Model Development Report: Cardiovascular Disease Prediction

Key Findings

1. **Top Model**: **XGBoost** achieved **73.67% accuracy** on test data, with an average precision, recall, and F1 score of **0.74**.
2. **Feature Selection**: Identified 15 key features, heavily influenced by clinical biomarkers such as **age, BMI, blood pressure, cholesterol, and glucose levels**.
3. **Overfitting Issues**:
    1. **KNN** showed significant overfitting (train: 76.6% → test: 70.4%).
    2. **Random Forest** had minor overfitting (train: 74.5% → test: 73.4%).
4. **Class Imbalance**: Models slightly preferred predicting **Class 0 (no disease)**, indicated by a higher recall for this class.

Selected Features

The following 15 features were retained after preprocessing and feature selection:

['weight', 'ap_hi', 'ap_lo', 'pulse_pressure', 'age_years', 'bmi', 'cholesterol', 'gluc', bp_category_Elevated', 'bp_category_Hypertension Stage 1', 'bp_category_Hypertension Stage 2', 'bp_category_Normal', 'bmi_category_Normal', 'bmi_category_Obese', 'is_obese']

Insights:

1. Excluded demographic/lifestyle factors (gender, smoking, alcohol, and active), showing that clinical biomarkers hold more predictive power.
2. The importance of **blood pressure** and **BMI categories** emphasizes the need for stratified risk assessment in cardiovascular predictions.

Model Performance Summary

Model	Test Accuracy	Precision	Recall	F1-Score	Overfitting Risk
XGBoost	**73.67%**	**0.74**	**0.74**	**0.74**	Low
Random Forest	73.44%	0.74	0.73	0.73	Mild
Logistic Regression	72.68%	0.73	0.73	0.72	Low
SVM	72.62%	0.73	0.73	0.72	Low
KNN	70.35%	0.70	0.70	0.70	**High**

XGBoost: Best Model Details

Hyperparameters

{clf__n_estimators': 200, clf__max_depth': 3, 'clf__learning rate': 0.1}

Classification Report

	precision	recall	f1-score	support
0	0.72	0.79	0.75	6907
1	0.76	0.68	0.72	6734
accuracy			0.74	13641

1. **Strengths**: Balanced performance across both classes.
2. **Weakness**: Slightly lower recall for Class 1 (cardiovascular disease cases).

Confusion Matrix Insights

Similar patterns across all models:

1. **True Positives (TP)**: ~72% of cases accurately identified.
2. **False Negatives (FN)**: ~14% of high-risk patients are misclassified.
3. **False Positives (FP)**: ~14% of low-risk patients incorrectly flagged.

Deployment Readiness

✅ **Preprocessor, feature selector, and XGBoost model successfully saved**.

✅ **Pipeline reproducibility confirmed** (test accuracy aligns with final evaluation).