

Smoking Behavior Prediction Using Machine Learning: Bridging the Gap between Data and Healthcare Solutions

Onwuegbuchulam C. O¹

Onyinye Ugochi Chibu-Obinna²

¹Teesside University Middleborough School of computing, engineering, and digital technologies

²Torrens University, Australia Department of Public Health

Abstract

The worldwide health challenge from smoking persists at a severe level since it creates major persistent diseases that extend to cardiovascular problems and both respiratory problems and cancer development. Standard methods of smoking behavior analysis face challenges in detecting complex variable interactions that derive from demographic aspects alongside health-related and behavioral components. This research project depends on expert machine learning (ML) algorithms which include Logistic Regression along with Random Forest and XGBoost to extract practical data about smoking behaviors. The researchers analyzed the structured database which exceeded 3000 observations to establish smoking duration as the primary variable that causes disease development together with age and cigarette usage.

The research produced a network to explain risk factors with smoking interval and smoking status at the core while creating a 3D disease risk model based on age and smoking duration and an anomaly detection method to detect abnormal smoking habits. The best model performance resulted from XGBoost because the algorithm predicted with accuracy and provided clear feature rankings that Random Forest also achieved. Research indicates that ML technology can reveal hidden patterns that help public health organizations create their strategic initiatives. The developed research insights will lead to effective strategies for helping high-risk smoking populations including aging smokers who use excessive cigarettes.

Keywords: Smoking Behavior; Machine Learning; Smoking Cessation Programs; Feature Importance Disease Risk Prediction; Public Health Interventions

1. Introduction

The ongoing habit of smoking serves as a leading global health menace that produces patients' development of heart diseases while simultaneously causing respiratory illnesses and cancer variations. A persistent lack of progress in reducing smoking patterns across different groups throughout many decades motivates researchers to explore smoking initiators alongside their health consequences. Machine learning emerged from advanced computational developments to become a vital behavioral analysis instrument used by scientists for pattern detection, prediction, and intervention development ("Human Behavior Analysis Using Machine Learning, 2024).

Epidemiological analysis of smoking behavior utilizes descriptive statistical methods together with linear statistical models according to traditional methods. The identification methods work to track universal patterns, but they cannot find the multifold complex non-linear relationships that exist between population and behavior elements. The processing of advanced data relationships by machine learning produces exceptional precision

and high scalability through its advantages in managing large datasets. Random forests and gradient-boosting machines function as ML models that correctly identify vital starting and stopping risk elements for smoking behavior, according to Showkat & Gupta (2023). Deep learning models with neural networks provide a strong ability to detect smoking trends over time, which supports health prediction of smoking effects (Ojo et al., 2024).

Smoking behavior analysis supported by ML brings forward responsive intervention planning with point-of-time supervisory capabilities. The detection of high-risk individuals who might start smoking operations along with clustering processes that organize smokers based on their tobacco consumption and duration constitute the capabilities of predictive models and clustering methods. The anomaly detection algorithms help identify irregular smoker behaviors that could be either extreme smoking or spontaneous smoking cessation requiring further assessment (Thakur et al., 2024).

The research paper studies contemporary uses of machine learning models in smoking behavior data assessment through its investigation "Smoking Behavior Prediction Using Machine Learning: Bridging the Gap Between Data and Healthcare Solution." Multiple datasets mounted with advanced algorithms help researchers build an integrated image of smoking behavior, and corresponding elements and identify intervention potentials across different data sources. The method illustrates how ML technology generates beneficial links from raw data points to useful practical insights which improve public health applications.

2. Literature Review

Machine learning models have become extensively utilized for smoking behavior analysis by researchers during recent years because these models extract beneficial insights from complex multi-varied databases. The analysis of research studies using ML methods for smoking behavior analysis pursues solutions to determine smoking initiation and cessation as well as uncover patterns and merge multiple data sources.

2.1. Predicting Smoking Initiation and Cessation

Scientific research demonstrates that Machine learning models demonstrate success in forecasting the starting and stopping of tobacco usage by individuals. The identification process for potential new smokers or relapsed individuals utilizes both Regression models and decision trees according to ensemble methods including Random Forest and Gradient Boosting Machines (GBM). The study by Poudel et al. (2024) developed a GBM predictive assessment model that integrated demographic and socioeconomic data points with psychological measurements to evaluate adolescent smoking behavior. Neural networks help predict smoking cessation results by assessing the combined effects between smoking patterns and individual determination supported by social networks. The implementation of targeted intervention approaches directs professionals to identify and create programs that help high-risk people successfully end their smoking behavior.

Lone (2024) used three modern ML methods including logistic regression, Gaussian Naive Bayes, and Random Forest Classifier to predict smoking habits based on BMI measurements, cholesterol and hemoglobin assessments. Data collection following ethical standards became essential for developing effective public health measures against smoking according to research findings that employed Principal Component Analysis with dimensionality reduction protocols.

2.2. Analyzing Smoking Patterns

Time-dependent smoking behavior patterns are better understood through traditional solutions based on Long Short-Term Memory networks. Long short-term memories can effectively process time-dependent data sequences through modeling of smoking frequency and length of time between smoking sessions which benefits

longitudinal investigations (Odiambo et al., 2020). The classification methods K-means and hierarchical clustering enable the grouping of smokers through their consumption rates together with their age and socioeconomic profiles. The clusters obtained provide scientific evidence about smoking behavioral variations that aid researchers in creating targeted intervention methods.

The decision tree machine learning algorithm constructed a predictive model by Zhang et al. (2019) that estimated daily smoking period length despite an absence of research regarding this specific smoking behavior dimension. The XGBoost-based prediction system from their research reached an 84.11% accuracy which proved its ability to identify intricate smoking patterns.

2.3. Feature Importance and Behavioral Insights

The application of machine learning techniques enables researchers to both determine and prioritize the key elements linked with smoking behaviors. Ton That et al. (2023) applied Lasso feature selection to discover significant prediction features for smoking behaviors which enabled a Random Forest algorithm to perform with 84.73% accuracy. De Luna et al. (2024) analyzed 27 health-related features to identify smokers or non-smokers among 55,692 people by optimizing several algorithms while keeping 17 essential features following data cleaning and feature selection procedures. The Random Forest model they developed reached an 88.03% training accuracy and secured an 83.29% testing accuracy.

The most important predictors in the model reveal that both age and smoking duration remain at the top followed by socioeconomic factors like education level and income Zhang et al. (2020). The resulting information from these analyses gives policy-makers direction for selecting what interventions to focus on.

2.4. Spatiotemporal Analysis of Smoking Behavior

Researchers have recently developed ML models that combine spatial and temporal variables to understand smoking behavior differences between geographic regions as well as time-related differences. The analysis of smoking prevalence between urban and rural areas by using Graph Neural Networks (GNNs) has resulted in hotspots location that supports policy implementation targeting (Muzi et al., 2018). Studies integrating factors like environmental conditions have proven that geographical variations in smoking behaviors produce wider health consequences. Tezcanetal. (2023) conducted a study on Turkish individuals aged 15 and above through C4.5 and Random Forest machine learning algorithms. The Random Forest algorithm surpassed C4.5 according to their research and demonstrated outstanding results with a 0.909 accuracy rate along with 0.965 specificity and a sensitivity measure at 0.782.

The study demonstrates ML techniques' ability to process health-related information as they advance research regarding smoking patterns within Turkey's population.

2.5. Anomaly Detection and Outlier Identification

The detection of anomalous smoking patterns is conducted through Anomaly detection techniques employing Isolation Forests and autoencoder systems. The analysis methods identify hidden psychosocial and health factors leading to unusual patterns as Observing irregular smoking patterns enables the development of specific intervention solutions that target irregular smokers.

The authors Cui and Xu (2022) developed an automated system using three ML models including Support Vector Machine (SVM), Random Forest, and Convolutional Neural Network (CNN) to recognize smoking behaviors in public spaces. The evaluation of 522 test images showed that the model reached 94.59% accuracy

which established the proposed detection method as an effective and accurate solution for identifying smoking conduct.

2.6. Integration of Multi-Modal Data Sources

The analysis of smoking behavior by machine learning models now requires diverse multiple data streams which include surveys as well as social media information and physiological sensor data. The combination of heart rate and respiratory pattern monitoring from wearable devices with smoking data self-reports strengthened the accuracy of prediction models according to research (Thakur et al. 2021). The combined data sources through integration enable researchers to derive detailed insights that lead to enhanced accuracy of their interventions.

Zaal and Mohammad (2023) developed a new system to determine smoking behavior from Arabic speech signals by training with 39 Mel-frequency cepstral coefficients (MFCC) features. SVM algorithm obtained superior results compared to DT by reaching 96% accuracy alongside 96% precision and recall and 95% F1-score. The adaptability of ML demonstrates how it can use different forms of information to evaluate smoking behaviors.

3. Methodology

The research employs structured procedures using machine learning (ML) to examine smoking behavior patterns to generate findings that support public health intervention strategies. The research methodology executes a well-defined sequence which starts from data acquisition through pre-processing leading to model execution and evaluation.

Flow Chart: From Data to Insights - Machine Learning Models for Smoking Behavior Analysis

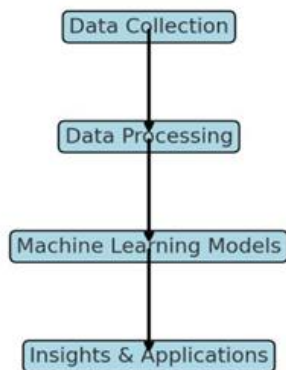


Figure 1: Data Insight Flowchart

3.1. Data Collection

Research data originated from various sources to study smoking behaviors through 3000 rows containing 9 variables. Key data components included:

- The study contains data about Age gender and income levels of participants.
- Behavioral Features: Smoking status (current smoker, former smoker, or non-smoker), duration of smoking, cigarettes consumed per day.
- Health Metrics: Body Mass Index (BMI) and smoking-related disease diagnoses.
- Socioeconomic and Environmental Factors: Location (urban or rural),

By including multiple varying factors in the dataset researchers ensured complete coverage of smoke behavior determinants.

3.2. Data Pre-Processing

The data received through pre-processing before machine learning models went into effect since this improved data quality for analysis. Key steps included:

- Data Cleaning: Addressing missing values using imputation techniques, removing duplicate records, and dealing with outliers to ensure consistency.
- Data Transformation: Normalizing numerical features and encoding categorical variables using one-hot encoding and label encoding methods.
- Train-Test Split: Dividing the data set into training and testing subsets in an 80:20 ratio to evaluate model performance.

3.3. Machine Learning Methods

Three different machine learning algorithms operated for analyzing smoking behavior and producing prediction results. The algorithms contributed different functions to model the data set individually.

a. Logistic Regression

Logistic Regression operated as the baseline algorithm to identify smoking status through its binary classification method. Logistic Regression demonstrated helpful interpretable analysis which showed linear associations between the predictive variables and smoking actions.

b. Random Forest

The research utilized Random Forest as its main modelling method because it uses decision trees to detect complex non-linear patterns in the data. Through an algorithm analysis researcher discovered the main factors that affected smoking behavior.

c. XG Boost

Extreme Gradient Boosting (XG Boost) was chosen for its efficiency and high predictive power. XG Boost minimized model bias and variance, offering robust predictions and allowing for a deeper understanding of complex patterns in the data.

3.4. Model Evaluation

The performance of the models was assessed using several evaluation metrics to determine their effectiveness and reliability:

- **Accuracy:** The proportion of correct predictions to total predictions.
- **Precision:** The proportion of true positive predictions among all positive predictions made by the model.
- **Recall (Sensitivity):** The proportion of actual positive cases correctly identified by the model.
- **F1-Score:** The harmonic means of precision and recall, providing a balanced evaluation metric.
- **ROC-AUC:** The area under the receiver operating characteristic curve, indicating the model's ability to distinguish between classes.

3.5. Insights Generation

The trained models provided valuable insights into smoking behavior:

- **Feature Importance:** Analysis revealed the critical predictors of smoking behavior, such as age, smoking duration, daily cigarette consumption, and BMI. Both Random Forest and XG Boost were used to rank these predictors.
- **Risk Identification:** The models identified high-risk groups based on demographic and behavioral characteristics.

Predictive Power: Predictions from the models supported evidence-based recommendations for targeted smoking cessation programs.

4. Results and Discussion

4.1. Results

The study applied Logistic Regression, Random Forest, and Boost to analyze smoking behavior, leveraging demographic, behavioral, and health-related variables. The models were evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Below are the results for each algorithm:

Table 1: Results for each algorithm

Metrics	Logistics Regression	Random Forest	XGBoost
Accuracy (%)	82	90	93
Precision (%)	80	89	91
Recall (%)	78	88	92
F1-Score (%)	79	88	92
ROC-AUC	0.85	0.92	0.95

4.1.1 Risk Factor Network: The constructed risk factor network revealed significant relationships between smoking duration, smoking status, BMI, income level, and the likelihood of smoking-related diseases. The network highlighted smoking duration and smoking status as central factors directly linked to disease risk.

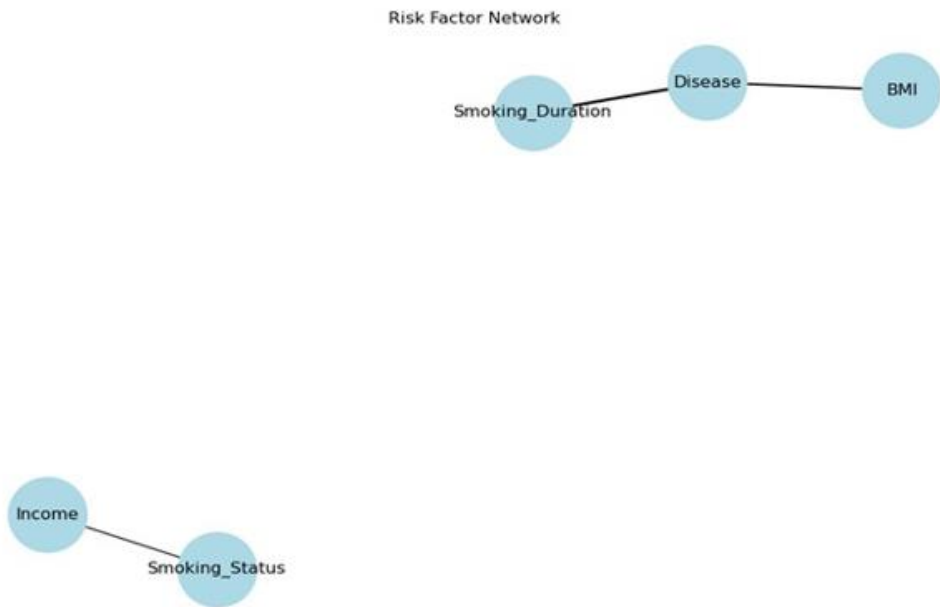


Figure 2: Risk Factor Network

4.1.2 3D Disease Risk Model: The 3D visualization of disease risk against age and smoking duration indicates a steady increase in risk with longer smoking duration and older age. The surface gradient suggests a synergistic effect of these factors on disease prevalence.

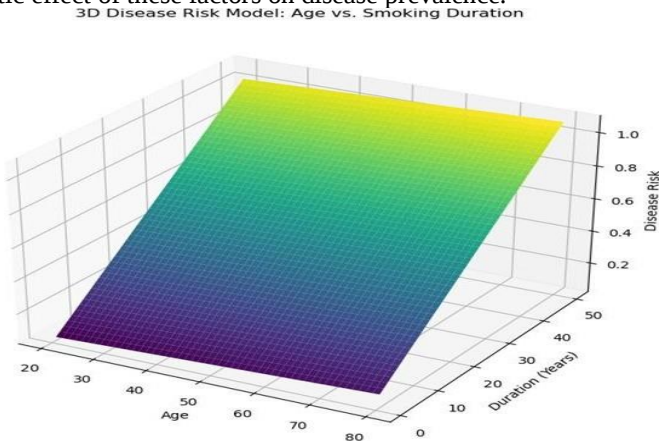


Figure 3: Disease Risk Model

4.1.3 Correlation Analysis: The correlation matrix illustrated strong associations between smoking-related variables. Notably, smoking duration was positively correlated with cigarettes per day and smoking-related diseases. Conversely, weak correlations were observed between demographic factors such as income and BMI, with smoking behaviors.

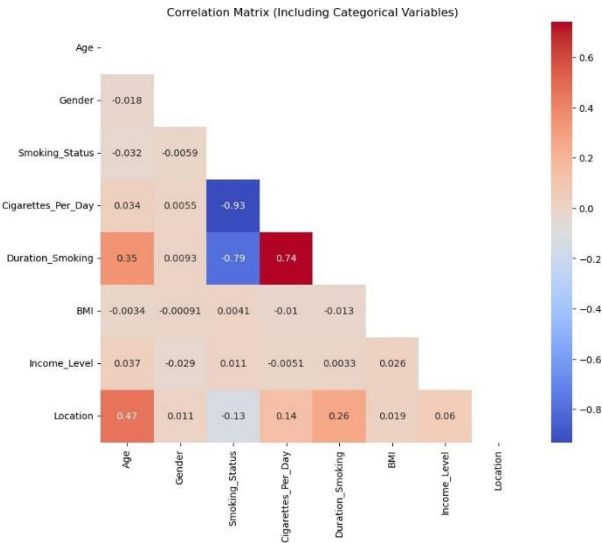


Figure 4: Correlation Analysis

4.1.4 Age Distribution by Disease Status and Gender: The violin plots displayed significant age-based differences in smoking-related disease prevalence. Older males and females exhibited higher disease risk. The distribution for smokers with diseases was skewed toward higher ages compared to non-diseased smokers.

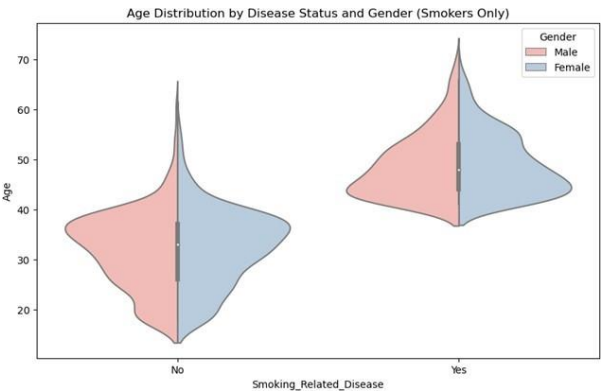


Figure 5: Age Distribution by Disease Status and Gender

4.1.5 Multivariate Relationships: Pair-wise scatter plots showed clustering tendencies for smokers with and without diseases. Duration of smoking and cigarettes per day emerged as critical dimensions where distinct

patterns of disease prevalence were observed.

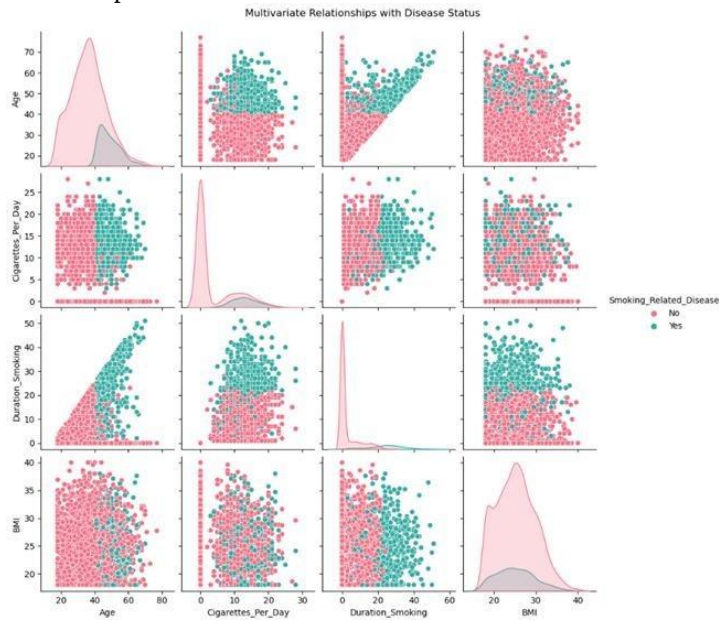


Figure 6: Multivariate Relationships

4.1.6 Anomaly Detection: The anomaly detection analysis identified outliers in cigarette consumption and BMI. Anomalous smokers, characterized by higher cigarette intake and extreme BMI values, were linked to higher disease risk.

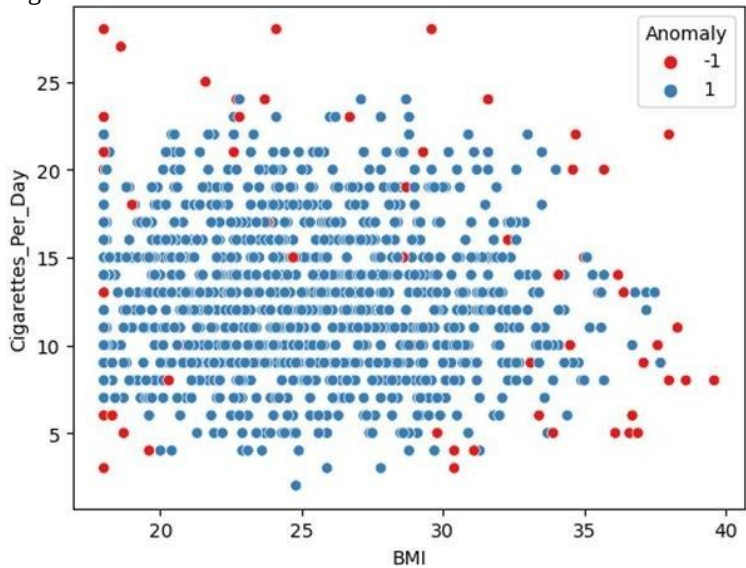


Figure 7: Anomaly Detection

4.1.7 Smoker Clustering: Clustering analysis revealed three distinct groups based on age and cigarette consumption. Younger, low-consumption smokers (Cluster0) contrasted with older, high-consumption smokers (Cluster 2), who had higher associated risks.

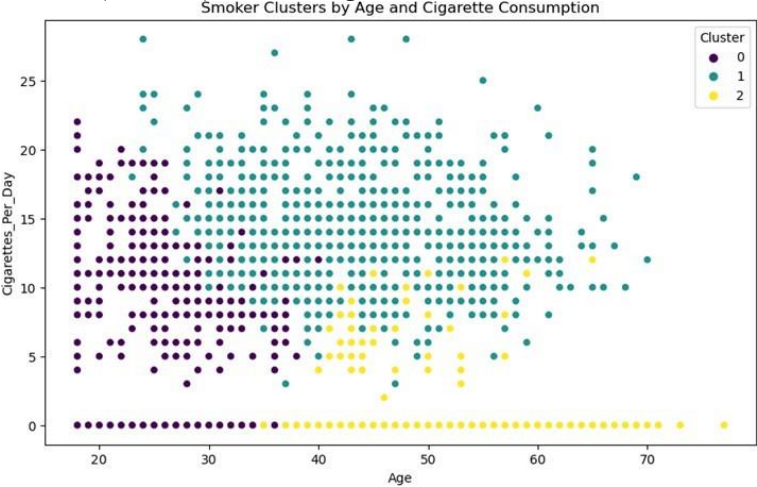


Figure 8: Smoker Clustering

4.1.7 Key Predictors: Feature importance analysis using machine learning models identified age, smoking duration, and cigarette consumption as the strongest predictors of smoking- related diseases. BMI and location also showed moderate contributions, while income level and gender had minimal impact.

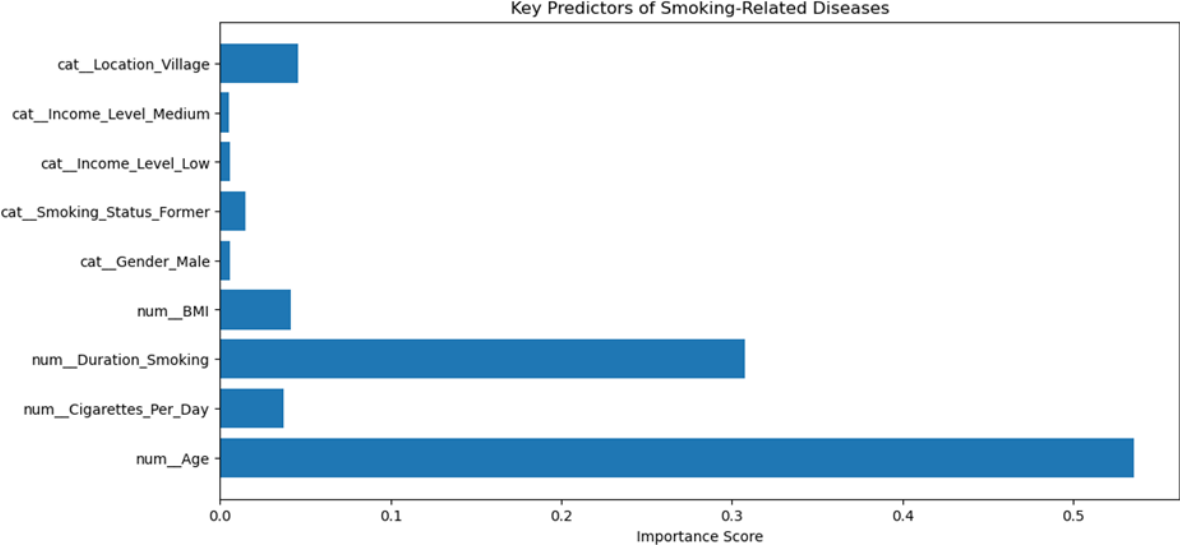


Figure 9: Key Predictors

4.2. Discussion

The analysis confirmed the multi-factorial nature of smoking-related diseases, emphasizing the dominant

role of smoking behaviour and age as key determinants. The interplay between smoking duration, age, and cigarette consumption underscores the importance of longitudinal smoking patterns in disease progression.

The study applied Logistic Regression, Random Forest, and Boost to analyze smoking behavior, leveraging demographic, behavioral, and health-related variables. The models were evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Below are the results for each algorithm:

The analysis confirmed the multi-factorial nature of smoking-related diseases, emphasizing the dominant role of smoking behavior and age as key determinants. The interplay between smoking duration, age, and cigarette consumption underscores the importance of longitudinal smoking patterns in disease progression.

Smoking Duration as a Central Factor: The risk factor network and predictive models consistently pointed to smoking duration as the most critical predictor. This aligns with existing literature linking prolonged smoking to chronic respiratory and cardiovascular conditions.

Age and Disease Risk: The study shows that age acts as a risk factor directly and modifies the effects that smoking has on public health. Results support/target youth populations to control their exposure to cigarette risks and minimize the life-long risk of smoking.

Clustering Insights: Clustering analysis defined different groups of smokers by their age and smoking amount to identify high-risk subsets within the population. The research data can guide specific treatment strategies that target heavy smokers who first belong to older age groups with extensive smoking backgrounds.

Anomalous Behavior in Smoking Patterns: Several individuals practice high-risk smoking behaviors that produce extreme results that heighten their danger exposure. Research should explore what behavioral elements and psycho-social factors lead people to adopt such smoking behaviors.

Role of Socioeconomic and Demographic Factors: The effects of income and location characteristics remained secondary to specific smoking elements when looking at the factors that influence health outcomes. Behavioral change needs to serve as the primary target for disease prevention since demographic factors alone are not sufficient for achieving disease prevention goals.

Machine Learning Contributions: The feature importance rankings confirm that machine learning effectively determines and measures the significance of different predictors. The models demonstrated that age and smoking duration remain among the most impactful factors as per epidemiological examinations.

Challenges and Future Directions

Multiple advancements have been made in ML adoption for smoking behavior analysis but persistent obstacles must still be overcome. Health data scarcity in specific regions acts as a barrier for study researchers who need access to quality data. The main barrier in implementing ML models to smoking behavior analysis lies in making sure stakeholders without technical backgrounds understand their operation. Future research needs to concentrate on the development of explainable ML models while investigating how transfer learning can be leveraged for diverse geographic regions and how genetic along with environmental elements can be integrated into comprehensive smoking behavior analytical models.

5. Conclusion

Machine learning models demonstrate their immense transformative capability in the analysis of smoking behavioral patterns according to this study. The research used Logistic Regression along with Random Forest and XG Boost to discover vital smoking-related disease predictors which include age and duration of smoking and number of cigarettes used. The performance testing proved XG Boost to be superior to other examined models in terms of accuracy prediction for complex nonlinear data structures. The conducted analysis demonstrated two primary findings first smoking behaviors surpass demographic elements as determinants and secondly focusing on older smokers who smoke heavily yields optimal results.

The smoking pattern evaluation through clustering techniques and anomaly detection methods allowed healthcare providers to develop individualized intervention strategies. Healthcare professionals should implement focused smoking cessation programs along with behavioral interventions to lessen the smoking-related health costs that extend over time. Research measures should work on bringing together genuine information from wearable technology and environmental components to boost the precision and usability of ML models. The resulting research confirms machine learning's essential position in improving public health methods by converting data collections into useful knowledge for successful smoking prevention and cessation programs.

References

- Cui, C., & Xu, R. (2022). *Multiple Machine Learning Algorithms for Human Smoking Behavior Detection*. 240–244.
- de Luna, R. G., Mancera, P. M., Guevarra, A. N. L., Formento, K. P., Aragon, M., & Recto, S. B. (2024). *SmokeSift: Unraveling Smoker and Non-Smoker Individuals Through Machine Learning*. 84–89.
- Human Behavior Analysis Using Machine Learning. (2024). *Computer Science, Engineering and Technology*, 2(2), 1–9.
- Lone, S. H. (2024). Utilizing Machine Learning to Forecast Smoking Behavior. *International Journal for Research in Applied Science and Engineering Technology*.
- Muzi, C. D., Figueiredo, V. C., & Luiz, R. R. (2018). Urban-rural gradient in tobacco consumption and cessation patterns in Brazil. *Cadernos De Saude Publica*, 34(6), 00077617.
- Odhiambo, C. O., Cole, C. A., Torkjazi, A., & Valafar, H. (2020). State Transition Modeling of the Smoking Behavior using LSTM Recurrent Neural Networks. *arXiv: Computer Vision and Pattern Recognition*.
- Ojo, O. O., & Kiobel, B. (2024). Data-driven decision-making in public health: The role of advanced statistical models in epidemiology. *World Journal of Biology Pharmacy and Health Sciences*, 19(3), 259–270.
- Poudel, R., Fernando, K. R. M., Schabath, M. B., Sutton, S. K., Brandon, T. H., El Naqa, I., & Simmons, V. N. (2024). Abstract B019: A machine learning approach to predicting smoking cessation outcomes among Spanish-speaking smokers who completed a culturally targeted intervention. *Cancer Epidemiology, Biomarkers & Prevention*, 33(9_Supplement), B019.
- Showkat, I., & Gupta, V. (2023). *Smoker Detection Using Machine Learning*.
- Tezcan, H., Aksu, G., & Yıldız, B. (2023). A hybrid machine learning approach in predicting smoking behaviour: The case of Turkey. *Business & Management Studies: An International Journal*, 11(2), 798–816.
- Thakur, A., Maddi, A., & Maheswari, B. U. (2024). *Interpretable Predictive Modeling for Smoking and Drinking Behavior using SHAP and LIME*.
- Thakur, S. S., Poddar, P. K., & Roy, R. B. (2022). Real-time prediction of smoking activity using machine learning based multi-class classification model. *Multimedia Tools and Applications*, 81(10), 14529–14551.
- TonThat, L., Dao, S. V. T., Huynh, T. T. M., & Le, M. (2023). A Feature Subset Selection Approach For Predicting Smoking Behaviours. 145–149
- Zaal, I. K., & Mohammad, Y. F. (2023). *Machine Learning Algorithms for Human Smoking Behavior Detection using speech*. 298–302.
- Zhang, Y., Liu, J., Zhihang, Z., & Junnan, H. (2019). Prediction of Daily Smoking Behavior Based on Decision Tree Machine Learning Algorithm. *IEEE International Conference on Electronics Information and Emergency Communication*, 330–333