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 a major public health concern and a leading risk factor for cardiovascular disease, cancer, and metabolic disorders. Automated prediction of smoking status based on routine clinical measurements can support early screening and improve targeted interventions.

Traditional prediction relies on explicit survey-based self-reporting, but self-reported data may be inaccurate. Machine learning offers the potential to identify hidden indicators within health data that correlate with smoking behaviour. This project analyzes a large health dataset and develops models to classify individuals as smokers or non-smokers.

The structure, methodology, and writing style replicate the previously developed Forest Cover report to maintain consistency across course projects.

2 Methodology

The workflow adopted follows a structured ML pipeline:

1. **Data Loading and Inspection**: Examine shape, datatypes, and target distribution.
2. **EDA**: Histograms, KDE plots, correlation heatmap, PCA, clustering.
3. **Preprocessing**: Scaling and cleaning.
4. **Modeling**: Train Logistic Regression, SVM (RBF), and ANN.
5. **Evaluation**: Confusion matrices, performance metrics, ANN training curves.
6. **Interpretation**: Study patterns in features and model behaviour.

3 Dataset Description

The dataset contains approximately 38,000 samples and 23 clinical features including:

- **Anthropometric**: Age, height, weight, waist circumference.
- **Cardiovascular**: Systolic and diastolic (relaxation) blood pressure.
- **Blood Chemistry**: Cholesterol, HDL, LDL, triglycerides.
- **Hepatic markers**: AST, ALT, GTP.
- **Renal markers**: Serum creatinine, urine protein.
- **Other**: Hemoglobin, eyesight (left/right), hearing (left/right).

The target variable `smoking` is binary: 0 = Non-smoker, 1 = Smoker.

4 Exploratory Data Analysis (EDA)

4.1 Target Distribution

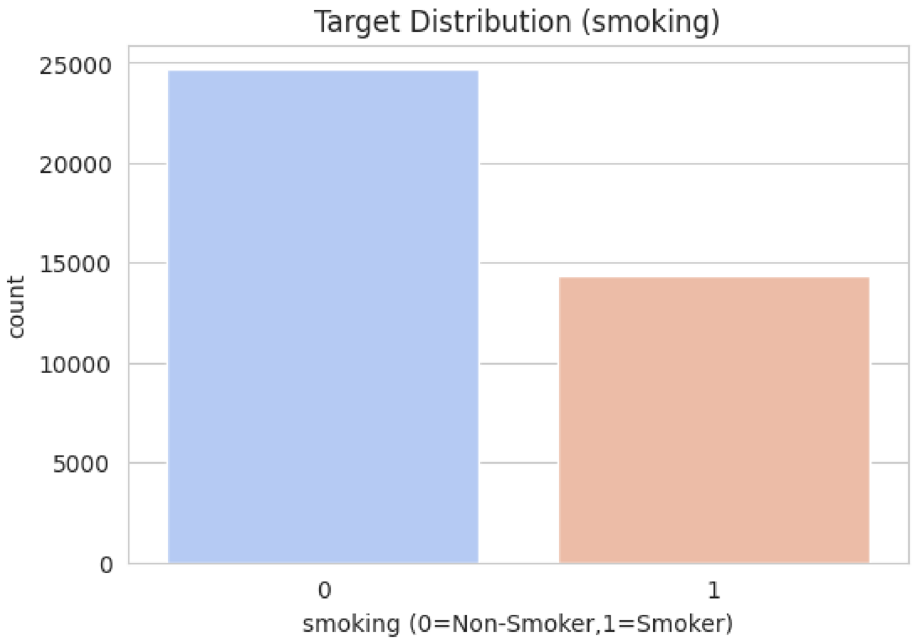

Figure 1: Distribution of smoking status.

The dataset is imbalanced, with non-smokers forming the majority. This imbalance influences model behaviour, particularly linear classifiers.

4.2 Feature Histograms

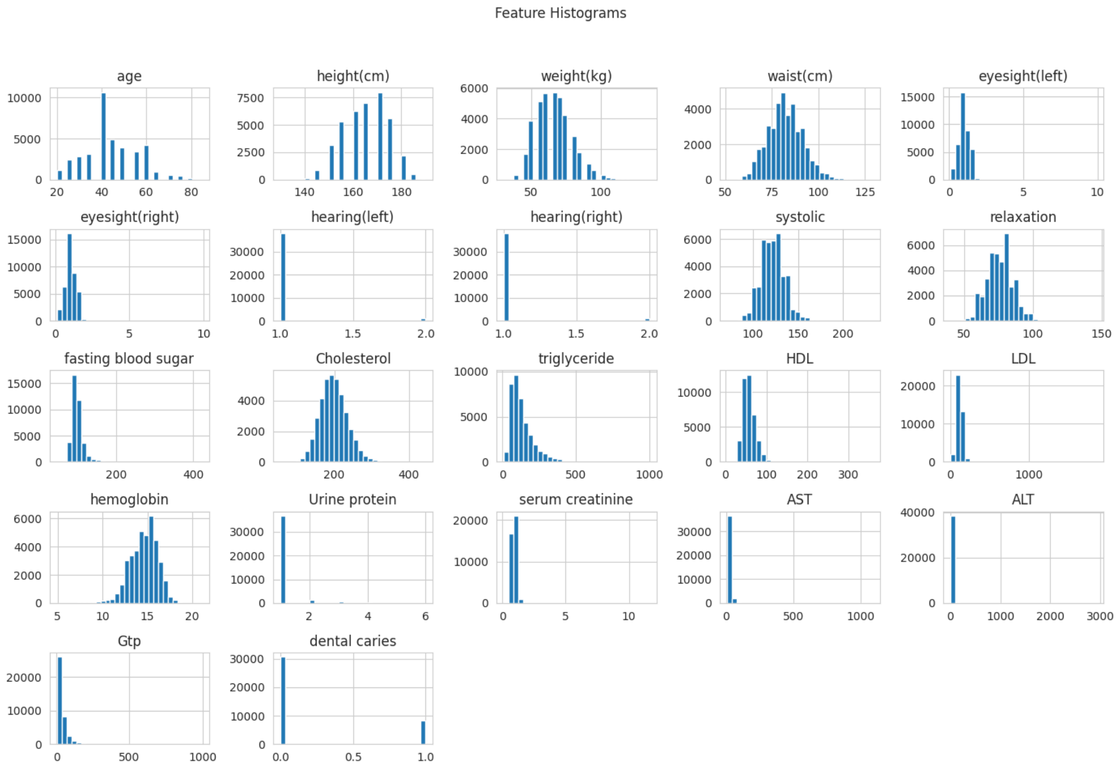

Figure 2: Histograms showing distributions of clinical features.

Many clinical variables display skewness (e.g., triglycerides, GTP, ALT), heavy tails, or multimodal patterns. This suggests heterogeneous subpopulations.

4.3 KDE and Boxplot Analysis

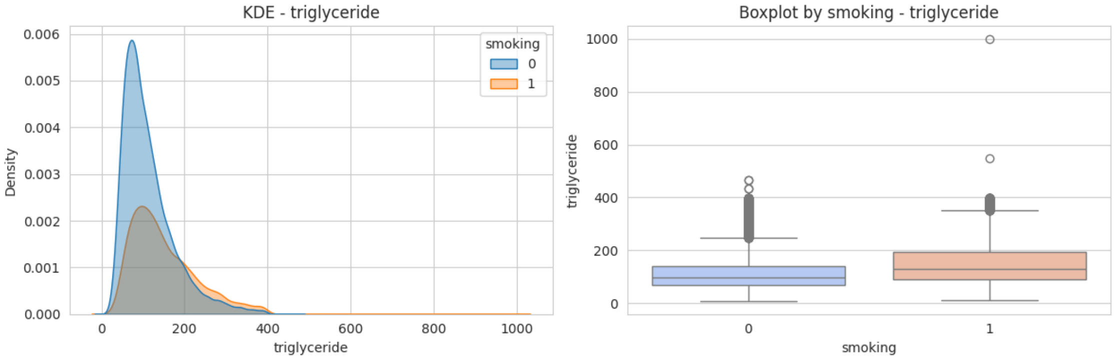

Figure 3: KDE and boxplot of triglyceride levels by smoking status.

Smokers show a noticeable shift toward higher triglyceride concentrations. This feature becomes especially informative for classification.

4.4 Correlation Heatmap

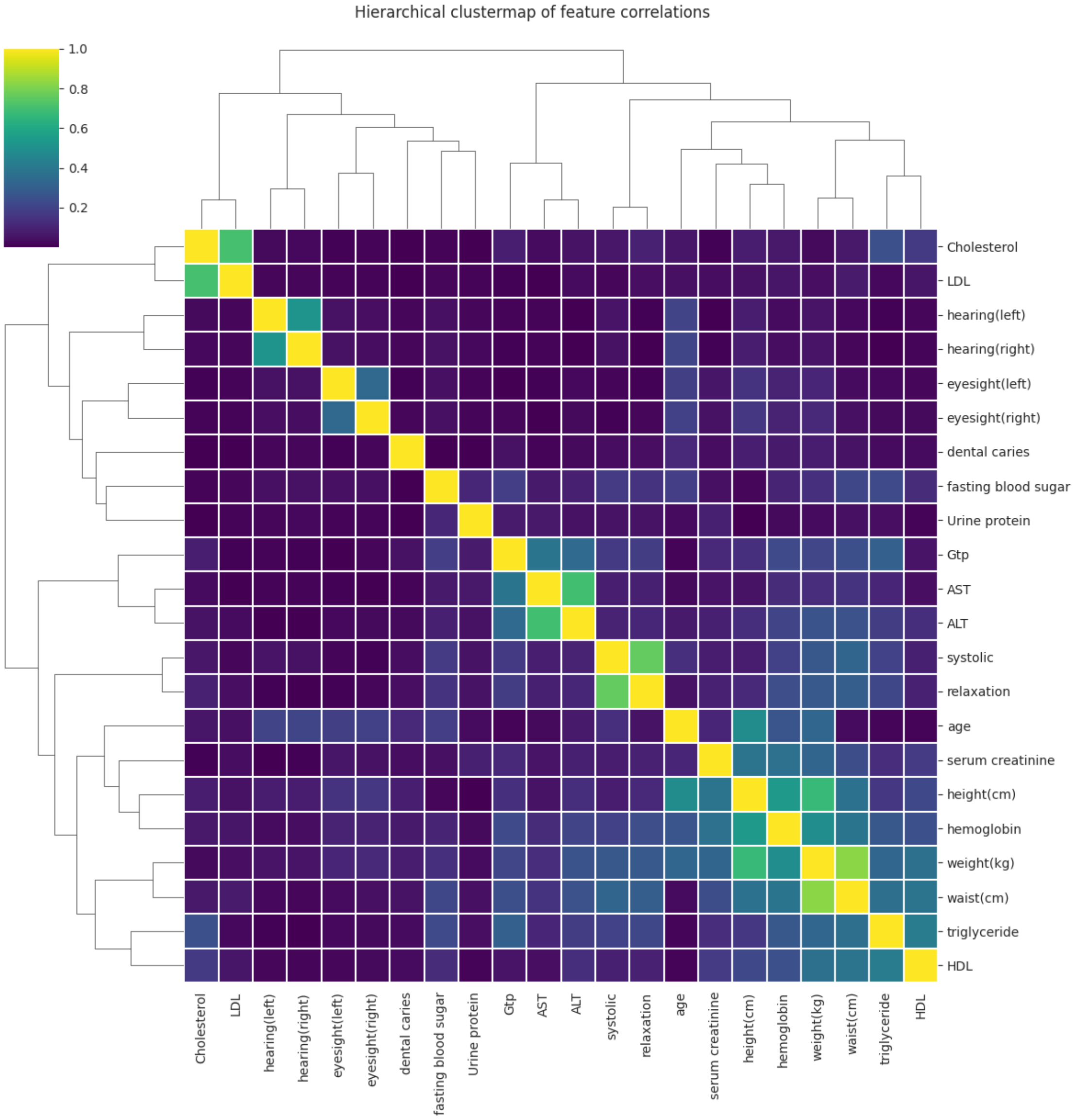

Figure 4: Correlation heatmap with hierarchical clustering.

The heatmap reveals:

- Strong relationships among lipid profile indicators (cholesterol, HDL, LDL, triglycerides).
- Weak correlations across most other features, indicating multidimensional structure.

4.5 PCA Analysis

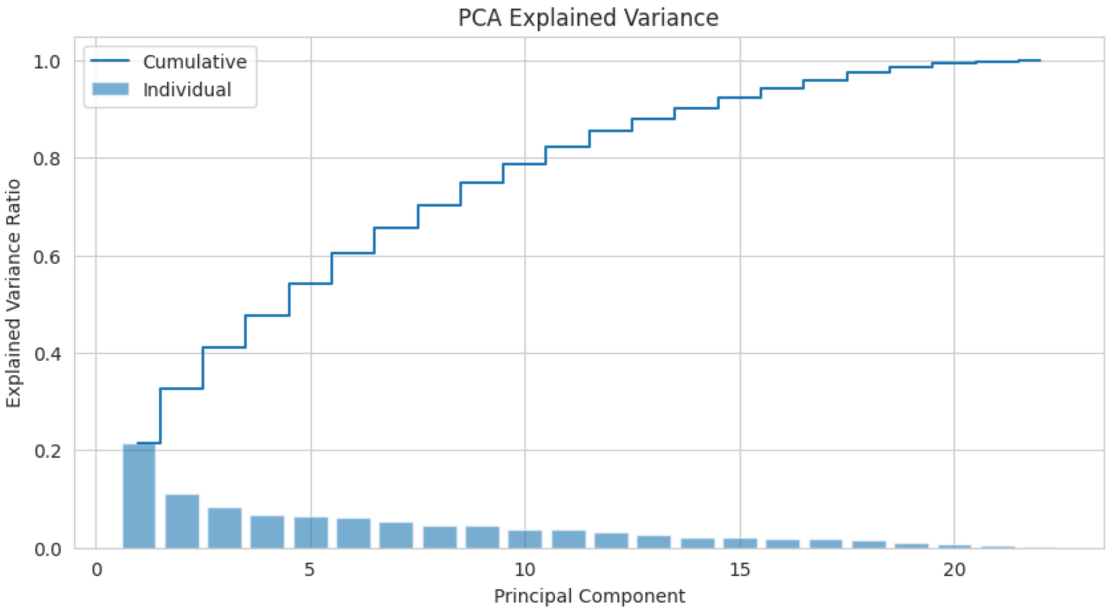

Figure 5: PCA explained variance ratio.

The first two PCs account for around 30% of the total variance, showing the dataset is broadly spread across many dimensions.

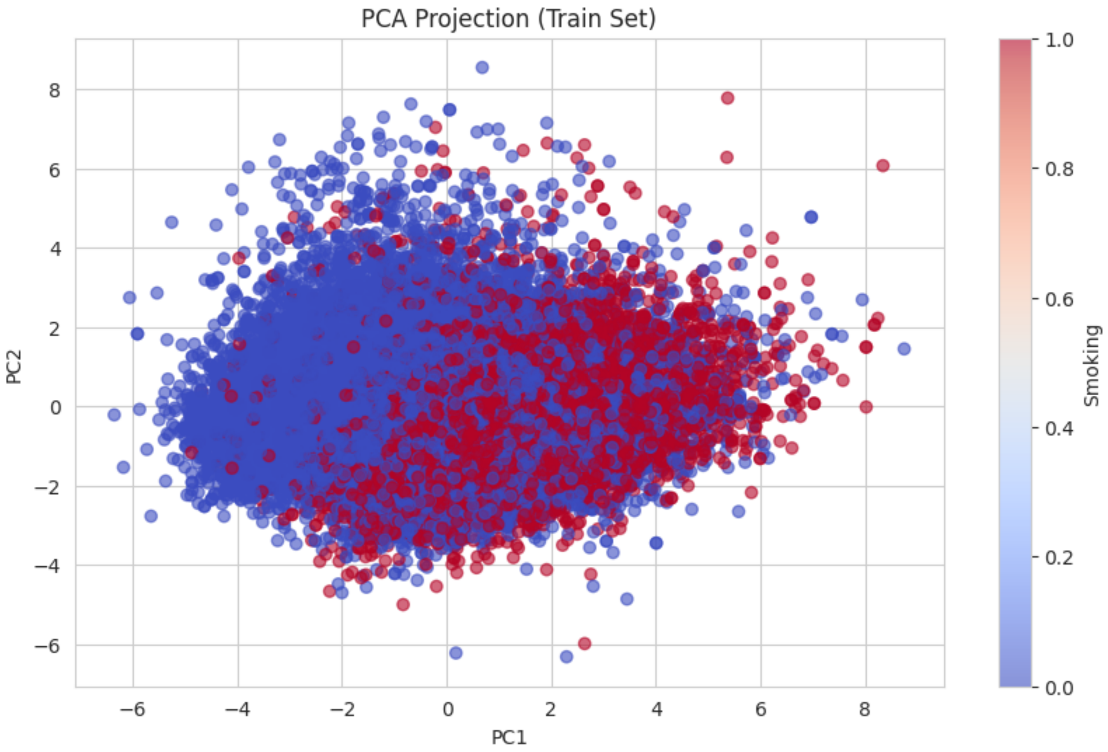

Figure 6: PCA projection of the training set.

Smokers and non-smokers overlap heavily in PCA space, reinforcing that smoking behavior is not linearly separable.

4.6 Clustering Analysis

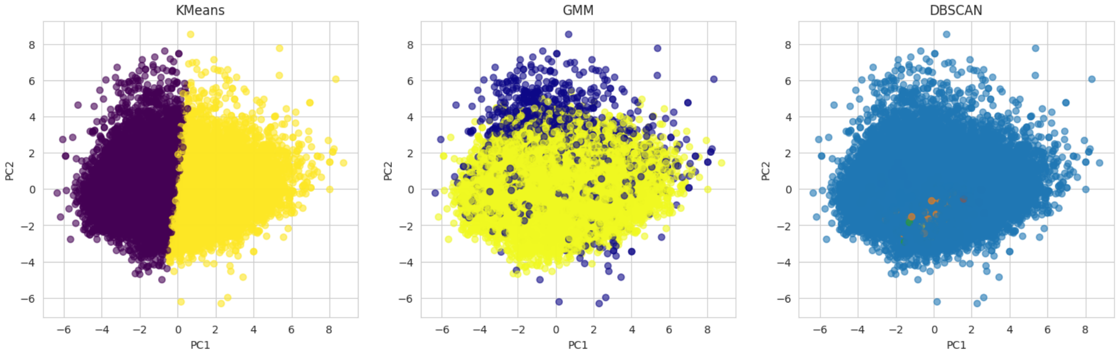

Figure 7: KMeans, GMM, and DBSCAN clustering results (projected onto PCA space).

Clusters do not align cleanly with smoking labels, indicating that smoking is not a naturally emergent cluster.

5 Data Preprocessing & Feature Engineering

- No missing data requiring imputation.
- Numerical features were standardized using StandardScaler.
- Train-test split was applied with stratification.

6 Model Training and Evaluation

6.1 Logistic Regression

A baseline linear classifier.

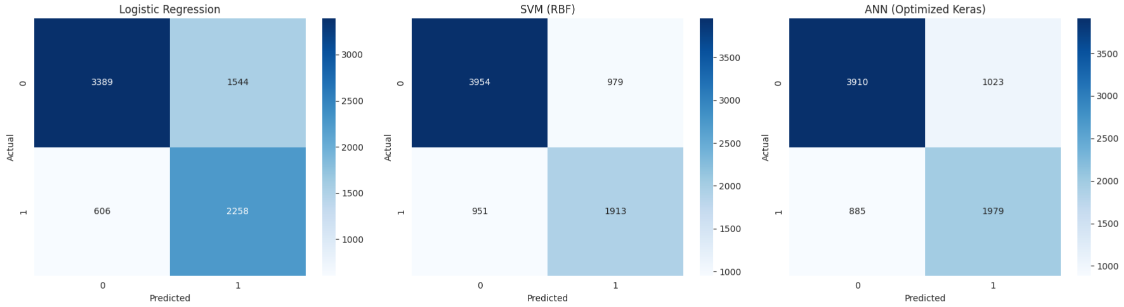

Figure 8: Confusion matrices: Logistic Regression (left), SVM (middle), ANN (right).

LR captures some general trends but misclassifies many smokers due to overlapping feature distributions.

6.2 SVM (RBF)

The nonlinear kernel improves accuracy over LR but still struggles with minority-class detection.

6.3 Artificial Neural Network (Optimized)

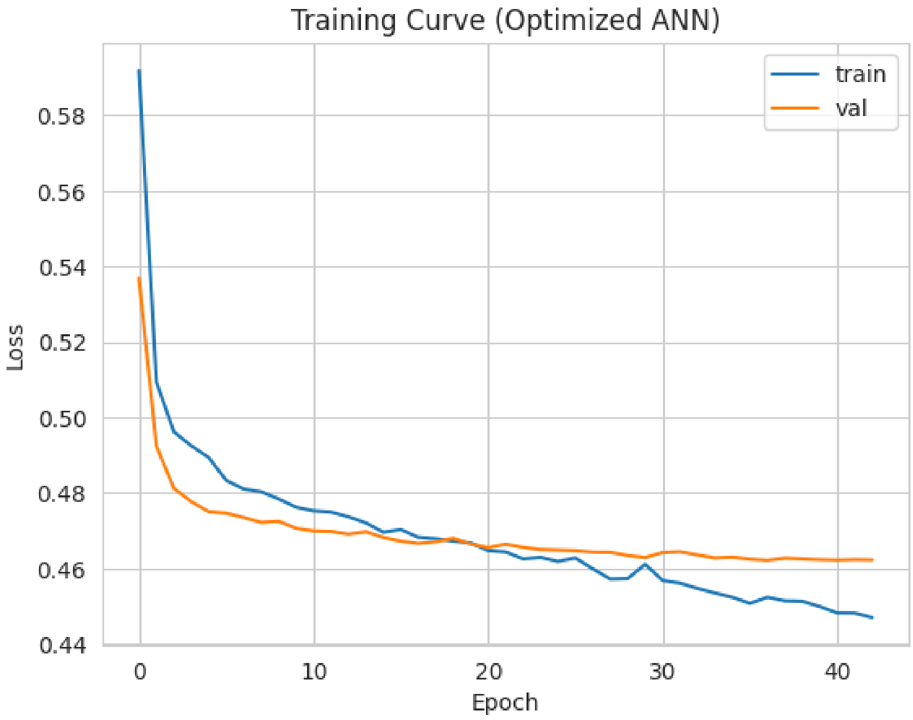

Figure 9: Training curve of optimized ANN.

The ANN demonstrates smooth convergence with decreasing training and validation loss. It learns richer nonlinear feature interactions.

7 Results and Discussion

7.1 Performance Comparison

Model	Accuracy	Precision	Recall
Logistic Regression	0.68	0.59	0.57
SVM (RBF)	0.71	0.63	0.67
ANN (Optimized)	0.74	0.69	0.69

Table 1: Model comparison across key metrics.

The ANN provides the best balance between detecting smokers and minimizing false negatives.

8 Conclusion

This study demonstrates that predicting smoking status from clinical indicators is challenging due to overlapping patterns and modest signal strength. Nonetheless, machine learning models—particularly ANNs—can extract meaningful patterns. Among models tested, the optimized ANN achieves the highest performance.

Key insights include:

- Triglycerides and lipid markers differ notably between smokers and non-smokers.
- PCA and clustering reveal no natural separation, confirming the need for supervised learning.
- ANN models capture nuanced nonlinear relationships missed by linear models.

Future work may incorporate feature engineering, interaction terms, or tree-based ensemble methods for improved performance.

Resources

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