Capstone Project Summary: Heart Disease Prediction Using Machine Learning

This capstone project aims to predict the presence of heart disease using the UCI Heart Disease dataset. The goal is to apply machine learning techniques to assist in early detection and diagnosis, which is vital for improving patient outcomes and managing healthcare resources efficiently.

The dataset includes 303 patient records with 13 features, such as age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, maximum heart rate, and exercise-induced angina. After loading the dataset, we conducted data preprocessing, including handling missing values, encoding categorical variables, and standardizing numerical features to prepare the data for modeling.

We began with Exploratory Data Analysis (EDA) to better understand the distribution and relationships of features. Histograms were used to visualize the frequency and spread of each feature, and a correlation heatmap helped identify patterns and potential multicollinearity. Dimensionality reduction techniques—Principal Component Analysis (PCA), t-SNE, and UMAP—were also applied to visualize how the data clusters in 2D space based on the target variable.

For classification, we trained multiple machine learning models: Logistic Regression, Random Forest, Gradient Boosting Classifier, and K-Nearest Neighbors (KNN). Hyperparameter tuning was performed for each model using `RandomizedSearchCV` with cross-validation to find the optimal parameters.

After evaluating the models using accuracy scores on the test set, **K-Nearest Neighbors (KNN)** outperformed the other models. It demonstrated strong predictive capability and was chosen as the final model.

Key Findings:

- KNN achieved the highest test accuracy after hyperparameter tuning.
- Features such as chest pain type, maximum heart rate, and ST depression (oldpeak) showed strong associations with heart disease.
- Dimensionality reduction techniques helped visualize class separation, especially with t-SNE and UMAP.

Next Steps:

- Consider collecting more data or using ensemble methods to improve robustness.
- Explore deployment options, such as building a user interface with tools like Gradio.
- Investigate longitudinal or time-based patient data for deeper clinical insights.

This project demonstrates how even simple models like KNN can be powerful with good preprocessing and tuning, and highlights the importance of interpretable, reproducible machine learning in healthcare applications.