Nahm FS. (2022). *Receiver operating characteristic curve: overview and practical use for clinicians.* Korean J Anesthesiol. Feb;75(1):25-36. doi: 10.4097/kja.21209. Epub 2022 Jan 18. PMID: 35124947; PMCID: PMC8831439.

Appendix

Appendix A - Code And Dataset

File Types	Name	Description
Code	DATA7703_model_modified.ipynb	It serves as a detailed log of the step-by-step processes involved in data preprocessing, feature engineering, the application of various machine learning models, hyperparameter tuning, and visualisation.
Dataset	smoking_driking_dataset_Ver01.csv	Initial dataset downloaded directly from Kaggle https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset