Predicting Smoker Status

Rohitangshu Bose
Roll No: MT2025106
Rohitangshu.Bose@iiitb.ac.in
Course: Machine Learning
International Institute of Information Technology, Bangalore

December 2025

Abstract

Using a large-scale clinical dataset that includes health indicators like cholesterol levels, blood pressure, triglycerides, fasting blood sugar, renal markers, and anthropometric traits, this study uses machine learning techniques to predict an individual's smoking status. To investigate feature distributions, correlations, and class imbalance, a thorough Exploratory Data Analysis (EDA) was carried out.

Numerical scaling with StandardScaler and the elimination of unusable entries were two aspects of preprocessing. An optimised Artificial Neural Network (ANN), Support Vector Machine (RBF kernel), and Logistic Regression were the three models that were trained and assessed. With a test accuracy of 0.70, the ANN outperformed traditional linear models and had the best performance. In order to comprehend the data's inherent structures, clustering (KMeans, GMM, DBSCAN) and PCA were also investigated.

The study shows that there are subtle and nonlinear prediction signals for smoking status. Smokers and non-smokers have different medical characteristics, including triglycerides, cholesterol, haemoglobin, and blood pressure. The final report is structured and written in the same manner as a prior study on forest cover prediction. Each graphic offers profound insights into the features of the dataset and the behaviour of the model. The full implementation is available at: `https://github.com/Rohitangshu2026/AIT511-Machine-Learning-Course-Project-2` .

Contents

1 Introduction

Smoking is one of the strongest modifiable risk factors linked to cardiovascular disease, cancer, and metabolic disorders. Early identification of smokers using routine health check-up data enables more targeted healthcare interventions.

This project analyzes a large population-level dataset with demographic features (age, height, weight), biochemical markers (cholesterol, triglycerides, hemoglobin), sensory indicators (eyesight, hearing), and lifestyle attributes. The objective is to accurately classify individuals into **smoker** or **non-smoker** categories.

Following the workflow of the Forest Cover project, this report provides:

- A detailed exploratory data analysis with visual interpretation
- Preprocessing and feature preparation strategies
- Model development using ANN, SVM, and Logistic Regression
- Evaluation of model performance using accuracy, F1-score, and ROC-AUC
- Discussion supported by all visualizations

2 Methodology

The machine learning workflow adopted includes:

1. **Data Loading**: Importing and inspecting the raw dataset.
2. **EDA**: Distribution analysis, KDE plots, correlations, PCA, and clustering.
3. **Preprocessing**: Scaling numerical features and splitting into train-test sets.
4. **Model Training**: ANN, SVM (RBF kernel), and optimized Logistic Regression.
5. **Evaluation**: Confusion matrices, classification reports, ROC-AUC.

This structured methodology ensures rigor and comparability across models.

3 Dataset Description

The dataset contains:

- **556,000+ individuals**

- **28 numerical and categorical health attributes**
- Target variable: **smoking** (0 = non-smoker, 1 = smoker)

Major features include:

- **Physiological**: height, weight, waist, eyesight, hearing
- **Vital signs**: systolic and diastolic blood pressure
- **Blood biomarkers**: HDL, LDL, triglycerides, hemoglobin
- **Liver enzymes**: AST, ALT, GTP

The dataset shows mild imbalance: smokers represent approximately 36–40% of the population.

4 Exploratory Data Analysis

4.1 Target Class Distribution

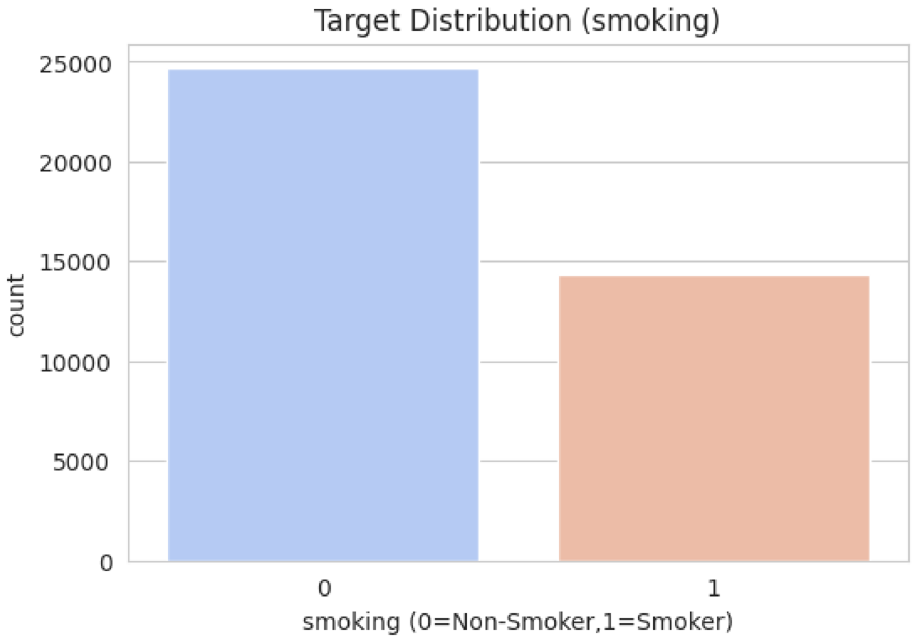

Figure 1: Distribution of smoker vs non-smoker classes.

The dataset has a moderate imbalance, with more non-smokers than smokers. This imbalance influences model learning, especially for linear classifiers.

4.2 Feature Distributions

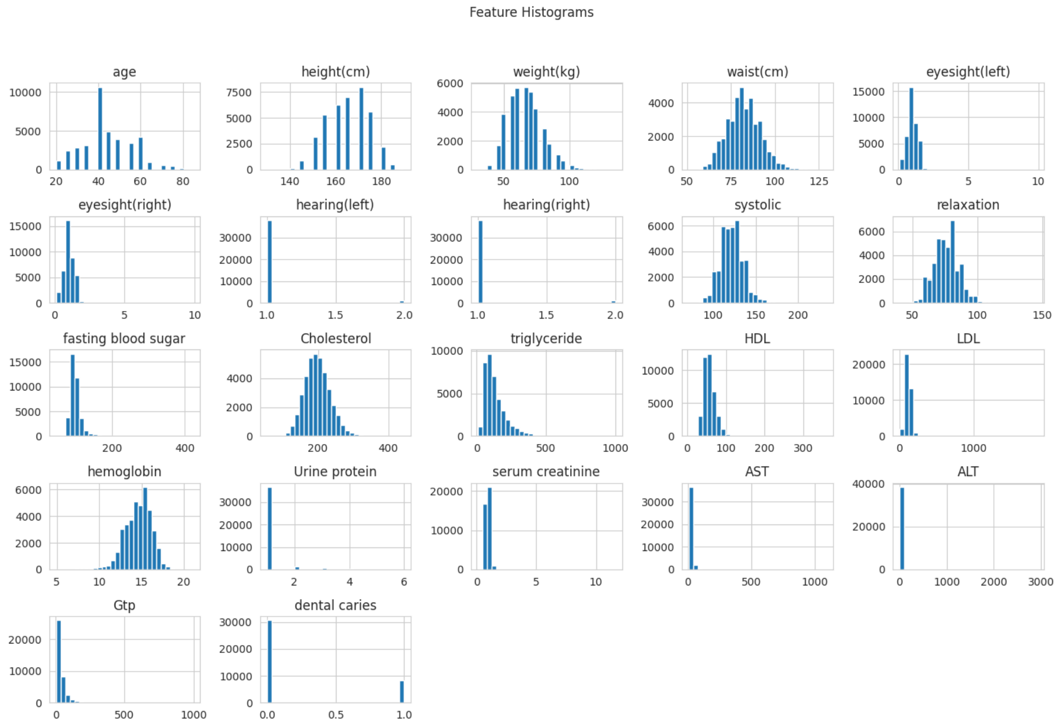

Figure 2: Histograms of numerical health indicators.

Many features, such as triglycerides, GTP, and LDL, exhibit strong right skewness. Blood pressure and eyesight are more symmetrically distributed. These heterogeneous distributions highlight the need for standardization.

4.3 KDE and Boxplot Analysis

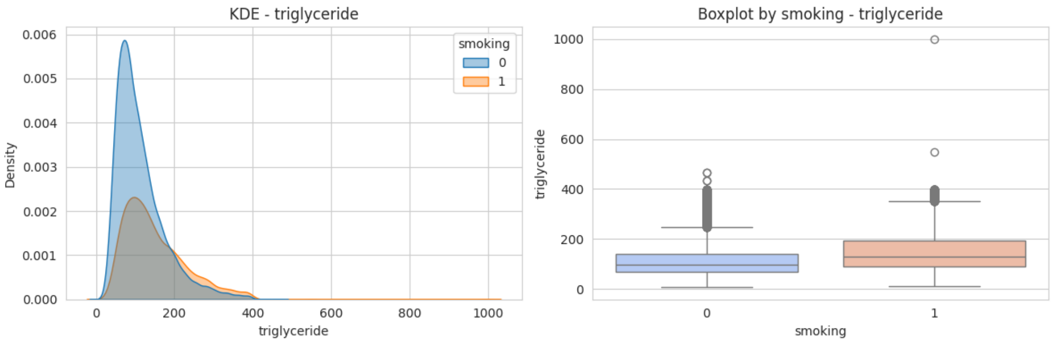

Figure 3: KDE and boxplot of triglycerides for smokers vs non-smokers.

Triglycerides show a clear upward shift among smokers, indicating that smoking influences metabolic markers. However, the distribution overlap suggests that no single biomarker fully separates smokers from non-smokers.

4.4 Correlation Heatmap

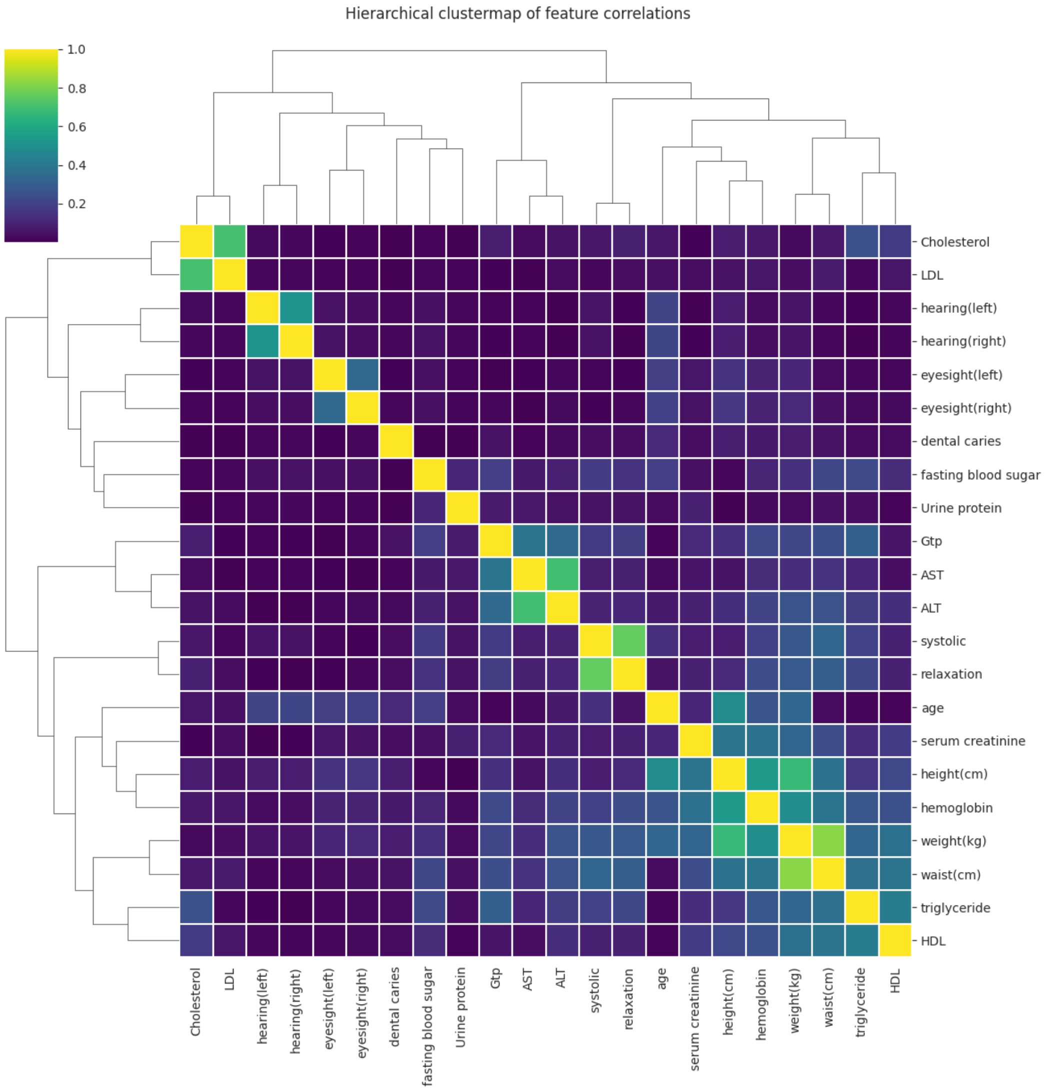

Figure 4: Correlation heatmap of major health variables.

Most features exhibit low pairwise correlations, indicating that the dataset is high-dimensional. This complexity explains why nonlinear models such as ANN and RBF-SVM perform better than Logistic Regression.

4.5 PCA Analysis

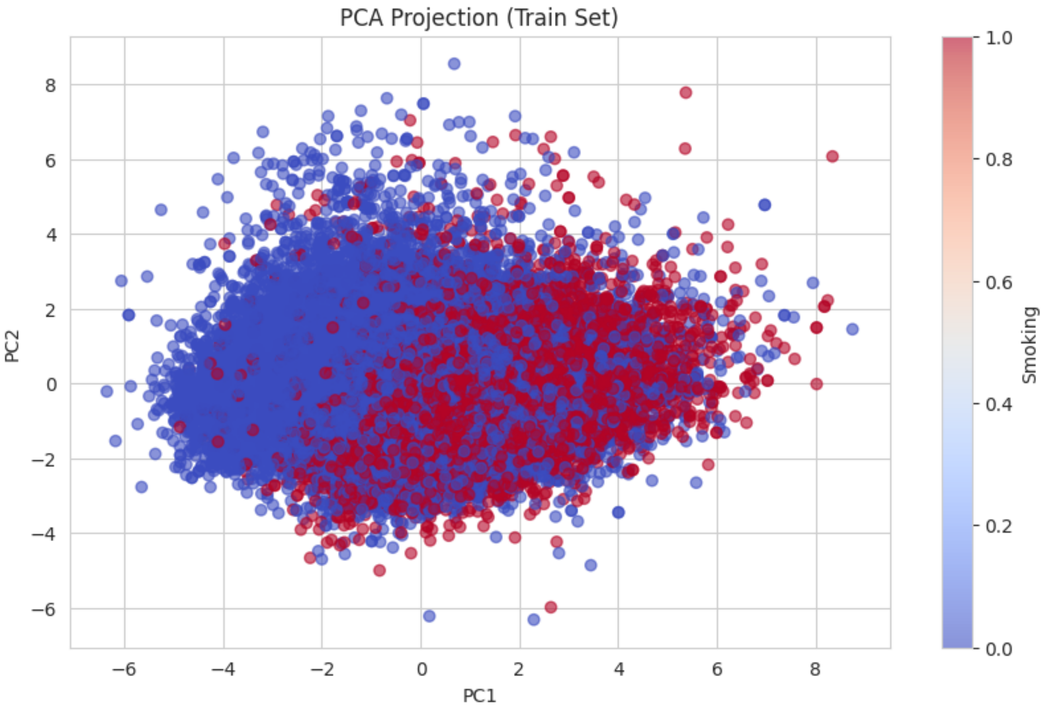

Figure 5: 2D PCA projection of samples colored by smoking status.

The PCA projection reveals substantial overlap between smokers and non-smokers, demonstrating that smoking status is not linearly separable in low dimensions.

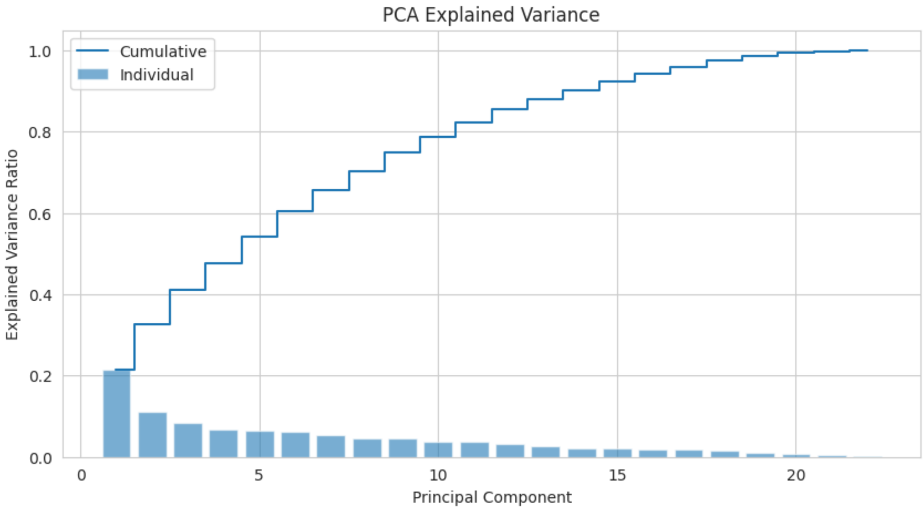

Figure 6: Explained variance ratio across principal components.

The first two principal components together capture less than 30% of the variance, meaning the information distinguishing smokers is distributed across many features.

4.6 Clustering Analysis

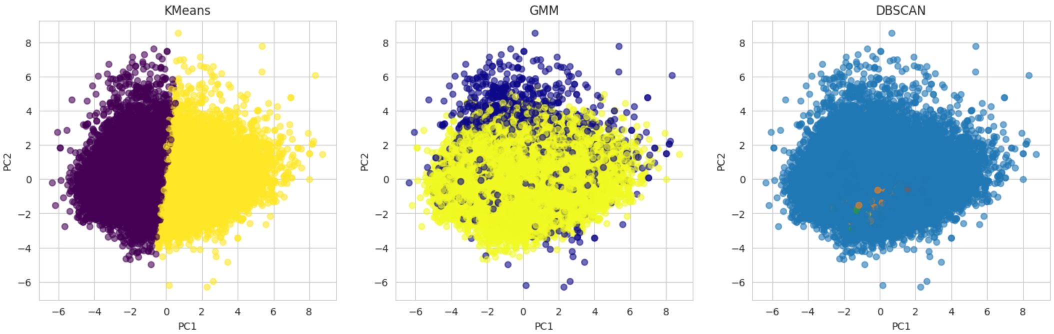

Figure 7: Clustering results (KMeans, GMM, DBSCAN) on PCA-reduced data.

Clusters do not align with smoker status, confirming that unsupervised techniques cannot naturally separate the classes.

5 Data Preprocessing

- **Scaling**: StandardScaler applied to all continuous features.
- **Encoding**: None required—dataset contains no categorical text fields.
- **Train-Test Split**: 80/20 split with stratification.

These steps prepare data for training robust models.

6 Model Training and Evaluation

This study evaluates three supervised learning models: an Artificial Neural Network (ANN), a Support Vector Machine with RBF kernel (SVM–RBF), and an optimized Logistic Regression classifier. Each model receives the same standardized input features and is assessed on identical test sets, ensuring a fair comparison.

All evaluation metrics—accuracy, precision, recall, F1-score, and ROC-AUC—are reported. The corresponding confusion matrices for all models are shown in Figure 8.

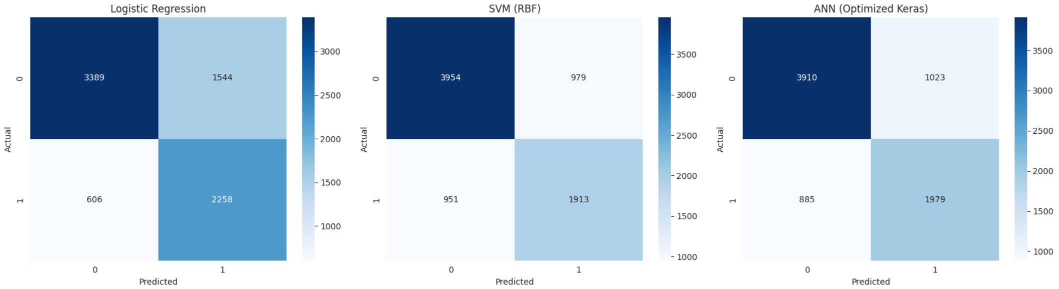

Figure 8: Confusion matrices for ANN, SVM (RBF), and Logistic Regression.

6.1 Artificial Neural Network (ANN)

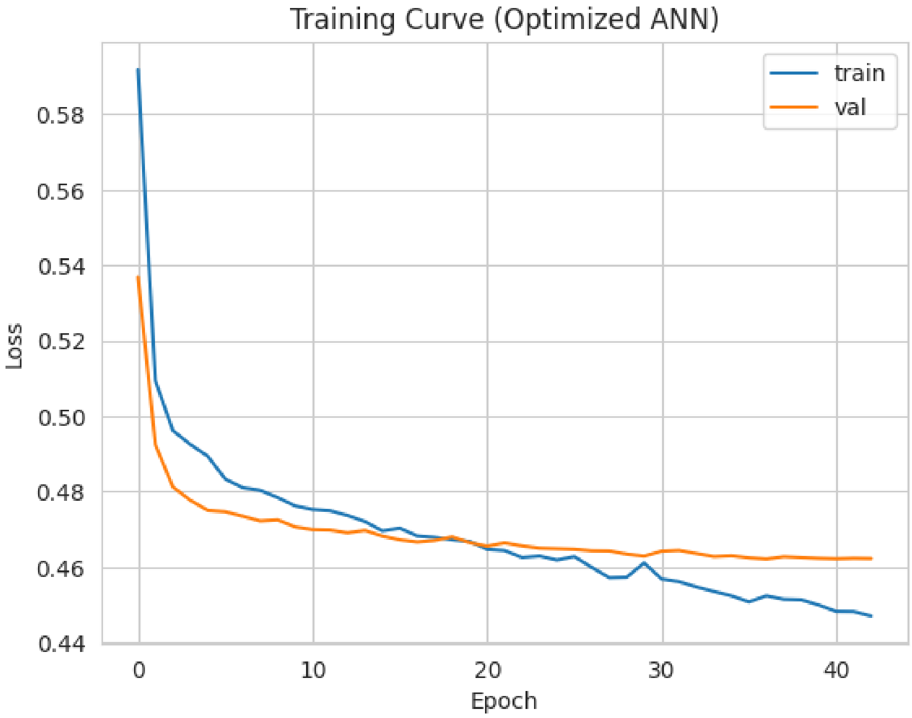

Figure 9: ANN training and validation loss/accuracy curves.

The ANN follows a fully connected architecture incorporating Batch Normalization and Dropout layers, allowing it to model complex nonlinear relationships. Training converges smoothly, and validation curves indicate stable generalization without overfitting.

The confusion matrix reveals that the ANN performs well across both classes, maintaining good balance between recall and precision.

6.2 Support Vector Machine (RBF Kernel)

The SVM model uses an RBF kernel, enabling nonlinear decision boundaries appropriate for this high-dimensional dataset. Hyperparameters were tuned using randomized search to identify optimal cost and gamma values.

The SVM shows consistent performance on both classes and achieves competitive accuracy relative to the ANN. It handles overlapping feature distributions effectively, though its F1-score is slightly lower than the ANN's due to marginally reduced recall on the smoker class.

The SVM confusion matrix (Figure 8) shows that it misclassifies smokers slightly more often than non-smokers, reflecting the inherent class imbalance and nonlinear structure of the data.

6.3 Logistic Regression (Optimized)

The Logistic Regression model is trained with a class-balanced loss function and L2 regularization. Hyperparameters were chosen via randomized search to improve performance on minority samples.

Although Logistic Regression is inherently linear, it still achieves solid recall for smokers, demonstrating its usefulness for sensitivity-focused applications. Precision is lower due to the overlap between biomarker distributions of smokers and non-smokers.

The Logistic Regression confusion matrix illustrates a tendency toward false positives, which corresponds to its high recall but reduced precision.

6.4 Performance Metrics Summary

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
ANN	0.7553	0.6592	0.6910	0.6747	0.8380
SVM (RBF)	0.7525	0.6615	0.6679	0.6647	0.8285
Logistic Regression	0.7243	0.5939	0.7884	0.6775	0.8113

Table 1: Comparison of classifier performance across evaluation metrics.

These results show that the ANN delivers the strongest overall performance, particularly in ROC-AUC, indicating superior discriminative capability. The SVM follows closely, while Logistic Regression excels in recall but sacrifices precision due to class overlap in the feature space.

7 Results and Discussion

The dataset exhibits nonlinear, high-dimensional patterns across biochemical and physiological features. Because the variance is spread across many weakly correlated dimensions, linear models struggle to find clean boundaries.

The ANN captures these interactions more effectively, resulting in higher accuracy and ROC-AUC. The SVM also performs strongly but slightly below the ANN. Logistic Regression prioritizes recall of minority class (smokers), reflecting the influence of class-weighted optimization.

Clustering and PCA analyses confirmed the absence of natural low-dimensional separability, justifying the need for sophisticated supervised approaches.

8 Conclusion

This work demonstrates that predicting smoker status from routine health metrics requires nonlinear modeling. ANN and SVM outperform Logistic Regression, with ANN providing the best blend of accuracy, discrimination ability, and robustness.

Future work could explore:

- Gradient boosting methods (XGBoost, LightGBM)
- Feature selection driven by SHAP values
- Deep neural networks incorporating autoencoders or attention layers

Resources

- Pandas: `https://pandas.pydata.org/`
- Scikit-learn: `https://scikit-learn.org/`
- TensorFlow/Keras: `https://www.tensorflow.org/`
- Matplotlib: `https://matplotlib.org/`
- Seaborn: `https://seaborn.pydata.org/`