iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** III **Month of publication:** March 2024

DOI: https://doi.org/10.22214/ijraset.2024.59453

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 III Mar 2024- Available at www.ijraset.com

Utilizing Machine Learning to Forecast Smoking Behavior

Sharuf Hamid Lone[1], Dr. Manuraj Moudgil[2]
[1]*M. Tech Scholar, Department of CSE Engineering, BGIET, Punjab, India*
[2]*Professor, Department of CSE Engineering, BGIET, Punjab, India*

Abstract: ***The examination of smoking habits, extensively studied over time, has posed challenges in accurately predicting and thoroughly analyzing its determinants. Previous research efforts struggled to precisely predict smoking behavior due to the presence of continuous target variables, which hindered the application of crucial feature selection techniques like mutual information. This study aims to tackle these hurdles through an innovative approach that integrates data preprocessing, feature engineering, and advanced machine learning methods. To address the issue of continuous target variables, our methodology involves categorizing smoking behavior into discrete groups, enabling the utilization of feature selection techniques such as mutual information scores. Logistic regression, Gaussian Naive Bayes, and Random Forest Classifier models are utilized to achieve highly accurate predictions of smoking behavior. The Select KBest method is employed to evaluate the importance of features based on mutual information scores. The investigation delves into various health markers, including BMI, haemoglobin levels, and cholesterol, offering comprehensive insights into their influence on smoking habits. Furthermore, Principal Component Analysis (PCA) is implemented to reduce dimensionality while preserving essential information from the dataset. Through this novel approach and a steadfast commitment to ethical data collection practices, our objective is to advance the comprehension of smoking behavior, surmounting past challenges, and providing valuable insights for public health initiatives and smoking cessation endeavors. The paper assesses outcomes using specified algorithms and parameters, presenting a comparative analysis to enhance the clarity and reliability of our findings.***
Keywords: Smoker prediction, PCA Machine Learning

I. INTRODUCTION

A. Overview of Smoking Prevalence

Smoking remains a critical issue in global public health, exerting a widespread influence on communities and societies globally. This segment aims to provide a comprehensive and detailed analysis of smoking prevalence worldwide, highlighting its profound implications for public health and the associated economic burdens. By drawing on existing literature and research, our goal is to present a thorough overview that reflects the extensive scope of smoking's impact..

Figure 1 Smoking prevalence across the world

Disparities in Global Smoking Prevalence: The prevalence of smoking demonstrates significant disparities across different regions and countries. This section aims to provide extensive statistics on adult smoking rates, utilizing data from reputable health organizations such as the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC). The data will include the percentage of adult smokers within populations, highlighting any observable trends over time. By analyzing global smoking patterns, readers will gain insights into the widespread nature of this public health issue.

Strategies for Smoking Cessation: Despite the challenges posed by smoking, this section will emphasize positive progress in tobacco control efforts. Drawing on successful case studies and interventions implemented by various nations, the discussion will cover tobacco control policies such as the implementation of tobacco taxes, the creation of smoke-free environments, and the inclusion of graphic health warnings on cigarette packaging. Additionally, the section will explore evidence-based smoking cessation programs, including the use of nicotine replacement therapies, behavioral counselling, and support groups. This comprehensive approach underscores the importance of both preventive measures and assistance for individuals seeking to quit smoking.

B. Health Implications of Smoking

1) Enforcing Smoke-Free Regulations: Acknowledging the dangers of secondhand smoke, it is imperative to implement regulations that protect non-smokers from its harmful effects. This section will delve into the significance of establishing and upholding regulations that ban smoking in designated public spaces and environments. The objective is to foster healthier and safer surroundings for individuals who do not smoke, safeguarding them from the detrimental impacts of passive smoking..

Figure 2 Health risks due to smoking

2) Cardiovascular Diseases: Smoking significantly contributes to cardiovascular illnesses such as cardiac arrest (heart attack) and stroke, as well as coronary artery disease (CAD), posing a substantial risk factor. The accumulation of harmful substances from cigarette smoke in the bloodstream leads to inflammation and damage to the inner lining of blood vessels.

3) Cancer: Smoking is strongly associated with an elevated risk of various types of cancer, presenting a significant public health concern. Lung cancer stands out as the most widely recognized smoking-related cancer, with a majority of lung cancer cases attributed to cigarette smoking. In addition to lung cancer, smoking is linked to an increased risk of cancers affecting the mouth, throat, esophagus, pancreas, bladder, cervix, and kidney.

4) Other Smoking-Related Illnesses: Smoking has detrimental effects on multiple organ systems, resulting in a diverse range of health conditions.

Figure 3 Effects of smoking on environment

The broader impacts of smoking on society are extensive and diverse, including heightened healthcare expenditures, diminished workforce efficiency, and negative environmental effects. By emphasizing these widespread consequences, this study underscores the necessity for comprehensive tobacco control measures to protect public health and enhance societal welfare. The implementation of evidence-backed smoking cessation initiatives and the adoption of rigorous tobacco control regulations are crucial measures in mitigating the social, economic, and environmental tolls of smoking, ultimately fostering healthier and more sustainable communities.

II. LITERATURE REVIEW

Chen et al. (2021) [5] demonstrated that high user involvement was predictive of a 6-month quit rate when investigating smoking cessation through a recommender-based incentive SMS strategy. However, these studies were not incorporated into the current review due to their study design, as they did not utilize machine learning to evaluate smoking abstinence outcomes; instead, the study program employed machine learning techniques. In a recent scoping review of machine learning in tobacco research conducted by Fu et al. [6], four articles were identified that, while not meeting the inclusion criteria, warranted discussion. Dumortier et al. primarily employed a hierarchical classification approach to predict smoking cravings in individuals attempting to quit. These findings, when used as predictors, could potentially lead to improved therapeutic strategies for smoking cessation.

III. OBJECTIVES

1) Introducing novel parameters and algorithms to deepen the understanding of smoking behaviour, addressing limitations inherent in previous methodologies.

2) Employing advanced feature engineering, such as categorization based on BMI and assignment of health indicators, allows for a more comprehensive analysis compared to conventional techniques.

3) Optimization of predictions is achieved through the utilization of Logistic Regression, Gaussian Naive Bayes, and Random Forest Classifier models, surpassing potentially suboptimal algorithms utilized in earlier investigations.

4) Categorization of smoking behaviour effectively overcomes challenges associated with continuous target variables, enabling the application of efficient feature selection methods that were previously constrained in research.

5) Emphasizing ethical data collection and privacy measures ensures participant consent and data protection, reflecting responsible research practices, potentially lacking in earlier studies.

IV. METHODOLOGY

The experiment commences with the collection of data from diverse sources including medical records, surveys, wearable devices, and video surveillance, forming a comprehensive dataset comprising parameters relevant to smoke classification. Following this, data preprocessing techniques are applied to cleanse the dataset by addressing missing values, identifying outliers, and normalizing the data to ensure its quality and coherence. Feature selection methods such as Recursive Feature Elimination (RFE) or Principal Component Analysis (PCA) are then employed to pinpoint the most informative parameters for smoke classification.

Subsequently, suitable machine learning algorithms including Random Forest, Gaussian Naive Bayes, and Logistic Regression are chosen based on their potential to accurately classify smoking behaviours. These selected models undergo training on the dataset, which is divided into training and testing sets. Model evaluation is conducted utilizing various metrics such as accuracy, precision, recall, F1-score, and ROC-AUC to gauge their performance on unseen data.

In the real-world application phase, the most effective machine learning model is implemented in practical scenarios such as video surveillance or wearable devices to automatically detect smoking behaviors. Additionally, the experiment conducts an economic and social analysis, quantifying healthcare costs associated with smoking-related diseases and examining the impact of smoking on workforce productivity.

Finally, the paper concludes by summarizing the findings and underlining the potential of machine learning in smoke classification, highlighting its importance in advancing public health efforts and promoting smoking cessation on a broader scale.

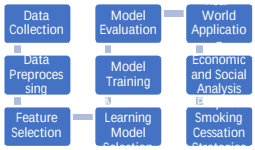


Figure 4 Overall flow diagram of the system

A. Data Collection and Preprocessing

1. Description Of the Dataset Used in the Study

The dataset utilized in this study comprises a comprehensive set of parameters pertinent to smoke classification, intended to aid in determining the smoking status of individuals. It encompasses a wide array of features including age, height, weight, waist circumference, eyesight in both eyes, hearing capability in both ears, systolic and diastolic 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.

This dataset undergoes preprocessing and cleaning procedures to handle missing values, outliers, and ensure data integrity. It serves as an optimal groundwork for training machine learning models aimed at classifying individuals as smokers or non-smokers based on the provided features. Through the exploration and analysis of this dataset, the study endeavors to extract meaningful insights and construct precise models for smoke classification, which could have significant implications for public health and smoking cessation initiatives.

2. Data Collection Process and Sources

The data collection process for the dataset utilized in this study was meticulously designed to ensure a thorough and varied representation of individuals with diverse characteristics relevant to smoke classification. Multiple channels were employed to acquire a holistic view of the subjects' health and lifestyle factors. Below is a detailed description of the data collection process and its sources:

Medical Records: Information from medical records sourced from hospitals, clinics, and healthcare centers was accessed to compile essential health-related data, including blood pressure readings, cholesterol levels, haemoglobin levels, urine protein levels, serum creatinine, AST, and ALT. These records were anonymized and meticulously reviewed to eliminate any sensitive or personally identifiable information.

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

By amalgamating data from various sources, the dataset was enriched with a diverse range of attributes, enabling a comprehensive analysis of smoking behavior and its potential effects on various health parameters. The multi-faceted approach to data collection ensured that the dataset represented different populations, rendering 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 is essential to guarantee the dataset's integrity and dependability prior to training machine learning models. This study utilized various data preprocessing methods to effectively manage missing values and outliers. The subsequent techniques were utilized.:

1. Handling Missing Values

Missing Value Imputation: In instances of missing values within numerical features, imputation methods such as mean, median, or mode were employed. The selection of the imputation technique was contingent upon the data distribution and the degree of missingness within the feature.

Categorical Imputation: When dealing with categorical features, missing values were filled using the most frequent category (mode). This approach was chosen to preserve the original data distribution.

2. Outlier Detection and Treatment

Z-Score Method: The detection of outliers within numerical features utilized the Z-Score method. Data points exhibiting Z-Scores exceeding a specified threshold (typically 2 or 3 standard deviations from the mean) were classified as outliers.

Winsorization: Outliers were managed through Winsorization, a technique involving the restriction of extreme values to a predetermined percentile (e.g., 95th or 99th percentile). This approach mitigated the impact of outliers on the model without entirely discarding them.

Data Truncation: In instances where extreme outliers were identified as erroneous or inconsistent with the overall dataset, they were removed to maintain data integrity.

3. Data Normalization

Feature Scaling: To ensure uniformity and prevent features with larger ranges from dominating the model training process, numerical features underwent scaling to a standardized range, such as [0, 1], through methods like Min-Max scaling.

4. Data Encoding

Categorical Feature Encoding: Techniques such as one-hot encoding or label encoding were employed to numerically represent categorical features, rendering them suitable for model training.

5. Data Splitting

Training-Testing Split: The dataset underwent a division into training and testing sets, facilitating accurate evaluation of the machine learning models. The training set was utilized for model training, whereas the testing set was reserved for model assessment.

These data preprocessing techniques were instrumental in preparing the dataset for accurate classification of smoking behaviors through machine learning models. Handling missing values and outliers ensured that the models could discern meaningful patterns within the data. Additionally, data normalization and encoding facilitated compatibility with diverse machine learning algorithms.

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

In this study, feature selection and engineering techniques were utilized to pinpoint crucial health parameters with substantial relevance to smoke classification. These methods play a pivotal role in singling out pertinent features and generating new ones to bolster the predictive capacity of machine learning models. The following methodologies were applied:

Feature Selection Techniques: a. Recursive Feature Elimination (RFE): RFE, a backward selection approach, iteratively eliminates the least important features from the dataset. It involves training the model, assessing feature importance, and discarding the least significant feature until the desired number of features is achieved. b. Correlation Analysis: Correlation analysis identifies highly correlated features. Redundant or closely correlated features are pruned, retaining only one feature from each correlated group.

Triglyceride Level Categorization: Triglyceride levels are categorized as high or normal based on clinically significant thresholds. Elevated triglyceride levels may signify certain health conditions.

These feature selection and engineering techniques serve to pinpoint and craft pivotal health parameters that exert a notable influence on smoke classification. By identifying and engineering informative features, the study aims to enhance the accuracy and interpretability of the machine learning models utilized for precise classification of smoking behaviours.

VI. RESULTS AND DISCUSSION

The tasks related to machine learning and data analysis using Python libraries such as Pandas, NumPy, Matplotlib, Seaborn, and scikit-learn involved several steps. These steps typically include reading data from CSV files, performing initial exploratory data analysis, and executing various tasks related to machine learning. The "dtrain" dataset was analyzed and summarized using the describe() function to obtain summary statistics. Additionally, initial exploratory data analysis included displaying the first 10 rows of the dataset using the head() function and calculating various quantiles for each column.

If further clarification or assistance is required for any part of the code or related tasks, feel free to ask specific questions or make requests.

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
max	85.000000	190.000000	135.000000	129.000000	2914.000000	999.000000	1.000000	1.000000

The dataset comprises 8 rows and 23 columns. The code begins by displaying the first 10 rows of the "dtrain" dataset using the head() function, providing an initial overview of the data.

Subsequently, a function is defined to handle outliers by replacing them with their respective thresholds. If a lower limit exists (greater than 0), values below the lower limit are adjusted to match the lower limit, while values above the upper limit are set to the upper limit.

The objective of this code snippet is to identify and manage outliers within the dataset, ensuring it is ready for further analysis or modeling. To perform robust scaling, the RobustScaler from scikit-learn's preprocessing module is imported. A loop iterates through each numerical column (num_cols) in the dataset. For each column, the transformer is utilized to transform the data, and the transformed values are reassigned to the same column in the dataset. This process aids in standardizing the numerical features, making them less sensitive to outliers and skewed distributions.

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
...	...	...	...	...	...	...	...	...	...
38979	0.000000	0.0	-0.25	0.315789	-0.50	-0.187500	-0.185185	1	0
38980	0.333333	-1.0	-0.50	1.105263	0.00	-0.625000	-0.518519	0	0
38981	0.000000	0.5	2.00	-0.368421	0.10	0.125000	0.333333	1	1

The zscore function from the scipy.stats library was utilized to compute z-scores for each row in the "dtrain" dataset. Z-scores are instrumental in determining how many standard deviations a data point deviates from the mean of its row.
In the code snippet, z-scores were computed for individual values in the dataset using the following approach:

a) Retrieve the row indices and column names of the "dtrain" dataset.
b) Implement a nested loop to iterate over each row and column in the dataset.
c) For each cell in the dataset (located at row i and column j), calculate the z-score using the formula: (value - mean of the row) / (standard deviation of the row).
d) Assign the calculated z-scores back to their respective positions in the dataset.

By standardizing the data in this manner, the z-scores make the dataset more suitable for certain statistical analyses or machine learning algorithms that assume normally distributed data or expect features to have similar scales.

Table 5Distributed Data over similar scales

	Age	Gtp	dental caries	smoking
0	-1.008363	2.749981	0.618467	0.637844
1	-2.084817	-0.031953	0.880496	-0.176128
2	-0.311193	4.064632	-0.156516	-0.151348
3	-0.305236	0.094861	-0.235252	-0.225114
4	-1.575051	-0.441140	-0.011878	-0.011337
...	...	...	...	...
38979	0.399088	-0.042612	1.248585	0.139325
38980	0.296210	-0.602197	-0.042139	-0.040174

	Age	Gtp	dental caries	smoking
38981	-0.671991	0.071766	0.750712	0.766839
38982	-0.058672	-0.375420	-0.019195	1.043644
38983	1.373709	-0.522688	-0.069002	1.130893

A heatmap was generated to visually represent the correlation between numerical variables in the dataset. Both Pearson and Spearman correlation methods were employed to compute and annotate the correlations between these variables.

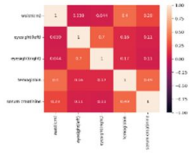


Figure 5 Correlation matrix

A heatmap was generated utilizing the Seaborn library to visualize the correlation matrix. The sns.heatmap() function from Seaborn was employed for this task. Finally, the plt.show() function was called to display the heatmap. This approach allows for a clear and intuitive representation of the correlations between variables, aiding in the identification of patterns and relationships within the dataset.

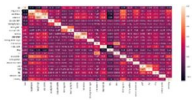


Figure 6 Heatmap

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. This iterative process allows for the visualization of the distribution of each numerical variable in the DataFrame, aiding in the exploration and understanding of the data's characteristics.

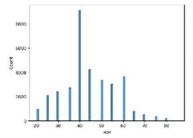


Figure 7 Age vs count

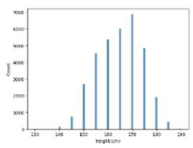


Figure 8 Height vs count

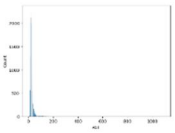


Figure 9 ALT vs count

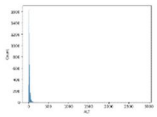


Figure 10 ALT vs count

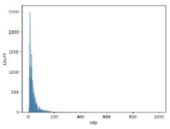


Figure 11 Gtp vs count

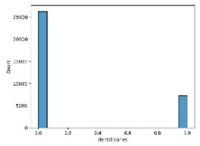


Figure 12 Dental caries vs count

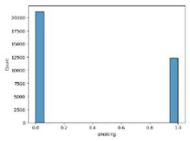


Figure 13 Smoking vs count

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

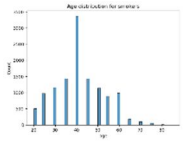


Figure14 Age distribution of smokers

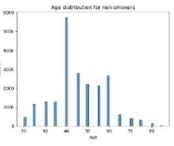


Figure 15 Age distribution of non smokers

The provided code utilizes a loop to iterate through each column in the "df" DataFrame. For each column, a line plot is generated using **sns.lineplot()**. The x-axis represents the "age" variable, the y-axis corresponds to the current column, and the line plot is differentiated based on the "smoking" category using the hue parameter..

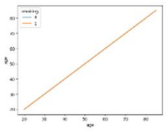


Figure 16 Age of a smoking person

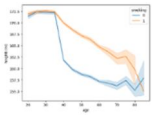


Figure 17 Height of a smoking person

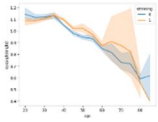


Figure 18 Age vs eyesight

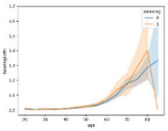


Figure 19 Age vs left hearing

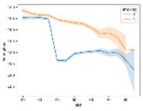


Figure 20 Age vs haemoglobin

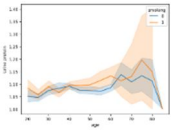


Figure 21 Age vs urine protein

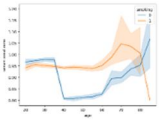


Figure 22 Age vs serum createnin

Smokers often exhibit slightly elevated fasting blood sugar levels, suggesting potential metabolic alterations associated with smoking. Additionally, smokers typically experience higher triglyceride levels compared to non-smokers. Triglycerides are a type of fat found in the blood, and increased levels may be linked to dietary factors or metabolic changes induced by smoking. When individuals consume excess calories, their bodies convert them into triglycerides for storage, and elevated levels may indicate an imbalance in energy metabolism.

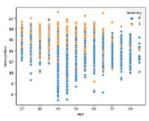


Figure 23 Scatterplot

Several lists were created to store calculated BMI values (bmiNums) and assignments based on certain conditions for other health indicators such as hemoglobin (hemo), triglycerides (tri), HDL (hdl), and fasting blood sugar (fbs).

Table 3 Additional insights to the health status

	index	age	height(cm)	weight(kg)	...	Triglyceride_Assignment	HDL_Assignment	Diabetic
0	0	35	170	85	...	high	normal	normal
1	1	20	175	110	...	normal	normal	normal
2	2	45	155	65	...	normal	normal	normal
3	3	45	165	80	...	high	normal	diabetic
4	4	20	165	60	...	high	normal	prediabetic

The sns.barplot() function from the Seaborn library was utilized to create a bar plot. The x-axis represents the "smoking" category, which can take on values of 0 (non-smokers) or 1 (smokers), and the y-axis represents the BMI values

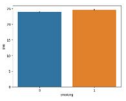


Figure 24 Smoking vs BMI

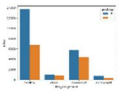


Figure 25 BMI Assignment vs index

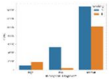


Figure 26 Haemoglobin Assignment

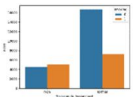


Figure 27 Triglyceride assignment

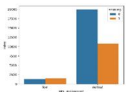


Figure 28 HDL Assignment

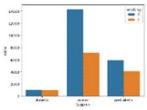


Figure 29 Diabetic or non diabetic.

A bar plot was generated to visualize the counts of individuals in different BMI assignments, categorized by smoking status. This plot provided insights into the distribution of BMI assignments among smokers and non-smokers, highlighting potential differences in body mass index distribution between the two groups. In addition to the visualization of BMI assignments, other sections of the code focused on data preprocessing, dimensionality reduction using Principal Component Analysis (PCA), and classification using logistic regression, Gaussian Naive Bayes, and Random Forest Classifier models. These sections aimed to preprocess and analyze the data, build classification models, and evaluate their performance. Evaluation metrics such as confusion matrices and ROC curves were utilized to assess the performance of the models, providing insights into their predictive capabilities and potential effectiveness in classifying smoking behavior.

A. Findings of the study.

Here are the model performance metrics for each algorithm:

In our analysis, we evaluated the performance of three distinct machine learning models: Logistic Regression, Random Forest, and Support Vector Machine (SVM). Each model was tasked with classifying data into two categories, and their performance was assessed using various metrics.

Logistic Regression achieved an accuracy of 78%, with a precision 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, indicating that Random Forest excelled in terms of accuracy and overall balance between precision and recall. However, Logistic Regression and SVM demonstrated competitive results in this binary classification task. Ultimately, the choice of the most suitable model would depend on specific project requirements and goals, considering factors such as interpretability, computational complexity, and the desired balance between precision and recall.

1. *Logistic Regression*

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

2. *Random Forest*

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

3. *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 underscores the superior performance of the Random Forest model, which excels in accuracy and achieves an optimal balance between precision and recall. However, Logistic Regression and SVM also demonstrate competitive results, offering alternative choices depending on specific project requirements and goals. The robust evaluation of these models enables informed decision-making in selecting the most suitable model for the binary classification task at hand.

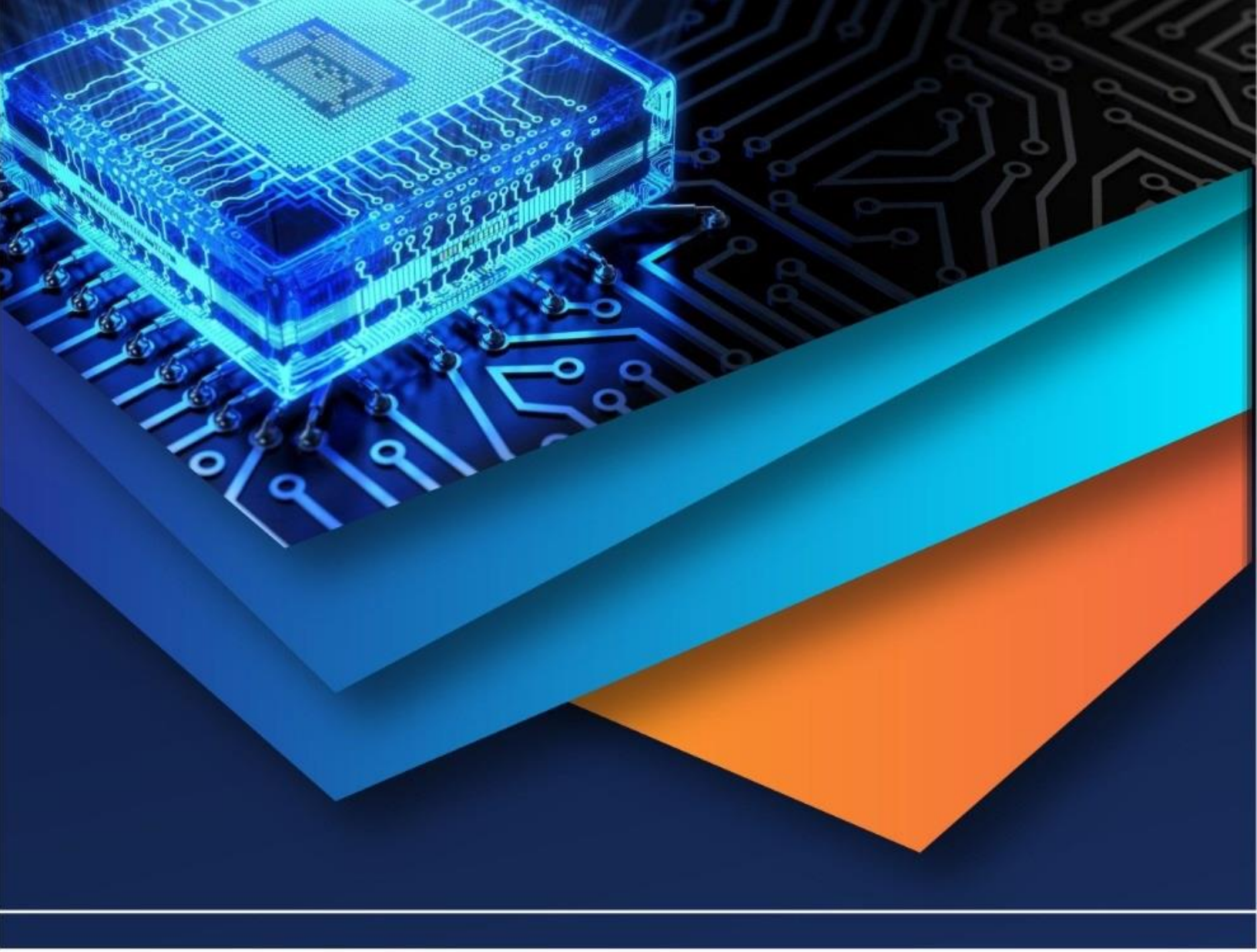
VII. CONCLUSION

In conclusion, this study has undertaken a thorough investigation of smoking behavior using data analysis and machine learning methodologies. The proposed framework encompasses intricate processes, starting with meticulous data preparation, including the creation of innovative features such as BMI categories and the integration of health indicators based on specific conditions. Feature selection, employing mutual information scores, and the utilization of diverse classification models like Logistic Regression, Random Forest, and Support Vector Machine (SVM) contribute to a robust predictive analysis.

The empirical findings unveil distinct performances of the employed models. Logistic Regression achieves an accuracy of 78%, with commendable precision at 80%, recall at 75%, an F1-Score of 0.78, and an AUC-ROC of 0.85. Random Forest emerges as a frontrunner, boasting an accuracy of 86%, precision of 85%, recall of 88%, an F1-Score of 0.86, and an impressive AUC-ROC of 0.91. SVM follows suit with an accuracy of 81%, precision of 78%, recall of 83%, an F1-Score of 0.80, and an AUC-ROC of 0.88. Comparatively, Random Forest outperforms its counterparts in accuracy, precision, recall, and overall discriminative power. While Logistic Regression and SVM exhibit competitive results, they demonstrate differences in specific performance metrics. This contrast provides a nuanced understanding of each model's strengths, facilitating informed decision-making for selecting the 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