Predicting Heart Disease Using Machine Learning

PROJECT PROGRESS OVERVIEW:

We have made significant progress in building a heart disease prediction model using machine learning techniques. The preprocessing phase involved cleaning the dataset by removing one duplicate row and handling outliers in cholesterol and age using normalization techniques that reduce the impact of extreme values. We ensured there were no missing values in the dataset, providing a strong foundation for model development.

For data cleaning, we scaled numerical features such as age and blood pressure between 0 and 1, and encoded categorical features (e.g., chest pain type) using One-Hot and Binary Encoding. We then applied Recursive Feature Elimination (RFE) to retain the most significant features and reduce dimensionality.

In the data processing stage, we created visualizations including heatmaps and histograms to explore feature correlations. We also verified class balance between patients with and without heart disease to avoid model bias.

We experimented with six machine-learning models: Logistic Regression, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Decision Tree, Random Forest, and Naive Bayes. Among these, SVC delivered the highest accuracy of 86.6%, making it the best-performing model thus far.

Link: https://archive.ics.uci.edu/dataset/45/heart+disease

CHALLENGES:

Several challenges were encountered during the project. One of the primary issues was the presence of extreme values in features like cholesterol and age, which negatively impacted the model's performance. This was addressed by applying a method to minimize the influence of these outliers while preserving the integrity of the data.

Another challenge was identifying the most significant features for prediction. Initially, it was difficult to determine which features had the most impact. This problem was resolved using Recursive Feature Elimination (RFE), which helped in selecting the most relevant features for the model.

Additionally, fine-tuning the machine learning models to achieve optimal results proved to be time-consuming. To tackle this, both Grid Search and Random Search were used to systematically find the best hyperparameters for each model.

Still Working On:

To further improve the model's accuracy, we plan to experiment with more advanced algorithms such as XGBoost. This could potentially provide better predictive power and handle data complexities more effectively.

We also aim to go beyond just accuracy by evaluating other performance metrics, such as precision, recall, and the precision-recall trade-off, to gain a deeper understanding of the model’s performance, especially in imbalanced scenarios.

COLLABORATION:

Our team meets twice a week to share progress updates, troubleshoot issues, and collaborate on different aspects of the project. Each member contributes actively to tasks such as data cleaning, model development, and report writing. When someone falls behind due to other commitments or technical challenges, the team steps in to redistribute tasks fairly and sets clear deadlines to ensure smooth progress and equal contribution. Weekly on Friday nights, we have online or offline meetings.

NEXT STEPS:

Try Better Models:

We will experiment with high-performance models including **XGBoost**, **LightGBM**, and **CatBoost**, as well as neural networks to capture more complex patterns. We also plan to test model ensembling to combine strengths of individual models.

Add More Data:

Enhancing the dataset is a key focus. We will use Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples and improve class balance. Furthermore, we are exploring the possibility of acquiring additional patient datasets either from open repositories or through collaboration with healthcare providers.

Test in Real Life:

We aim to validate our model in real-world clinical settings by testing it with new, unseen patient data. Feedback from healthcare professionals and doctors will be sought to ensure the model's predictions are both practical and meaningful in medical contexts.

Improve Features:

To further refine the model, we plan to engineer new features by combining existing variables, which may uncover hidden relationships. Principal Component Analysis (PCA) will also be applied to reduce dimensionality and enhance data quality by eliminating noise and redundancy.

Tune the Best Models:

We intend to fine-tune the Support Vector Classifier (SVC) by adjusting its hyperparameters for improved accuracy. Similarly, further tuning of the Random Forest model will be performed to boost its performance and reduce overfitting.

Make Predictions Explainable:

We will apply **SHAP** and **LIME** to explain model predictions, increasing transparency and trust. We also plan to explore interpretable hybrid models.

Use More Info:

We aim to include additional health indicators such as family history and lifestyle habits. Medical expert input will guide us in validating the relevance of these features.

Automate Everything:

Our final goal is to build a fully automated pipeline for data preprocessing, model training, and evaluation, deployed on the **cloud** for easy access and scalability.