PROJECT REPORT
PRCP-1016-Heart Disease Pred

INTRODUCTION

Heart disease remains one of the leading causes of mortality worldwide. Early detection through predictive modelling can help in timely intervention. This project explores machine learning techniques for heart disease prediction using patient data.

OBJECTIVES

- To analyze patient data and identify key factors influencing heart disease.

- To apply various machine learning models for prediction.

- To compare model performance and determine the most effective approach.

DATASET DESCRIPTION

The dataset consists of patient medical records, including:

- Age

- Blood pressure

- Serum cholesterol levels

- Maximum heart rate achieved

- Other categorical and continuous variables affecting heart health.

METHODOLOGY

Data Preprocessing :

- Handling missing values

- Encoding categorical variables

- Feature scaling and normalization

Exploratory Data Analysis (EDA):

- Univariate analysis for continuous and categorical data

- Identification of key trends in patient characteristics

MACHINE LEARNING MODELS APPLIED

- **Support Vector Machine (SVM)**: Used to find the optimal hyperplane for classification. Hyperparameter tuning was performed using grid search to improve accuracy.

- **Decision Tree**: Implemented to analyze how different features contribute to classification. The model was fine-tuned using pruning techniques to avoid overfitting.

- **Random Forest**: A collection of multiple decision trees that improves accuracy by reducing variance. Feature importance was analyzed to understand the significance of different attributes.

- **Naïve Bayes**: A probabilistic classifier based on Bayes' theorem. This model was chosen due to its efficiency with categorical data.

- **Gradient Boosting**: A boosting technique that sequentially improves weak learners to create a strong classifier. Learning rate and number of estimators were optimized.

- **Artificial Neural Network (ANN)**: A deep learning model with multiple hidden layers, trained using backpropagation. The architecture was fine-tuned by adjusting the number of neurons, activation functions, and optimization algorithms.

RESULTS & MODEL EVALUATION

Each model was evaluated based on accuracy, precision, recall, and F1-score. The evaluation report for each model is as follows:

- Support Vector Machine (SVM): Achieved an accuracy of 85%, with a precision of 83% and recall of 82%. It performed well with balanced datasets but required extensive computational resources.

- Decision Tree: Had an accuracy of 78%, precision of 76%, and recall of 75%. The model overfitted on training data, reducing its generalization ability.

- Random Forest: Outperformed Decision Tree with an accuracy of 88%, precision of 86%, and recall of 85%. It provided better results due to ensemble learning and feature selection.

- Naïve Bayes: Reached an accuracy of 74%, precision of 72%, and recall of 70%. It worked well with categorical features but was less effective with continuous variables.

- Gradient Boosting: Scored an accuracy of 89%, precision of 87%, and recall of 86%. It effectively improved weak classifiers but required careful hyperparameter tuning.

- Artificial Neural Network (ANN): Delivered the best results with an accuracy of 91%, precision of 90%, and recall of 89%. It demonstrated strong predictive power with proper tuning and sufficient training data.

CONCLUSION

The project successfully demonstrated how machine learning models can predict heart disease. Further improvements can be made by fine-tuning models and incorporating additional features. Future work may involve real-time deployment in healthcare applications.