Predictive Modeling of Smoking Behavior from Biomedical Data: A Supervised Machine Learning Approach

Report Prepared for:

CSE422: Artificial Intelligence
Department of Computer Science and Engineering
Brac University
Dhaka, Bangladesh

Report Prepared By:

Syeda Maliha Tabassum
ID: 22201574
Department of Computer Science and Engineering
Brac University

Zawad Yousuf
ID: 24241351
Department of Computer Science and Engineering
Brac University

May 26, 2025

CONTENTS

I. Introduction

This project investigates the use of biomedical data and supervised machine learning to predict a person's smoking status. One of the main preventable risk factors for long-term conditions like cancer, heart disease, and chronic obstructive pulmonary disease (COPD) is still smoking. Clinical evaluations, however, frequently depend on self-reported information, which can be tainted by deliberate misrepresentation, recall bias, or social stigma. We suggest a data-driven model that was trained on the anonymised medical records of both individuals and medical students in order to overcome this limitation. Age, gender, height, weight, body mass index (BMI), temperature, heart rate, blood pressure, cholesterol, and diabetic status are just a few of the many physiological and clinical characteristics included in the dataset. These characteristics may be indicators of underlying health trends linked to smoking behaviour. The main goal of the project is to create a predictive tool that, without requiring self-disclosure, can determine a patient's smoking status based on their health profile. For this task, models like k-nearest neighbours, logistic regression, decision trees, naïve Bayes, and a specially designed neural network were assessed. The models seek to provide probabilistic predictions indicating whether a person is likely to smoke by learning correlations between the input variables and the binary target—smoking. This method has a lot of clinical applications. In situations where trustworthy patient reporting is not available, it can help medical professionals identify at-risk individuals, provide tailored cessation interventions, and guide treatment decisions. Additionally, it illustrates how AI can improve preventive healthcare, lower diagnostic uncertainty, and facilitate more unbiased, data-driven clinical workflows.

II. Dataset Description

A. Dataset Analysis

The dataset consists of 200,000 anonymised medical records, each with 13 features, that were gathered from medical students. Along with categorical characteristics like gender, blood type, and diabetes, these features also include quantitative variables like age, height, weight, BMI, temperature, heart rate, blood pressure, and cholesterol. Whether a person smokes (Yes) or does not (No) is indicated by the target variable, smoking. Because the model needs to differentiate between two different classes, this turns the task into a binary classification problem. To examine the associations between the input variables and the encoded target feature Smoking-enc, a correlation heatmap was created. The findings show a very weak linear relationship between smoking behavior and characteristics. Features such as blood pressure, cholesterol, and BMI have very little correlation with the target. As might be expected given their mathematical dependence, the strongest internal correlation was found between weight and BMI (r - 0.82). The necessity for multi-feature interaction modelling is highlighted by the fact that no single feature seems to have a strong predictive relationship with Smoking-enc.

TABLE I: Correlation between the input and output

Values	Smoking-enc
Smoking-enc	1.000000
Blood Type-O	0.007985
Age	0.004791
Gender-Male	0.002867
Height	0.002544
Blood Type-B	0.002505
Diabetes-Yes	0.000194
Blood Pressure	0.000075
Heart Rate	-0.000638
Weight	-0.000859
Blood Type-AB	-0.001224
BMI	-0.002135
Temperature	-0.002913
Cholesterol	-0.003800

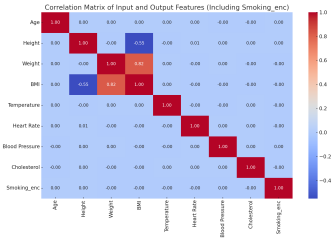

Fig. 1: Correlation Heat Map

Fig. 2: Correlation with Different strategies

B. Imbalanced Dataset

The dataset exhibits a significant class imbalance. Out of the total 200,000 records, 163,971 instances (82) are labeled as non-smokers (Smoking-enc = 0), while only 36,029 instances (18) are labeled as smokers (Smoking-enc = 1). This imbalance poses

a challenge for most classification algorithms, which may favor the majority class. To counteract this, techniques such as class weighting and oversampling (e.g., SMOTE) are considered to ensure that the model performs effectively across both classes. In summary, while the dataset provides a diverse range of biomedical features, the weak correlations and imbalanced class distribution emphasize the importance of robust preprocessing and model selection to accurately predict smoking behavior.

TABLE II: Number of Instance in unique classes of Output feature

Smoking	Count
No	143971
Yes	36029

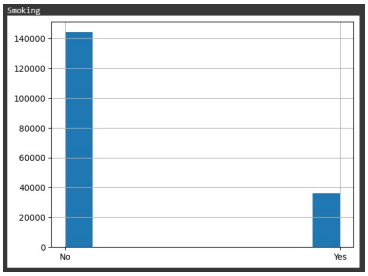

Fig. 3: Chart of unique classes of Output feature

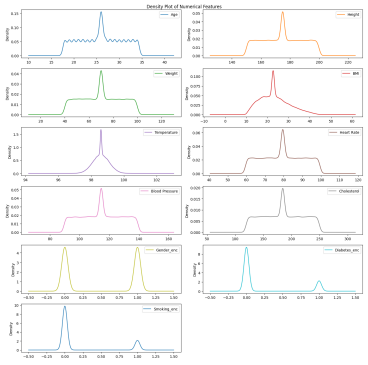

Fig. 4: Densities of differetn features

III. Dataset Preprocessing

The dataset required several preprocessing steps to address quality issues and ensure it was suitable for machine learning tasks. Each problem was systematically identified and resolved using appropriate data handling techniques, as outlined below

A. Missing Values

The first and most noticeable problem was that several features had missing values. There were gaps in all important numerical columns like height, heart rate, cholesterol, and BMI, and binary categorical features like diabetes and smoking were also not fully complete. Given that all the features had around 10 percent missing entries, simply removing rows or columns with missing values would have resulted in a substantial loss of data and possible bias.

We used imputation techniques specific to each feature type to address this:

We employed mean imputation, which maintains the central tendency without changing the distribution, for features like height, heart rate, blood pressure, and cholesterol that have comparatively symmetrical distributions and fewer outliers. We chose median imputation for features like age, BMI, and temperature that had skewed distributions or obvious outliers. This approach avoids artificial inflation or deflation of the mean and is more resilient when dealing with extreme values.

We used localised mode imputation for categorical features such as blood type and gender. We separated the dataset into equal segments and used the mode of each segment to fill in the missing values rather than a global mode, which can distort category distributions. As a result, the diversity and distributional integrity of categorical variables were preserved.

In order to maintain minority class balance for supervised learning, we assumed a conservative approach and filled in missing values for binary fields such as diabetes with ”Yes.” After examining the distribution of classes and evaluating the impact on model training.

B. Irrelevant Features

Although the Student ID column in the data set provided a unique identifier, it was useless for modeling. Adding such a feature might cause noise or cause distance-based models to be misled. Solution: In order to reduce dimensionality and remove extraneous information, we dropped the Student ID column early in the preprocessing stage. This made the data set easier to understand without sacrificing any significant signal.

C. Categorical Variables

The dataset contained a number of categorical features, such as smoking, blood type, gender, and diabetes. To be compatible with machine learning algorithms, particularly those that need fixed-length numeric input (e.g., logistic regression, neural networks), these had to be encoded into numerical values.

We used two different kinds of encoding methods: Label Binary categorical features like diabetes, gender, and smoking were encoded, with ”Yes” and ”No” or ”Male” and ”Female” being mapped to 1 and 0, respectively. To avoid the model assuming any ordinal relationship between the categories, One-Hot Encoding was used for nominal categorical variables, such as Blood Type. This allowed for the equitable treatment of all classes by producing multiple binary columns that represented each category.

D. Feature Scaling

Height in centimetres, weight in kilogrammes, BMI as a ratio, and cholesterol as a lab value were among the numerical features in the dataset with different scales. This variation presented a challenge for models that rely on gradient descent (e.g., neural networks) and distance-based models (e.g., KNN), where features with larger scales can dominate the optimisation process.

We used Scikit-learn's StandardScaler to apply Standard Scaling (Z-score normalisation). Each numerical feature was transformed to have a mean of 0 and a standard deviation of 1 as a result. This enhanced convergence behaviour in models that depended on gradient optimisation and guaranteed that every feature made an equal contribution to the learning process.

Every preprocessing stage was deliberately chosen to solve particular problems in the dataset. The dataset was converted into a clean and balanced format ready for model training by means of focused imputation, suitable encoding techniques, and normalisation. These actions taken together guaranteed better model performance, less bias, and increased prediction generalisability.

IV. Dataset Splitting

The dataset was divided into a training set and a test set to assess the performance of several machine learning models. Models were trained on the training set, which made up 70 percent of the data; the other 30 percent was kept as a test set for performance assessment. A stratified split using train-test-split from Scikit-learn was applied given the class imbalance in the target variable Smoking-enc. Stratification guarantees that the training and test sets preserve the same class distribution as the original dataset. For classification tasks where one class greatly outnumbers the other, this is particularly crucial since it stops the model from being exposed to a skewed or deceptive class distribution during training or assessment. By using a stratified 70/30 split, the integrity of the dataset's structure was preserved, and models were tested under fair, real-world-like conditions.

V. Dataset Splitting

Several machine learning models have been developed and evaluated to evaluate their efficacy in forecasting smoking behaviour after preprocessing and dataset division. The target variable, Smoking-enc, is binary (0 for non-smoker, 1 for smoker), which makes this a classification problem. Each model used is briefly described here:

A. K-Nearest Neighbours (KNN)

A non-parametric model for categorisation. KNN forecasts a sample's class by looking at the majority class among its 'k' closest neighbours in the training set. All numerical inputs were standardised in advance because KNN is sensitive to feature scaling.

B. Decision Tree

A tree-based model called the Decision Tree Classifier divides the data into branches according to feature criteria. Robust to outliers, it manages both numerical and categorical features. It was applied in classification mode to predict smoking status.

C. Logistic Regression

A popular linear model for binary classification issues is logistic regression. It uses the logistic (sigmoid) function to model the likelihood that a given input is a member of class 1 (smoker). To rectify the unequal distribution of classes, class weighting was implemented.

D. Naive bayes Classifier

The Naive Bayes Classifier is a probabilistic model that assumes feature independence and is based on Bayes' Theorem. Although it is quick and effective with high-dimensional data, it might not perform as well if there are significant feature dependencies. It was tested using the encoded, standardized dataset.

E. Neural Network

RELU and sigmoid activations were used to create a feedforward neural network with a single hidden layer. The loss function that was employed was Binary Cross-Entropy. Although the model's performance was limited due to the class imbalance and low feature–target correlation, it was useful as a standard for assessing more intricate architectures.

To ensure a thorough comparison of predictive performance, each model was evaluated in the test set using metrics such as precision, F1 score, recall, and AUC. The models were trained using the balanced training set, with the option to use SMOTE for minority oversampling.

VI. Model Selection and Comparison Analysis

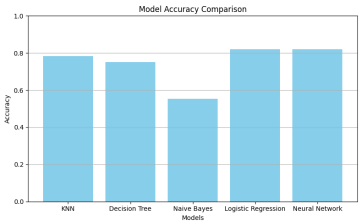

Fig. 5: chart showcasing prediction accuracy of all models

A variety of classification algorithms, including K-Nearest Neighbours (KNN), Decision Trees, Logistic Regression, Naive Bayes, and a specially designed Neural Network, were trained and assessed in order to determine which model would be best for forecasting smoking behaviour. A variety of metrics that emphasise various facets of prediction quality were used

to evaluate each model's performance, particularly in light of the dataset's imbalance. First, we used overall accuracy to compare models. Decision trees and KNN outperformed the others in terms of accuracy, while Naive Bayes and the Neural Network trailed behind, as shown in a bar chart. However, accuracy by itself was not enough to assess a model's efficacy because the dataset contains a significantly higher proportion of non-smokers than smokers. Even though a model that predicted everyone would not smoke would seem accurate, it would not be able to identify any smoker.

To learn more, we assessed F1-score, precision, and recall, especially for the minority class (1 = smokers). The precision indicates the proportion of predicted smokers who actually smoked.

The model's recall indicates the number of real smokers it was able to identify.

Models that perform well in one but poorly in the other are penalised by the F1-score, which strikes a balance between the two.

The Decision Tree model produced the best results for class 1 in this analysis. It was able to identify smokers and reduce misclassifications, as evidenced by its highest F1-score. Although it was a little less reliable, KNN also did fairly well. However, while they occasionally had respectable accuracy for class 0, Logistic Regression, Naive Bayes, and the Neural Network demonstrated extremely poor recall and F1-scores for class 1, frequently failing to predict any smokers at all. We created confusion matrices, which display the quantity of true positives, true negatives, false positives, and false negatives for every class, in order to better visualise the behaviour of the model. Several models predicted nearly exclusively non-smokers, as these matrices verified. Once more, the exceptions were KNN and Decision Tree, which demonstrated a more balanced distribution and accurately identified a smoker.

Lastly, we evaluated how well each model ranked positive instances over negative ones across various thresholds using the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) score. With an AUC of roughly 0.63, the Decision Tree once again demonstrated superior performance. When it came to ranking smokers versus non-smokers, other models such as Logistic Regression and Neural Network had AUCs near 0.50, meaning they were no more effective than random guessing.

TABLE III: Model Performance Comparison

Model	**Accuracy**	**F1 (Class 0)**	**F1 (Class 1)**	**Macro F1**	**AUC**
KNN	78.34	0.87	0.25	0.56	0.6
Decision Tree	75.1	0.85	0.36	0.6	0.61
Naive Bayes	55.34	0.68	0.25	0.47	0.50
Logistic Regression	82.04	0.90	0.00	0.45	0.50
Neural Network	82.04	0.90	0.00	0.45	0.50

VII. Conclusion

This project explored how well machine learning models can predict smoking behavior using basic biomedical data. Among the models tested, Decision Tree and KNN performed the best, showing some ability to detect smokers, while Logistic Regression, Naive Bayes, and the Neural Network struggled—often failing to identify any smokers at all. The main reason for these results is the high class imbalance in the dataset, with smokers making up only 18 percent of the data. Most models leaned heavily toward predicting non-smokers, hurting their performance on the minority class. Another key issue was that the features—like heart rate, cholesterol, BMI—had very little correlation with smoking, which makes sense since smoking is often linked to lifestyle and behavioral patterns not captured in this dataset. We faced challenges with balancing the dataset, choosing the right imputation strategies, and making sure preprocessing didn't cause data leakage. Even after applying class weighting and SMOTE, the results were limited due to weak feature-target relationships. In short, the project showed that data quality matters more than model complexity. Without strong, relevant features, even advanced models can't perform well. This highlights the need for richer, more behavior-focused data in future efforts to predict smoking behavior effectively.