Smoke Signals: Predict Smokers by Vital Signs

Saketh Ragirolla 2021092
Saumil Lakra 2021097
Ishwar Babu 2021532
Rahul Oberoi 2021555

1. Abstract

The project aims to develop a Machine Learning model to predict whether a given individual is a smoker or not given his/her vital biological signals. By applying classical algorithms, the project seeks to uncover patterns in smoker classification tasks, creating a model that is both accurate and easily interpretable, with the potential for continuous improvement to achieve even better prediction outcomes. Git

2. Introduction

Smoking is the most common form of recreational drug abuse. Most of the this drug users are from developing countries. It is the leading cause of cancer, stroke, heart attack, and lung disease and it can also escalate the risk of depression and anxiety. This can decrease the risk of tobacco use. In our dataset, after analyzing we have inferred that young people smoke more. Timely and accurate diagnosis has to be done to prevent such serious health consequences.

3. Literature Survey

3.1. Consequences of Smoking for Body Weight, Fat Distribution, and Insulin Resistance [1]

The paper tries to study the relations between smoking, body weight, body fat distribution(fat along the waistlines) and insulin resistance. Research has shown that smoking leads to a reduction in body weight during the period of active smoking, but this is often followed by weight gain after stopping the habit of smoking. Smoking influences the distribution of body fat, increasing the accumulation of fat around the abdomen region. The paper doesn't use any machine learning model to identify the correlation.

3.2. Predicting Smoking Status Using Machine Learning Algorithms and Statistical Analysis [2]

The study aims to predict the smoking status of an individual from medical data. The research was conducted using the following machine learning models: Naive Bayes, Multilayer Perceptron, Logistic Regression, J48(?) and Decision Table. Among the five algorithms tested, Logistic Regression emerged as the most effective model, outperforming and, achieving an accuracy of approximately 83%. For our study, we have considered Gaussian Naive Bayes as baseline model for comparison.

3.3. Artificial neural network machine learning prediction of the smoking behavior and health risks perception of Indonesian health professionals [3]

The research paper aims to predict the smoking status of Health Professionals (HPs) using Artificial Neural Networks(ANN) machine learning method. The study was aimed for second-hand smoker, to study the affect that it indirectly has on them. It can cause various diseases such as stroke, respiratory diseases. The scores obtained are as follows: Precision(89%), Accuracy(81%), Recall(85%),and AUC(70%). For our study, we have taken 55,692 data points which after balancing becomes 40,000 rows.

4. Dataset2

The dataset for this project was obtained from Kaggle and originally contained 55,692 rows and 27 columns. The dataset had a higher number of non-smokers compared to smokers, so 20,000 non-smoker entries were randomly removed to balance the classes for analysis.

4.1. Plots

4.1.1 Correlation Heat Map

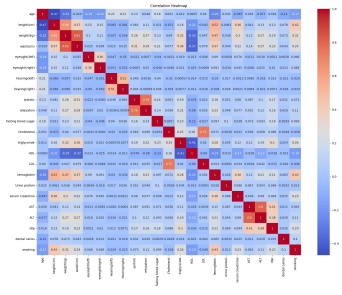

Figure 1. Correlation Heatmap

From the correlation heatmap, it is quite clear that there are positive correlations between "weight and waist", "cholestrol and LDL", "systolic and diastolic blood pressure" etc.

4.1.2 Box Plot

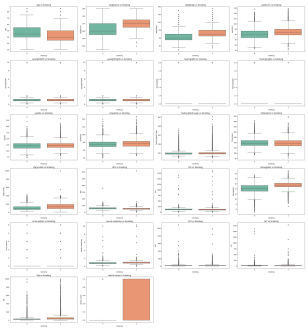

Figure 2. Box Plot

From the box plot, we can infer that young people tend to be smokers, HDL (Good Cholestrol) is lower in smokers, additionally we can get an idea about the outliers in the numerical features from this plot.

4.1.3 Pair Plot

Figure 3. Pair Plot

From the pair plots, we can see that there are some clusters which form for features such as "hemoglobin vs systolic" and "hemoglobin vs relaxation", non smokers tend to have lower values for these

4.1.4 Violin Plot

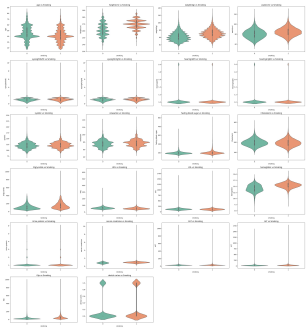

Figure 4. Violin Plot

From the violin plots, we can infer that the weight distribution of smokers is a bit more concentrated towards the heavier side, triglyceride levels of non-smokers are also lower (higher proportion of people) than that of smokers.

4.1.5 Histograms

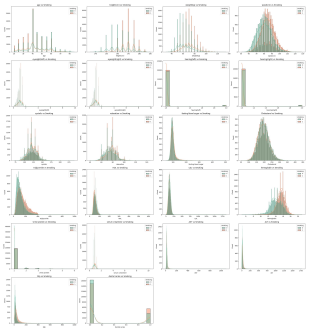

Figure 5. Histogram

From the histograms, we can infer that some features (like triglycerides, HDL, LDL, CRP, and dental caries) show mild differences between smokers and non-smokers. Specific variables like triglycerides, cholesterol levels, and dental health may offer insights into the adverse effects of smoking.

5. Methodology

5.1. Exploratory Data Analysis (EDA)

Smoking is our target label. 0 means a non-smoker and 1 means a smoker. After analyzing our dataset, we found out the there are no NULL values in the dataset. Oral column in the dataset had all of its entries are Y(yes), so we have dropped that column.

5.2. Dataset Balancing

An unbalanced dataset can lead to poor performance especially on the minority class because while training the model focuses more on the majority class. In our dataset, Non-smoker class had more count than the smoker class, so to solve this problem we have brought down the count of majority class Non-smoker) to match with the count of minority class (Smoker) by randomly removing rows from the majority class.

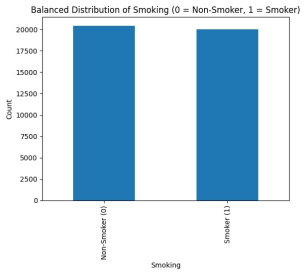

Figure 6. Balanced Distribution

5.3. Data Encoding & Normalization

The dataset contains 3 categorical columns gender, tartar and smoking (we are not considering oral column because we have already dropped it), out of which gender and tartar column are non-numeric. To make them understandable by our machine

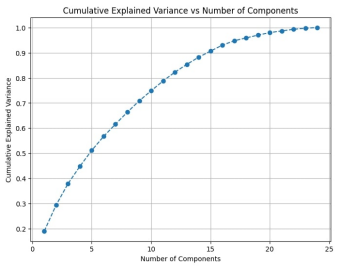

Figure 7. Cumulative Explained Variance vs Number of Features

learning algorithm, we have encoded them using label encoding.

Numeric data has been normalized using Standard Scaler. Standardization helps to rescale the data such that it lies between a certain range. This is helpful when the dataset has features that are spread across different scales. In our dataset, height is in centimeters whereas weight is in kilogram(kg). So to handle this, normalization is done.

5.4. Outlier Detection

We noticed the outliers from the box plots and have decided to remove the outliers from the following columns with a threshold z value of 3 (This is currently arbitrary and we shall experiment with it in the future).

5.5. Principal Component Analysis(PCA)

We have used cumulative variance to find the ideal n(n=17). This n covered 95% of the total variance.

5.6. Models Used

5.6.1 Gaussian Naive Bayes

Gaussian Naive Bayes was chosen because, after standardizing the dataset, the feature values approximately follow a Gaussian (normal) distribution. Additionally, it is a simple and efficient classifier and provides a solid baseline for comparison with other, more complex models.

5.6.2 Decision Trees

We used Decision Trees for their ability to classify both smokers and non-smokers based on health indicators such as blood test results. Decision Trees are interpretable, making it easier to understand the classification decisions making them versatile for this dataset. However, they tend to overfit, hence the use of Random Forests.

5.6.3 Random Forest

Random Forest was selected due to its robustness and its ability to handle imbalanced datasets effectively. This model is also known for its ability to capture non-linear relationships between features, which is critical when predicting smoking behavior from a diverse set of health indicators.

5.6.4 Logistic Regression

Logistic Regression was included because of its simplicity and effectiveness in binary classification tasks like predicting smoker vs. non-smoker status. Although it is a linear model, Logistic Regression provides a strong baseline and performs well when the data has a linear decision boundary. Additionally, it is highly interpretable, making it useful for understanding the relationships between features and smoking status.

5.6.5 Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) was used for its ability to learn complex non-linear relationships in the data. In this study, MLP was chosen because of its capacity to capture hidden patterns between health indicators and smoking status, even when relationships are not linear.

5.6.6 Support Vector Classifier

Support Vector Classifier (SVC) was chosen because of its strength in handling high-dimensional data. In this study, SVC was used to classify smokers and non-smokers by finding the optimal hyperplane that maximizes the margin between the two classes. This model is particularly effective when the data is linearly separable.

5.6.7 AdaBoost

AdaBoost was chosen for its ability to improve the performance of weak classifiers by sequentially applying them to different subsets of data, where misclassified samples are given higher weights in subsequent iterations. By combining multiple weak learners, AdaBoost builds a strong predictive model that is robust to noise and performs well even with imbalanced datasets.

6. Hyperparameter Tuning using Grid Search

Grid Search was applied to optimize models of Gaussian Naive Bayes, Random Forest, Linear SVC, Logistic Regression, ADABoost and Decision Tree with data types of Vanilla, K-Fold, Outlier Removed and PCA. Reduced parameter grids were utilized to balance computational efficiency with thorough exploration. The outlier-removed dataset, PCA-transformed data, was integrated into the Grid Search pipeline to ensure comprehensive evaluation and to address challenges. This systematic approach enabled the identification of well-calibrated models tailored to the task of classifying smokers and non-smokers.

7. Results and Analysis

This project aims to develop a Machine Learning model to accurately predict whether a given individual is a smoker or not given his/her vital biological signals. By applying classical algorithms, the project seeks to uncover patterns in smoker classification tasks, creating a model that is both accurate and easily interpretable, with the potential for continuous improvement to achieve even better prediction outcomes.

Table 1:

Model	Method	Macro F1	W. F1	W. Precision	W. Recall	Accuracy	Train Loss	Val Loss	Test Loss
GNB	Vanilla	0.70|0.73	0.71|0.73	0.75|0.74	0.70|0.73	0.70|0.73	1.29|1.30	1.30|1.33	1.38|1.29
GNB	K-Fold	0.70|0.73	0.71|0.73	0.75|0.75	0.70|0.74	0.70|0.74	1.28|1.31	1.29|1.33	-|-
GNB	No Outlier	0.74	0.74	0.75	0.74	0.74	1.33	1.41	1.27
GNB	No Outlier + PCA	0.72	0.72	0.73	0.72	0.72	0.63	0.64	0.63
DT	Vanilla	0.76|0.76	0.78|0.76	0.78|0.76	0.78|0.76	0.78|0.76	2.22|2.22	7.76|8.89	8.12|8.63
DT	K-Fold	0.75|0.75	0.77|0.75	0.77|0.75	0.77|0.75	0.77|0.75	0.00|0.00	8.30|9.31	-|-
DT	No Outlier	0.76	0.76	0.76	0.76	0.76	2.22	8.74	8.44
DT	No Outlier + PCA	0.75	0.75	0.75	0.75	0.75	2.22	9.36	9.04
DT	Grid Search	0.77	0.77	0.80	0.77	0.77	0.48	0.51	0.48
RF	Vanilla	0.82|0.82	0.83|0.82	0.83|0.83	0.83|0.82	0.83|0.82	0.10|0.11	0.37|0.40	0.39|0.39
RF	K-Fold	0.81|0.81	0.82|0.81	0.82|0.82	0.82|0.81	0.82|0.81	0.10|0.11	0.40|0.41	-|-
RF	No Outlier	0.82	0.82	0.83	0.82	0.82	0.11	0.40	0.38
RF	No Outlier + PCA	0.80	0.80	0.81	0.81	0.81	0.12	0.43	0.41
RF	Grid Search	0.83	0.83	0.84	0.83	0.83	0.11	0.40	0.38
LR	Vanilla	0.73|0.75	0.74|0.75	0.75|0.79	0.74|0.76	0.74|0.76	0.47|0.48	0.47|0.49	0.48|0.48
LR	K-Fold	0.73|0.76	0.75|0.76	0.75|0.79	0.75|0.76	0.75|0.76	0.47|0.48	0.47|0.49	-|-
LR	No Outlier	0.76	0.76	0.79	0.77	0.77	0.48	0.49	0.47
LR	No Outlier + PCA	0.74	0.74	0.75	0.74	0.74	0.52	0.53	0.51
LR	Grid Search	0.77	0.77	0.80	0.77	0.77	0.48	0.49	0.47
MLP	Vanilla	0.75|0.77	0.76|0.77	0.76|0.77	0.76|0.77	0.76|0.77	0.40|0.40	0.46|0.49	0.47|0.47
MLP	K-Fold	0.74|0.77	0.76|0.77	0.76|0.78	0.76|0.77	0.76|0.77	0.40|0.40	0.47|0.49	-|-
MLP	No Outlier	0.77	0.77	0.77	0.77	0.77	0.40	0.49	0.48
MLP	No Outlier + PCA	0.75	0.75	0.75	0.75	0.75	0.44	0.51	0.50
MLP	Grid Search	0.77	0.77	0.78	0.78	0.78	0.43	0.47	0.46
SVC	Vanilla	0.73|0.75	0.74|0.75	0.75|0.80	0.74|0.76	0.74|0.76	0.47|0.48	0.47|0.49	0.48|0.48
SVC	K-Fold	0.73|0.75	0.75|0.75	0.75|0.80	0.75|0.76	0.76|0.76	0.47|0.48	0.47|0.49	-|-
SVC	No Outlier	0.76	0.76	0.80	0.76	0.76	0.48	0.49	0.47
SVC	No Outlier + PCA	0.75	0.75	0.75	0.75	0.75	0.52	0.53	0.51
SVC	Grid Search	0.77	0.77	0.80	0.77	0.77	0.48	0.49	0.47
AB	Vanilla	0.74|0.78	0.75|0.78	0.76|0.79	0.75|0.78	0.75|0.78	0.67|0.68	0.67|0.68	0.67|0.68
AB	K-Fold	0.75|0.76	0.77|0.74	0.75|0.77	0.76|0.78	0.75|0.77	0.67|0.68	0.67|0.68	-|-
AB	No Outlier	0.77	0.77	0.79	0.78	0.78	0.68	0.68	0.68
AB	No Outlier + PCA	0.75	0.75	0.75	0.75	0.75	0.68	0.68	0.68
AB	Grid Search	0.78	0.79	0.78	0.78	0.78	0.68	0.68	0.68

GNB: Gaussian Naive Bayes DT: Decision Tree RF: Random Forest LR: Logistic Regression AB: AdaBoost

Imbalanced Dataset |Balanced Dataset

Table 2:

Feature	Description	Data-Type	Unit of Measurement	Value Range
gender	Gender	Categorical Code	-	1: Female, 2: Male

Continued on next page

Table 2: (Continued)

age	Age (5-year gap)	integer	Years	Any Integer Value ≥ 0
height	Height	integer	Centimeters (cm)	Any Integer Value > 0
weight	Weight	float	Kilograms (kg)	Any Floating Value > 0
waist	Waist Circumference Length	integer	Centimeters (cm)	Any Integer Value > 0
eyesight (right)	Eyesight (Right)	float	-	0: Normal, 1: Impaired
hearing (right)	Hearing (Right)	float	-	0: Normal, 1: Impaired
systolic	Systolic Blood Pressure	integer	Millimeter(s) of Mercury (mmHg)	Any Integer Value > 0
relaxation	Diastolic Blood Pressure	integer	Millimeter(s) of Mercury (mmHg)	Any Integer Value > 0
fasting blood sugar	Fasting Blood Sugar	integer	mg/dL	Any Integer Value > 0
cholesterol	Total Cholesterol	integer	mg/dL	Any Integer Value > 0
triglyceride	Triglyceride	integer	mg/dL	Any Integer Value > 0
HDL	HDL Cholesterol	integer	mg/dL	Any Integer Value > 0
LDL	LDL Cholesterol	integer	mg/dL	Any Integer Value > 0
hemoglobin	Hemoglobin	float	g/dL	Any Floating Value > 0
urine protein	Urine Protein	Categorical Code	-	0: Normal, 1: Abnormal
serum creatinine	Serum Creatinine	float	mg/dL	Any Floating Value > 0
AST	AST (Glutamic Oxaloacetic Transaminase)	integer	U/L	Any Integer Value > 0
ALT	ALT (Glutamic Oxaloacetic Transaminase)	integer	U/L	Any Integer Value > 0
Gtp	-GTP (Gamma-Glutamyl Transferase)	integer	U/L	Any Integer Value > 0
smoking	Smoking Status	Categorical Code	-	0: Non-Smoker, 1: Smoker

8. Conclusion

This project shows the importance of structured machine learning pipeline from preprocessing to hyperparameter tuning, in achieving reliable predictions. Ensemble learning methods like Random Forest performed best to classify smokers and non-smokers.

Week	Tasks	Team
1 - 2	Data Cleaning	Saketh, Saumil
3	Data Pre-Processing	Ishwar, Rahul
4	Feature Extraction and Analysis	Saketh, Ishwar
5	Exploratory Data Analysis & Data Visualization	Saumil, Rahul
6-8	Deploying ML models like Support Vector Classifier, Logistic Regression, etc.	All members
9-10	Hyper parameter Tuning using Grid Search	All members
10-11	Reporting the scores	All members

Table 3. Individual Tasks

References

[1] Arnaud Chiolero, David Faeh, Fred Paccaud, and Jacques Cornuz. Consequences of smoking for body weight, body fat distribution, and insulin resistance. *The American Journal of Clinical Nutrition*, 87(4):801–809, Apr 2008.

[2] Charles Frank, Asmail Habach, Raed Seetan, and Abdullah Wahbeh. Predicting smoking status using machine learning algorithms and statistical analysis. *Advances in Science, Technology and Engineering Systems Journal*, 3:184–189, 2018.

[3] D Nuryunarsih, O Okatiranti, and L Herawati. Artificial neural network machine learning prediction of the smoking behavior and health risks perception of indonesian health professionals. *Environmental Analysis Health and Toxicology*, 38(1):e2023003–0, Mar 2023. PMID: 37100398; PMCID: PMC10195675.