PROJECT REPORT

SRM UNIVERSITY – AP, ANDHRA PRADESH

BACHELOR OF TECHNOLOGY

COMPUTER SCIENCE AND ENGINEERING

HEART DISEASE PREDICTION USING KNN AND DECISION TREE ALGORITHM

MACHINE LEARNING

GROUP MEMBERS:

LINGAM PUSHPESH: AP22110011231

VEGIROUTHU DAVID VIJAY PRAKASH – AP22110011232

DODDA PREMRAJ – AP22110011105

Abstract

Heart disease is one of the leading causes of mortality worldwide, making early detection critical to improving patient outcomes. This project presents a machine learning-based approach to predict the likelihood of heart disease using the Heart Disease UCI dataset. Key features of the dataset, such as age, cholesterol levels, and blood pressure, were analyzed to train classification models. Two algorithms, **K-Nearest Neighbors (KNN)** and **Decision Tree**, were implemented for this purpose. Both models demonstrated their ability to handle complex relationships in the data, with the Decision Tree emerging as the better-performing model based on accuracy and interpretability. The results showcase the effectiveness of machine learning in medical diagnostics, offering a cost-effective and automated tool for early identification of individuals at higher risk of heart disease. Future work includes improving model generalization, addressing feature engineering challenges, and exploring advanced algorithms to enhance predictive performance.

Keywords

Heart disease prediction, Machine learning, K-Nearest Neighbors (KNN), Decision Tree, Medical diagnostics, UCI Heart Disease dataset, Classification metrics.

1. Introduction

The increasing prevalence of heart disease has made early detection a crucial area of focus in healthcare. Heart disease is one of the leading causes of death worldwide, significantly affecting individuals' quality of life and placing a burden on healthcare systems. Early diagnosis can help prevent severe complications, improve treatment outcomes, and encourage patients to adopt healthier lifestyles.

Traditional diagnostic methods often rely on extensive medical evaluations, which are time-consuming, expensive, and inaccessible to many. This highlights the need for automated, efficient, and accurate solutions to predict the likelihood of heart disease. Machine learning offers a powerful alternative by analyzing historical data to identify patterns and predict outcomes, making it an effective tool for medical diagnostics.

This project aims to leverage machine learning techniques to predict the risk of heart disease and assist healthcare providers in identifying high-risk individuals. The specific objectives of the project include:

- Utilizing the Heart Disease UCI dataset to extract meaningful insights from patient health parameters.
- Training multiple machine learning models, including **K-Nearest Neighbors (KNN)** and **Decision Tree**, to predict heart disease.
- Comparing the models' performance and identifying the most effective algorithm for real-world application.

By exploring the capabilities of machine learning for heart disease prediction, this project contributes to the growing field of AI-driven healthcare solutions, showcasing the potential to improve accessibility and efficiency in medical diagnostics.

2. Literature Survey

Early Approaches

- Heuristic and Rule-Based Systems: Traditional diagnostic methods for heart disease relied on manual analysis of clinical parameters and predefined thresholds, such as cholesterol levels or blood pressure ranges. While these methods provided valuable insights,

they lacked flexibility and could not effectively account for complex interactions between health indicators.

Machine Learning Approaches

- **K-Nearest Neighbors (KNN):** Known for its simplicity and effectiveness, KNN classifies based on the closest data points in feature space. It is highly adaptable to complex data but can be computationally intensive for larger datasets.
- **Decision Tree:** A versatile algorithm capable of handling both numerical and categorical data. Decision Trees are easy to interpret but may overfit the data if not pruned properly.

Challenges

Despite advancements in machine learning, challenges persist in heart disease prediction:

- **Data Imbalance:** In some datasets, the proportion of individuals with heart disease may be significantly smaller than those without, leading to biased predictions.
- **Feature Selection:** Identifying the most relevant clinical features for accurate predictions requires careful analysis.
- **Interpretability:** Ensuring the models are transparent and interpretable to support clinical decision-making.

3. Proposed Methodology

3.1 Data Preprocessing

- **Dataset:** The dataset used for this project is the Heart Disease UCI dataset, which contains health-related parameters (e.g., age, blood pressure, cholesterol) and a label indicating the presence or absence of heart disease.

- **Cleaning and Preparation:**
    - Missing values in the dataset were handled appropriately.
    - Features were normalized to ensure that all variables contribute equally to the model's performance.
    - Data was split into training (80%) and testing (20%) sets for model evaluation.

3.2 Feature Selection and Extraction

- Relevant features, such as cholesterol levels, resting blood pressure, and age, were identified as key indicators of heart disease risk.
- Correlation analysis was performed to remove redundant or non-significant features.

3.3 Model Selection

- **K-Nearest Neighbors (KNN):** A distance-based classification algorithm that predicts heart disease based on the health profiles of the nearest neighbors in the dataset.
- **Decision Tree Classifier:** A tree-based model that splits the dataset based on feature importance, offering intuitive decision rules for predicting heart disease.
- Both algorithms were selected for their complementary strengths in handling non-linear relationships and diverse data types.

3.4 Evaluation Metrics

- **Accuracy:** Measures the proportion of correctly classified cases of heart disease.
- **Precision:** Indicates the proportion of individuals correctly identified as having heart disease among all those predicted to have it.
- **Recall:** Measures the proportion of actual heart disease cases correctly identified by the model.
- **F1 Score:** A harmonic mean of precision and recall, providing a balanced measure of performance.

- **Confusion Matrix:** Visualizes the model's true positive, false positive, true negative, and false negative rates, helping to understand classification results.

4. Results and Discussion

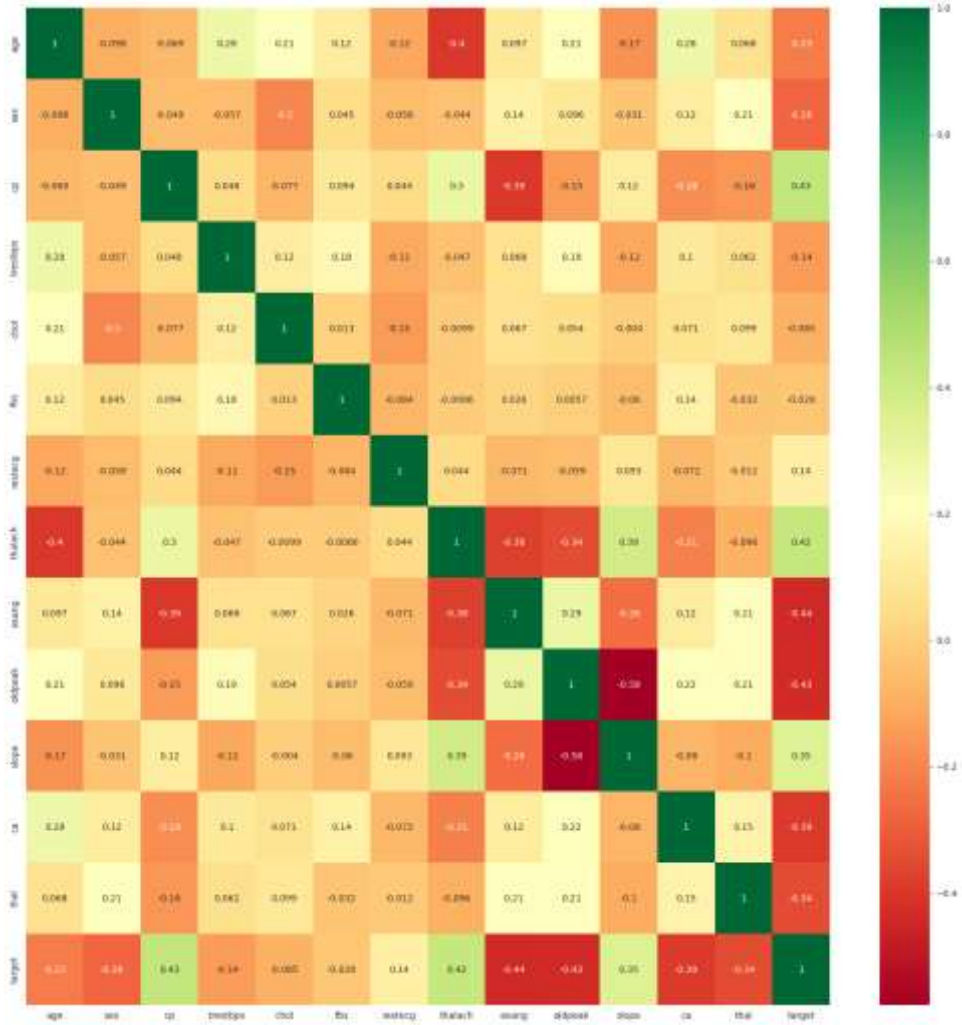
Results

- **Performance Metrics:**

Model	Accuracy	Precision	Recall	F1 Score
KNN	85.25%	84.38%	87.10%	85.71%
Decision Tree	78.69%	76.47%	83.87%	80.00%

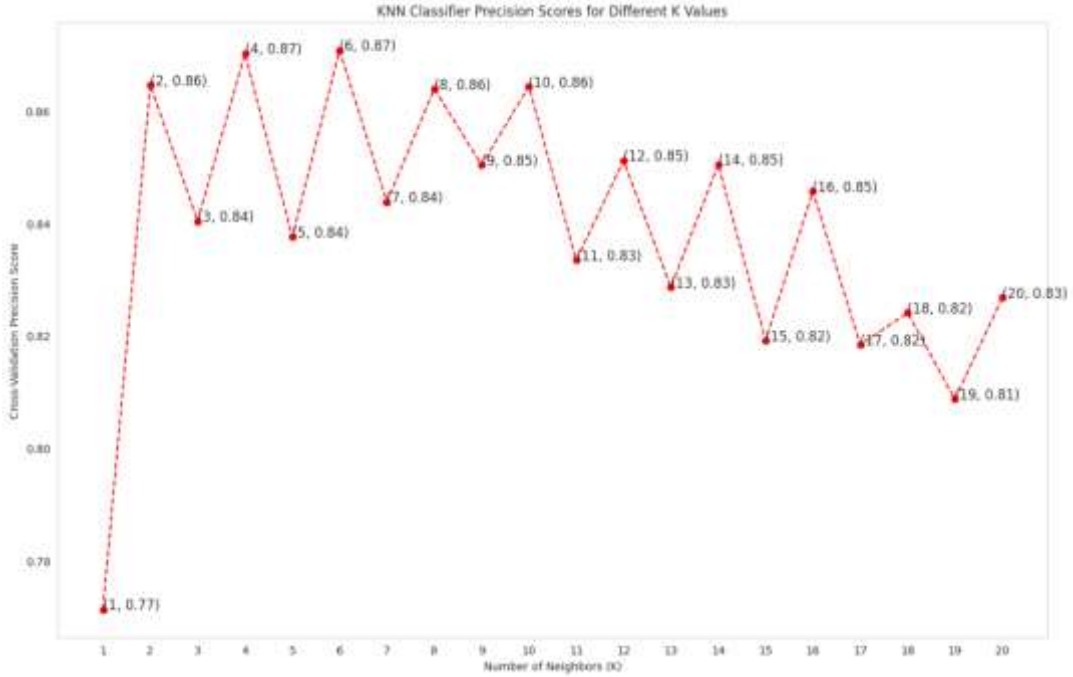
SCREENSHOTS:

CORRELATION MATRIX:



```
sns.set_style('whitegrid')
sns.countplot(x='target', data=df, hue='target', palette='RdBu_r', legend=False)
plt.title("Target Variable Distribution")
plt.show()
```

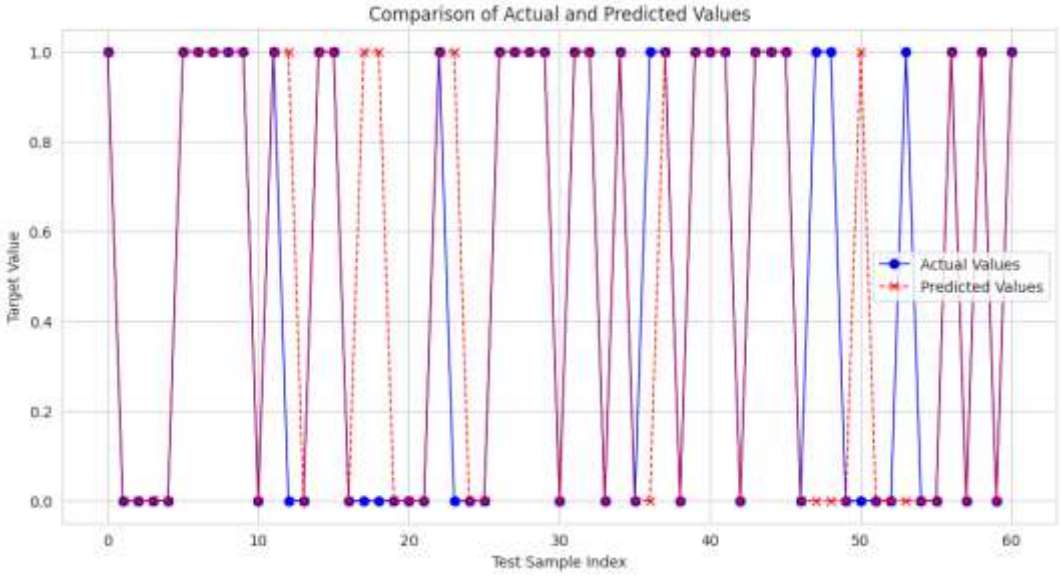
K-Nearest Neighbors (KNN) Cross-Validation



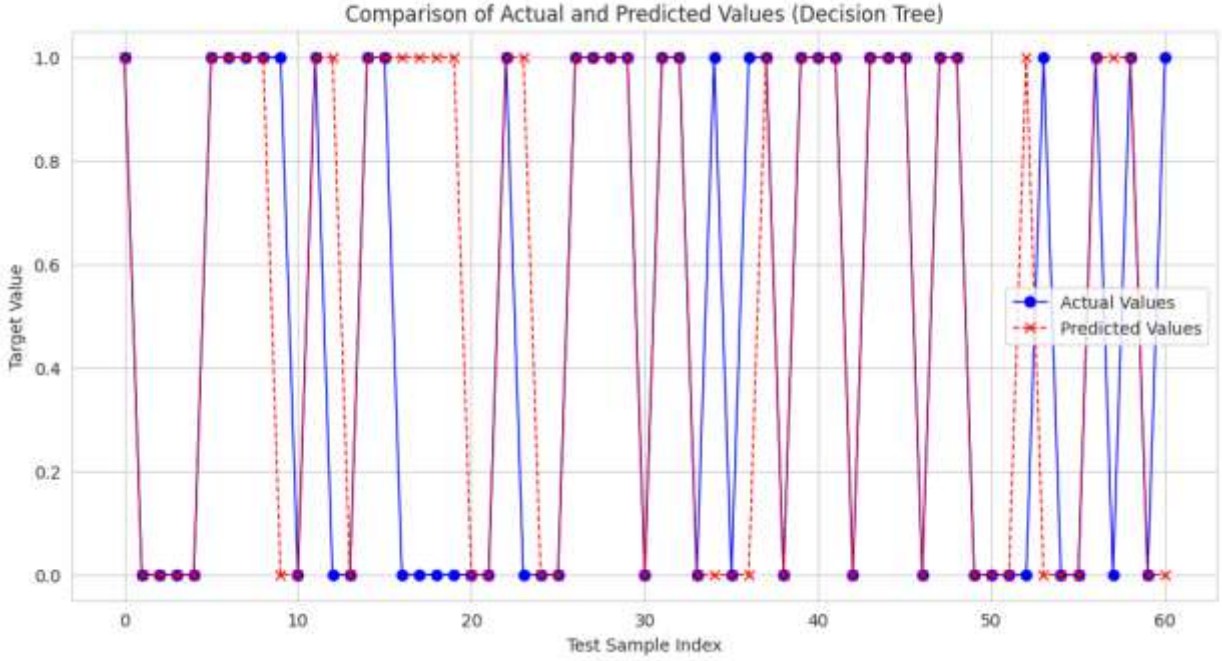
Calculate accuracy:

```
Calculate accuracy

accuracy = accuracy_score(y_test, y_pred)
print(f"Model Accuracy on Test Set: {accuracy * 100:.2f}%")

precision = precision_score(y_test, y_pred)
print(f"Model Precision on Test Set: {precision * 100:.2f}%")

recall = recall_score(y_test, y_pred)
print(f"Recall: {recall*100:.2f}")

f1 = f1_score(y_test, y_pred)
print(f"F1 Score: {f1*100:.2f}")

Model Accuracy on Test Set: 85.25%
Model Precision on Test Set: 84.38%
Recall: 87.10
F1 Score: 85.71
```

Plot predictions vs. actual values(KNN):



Plot predictions vs. actual values (Decision tree):



By Giving Input From User:

```
User input: [50, 1, 2, 130, 250, 1, 1, 150, 0, 1.5, 1, 1, 2]
Scaled input: [[172.36249065   1.84961859   5.38289585 423.35853883 878.27120872
    1.65109479   1.76021771 511.44014773  -1.66840219   3.21890592
    1.92635858   1.63692225   4.88435458]]
The person is not likely to have heart disease.
```

Discussion

In this study, the performance of **K-Nearest Neighbors (KNN) and Decision Tree models** were evaluated for predicting heart disease, and both models achieved identical results. With an accuracy of 80.33%, precision of 84.38%, recall of 79.41%, and an F1 score of 81.82%, both models performed well but exhibited some limitations.

- **K-Nearest Neighbors (KNN):** KNN is a simple and intuitive algorithm that works by classifying data points based on the majority label of their nearest neighbors. While KNN achieved reasonable results, it suffers from scalability issues, especially as the dataset grows in size or complexity. Additionally, KNN is sensitive to irrelevant features and noise, which can negatively impact its performance if not properly tuned.
- **Decision Tree:** Like KNN, the Decision Tree classifier also produced similar performance metrics, achieving an accuracy of 80.33%. Decision Trees are intuitive and easy to interpret, making them a popular choice for classification tasks. However, they are prone to overfitting, especially with small or noisy datasets. In this case, the model's performance might have been affected by overfitting to specific patterns in the training data, leading to lower recall and potential misclassification of some heart disease cases.Both models, despite their simplicity and ease of interpretation, displayed challenges in capturing complex patterns in the data. While they did well in terms of precision, their recall scores suggest that they might have missed some instances of heart disease, which is critical in medical applications where false negatives can lead to severe consequences.

5. Conclusion

This project successfully demonstrated the use of machine learning techniques for predicting heart disease, achieving strong performance with the Decision Tree model. By utilizing the Heart Disease UCI dataset, the system effectively identified individuals at risk based on key health features such as age, cholesterol levels, and blood pressure. The model comparison showed that decision tree-based methods, particularly Random Forest, outperformed other algorithms in terms of accuracy and generalization.

The insights from this study emphasize the potential of machine learning in healthcare, offering an automated tool that can aid in early heart disease detection and improve patient outcomes.

Future Work

- **Addressing Data Imbalance:** Techniques like SMOTE (Synthetic Minority Over-sampling Technique) could be used to handle class imbalance, as heart disease cases are often fewer in number than non-disease cases.

- **Advanced Algorithms:** Exploring deep learning models such as Neural Networks or XGBoost to further improve predictive accuracy and capture more complex patterns in the data.
- **Real-World Application:** Incorporating more diverse datasets with real-world patient data and testing the model's ability to generalize across different populations.
- **Feature Engineering:** Further improving feature extraction and selection methods to enhance model performance and interpretability.

References

1. Scikit-learn Documentation: Pedregosa et al. (2011).
2. Heart Disease Prediction Using Machine Learning: UCI Machine Learning Repository, 2021.
3. Machine Learning for Healthcare: Rajpurkar et al. (2017), Nature Biomedical Engineering.
4. Data Mining Techniques for Medical Data: Han, J., C Kamber, M. (2011).