Predictive Modeling of Cardiovascular Disease Risk

This project aims to develop machine learning models capable of predicting the risk of cardiovascular disease (CVD) using structured patient data. Cardiovascular disease remains one of the leading causes of death worldwide, and early detection through data-driven tools is critical in reducing its burden. By analyzing anonymized clinical records of over 60,000 individuals, we explore patterns in biometric, demographic, and lifestyle variables that signal an elevated risk of heart disease. Through this work, we demonstrate how machine learning can complement clinical expertise and support targeted preventive care strategies.

Dataset Overview and Feature Engineering

The dataset used for this project was derived from a publicly available collection of patient records. It contained several key features: age (in days), gender, height, weight, systolic and diastolic blood pressure, cholesterol, glucose levels, smoking and alcohol habits, physical activity, and a binary target variable indicating the presence or absence of cardiovascular disease.

Initial preprocessing addressed data formatting challenges. The original dataset used a semicolon delimiter, which was corrected during the import stage. Missing or invalid data entries were removed, and duplicate records were checked. Feature engineering played a critical role in enriching the dataset. Two significant derived features were introduced:

- **Age in years**: Converted from the original age (in days).
- **Body Mass Index (BMI)**: Calculated from height and weight to capture obesity-related risk.

Categorical features such as cholesterol and glucose were label-encoded for modeling. Outliers in blood pressure and BMI were assessed and clipped within realistic medical ranges to avoid skewing model training.

Exploratory Data Analysis (EDA)

EDA revealed several trends in the dataset. Most patients were middle-aged, with a median age of around 53 years. Men made up a slight majority. Cardiovascular disease incidence was noticeably higher among patients with high systolic blood pressure (>140 mmHg), high cholesterol levels, sedentary habits, and elevated BMI (above 30).

Visualizations such as histograms, boxplots, and correlation heatmaps supported the initial analysis. Patients with CVD typically had higher BMI and blood pressure compared to those without. A heatmap showed strong correlations between systolic/diastolic pressure and cardiovascular risk, confirming their importance as predictive features.

Machine Learning Models and Evaluation

We trained and evaluated four machine learning classifiers:

1. **Logistic Regression**: Served as a baseline. It is fast and interpretable but limited in capturing nonlinear patterns.
2. **Random Forest**: Provided better performance and handled feature interactions well. The model also provided feature importance rankings.
3. **XGBoost**: Delivered the highest performance across most metrics. It effectively handled class imbalance and overfitting.
4. **K-Nearest Neighbors (KNN)**: Provided a simple non-parametric comparison but struggled with scalability and accuracy in this high-dimensional context.

Models were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. XGBoost achieved the best results, with an accuracy of ~0.78 and AUC > 0.84, followed closely by Random Forest. The confusion matrix showed that both models were reasonably balanced in identifying true positives and true negatives.

Feature Importance and Key Insights

Tree-based models highlighted that the most influential predictors of cardiovascular disease were:

- Systolic blood pressure
- Age (in years)
- Cholesterol level
- BMI
- Physical inactivity

Aging was directly associated with increasing risk, particularly beyond age 50. Obese individuals with high blood pressure and poor lifestyle habits showed disproportionately high CVD rates. Feature importance visualizations and partial dependence plots further emphasized the marginal effect of each variable on the predicted probability.

Visual Analytics and Interpretability

To supplement modeling, visual analytics were used extensively:

- **Feature Importance Plots**: Explained model behavior.
- **Age Distribution by Risk**: Highlighted age-based thresholds.
- **BMI Boxplots by Class**: Confirmed obesity as a contributing factor.
- **ROC Curves**: Assessed classifier performance under different thresholds.

Together, these tools helped balance predictive accuracy with clinical interpretability—a key requirement in healthcare applications.

Limitations and Future Enhancements

While the models performed well, some limitations remain:

- **Lack of contextual variables**: The dataset does not include data on family history, genetic markers, mental health, medication, or socioeconomic status.
- **No temporal data**: Longitudinal data could improve prediction by tracking changes over time.
- **Single snapshot**: All variables represent a single point in time, which limits dynamic risk modeling.

To enhance model quality and usefulness:

- Incorporate wearable device data for real-time monitoring.
- Use explainable AI methods such as SHAP values to make predictions transparent.
- Deploy models as part of clinical decision support tools embedded in hospital systems.

Conclusion

This study demonstrates how structured health data, combined with machine learning, can predict cardiovascular disease risk with high accuracy. The models—especially XGBoost—successfully identify the key risk indicators and provide a solid foundation for clinical screening tools. With further refinement, inclusion of more contextual features, and validation on diverse populations, these models can evolve into powerful digital health tools.

In an era where proactive healthcare is increasingly important, such data-driven solutions have the potential to reduce long-term health burdens, improve patient outcomes, and allocate healthcare resources more efficiently.

All code, notebooks, and visualizations are available at:
https://github.com/Santhakumarramesh/Cardiovascular-Risk-Prediction