A Machine Learning Analysis of Impact of the Covid-19 Pandemic on Alcohol Consumption Habit Changes Among Healthcare Workers in the U.S.

Mostafa Rezapour (rezapom@wfu.edu)
Wake Forest University https://orcid.org/0000-0001-9569-118X

Research Article

Keywords: The Covid-19 Pandemic, Healthcare Workers, Mental Health, Machine Learning

Posted Date: May 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1651579/v1

A Machine Learning Analysis of Impact of the Covid-19 Pandemic on Alcohol Consumption Habit Changes Among Healthcare Workers in the U.S.

Mostafa Rezapour[1]

[1]Department of Mathematics and Statistics, Wake Forest University, Winston-Salem, North Carolina, United States.

Contributing authors: rezapom@wfu.edu;

Abstract

In this paper, we discuss the impact of the Covid-19 pandemic on alcohol consumption habit changes among healthcare workers in the United States. We utilize multiple supervised and unsupervised machine learning methods and models such as Decision Trees, Logistic Regression, Naive Bayes classifier, k-Nearest Neighbors, Support Vector Machines, Multilayer perceptron, Random Forests, XGBoost, CatBoost, LightGBM, Synthetic Minority Oversampling, Chi-Squared Test and mutual information method on a mental health survey data obtained from the University of Michigan Inter-University Consortium for Political and Social Research to find out relationships between COVID-19 related negative effects and alcohol consumption habit changes among healthcare workers. Our findings suggest that COVID-19-related school closures, COVID-19-related work schedule changes and COVID-related news exposure may lead to an increase in alcohol use among healthcare workers in the United States.

Keywords: The Covid-19 Pandemic, Healthcare Workers, Mental Health, Machine Learning

1 Introduction

COVID-19, caused by a virus named SARS-CoV-2, was first discovered in December 2019 in Wuhan, China [1]. The Covid-19 pandemic has caused an unparalleled health crisis around the world and has brought about a negative effect on the mental health of healthcare workers [2–4]. A rise in the rate of positive COVID-19 tests and the number of hospitalizations, reduction of proper personal protection equipment, working under severe pressures, and an increase in fears related to contracting the virus and transmitting it to others can contribute to a mental health decline among health care workers [5–7]. Many research articles have studied the impact of the Covid-19 pandemic on the mental health decline of healthcare staff [8–13]. In this paper, we apply several machine learning models and techniques to investigate relationships between COVID-19 related mental health decline of healthcare workers and their alcohol use habit changes.

Machine learning (ML) and artificial intelligence (AI) have been employed by various healthcare providers, scientists, and clinicians in medical industries in the fight against the novel disease. Researchers have utilized several advanced machine learning models and algorithms to tackle various issues related to the virus, and to better understand the pattern of viral spread. Kushwaha et al. [14] discusses how ML algorithms can be used to understand the nature of the virus and predict the upcoming related issues. Lalmuanawma et al. [15] reviews the importance of AI and ML in screening, predicting, forecasting, contact tracing, and drug development for SARS-CoV-2 and its related epidemic. Benvenuto et al. [16] discusses an autoregressive integrated moving average model that can predict the spread of COVID-19. Using several datasets of the COVID-19 outbreak inside and outside Wuhan, Kucharski et al. [17] introduces a model that can explore the possible viral spread outside Wuhan. Machine learning has been employed to reveal the negative impacts of the Covid-19 pandemic on the students' mental health decline [18]. Machine learning methods and statistical tests have also been utilized to investigate the relationship between the COVID-19 vaccines and boosters and the total case count for the Coronavirus across multiple states in the USA as well as the relationship between several, selected underlying health conditions with COVID-19 [19].

In this study, like [2], we use survey data obtained from the University of Michigan's Inter-university Consortium for Political and Social Research (ICPSR). The data was collected by Deirdre Conroy [20, 21] from the University of Michigan Department of Psychiatry and Cathy Goldstein from University of Michigan Department of Neurology. According to the ICPSR, "The rationale for this study was to examine whether sleep, mood, and health related behaviors might differ between healthcare workers who transitioned to conducting care from home and those who continued to report in-person to their respective hospitals or healthcare facilities." The original data contained 916 survey responses. The average survey response answered 94.5% of the survey questions, 29 questions in total, with many questions leaving room for

respondents to write how their mood or habits had changed since COVID-19 protocols were in effect. The data was stripped of any identifying information about the respondents and contained both categorical and numeric columns [22]. Rezapour and Hansen [2] consider Question 29a in the survey, which reads "Please tell us how your mood has changed?" as a target variable in their analysis. They utilize multiple multiclass classification machine learning models and techniques to find the most important factors (features) in predicting the mental health decline of healthcare workers. However, this paper focuses on Question 18a: "Please tell us how the amount of alcohol that you are consuming has change?" as the target variable. In this paper, we employ many supervised and unsupervised machine learning models and techniques to analyze the hidden effects of COVID-19 on alcohol consumption changes among healthcare workers.

The remainder of this paper is structured as follows: Section 2 discusses the methodology, describes the experimental framework used to find the top predictors of alcohol consumption habit changes among healthcare workers in the United States. Section 3 discusses and analyzes the top predictors of alcohol consumption habit changes among healthcare workers in the United States and compare the results with some other research articles. Finally, Section 4 concludes the paper by summarizing our overall findings.

2 Methods

The data utilized in this study was taken from the University of Michigan's Inter-University Consortium for Political and Social Research, and collected by Deirdre Conroy [20]. Conroy et al. [21] confirms that all experiments were performed in accordance with relevant guidelines and regulations (see the Methods section in [21]). The data is from a survey conducted within the University of Michigan Medical Center [21]. All survey data was collected in accordance with relevant guidelines and regulations. The survey was undertaken after clearing University of Michigan Institutional Review Board (HUM00180147) approval, at which point the Qualtrics survey link was sent via email listservs that would reach large numbers of health care providers [21]. It is noted that no compensation for participation was provided [21]. Additionally, all participants have provided full and informed consent by completing the survey [21].

Data availability: The data that support the findings of this study are available from the University of Michigan's Inter-University Consortium for Political and Social Research, which is collected by Deirdre Conroy [20, 21], at https://www.openicpsr.org/openicpsr/project/127081/version/V1/view?path=/openicpsr/127081/fcr:versions/V1&type=project [22].

Python codes availability: Python codes for the data preparation process and other supervised and unsupervised machine learning analysis are available at https://github.com/MostafaRezapour/Hidden-Effects-of-COVID-19-on-Healthcare-Workers-A-Machine-Learning-Analysis [41].

2.1 Computational Process

The COVID Isolation on Sleep and Health in Healthcare Workers data is a tabular dataset with 915 rows (datapoints or participants) and 64 columns (features or attributes) such as *StartDate, EndDate, IPAddress*, etc. [23]. The dataset contains 14678 missing values with 480 missing values for Question 28a, 311 missing values for Question 29a, and so on. To prepare the dataset for a machine learning analysis with Question 18a as the target variable, we first remove all the rows with no value (answer) in the column corresponding to Question 18a. We then remove unrelated columns such as *StartDate, EndDate, Status*, etc. reducing the number of missing values to 1612. Most of the variables in the dataset are categorical (nominal or ordinal), but since this is a COVID-related analysis, we would prefer not to use machine learning techniques or models (e.g. KNN) to treat the missing values for the categorical variables. Instead, we directly remove the rows with too many missing values from the dataset, and as the result, we end up with 273 clean rows (with no missing value). Since the mental health dataset contains categorical and ordinal variables, we first encode them to numbers, and we then use several encoding techniques such as one-hot-encoding or dummy variable encoding as well as several packages in Python such as *OneHotEncoder, LabelEncoder* or *OrdinalEncoder* from *sklearn.preprocessing* to prepare the dataset for analysis.

In this paper, our main goal is to use multiple supervised and unsupervised machine learning models and techniques to find the top predictors (features) of alcohol consumption habit changes among healthcare workers in the U.S. during the severe phase of Covid-19 pandemic. Only for supervised machine learning models with high accuracy (at least 95%), we discuss, export the feature importance scores and proceed feature selection phase that lead us to a reliable list of top predictors.

We now provide a brief rerview and results for all supervised and unsupervised machine learning methods that are employed in this paper. In the supervised learning process, we train a model on a training data and evaluate its accuracy on a test data. We often split the data set into two parts, one part (for instance 75% of all observations) as the training data set, and the other part (for instance 25% of all data set) as the test data set. We train each model only on the training data set and then test the model on the test data set.

2.1.1 Unsupervised Machine Learning and Feature Selection

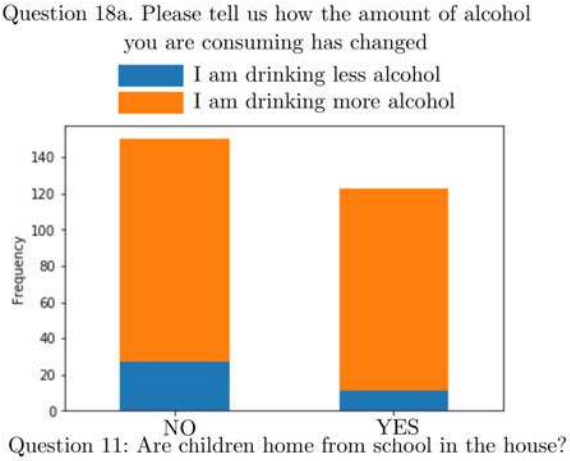
Here, we employ two unsupervised learning methods, Chi-squared test and mutual information, to find out the relationship between the target variable and the rest of variables.

Chi-Squared Test:

Since the target variable and the majority of input variables are categorical (nominal or ordinal), we first apply a Chi-squared test under the null hypothesis H_0 : Question 18a is independent of Question i, where $i \in \{2, 3, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29\}$

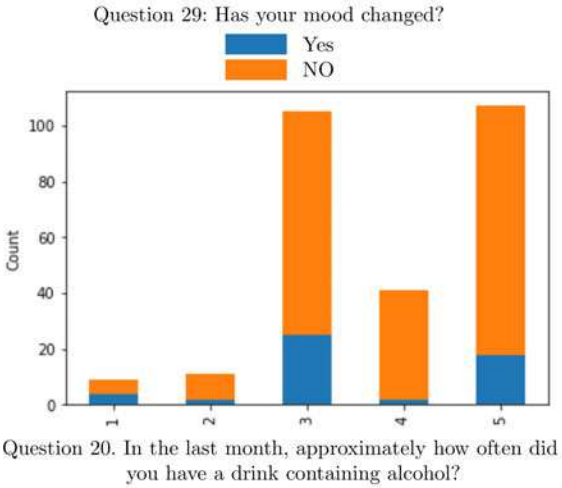
(categorical variables). Our goal is to determine whether the target variable, alcohol consumption habit changes, is independent of the input variables. By a significance level of $\alpha = 0.05$, it turns out that the null hypothesis H_0 is rejected for Questions 11: "Are children home from school in the house?" (p-value= 0.048), Questions 15: "Have you varied your work schedule?" (p-value= 0.037), and Questions 20: "In the last month, approximately how often did you have a drink containing alcohol?" (p-value$<$ 0.001). It is obvious that there is a significant association between Question 18a and Question 20. But an interesting result is the existence of sufficient statistical evidence that leads to rejection of hypothesis that Question 18a is independent of Question 11. It raises a question about the relationship between COVID-related school closures and alcohol consumption habit changes among healthcare workers. By means of bar graph, Figure 1 displays the relationship between alcohol consumption habit changes and COVID-related school closures. One may consider parenting stress associated with COVID-related school closures as a hidden effect of the COVID-19 pandemic.

Fig. 1: Bar graph grouped by Question 18a and Question 11



Another question that arises is whether there is a relationship between Question 29: "Has your mood changed?" and Question 20: "In the last month, approximately how often did you have a drink containing alcohol?." Since both variables are categorical (nominal or ordinal), we employ Chi-squared test under the null hypothesis H_0 : Question 20 is independent of Question 29 with a significance level of $\alpha = 0.05$ to answer the question. It turns out that the null hypothesis is rejected with a p-value equals 0.025, which indicates that there is a statistically significant association between the shifts in mood related to Covid-19 and alcohol consumption frequency (see Figure 2).

Fig. 2: Bar graph grouped by Question 29 and Question 20



Mutual Information:

The mutual information method utilizes information gain to score features based on their importance. It can be used to measure the dependence of two variables. We use *mutual-info-classif* from *sklearn.feature-selection* in Python to find the most related features to Question 18a. It turns out that Question 20: "In the last month, approximately how often did you have a drink containing alcohol?," Question 11: "Are children home from school in the house?" and Question 15: "Have you varied your work schedule?" obtain the highest scores among all features. Like Chi-squared test, the mutual information method also emphasizes that there exists a relationship between the parenting stress associated with COVID-related school closures and alcohol consumption changes. The parenting stress might contribute to an increase in alcohol consumption and a decline in the mental health of healthcare workers too.

The findings in [2], which indicate that alcohol consumption is an important feature (factor) that may contribute to a decline the mental health of healthcare workers, agree with our Chi-squared test and Mutual Information results.

2.1.2 Supervised Machine Learning and Feature Selection

We now apply supervised machine learning models and techniques to find the most important factors of alcohol consumption changes among the healthcare workers during the severe phase of the COVID-19 pandemic. Notice that Question 18a has two classes, which implies that binary classifiers must be employed for analysis. We consider an accuracy score threshold to determine whether a feature score analysis for a trained model is worthwhile or not. If accuracy score of a supervised model is above (or close to) 95%, we analyze its feature importance scores to find the most reliable list of top predictors.

Logistic Regression:

Logistic regression [24, 25] is one of the simplest supervised learning classification algorithms that can be used for a binary classification. Using logistic regression model, we can find out the relationship between the target variable and input variables because not only does it give inference about the importance of each feature, but it also gives the direction of association. It turns out that Logistic regression accuracy on the test set is 89.01% (see Figure 3), which prevents us from diving into further details about feature importance scores of this model.

Support Vector Machines:

Support vector machines (SVM) [26, 27] is another supervised learning algorithm that can be used for both regression and classification. It is very effective in cases where the number of features is greater than the number of datapoints. SVM algorithm can be used for binary classification where it finds a separating line (or hyperplane for higher dimensions) between datapoints of two classes. A SVM algorithm finds the optimal separating hyperplane by finding the closest points (support vectors) to the hyperplane, and then maximizing the distance (margin) between the hyperplane and the support vectors. The interesting strategy that SVM algorithm utilizes for binary classification is that if the data is not linearly separable, then SVM transforms the data to a higher dimensional space, where the transformed data is linearly separable. It turns out that SVM accuracy on test set is 86.81%, which is below our threshold for feature importance analysis (see Figure 3).

Multilayer Perceptron:

Multilayer perceptron (MLP) [28–33] is a class of feedforward artificial neural-networks that contains three types of layers (the input layer, hidden layers and output layer). In an MLP, the data flows from input to output layer, and it can be used for binary classification problems whose data is not linearly separable. One of advantages of using MLP is that it does not make any assumption regarding the underlying probability density functions. Despite MLP achieves a better accuracy on test set, 90.11%, it does not satisfy our criteria for feature importance analysis (see Figure 3).

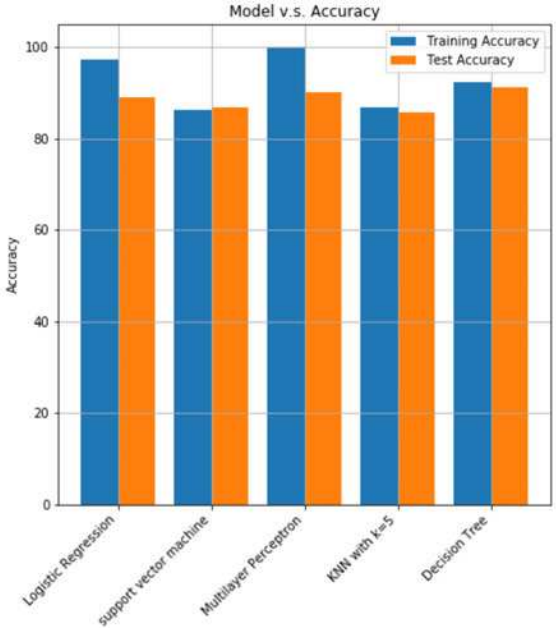
K-Nearest Neighbors:

The k-nearest neighbors (KNN) algorithm [34] is a simple supervised learning algorithm that can be used for both classification and regression. For a classification, the KNN algorithm finds the distance (e.g. Euclidean) between a query and all datapoints, and it then makes a sample space with the k nearest datapoints to the query, and it finally returns the majority class as the prediction for the class of the query. The KKN algorithm is often slow, computationally expensive, and it may not work well with large datasets. Like the previous supervised models, KNN accuracy on test set, 85.71%, is not good enough for going through further details of the results obtained by the model (see Figure 3).

Decision Tree:

Decision trees (DT) [34] are nonparametric supervised learning models that can be used for both classification and regression. A DT contains nodes (the root node, intermediate nodes, and leaf nodes) and branches. For categorical decision trees, certain metrics, such as the Gini index or the Entropy, are used to split the training data to into smaller and smaller subsets containing data points that are more homogenous. A DT can be used for finding out the relationship between the target variable and the input variable, or feature selection. It turns out that DT accuracy on the test set is 91.21% (see Figure 3), but it is not satisfactory.

Fig. 3: Accuracy scores of some supervised machine learning models



Extreme Gradient Boosting (XGBoost):

XGBoost [35, 36] is a decision-tree based ensemble supervised learning algorithm that follows the principle of gradient boosting, but it is more regularized. XGBoost was initially introduced to improve GBM's training time, and it combines the estimates of a set of decision trees (weaker learner) to predict the output of a target variable. In the learning process, XGBoost minimizes a regularized loss function. In the learning process, new decision-trees are added iteratively, one by one, in order to correct the prediction of the previous trees in the model. XGBoost algorithm has several hyperparameters such as number of trees, depth of each tree and number of datapoints used to train the model. To control number of datapoints used for training process, we combine a XGBoost and k-fold cross-validation [57]. Figure 4 displays accuracy scores

of a XGboost with 100 trees with depth 3, and k-fold cross-validation, where k changes between 2 and 40.

Fig. 4: Accuracy scores of XGboost (100 trees, depth=3) k-fold cross validation when k is changing between 2 and 40

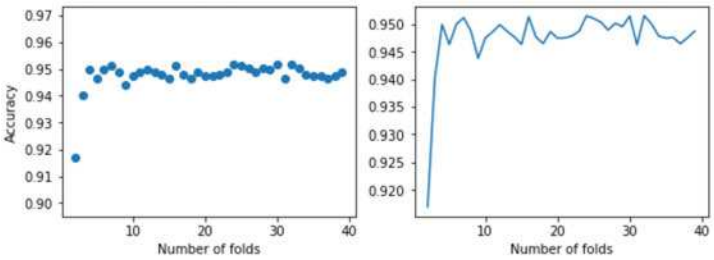


Figure 4 indicates that a XGboost with 100 trees with depth 3 and $k = 4$ cross-validation returns an accuracy of the model is equal to 95%. It also reaches an accuracy of 95.2% if $k = 32$. Figure 5 displays the feature importance scores of the model. Figure 6 spells out why the feature importance score of Question 18 is zero, and it is because all survey respondents answers to Question 18 is "Yes". Table 1 displays the top predictors, the most important features, obtained by the XGBoost model.

Fig. 5: Feature scores of XGboost with 100 trees, depth equals 3, and $k = 9$ cross-validation

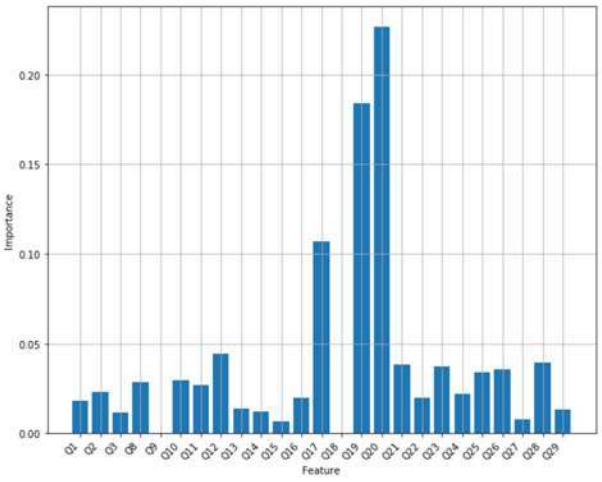
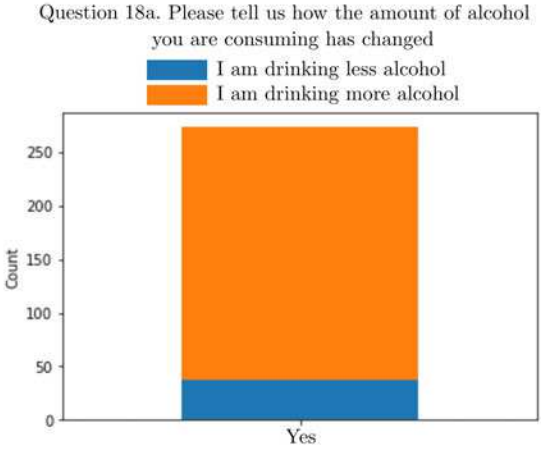


Table 1: The top predictors obtained by the XGBoost model

Question 19: "In January 2020, approximately how often did you have a drink containing alcohol?"
Question 20: "In the last month, approximately how often did you have a drink containing alcohol?"
Question 17: "Has the number of naps you are taking changed?"

Fig. 6: Q18a versus Q18 bar graph group



Question 18. Has the amount of alcohol you are consuming changed?

However, to get a good understanding of other features and find out the relation between alcohol consumption habit changes and other variables, we remove Questions 19 and 20, and retrain a new XGBoost on the remaining. As the result, the feature importance scores of XGBoost in absence of Questions 19 and 20 are shown in Figure 7.

Fig. 7: Feature scores of XGboos with 100 trees, depth equals 3, k=9 cross-validation after Q19 and Q20 are removed

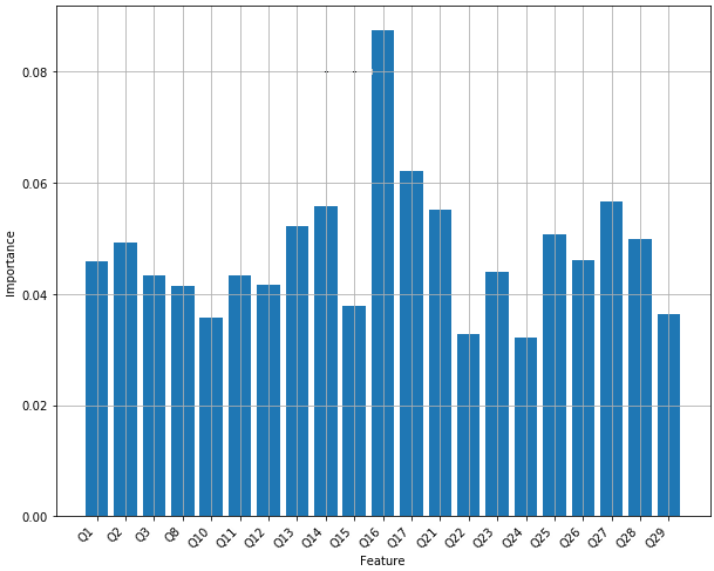


Figure 7 indicates that Question 16: "Have your sleep patterns changed?" is the most important feature of the new model. It seems that there is a relationship between the sleep pattern and alcohol consumption habit changes among healthcare workers.

Light Gradient Boosted Machine (LightGBM):

LightGBM [38] is another leaf-wise decision-tree based algorithm, and it works based on two novel techniques, Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). The GOSS reduces the complexity by down sampling to remove examples with small gradients. The EFB improves the speed of the algorithm and reduces the complexity of it by down sampling features. In other words, the EFB bundles exclusive features into a single feature to reduce the complexity. LightGBM may outperform XGBoost, and it is often faster than XGBoost. One of the most important hyperparameters of LightGBM is number of trees. To find the most robust and accurate model, we tune number of trees between 2 and 40 while the number of leaves and the depth of each tree is left with no limit. Figure 8 displays the accuracy scores of LightGBM as number of trees changes. It turns out that LightGBM with 20 trees performs better with an accuracy score of 94.51%. Figure 9 displays the feature importance scores of LightGBM with 20 trees.

Fig. 8: Accuracy of LightGBM for many different numbers of trees

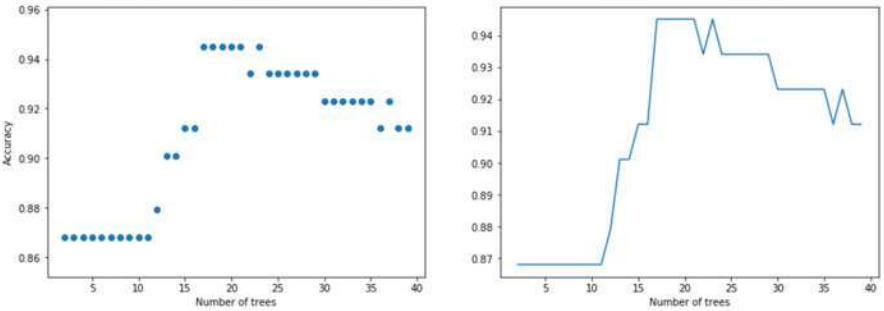
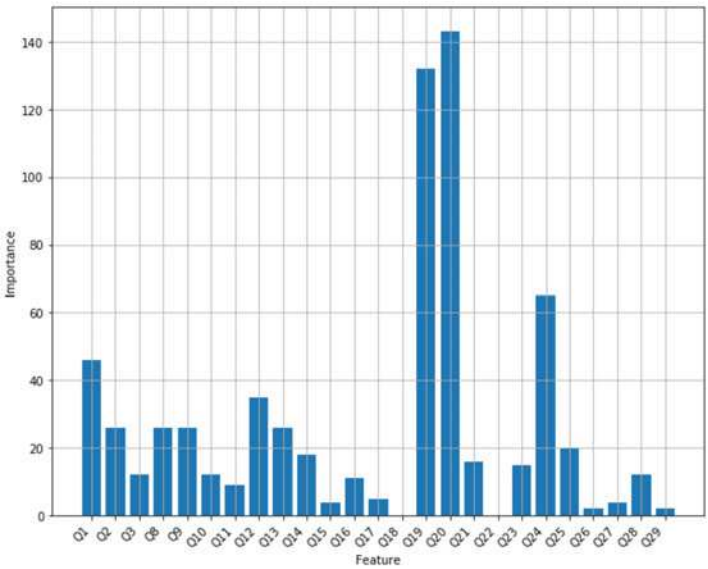
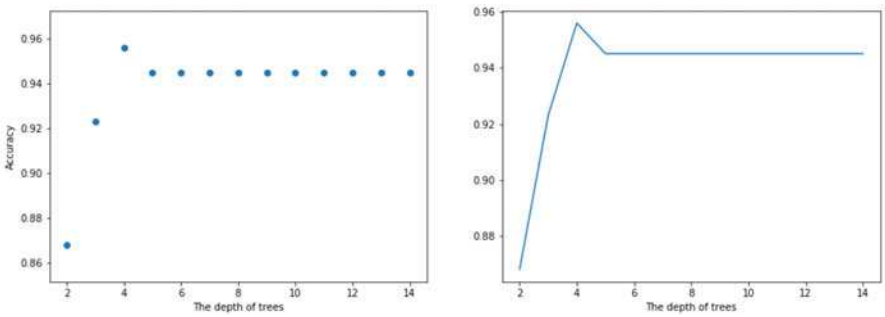


Fig. 9: Feature scores of LightGBM



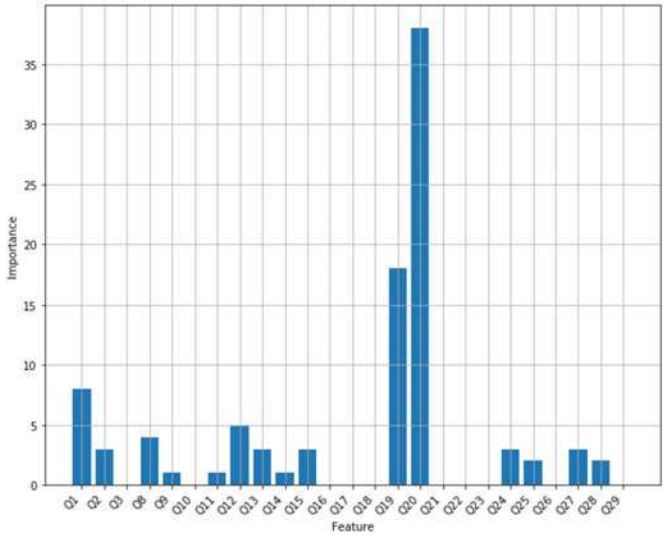
We now control the number of leaves and the depth of trees to see whether we can improve the performance of LightGBM. Figure 10 displays the accuracy scores of LightGBM with 20 trees as the depth of each tree and number of leaves change. Note that, we define number of leaves to be twice as large as the maximum depth of trees.

Fig. 10: The accuracy scores of a 20-tree-LightGBM with different maximum depths



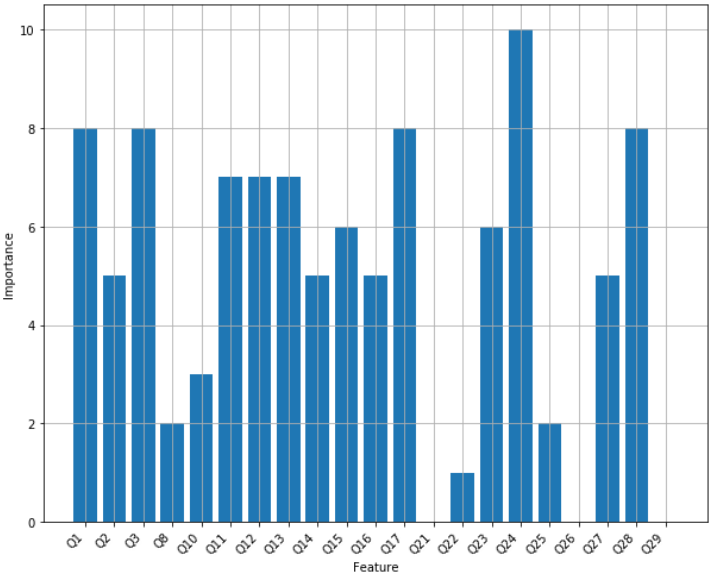
It turns out that LightGBM with 20 trees, where the depth of each tree is 4, obtains an accuracy of 95.6%. Figure 11 displays the feature importance scores of the LightGBM with 20 trees with depth 4.

Fig. 11: Feature scores of LightGBM with 20 trees (depth equals 4)



Figures 10 and 11 indicate that LightGBM is significantly dependent on Questions 19 and 20 in the training process. To allow the model to explore further in details of other features, we remove questions 19 and 20 and retrain LightGBM on the remaining features. Figure 12 displays the feature scores of LightGBM with 20 trees in absence of Questions 19 and 20.

Fig. 12: Feature scores of LightGBM with 20 trees (depth equals 4) in absence of Questions 19 and 20



CatBoost:

CatBoost [40] is another algorithm for gradient boosting on decision trees. It is a weighted sampling version of Stochastic Gradient Boosting that uses one-hot-encoding for categorical data. CatBoost utilizes two feature importance methods, the prediction-value-change and the loss-function-change. The prediction-value-change sorts features based on obtained prediction changes if a feature value changes. On the other hand, the loss-function-change sorts features based on the difference between loss value of the model with and without a feature. Number of trees and the depth of each tree are two important hyperparameters of CatBoost. Figure 13 displays accuracy scores of CatBoost model while number of trees changes and the depth of each tree is fixed to be 6. It turns out that CatBoost with 7 trees reaches the maximum accuracy of 95.4%. Figure 14 displays feature scores of CatBoost with 7 trees.

Fig. 13: Accuracy scores of CatBoost model as the number of trees changes between 4 and 19

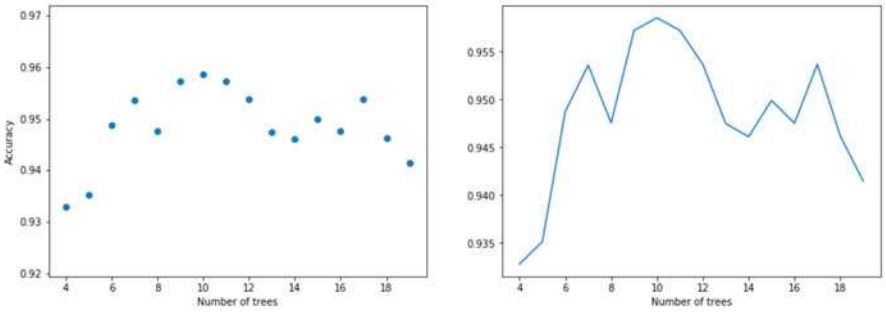
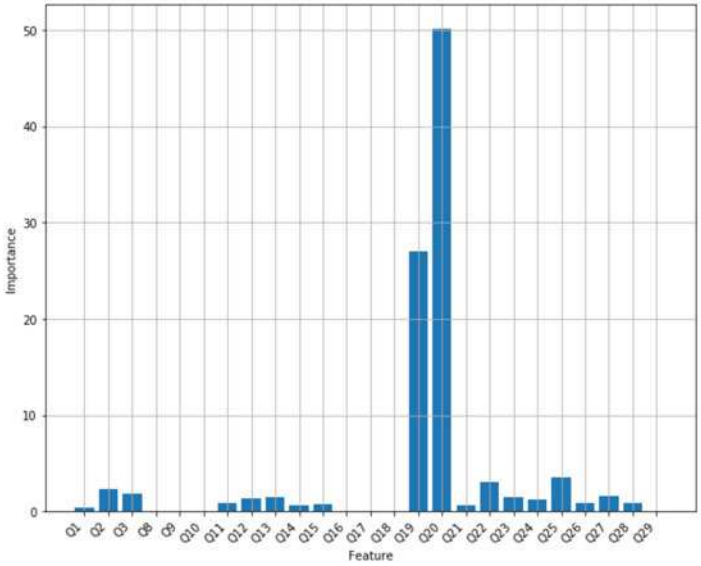
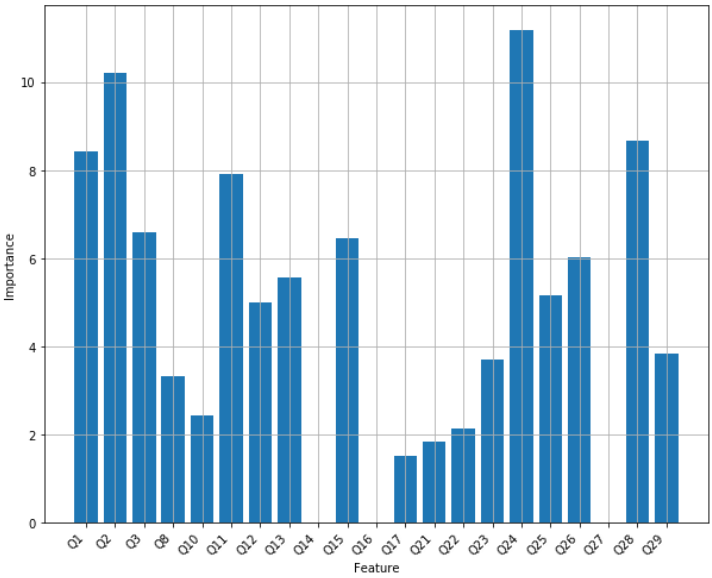


Fig. 14: Feature scores of CatBoost with 11 trees and depth equals 6



Like XGBoost and LightGBM, CatBoost also makes its prediction mostly based on Questions 19 and 20. To find out the relationship between the target variable and other input variables, we remove Questions 19 and 20, and retrain CatBoost with 11 trees on the remaining features. Figure 15 displays feature scores of CatBoost in absence of Questions 19 and 20.

Fig. 15: Feature scores of CatBoost with 11 trees and depth equals 6 in absence of Questions 19 and 20



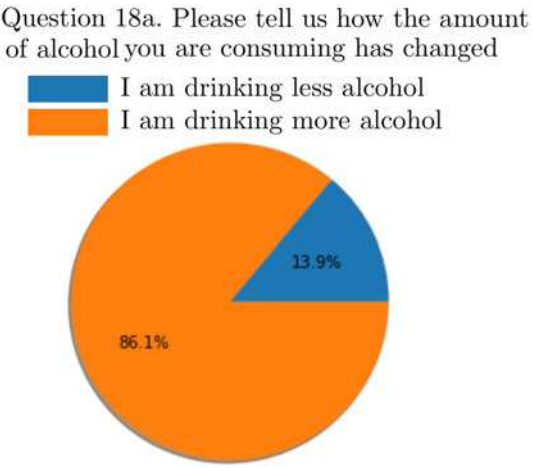
The most important features, or the top predictors, of the retrained CatBoost model are:

- Question 24: 'How many hours of COVID-related news or social media are you consuming on average per day?"
- Question 2: "What is your gender?"
- Question 28:" Has the amount of food you have been eating per day changed?"
- Question 1: "What is your age?"

Synthetic Minority Oversampling Technique (SMOTE):

SMOTE [39] is a statistical oversampling technique for increasing the number of cases in minority classes to alleviate the challenges of an imbalanced dataset. As Figure 16 illustrates, there is a significant difference in the distribution of the cases in classes of Question 18a. The dataset is biased toward the class "I am drinking more alcohol" that may cause a poor performance.

Fig. 16: The distribution of examples between two classes of the target variable (Question 18a)



We apply SMOTE with a random forest containing 100 trees (the depth of each tree is 8). We first apply SMOTE to the dataset, then split the new dataset into training and test sets and finally we train the model on the new training set and calculate accuracy of the model on the test set. The accuracy score of the model is 96.79%. Figure 17 displays feature importance scores of the model. As Figure 17 illustrates, Questions 19, 20, 15 and 11 are the most important features of the model.

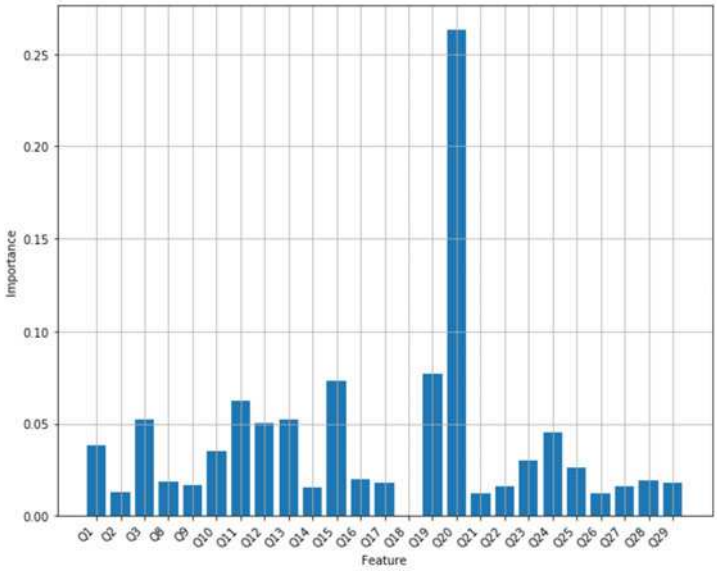
Figure 18 displays the accuracy scores of all supervised machine learning models that discussed in this section. Top predictors for mental health analysis have been identified from different approaches. Among all the approaches, the random forest using SMOTE is the model that has identified the maximum top predictors. Figure 19 summarizes the methology and the results in this section.

3 Discussion

In this section, we discuss the top predictors of alcohol consumption habit changes among healthcare workers in the United States that have obtained in Section 2 by several supervised and unsupervised machine learning models and methods. We also discuss some research articles with non machine learning approaches whose results agree with our results.

Question 20: “In the last month, approximately how often did you have a drink containing alcohol?” is the most important feature for all supervised learning algorithms, Chi-Squared test and Mutual-Information method. Moreover, Question 19: “In January 2020, approximately how often did you have a

Fig. 17: Feature scores of SMOTE Random Forest with 100 trees and depth equals 8

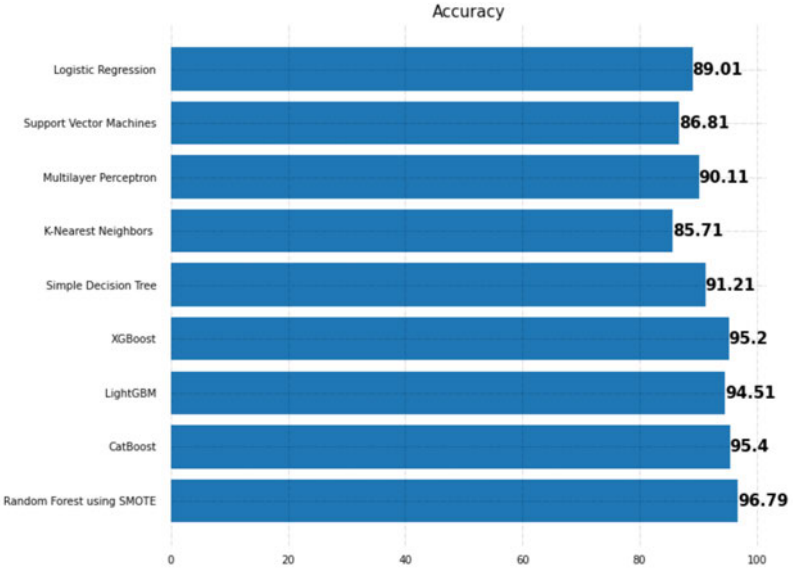


drink containing alcohol?" is the second important feature in supervised learning algorithms. But, due to the nature of Question 18a, the relationship among Question 18a, Question 19 and Question 20 is clear.

One of the most important questions, or features, that appears in unsupervised learning methods as well as some supervised learning algorithms is Questions 11: "Are children home from school in the house?" It raises a question about the relationship between COVID-19 associated school closure and alcohol consumption habit changes among healthcare workers. There may be a strong relationship between COVID-related school closure, parenting stress and alcohol consumption that requires the deep research of the topic. Based on our findings, which agree with the findings in [50], we may make a conclusion about the relationship between parenting stress and alcohol consumption: the school closure associated with the COVID-19 pandemic can be associated with an increase in stress and anxiety among healthcare workers, which may lead to an increase in alcohol consumption as a coping mechanism. To alleviate hidden effects of COVID-19, e.g. parenting stress, on healthcare workers, some operational strategies might be used in schools to reduce the spread of COVID-19 and maintain safe operations in school during the COVID-19 pandemic.

Both Chi-Squared test and Mutual-Information method as well as some supervised learning algorithms imply that Questions 15: "Have you varied your work schedule?" is not independent of alcohol consumption habit changes among healthcare workers. As the COVID-19 pandemic continues, healthcare workers are more in need to work shifts, and consequently healthcare workers who work shifts may consume more alcohol than dayworkers. Our

Fig. 18: Feature scores of SMOTE Random Forest with 100 trees and depth equals 8

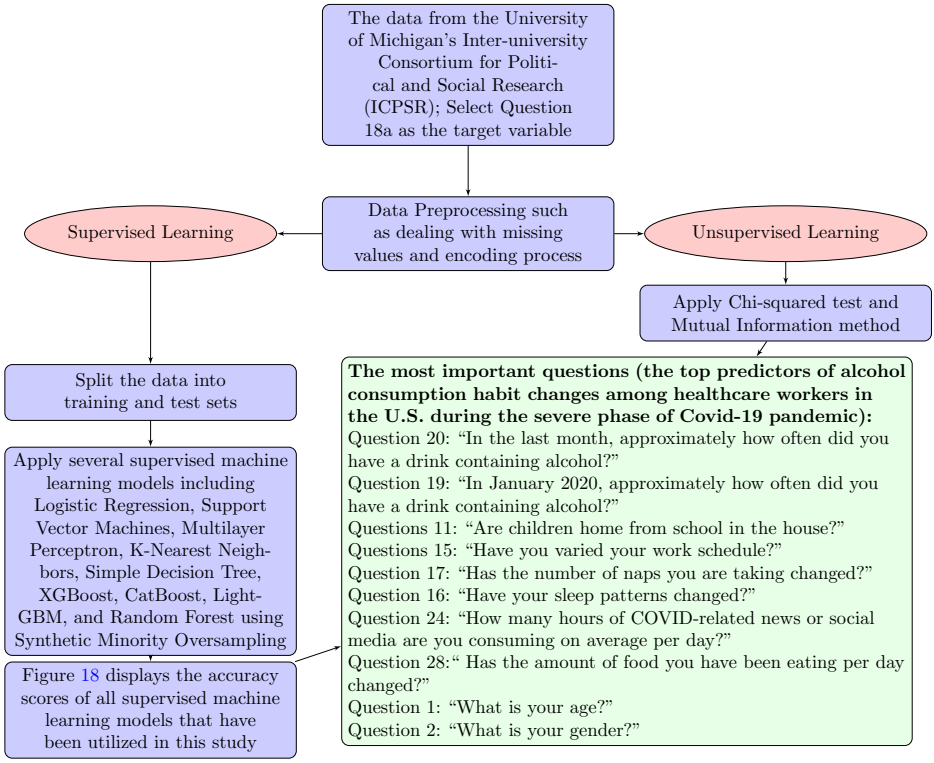


results support a study from The Ohio State University College of Nursing [42] that reveals how the COVID-19 pandemic associated long shifts has impacted nurses working. In another study by Cooper et al. [52], it is shown that work stressors lead to increased distress, which in turn promotes problematic alcohol use. The findings in [52] also reveals the relationship between work-related stress and an increase in alcohol consumption among healthcare professionals during the COVID-19 pandemic.

Next question that is selected by XGBoost and LightGBM is Question 17: "Has the number of naps you are taking changed?." Moreover, Question 16: "Have your sleep patterns changed?" is selected by XGBoost as an important feature. One way that we can justify the relationship between alcohol consumption changes and sleep pattern changes is to give close and thoughtful attention to the findings of [43]. According to a study titled Alcohol Alert by the National Institute on Alcohol Abuse and Alcoholism [43], alcohol consumption can change sleep patterns. One way that healthcare workers may choose to cope with the COVID-19 associated parenting stress and long shifts is by turning to alcohol, which leads to a change in sleep pattern. So, to reduce sleep problems associated with COVID-19 among healthcare workers, some psychology-based strategies may be used in hospitals to decrease work-related stress and the usage of alcohol in healthcare workers and keep them safe.

LightGBM and CatBoost select Question 24: "How many hours of COVID-related news or social media are you consuming on average per day?" as one the most important features in their training process. Buchanan et al. [44]

Fig. 19: A summary of the methology and the results



examines the emotional consequences of exposure to COVID-related news. The findings of [29] can help us have a better understanding on the relationship between COVID-related news exposure and alcohol use changes. Due to their job, healthcare workers seek out COVID-19 information as a means of coping with challenging situations, and as a result they put themselves in stressful situations. Strainback et al. [45] also explores COVID-related news effects on mental health. Their findings show that COVID-19 media consumption leads to psychological distress, which may cause an increase in alcohol consumption. Thus, we can make a conjecture: the stress resulting from COVID-19 related news exposure may produce changes in drinking behavior.

Question 28:" Has the amount of food you have been eating per day changed?" is another important feature that is selected by LightGBM and CatBoost. An increase in alcohol consumption can cause changes in a healthcare worker diet. People who drink alcohol more frequently may choose less healthy food options that are high in fat and sugar [46]. Moreover, each gram of pure alcohol has 29 kilojoules of energy [47], and when it is mixed with sugary drinks, it contains even more calories. Since alcohol is absorbed directly in the bloodstream [48] that can lead to immediate changes on the amount of food that people eat.

Finally, Question 1: "What is your age?" and Question 2: "What is your gender?" are selected as important features by LightGBM and CatBoost. Lavretsky et al. [49] examines stress-related changes associated with aging and sex differences. Their results can help us have a better understanding about the relationship between alcohol consumption changes due to COVID-related stress, aging and gender differences. Novais et al. [51] also examines the effects of age and sex differences in the stress response. Their findings agree with the findings of [49] and lead to the same result. Therefore, regarding the relationship between age and alcohol consumption changes, we may utilize our findings and the findings of [53] to make a conclusion: the susceptibility to stress differs between young and old people, and may lead to different stress responses e.g., turning to alcohol to cope with stressful situations. Regarding the relationship between gender differences and alcohol consumption changes, we may give a close attention to the results of [54] and make a conclusion: men and women tend to react differently with stress, both psychologically and biologically, and may lead to different stress responses.

4 Conclusions

In this paper, we have examined the effects of the COVID-19 pandemic on alcohol consumption habit changes among healthcare workers using a mental health survey data obtained from the University of Michigan Inter-University Consortium for Political and Social Research. We have utilized several supervised and unsupervised machine learning methods and models such as Decision Trees, Logistic Regression, Naive Bayes classifier, k-Nearest Neighbors, Support Vector Machines, Multilayer perceptron, Random Forests, XGBoost, CatBoost, LightGBM, Synthetic Minority Oversampling, Chi-Squared Test and mutual information method to find out how the COVID-19 pandemic causes changes in alcohol use and stress. Through the interpretation of many supervised and unsupervised methods applied to the dataset, we have concluded that some effects of the COVID-19 pandemic such as school closure, work schedule change and COVID-related news exposure may lead to an increase in alcohol use.

In future work, we are interested in analyzing more data related to healthcare workers to examine the relationship between their mental health and other factors related or unrelated to COVID-19. We are also interested in implementing other optimization methods such as those that are introduced in [55] and [56] to see whether accuracy or speed of some models, which were utilized in this paper, can be improved.

References

[1] Huang, Chaolin, et al. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China." The lancet 395.10223 (2020): 497-506.

[2] Rezapour, Mostafa, and Lucas Hansen. "A Machine Learning Analysis of COVID-19 Mental Health Data." arXiv preprint arXiv:2112.00227 (2021).

[3] Spoorthy, Mamidipalli Sai, Sree Karthik Pratapa, and Supriya Mahant. "Mental health problems faced by healthcare workers due to the COVID-19 pandemic–A review." Asian journal of psychiatry 51 (2020): 102119.

[4] Vizheh, Maryam, et al. "The mental health of healthcare workers in the COVID-19 pandemic: A systematic review." Journal of Diabetes & Metabolic Disorders 19.2 (2020): 1967-1978.

[5] Lai, Jianbo, et al. "Factors associated with mental health outcomes among health care workers exposed to coronavirus disease 2019." JAMA network open 3.3 (2020): e203976-e203976.

[6] Lima, Carlos Kennedy Tavares, et al. "The emotional impact of Coronavirus 2019-nCoV (new Coronavirus disease)." Psychiatry research 287 (2020): 112915.

[7] Greenberg, Neil, et al. "Managing mental health challenges faced by healthcare workers during covid-19 pandemic." bmj 368 (2020).

[8] Vizheh, Maryam, et al. "The mental health of healthcare workers in the COVID-19 pandemic: A systematic review." Journal of Diabetes & Metabolic Disorders 19.2 (2020): 1967-1978.

[9] Di Tella, Marialaura, et al. "Mental health of healthcare workers during the COVID-19 pandemic in Italy." Journal of evaluation in clinical practice 26.6 (2020): 1583-1587.

[10] Chirico, Francesco, Gabriella Nucera, and Nicola Magnavita. "Protecting the mental health of healthcare workers during the COVID-19 emergency." BJPsych International 18.1 (2021).

[11] Greenberg, Neil, et al. "Managing mental health challenges faced by healthcare workers during covid-19 pandemic." bmj 368 (2020).

[12] Yang, Lei, et al. "Urgent need to develop evidence-based self-help interventions for mental health of healthcare workers in COVID-19 pandemic." Psychological Medicine 51.10 (2021): 1775-1776.

[13] Chatzittofis, Andreas, et al. "Impact of the COVID-19 pandemic on the mental health of healthcare workers." International journal of environmental research and public health 18.4 (2021): 1435.

[14] Kushwaha, Shashi, et al. "Significant applications of machine learning for COVID-19 pandemic." Journal of Industrial Integration and Management

5.04 (2020): 453-479.

[15] Lalmuanawma, Samuel, Jamal Hussain, and Lalrinfela Chhakchhuak. "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review." Chaos, Solitons & Fractals 139 (2020): 110059.

[16] Benvenuto, Domenico, et al. "Application of the ARIMA model on the COVID-2019 epidemic dataset." Data in brief 29 (2020): 105340.

[17] Kucharski, Adam J., et al. "Early dynamics of transmission and control of COVID-19: a mathematical modelling study." The lancet infectious diseases 20.5 (2020): 553-558.

[18] Rezapour, Mostafa, and Scott K. Elmshaeuser. "Artificial Intelligence-Based Analytics for Impacts of COVID-19 and Online Learning on College Students' Mental Health." arXiv preprint arXiv:2202.07441 (2022).

[19] Rezapour, Mostafa, and Colin A. Varady. "A machine learning analysis of the relationship between some underlying medical conditions and COVID-19 susceptibility." arXiv preprint arXiv:2112.12901 (2021).

[20] Conroy, Deirdre, and Goldstein, Cathy. COVID Isolation on Sleep and Health in Healthcare Workers. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2020-11-20. https://doi.org/10.3886/E127081V1

[21] Conroy, Deirdre A., et al. "The effects of COVID-19 stay-at-home order on sleep, health, and working patterns: a survey study of US health care workers." Journal of Clinical Sleep Medicine 17.2 (2021): 185-191.

[22] https://www.openicpsr.org/openicpsr/project/127081/version/V1/view?path=/openicpsr/127081/fcr:versions/V1&type=project

[23] https://www.openicpsr.org/openicpsr/project/127081/version/V1/view?path=/openicpsr/127081/fcr:versions/V1&type=project

[24] Menard, Scott. Applied logistic regression analysis. Vol. 106. Sage, 2002.

[25] Wright, Raymond E. "Logistic regression." (1995).

[26] Steinwart, Ingo, and Andreas Christmann. Support vector machines. Springer Science & Business Media, 2008.

[27] Hearst, Marti A., et al. "Support vector machines." IEEE Intelligent Systems and their applications 13.4 (1998): 18-28.

[28] Wang, Sun-Chong. "Artificial neural network." Interdisciplinary computing in java programming. Springer, Boston, MA, 2003. 81-100.

[29] Hecht-Nielsen, Robert. "Theory of the backpropagation neural network." Neural networks for perception. Academic Press, 1992. 65-93.

[30] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[31] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436-444.

[32] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.

[33] Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." Neural networks 61 (2015): 85-117.

[34] Bishop, Christopher M. "Pattern recognition." Machine learning 128.9 (2006).

[35] 20. Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016

[36] Chen, Tianqi, et al. "Xgboost: extreme gradient boosting." R package version 0.4-2 1.4 (2015): 1-4.

[37] Dorogush, Anna Veronika, Vasily Ershov, and Andrey Gulin. "CatBoost: gradient boosting with categorical features support." arXiv preprint arXiv:1810.11363 (2018).

[38] Ke, Guolin, et al. "Lightgbm: A highly efficient gradient boosting decision tree." Advances in neural information processing systems 30 (2017): 3146-3154.

[39] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.

[40] Dorogush, Anna Veronika, Vasily Ershov, and Andrey Gulin. "CatBoost: gradient boosting with categorical features support." arXiv preprint arXiv:1810.11363 (2018).

[41] https://github.com/MostafaRezapour/Hidden-Effects-of-COVID-19-on-Healthcare-Workers-A-Machine-Learning-Analysis

[42] Melnyk, Bernadette Mazurek, et al. "Associations Among Nurses' Mental/Physical Health, Lifestyle Behaviors, Shift Length, and Workplace

Wellness Support During COVID-19: Important Implications for Health Care Systems." Nursing Administration Quarterly 46.1 (2022): 5-18.

[43] https://pubs.niaaa.nih.gov/publications/aa41.htm

[44] Buchanan, Kathryn, et al. "Brief exposure to social media during the COVID-19 pandemic: Doom-scrolling has negative emotional consequences, but kindness-scrolling does not." Plos one 16.10 (2021): e0257728.

[45] Stainback, Kevin, Brittany N. Hearne, and Monica M. Trieu. "COVID-19 and the 24/7 News Cycle: Does COVID-19 News Exposure Affect Mental Health?." Socius 6 (2020): 2378023120969339.

[46] Lavin, Jacquie, Carolyn Pallister, and Leigh Greenwood. "The government must do more to raise awareness of the links between alcohol and obesity, rather than treating them as separate issues." Perspectives in public health 136.3 (2016): 123-124.

[47] Lourenço, S., A. Oliveira, and C. Lopes. "The effect of current and lifetime alcohol consumption on overall and central obesity." European Journal of Clinical Nutrition 66.7 (2012): 813-818.

[48] Youngerman, Barry, and Mark J. Kittleson. The truth about alcohol. Infobase Publishing, 2005.

[49] Lavretsky, Helen, and Paul A. Newhouse. "Stress, inflammation and aging." The American journal of geriatric psychiatry: official journal of the American Association for Geriatric Psychiatry 20.9 (2012): 729.

[50] Avery, Ally R., et al. "Stress, anxiety, and change in alcohol use during the COVID-19 pandemic: findings among adult twin pairs." Frontiers in psychiatry 11 (2020).

[51] Novais, Ashley, et al. "How age, sex and genotype shape the stress response." Neurobiology of stress 6 (2017): 44-56.

[52] Cooper, M. Lynne, Marcia Russell, and Michael R. Frone. "Work stress and alcohol effects: A test of stress-induced drinking." Journal of health and social behavior (1990): 260-276.

[53] Hamilton, Jessica L., et al. "Pubertal timing and vulnerabilities to depression in early adolescence: Differential pathways to depressive symptoms by sex." Journal of adolescence 37.2 (2014): 165-174.

[54] Verma, Rohit, Yatan Pal Singh Balhara, and Chandra Shekhar Gupta. "Gender differences in stress response: Role of developmental and biological determinants." Industrial psychiatry journal 20.1 (2011): 4.

[55] Rezapour, Mostafa, and Thomas J. Asaki. "Adaptive trust-region algorithms for unconstrained optimization." Optimization Methods and Software (2019): 1-23.

[56] Erway, Jennifer B., and Mostafa Rezapour. "A New Multipoint Symmetric Secant Method with a Dense Initial Matrix." arXiv preprint arXiv:2107.06321 (2021).

[57] Refaeilzadeh, Payam, Lei Tang, and Huan Liu. "Cross-validation." Encyclopedia of database systems 5 (2009): 532-538.

Figures

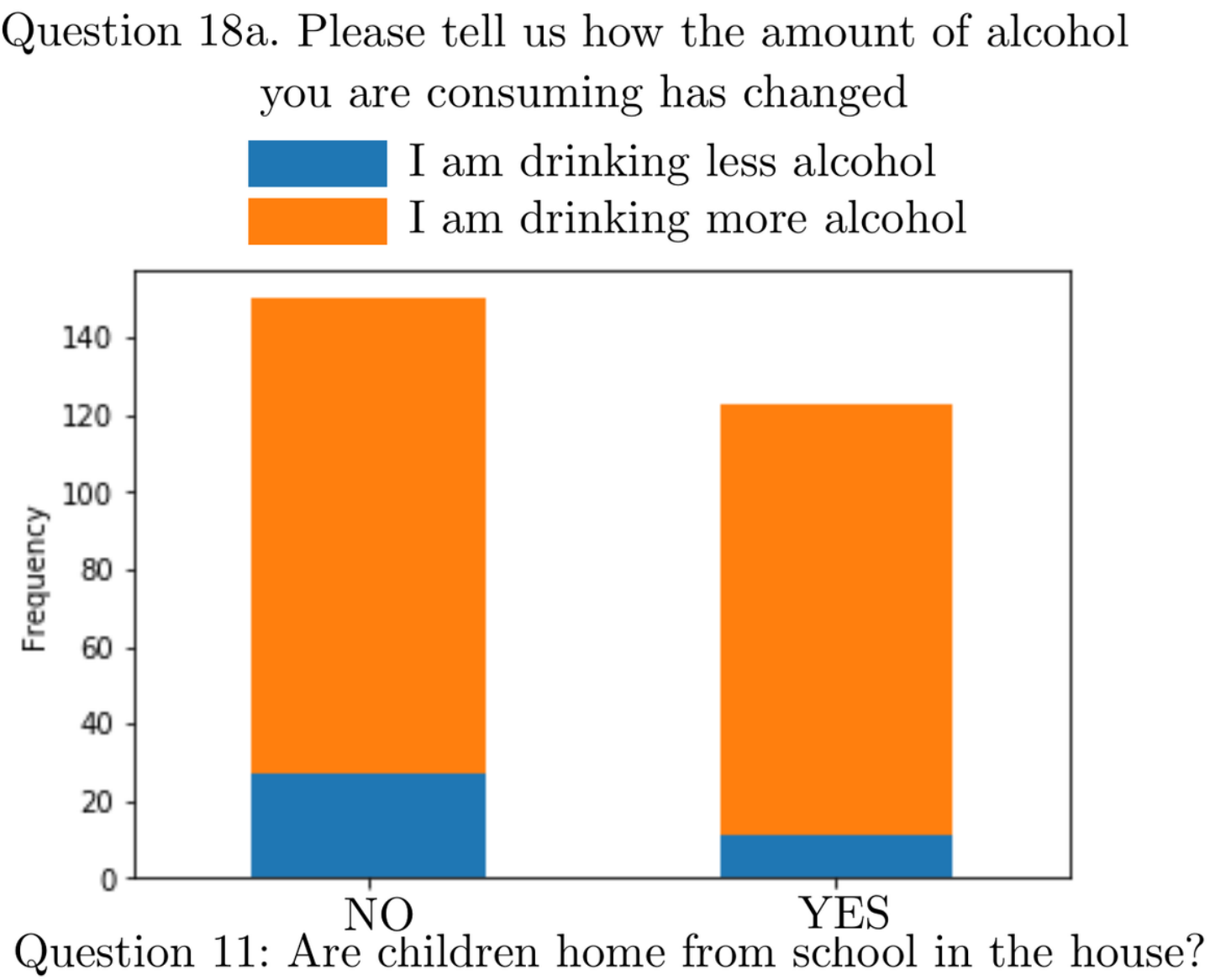


Figure 1

Bar graph grouped by Question 18a and Question 11

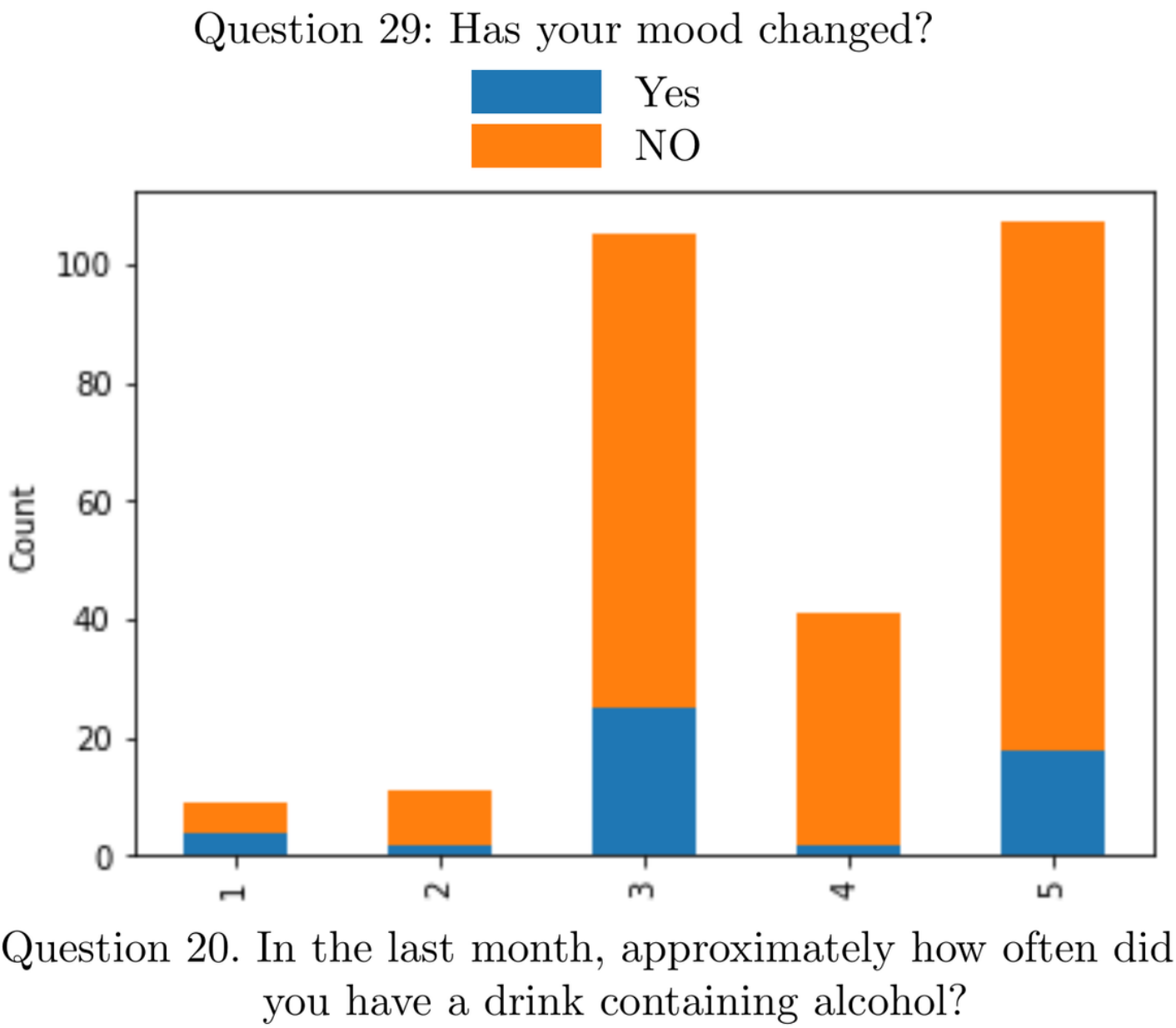


Figure 2

Bar graph grouped by Question 29 and Question 20

Figure 3

Accuracy scores of some supervised machine learning models

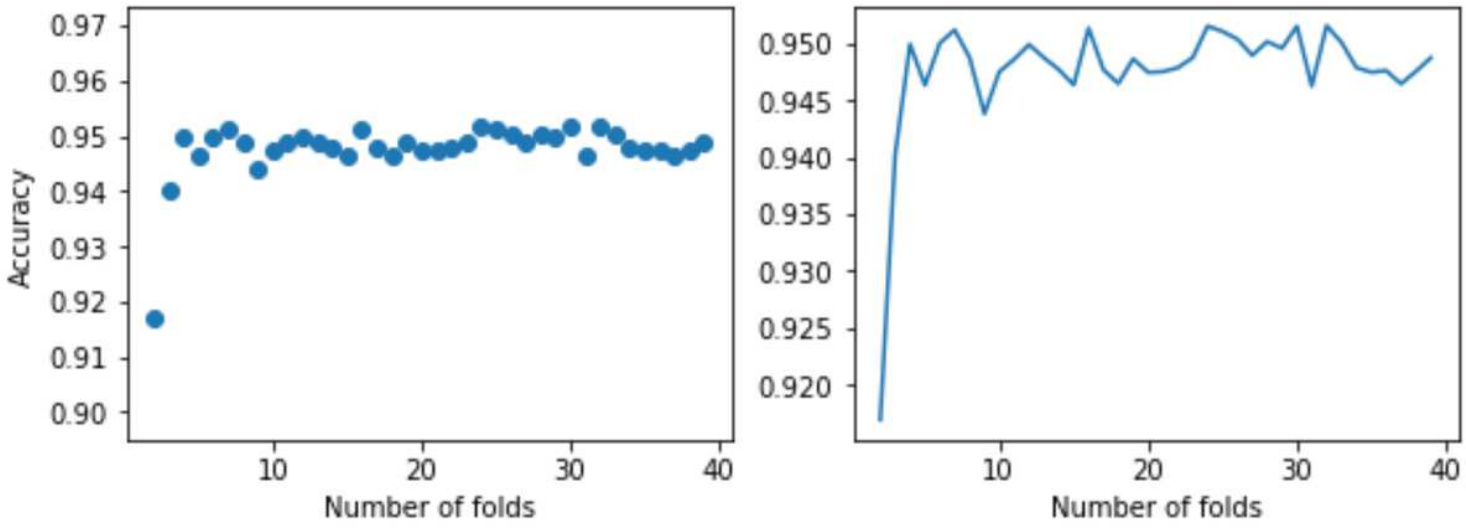


Figure 4

Accuracy scores of XGboost (100 trees, depth=3) k-fold cross validation when k is changing between 2 and 40

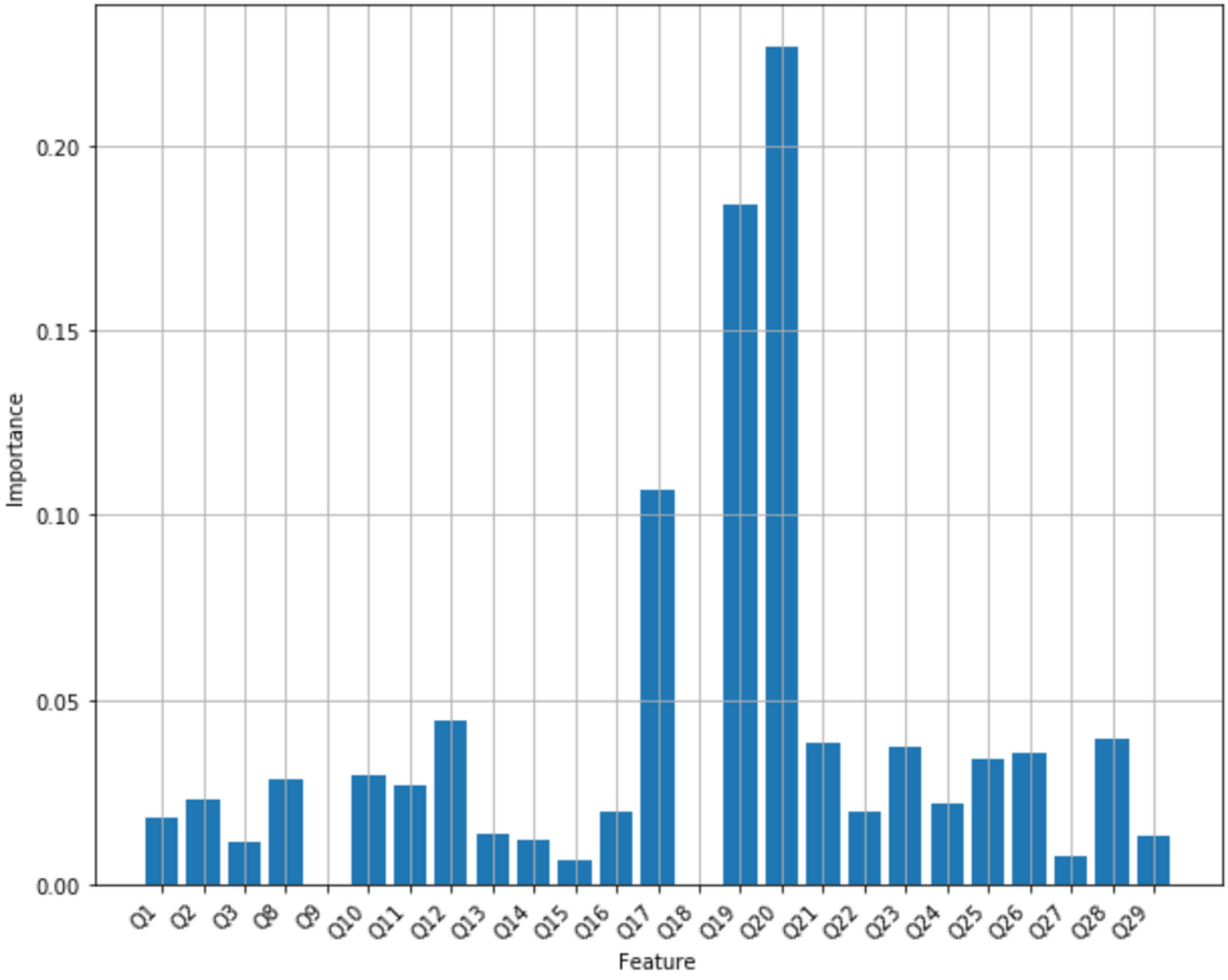


Figure 5

Feature scores of XGboost with 100 trees, depth equals 3, and cross-validation

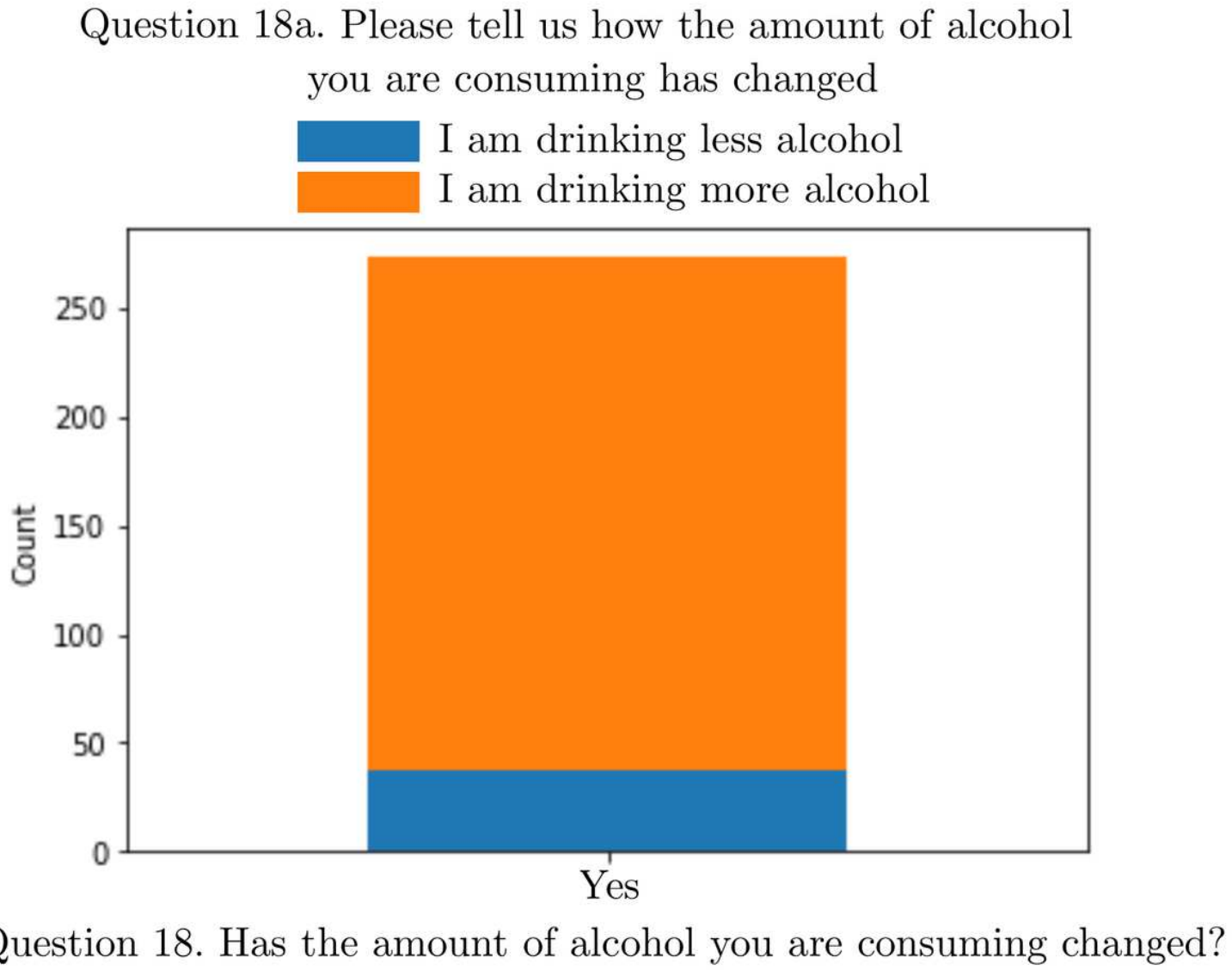


Figure 6

Q18a versus Q18 bar graph group

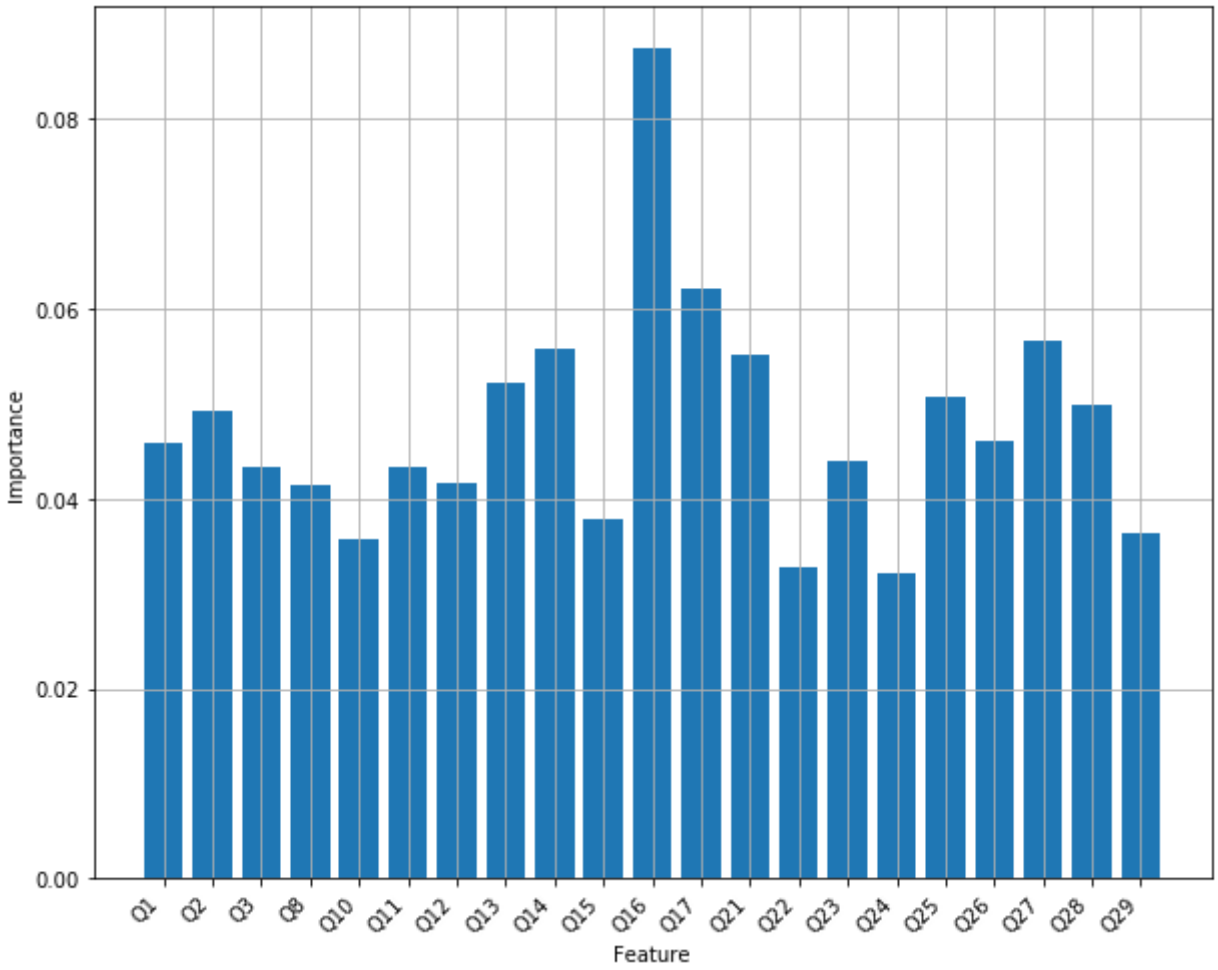


Figure 7

Feature scores of XGboos with 100 trees, depth equals 3, k=9 cross-validation after Q19 and Q20 are removed

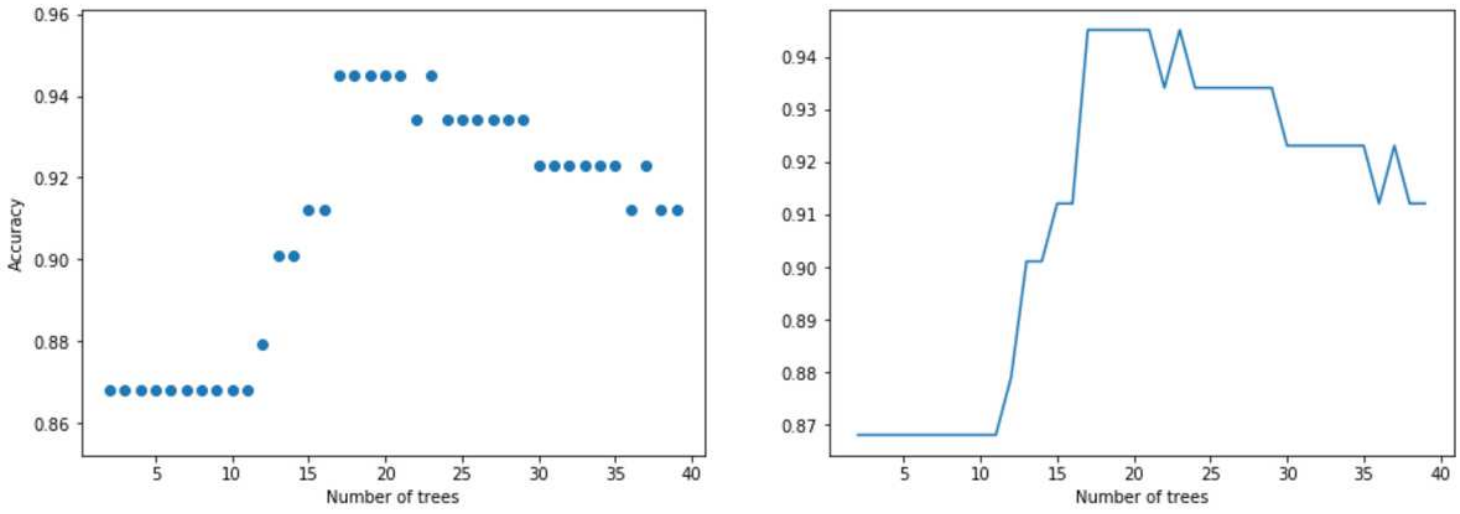


Figure 8

Accuracy of LightGBM for many different numbers of trees

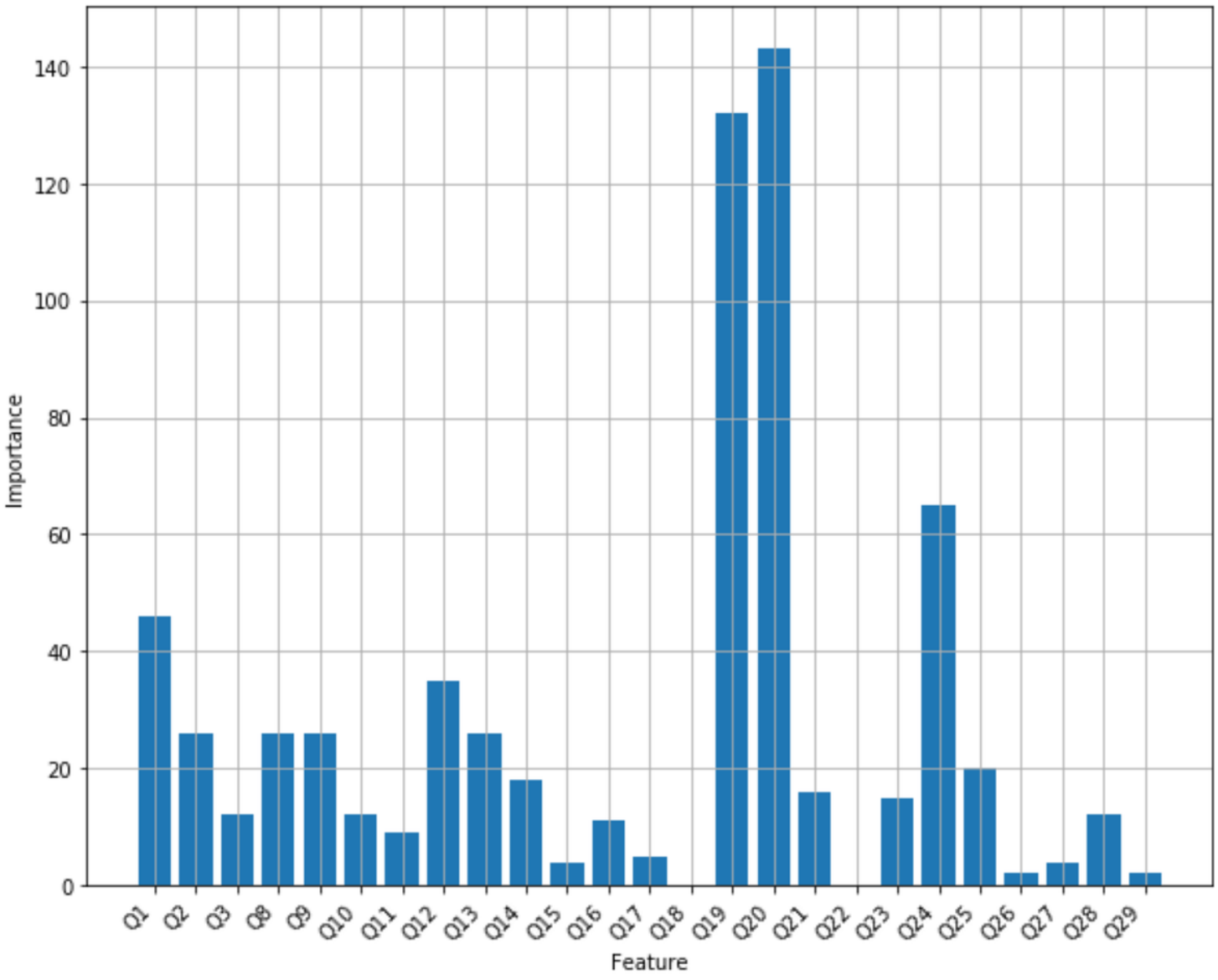
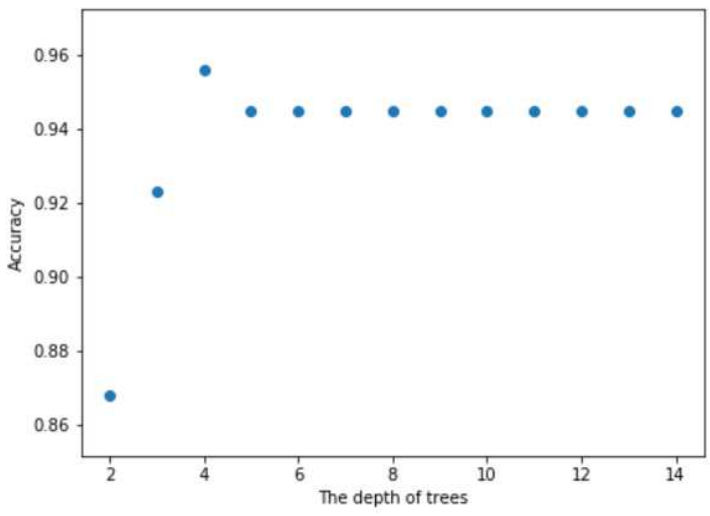


Figure 9

Feature scores of LightGBM



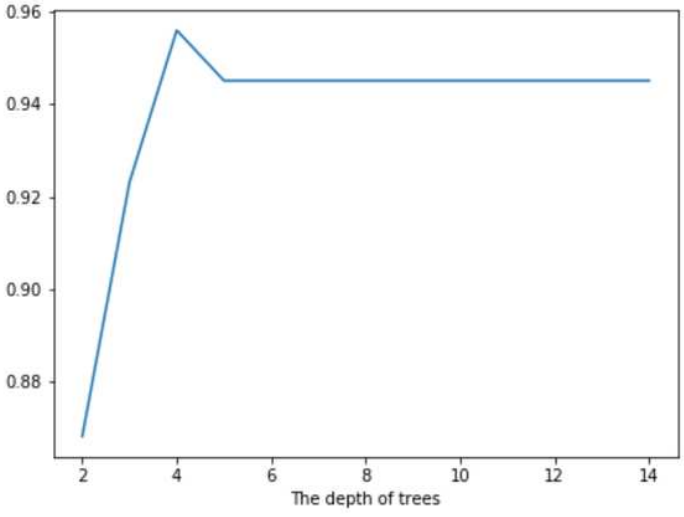


Figure 10

The accuracy scores of a 20-tree-LightGBM with different maximum depths

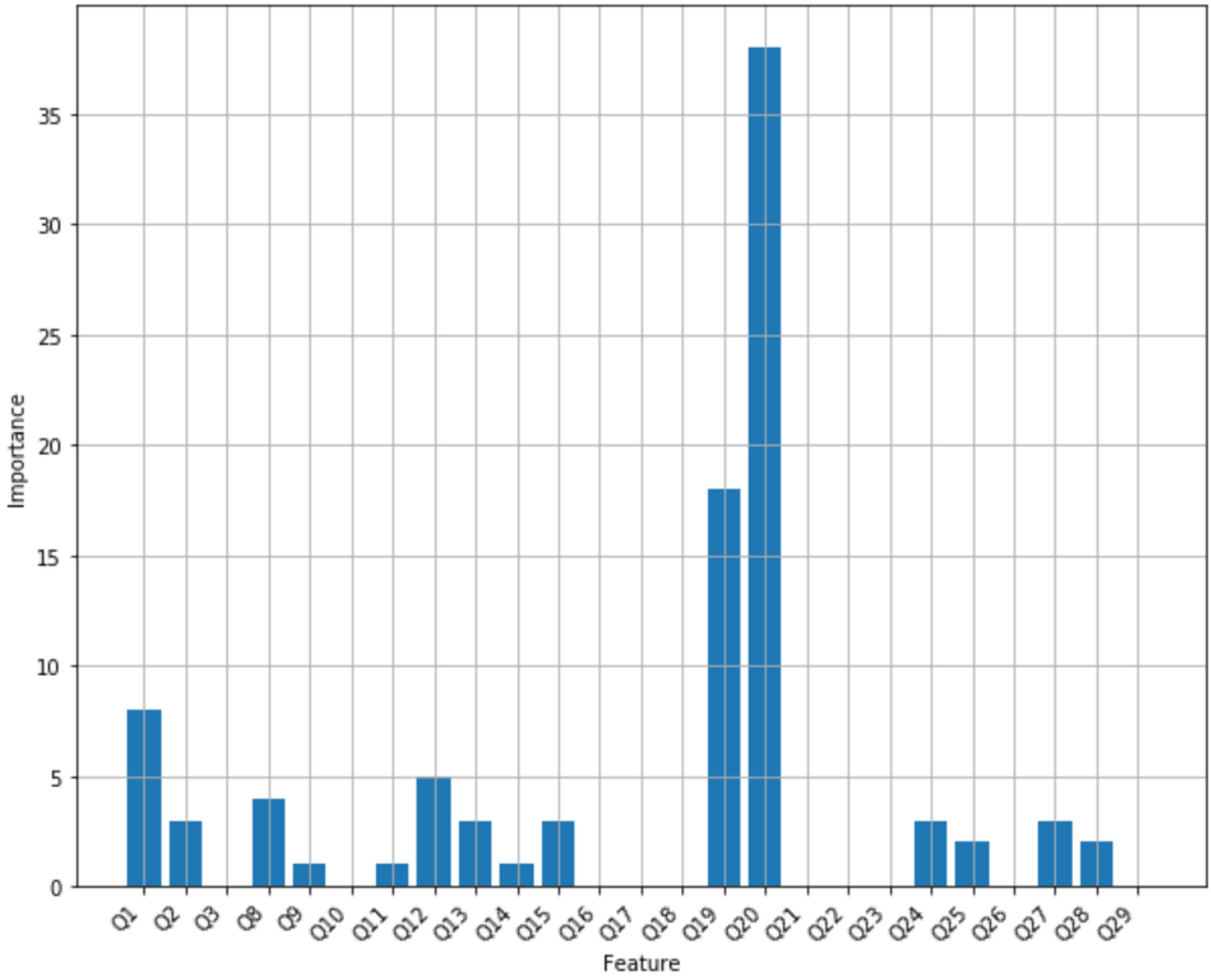


Figure 11

Feature scores of LightGBM with 20 trees (depth equals 4)

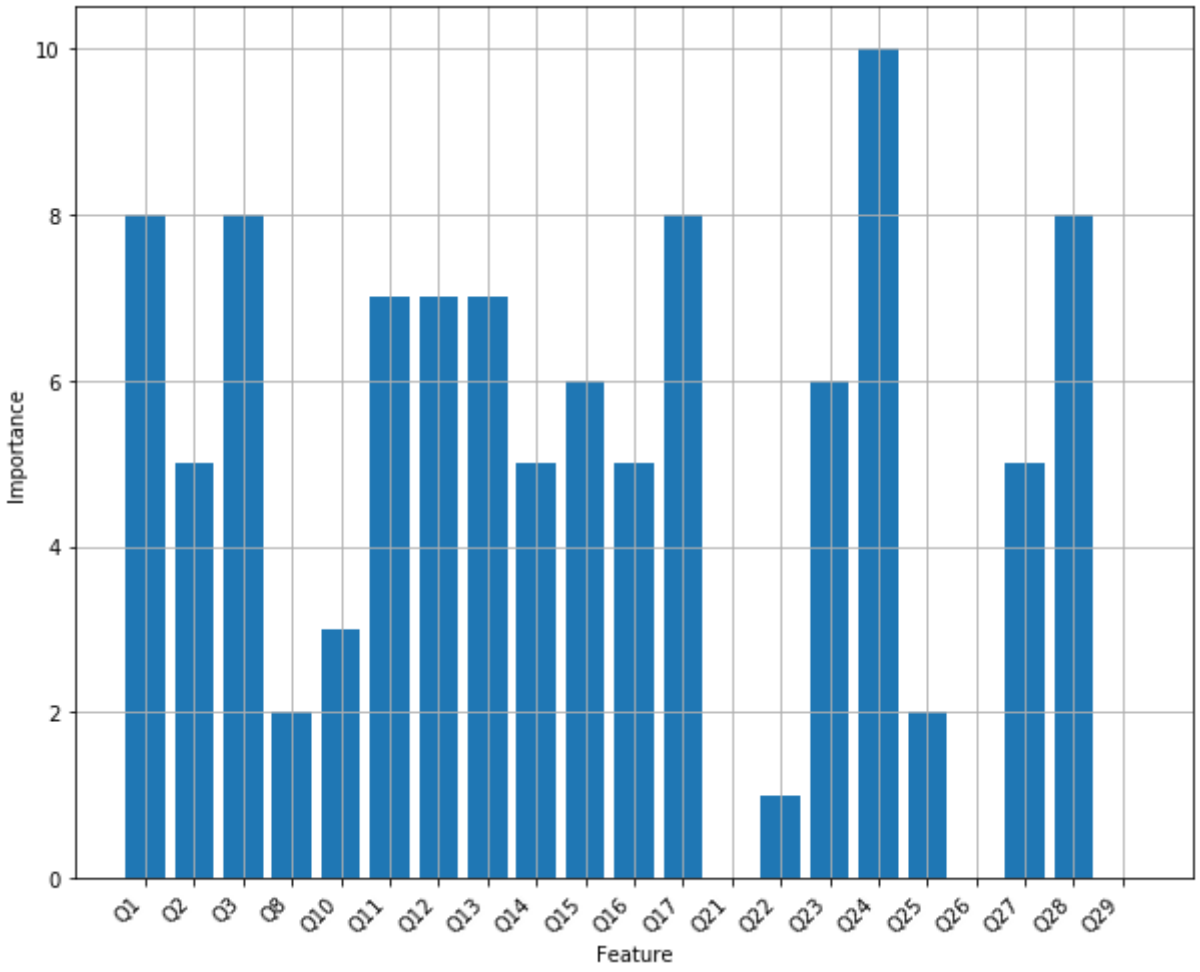


Figure 12

Feature scores of LightGBM with 20 trees (depth equals 4) in absence of Questions 19 and 20

Figure 13

Accuracy scores of CatBoost model as the number of trees changes between 4 and 19

Figure 14

Feature scores of CatBoost with 11 trees and depth equals 6

Figure 15

Feature scores of CatBoost with 11 trees and depth equals 6 in absence of Questions 19 and 20

Figure 16

The distribution of examples between two classes of the target variable (Question 18a)

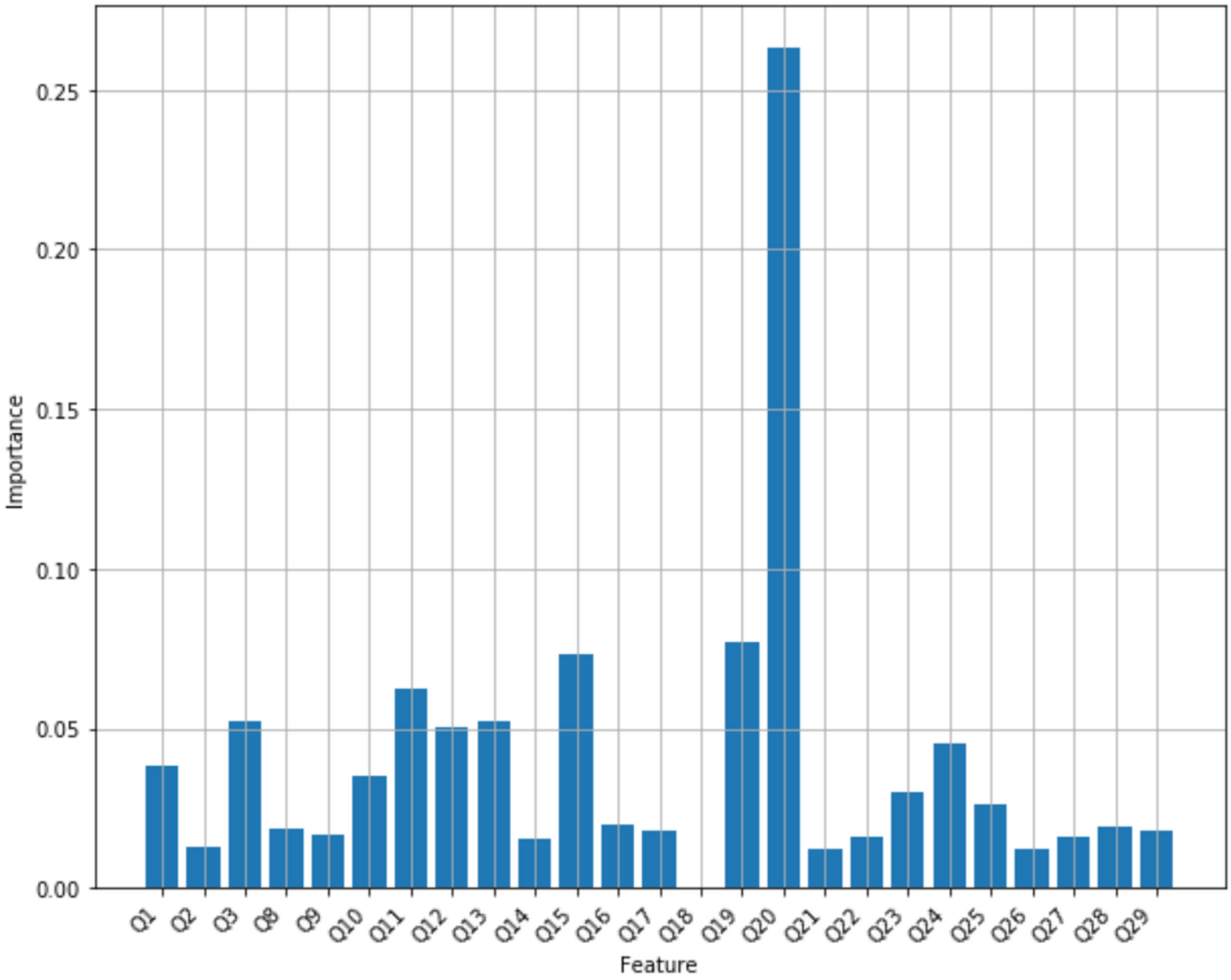


Figure 17

Feature scores of SMOTE Random Forest with 100 trees and depth equals 8

Figure 18

Feature scores of SMOTE Random Forest with 100 trees and depth equals 8

Figure 19

A summary of the methology and the results