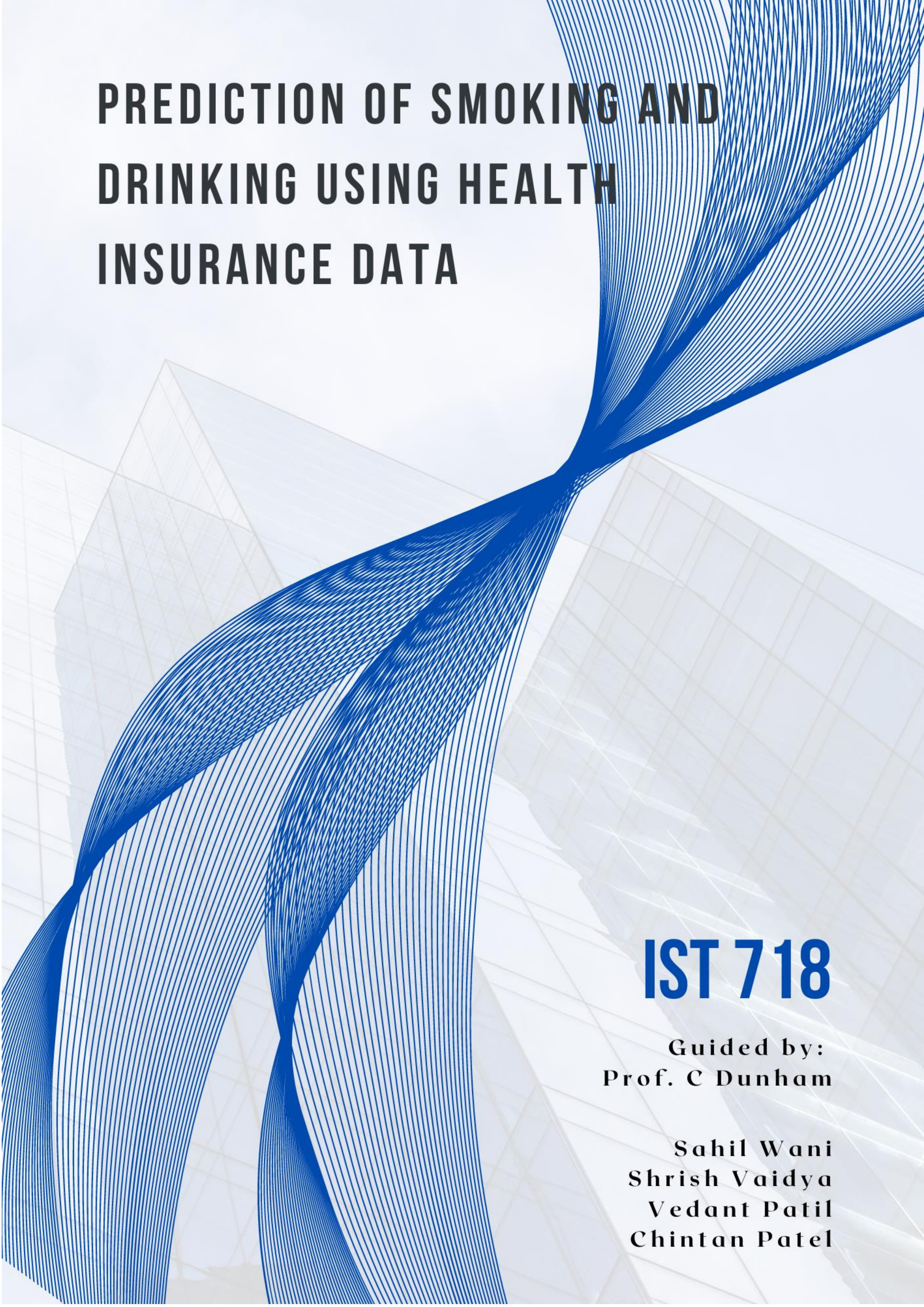


PREDICTION OF SMOKING AND DRINKING USING HEALTH INSURANCE DATA

IST 718

Guided by:
Prof. C Dunham

Sahil Wani
Shrish Vaidya
Vedant Patil
Chintan Patel

1. Introduction

Unveiling the true Power of Big Data represents a pioneering effort in Big Data Analytics, aiming to reshape healthcare sector, specifically in addressing the impacts of alcohol consumption and smoking. By leveraging predictive analytics, the project seeks to predict the likelihood of individuals adopting these habits. Focused on unraveling the underlying determinants, including personal life, as well as physiological indicators such as height, weight, hemoglobin, gender, urine creatine, and other 24 features, the interdisciplinary team strives to proactively intervene. This proactive approach aims to prevent the initiation of harmful behaviors by examining how these factors are influenced by unhealthy lifestyle choices. The research extends its predictive modeling to assess current smoking behavior, past history, or non-smoking status, incorporating lifestyle indicators and health metrics for comprehensive insights.

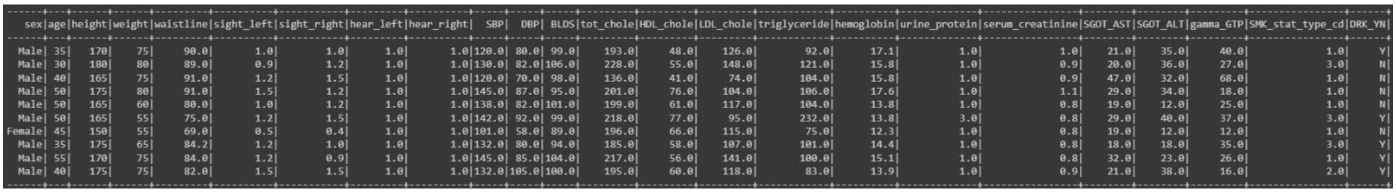
Beyond its implications for healthcare, project holds potential benefits for health management organizatins, gyms, and charities dedicated to combating smoking and drinking issues. The project's predictive insights offer employers cost savings on health insurance, aid gyms in optimizing health routines, and support charities in enhancing their interventions. Emphasizing a holistic approach, the project goes beyond statistical analysis, positioning itself as a trailblazer in predictive healthcare. By laying the foundation for a transformative paradigm, the research aims to instigate a proactive, preventative healthcare approach, offering individuals a second chance and contributing to a healthier and more informed society. (Md. Ileas Pramanik.)

2. Project Goal

The overarching objective of this research project is to construct a robust predictive model capable of accurately classifying individuals into distinct categories based on their smoking habits and alcohol consumption patterns. This advanced model not only classifies individuals as abstainers, current smokers, or former smokers but also forecasts their likelihood of engaging in alcohol consumption. Through a meticulous analysis of the intricate interplay between these critical lifestyle choices, our aim is to unveil underlying patterns and trends that significantly influence individual behavior. The invaluable insights derived from this project will play a pivotal role in shaping targeted health interventions, facilitating tailored awareness campaigns, and devising personalized healthcare strategies. Furthermore, the developed model stands to benefit insurance companies and health management organizations by furnishing crucial insights into individual health risks and predicting potential healthcare costs, thereby contributing to more informed decision-making in the realm of public health. (Robert H. Brook)

3. Dataset

The project's dataset, which was obtained from Kaggle(“Smoking and Drinking Dataset…”), is the Smoking and Drinking Dataset with Body Signal which was gathered from the Korean National Health Insurance Service. There are 24 columns and 9,91,346 rows in the data. The dataset offers a comprehensive analysis of health metrics linked to lifestyle habits, focusing on individuals' smoking and drinking behaviors. It includes various biological and physiological parameters, providing valuable insights into the effects of lifestyle choices on health and well-being. This data is especially vital for health professionals, policymakers, and researchers endeavouring to understand, predict, and manage health outcomes related to alcohol consumption and tobacco use.



```
------+---+------+------+---------+----------+-----------+---------+----------+-----+-----+-----+---------+---------+---------+------------+----------+-------------+----------------+--------+--------+---------+----------------+------+
   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|SMK_stat_type_cd|DRK_YN|
------+---+------+------+---------+----------+-----------+---------+----------+-----+-----+-----+---------+---------+---------+------------+----------+-------------+----------------+--------+--------+---------+----------------+------+
  Male| 35|   170|    75|     90.0|       1.0|        1.0|      1.0|       1.0|120.0| 80.0| 99.0|    193.0|     48.0|    126.0|        92.0|      17.1|          1.0|             1.0|    21.0|    35.0|     40.0|             1.0|     Y|
  Male| 30|   180|    80|     89.0|       0.9|        1.2|      1.0|       1.0|130.0| 82.0|106.0|    228.0|     55.0|    148.0|       121.0|      15.8|          1.0|             0.9|    20.0|    36.0|     27.0|             3.0|     N|
  Male| 40|   165|    75|     91.0|       1.2|        1.5|      1.0|       1.0|120.0| 70.0| 98.0|    136.0|     41.0|     74.0|       104.0|      15.8|          1.0|             0.9|    47.0|    32.0|     68.0|             1.0|     N|
  Male| 50|   175|    80|     91.0|       1.5|        1.2|      1.0|       1.0|145.0| 87.0| 95.0|    201.0|     76.0|    104.0|       106.0|      17.6|          1.0|             1.1|    29.0|    34.0|     18.0|             1.0|     N|
  Male| 50|   165|    60|     80.0|       1.0|        1.2|      1.0|       1.0|138.0| 82.0|101.0|    199.0|     61.0|    117.0|       104.0|      13.8|          1.0|             0.8|    19.0|    12.0|     25.0|             1.0|     N|
  Male| 50|   165|    55|     75.0|       1.2|        1.5|      1.0|       1.0|142.0| 92.0| 99.0|    218.0|     77.0|     95.0|       232.0|      13.8|          3.0|             0.8|    29.0|    40.0|     37.0|             3.0|     Y|
Female| 45|   150|    55|     69.0|       0.5|        0.4|      1.0|       1.0|101.0| 58.0| 89.0|    196.0|     66.0|    115.0|        75.0|      12.3|          1.0|             0.8|    19.0|    12.0|     12.0|             1.0|     N|
  Male| 35|   175|    65|     84.2|       1.2|        1.0|      1.0|       1.0|132.0| 80.0| 94.0|    185.0|     58.0|    107.0|       101.0|      14.4|          1.0|             0.8|    18.0|    18.0|     35.0|             3.0|     Y|
  Male| 55|   170|    75|     84.0|       1.2|        0.9|      1.0|       1.0|145.0| 85.0|104.0|    217.0|     56.0|    141.0|       100.0|      15.1|          1.0|             0.8|    32.0|    23.0|     26.0|             1.0|     Y|
  Male| 40|   175|    75|     82.0|       1.5|        1.5|      1.0|       1.0|132.0|105.0|100.0|    195.0|     60.0|    118.0|        83.0|      13.9|          1.0|             0.9|    21.0|    38.0|     16.0|             2.0|     Y|
------+---+------+------+---------+----------+-----------+---------+----------+-----+-----+-----+---------+---------+---------+------------+----------+-------------+----------------+--------+--------+---------+----------------+------+
```

Figure 1- Data Preview

4. Exploratory Data Analysis

While exploring the data and gaining the insights from data we checked the data for null values, duplicate values and removed it. We observed that there were lot of outliers in column features like hdl_cholesterol, ldl_cholesterol, haemoglobin, waistline, systolic blood pressure (SBP) and diastolic blood pressure (DBP). We employed the z-score method for outlier detection and removal in a dataset with numerical columns ("Z-Score for Outlier Detection in Python."). For each numerical column, the mean and standard deviation are calculated, and z-scores are computed based on these statistics with help of formula $Z = \frac{x - \mu}{\sigma}$ where x= data point, μ = mean of the data column and σ = std. deviation. Rows with z-scores exceeding a specified threshold (here, 3) are considered outliers and subsequently removed from the dataset.

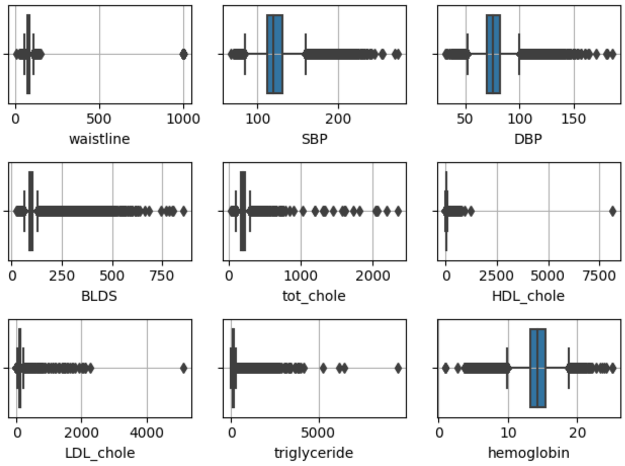


Figure 2: Before outlier deduction

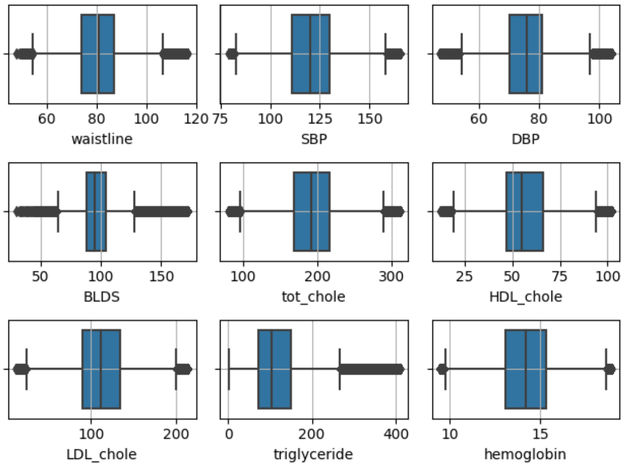


Figure 3: After outlier deduction

During the preliminary analysis of the data, we observed that even though there is almost similar number of male and female in the data, it has significant disparity in smoking habits between males and females, with a notably higher count of females who have never smoked. Such an imbalance can lead to biased outcomes in predictive modeling. To mitigate this, the 'sex' feature was excluded from the training dataset, aiming to enhance the model's performance by preventing it from relying on this skewed distribution.

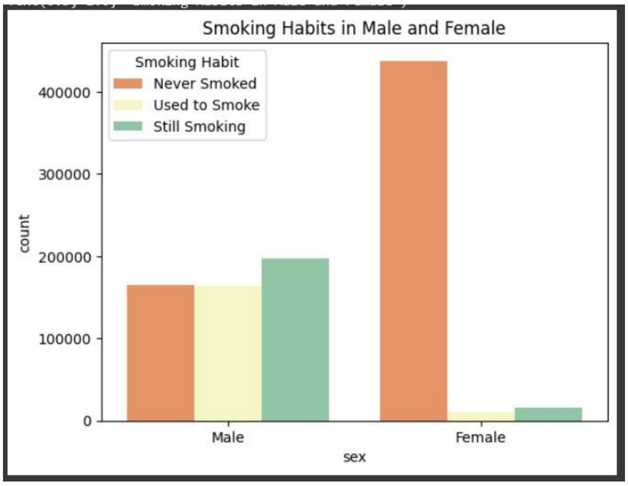


Figure 4 smoking habits of male and female

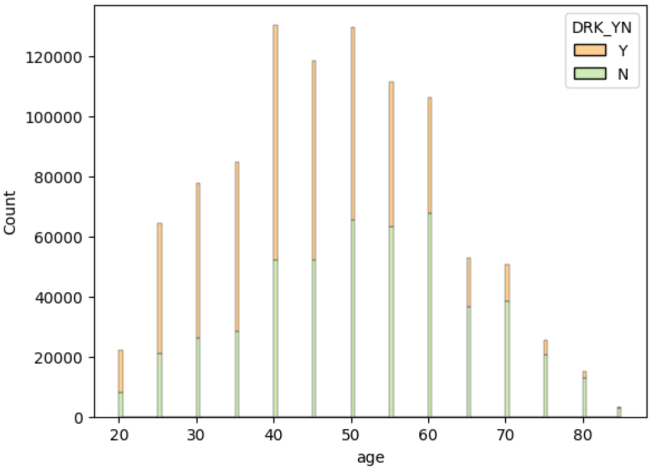


Figure 5: Drinking habits vs age

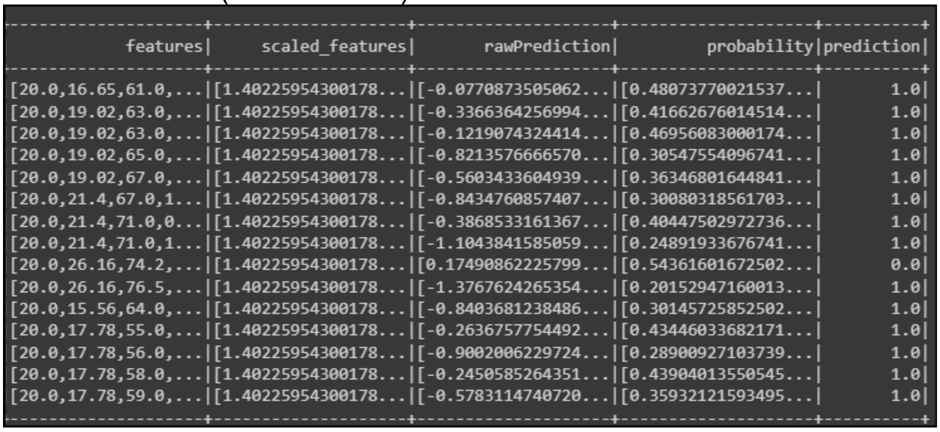
In figure 5 we can observe the drinking habit of people distributed age wise. The bar chart presents a comparison of two categories, designated as "Y" (Yes) and "N" (No), across various age groups for the variable "DRK_YN". It quantitatively illustrates the count of individuals within each age bracket who have either drinking (Y) or not drinking (N). Overall, there is inverse relation between drinking habits and age, as age increases number of people who drinks significantly reduces. Age bracket 30-60 has most drinkers.

5. Methodology

This study employed a comprehensive analytical framework to model and predict the target variable from our dataset. The dataset was first pre-processed to handle missing values, vectorize numerical features, scale the features and index categorical variables. Following preprocessing, we partitioned the data into training (80%) and testing (20%) sets to ensure the generalizability of our models.

For our first prediction of drinking habit (yes/no) we applied five different predictive modeling techniques for prediction:

- **Logistic Regression**: As a foundational modeling approach, logistic regression was applied due to its efficiency and interpretability when examining the relationship between a binary dependent variable and multiple independent variables. It was particularly useful for understanding the impact of each predictor on the likelihood of the outcome.(Keerthana V.)



```
--------------------+--------------------+--------------------+--------------------+----------+
            features|     scaled_features|       rawPrediction|         probability|prediction|
--------------------+--------------------+--------------------+--------------------+----------+
[20.0,16.65,61.0,...|[1.40225954300178...|[-0.0770873505062...|[0.48073770021537...|       1.0|
[20.0,19.02,63.0,...|[1.40225954300178...|[-0.3366364256994...|[0.41662676014514...|       1.0|
[20.0,19.02,63.0,...|[1.40225954300178...|[-0.1219074324414...|[0.46956083000174...|       1.0|
[20.0,19.02,65.0,...|[1.40225954300178...|[-0.8213576666570...|[0.30547554096741...|       1.0|
[20.0,19.02,67.0,...|[1.40225954300178...|[-0.5603433604939...|[0.36346801644841...|       1.0|
[20.0,21.4,67.0,1...|[1.40225954300178...|[-0.8434760857407...|[0.30080318561703...|       1.0|
[20.0,21.4,71.0,0...|[1.40225954300178...|[-0.3868533161367...|[0.40447502972736...|       1.0|
[20.0,21.4,71.0,1...|[1.40225954300178...|[-1.1043841585059...|[0.24891933676741...|       1.0|
[20.0,26.16,74.2,...|[1.40225954300178...|[0.17490862225799...|[0.54361601672502...|       0.0|
[20.0,26.16,76.5,...|[1.40225954300178...|[-1.3767624265354...|[0.20152947160013...|       1.0|
[20.0,15.56,64.0,...|[1.40225954300178...|[-0.8403681238486...|[0.30145725852502...|       1.0|
[20.0,17.78,55.0,...|[1.40225954300178...|[-0.2636757754492...|[0.43446033682171...|       1.0|
[20.0,17.78,56.0,...|[1.40225954300178...|[-0.9002006229724...|[0.28900927103739...|       1.0|
[20.0,17.78,58.0,...|[1.40225954300178...|[-0.2450585264351...|[0.43904013550545...|       1.0|
[20.0,17.78,59.0,...|[1.40225954300178...|[-0.5783114740720...|[0.35932121593495...|       1.0|
--------------------+--------------------+--------------------+--------------------+----------+
```

```
Area Under ROC Curve for Logistic Regression: 0.77
```

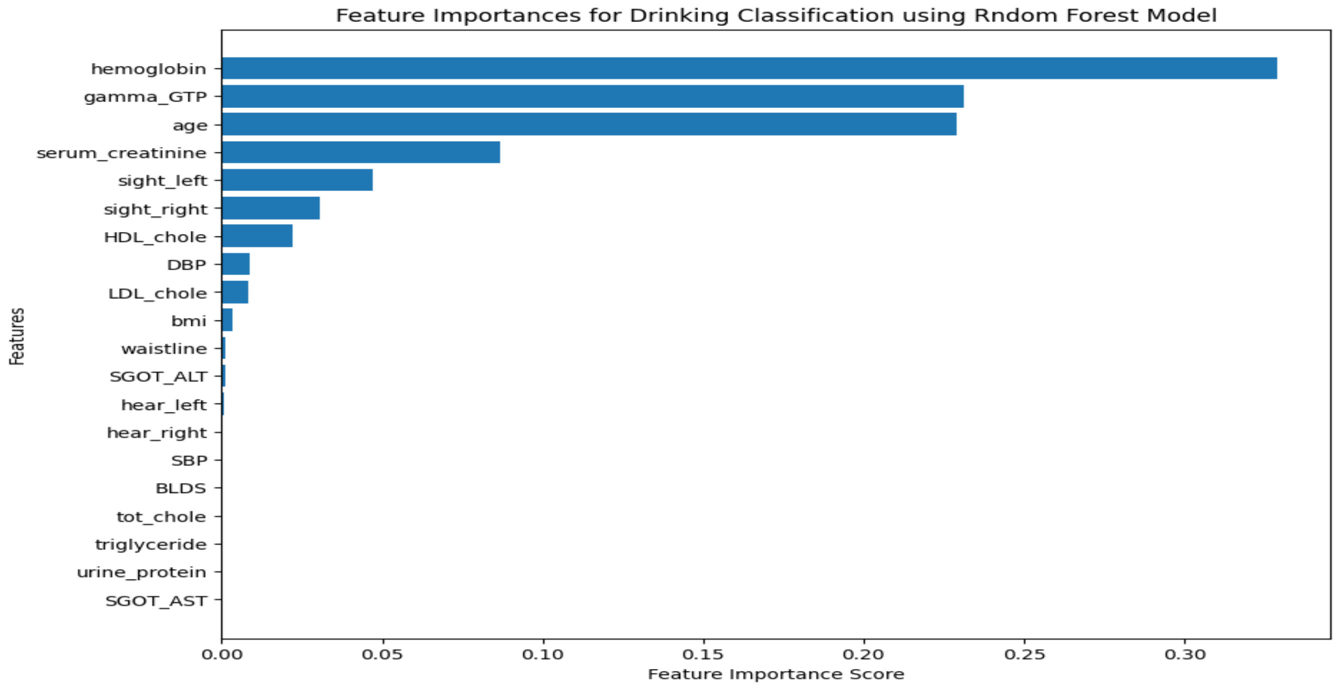
- **Decision Tree**: The decision tree algorithm was utilized for its interpretability and ease of visualization, allowing for an intuitive understanding of the decision-making process. It constructs and map out the possible outcomes of a series of related choices. This helped in understanding the data's branching patterns and making straightforward predictions. Cross-validation was used in the Decision Tree model for two key reasons: first, to adjust the model's complexity; second, to confirm the model's ability to predict outcomes across various data samples. To ensure that the model's performance was not influenced by the specifics of a single data split and that the ideal depth of the tree was chosen based on a constant error rate, we systematically divided the dataset into training and testing folds using 5-fold cross-validation.

```
Area Under ROC Curve for Drink Prediction: 0.64
```

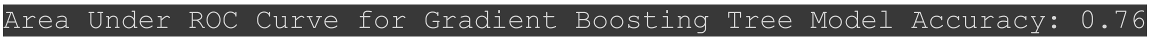
The feature importance graph for the Decision Tree model provided a visual representation of the relative importance of each predictor variable, highlighting which features most significantly influence the model's decisions. illustrates that gamma_GTP, haemoglobin, and age are the most significant predictors for drinking classification, with gamma_GTP being the most influential feature by a considerable margin.

- **Random Forest:** We used a random forest method to improve the predicted accuracy and get over the drawbacks of using only one decision tree. This ensemble approach builds many trees on different dataset subsamples, averaging the outcomes to increase prediction accuracy and reduce overfitting. Cross-validation was an essential step in the creation of the Random Forest algorithm model. We used a 5-fold cross-validation technique, which enabled us to assess the accuracy and stability of the forest's performance across several data sets. A more precise and trustworthy prediction model was produced as a result of this procedure, which also optimized the number of trees and the maximum amount of characteristics taken into account for splitting at each leaf node.

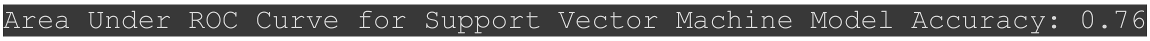
```
Area Under ROC Curve for Random Forest Model Accuracy: 0.75
```



- **Gradient Boosting:** Due to its ability to handle non-linear correlations and heterogeneous information, gradient boosting was selected. By maximizing a differentiable loss function, we were able to improve on regions where earlier models were lacking, and we gradually constructed the model from weak learners. In order to reduce prediction errors, parameters such as the learning rate, number of boosting stages, and tree depth were tuned using a thorough search process. For gradient boosting cross-validation was rigorously applied to assess model robustness and prevent overfitting. Specifically, 5-fold cross-validation was used to ensure that each model's performance was validated across multiple subsets of the data, providing a comprehensive evaluation of their predictive capabilities (Louis Chan).

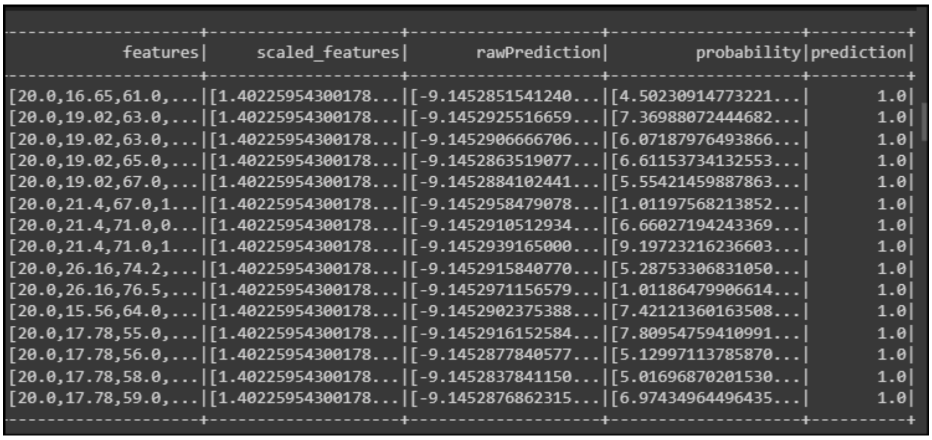

```
Area Under ROC Curve for Gradient Boosting Tree Model Accuracy: 0.76
```

- **Support Vector Machine (SVM)**: It works by finding the hyperplane that best separates the classes in the feature space. The optimal hyperplane was determined using kernel functions to transform the data into a higher dimension where it could be linearly separated. We used cross-validation for SVM (Support Vector Machine) as a crucial step to guarantee that overfitting was minimized and the model's generalization performance was appropriately evaluated. The data was divided into five parts using a 5-fold cross-validation technique. Of these parts, one was utilized as the validation set and the other five parts were used for training. This method helped optimize the hyperparameters, such as the max iterations and regularization parameter, and gave a reliable indicator of SVM's performance.


```
Area Under ROC Curve for Support Vector Machine Model Accuracy: 0.76
```

For our second prediction goal we aim to predict an individual's smoking status, categorized into three classes: 'never smoked', 'used to smoke', and 'still smoking'. The data underwent preliminary processing same as above for drinking habit prediction. The predictive modeling for the multiclass classification problem was conducted using the following methods:

- **Logistic Regression**: Multinomial logistic regression was employed to model the probability of the three distinct categories of smoking habits. This extension of binary logistic regression is adept at handling multiple classes and provided interpretability in terms of odds ratios for the predictors. We created pipeline of vectorized and scaled features as input for the model and ran the model on that pipeline (Keerthana V.).

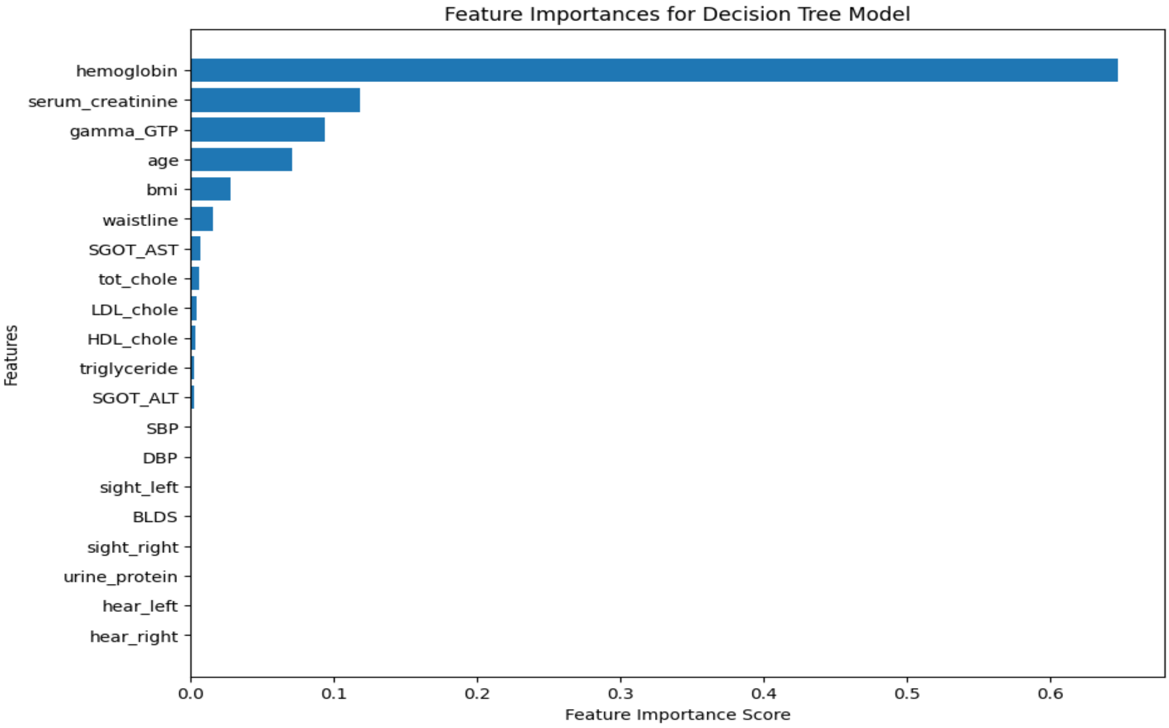


```
--------------------+--------------------+--------------------+--------------------+----------+
            features|     scaled_features|       rawPrediction|         probability|prediction|
--------------------+--------------------+--------------------+--------------------+----------+
[20.0,16.65,61.0,...|[1.40225954300178...|[-9.1452851541240...|[4.50230914773221...|       1.0|
[20.0,19.02,63.0,...|[1.40225954300178...|[-9.1452925516659...|[7.36988072444682...|       1.0|
[20.0,19.02,63.0,...|[1.40225954300178...|[-9.1452906666706...|[6.07187976493866...|       1.0|
[20.0,19.02,65.0,...|[1.40225954300178...|[-9.1452863519077...|[6.61153734132553...|       1.0|
[20.0,19.02,67.0,...|[1.40225954300178...|[-9.1452884102441...|[5.55421459887863...|       1.0|
[20.0,21.4,67.0,1...|[1.40225954300178...|[-9.1452958479078...|[1.01197568213852...|       1.0|
[20.0,21.4,71.0,0...|[1.40225954300178...|[-9.1452910512934...|[6.66027194243369...|       1.0|
[20.0,21.4,71.0,1...|[1.40225954300178...|[-9.1452939165000...|[9.19723216236603...|       1.0|
[20.0,26.16,74.2,...|[1.40225954300178...|[-9.1452915840770...|[5.28753306831050...|       1.0|
[20.0,26.16,76.5,...|[1.40225954300178...|[-9.1452971156579...|[1.01186479906614...|       1.0|
[20.0,15.56,64.0,...|[1.40225954300178...|[-9.1452902375388...|[7.42121360163508...|       1.0|
[20.0,17.78,55.0,...|[1.40225954300178...|[-9.1452916152584...|[7.80954759410991...|       1.0|
[20.0,17.78,56.0,...|[1.40225954300178...|[-9.1452877840577...|[5.12997113785870...|       1.0|
[20.0,17.78,58.0,...|[1.40225954300178...|[-9.1452837841150...|[5.01696870201530...|       1.0|
[20.0,17.78,59.0,...|[1.40225954300178...|[-9.1452876862315...|[6.97434964496435...|       1.0|
--------------------+--------------------+--------------------+--------------------+----------+
```

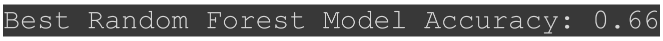
```
Logistic regression model accuracy: 0.67
```

- **Decision Tree:** To definitively differentiate between the three smoking statuses, a Decision Tree classifier was created. Understanding the interpretability of the model was essential to determining the decision rules that give rise to each smoking type. The tree was trimmed to avoid overfitting after it was cultivated according to the Gini impurity index. Three-fold cross-validation was used to assess the decision tree's effectiveness and determine the ideal tree depth, guaranteeing that the model was tested on a range of dataset subsamples.

```
Best Decision Tree Model Accuracy: 0.68
```



- **Random Forest:** This technique was chosen because it is better at handling class imbalances and because it is ensemble-based, combining the output of several decision trees to produce a single prediction. This approach yields a measure of feature relevance and is less prone to overfitting. Using 2-fold cross-validation, the number of trees, the maximum depth of the trees, and the maximum number of features evaluated for the best split were all adjusted to balance the model's bias and variance.



```
Best Random Forest Model Accuracy: 0.66
```

6. Conclusion

This project's exploration into Machine Learning methodologies has yielded substantial insights into behavioural patterns related to health. By deploying a suite of statistical models—namely Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, and SVM—we have uncovered significant correlations and influential factors within health indicators. The systematic application of cross-validation has underscored the predictive reliability of these models.

A pivotal outcome of our analysis was the identification of 'Haemoglobin' as a critical feature in predicting both smoking and drinking behaviours, as revealed by feature importance graphs generated from the Decision Tree and Random Forest models. This finding underscores the intertwined nature of health variables, such as the relationship between weight, waistline, blood pressure, and the roles of age, height, and cholesterol.

The predictive performance of the models varied, with Logistic Regression emerging as the most effective for drinking behaviour prediction, demonstrating an area under the curve of 77%. In contrast, the Decision Tree model exhibited the highest accuracy for smoking behaviour prediction at 68%. These outcomes highlight the necessity of selecting the appropriate model to match the data characteristics and the analytical objectives (Sarang Narkhede).

The potential for Machine Learning to navigate the complexities of health data has been clearly established through this research. Nevertheless, it also brings to light the criticality of model selection to fulfil specific data and goal-oriented needs. These insights pave the way for future research to expand upon our findings, whether through the integration of new predictive variables, the adoption of more intricate model architectures, or the application of our models to diverse datasets to test their generalizability.

Conclusively, this initiative has not only met its objective of crafting robust models for the prediction of health-related behaviours but also laid the groundwork for subsequent research endeavours. The advances made herein could propel healthcare analytics forward, enhancing decision-making processes, fostering better health practices, and contributing to improved business outcomes.

This conclusion encapsulates the project's achievements and key findings, while also pointing towards future research directions, maintaining the integrity of the information provided in the initial summary.

Citations

“Smoking and Drinking Dataset with body signal”. Kaggle.com, https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset

"Z-Score for Outlier Detection in Python." GeeksforGeeks, https://www.geeksforgeeks.org/z-score-for-outlier-detection-python/

Keerthana V. "My Great Learning. "Multinomial Logistic Regression." *My Great Learning, n.d.,* www.mygreatlearning.com/blog/multinomial-logistic-regression/

Louis Chan. "Understanding Gradient Boosting: A Data Scientist's Guide." *Towards Data Science,* n.d., https://towardsdatascience.com/understanding-gradient-boosting-a-data-scientists-guide-f5e0e013f441

Md. Ileas Pramanik. "Healthcare Informatics and Analytics in Big Data." *Expert Systems with Applications*, vol. 152, 2020, 113388. *ScienceDirect*, https://doi.org/10.1016/j.eswa.2020.113388

Robert H. Brook. "A Framework for Analyzing...". Volume 1, *Rand Corporation, 2006*,. www.rand.org/content/dam/rand/pubs/reports/2006/R2374.1.pdf.

Sarang Narkhede. "Understanding AUC-ROC Curve." Towards Data Science, n.d., https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5.