**Artificial Intelligence**

**Group project**

**Topic: Heart Disease Prediction Using Machine Learning to Identify Risk Factors and Improve Diagnosis.**

**Team Members:**

1. **Sravani Seelam (700739823)**
2. **Santhoshini Arkatala (700731563)**
3. **Sai Kumar Reddy Kambam (700739712)**
4. **Samyuktha cherala (700739890)**
5. **Haritha Danada (700735687)**

**Introduction:**

Heart disease is a major health concern worldwide and a leading cause of death. Machine learning can help in predicting heart diseases with high accuracy, based on patient data and other relevant factors. The Heart-Disease-Prediction-using-Machine-Learning project aims to use machine learning algorithms to predict the presence of heart disease in a patient. The project involves analysis of a heart disease patient dataset with proper data processing, followed by training and testing of different machine learning algorithms.

The project uses a variety of machine learning algorithms, including Logistic Regression, Naive Bayes, SVM, KNN, Decision Tree, Random Forest, XGBoost, and Artificial Neural Network with 1 Hidden layer. The target variable is a binary variable, predicting whether heart disease is present or not. The project achieved an accuracy of 90.16% with the Random Forest algorithm.

The project has important implications for improving heart disease prediction and prevention and can potentially save many lives.

**Problem Detection:**

Heart disease is a major health problem that affects millions of people worldwide. It is caused by a variety of factors, including genetics, lifestyle, and environmental factors. Early detection and prevention of heart disease can significantly improve people's health and quality of life. The Heart-Disease-Prediction-using-Machine-Learning project aims to use machine learning algorithms to predict the presence of heart disease in a patient. The project uses a dataset of heart disease patients, which includes various parameters such as age, sex, blood pressure, cholesterol level, etc.

The goal is to develop a machine learning model that can accurately predict whether a patient has heart disease or not, based on these parameters. This is a binary classification problem, where the target variable is a binary variable predicting whether heart disease is present or not.

The project has important implications for improving heart disease prediction and prevention and can potentially save many lives.

**AI Algorithms and Techniques Used:**

The Heart-Disease-Prediction-using-Machine-Learning project uses a variety of machine learning algorithms for the classification task.Logistic Regression, Naive Bayes, SVM, KNN, Decision Tree, Random Forest, XGBoost, and Artificial Neural Network with 1 Hidden layer are the algorithms used in the project.

1. **Logistic Regression**:   
   Logistic regression is a simple and widely used classification algorithm. It is a linear model that uses a logistic function to model the probability of the target variable. It is easy to implement and interpret, but it may not perform well when the relationship between the input features and target variable is non-linear.
2. **Naive Bayes**:   
   Naive Bayes is a probabilistic classification algorithm that is based on Bayes' theorem. It assumes that the input features are independent of each other, given the target variable. Naive Bayes is fast and requires a small amount of training data, but it may not perform well when the independence assumption is violated.
3. **Support Vector Machine (SVM)**:   
   SVM is a popular classification algorithm that finds the best hyperplane to separate the data into different classes. It can handle non-linearly separable data by using kernel functions. SVM is effective in high-dimensional spaces, but it may not perform well when the data is noisy, or the classes are overlapping.
4. **K-Nearest Neighbors (KNN)**:   
   KNN is a non-parametric classification algorithm that classifies new data points based on the class of their nearest neighbors in the training data. KNN is simple and easy to implement, but it may not perform well when the number of input features is large, or the data is imbalanced.
5. **Decision Tree**:   
   Decision tree is a popular classification algorithm that builds a tree-like model of decisions and their possible consequences. It is easy to interpret and can handle both categorical and numerical input features. However, decision trees may overfit the training data and may not generalize well to new data.
6. **Random Forest**:   
   Random Forest is an ensemble learning method that builds multiple decision trees and combines their predictions. It reduces overfitting and improves the accuracy of the predictions. Random forest is easy to use and can handle high-dimensional input features, but it may not perform well when the input features are highly correlated.
7. **XGBoost**:   
   XGBoost is a gradient boosting algorithm that uses decision trees as base learners. It is a powerful and scalable algorithm that can handle large datasets and high-dimensional input features. XGBoost is effective in reducing bias and variance, but it may require more computational resources than other algorithms.

Artificial Neural Network with 1 Hidden layer is a deep learning algorithm that uses a layered structure of artificial neurons to learn patterns in the input data.

Each algorithm has its strengths and weaknesses, and the choice of algorithm depends on the specific requirements of the application and the characteristics of the data.

**Comparison with Other AI Algorithms:**

The heart disease prediction using machine learning project uses a diverse set of machine learning algorithms for the classification task.

Here is the comparison of the algorithms used in the project with other popular algorithms in the same domain.

**Gradient boosting**: it is an ensemble learning method that combines multiple weak learners to create a strong learner. It is like a random forest, but it builds trees sequentially and adjusts the weights of misclassified samples.

**Support vector machines with nonlinear kernel**: SVM can handle nonlinearly separable data by using kernel functions. On linear kernels such as polynomial and radial basis functions can capture complex relationships between input features and target variables.

**Neural networks with more hidden layers**: Deep neural networks with multiple hidden layers can learn more complex and abstract features from the input data. They can potentially improve the accuracy of the predictions, but they require more data and computational resources.

The choice of algorithms depends on the specific requirement of the applications and the characteristics of the data.

**Diagram Explanation:**

**Data Processing Steps Diagram**

The data processing steps diagram shows the different steps involved in processing the heart disease patient dataset for Heart Disease Prediction project. The diagram starts with loading the dataset, which is followed by data cleaning, data preprocessing, data splitting, feature scaling, feature selection, model training, model evaluation, model tuning, and finally, the final model.

In the data cleaning step, the dataset is checked for any missing or inconsistent data, and any such data is removed or corrected. In the data preprocessing step, the dataset is prepared for use in machine learning algorithms by converting categorical variables to numerical variables, and scaling the features to a common range. In the data splitting step, the dataset is split into training and testing sets to evaluate the performance of the machine learning algorithms.

In the feature scaling step, the features are scaled to a common range to ensure that no feature has a disproportionate impact on the results. In the feature selection step, the most relevant features are selected for use in the machine learning algorithms to improve their performance. In the model training step, the machine learning algorithms are trained on the training data to predict the presence of heart disease.

In the model evaluation step, the performance of the machine learning algorithms is evaluated on the testing data using metrics such as accuracy, precision, recall, and F1-score. In the model tuning step, the hyperparameters of the machine learning algorithms are adjusted to improve their performance. Finally, in the final model step, the best performing machine learning algorithm is selected for use in predicting the presence of heart disease in a patient.

**Diagram

Description automatically generated**

**Machine Learning Algorithms Comparison Diagram**

The machine learning algorithms comparison diagram shows a pie chart that compares the different machine learning algorithms used in Heart Disease Prediction project. The pie chart is divided into eight segments, each representing a different machine learning algorithm. The size of each segment represents the proportion of the dataset that was correctly classified by that algorithm.

The eight machine learning algorithms used in the project are logistic regression, naive Bayes, support vector machine, k-nearest neighbors, decision tree, random forest, XGBoost, and artificial neural network. The pie chart shows that the random forest algorithm achieved the highest accuracy, correctly classifying 30% of the dataset. The decision tree algorithm achieved the second-highest accuracy, correctly classifying 20% of the dataset. The remaining algorithms achieved accuracies ranging from 10% to 15%.

The machine learning algorithms comparison diagram is useful for comparing the performance of different machine learning algorithms and selecting the best algorithm for a given problem. In the case of Heart Disease Prediction project, the random forest algorithm was selected as the best performing algorithm for predicting the presence of heart disease in a patient..

**Chart, pie chart

Description automatically generated**

**System Components Diagram**

The system components diagram shows the different components of Heart Disease Prediction system and how they are related to each other. The diagram consists of four main components: the heart disease prediction application, data processing, machine learning algorithms, and model evaluation.

The heart disease prediction application is the main component of the system and is responsible for predicting the presence of heart disease in a patient. It uses the machine learning algorithms and the final model to make predictions based on input features.

The data processing component is responsible for cleaning, preprocessing, and splitting the heart disease patient dataset for use in machine learning algorithms. It also performs feature scaling and feature selection to improve the performance of the machine learning algorithms.

The machine learning algorithms component consists of eight different algorithms that are used to predict the presence of heart disease. These algorithms include logistic regression, naive Bayes, support vector machine, k-nearest neighbors, decision tree, random forest, XGBoost, and artificial neural network.

The model evaluation component is responsible for evaluating the performance of the machine learning algorithms using metrics such as accuracy, precision, recall, and F1-score. It selects the best performing algorithm for use in the heart disease prediction application.

The system components diagram is useful for understanding the different components of Heart Disease Prediction system and how they work together to predict the presence of heart disease in a patient. It provides a high-level overview of the system architecture and can be used to identify areas for improvement and optimization.

Diagram

Description automatically generated

**Result screenshots:**

Below pictures show the accuracies obtained for the algorithms.

Text

Description automatically generated

**Results and Conclusion:**

The Heart-Disease-Prediction-using-Machine-Learning project achieved an accuracy of 90.16% with the Random Forest algorithm. The project demonstrates the effectiveness of machine learning algorithms in predicting heart disease and improving patient outcomes.

However, the project has some limitations and future work can be done to improve the accuracy and generalization of the model. The project can potentially save many lives by enabling early detection and prevention of heart disease.

In conclusion, the project highlights the importance of machine learning in healthcare and the potential for AI to improve patient outcomes.

**References:**

1. The Heart-Disease-Prediction-using-Machine-Learning project used various resources and references for its implementation and analysis.
2. Some of the references used in the project include:
   1. UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Heart+Disease
   2. Scikit-learn documentation: https://scikit-learn.org/stable/documentation.html
   3. Keras documentation: https://keras.io/documentation/
3. The project also acknowledges the contributions of individuals or organizations who helped in the project implementation and analysis.
4. Some of the acknowledgments include:
   1. Kaggle community for providing the dataset and platform for analysis
   2. Open-source libraries and frameworks used in the project
   3. Mentors, advisors, or colleagues who provided guidance and support throughout the project