Predicting Alcoholic Status Using a Person's Vitals

Maeve Horan-Portelance, John Wu, Justin Noche, Kanchan Raju, Shashvat Patel

STATS 101C

Lecture 2, Group 7

Professor Almohalwas

17 December 2023

I. Abstract

The objective of this project was to develop a machine learning model that best predicts an individual's alcoholic status by using their vital health statistics. The data set used consists of both categorical and numeric vital predictors, with a training set of length 70,000 and a testing set of length 30,000. This report details the process of developing our optimal model including exploratory data analysis, variable selection, missing value imputation, model selection, and conclusions and limitations. The final model our group arrived at was an XGBoost model using hmisc imputation comprising all variables from the original data set. Our final score on Kaggle was 0.7315, and our rank in the class leaderboard was 19th.

II. Introduction

Alcohol consumption is a pervasive aspect of our socio-cultural framework, with significant implications for an individual's health and well-being. The ability to accurately predict an individual's alcoholic status is of paramount importance, as it enables targeted interventions, personalized healthcare, and the formulation of public health strategies. Furthermore, every human is biologically different and therefore breaks down alcohol differently, so it is inappropriate to apply just one standardized metric for alcoholic status assessment. Excessive alcohol consumption has been linked to a myriad of health issues, including liver diseases, cardiovascular problems, and mental health disorders. In fact, the World Health Organization (WHO) reports that binge drinking is a leading cause of 200+ diseases and injuries (2022). Traditional approaches to assessing alcohol consumption often rely on self-reported data, which can be subject to biases and inaccuracies. However in recent years, the integration of machine learning techniques into healthcare research has shown promise in

enhancing our understanding and predictive capabilities in various medical domains. Machine learning models provide the advantage of offering a data-driven approach that can leverage a diverse set of vital health predictors to provide more accurate and reliable predictions.

This paper addresses the crucial task of predicting alcoholic status using a dataset collected by the National Health Insurance Service in Korea, comprising 26 predictor variables (i.e. age, weight, hemoglobin levels, smoking status) and 1 response variable (alcoholic status). As we delve into the intricacies of our approach and present our findings, it becomes evident that accurate prediction of alcoholic status through machine learning has significant implications for public health and individual well-being. By advancing our understanding of the relationships between vital health statistics and alcohol consumption, this research contributes to the ongoing efforts to harness the power of data-driven methodologies in healthcare, paving the way for more targeted and effective interventions in the realm of alcohol-related health issues and beyond.

III. Exploratory Data Analysis

In order to gain a better understanding of the original data set, we performed an exploratory analysis of the 26 predictor variables to get a glimpse into which could have the most significant impact on alcoholic status. We utilized the ggplot2 library in R-studio to create visualizations of individual predictors.

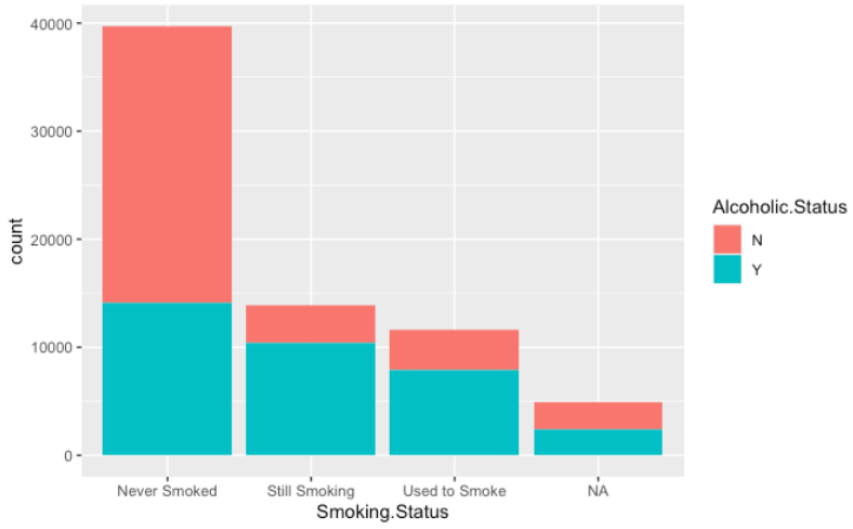


Figure 1: Stacked bar-chart of Alcoholic.Status vs. Smoking.Status.

Figure 1 reveals a significant correlation between smoking status and alcoholic status. Those who have never smoked appear significantly less likely to have an alcoholic status of "N". Furthermore, individuals who are still smoking or used to smoke are more likely to have an alcoholic status of "Y".

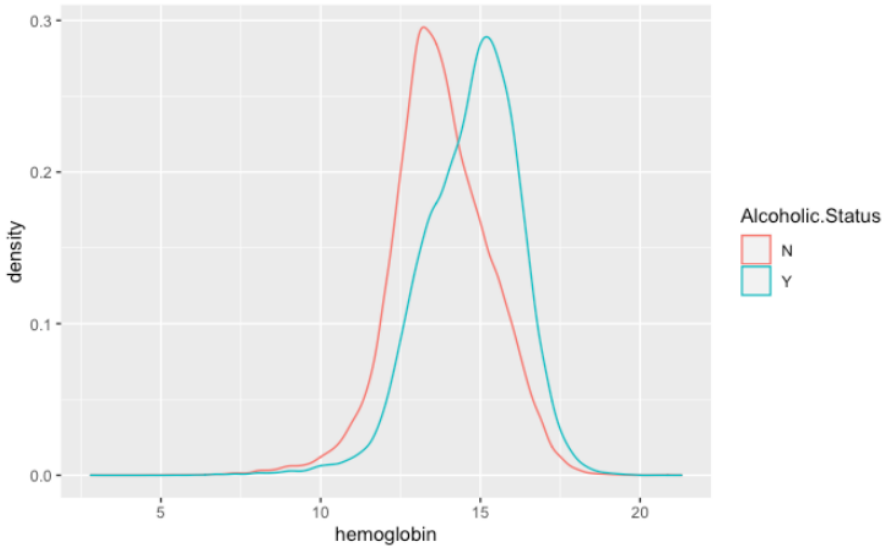


Figure 2: Density plot of Alcoholic.Status vs. hemoglobin.

According to Figure 2, individuals with an alcoholic status of "Y" tend to have higher hemoglobin levels than individuals with an alcoholic status of "N," which indicates that hemoglobin could be a potentially strong predictor of alcoholic status.

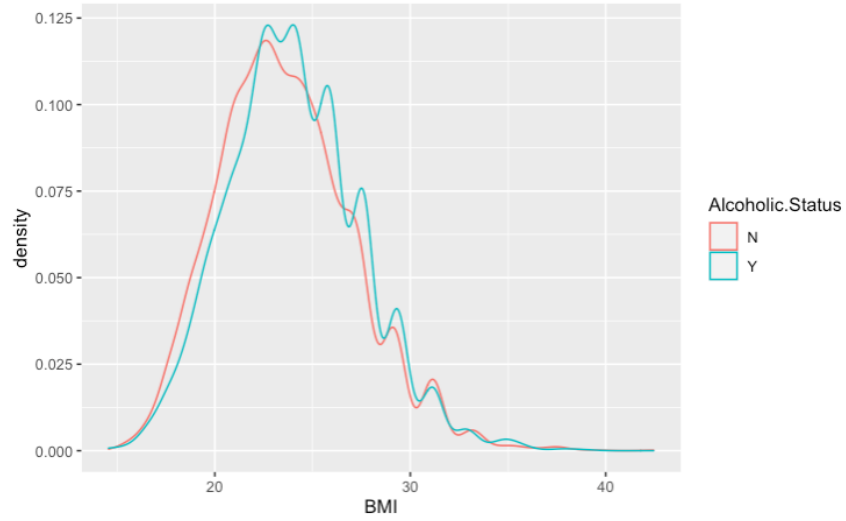


Figure 3: Density plot of Alcoholic.Status vs. BMI

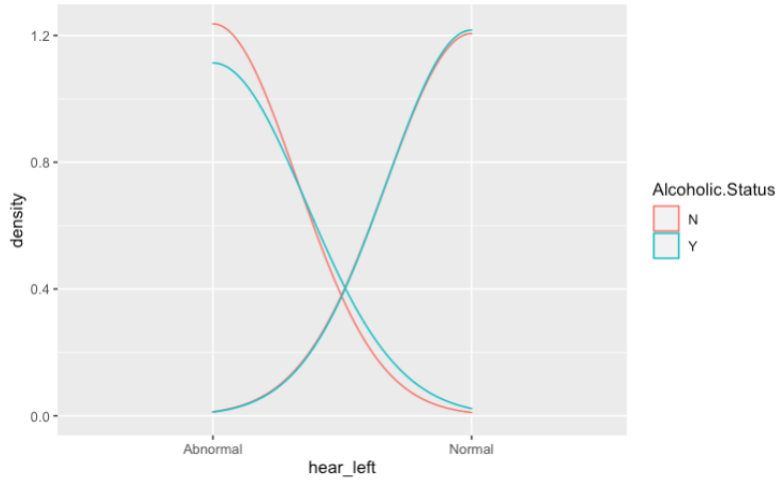


Figure 4: Density plot of Alcoholic.Status vs. hear_left

Figures 3 and 4 provide examples of variables which may not be significant predictors of alcoholic status. Visually, the trends in these variables do not appear to differ significantly between individuals who are considered alcoholics and those who do not, which indicates that these variables may not be extremely favorable while constructing the final model.

IV. Missing Value Imputation

A crucial step in the model-building process was the imputation of missing values. NA imputation is a necessary element of data-cleaning for several reasons: maintaining data integrity, enabling proper model performance, preserving relationships between predictors, maintaining sample size, improving predictive power, and more. The training and testing sets contained 130,776 and 56,035 missing values, respectively. A breakdown of missing values by variable can be found below:

Figure 5: Count of NA values by variable

sex	age	height	weight	waistline	sight_left	sight_right	hear_left	hear_right
4962	4877	4941	4972	4940	4877	4900	4833	4887
SBP	**DBP**	**BLDS**	**tot_chole**	**HDL_chole**	**LDL_chole**	**triglyceride**	**hemoglobin**	**urine_protein**
4919	4895	4821	4864	4816	4914	4877	4961	4899
Serum_ creatinine	**SGOT_ AST**	**SGOT_ ALT**	**Gamma_ GTP**	**BMI**	**BMI.Category**	**AGE.Category**	**Smoking. Status**	**Alcoholic.Status**
4847	4887	4893	4961	4967	4874	8313	4879	–

Figure 6: Percentage of NA values by variable

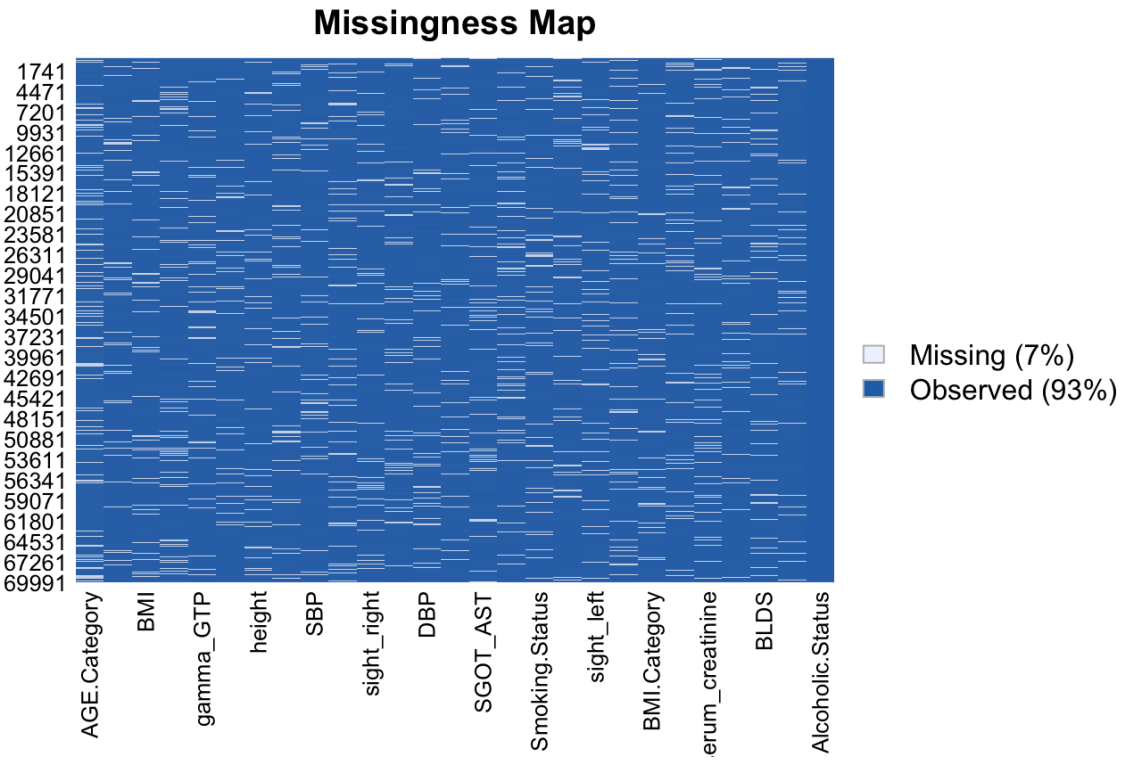
sex	age	height	weight	waistline	sight_left	sight_right	hear_left	hear_right
7.08%	6.97%	7.06%	7.10%	7.06%	6.97%	7.00%	6.90%	6.98%
SBP	**DBP**	**BLDS**	**tot_chole**	**HDL_chole**	**LDL_chole**	**triglyceride**	**hemoglobin**	**urine_protein**
7.03%	6.99%	6.88%	6.95%	6.88%	7.02%	6.97%	7.09%	6.99%
Serum_ creatinine	**SGOT_ AST**	**SGOT_ ALT**	**Gamma_ GTP**	**BMI**	**BMI.Category**	**AGE.Category**	**Smoking. Status**	**Alcoholic.Status**
6.92%	6.98%	6.99%	7.09%	7.10%	6.83%	11.87%	6.97%	–

The majority of variables had approximately 7% of values missing, with the exception of AGE.Category which had a significantly larger proportion of missing values. According to industry practices for missing value imputation, the typically accepted maximum threshold for missing values in a large dataset is 5%. Because all predictors exceeded this threshold, missing value imputation is especially imperative for this project.

In addition to discovering what proportion of values in our dataset were missing, it was also necessary to determine whether these values were missing in a random or systematic manner. The desirable case for missing value imputation is that the values are MCAR, or Missing Completely at Random. However, often real world data can be MNAR, Missing Not at

Random, in which case further investigation is required on the data gathering process. For this reason, we created a plot of missingness for the training dataset:

Figure 7: Missingness plot of predictor variables in training dataset



The missingness plot did not reveal any noticeable patterns in the data values which were absent. Therefore, we could proceed with the assumption that all missing values were in fact MCAR, which allowed us to employ the following missing value imputation methods for missing-at-random data:

Imputation Methods

i. 'MICE' Package

MICE, which stands for Multivariate Imputation by Chained Equations, provides advanced features for missing value treatment. It uses the mice() function to build an imputed and the complete() function to generate the completed data and provide multiple copies of the

data frame with a variety of imputations. Furthermore, MICE has the ability to choose and apply a unique imputation model to each predictor variable.

ii. 'Hmisc' Package

Hmisc, which stands for Harrell Miscellaneous, consists of two functions for powerful imputation. impute() imputes NAs using a specified statistical function (such as mean, median, or mode). aregImpute() performs mean imputation using additive regression, bootstrapping, and predictive mean matching.

iii. 'Amelia' Package

Named after Amelia Earhart, the first female aviator to fly solo across the Atlantic Ocean, this method utilizes multiple imputation similar to the MICE package. In addition to the MCAR assumption, it also assumes that all variables have a multivariate normal distribution so that it can use means and covariances to summarize the imputed data.

For each imputation method, multiple datasets were constructed utilizing the many different outputs given from the R code. Between the 10 MICE, 10, Hmisc, and 10 Amelia datasets, the best-performing dataset was Hmisc 6. However, for different models, different datasets/imputation methods performed better, meaning one method is not universally dominant over the other.

V. Model Selection

In our modeling process, we systematically explored a variety of regression and classification techniques to identify the most effective approach for predicting an individual's alcoholic status. The considered techniques included K Nearest Neighbors, Logistic Regression GLM, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Tree Model, Random

Forest, Boosting, and Support Vector Machine (SVM). This comprehensive approach allowed us to tailor our modeling strategy to the unique characteristics of our dataset and the specific demands of the predictive task.

Determining the size and scope of our models involved exploring different subsets of predictors and evaluating their impact on model performance. Constructing various models using the training set, we employed the chosen classification technique and imputation method to develop models that could effectively handle both numerical and categorical predictor variables. The models constructed were then applied to the testing set to assess their predictive performance on new, unseen data. Subsequently, we quantified and compared the performance of our models by submitting the predictions to Kaggle, obtaining accuracy scores for evaluation.

i. Random Forest

The Random Forest algorithm, implemented in the "randomForest" package in R, is a learning method that builds an ensemble of decision trees during the training phase, with each decision tree trained on a bootstrapped sample of the original dataset (sampling with replacement). At each node of the tree, a random subset of predictors is considered for splitting, and each decision tree predicts the alcoholic status of an observation based on the majority class in the terminal node to which the observation belongs. We used cross-validation techniques to determine the optimal tuning parameters, "ntree" and "mtry," and our best-performing random forest model used "ntree" = 500, and "mtry" = default (the square root of the total number of predictors, or 5.099). Our best-performing random forest model, using hmisc-imputed data, scored a 0.72943 on Kaggle.

ii. Logistic Regression

Logistic regression models the relationship between a binary dependent variable and one or more independent predictor variables by use of the logistic function, mapping the linear combination of input features to a value between 0 and 1. A threshold is then applied to the predicted probabilities–observations with probabilities above 0.5 are classified as "Yes" for alcoholic status, and observations with probabilities below 0.5 are classified as "No." Our best-performing logistic regression model, using hmisc-imputed data, scored a 0.7236 on Kaggle.

iii. XGBoost

XGBoost, or Extreme Gradient Boosting, is an efficient and scalable implementation of gradient boosting machines, ensemble learning methods used for classification. XGBoost builds an ensemble of decision trees sequentially, with each subsequent tree focusing on the samples that were misclassified by the previous tree. It utilizes gradient boosting to minimize the classification error, incorporates regularization techniques to control overfitting by penalizing complexity, and uses shrinkage to scale each tree's contribution. Our best-performing XGBoost submission, using the hmisc-imputed data, scored a 0.7315 on Kaggle.

iv. Others

In addition to Random Forest, Logistic Regression, and XGBoost, we also attempted K Nearest Neighbors, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Tree Model, and Support Vector Machine (SVM). Although each of these models utilized a different method to predict alcoholic status, none performed as well as XGBoost, which emerged as the strongest performer. It incorporates regularization, handles missing data effectively, provides feature importance scores, and is computationally efficient, making it suitable for both categorization and regression tasks.

The final predictor variables in our highest-scoring submission comprised a combination of both categorical and numerical predictors, including sex, age, height, weight, waistline, sight_left, sight_right, hear_left, hear_right, SBP, DBP, BLDS, tot_chole, HDL_chole, LDL_chole, triglyceride, hemoglobin, urine_protein, serum_creatinine, SGOT_AST, SGOT_ALT, gamma_GTP, BMI, BMI.Category, AGE.Category, Smoking.Status, and Alcoholic.Status.

While our chosen XGBoost models, particularly with Hmisc imputation, demonstrated strong predictive performance with a Kaggle score of 0.7315, we acknowledge the potential for further improvement. Exploring specific predictor subsets and considering additional predictors not present in the original dataset presents opportunities for refinement and application in healthcare and public health interventions.

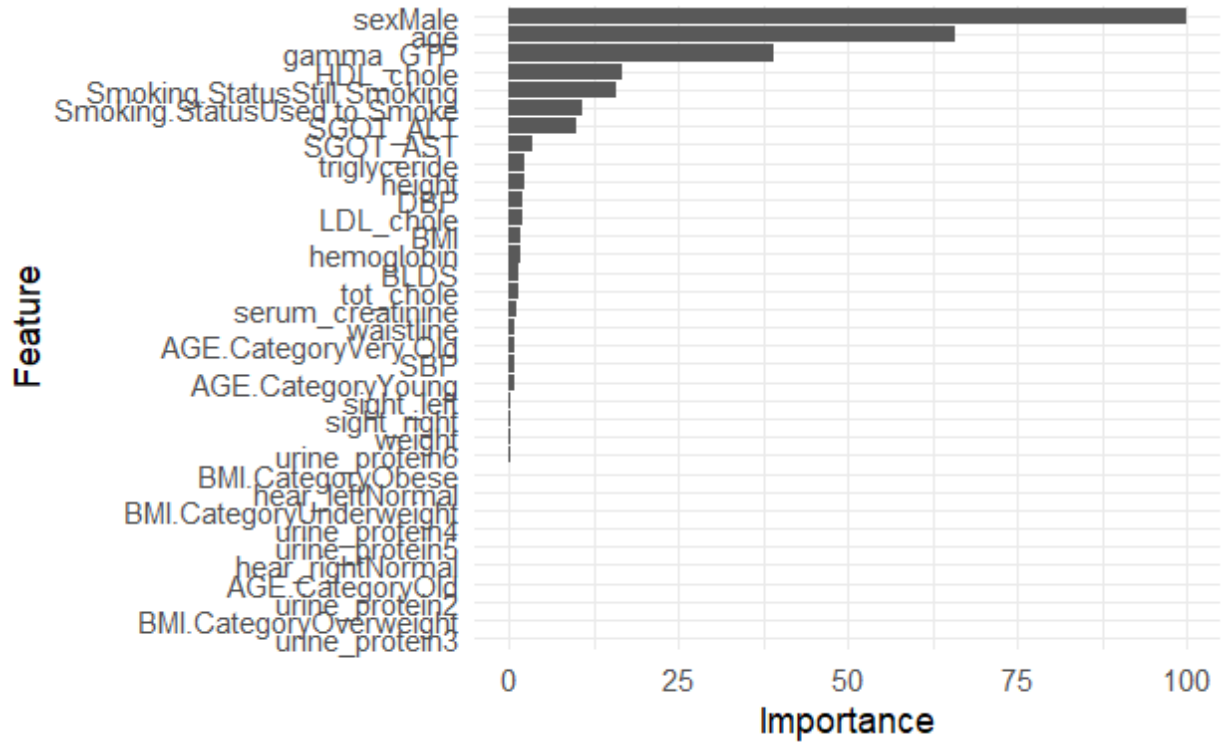
VI. Feature Importance

To assess further which variables had the greatest predictive value in our final model, we performed a variable importance analysis on the training dataset. Gradient boosting models assign importance scores based on the contribution of each feature to reducing the loss function (e.g., mean squared error). Higher scores indicate features that contribute more to the model's performance.

Figure 8: Chart of feature importance for the final model

Feature	Importance
sexMale	100.000000
age	65.7998550
gamma_GTP	39.1037312
HDL_chole	16.6306801
Smoking.StatusStill Smoking	15.9044323
Smoking.StatusUsed to Smoke	10.8375315
SGOT_ALT	9.8246254
SGOT_AST	3.4697910
triglyceride	2.1833472
height	2.1408432

Figure 9: Graph of feature importance for final model.



According to the feature importance statistics in Figures 8 and 9, the features with the greatest predictive power in our final model were: sex (Male), age, gamma_GTP, HDL_chole, Smoking.Status (Still Smoking and Used to Smoke), SGOT_ALT, and SGOT_AST.

VII. Conclusions and Limitations

To recall, our highest scoring model was an XGBoost using Hmisc imputation. We earned a public score of 0.7315 on kaggle and placed 19th on the class leaderboard. Although we were able to improve our model performance significantly throughout the course of the competition, we acknowledge that there is significant room for improvement in our model. For example, one major aspect to reconsider is that our highest scoring model utilized all 26 given predictors. There is further room for exploration regarding whether a specific subset of predictors could have created a stronger model, especially given that the results of our feature importance analysis identified a specific subset of predictors with much greater significance than the rest. Furthermore, there could also be additional predictors of alcoholic status not present in the original data set with the power to improve our model, for future exploration beyond the scope of this project. Another shortcoming is that with any imputation method, there are always risks and assumptions required. Ultimately, we cannot predict the exact behavior of the missing real world data values and thus every imputed model has its inherent limitations.

In conclusion, this research aimed to devise an effective machine learning model for predicting an individual's alcoholic status based on vital health statistics. Through a meticulous process encompassing exploratory data analysis, variable selection, missing value imputation, and model selection, our group identified one possible model for this data. This project not only

contributes to the growing body of literature on predictive modeling in health contexts, but it also highlights the iterative and dynamic nature of machine learning model development. As the scientific community continues to refine its understanding of the interplay between vital health statistics and alcoholic status, the potential for more accurate and reliable models emerges, paving the way for valuable applications in healthcare and public health interventions.

VIII. References

Almohalwas, Akram Mousa. “Predicting Alcoholic Status Using Person's Vitals Data Set for Predictive Analysis.” Dec. 17, 2023.

“Alcohol.” *World Health Organization*, World Health Organization, 2022, www.who.int/news-room/fact-sheets/detail/alcohol.