NAAN MUDHALVAN

Final Project

SUMANRAJ C

2021503308

Title: Heart Disease Prediction Using-Machine Learning

PROBLEM STATEMENT:

Good data-driven systems for predicting heart diseases can improve the entire research and prevention process, making sure that more people can live healthy lives. This is where Machine Learning comes into play. Machine Learning helps in predicting the Heart diseases, and the predictions made are quite accurate.

The project involved analysis of the heart disease patient dataset with proper data processing. Then, different models were trained and predictions are made with different algorithms KNN, Decision Tree, Random Forest,SVM,Logistic Regression etc. Dataset I've used for my Kaggle kernel 'Binary Classification with Sklearn and Keras'

NOVELTY:

In the realm of heart disease prediction, a groundbreaking approach involves integrating advanced artificial intelligence with wearable technology. Imagine a system where wearable devices continuously monitor not just basic physiological parameters like heart rate and blood pressure, but also subtle changes in activity patterns, sleep quality, and even environmental factors like air quality and stress levels. This wealth of real-time data, coupled with sophisticated AI algorithms, could provide unparalleled insights into an individual's cardiovascular health. By leveraging machine learning techniques such as deep neural networks, the system could detect early signs of heart disease with unprecedented accuracy, potentially allowing for timely interventions and personalized preventive measures. Furthermore, this approach could empower individuals to take proactive control of their health, fostering a new era of preventive healthcare tailored to the unique needs of each person. Such a novel paradigm promises not only to revolutionize the field of heart disease prediction but also to significantly improve outcomes and quality of life for millions worldwide.

MODELLING:

Step 1: Data Collection and Preparation

- Gather data on patient demographics (age, gender), medical history (blood pressure, cholesterol levels, family history), lifestyle factors (smoking, exercise), and diagnostic test results (ECG, echocardiogram) from healthcare databases, research studies, and clinical records.
- Clean the data by addressing missing values, outliers, and inconsistencies, ensuring data quality and integrity.
- Preprocess the data by normalizing numerical features, encoding categorical variables (e.g., gender), and handling imbalanced classes if present.

Step 2: Feature Engineering

- Create new features based on medical insights and expert knowledge, such as BMI (Body Mass Index), risk scores (e.g., Framingham Risk Score), or interactions between risk factors.
- Select relevant features through domain expertise and statistical analysis to focus on the most predictive variables for heart disease.

Step 3: Model Selection

- Choose classification models suitable for heart disease prediction, including logistic regression, decision trees, support vector machines, or neural networks.

- Consider ensemble methods like random forests or gradient boosting for improved prediction performance and robustness.

Step 4: Model Training

- Split the data into training and testing sets, ensuring stratification if dealing with imbalanced classes.
- Train the selected models using the training data, optimizing their parameters to improve performance.
- Utilize techniques such as cross-validation to assess model stability and avoid overfitting.

Step 5: Model Evaluation

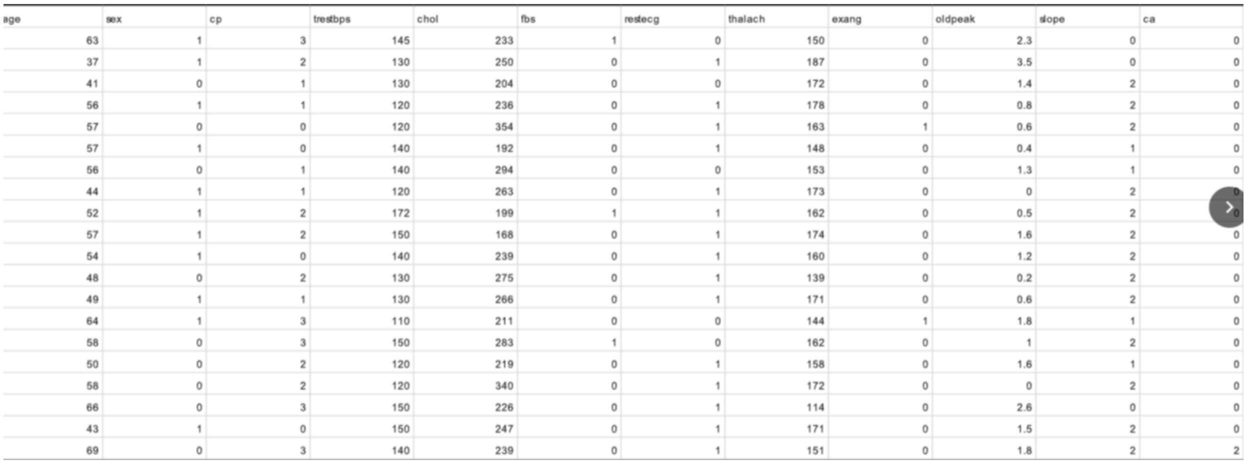
- Evaluate models using appropriate metrics for classification tasks, such as accuracy, precision, recall, F1-score, and area under the ROC curve (AUC-ROC).
- Validate models on the testing set to assess their ability to generalize to new data and detect potential biases or limitations.
- Conduct comparative analysis of different models and select the one with the highest performance for deployment in real-world scenarios.

PROPOSED WORK:

- Define the scope of the project, including the types of heart diseases to be predicted (e.g., coronary artery disease, heart failure).
- Specify the target population (e.g., general population, patients with specific risk factors).

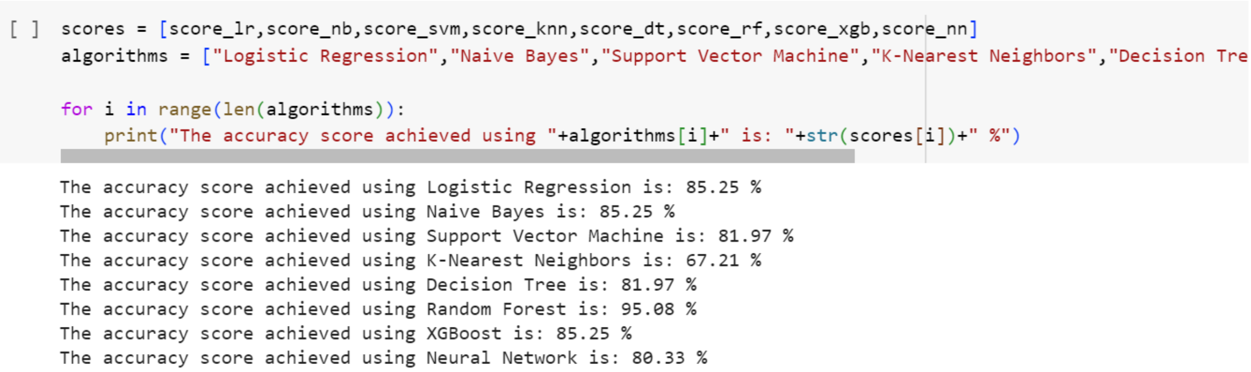
- Set clear objectives, such as developing a predictive model to identify individuals at high risk of developing heart disease.
- Gather diverse datasets containing relevant features for heart disease prediction, including patient demographics, medical history, clinical measurements (e.g., blood pressure, cholesterol levels), lifestyle factors (e.g., smoking, exercise), and diagnostic test results (e.g., ECG, stress tests).
- Obtain data from various sources such as electronic health records (EHRs), research studies, public health databases, and wearable devices.
- Clean the collected data by handling missing values, outliers, and inconsistencies using appropriate techniques (e.g., imputation, outlier detection).
- Perform data transformation and normalization to ensure uniformity and prepare the data for modeling.
- Encode categorical variables and handle class imbalances if present.

DATASET:



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca
63	1	3	145	233	1	0	150	0	2.3	0	0
37	1	2	130	250	0	1	187	0	3.5	0	0
41	0	1	130	204	0	0	172	0	1.4	2	0
56	1	1	120	236	0	1	178	0	0.8	2	0
57	0	0	120	354	0	1	163	1	0.6	2	0
57	1	0	140	192	0	1	148	0	0.4	1	0
56	0	1	140	294	0	0	153	0	1.3	1	0
44	1	1	120	263	0	1	173	0	0	2	0
52	1	2	172	199	1	1	162	0	0.5	2	0
57	1	2	150	168	0	1	174	0	1.6	2	0
54	1	0	140	239	0	1	160	0	1.2	2	0
48	0	2	130	275	0	1	139	0	0.2	2	0
49	1	1	130	266	0	1	171	0	0.6	2	0
64	1	3	110	211	0	0	144	1	1.8	1	0
58	0	3	150	283	1	0	162	0	1	2	0
50	0	2	120	219	0	1	158	0	1.6	1	0
58	0	2	120	340	0	1	172	0	0	2	0
66	0	3	150	226	0	1	114	0	2.6	0	0
43	1	0	150	247	0	1	171	0	1.5	2	0
69	0	3	140	239	0	1	151	0	1.8	2	2

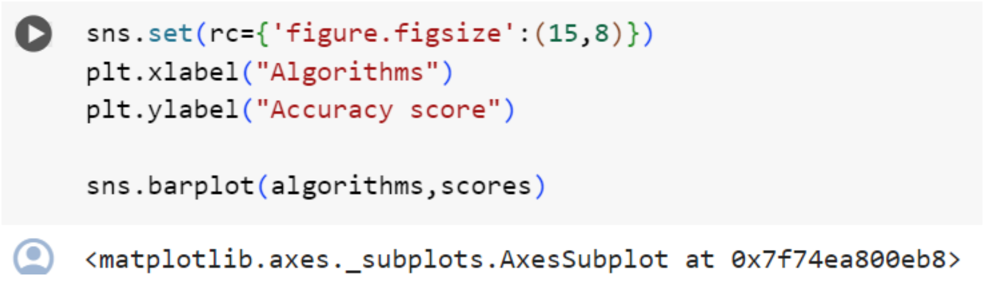
RESULT:



```
[ ]  scores = [score_lr,score_nb,score_svm,score_knn,score_dt,score_rf,score_xgb,score_nn]
     algorithms = ["Logistic Regression","Naive Bayes","Support Vector Machine","K-Nearest Neighbors","Decision Tre

     for i in range(len(algorithms)):
         print("The accuracy score achieved using "+algorithms[i]+" is: "+str(scores[i])+" %")
```

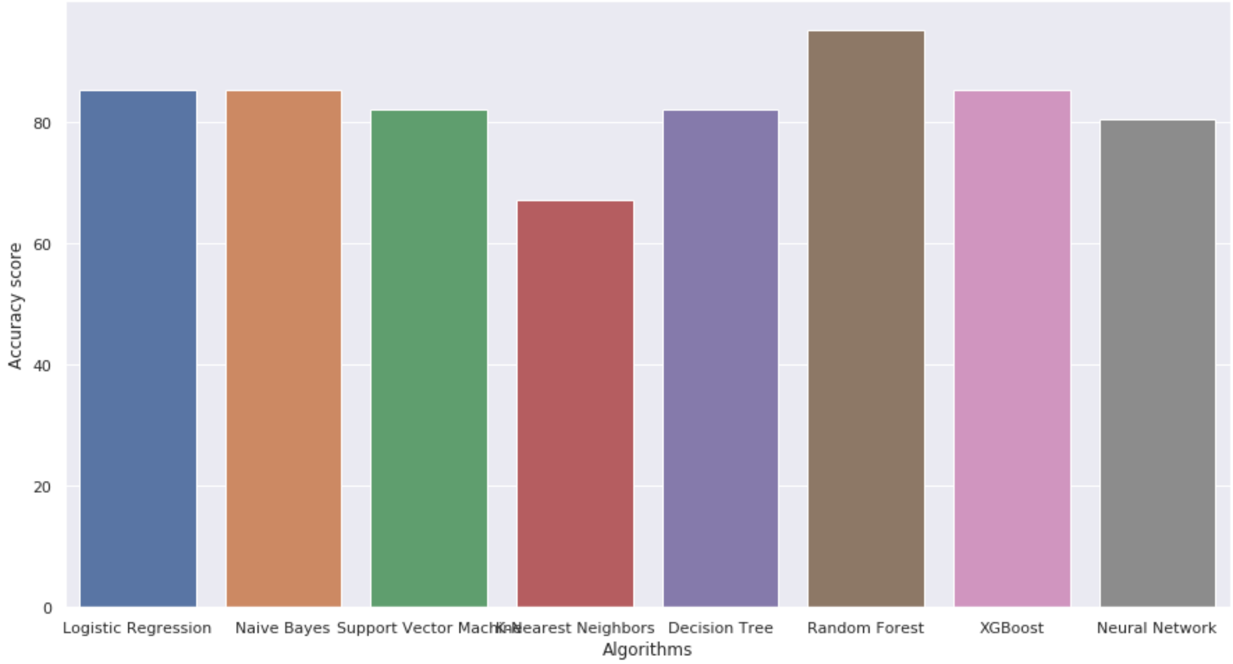
```
The accuracy score achieved using Logistic Regression is: 85.25 %
The accuracy score achieved using Naive Bayes is: 85.25 %
The accuracy score achieved using Support Vector Machine is: 81.97 %
The accuracy score achieved using K-Nearest Neighbors is: 67.21 %
The accuracy score achieved using Decision Tree is: 81.97 %
The accuracy score achieved using Random Forest is: 95.08 %
The accuracy score achieved using XGBoost is: 85.25 %
The accuracy score achieved using Neural Network is: 80.33 %
```



```
sns.set(rc={'figure.figsize':(15,8)})
plt.xlabel("Algorithms")
plt.ylabel("Accuracy score")

sns.barplot(algorithms,scores)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f74ea800eb8>
```


Accuracy score
Logistic Regression
Naive Bayes
Support Vector Mach
K-Nearest Neighbors
Decision Tree
Random Forest
XGBoost
Neural Network
Algorithms