Smoking Classification Using Bio-Signal Data

1. Project Objective

The aim of this project is to develop a machine learning model that can accurately classify whether an individual is a smoker or not using basic health and biological signal data. This smart classification system is intended to support healthcare data analytics in identifying smoking habits based on measurable indicators from health checkups.

---

2. Dataset Overview

The dataset used for this case study consists of approximately 55,000 records and 27 features. These include both categorical and numerical health indicators such as:

- **Demographics**: Gender, Age
- **Body Measurements**: Height (cm), Weight (kg), Waist circumference (cm)
- **Sensory Test Results**: Eyesight (left/right), Hearing (left/right)
- **Blood Pressure**: Systolic and Diastolic
- **Biochemical Tests**: Fasting blood sugar, Cholesterol, Triglyceride, HDL, LDL
- **Liver Enzymes**: AST, ALT, GTP
- **Kidney and Blood Indicators**: Hemoglobin, Urine Protein, Serum Creatinine
- **Oral Health**: Dental caries, Tartar, Oral exam
- **Target Variable**: Smoking status (0 = Non-smoker, 1 = Smoker)

---

3. Data Preprocessing

3.1 Null Handling and Encoding

- Checked for null or missing values in all columns; appropriate handling applied where needed.
- Converted categorical variables to numerical:
    - Gender: M = 1, F = 0
    - Oral, Tartar: Y = 1, N = 0

3.2 Outlier Removal

- Used the **IQR (Interquartile Range)** method on 22 numerical columns to remove statistical outliers.
- Columns treated include age, height, weight, waist, blood sugar, cholesterol, liver enzymes, and kidney function markers.

- Resulted in significant reduction of dataset size but improved the reliability of model training.

---

4. Exploratory Data Analysis (EDA)

4.1 Distribution Insights

- **Age**: Most individuals are in the 30–60 range, peaking at 40–50.
- **Gender**: Male-dominated (35,401 males vs. 20,291 females).
- **Height and Weight**: Heights peak around 170 cm; most weights fall between 60–80 kg.
- **Waist Circumference**: Most common range is 80–90 cm.
- **Blood Pressure**: Systolic values cluster between 100–140 mmHg; diastolic between 70–90 mmHg.
- **Biochemical Tests**:
    - Cholesterol: Majority in the 175–225 range.
    - Triglyceride: Clustered around 80–150.
    - HDL: Most values between 40–60.
    - LDL: Peaks around 110–130.
- **Oral Health**: About 11,800 people have dental caries; 30,940 have tartar.

4.2 Smoking Distribution

- Non-smokers: 63% of dataset (≈35,237)
- Smokers: 37% (≈20,455)

---

5. Correlation Analysis

5.1 Full Correlation Matrix

- Computed pairwise Pearson correlation between all numerical features.

5.2 Key Correlations with Smoking:

- Positive:
    - Gender (0.51)
    - Hemoglobin (0.40)
    - Height (0.40)
    - Weight (0.30)
    - Triglyceride (0.25)
    - GTP (0.24)

    - Waist (0.23)
    - Serum Creatinine (0.22)
- Negative:
    - HDL (-0.18)
    - Age (-0.16)

These suggest that smoking habits correlate strongly with physical measurements and some blood and liver-related health markers.

---

6. Feature Selection

Method:

- Used SelectKBest with ANOVA F-test to rank features.
- Selected top 10 predictors of smoking:
    - Hemoglobin, Height, Weight, Waist, Gender
    - HDL, GTP, Creatinine, Triglyceride, LDL

---

7. Model Building and Evaluation

Models Trained:

- Logistic Regression
- Decision Tree Classifier
- Random Forest Classifier

Training Details:

- Train-test split: 80/20
- Standard scaling applied to continuous features
- Evaluation metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC

Model Performance:

Model	Accuracy	ROC-AUC
Logistic Regression	~0.79	~0.84
Decision Tree	~0.76	~0.81
Random Forest	~0.84	~0.88

Random Forest delivered the highest accuracy and AUC, indicating strong generalization and reliability.

---

8. Final Conclusion

- The Random Forest Classifier is best suited for this classification task.
- Gender, Hemoglobin, and liver enzymes (GTP) are highly influential features in predicting smoking.
- The model is effective and interpretable for deployment in healthcare analytics systems.

---

9. Recommendations

- Address gender imbalance in future data collection.
- Explore SHAP/LIME for model interpretability in medical contexts.
- Consider deeper models (e.g., XGBoost) or ensemble stacking to improve performance.

---

10. Project Files & References

- Script: casestudy.py
- Dataset: Provided in tabular format (55K rows)
- Libraries used: pandas, numpy, seaborn, matplotlib, sklearn