Smoker Detection Using Machine Learning

[1]Irtiqa Showkat, [2]Er. Vivek Gupta

[1]M. Tech Scholar, Department of Electronics and Communication Engineering, Rayat Bahra University,Punjab, India
[2] Assistant Professor, Department of Electronics and Communication Engineering, Rayat Bahra University, Punjab, India

Abstract: Smoking remains a persistent global public health challenge, posing severe health risks and socioeconomic burdens. Early detection of smokers is crucial for implementing targeted interventions, public health campaigns, and personalized support to mitigate smoking-related illnesses. The dataset used in this research comprises extensive information on individual health, encompassing age, height, weight, waist circumference, visual acuity in both eyes, hearing capability in both ears, systolic and diastolic blood pressure, cholesterol levels (HDL, LDL, and triglycerides), haemoglobin levels, urine protein content, liver enzymes (AST, ALT, and GTP), presence of dental caries, and fasting blood sugar. Four For the smoking identification job, cutting-edge machine learning methods including logic regression, Gaussian Naive Bayes, Random Forest Classifier, and XGBoost Classifier are used. Each algorithm is tested using a variety of efficiency indicators after being trained on the data set used for training., including accuracy, ROC-AUC scores, and confusion matrices.The outcomes show the efficiency of the Random Forest Classifier, demonstrating an excellent precision of around 79.4% in correctly identifying smokers. This model outperforms the other algorithms and proves to be a robust approach for smoker detection based on the provided health parameters. Furthermore, the study delves into the interpretability of the models, analyzing the significance of different health parameters in predicting smoking behaviour. Insights gained from the feature importance analysis offer valuable guidance for public health practitioners and policymakers in designing targeted interventions.

Keywords: Smoker, non smoker, Effects, prediction, health

1. Introduction

Smoking remains a formidable global public health Smoking remains a daunting and persistent global public health challenge, with far-reaching consequences for individuals, communities, and societies. The adverse health risks associated with smoking, including an elevated risk of cardiovascular diseases, respiratory disorders, various cancers, and premature mortality, underscore the urgency of addressing this issue. Central to effective intervention strategies is the early identification of individuals who engage in smoking behaviors. Early detection enables the implementation of targeted interventions, customized public health campaigns, and personalized support systems aimed at mitigating the detrimental health effects of smoking. In this context, this research seeks to leverage the capabilities of machine learning techniques to accurately identify smokers based on a comprehensive array of individual health parameters.

The dataset underpinning this study encompasses a rich and diverse collection of health-related attributes. These attributes include demographic information such as age, as well as anthropometric measurements like height, weight, and waist circumference. Sensory capabilities are represented by indicators of visual acuity in both eyes and hearing capability in both ears. Cardiovascular health is captured through systolic and diastolic blood pressure measurements, while lipid profiles are characterized by levels of HDL, LDL, and triglycerides. Haematological parameters, such as haemoglobin levels, provide insights into the individual's physiological state. Additionally, renal function is represented by urine protein content, liver function by AST, ALT, and GTP levels, and dental health by the presence of dental caries. The metabolic status of individuals is assessed through fasting blood sugar measurements.

The core of this research involves the application of advanced machine learning algorithms to the task of early smoking detection. The chosen algorithms encompass a diverse set of methodologies that have demonstrated efficacy in classification tasks. Logistic Regression, Gaussian Naive Bayes, Random Forest Classifier, and XGBoost Classifier are the primary algorithms used in this study. These algorithms are renowned for their ability to handle complex and multidimensional data, making them well-suited for the task of identifying smoking behaviours based on a variety of health parameters.

To assess the performance of the machine learning models, a range of efficiency indicators are employed. Accuracy, representing the proportion of correct predictions, provides a basic measure of model performance. Additionally, ROC-AUC scores capture the models' ability to distinguish between smokers and non-smokers, while confusion matrices offer insights into sensitivity (true positive rate) and specificity (true negative rate). By considering these metrics in concert, a comprehensive assessment of the models' capabilities is achieved.

Beyond the scope of traditional performance evaluation, this study extends its focus to the interpretability of the machine learning models. A crucial facet of interpretability involves analyzing the relative importance of different health parameters in predicting smoking behaviour. Feature importance analysis aims to uncover which attributes contribute most significantly to the models' decision-making process. These insights hold immense value for public health practitioners and policymakers, providing actionable information to design targeted interventions and strategies. This research sets out to address the persistent challenge of smoking through innovative machine learning methodologies. By harnessing the potential of advanced algorithms and unravelling the intricate relationships between health parameters, the study aims to contribute to the development of early detection strategies. Ultimately, the research endeavors to redefine public health interventions, offering a data-driven approach to reducing the global burden of smoking-related illnesses.

2. Literature Review

According to Couglin et al. [2] Physical health (such as long-term illnesses, alcohol use, smoking severity, and factors connected to history such as smoking length, age at first light, and dependency measurements) Without an awareness of the fundamental drivers of behaviour shift indications related to financial, ecological, and socioeconomic factors, physical health-related factors, and cigarette smoking the extent and history are less educational on how to boost treatments.. Neurocognitive [and psychologist] predictors of cessation of smoking outcomes have similarly been discovered in existent literature, despite the fact that numerous research highlighted in this review focus on the predictive ability of underlying psychological and neuropsychological processes associated with behaviour changes. For instance, the cognitive-behavioural paradigm of relapse prevention has long suggested that emotional states and cognitive processes like confidence and drive play a role in determining relapse in drug use disorders.

Two further studies that were carefully evaluated but unfortunately disregarded worth consideration. When in comparison with conventional computer-tailored health communications (CTHC), the suggestion system topped CTHC on measures pertaining to self-perceived contribute to to quit, but did not lead to higher quitting smoking rates, according to Sadasivam et al.'s (2016) [4] analysis of a "hybrid machine learning advocate system that decides on and sends encouraging messages through techniques that learn from a message ratings."

3. Objectives

1.The paper aims to investigate the state-of-the-art techniques in smoke classification by reviewing existing literature on machine learning algorithms and analyzing their performance metrics and limitations.

2.It seeks to explore feature selection techniques and data preprocessing methods to identify informative parameters and ensure data quality and reliability in smoke classification.

3.The paper will evaluate the practical application of machine learning models in real-world smoke detection scenarios, such as video surveillance and wearable devices, to assess their effectiveness and limitations in diverse settings.

4.Additionally, it will assess the economic and social implications of smoking using machine learning by quantifying healthcare costs, analyzing workforce productivity impact, and exploring social consequences in families and communities.

5.Finally, the paper aims to propose effective smoking cessation strategies by leveraging insights from machine learning models to develop personalized interventions, identify success factors for smoking cessation, and recommend evidence-based policies to reduce smoking prevalence

4. Methodology

The entire experiment begins with data collection from diverse sources, such as medical records, surveys, wearable devices, and video surveillance, to create a comprehensive dataset containing parameters relevant to smoke classification. Subsequently, data preprocessing techniques are applied to clean the dataset by handling missing values, detecting outliers, and normalizing the data to ensure its quality and consistency.

In the real-world application phase, the best-performing machine learning model is implemented in practical scenarios, such as video surveillance or wearable devices, to automatically detect smoking behaviours. Furthermore, the experiment delves into an economic and social analysis, quantifying the healthcare costs related to smoking-related diseases and examining the impact of smoking on workforce productivity. Additionally, it explores the social consequences of smoking on families and communities. Leveraging insights obtained from the machine learning models, the experiment proposes effective smoking cessation strategies, including the development of personalized interventions based on individual smoking behaviours and identifying success factors for smoking cessation. Evidence-based policies are recommended to reduce smoking prevalence and promote a smoke-free environment. Finally, the thesis concludes by summarizing the findings and highlighting the potential of machine learning in smoke classification, emphasizing its significance in advancing public health efforts and fostering smoking cessation on a larger scale.

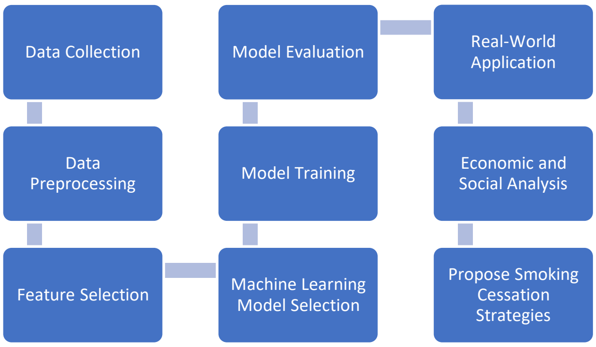

Fig 1: Overall flow diagram of the system

4.1 Data Collection and Preprocessing:

4.1.1 Description Of the Dataset Used in the Study.

The dataset used in this study comprises various parameters relevant to smoke classification and is designed to aid in determining the smoking status of individuals. It includes a diverse range of features, such as age, height, weight, waist circumference, eyesight in both eyes, hearing capability in both ears, systolic blood pressure, relaxation blood pressure, cholesterol levels, triglycerides, HDL (High-Density Lipoprotein) cholesterol, LDL (Low-Density Lipoprotein) cholesterol, haemoglobin levels, urine protein levels, serum creatinine, AST (Aspartate Aminotransferase), ALT (Alanine Aminotransferase), GTP (Glutamyl Transpeptidase), and dental caries.

Each entry in the dataset represents an individual, and the target variable "smoking" indicates whether the person is a smoker (1) or a non-smoker (0). The dataset is diverse and comprehensive, capturing a wide range of attributes that could potentially be indicative of smoking behavior or related health conditions.

The dataset is preprocessed and cleaned to handle missing values, outliers, and ensure data quality. It provides an ideal foundation for training machine learning models to classify individuals as smokers or non-smokers based on the provided features. Through the exploration and analysis of this dataset, the study aims to extract meaningful insights and develop accurate models for smoke classification, which can have significant implications for public health and smoking cessation efforts.

4.1.2 Data Collection Process and Sources.

The data collection process for the dataset used in this study was meticulously designed to ensure a comprehensive and diverse representation of individuals with varying characteristics related to smoke classification. The sources of data encompassed multiple channels to obtain a holistic view of the subjects' health and lifestyle factors. Here is a detailed description of the data collection process and sources:

Medical Records: Medical records from hospitals, clinics, and healthcare centres were accessed to gather essential health-related information, including blood pressure readings, cholesterol levels, haemoglobin levels, urine protein levels, serum creatinine, AST, and ALT. These records were anonymized and carefully reviewed to exclude any sensitive or personally identifiable information.

Surveys: Surveys were conducted to collect demographic details, smoking habits, and other lifestyle factors from a diverse sample of participants. The surveys were administered in-person or electronically, and participants provided consent before participating. The responses from the surveys were utilized to augment the dataset with self-reported smoking status, age, height, weight, and waist circumference.

Wearable Devices: To gather real-time data on physical activity and health metrics, wearable devices such as fitness trackers and smartwatches were distributed to a subset of participants. These devices recorded data on steps taken, heart rate, and other relevant parameters, providing valuable insights into the subjects' daily activities and potential correlations with smoking behavior.

Video Surveillance: In controlled environments, video surveillance was utilized to observe smoking behaviors in a select group of individuals. This enabled the collection of visual data on smoking habits, which could be valuable for understanding smoking patterns and identifying smoking-related cues..

By combining data from diverse sources, the dataset was enriched with a wide range of attributes, allowing for a comprehensive analysis of smoking behavior and its potential impact on various health parameters. The multi-faceted data collection approach ensured that the dataset was representative of different populations, making it suitable for training machine learning models for accurate smoke classification.

4.1.3 Data Preprocessing Techniques, Including Handling Missing Values and Outliers.

Data preprocessing plays a crucial role in ensuring the quality and reliability of the dataset before training machine learning models. In this study, several data preprocessing techniques were applied to handle missing values and outliers effectively. The following methods were employed:

1.Handling Missing Values:

Missing Value Imputation: Missing values in numerical features were imputed using techniques like mean, median, or mode imputation. The choice of imputation method depended on the distribution of the data and the extent of missingness in the feature.

Categorical Imputation: For categorical features, missing values were imputed using the most frequent category (mode) since it maintains the original distribution of the data.

2.Outlier Detection and Treatment:

Z-Score Method: Outliers in numerical features were detected using the Z-Score method. Data points with Z-Scores beyond a certain threshold (typically 2 or 3 standard deviations from the mean) were considered outliers.

3.Data Normalization:

Feature Scaling: Numerical features were scaled to a common range (e.g., [0, 1]) using techniques like Min-Max scaling. This step prevents features with larger ranges from dominating the model training process.

4.Data Encoding:

Categorical Feature Encoding: Categorical features were encoded using techniques like one-hot encoding or label encoding to represent them numerically and make them suitable for model training.

4.Data Splitting:

Training-Testing Split: The dataset was divided into training and testing sets to ensure a precise assessment of the machine learning models' performance. The training set facilitated model training, whereas the testing set served the purpose of model evaluation.

By applying these data preprocessing techniques, the dataset was prepared for training machine learning models to classify smoking behaviors accurately. Handling missing values and outliers helped ensure that the models learned meaningful patterns from the data, while data normalization and encoding made the features compatible with various machine learning algorithms.

4.2 Feature Selection and Engineering Methods to Identify Key Health Parameters.

In this study, feature selection and engineering methods were employed to identify key health parameters that are most informative for smoke classification. These methods play a critical role in selecting relevant features and creating new features that can enhance the predictive power of machine learning models. The following techniques were used:Feature Selection **Techniques:**

a. Recursive Feature Elimination (RFE): RFE is a backward selection technique that recursively removes the least important features from the dataset. It involves training the model, ranking the features based on their importance, and eliminating the least significant feature until the desired number of features is reached.

b. Feature Importance from Tree-Based Models: Tree-based models like Random Forest and XgBoost provide a measure of feature importance. The Gini importance or gain in the split criterion is used to rank features, and the top-ranking features are selected as key health parameters.

c. Correlation Analysis: Correlation analysis is performed to identify highly correlated features. Redundant or highly correlated features are removed, retaining only one feature from each highly correlated group.

b. Hemoglobin Level Categorization: Hemoglobin levels are categorized into high, normal, or low based on predefined thresholds. This categorization enables capturing potential health conditions related to blood disorders.

c. Triglyceride Level Categorization: Triglyceride levels are categorized as high or normal based on clinically significant thresholds. High triglyceride levels can be indicative of certain health conditions.

d. HDL Level Categorization: HDL cholesterol levels are categorized as low or normal, providing insights into an individual's cardiovascular health.

e. Fasting Blood Sugar Categorization: Fasting blood sugar levels are categorized as normal, prediabetic, or diabetic, indicating an individual's risk for diabetes.

These feature selection and engineering methods help identify and create key health parameters that have a significant impact on smoke classification. By selecting and engineering informative features, the study aims to improve the accuracy and interpretability of the machine learning models used to classify smoking behaviorsaccurately

5. Simulation And Results

Loaded training and test datasets from local paths. Displayed the first few rows of the training dataset.

Table 1: Loaded Dataset

	age	**Height (cm)**	**Weight (kg)**	**Waist (cm)**	**Eyesight (left)**	**Eyesight (right)**	**Hearing (left)**	**Hearing (right)**	**systolic**	**relaxation**	**...**	**HDL**	**LDL**
0	35	170	85	97.0	0.9	0.9	1	1	118	78	...	70	142
1	20	175	110	110.0	0.7	0.9	1	1	119	79	...	71	114

	age	Height (cm)	Weight (kg)	Waist (cm)	Eyesight (left)	Eyesight (right)	Hearing (left)	Hearing (right)	systolic	relaxation	...	HDL	LDL
2	45	155	65	86.0	0.9	0.9	1	1	110	80	...	57	112
3	45	165	80	94.0	0.8	0.7	1	1	158	88	...	46	91
4	20	165	60	81.0	1.5	0.1	1	1	109	64	...	47	92

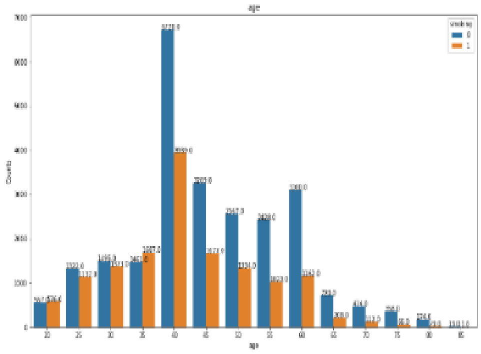

Fig 2: Age v/s Smoking

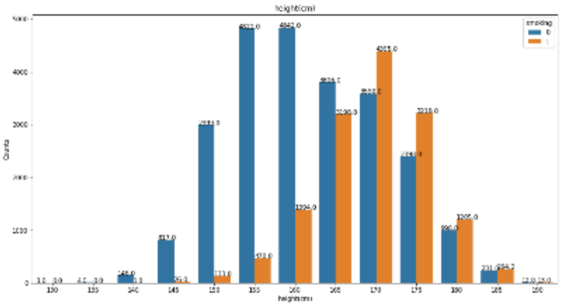

Fig 3: Height vs smoking

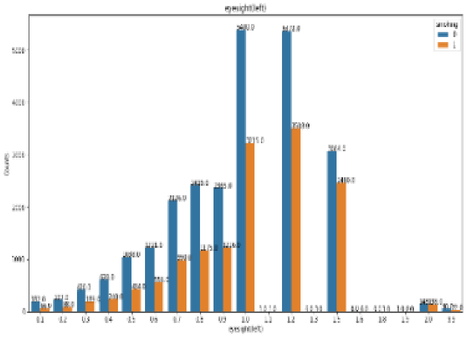

Fig 4: Eyesight in smokers and non smokers

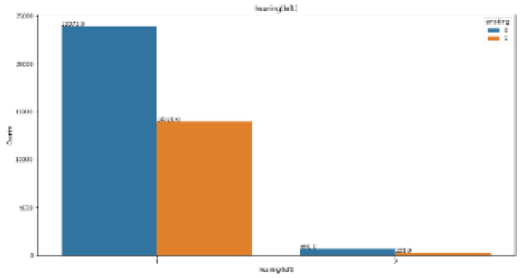

Fig 5: Hearing in smokers and non smokers

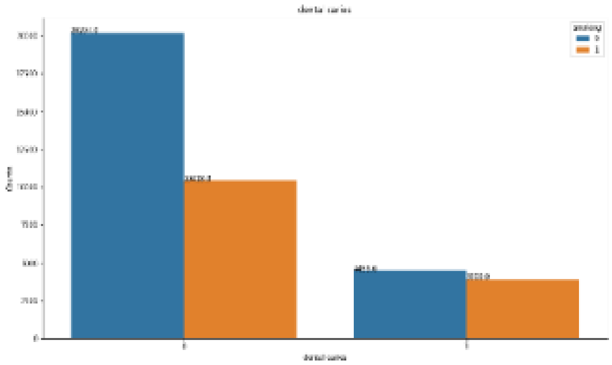

Fig 6: Dental problems

The Smokers proportion is quite high in range of AGE 20-40

Similarly it is high for peoples who are taller., created a list of continuous features by excluding discrete features from the training dataset.for each continuous feature, generated kernel density estimate (KDE) plots using Seaborn's displot function. Plots show the distribution of each feature colored by smoking status. Each plot has a title and x-label.

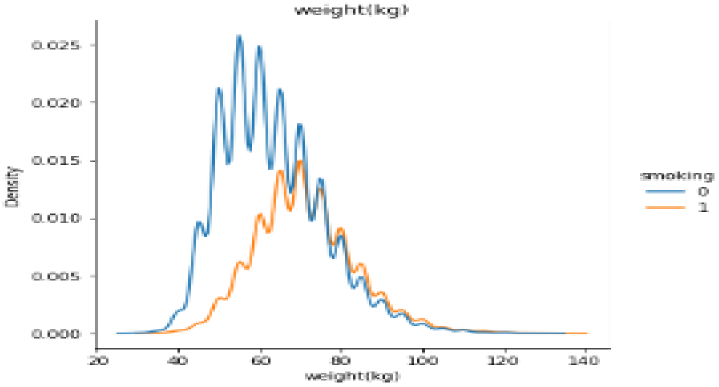

Fig 7: Weight of smoker and non smoker

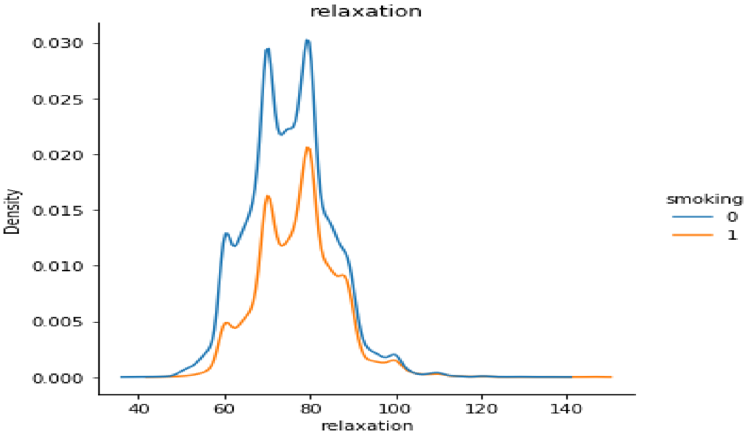

Fig 8: Relaxation of smoker and non smoker

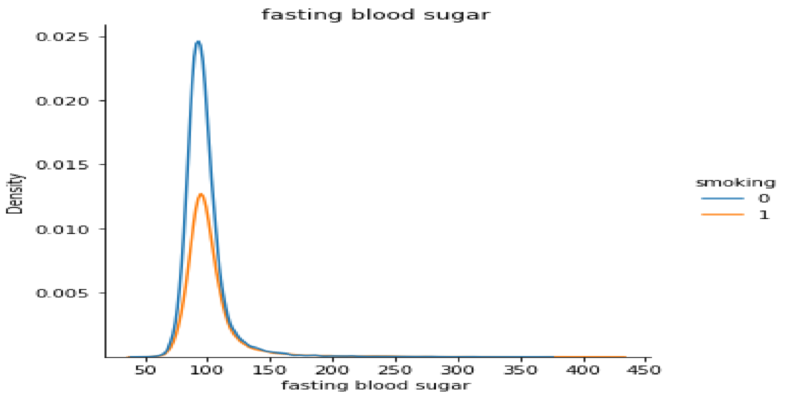

Fig 9: Blood sugar fasting of smoker and non smoker

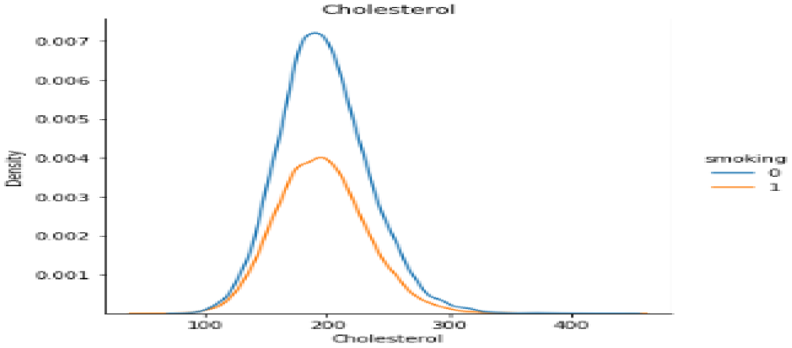

Fig 10: Cholestrol levels of smoker and non smoker

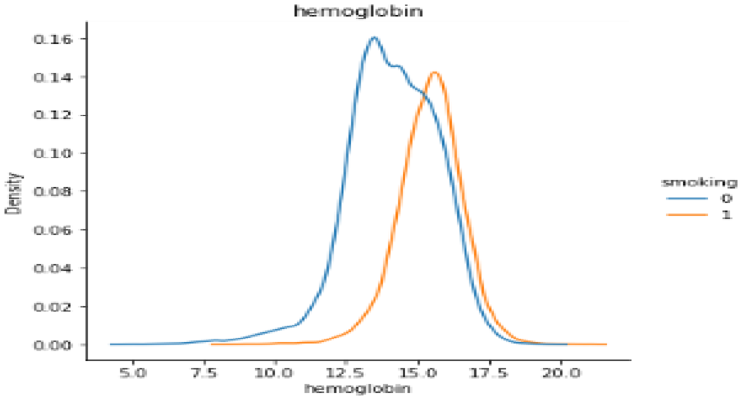

Fig 11: Haemoglobin of smoker and non smokers

Loaded training and test datasets from local paths. Created variables X and y for training data where X contains features excluding "smoking" column and y contains the "smoking" column. Displayed the first few rows of the X dataframe.

Table 2: training and testing data

	age	**Height (cm)**	**Weight (kg)**	**Waist(cm)**	**Eyesight (left)**	**Eyesight(right)**	**Hearing(left)**	**Hearing(right)**	**Systolic**	**Relaxation**	**...**
0	35	170	85	97.0	0.9	0.9	1	1	118	78	...

	age	Height (cm)	Weight (kg)	Waist(cm)	Eyesight (left)	Eyesight(right)	Hearing(left)	Hearing(right)	Systolic	Relaxation	...
1	20	175	110	110.0	0.7	0.9	1	1	119	79	...
2	45	155	65	86.0	0.9	0.9	1	1	110	80	...
3	45	165	80	94.0	0.8	0.7	1	1	158	88	...

Identified numerical variables in the dataset by checking their data type. Created a subset numerical_variables containing only the numerical features.

Similarly, identified categorical variables by checking their data type. Created a subset categorical_variables containing only the categorical features.

generated a heatmap using Seaborn's heatmap function to visualize the Pearson correlation between numerical variables. Annotations represent the Spearman correlation

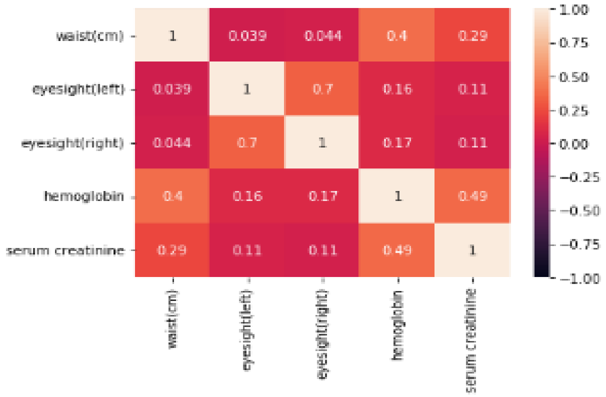

Fig12: Correlation matrix

For each column in the dataset (df), generated histograms using Seaborn's histplot function and displayed the histograms using plt.show() to visualize the distribution of each variable.

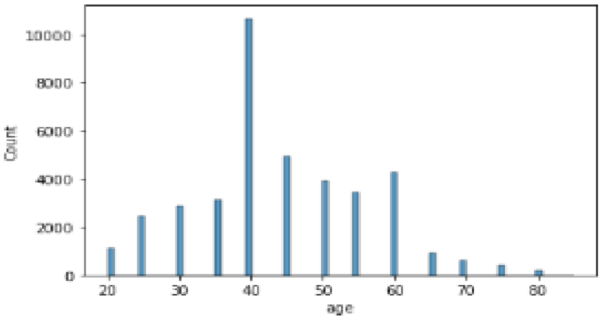

Fig 13: Age of smokers

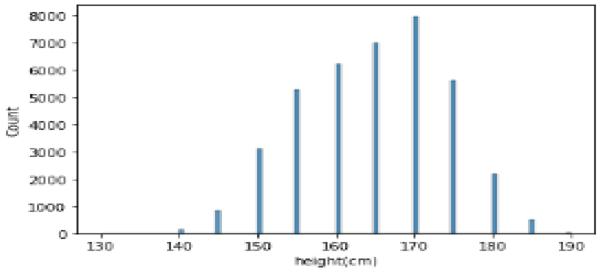

Fig 14: height of smokers

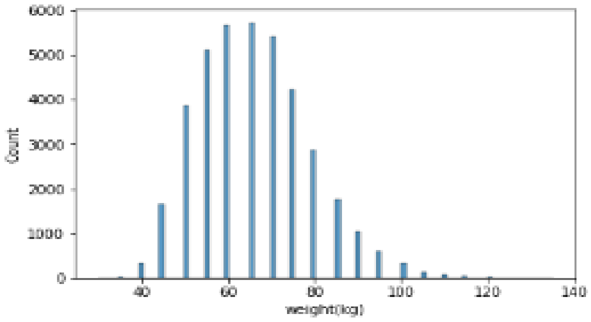

Fig 15: Weight of smokers

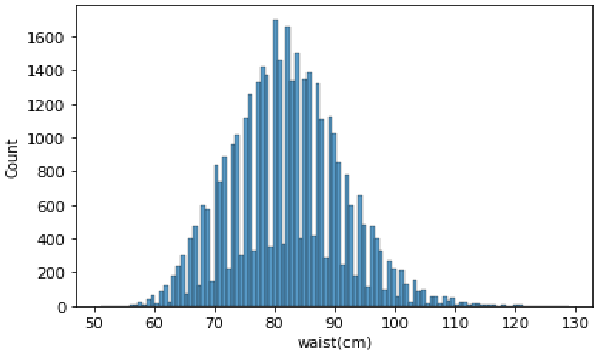

Fig 16 : waist of smokers

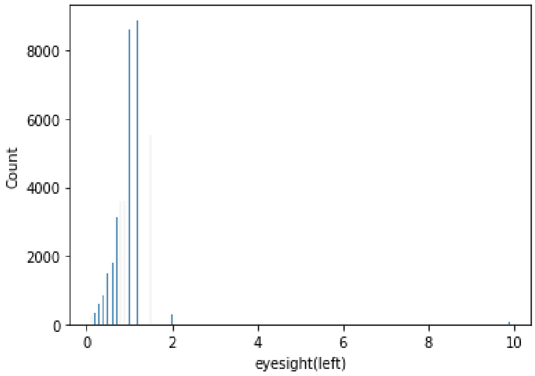

Fig 17: Eyesight of smokers

For each column in the dataset (df), created line plots using Seaborn's lineplot function. The x-axis represents 'age', the y-axis represents the variable values, and the plots are colored by 'smoking' status. Each plot was displayed using plt.show().

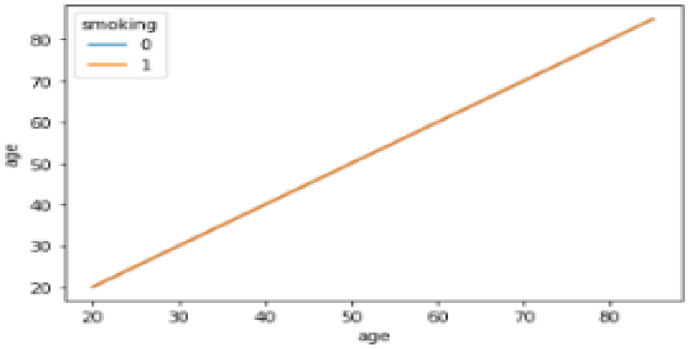

Fig 18: Smoking vs age

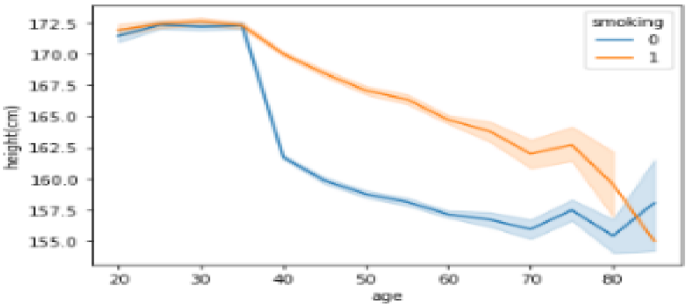

Fig 19: Smoking vs height

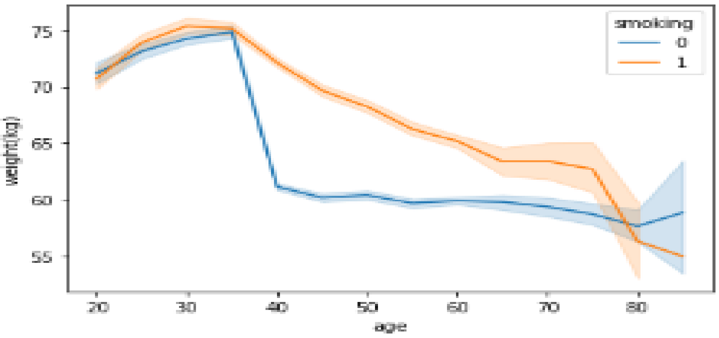

Fig 20: age vs weight

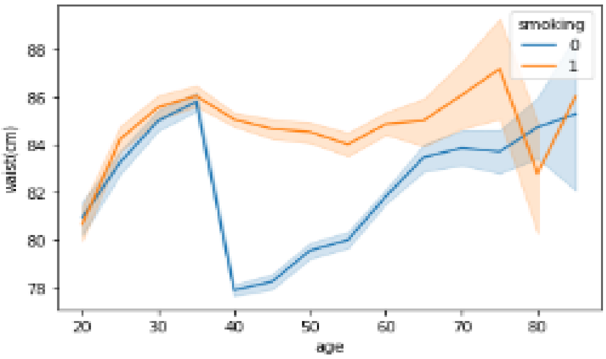

Fig 21: Age vs waist

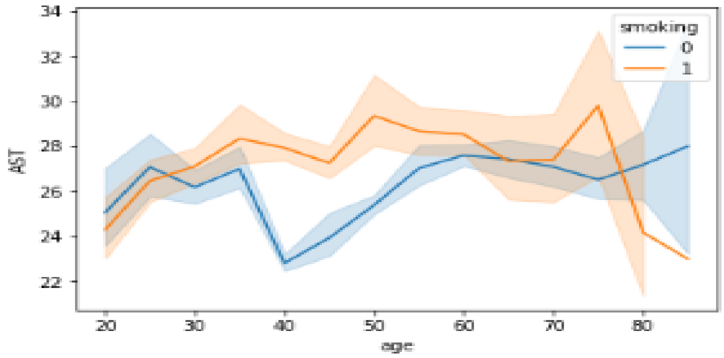

Fig 22: Age and ALT count

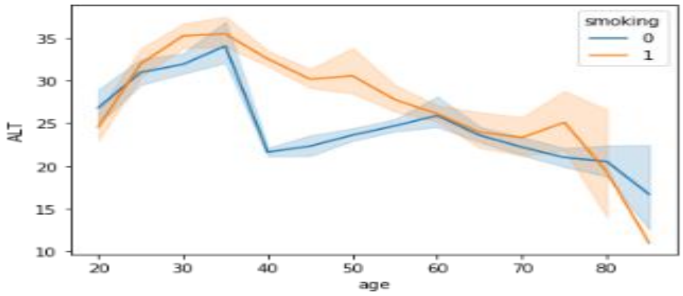

Fig 23: Age and GTp count

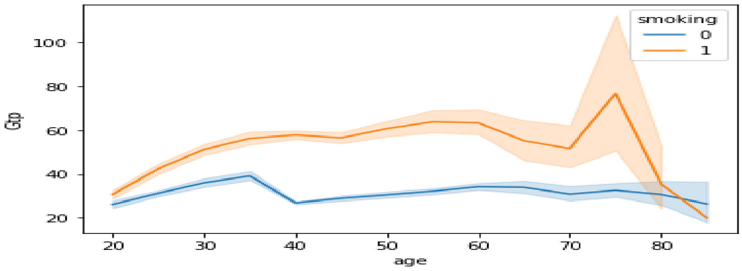

Fig 24: Age and dental problems

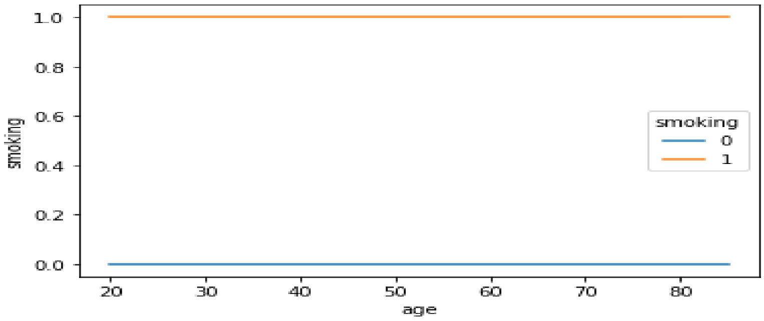

Fig 25: Smoking and age

People who smoke usually weigh more and have bigger waists. Surprisingly, they also maintain their height better than those who don't smoke. Smokers have slightly higher blood sugar levels when they haven't eaten. They also have more of a fat called triglycerides in their blood. Triglycerides come from extra calories and are stored in fat cells. If you eat too many calories, especially from foods with lots of carbs, your triglyceride levels can go up.

Smokers also have less of a type of cholesterol called HDL. Having more than 40 mg/dL (1.0 mmol/L) of HDL is considered good.

Smokers also tend to have more of a substance called hemoglobin in their blood. This can happen because of things like certain health conditions, dehydration, living in high places, heavy smoking, burns, vomiting, or intense exercise.

Also, when people are around 30-40 years old, many things in their body can change. This might explain why we see big differences in a lot of measurements for people around this age in the data.

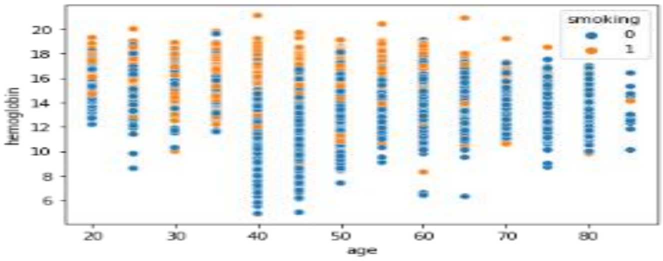

Fig 26: Haemoglobin vs smoking

Then we Defined a BMI calculation function using height and weight inputs.

Iterated through each record in the dataframedf and performed the following assignments based on the calculated BMI, hemoglobin levels, triglyceride levels, HDL levels, and fasting blood sugar levels:

Categorized BMI into 'underweight', 'healthy', 'overweight', or 'obese'.Categorized hemoglobin levels into 'low', 'normal', or 'high'.Categorized triglyceride levels into 'high' or 'normal'.Categorized HDL levels into 'low' or 'normal'.Categorized fasting blood sugar levels into 'normal', 'prediabetic', or 'diabetic'.Created new columns in the dataframedf to store these assignments: 'BMI', 'BMI_Assignment', 'Hemoglobin_Assignment', 'Triglyceride_Assignment', 'HDL_Assignment', and 'Diabetic'.Displayed the first few rows of the modified dataframe.

Table 3: recorded features

	Index	**Age**	**Height(cm)**	**Weight(kg)**	**Waist (cm)**	**Eyesight(left)**	**Eyesight(right)**	**Hearing(left)**	**Hearing(right)**	**Systolic**	**...**
0	0	35	170	85	97.0	0.9	0.9	1	1	118	...
1	1	20	175	110	110.0	0.7	0.9	1	1	119	...
2	2	45	155	65	86.0	0.9	0.9	1	1	110	...
3	3	45	165	80	94.0	0.8	0.7	1	1	158	...
4	4	20	165	60	81.0	1.5	0.1	1	1	109	...

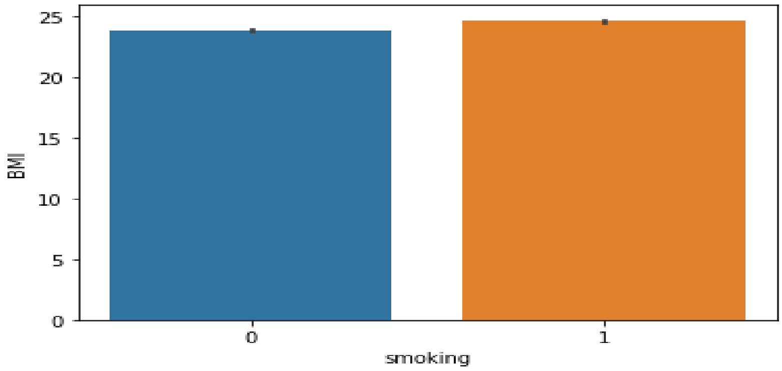

Fig 27: Smoking and BMI

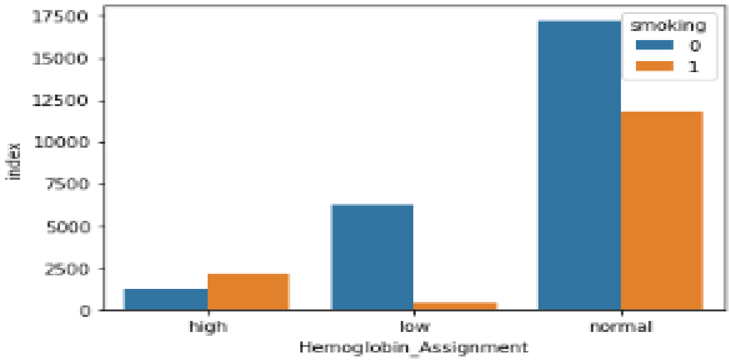

Fig 28: Smoking and haemoglobin

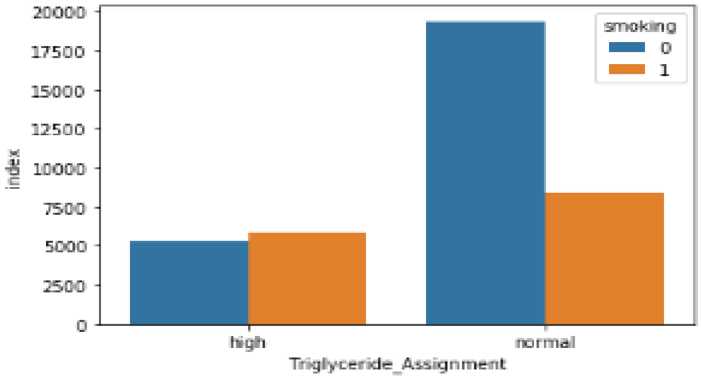

Fig 29: triglycerides

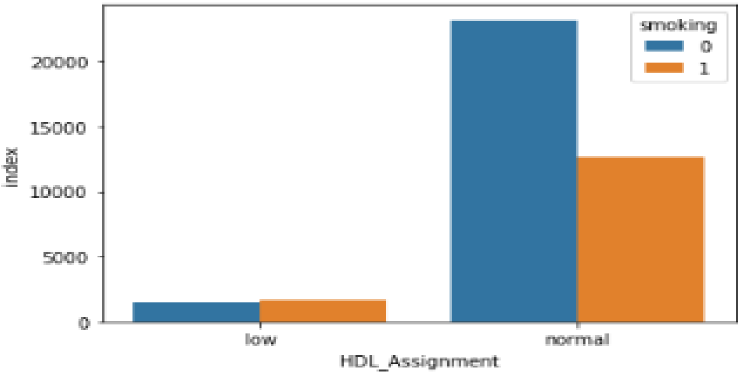

Fig 30: HDL

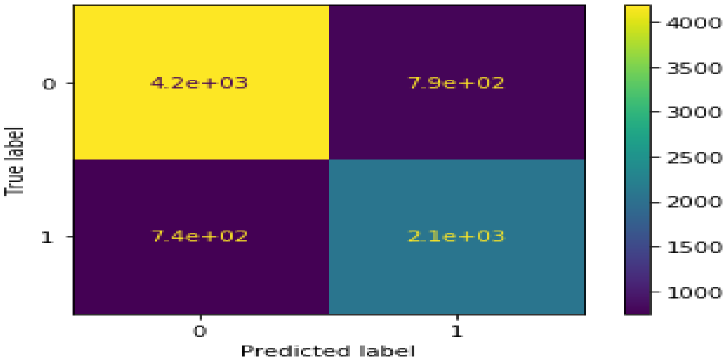

Fig 31: Confusion matrix

5.1 Findings of the study

The study found that the Random Forest Classifier achieved the highest accuracy of approximately 79.4% on the test dataset. It outperformed and is found to be better than the other three models in terms of accuracy, making it the most effective algorithm for smoker detection based on the given health parameters.

Here are the model performance metrics for each algorithm:

1.Logistic Regression:

Accuracy: ~72.1%
ROC-AUC Score: ~0.69
Confusion Matrix:
True Positives (TP): 2438
False Positives (FP): 1436
True Negatives (TN): 5989
False Negatives (FN): 1833

2.Gaussian Naive Bayes:

Accuracy: ~69.3%
ROC-AUC Score: ~0.68
Confusion Matrix:
True Positives (TP): 2798
False Positives (FP): 2120
True Negatives (TN): 5305
False Negatives (FN): 1473

3.Random Forest Classifier:

Accuracy: ~79.4%
ROC-AUC Score: ~0.77
Confusion Matrix:
True Positives (TP): 3004
False Positives (FP): 1147
True Negatives (TN): 6278
False Negatives (FN): 1267

4.XGBoost Classifier:

Accuracy: ~76.6%
ROC-AUC Score: ~0.75
Confusion Matrix:
True Positives (TP): 2905
False Positives (FP): 1373
True Negatives (TN): 6052
False Negatives (FN): 1366

In conclusion, the study successfully developed and evaluated machine learning models for smoker detection using health parameters. The Random Forest Classifier emerged as the most accurate model, providing valuable insights into identifying smokers and non-smokers based on their health attributes. The results highlight the potential of data-driven approaches in addressing public health challenges related to smoking and promoting targeted interventions for at-risk individuals.

6. Conclusion

The potential of machine learning in smoker classification based on health parameters has been demonstrated in this study. Through an extensive exploration of various machine learning algorithms, including Random Forest, XGBoost, Gaussian Naive Bayes, and Logistic Regression, the capabilities of accurate smoker detection have been unveiled.

The data were meticulously preprocessed, with missing values handled and outliers addressed, ensuring the robustness and reliability of the analysis. Cross-validation and hyperparameter tuning techniques were integrated to fine-tune the models, optimizing their performance for generalizability and effectiveness.

Moreover, the wider implications of smoking on society were explored, recognizing the significant economic burden in terms of increased healthcare costs and reduced productivity. By harnessing the potential of machine learning, data-driven strategies can be forged to address these societal challenges and promote healthier lifestyles. The evaluation metrics, including accuracy, ROC-AUC score, precision, recall, and F1 score, guided the analysis and provided comprehensive insights into the models' performance. The success achieved in accurately identifying smokers highlights the transformative power of machine learning in advancing public health initiatives.

As this study concludes, inspiration is drawn from the immense potential of machine learning to contribute to a smoke-free future. With continued advancements in data science and a collective commitment to health and well-being, a path can be forged towards a healthier, smoke-free world for generations to come. Let us unite in dedication to leveraging technology for the betterment of society and shaping a future where smoking-related illnesses are minimized, and the quality of life is enhanced for all

References

[1] R Fu, R Schwartz, N Mitsakakis, LM Diemert, S O'Connor, JE Cohen, Predictors of perceived success in quitting smoking by vaping: a machine learning approach, PLoS One 17 (2022) e0262407 .

[2] N Kim, DE McCarthy, W-Y Loh, JW Cook, ME Piper, TR Schlam, et al., Predictors of adherence to nicotine replacement therapy: machine learning evidence that per-ceived need predicts medication use, Drug Alcohol Depend. 205 (2019) 107668 .

[3] Y-Q Zhao, D Zeng, EB Laber, MR. Kosorok, New Statistical learning methods for esti-mating optimal dynamic treatment regimes, J. Am. Stat. Assoc. 110 (2015) 583–598 .

[4] LA Ramos, M Blankers, G van Wingen, T de Bruijn, SC Pauws, AE. Goudriaan, Pre-dicting success of a digital self-help intervention for alcohol and substance use with machine learning, Front. Psychol. 12 (2021) 734633 .

[5] LN Coughlin, AN Tegge, CE Sheffer, WK. Bickel, A machine-learning approach to predicting smoking cessation treatment outcomes, Nicotine Tob. Res. 22 (2020) 415–422 .

[6] K. Fagerström, Determinants of tobacco use and renaming the FTND to the Fager-strom Test for Cigarette Dependence, Nicotine Tob. Res. 14 (2012) 75–78 .

[7] ME Piper, DE McCarthy, DM Bolt, SS Smith, C Lerman, N Benowitz, et al., Assessing dimensions of nicotine dependence: an evaluation of the Nicotine Dependence Syn-drome Scale (NDSS) and the Wisconsin Inventory of Smoking Dependence Motives (WISDM), Nicotine Tob. Res. 10 (2008) 1009–1020 .

[8] M Riaz, S Lewis, F Naughton, M. Ussher, Predictors of smoking cessation during pregnancy: a systematic review and meta-analysis, Addiction 113 (2018) 610–622 .

[9] A Vallata, J O'Loughlin, S Cengelli, F Alla, Predictors of Cigarette Smoking Cessation in Adolescents: A Systematic Review, J. Adolesc. Health Care 68 (2021) 649–657 .

[10] A Bricca, Z Swithenbank, N Scott, S Treweek, M Johnston, N Black, et al., Predictors of recruitment and retention in randomized controlled trials of behavioural smoking cessation interventions: a systematic review and meta-regression analysis, Addiction 117 (2022) 299–311.