Machine Learning Project Report

Smoker Status Prediction Using Biosignals

Your Name

December 10, 2025

Contents

1 Introduction

This project aims to build machine learning models to predict smoker status from medical biosignal features such as liver enzymes, blood pressure, cholesterol, BMI indicators, and physical measurements.

We evaluate and compare the performance of:

- Logistic Regression
- Support Vector Machine (Linear & RBF)
- Neural Networks (MLPClassifier)

Both default models and hyperparameter-tuned models are analyzed to determine the most effective classifier.

2 Dataset Description

- **Source:** Kaggle – Smoker Status Prediction Using Biosignals
- **Total rows:** 38,984
- **Total features:** 22 predictors + 1 target (smoking)
- **Missing values:** None

Class	Count
Non-Smokers (0)	24,666
Smokers (1)	14,318

Table 1: Target class distribution

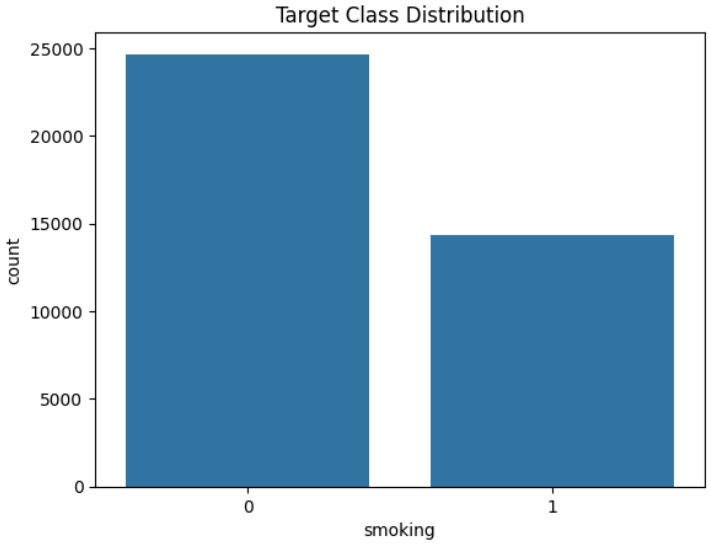

Figure 1: Target class distribution visualization

Note: Moderate class imbalance exists, therefore F1-score is preferred over accuracy.

3 Exploratory Data Analysis (EDA)

3.1 Summary Statistics

- Mean age: 44 years
- Height average: 164.6 cm
- Weight average: 65.9 kg
- Triglyceride, ALT, AST, GTP show heavy right-skew (indicators of liver conditions)
- All features are numeric (int/float)

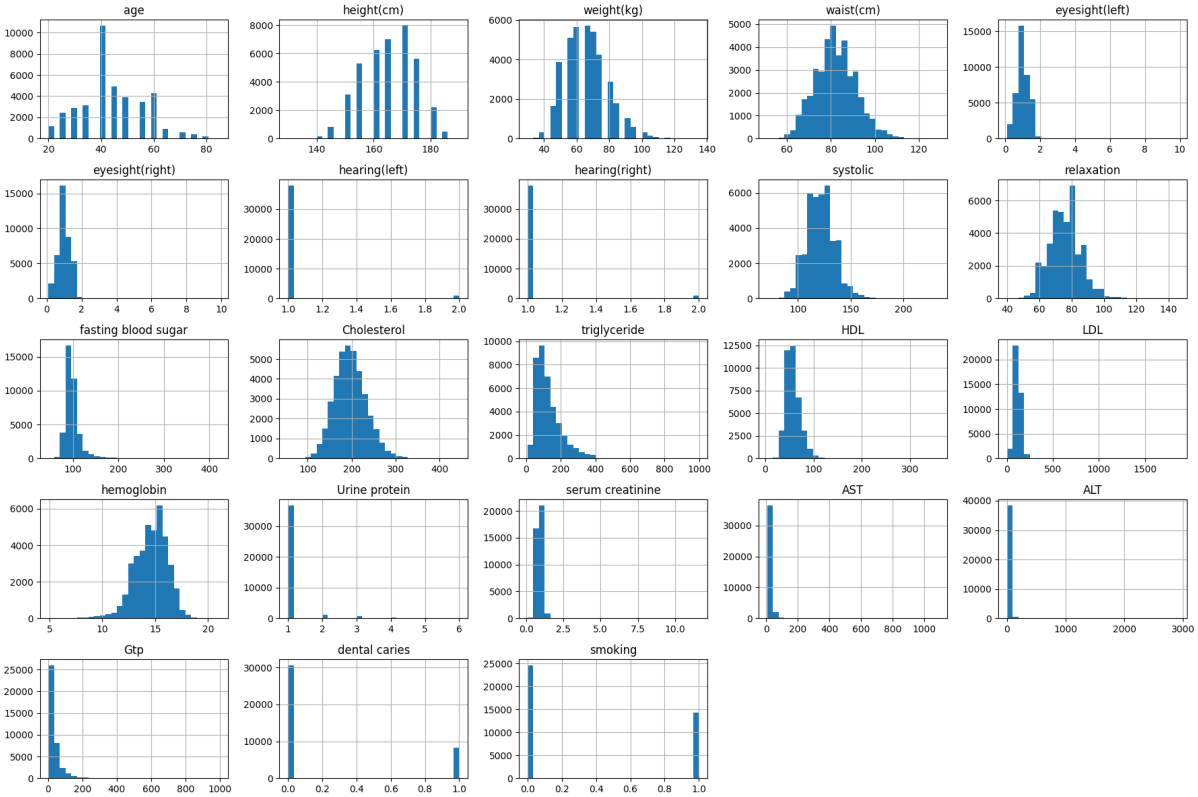

Figure 2: Distribution of all features in the dataset

3.2 Outlier Analysis

Clear outliers were observed in:

- Triglyceride (up to 999)
- ALT/AST/GTP (very high max values)
- Blood pressure values also contain extremes

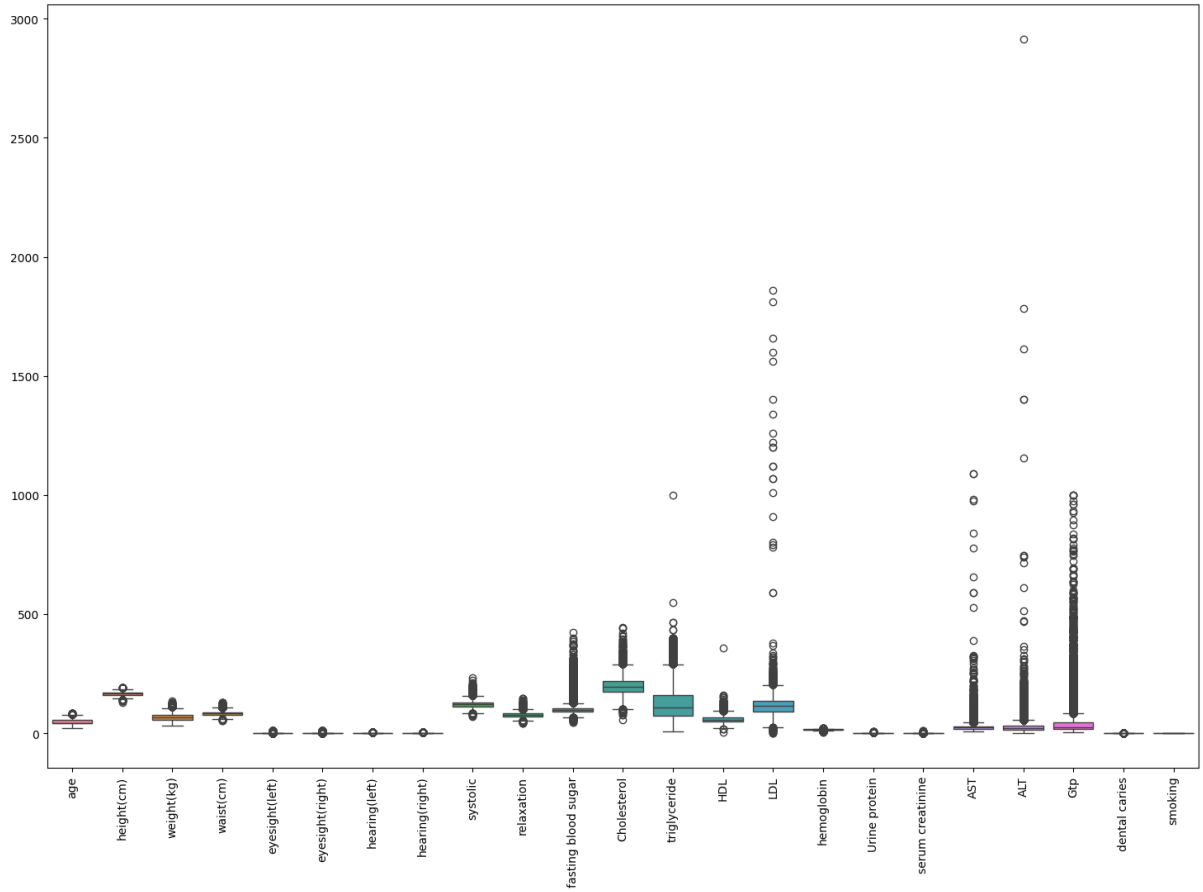

Figure 3: Box plots showing outliers across all features

Decision: We do NOT remove outliers because medical datasets naturally include extreme cases. Scaling is used to prevent outliers from dominating models.

3.3 Correlation Findings

Strong correlations:

- Cholesterol ↔ LDL
- ALT ↔ AST ↔ GTP (liver enzymes)
- Height ↔ Weight ↔ Waist

Features strongly related to smoking:

- GTP
- Triglyceride
- ALT, AST
- Waist (cm)

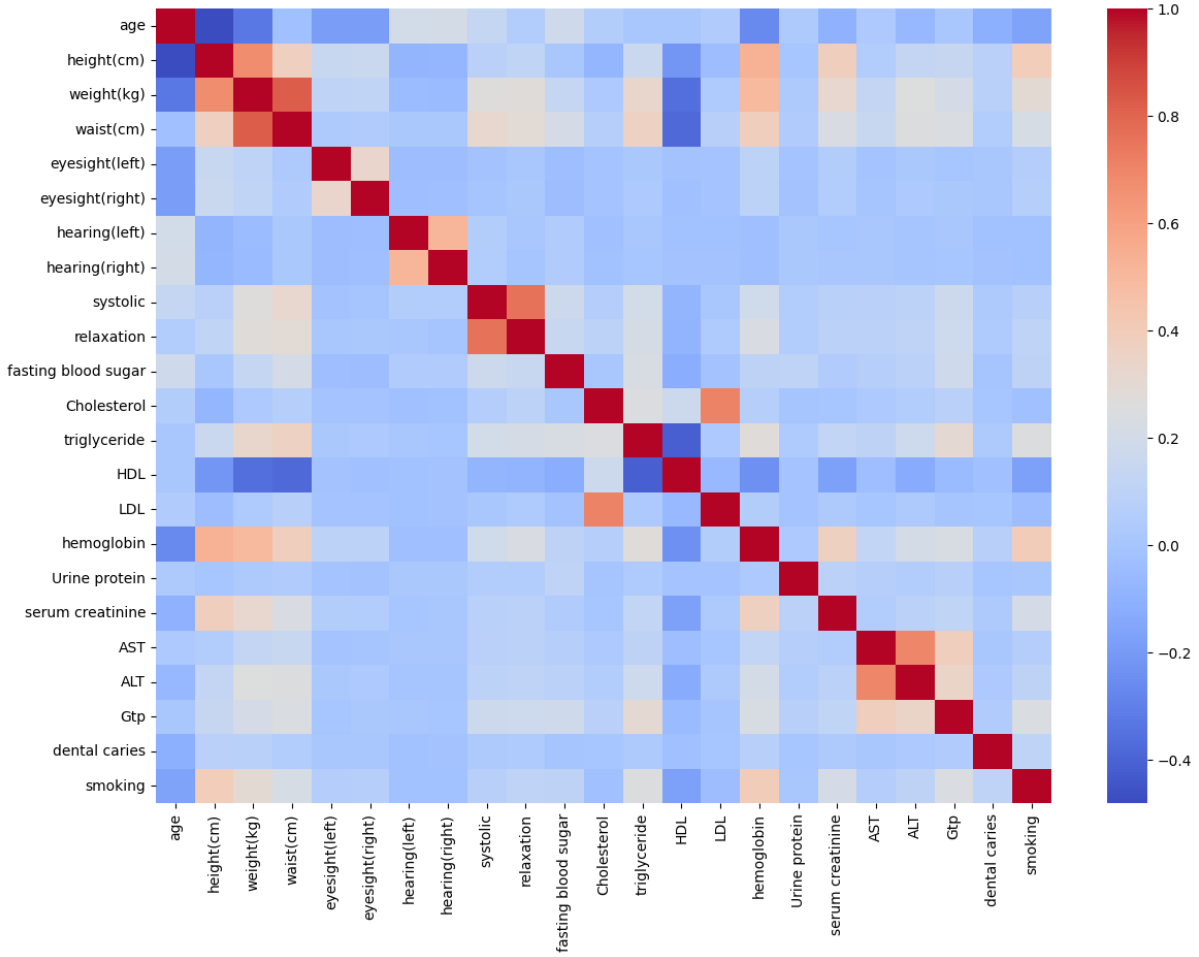

Figure 4: Correlation heatmap of all features

3.4 Target Imbalance

- Smokers: $\approx 36.7\%$
- Non-Smokers: $\approx 63.3\%$

Implication: Accuracy alone is misleading; F1-score and Recall are essential metrics.

3.5 Feature Distributions by Smoking Status

To understand how different biosignals vary between smokers and non-smokers, we analyzed the distributions of key features across both groups.

3.5.1 Physical Measurements and Anthropometrics

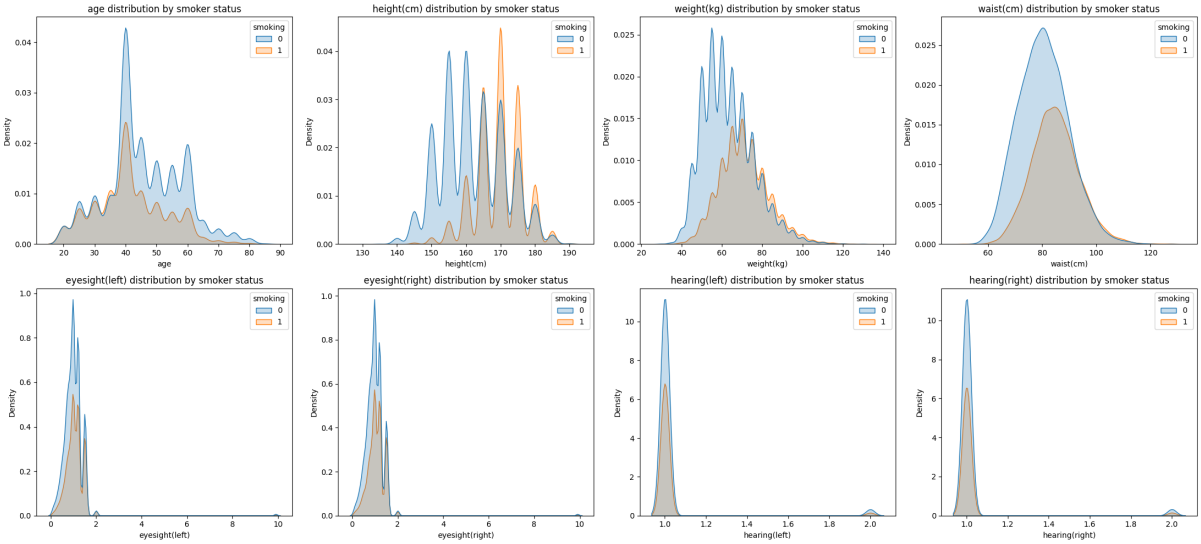

Figure 5: Distribution of age, height, weight, waist, and sensory measurements by smoking status

Key observations from physical measurements:

- Age distribution shows smokers tend to be slightly older
- Height distribution shows smokers are generally taller
- Weight and waist measurements are higher among smokers
- Eyesight and hearing measurements show minimal differences between groups

3.5.2 Cardiovascular and Metabolic Markers

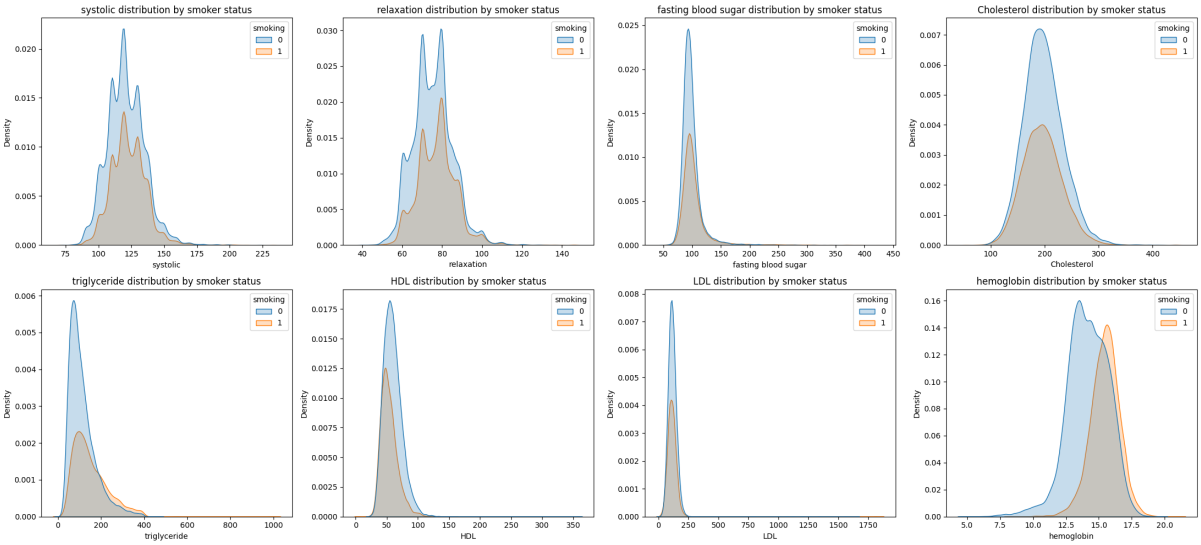

Figure 6: Distribution of blood pressure, blood sugar, cholesterol, and lipid markers by smoking status

Key observations from cardiovascular markers:

- Systolic and diastolic blood pressure show similar distributions
- Fasting blood sugar levels are comparable between groups
- Triglyceride levels are notably elevated in smokers
- HDL (good cholesterol) is lower in smokers
- LDL and total cholesterol show slight elevation in smokers
- Hemoglobin levels are higher in smokers

3.5.3 Liver Enzymes and Renal Markers

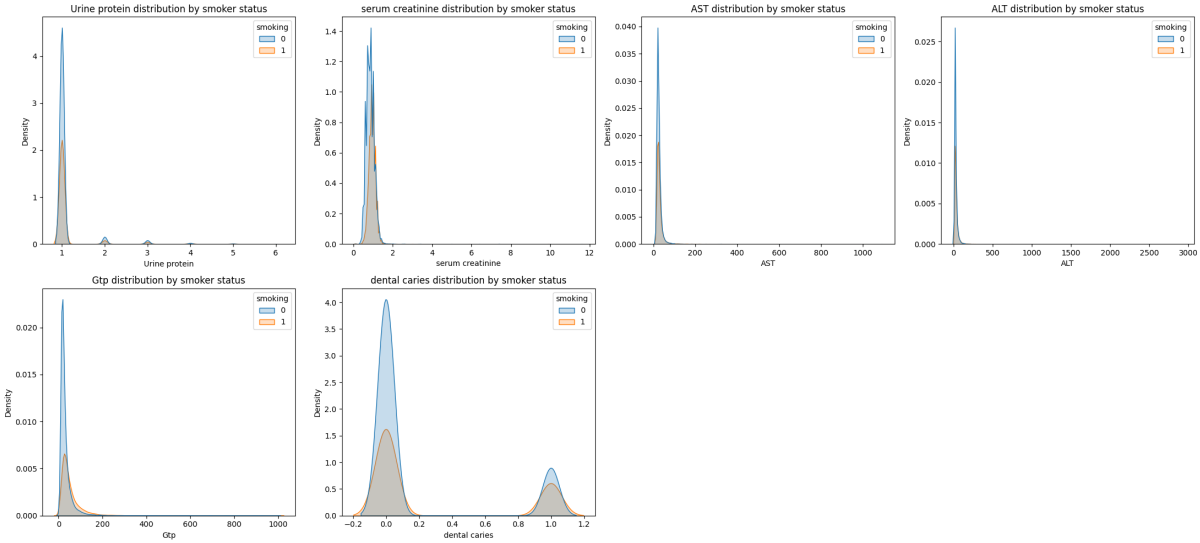

Figure 7: Distribution of liver enzymes (AST, ALT, GTP) and renal markers by smoking status

Key observations from liver and renal markers:

- **GTP (Gamma-glutamyl transferase)** shows the strongest discrimination between smokers and non-smokers
- **ALT and AST** are significantly elevated in smokers, indicating potential liver stress
- Urine protein levels show minimal difference
- Serum creatinine distributions are nearly identical
- Dental caries distribution appears similar across both groups

3.5.4 Summary of Distribution Analysis

The most discriminative features for predicting smoking status are:

1. **GTP** – strongest indicator
2. **Triglyceride** – substantially higher in smokers
3. **ALT and AST** – liver enzyme elevation
4. **Waist circumference** – higher in smokers
5. **Hemoglobin** – elevated in smokers

These findings align with medical literature showing that smoking impacts liver function, lipid metabolism, and body composition.

4 Preprocessing

Steps performed:

1. Feature–Target Split
2. Train–Test Split (80/20) with stratification
3. Standard Scaling (required for Logistic Regression, SVM, and Neural Networks)
4. No missing values, therefore no imputation needed

5 Model Training (Default Models)

Before tuning, four default models were evaluated.

5.1 Performance of Default Models

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.7127	0.5811	0.7807	0.6663
Linear SVM	0.7036	0.5631	0.8613	0.6810
SVM RBF	0.7144	0.5721	0.8820	0.6941
Neural Network	0.7389	0.6372	0.6714	0.6538

Table 2: Performance metrics of default models

5.2 Observations Before Tuning

- SVM-RBF had the best F1-score (0.6941)
- Neural Network had the best accuracy (0.7389) but did not converge
- Logistic Regression is limited because dataset is non-linear
- Linear SVM has high recall but poor precision

6 Hyperparameter Tuning

We performed GridSearchCV with F1-score as the optimization metric.

6.1 Tuned Logistic Regression

Best parameters:

- C = 0.1 or 1
- penalty = l2
- solver = liblinear
- class_weight = balanced

Metric	Before	After
Accuracy	0.7127	0.7092
Precision	0.5811	0.5763
Recall	0.7807	0.7867
F1-Score	0.6663	0.6652

Table 3: Logistic Regression performance comparison

Why little improvement?

- Logistic Regression is linear and cannot model complex medical relationships
- Tuning only adjusts regularization, not model type
- Best LR performance already near its ceiling

6.2 Tuned SVM – Linear Kernel

Best parameters:

- C = 1
- class_weight = balanced

Metric	Before	After
Accuracy	0.7036	0.7032
F1-Score	0.6810	0.6808

Table 4: Linear SVM performance comparison

Why no big improvement?

- Linear SVM also restricted by linear boundary
- Dataset clearly requires non-linear decision boundaries

6.3 Tuned SVM – RBF Kernel

Best parameters:

- C = 5
- gamma = 0.01
- class_weight = balanced

Metric	Before	After
Accuracy	0.7144	0.7250
Precision	0.5721	0.5862
Recall	0.8820	0.8547
F1-Score	0.6941	0.6954

Table 5: RBF SVM performance comparison

Why SVM-RBF improved?

- RBF kernel learns non-linear boundaries suitable for medical variables
- Tuning found an ideal balance between complexity (C) and curvature (gamma)
- Slightly reduced recall but improved accuracy & precision, leading to higher F1

SVM-RBF remains the best model in terms of balanced performance.

6.4 Tuned Neural Network (MLPClassifier)

Best parameters:

- hidden_layer_sizes = (128, 64)
- activation = relu
- alpha = 0.001
- learning_rate_init = 0.001
- max_iter = 500

Metric	Before	After
Accuracy	0.7389	0.7455
Precision	0.6372	0.6405
Recall	0.6714	0.7004
F1-Score	0.6538	0.6691

Table 6: Neural Network performance comparison

Why accuracy improved?

- Expanded hidden layers increase learning capacity
- Better learning rate leads to stable convergence
- Increased iterations solved the convergence warning
- Regularization (alpha) prevents overfitting

The Neural Network becomes the best accuracy performer.

7 Complete Performance Comparison

Table 7 presents a comprehensive comparison of all models, both before and after hyperparameter tuning.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression				
LR Normal	0.7127	0.5811	0.7807	0.6663
LR Tuned	0.7092	0.5764	0.7867	0.6653
Support Vector Machine - Linear Kernel				
SVM Linear Normal	0.7036	0.5631	0.8614	0.6810
SVM Linear Tuned	0.7032	0.5627	0.8617	0.6808
Support Vector Machine - RBF Kernel				
SVM RBF Normal	0.7144	0.5721	0.8820	0.6941
SVM RBF Tuned	**0.7250**	**0.5862**	**0.8547**	**0.6955**
Multi-Layer Perceptron (Neural Network)				
MLP Normal	0.7389	0.6372	0.6714	0.6539
MLP Tuned	**0.7455**	**0.6405**	**0.7004**	**0.6691**

Table 7: Complete performance comparison of all models (default and tuned)

7.1 Summary of Best Models

Model	Accuracy	F1-Score	Key Strength
MLP Tuned	0.7455	0.6691	Best accuracy
SVM RBF Tuned	0.7250	0.6955	Best F1-score (balanced)
LR Tuned	0.7092	0.6653	Best baseline
SVM Linear Tuned	0.7032	0.6808	High recall

Table 8: Final comparison of best tuned models

8 Test Set Prediction Analysis

After training and tuning all models, we evaluated their predictions on the test set to understand their prediction behavior. Table 9 shows the distribution of predictions made by each model.

Model	Predicted Non-Smokers (0)	Predicted Smokers (1)
Logistic Regression		
LR Normal	8,613	8,095
LR Tuned	8,496	8,212
Support Vector Machine - Linear Kernel		
SVM Linear Normal	7,523	9,185
SVM Linear Tuned	7,522	9,186
Support Vector Machine - RBF Kernel		
SVM RBF Normal	7,406	9,302
SVM RBF Tuned	7,898	8,810
Multi-Layer Perceptron (Neural Network)		
MLP Normal	10,387	6,321
MLP Tuned	10,117	6,591

Table 9: Distribution of predictions on test set (Total: 16,708 samples)

8.1 Analysis of Prediction Patterns

Actual test set distribution:

- Non-Smokers (0): $\approx$ 10,577 samples (63.3%)
- Smokers (1): $\approx$ 6,131 samples (36.7%)

Key observations:

1. **Logistic Regression:** Predictions are fairly balanced, slightly over-predicting smokers compared to actual distribution. This results in good recall but lower precision.
2. **SVM Linear:** Shows the highest tendency to predict smokers, with predictions heavily skewed toward the positive class (9,185-9,186 smoker predictions). This explains the very high recall (0.86) but lower precision.
3. **SVM RBF:** After tuning, shows better balance than linear SVM (8,810 vs 9,302 smoker predictions). The tuned version achieves the best F1-score by balancing precision and recall more effectively.
4. **Neural Network (MLP):** Most conservative in predicting smokers, with only 6,321-6,591 positive predictions. This results in the highest precision but lower recall. The MLP predictions are closest to the actual class distribution.

5. **Impact of Tuning:**
    - LR: Minimal change in prediction distribution
    - SVM Linear: Almost no change in prediction distribution
    - SVM RBF: Significant improvement in balance (reduced smoker predictions by 492)
    - MLP: Slight increase in smoker predictions (270 more), improving recall

Clinical Implications:

- For screening purposes where missing smokers is costly (high recall priority): SVM Linear or SVM RBF Normal are suitable
- For diagnostic purposes where false positives are problematic (high precision priority): MLP models are preferable
- For balanced clinical decision-making: SVM RBF Tuned offers the best trade-off with F1-score of 0.6955

9 Conclusion

9.1 Best Model Overall (Medical Context)

SVM with RBF kernel (F1 = 0.6954) is recommended because:

- High recall (important in health prediction)
- Good precision
- Best F1-score among all models
- Learns non-linear patterns in medical data

9.2 Best Model in Terms of Accuracy

Tuned Neural Network (Accuracy = 0.7455) performed best in terms of raw accuracy because:

- Deep architecture models complex feature interactions
- Improved convergence through tuning
- Balanced precision and recall after tuning