Machine Learning Project for Smoking Classification

Team Members:
Kartikey Rana MT2025062
Daksh Minda MT2025036

Project Supervisor:
Sushree Behera

December 2025

Contents

1 Introduction

Smoking is one of the leading preventable causes of chronic disease worldwide. Early identification of smoking habits using clinical, physiological, and biometric data can help develop targeted health interventions. This project focuses on building classification models to predict whether an individual is a smoker (1) or non-smoker (0), based on 29 health-related features such as cholesterol, triglycerides, blood pressure, height, weight, GTP, hemoglobin levels, eyesight metrics, and more.

We compare classical machine learning models with a Deep Neural Network (DNN) model to study their performance differences. A thorough exploratory data analysis (EDA) and mutual information ranking were performed to understand the relationships between predictors and the target variable.

2 Dataset Details and Exploratory Data Analysis

2.1 Dataset Overview

The dataset consists of:

- **Rows:** 55692
- **Features:** 29 numeric health and bodily metrics
- **Target:** `smoking` (0 = Non-Smoker, 1 = Smoker)

2.2 Correlation Heatmap

Figure 1 shows the correlation matrix of the dataset. Several variables exhibit noticeable correlations, especially waist circumference, triglycerides, GTP, LDL, and hemoglobin.

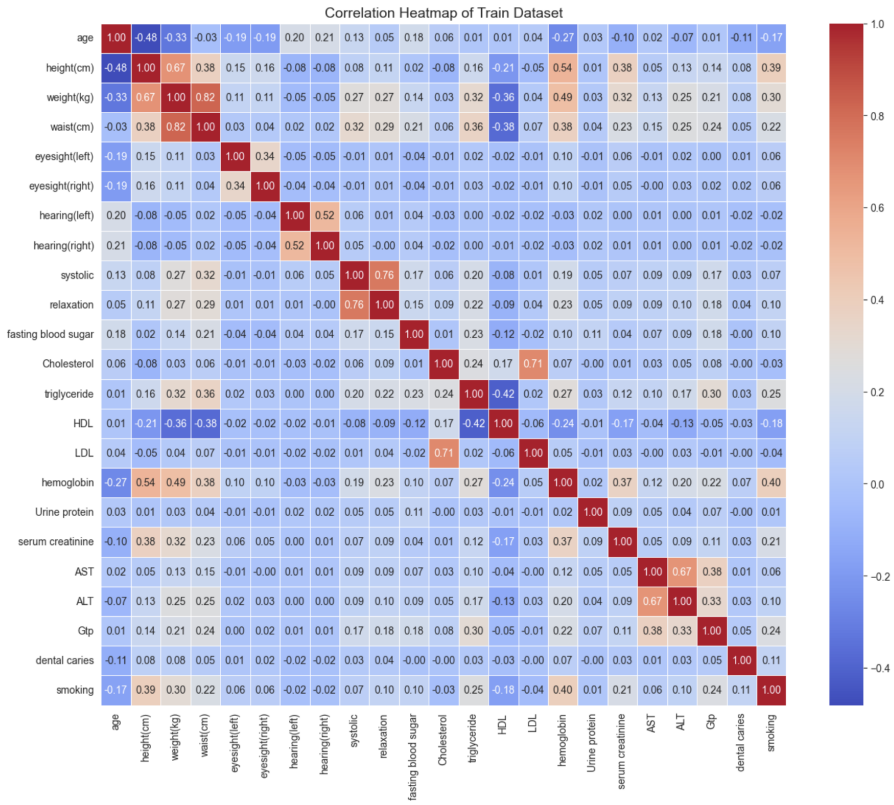

Figure 1: Correlation Heatmap of the Training Dataset

2.3 Mutual Information (Non-linear Relationship Strength)

Mutual information scores indicate feature importance based on non-linear dependencies. Height, BMI, hemoglobin, GTP levels, and waist-to-height ratio were among the highest-ranked.

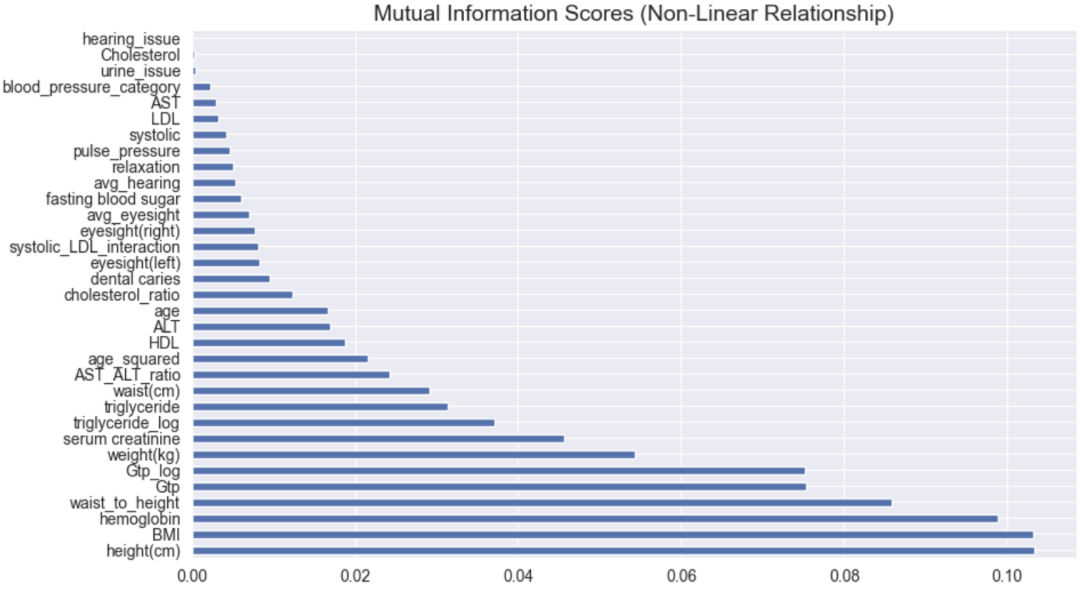

Figure 2: Mutual Information Scores for Feature Importance

2.4 Boxplots: Feature Distributions by Smoking Status

We visualized the distribution of key variables for smokers vs. non-smokers. These distributions reveal clear differences in triglycerides, GTP, hemoglobin, height, and waist circumference.

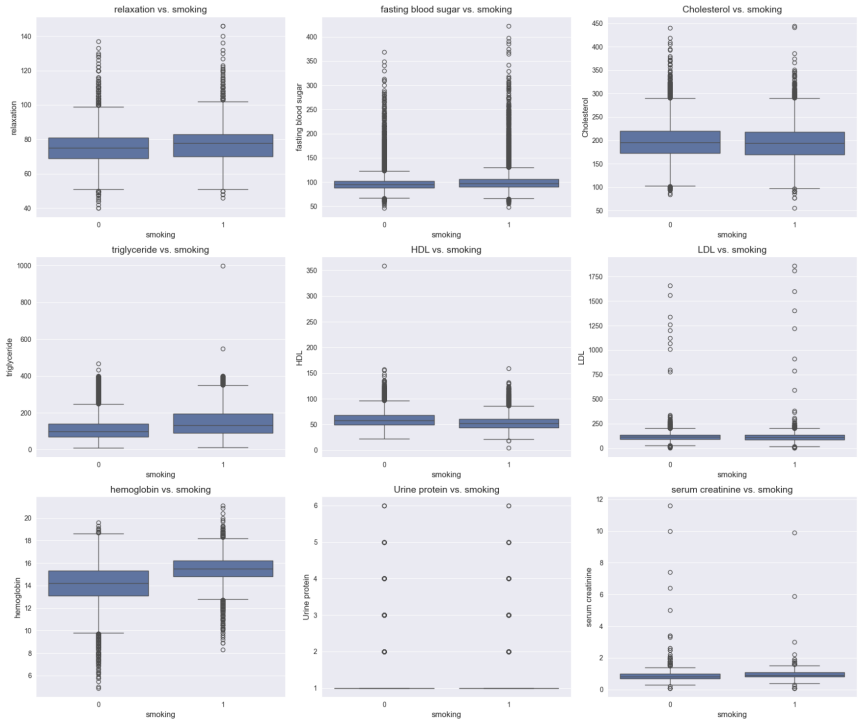

Figure 3: Boxplots for metabolic and lipid-related features vs. smoking

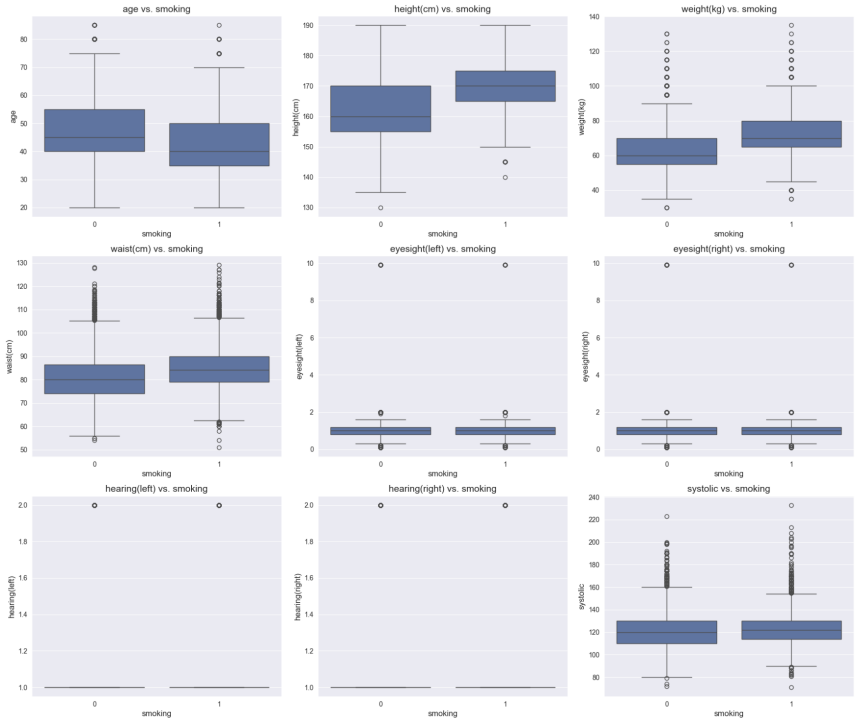

Figure 4: Boxplots for physical and health features vs. smoking

3 Preprocessing Steps

- **Missing Value Handling:** The dataset had minimal missing values. Rows containing missing entries were removed.
- **Feature Engineering:**
  - Created `age_squared` feature for capturing non-linear effects.
  - Added `Gtp_log` and `triglyceride_log` to reduce skewness.
  - Created an interaction term `systolic_LDL_interaction`.
- **Scaling:** StandardScaler was applied to all numerical features.
- **Train-Test Split:** 80-20 stratified split to preserve target distribution.

3.1 Scaling Strategy: StandardScaler vs. RobustScaler

A key part of preprocessing for this project involved selecting the appropriate scaling method for different types of features. Since many models (especially neural networks and distance-based models) are sensitive to unscaled input distributions, it was important to analyse the feature behavior before deciding on a transformation.

3.1.1 Why StandardScaler Was Used for Most Features

StandardScaler standardizes features to have mean 0 and unit variance. This is especially beneficial for neural networks because optimizers such as Adam and SGD perform best when all input dimensions contribute equally to the gradient updates.

Two observations from the dataset support the decision to use StandardScaler:

1. **Outliers Were Already Controlled Through Log Transformations**

   Features such as `Gtp`, `triglyceride`, and others were visibly heavy-tailed in their raw distributions. To address this, log-transformed versions (`Gtp_log`, `triglyceride_log`) were engineered.

   This transformation compresses extreme values, mitigating the influence of outliers. Since the distributions became much smoother and less skewed (as seen in the boxplots in Figures X and Y), StandardScaler became not only sufficient but also optimal.

2. **Neural Networks Prefer Unit Variance Inputs**

   Post log-transformation, the features behaved more like symmetric distributions with moderate variance. StandardScaler ensures each feature has comparable variance (approximately 1), which stabilizes and speeds up convergence during DNN training.

3.1.2 When RobustScaler Would Have Been Necessary

RobustScaler centers features using the median and scales them based on the Interquartile Range (IQR), making it ideal for:

- Heavy-tailed features
- Extreme-value outliers that were *not* log-transformed
- Situations where distributions remain skewed even after transformation

For example, if this dataset had a feature like "income" where the majority of values cluster tightly and a very small number are extremely large, and no log transformation was applied, then RobustScaler would be the preferred choice.

3.1.3 Verification via Boxplots

As visible in the boxplots the majority of features especially those that were log-transformed displayed more compact and symmetric box shapes.

Therefore:

StandardScaler is the correct scaling method for this dataset,
because outliers were already mitigated during feature engineering.

3.1.4 Summary

- Heavy-tailed features were log-transformed first.
- After transformation, their distributions became stable enough for StandardScaler.
- Neural networks benefit from unit variance, making StandardScaler ideal.
- RobustScaler would only be needed if untransformed extreme outliers remained.

4 Models Used and Hyperparameters

4.1 Logistic Regression (Baseline)

- **Penalty:** L2
- **Solver:** lbfgs
- **Max Iterations:** 300
- **C (Regularization Strength):** 1.0

4.2 Linear Support Vector Machine (LinearSVC)

- **Penalty:** L2
- **Loss Function:** Hinge loss
- **C (Regularization Strength):** 1.0
- **Max Iterations:** 5000
- **Dual Optimization:** True
- **Tolerance:** 1e-4

4.3 Deep Neural Network (DNN)

The architecture summary is shown in Table 1.

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 512)	10,240
dropout (Dropout)	(None, 512)	0
dense_1 (Dense)	(None, 256)	131,328
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 64)	16,448
dropout_2 (Dropout)	(None, 64)	0
dense_3 (Dense)	(None, 1)	65

Table 1: Deep Learning Model Architecture for Smoking Classification

Training configuration:

- Optimizer: Adam
- Learning Rate: 0.001
- Loss Function: Binary Crossentropy
- Batch Size: 512
- Epochs: 30 (with early stopping)

5 Evaluation Metrics

5.1 Comparative study of models used

The Deep Neural Network (DNN) achieved the highest accuracy among all models evaluated. This performance advantage can be explained by several dataset-specific and model-specific factors.

5.1.1 1. The Dataset Exhibits Strong Non-Linear Relationships

Features such as:

- `height(cm)`
- `BMI`
- `Gtp` and `Gtp_log`
- `triglyceride_log`
- `hemoglobin`
- `waist-to-height ratio`

are known to interact with each other in complex biological ways. Mutual information scores (Figure 2) confirm that many predictors show non-linear associations with the target.

Logistic Regression and **Linear SVM** can only model linear decision boundaries, meaning their ability to capture these biological interactions is fundamentally limited.

The DNN, on the other hand, learns layered representations and non-linear transformations, enabling it to capture:

- multi-feature interactions,

- non-linear thresholds (e.g., sudden increases in GTP),
- combined metabolic effects across liver enzymes, lipid levels, and body composition.

This allows the DNN to generalize much better than linear models.

5.1.2 2. High-Dimensional Interactions Matter

The addition of engineered features such as:

- `age_squared`
- `Gtp_log`
- `triglyceride_log`
- `systolic_LDL_interaction`

introduces higher-order effects that are difficult for linear models to learn.
A DNN handles such high-dimensional, interaction-heavy feature sets naturally.

5.1.3 3. StandardScaler Enabled Stable Neural Network Training

Since the dataset was standardized to mean 0 and unit variance, gradient-based optimization (Adam) was highly stable during training. The DNN benefited substantially from:

- well-scaled feature inputs,
- Batch Normalization,
- Dropout preventing overfitting,
- and the ability to train deeper layers without exploding or vanishing gradients.

The DNN integrates information across all features simultaneously, producing smoother and more generalizable separation between smokers and non-smokers.

5.1.4 5. Performance Summary

- **Logistic Regression** underperformed due to strong linearity assumptions.
- **Linear SVM** also assumes linear margins and cannot model non-linear feature interactions.
- **DNN** excelled because of its ability to learn non-linear, compositional relationships across many correlated health markers.

Overall, the DNN outperformed other models because smoking prediction relies on complex biological patterns that cannot be captured by simple linear or margin-based classifiers.

5.2 Confusion Matrix and ROC Curve

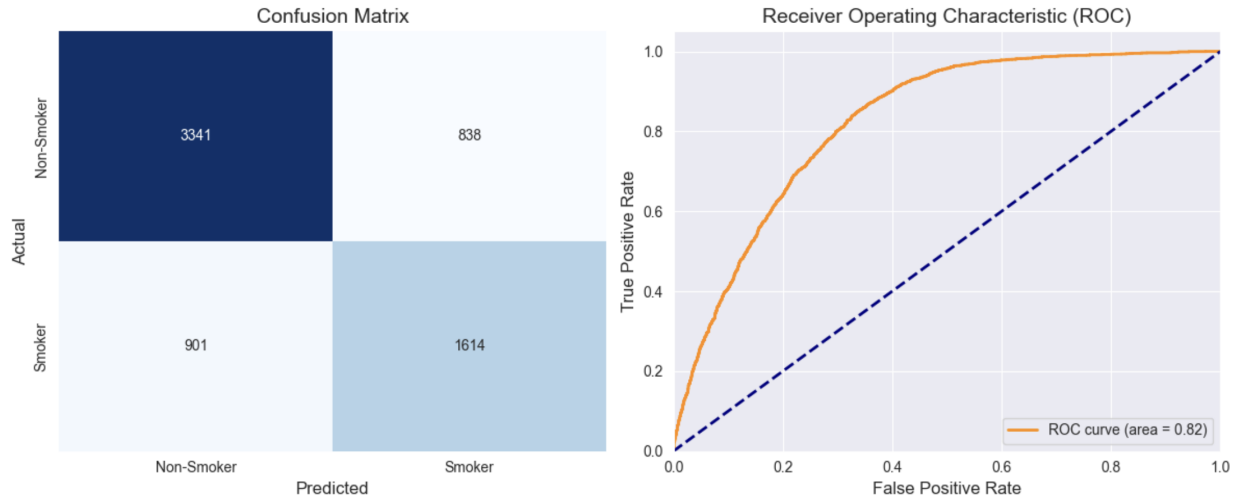

Figure 5: Confusion Matrix and ROC Curve (AUC = 0.82)

5.3 Accuracy Scores

Table 2: Model Performance Comparison

Model	**Accuracy**
Logistic Regression	0.73.5
Linear SVM	0.73
Deep Neural Network (DNN)	**0.75**

6 Comparative Analysis

Table 3: Comparison of Models and Their Strengths

Model	**Why it Works**	**Limitations**	**Reason for Performance Ranking**
Logistic Regression	Simple, interpretable, and effective for linearly separable patterns.	Assumes linear decision boundaries; cannot capture complex metabolic interactions.	Lowest performance because smoking prediction involves non-linear relationships between biological markers (e.g., GTP, triglycerides, BMI).
Linear SVM	Maximizes margin and handles moderate feature interactions better than Logistic Regression.	Still fundamentally linear; fails to capture high-order nonlinear relationships without kernels (which were not used).	Performs slightly better than Logistic Regression but worse than DNN because decision boundaries remain linear while dataset patterns are non-linear.
Deep Neural Network (DNN)	Learns multi-level non-linear feature interactions; handles high-dimensional patterns effectively.	Requires careful scaling, regularization, and more computational resources.	Best performance due to ability to model complex biological relationships (e.g., enzyme levels, lipid interactions, log-transformed features).

7 Conclusion

This project demonstrates that smoking behavior can be predicted with considerable accuracy using machine learning and deep learning techniques. Among all models tested, the Deep Neural Network performed the best with an accuracy of 75% and strong ROC performance.

Key takeaways:

- Feature engineering significantly improved model performance (log transformations, interaction terms).
- Mutual information revealed strong predictors such as GTP, height, BMI, and hemoglobin.
- DNNs outperform classical ML models due to their ability to learn deep, non-linear patterns.

Future improvements may include hyperparameter optimization (Optuna), class imbalance handling (SMOTE), or transformer-based architectures.

GitHub Repository Link: `https://github.com/kartikey09/Machine-Learning-Project-2`