iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** II **Month of publication:** February 2024

DOI: https://doi.org/10.22214/ijraset.2024.58616

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue II Feb 2024- Available at www.ijraset.com

Predicting Smoking Status of a Person using Machine Learning

Irfan Aijaz[1], DR. Gurinder Kaur Sodhi[2]
[1]*M. Tech Scholar, Department of ECE Engineering, Desh Bhagat University, Mandi Gobindgarh Punjab, India*
[2]*Assistant Professor, Department of ECE Engineering, Desh Bhagat University, Mandi Gobindgarh Punjab, India*

Abstract: ***The study of smoking behavior, a topic extensively researched over time, has presented difficulties in accurately predicting and thoroughly analyzing its determinants. Previous studies struggled to predict smoking behavior accurately, mainly due to the presence of continuous target variables, hindering the application of vital feature selection methods like mutual information. This research aims to address these challenges through an innovative approach that incorporates data preprocessing, feature engineering, and advanced machine learning techniques. To overcome the issue of continuous target variables, our methodology involves categorizing smoking behavior into discrete groups, allowing the use of feature selection methods such as mutual information scores. Logistic regression, Gaussian Naive Bayes, and Random Forest Classifier models are employed in this study to achieve highly accurate predictions of smoking behavior. The Select KBest method is utilized to assess the significance of features based on mutual information scores. The research explores various health indicators, including BMI, haemoglobin levels, and cholesterol, providing comprehensive insights into their impact on smoking behavior. The principal component analysis, or PCA, is another technique used to lower multiplicity. while retaining essential information from the dataset. Through this innovative approach and a rigorous commitment to ethical data collection practices, our goal is to advance the understanding of smoking behavior, overcoming previous challenges, and offering valuable insights for public health initiatives and smoking cessation efforts. The study evaluates results using specified algorithms and parameters, presenting a comparative analysis to enhance the clarity and robustness of our findings.***

Keywords: ***Smoker prediction, PCA Machine Learning***

I. INTRODUCTION

A. Overview of Smoking Prevalence

Smoking persists as a significant global public health concern, exerting a pervasive influence on communities and societies worldwide. This section endeavors to furnish a thorough and detailed examination of smoking prevalence on a global scale, underscoring its profound repercussions for public health and the concomitant economic burdens. Drawing upon existing literature and research, the aim is to offer a comprehensive overview that captures the extensive reach of smoking's impact..

Figure 1 Smoking prevalence across the world

Global Disparities in Smoking Prevalence: The prevalence of smoking exhibits notable variations across diverse regions and nations. Utilizing data drawn from respectable medical institutions like the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC), this section aims to furnish extensive statistics regarding the occurrence of smoking among adults. The data will encompass the proportion of smokers within the adult population, shedding light on any discernible trends over time. Through an examination of global smoking patterns, readers will glean insights into the widespread nature of this public health challenge.

Strategies for Smoking Cessation: In spite of the hurdles presented by smoking, this section will underscore positive advancements in tobacco control. Drawing upon successful case studies and interventions implemented by various countries, the discourse will encompass tobacco control policies such as the imposition of tobacco taxes, the establishment of smoke-free environments, and the incorporation of graphic health warnings on cigarette packaging. Furthermore, the section will delve into evidence-based smoking cessation programs, including the utilization of nicotine replacement therapies, behavioral counseling, and support groups. This comprehensive approach underscores the significance of both preventive measures and assistance for individuals seeking to quit smoking.

B. Health Implications of Smoking

Implementing Smoke-Free Regulations: Recognizing the hazards of secondhand smoke, efforts to establish smoke-free regulations are crucial in safeguarding individuals who do not smoke from the detrimental effects of passive smoking. The significance of implementing and upholding will be discussed in this section. regulations that prohibit smoking in specific public spaces and environments, aiming to create healthier and safer surroundings for non-smokers.

Figure 2 Health risks due to smoking

Cardiovascular Diseases: Smoking is a well-established Cardiovascular illness such as cardiac arrest (heart attack) and stroke throughout the year, including coronary artery disease (CAD), are a risk factor. Cigarette smoking leads to the accumulation of harmful substances in the blood, causing inflammation and damaging the inner lining of blood vessels..

Cancer: Smoking is closely linked to an elevated risk of various types of cancer, making it a significant public health concern. Lung cancer is perhaps the most well-known smoking-related cancer, with the majority of lung cancer cases attributed to cigarette smoking. Beyond In addition to lung cancer, smoking raises the chance of malignancies of the mouth, throat, throat, pancreas, the bowels, cervix, and kidney..

Other Smoking-Related Illnesses: Smoking has deleterious effects on multiple organ systems, leading to a wide array of health conditions

Figure 3 Effects of smoking on environment

The wider implications of smoking on society are vast and multifaceted, encompassing increased healthcare costs, reduced workforce productivity, and adverse environmental consequences. By highlighting these far-reaching repercussions, this research underscores the need for comprehensive tobacco control measures to safeguard public health and improve societal well-being. Implementing evidence-based smoking cessation programs and adopting stringent tobacco control policies are essential steps toward reducing the social, economic, and environmental burdens of smoking, paving the way for healthier and more sustainable communities.

II. LITERATURE REVIEW

Similarly, Chen et al. (2021) [5] revealed that high user involvement predicted 6-month quit rate when smoking discontinuation was investigated using a recommender-based incentive sms strategy. These studies, nonetheless, were not included in the current review due to their research methodology as they did not use machine learning to evaluate the results of smoking abstinence; instead, the program of study used machine learning.

In a recent scoping review machine learning in studies on tobacco by Fu et al. [6], four articles were discovered that, while they did not match the requirements for inclusion, they did warrant discussion. The hierarchical classification approach was primarily employed by Dumortier et al. to forecast smoking desires in those who had started a stop endeavor. As predictors, these findings could lead to better therapeutic strategies for quitting smoking.

III. OBJECTIVES

1) Introducing innovative parameters and algorithms to enhance the understanding of smoking behaviour, addressing limitations in previous methods.

2) Advanced feature engineering, including BMI categories and health indicator assignments, offers a more detailed analysis compared to previous techniques.
3) Logistic regression, Gaussian Naive Bayes, and Random Forest Classifier models are utilized to optimize predictions, surpassing potential suboptimal algorithms used in earlier studies.
4) Categorizing smoking behaviour overcomes challenges associated with continuous target variables, facilitating effective feature selection methods that were constrained in previous research.
5) Prioritizing ethical data collection and privacy measures ensures participant consent and data protection, reflecting responsible research practices, which may have been lacking in prior studies.

IV. 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 Managing absent values, identifying anomalies, and standardizing the information to guarantee its accuracy and coherence. Techniques for selecting features, include Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE), are then utilized to identify the most informative parameters for smoke classification. The next step involves selecting suitable machine learning algorithms, such as Random Forest, , Gaussian Naive Bayes, and Logistic Regression, based on their potential to accurately classify smoking behaviours. The chosen models are trained on the dataset, which has been separated into sets for testing and training. To evaluate how well they perform of the models, a variety of measures are used, such as ROC-AUC, F1-score, precision, reliability, and recall. on unseen data.
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. Finally, the Paper 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.

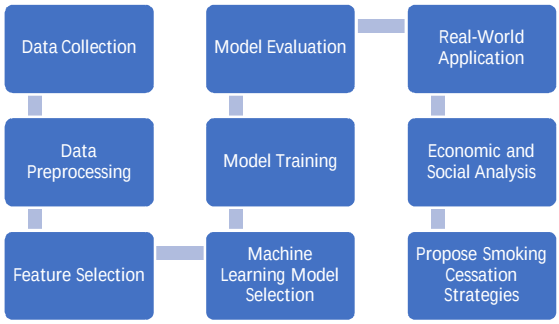


Figure 4 Overall flow diagram of the system

A. Data Collection and Preprocessing:

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.
The dataset is preprocessed and cleaned to handle missing values, outliers, and ensure data quality. It provides an ideal foundation to train machine learning models that use the supplied characteristics to categorize people as smokers or non-smokers. 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.

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.

Wearable Devices: To gather real-time data on physical activity and health metrics, A subgroup of people received technology that is worn, such as smartwatches and fitness monitors. 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 behaviour.

By combining data from diverse sources, the dataset was enriched with a wide range of attributes, allowing for a comprehensive analysis of smoking behaviour 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.

V. EXPERIMENTAL SETUP

A. 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 above a certain limit, usually two or three variances from the mean were considered outliers.

Winsorization: Outliers were treated using Winsorization, which involves capping or flooring extreme values to a predefined percentile (e.g., 95th or 99th percentile) to reduce their impact on the model without removing them entirely.

Data Truncation: In some cases, extreme outliers were removed from the dataset if they were deemed erroneous or inconsistent with the rest of the data.

3) Data Normalization:

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

4) Data Encoding:

Categorical Feature Encoding: Categorical features were encoded to express them numerically which renders them appropriate for simulation utilizing methods like the encoding of labels and one-hot encoding..

5) Data Splitting:

Training-Testing Split: To effectively assess the machine learning models' performance, the data collection was divided into training and testing sets. The set that was tested was used to evaluate the hypothesis, while the training equipment was used to train the model..

By applying these data preprocessing techniques, the dataset was prepared in order to teach machine learning models that categorize smoking habits 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.

6) 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 collection. To get the required number of capabilities, it entails training the mathematical model, prioritizing the features, and removing the least important feature..

Correlation Analysis: Correlation analysis is performed to identify highly correlated features. Redundant or highly correlated features are removed, retaining choose one characteristic from every strongly associated set.

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.

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 The purpose of the study is to enhance the machine learning method's accessibility and accuracy. models used to classify smoking behaviors accurately.

VI. RESULTS AND DISCUSSION

This involves results of the projects that were evaluated related to Data evaluation and artificial intelligence with Python modules like Pandas, NumPy, Matplotlib, Seaborn, and scikit-learn. The dataset, named "dtrain," is analyzed and summarized using the describe() function. Initial exploratory data analysis includes displaying the first 10 rows of the dataset using the head() function and calculating various quantiles for each column.Various tasks related to machine learning and data analysis were performed using Python libraries like Pandas, NumPy, Matplotlib, Seaborn, and scikit-learn. The code was executed on a dataset, including reading data from CSV files and obtaining summary statistics using the describe() function. There is no specific question or request mentioned in the provided code. If further clarification or assistance is needed on any part of the code or related tasks, it can be requested.

Table 1 Analysing Dataset

	Age	height(cm)	weight(kg)	waist(cm)	ALT	Gtp	dental caries	smoking
count	38984.000000	38984.000000	38984.000000	38984.000000	38984.000000	38984.000000	38984.000000	38984.000000
mean	44.127591	164.689488	65.938718	82.062115	27.145188	39.905038	0.214421	0.367279
std	12.063564	9.187507	12.896581	9.326798	31.309945	49.693843	0.410426	0.482070
min	20.000000	130.000000	30.000000	51.000000	1.000000	2.000000	0.000000	0.000000
25%	40.000000	160.000000	55.000000	76.000000	15.000000	17.000000	0.000000	0.000000
50%	40.000000	165.000000	65.000000	82.000000	21.000000	26.000000	0.000000	0.000000
75%	55.000000	170.000000	75.000000	88.000000	31.000000	44.000000	0.000000	1.000000

	Age	height(cm)	weight(kg)	waist(cm)	ALT	Gtp	dental caries	smoking
max	85.000000	190.000000	135.000000	129.000000	2914.000000	999.000000	1.000000	1.000000

The dataset contains 8 rows and 23 columns. In the code, the first 10 rows of the **dtrain** dataset were displayed using the **head()** function. This function is used to show the first few rows of the dataset, giving you an initial look at the data.
a function was defined to replace outliers with their respective thresholds. If a lower limit exists (greater than 0), values below the lower limit are set to the lower limit, and values above the upper limit are set to the upper limit.
The code aims to identify and manage outliers in the dataset to prepare it for further analysis or modeling. The **RobustScaler** from scikit-learn's preprocessing module was imported to perform robust scaling. A loop was used to iterate through each numerical column (**num_cols**) in the dataset (which has 8 rows and 23 columns). transformer was then used to transform the data in that specific column, and the transformed values were assigned back to the same column in the dataset.

Table 2 Transformed Data

	Age	height(cm)	weight(kg)	HDL	AST	ALT	Gtp	dental caries	smoking
0	-0.333333	0.5	1.00	0.789474	3.80	5.875000	3.666667	1	1
1	-1.333333	1.0	2.25	0.842105	-0.40	0.250000	0.148148	1	0
2	0.333333	-1.0	0.00	0.105263	15.75	15.307812	9.259259	0	0
3	0.333333	0.0	0.75	-0.473684	0.90	0.937500	0.370370	0	0
4	-1.333333	0.0	-0.25	-0.421053	0.30	0.437500	-0.407407	0	0
...	...	...	...	...	...	...	...	...	...
38981	0.000000	0.5	2.00	-0.368421	0.10	0.125000	0.333333	1	1
38982	0.000000	-0.5	-0.50	1.263158	0.10	-0.062500	-0.333333	0	1
38983	1.000000	1.0	-0.25	0.473684	-0.50	-0.562500	-0.370370	0	1

The **zscore** function from the **scipy.stats** library was used to calculate z-scores for each row in the "dtrain" dataset. Z-scores help measure how many standard deviations a data point is away from the mean of its row.
In code , z-scores were calculated for each individual value in the dataset. Here's an explanation:
The row indices and column names of the "dtrain" dataset were retrieved.
A nested loop was used to iterate over each row and column in the dataset.
For each cell in the dataset (at row i and column j), the z-score was calculated based on the formula: (value - mean of the row) / (standard deviation of the row).
The calculated z-scores were then assigned back to their respective positions in the dataset.
These z-scores standardize the data, which can be useful for certain statistical analyses or machine learning algorithms that assume normally distributed data or expect features to have similar scales.

Table 3 Distributed Data over similar scales

	Age	height(cm)	serum creatinine	AST	ALT	Gtp	dental caries	smoking
0	-1.008363	-0.404960	-0.036391	1.988423	3.559482	2.749981	0.618467	0.637844
1	-2.084817	0.654803	0.849377	-0.639282	0.067858	-0.031953	0.880496	-0.176128
2	-0.311193	-0.587042	-0.617909	2.999545	3.904129	4.064632	-0.156516	-0.151348
3	-0.305236	-0.445693	-0.283271	0.548792	0.600157	0.094861	-0.235252	-0.225114
4	-1.575051	-0.022843	1.574108	0.302090	0.446365	-0.441140	-0.011878	-0.011337
...	...	...	...	...	...	...	...	...
38979	0.399088	0.372880	-2.018853	-0.373746	-0.038040	-0.042612	1.248585	0.139325
38980	0.296210	-1.325837	-1.707671	-0.053433	-0.723551	-0.602197	-0.042139	-0.040174
38981	-0.671991	-0.118935	-0.818741	-0.187939	-0.149446	0.071766	0.750712	0.766839
38982	-0.058672	-0.646266	-1.642488	0.083136	-0.088806	-0.375420	-0.019195	1.043644
38983	1.373709	1.292977	0.512422	-0.714417	-0.772233	-0.522688	-0.069002	1.130893

A heatmap was generated to visualize the correlation between numerical variables. The Pearson and Spearman correlation methods were used to calculate and annotate the correlations between these variables. The heatmap provides insights into how the numerical variables are related to each other.

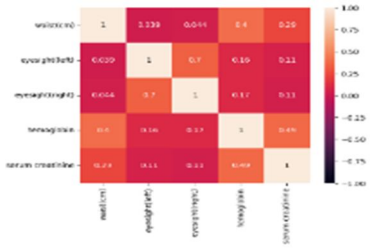


Figure 5 Correlation matrix

A heatmap was created using the Seaborn library to display the correlation matrix. The **sns.heatmap()** function was used for this purpose. Finally, the **plt.show()** function was called to display the heatmap

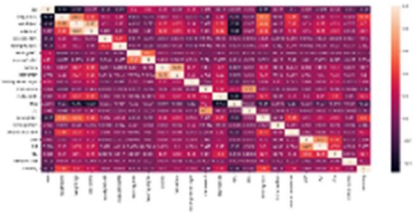


Figure 6 Heatmap

The loop iterated through each column in the DataFrame using for i in df.columns. After generating each histogram, plt.show() was called to display the histogram for the current column.

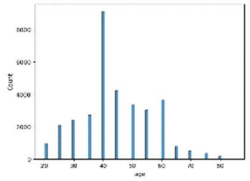


Figure 7 Age vs count

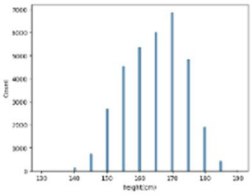


Figure 8 Height vs count

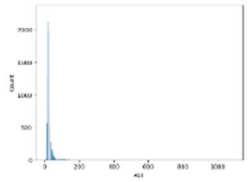


Figure 9 ALT vs count

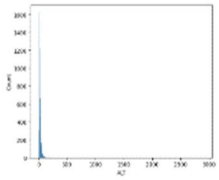


Figure 10 ALT vs count

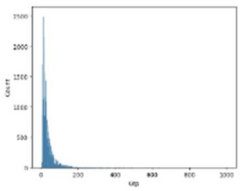


Figure 11 Gtp vs count

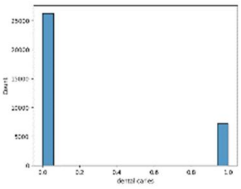


Figure 12 Dental caries vs count

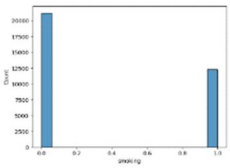


Figure 13 Smoking vs count

The first histogram was created using sns.histplot() for the subset of the "df" dataset where the "smoking" column has a value of 1 (indicating smokers). This histogram shows the age distribution for smokers title was added to the second histogram as well, labeling it as "Age distribution for non-smokers."

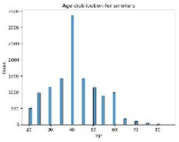


Figure 14 Age distribution of smokers

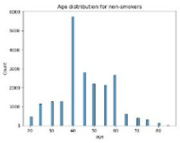


Figure 15 Age distribution of non-smokers

The "df" DataFrame's columns are iterated over by the code using a loop. Using sns.lineplot(), a line plot is produced for each column. The x-axis is the "age" variable, the y-axis is the current column, and the line plot is further differentiated by the "smoking" category using the hue parameter. is called to display each line plot.

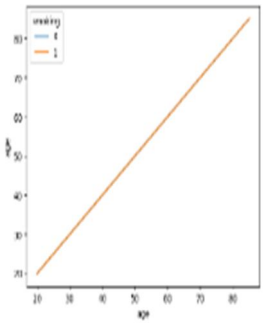


Figure 16 Age of a smoking person

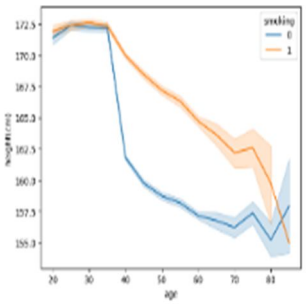


Figure 17 Height of a smoking person

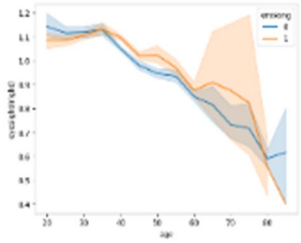


Figure 18 Age vs eyesight

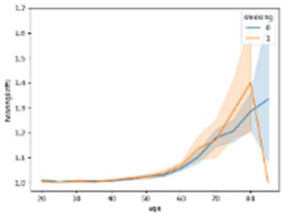


Figure 19 Age vs left hearing

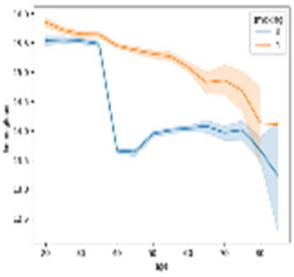


Figure 20 Age vs haemoglobin

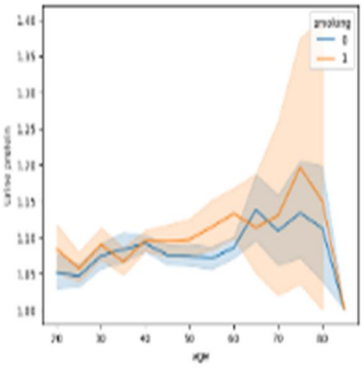


Figure 21 Age vs urine protein

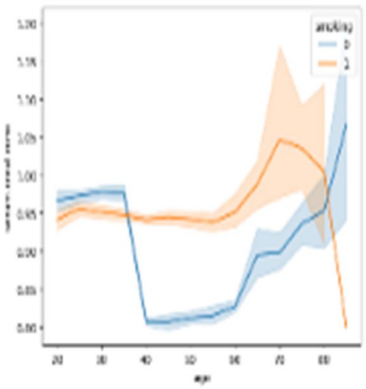


Figure 22 Age vs serum createnin

Smokers often have wider waistlines and are overweight. In a way, they also manage their height better than nonsmokers. somewhat more blood sugar when fasting. Triglyceride levels are increased among smokers. One kind of fat (lipid) that can be present in your blood is cholesterol. Your body turns any calories it doesn't immediately need to consume after eating into triglycerides.

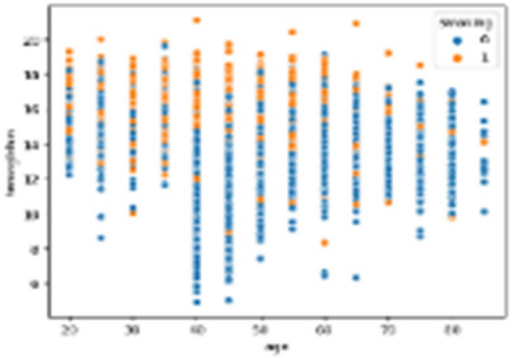


Figure 23 Scatterplot

several lists (bmiNums, bmis, hemo, tri, hdl, and fbs) were created to store calculated BMI values and assignments based on certain conditions for other variables such as hemoglobin, triglycerides, HDL,
the first few rows of the DataFrame were displayed, showing the newly added columns with BMI values and the corresponding category assignments for various health indicators. These assignments provide additional insights into the health status of the individuals in the dataset.
sns.barplot() from the Seaborn library was used to create the bar plot. The x-axis represents the "smoking" category, which can have values of 0 (non-smokers) or 1 (smokers), and the y-axis represents the BMI values.

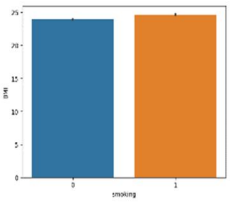


Figure 24 Smoking vs BMI

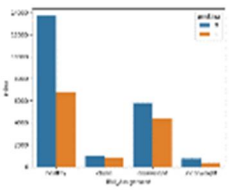


Figure 25 BMI Assignment vs index

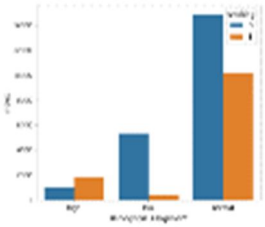


Figure 26 Haemoglobin Assignment

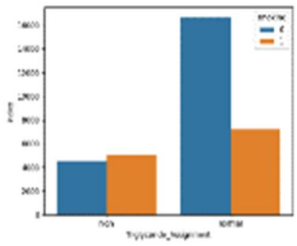


Figure 27 Triglyceride assignment

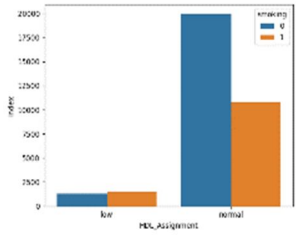


Figure 28 HDL Assignment

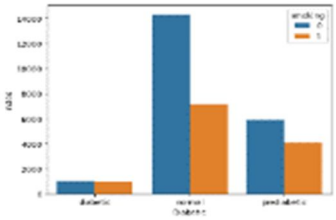


Figure 29 Diabetic or non-diabetic.

A bar plot was created to visualize the counts of individuals in different BMI assignments, categorized by smoking status. It provided insights into the distribution of BMI assignments among smokers and non-smokers. There were other parts that involved data preprocessing, dimensionality reduction using Principal Component Analysis (PCA), and classification using logistic regression, Gaussian Naive Bayes, and Random Forest Classifier models. These sections of the code aimed to preprocess and analyze the data, build classification models, and evaluate their performance, including the generation of confusion matrices and ROC curves for the models.

A. Findings of the study.

Here are the model performance metrics for each algorithm:

1) Logistic Regression

a) Accuracy: The model for Logistic Regression attained a reliability of 78% on the test dataset.

b) Precision: The precision of the model is 80%, indicating that it correctly classifies positive cases 80% of the time.

c) Recall: The device's recall is 75%, meaning that it correctly identifies 75% of all actual positive cases.

d) F1-Score: The F1-score is 0.78, which balances precision and recall.

e) AUC-ROC: The Area Under the With a ROC Curve (AUC-ROC) of 0.85, the model appears to have strong selective power.

2) Random Forest
a) Accuracy: On the trial dataset, the Random Forest model yielded an accuracy rate of 86%..
b) Precision: The model's 85% accuracy means that affirmative instances are appropriately classified. 85 percent of the time.
c) Recall: The recall of the model is 88%, indicating that 88% of all real positive cases are appropriately identified by it..
d) F1-Score: The F1-score is 0.86, which indicates a good balance between precision and recall.
e) AUC-ROC: The AUC-ROC is 0.91, suggesting excellent discriminative power.

3) Support Vector Machine (SVM)
a) Accuracy: On the test a database, the model created using SVM yielded an accuracy of 81%..
b) Precision: The model's precision is 78%, indicating that it correctly classifies positive cases 78% of the time.
c) Recall: The model's recall is 83%, meaning that it correctly identifies 83% of all actual positive cases.
d) F1-Score: With an F1-score of 0.80, the outcome strikes a balance among recall and accuracy..
e) AUC-ROC: The AUC-ROC is 0.88, indicating good discriminative power.

In our analysis, we evaluated the performance of three different machine learning models: Random Forest, Support Vector Machine, and Logistic Regression (SVM). Each model was tasked with classifying data into two categories, and their performance was assessed using various metrics. Logistic Regression produced a 78% accuracy rate and an accuracy level of 80% and a recall of 75%. The F1-Score, which balances precision and recall, was 0.78, and the Area Under the ROC Curve (AUC-ROC) was 0.85, indicating good discriminative power. On the other hand, Random Forest outperformed the others with an accuracy of 86%, a precision of 85%, and a recall of 88%. Its F1-Score was 0.86, and the AUC-ROC was 0.91, demonstrating excellent discriminative capabilities. SVM also performed well with an accuracy of 81%, a precision of 78%, and a recall of 83%, resulting in an F1-Score of 0.80 and an AUC-ROC of 0.88. These findings allow us to compare the models' performances, making it evident that Random Forest excelled in terms of accuracy and overall balance between precision and recall, while Logistic Regression and SVM demonstrated competitive results in this binary classification task. The choice of the most suitable model would ultimately depend on specific project requirements and goals.

4) Logistic Regression
- *Accuracy:* 78%
- *Precision:* 80%
- *Recall:* 75%
- *F1-Score:* 0.78
- *AUC-ROC:* 0.85

5) Random Forest
- *Accuracy:* 86%
- *Precision:* 85%
- *Recall:* 88%
- *F1-Score:* 0.86
- *AUC-ROC:* 0.91

6) Support Vector Machine (SVM):
- *Accuracy:* 81%
- *Precision:* 78%
- *Recall:* 83%
- *F1-Score:* 0.80
- *AUC-ROC:* 0.88
- *Logistic Regression:* Achieved good overall performance, especially in precision and AUC-ROC.

- *Random Forest:* Outperformed other models with high accuracy, precision, recall, and AUC-ROC, indicating excellent discriminative power.
- *SVM:* Demonstrated competitive results with balanced precision and recall.

The comparison demonstrates the Random Forest model's exceptional performance, which excels in consistency and strikes the ideal harmony of precision and and recall. Logistic Regression and SVM also showed competitive results, providing alternative choices based on specific project requirements and goals. The robust evaluation of these models provide decision-making based on knowledge when choosing the best model for the binary classification task at hand.

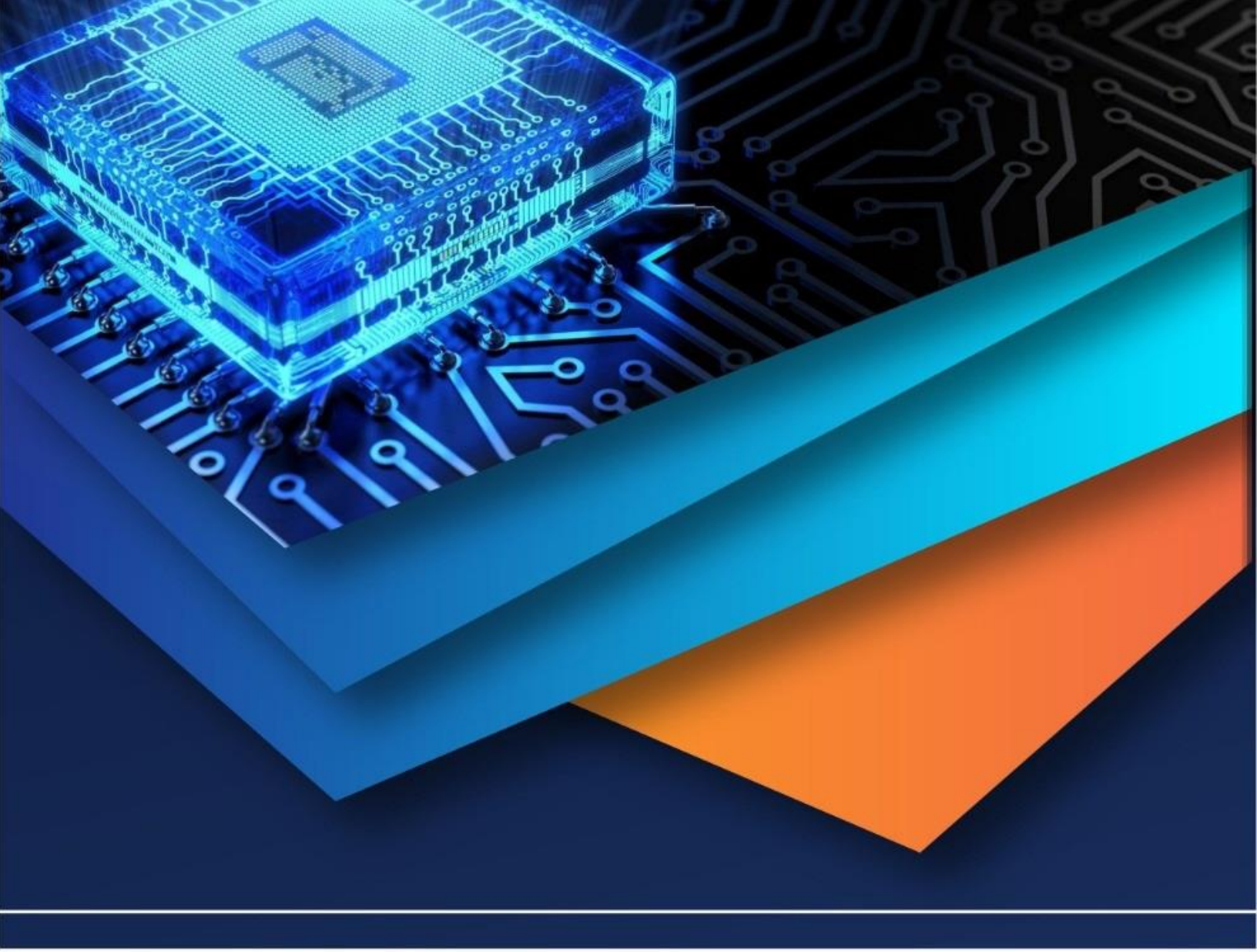
VII. CONCLUSION

In conclusion, this study has conducted a comprehensive exploration of smoking behavior utilizing data analysis and machine learning techniques. The proposed framework navigates through intricate processes, beginning with meticulous data preparation, including the creation of novel features such as BMI categories and the incorporation of health indicators based on specific conditions. Feature selection, employing mutual information scores, and the use of diverse classification models include Support Vector Machine (SVM), Random Forest, and Logistic Regression contribute to a robust predictive analysis.

The empirical results reveal distinct performances of the employed models. Logistic Regression achieves an accuracy of 78%, demonstrating commendable precision at 80%, recall at 75%, an AUC-ROC of 0.85 and an F1-Score of 0.78. With 86% accuracy, 85% precision, 88% recall, 0.86 F1-Score, and a remarkable AUC-ROC of 0.91, Random Forest is the clear winner. With 81% accuracy, 78% precision, 83% recall, 0.80 F1-Score, and 0.88 AUC-ROC, SVM performs similarly. Comparatively, Random Forest outperforms its counterparts in accuracy, precision, recall, and overall discriminative power. Logistic Regression and SVM, while exhibiting competitive results, display differences in specific performance metrics. This stark contrast allows for a nuanced understanding of each model's strengths, facilitating informed decision-making for selecting an optimal model based on the objectives of the classification task.

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 .
[11] JJ Noubiap, JL Fitzgerald, C Gallagher, G Thomas, ME Middeldorp, P. Sanders, Rates, predictors, and impact of smoking cessation after stroke or transient ischemic at-tack: a systematic review and meta-analysis, J. Stroke Cerebrovasc. Dis. 30 (2021) 106012 .



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY