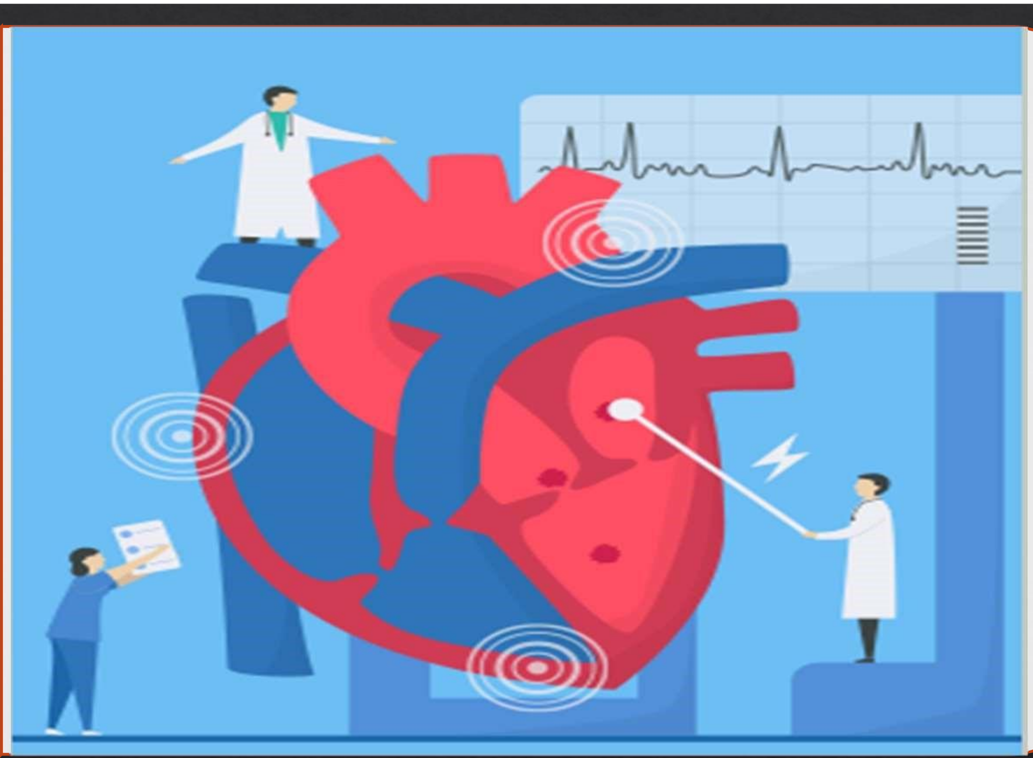
Predicting Heart Disease



TEAM MEMPERS

1- Ahmed Abdelaal Ahmed Abohadeed	(Section1)
2- Ahmed Mohammed Gouda	(Section 1)
3- Ali Nabil Ali	(Section 4)
4- Alsayed sameh alsayed	(Section 1)
5- Idris Tarek Elsayed Hashiesh	(Section 1)
6-Shereen ebrahiem ebrahiem saad	(Section 3)
7- Rania Yasser Mohamed	(Section 3)
8- Khloud Elsayed Elsayed Mohamed	(Section 3)
9- Eman hamada abdelhady mohamed	(Section 2)

Introduction

- Heart disease is a prevalent and critical health issue that affects millions of people worldwide. It encompasses a range of conditions that affect the heart's structure and function, including coronary artery disease, heart failure, and arrhythmias. According to the World Health Organization, heart disease is the leading cause of death globally, accounting for a significant burden on healthcare systems and societies.The impact of heart disease extends beyond mortality rates. It can significantly impair an individual's quality of life, leading to limitations in physical activity, increased healthcare costs, and decreased productivity. Therefore, early detection and accurate prediction of heart disease play a crucial role in managing this condition effectively.

- Machine learning, a subfield of artificial intelligence, has emerged as a powerful tool in healthcare for predicting and diagnosing various diseases, including heart disease. By leveraging the vast amounts of available patient data and applying advanced algorithms, machine learning models can analyze complex patterns and relationships in the data to make predictions.

- In this presentation, we will explore how machine learning techniques can be utilized to predict the presence or absence of heart disease. We will discuss the dataset used, the preprocessing steps taken to prepare the data, the selection and engineering of relevant features, and the evaluation of different machine learning algorithms. By harnessing the power of machine learning, we aim to contribute to the early detection and accurate prediction of heart disease, ultimately improving patient care and global health outcomes.

Dataset Description:

. The dataset used for predicting heart disease is sourced from Kaggle and can be found at the following link: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

. This dataset contains information on various personal key indicators that are potentially associated with heart disease. Let's explore the dataset in more detail:

Number of Rows: The dataset consists of 319,795 rows, indicating the presence of 319,795 records or instances in the dataset.

Number of Columns: There are 18 columns in the dataset, providing 18 different features that can be used to predict heart disease.

List of Features:

1- HeartDisease: This is the target variable indicating the presence or absence of heart disease. It serves as the label for our machine learning model.

2- BMI: Body Mass Index, a measure of body fat based on height and weight.

3- Smoking: Indicates whether the individual is a smoker or not.

4- AlcoholDrinking: Indicates whether the individual consumes alcohol or not.

5- Stroke: Indicates whether the individual has had a stroke or not.

6- PhysicalHealth: Represents the individual's overall physical health condition.

7- MentalHealth: Represents the individual's mental health condition.

9- DiffWalking: Indicates any difficulties in walking experienced by the individual.

10- Sex: Gender of the individual.

10.AgeCategory: Age category or range in which the individual belongs.

11.Race: Ethnicity or race of the individual.

12.Diabetic: Indicates whether the individual has diabetes or not.

13.PhysicalActivity: Level of physical activity engaged in by the individual.

14.GenHealth: Represents the individual's general health condition.

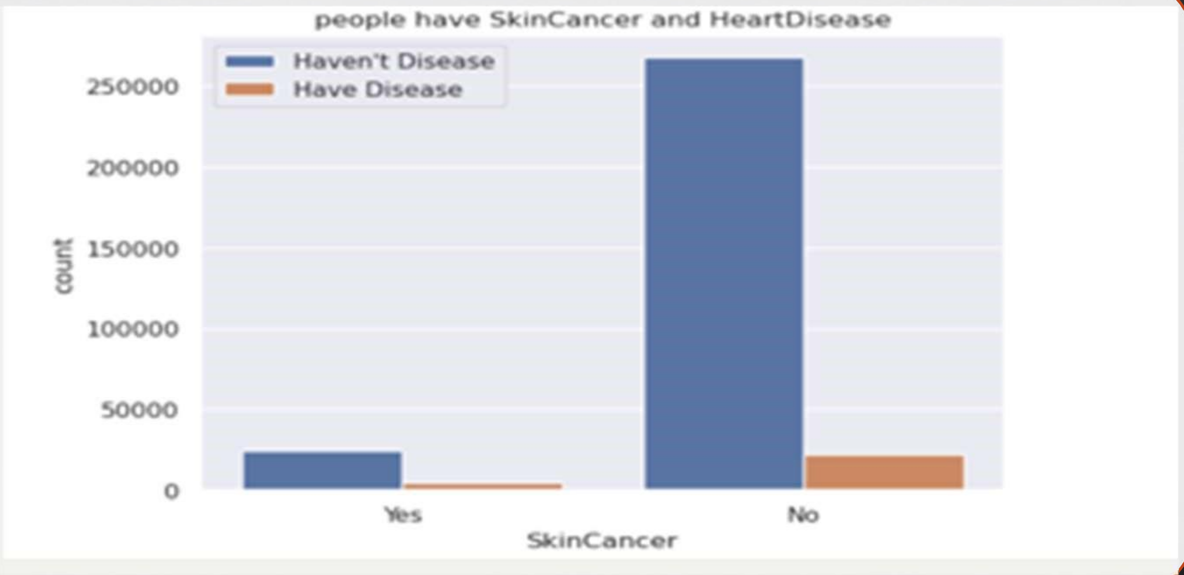
15.SleepTime: Amount of sleep time the individual typically gets.

16.Asthma: Indicates whether the individual has asthma or not.

17.KidneyDisease: Indicates whether the individual has kidney disease or not.

18.SkinCancer: Indicates whether the individual has skin cancer or not.

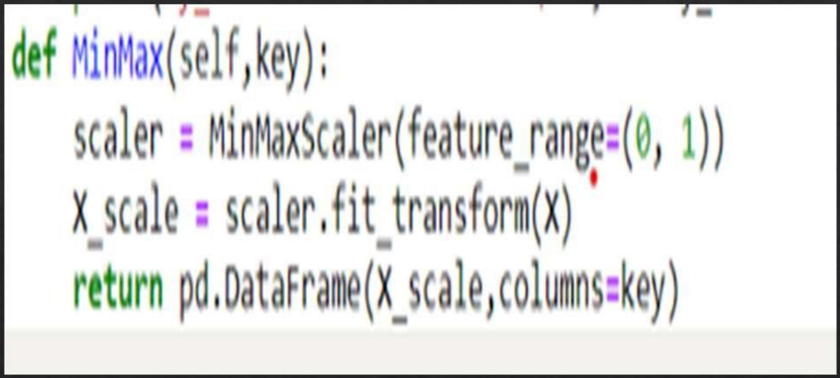
This dataset provides a diverse set of features related to personal key indicators that may have an impact on heart disease. By leveraging this dataset and applying machine learning algorithms, we can build a predictive model to identify patterns and relationships between these features and the presence or absence of heart disease.



Data Preprocessing:

Data preprocessing is an essential step in preparing the dataset for a machine learning model. It involves transforming and cleaning the data to ensure optimal performance and accurate predictions. Let's describe the steps taken to preprocess the data for our heart disease prediction model:

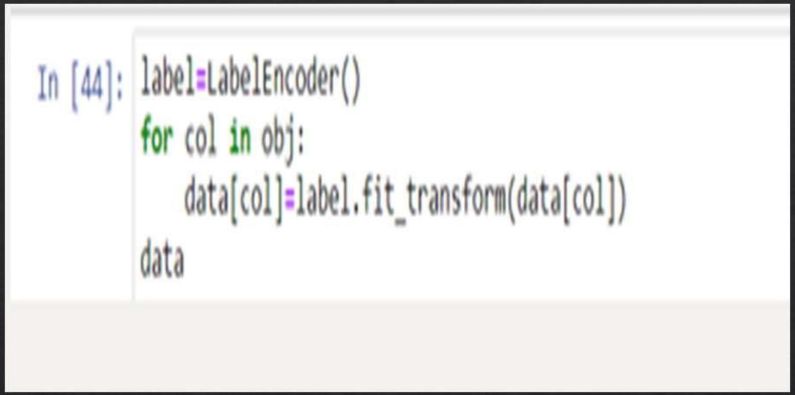
1.Data Scaling: To normalize the features and bring them to a similar scale, we can use data scaling techniques such as Min-Max scaling. The following code demonstrates the use of MinMaxScaler to scale the features:



```
def MinMax(self,key):
    scaler = MinMaxScaler(feature_range=(0, 1))
    X_scale = scaler.fit_transform(X)
    return pd.DataFrame(X_scale,columns=key)
```

2. Label Encoding:

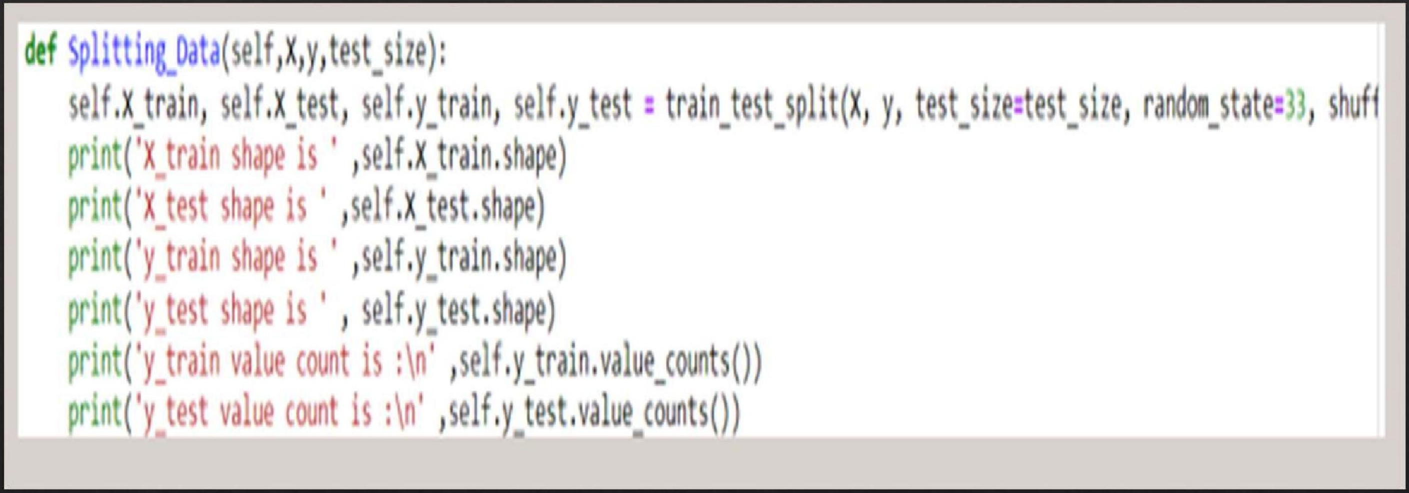
If any categorical variables are present in the dataset, they need to be encoded into numeric values for the machine learning algorithms to process them. Label encoding can be used to convert categorical variables into numerical labels. Here's an example of label encoding using LabelEncoder:



```
In [44]: label=LabelEncoder()
         for col in obj:
             data[col]=label.fit_transform(data[col])
         data
```

3.Data Splitting:

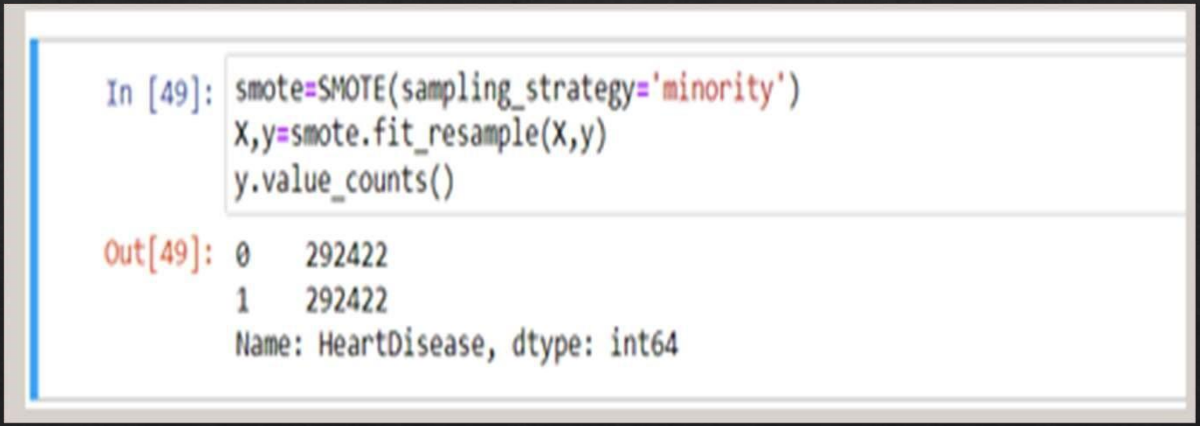
◈ It is crucial to split the dataset into training and testing sets to evaluate the model's performance. The following code demonstrates data splitting using train_test_split from scikit-learn:



```
def Splitting_Data(self,X,y,test_size):
    self.X_train, self.X_test, self.y_train, self.y_test = train_test_split(X, y, test_size=test_size, random_state=33, shuff
    print('X_train shape is ' ,self.X_train.shape)
    print('X_test shape is ' ,self.X_test.shape)
    print('y_train shape is ' ,self.y_train.shape)
    print('y_test shape is ' , self.y_test.shape)
    print('y_train value count is :\n' ,self.y_train.value_counts())
    print('y_test value count is :\n' ,self.y_test.value_counts())
```

4.Handling Imbalanced Data:

- If the dataset suffers from class imbalance, where one class has significantly fewer samples than the other, it can impact the model's performance. Techniques such as Synthetic Minority Over-sampling Technique (SMOTE) can be used to balance the data. Here's an example of balancing the data using SMOTE:

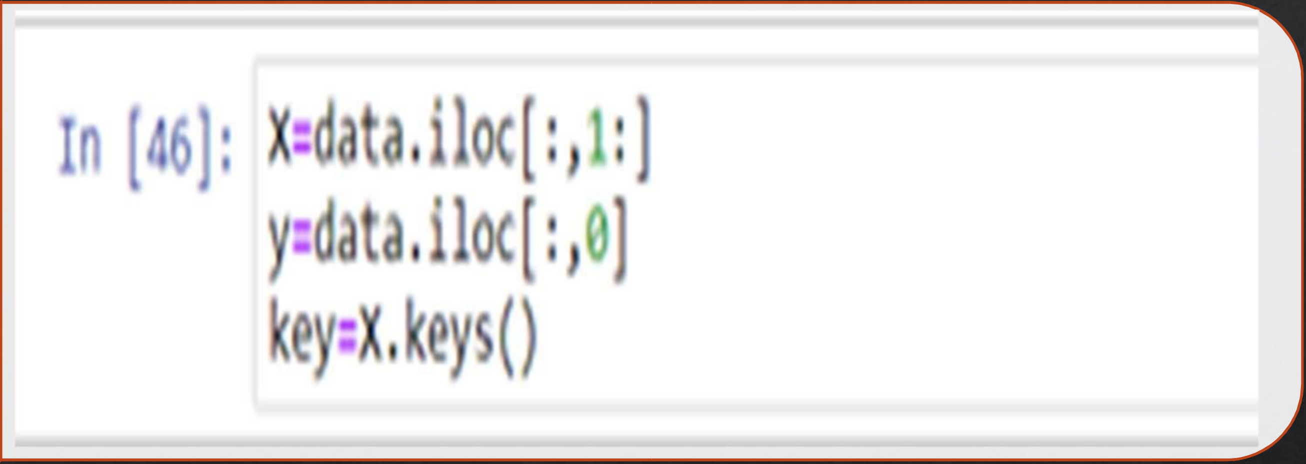


```
In [49]: smote=SMOTE(sampling_strategy='minority')
         X,y=smote.fit_resample(X,y)
         y.value_counts()

Out[49]: 0    292422
         1    292422
         Name: HeartDisease, dtype: int64
```

5.Feature Extraction:

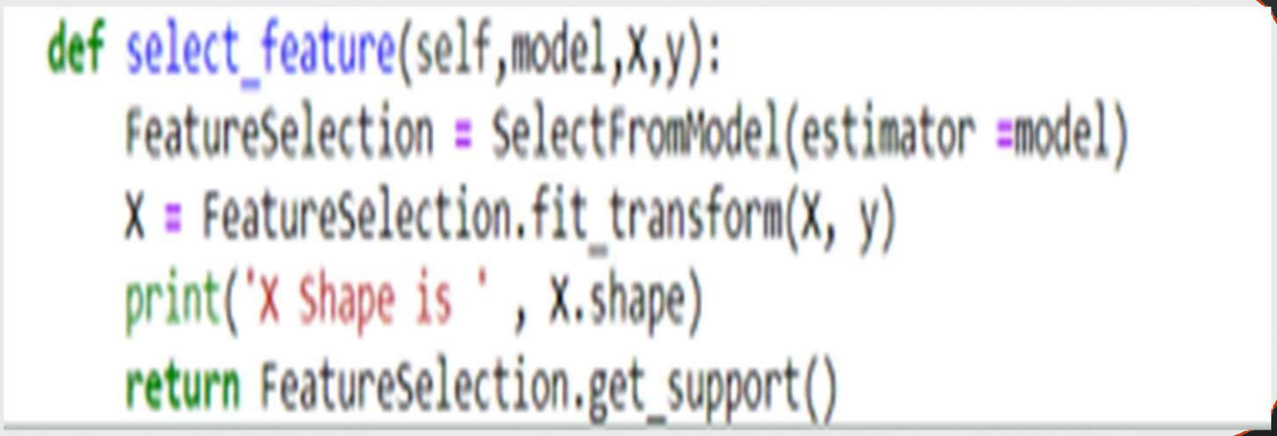
In feature extraction, we separate the target variable from the features. The following code extracts the features and target variable from the dataset:



```
In [46]: X=data.iloc[:,1:]
         y=data.iloc[:,0]
         key=X.keys()
```

Feature Selection and Engineering:

By applying feature selection techniques, we can enhance the model's performance by focusing on the most relevant features, reducing noise, and potentially mitigating overfitting.



```
def select_feature(self,model,X,y):
    FeatureSelection = SelectFromModel(estimator =model)
    X = FeatureSelection.fit_transform(X, y)
    print('X Shape is ' , X.shape)
    return FeatureSelection.get_support()
```

Machine Learning Algorithms:

- in machine learning, various algorithms can be employed to train models and make predictions based on the provided data. Each algorithm has its own characteristics and is suitable for different types of problems. Here are machine learning algorithms which we used in the model:
- 1. RandomForestClassifier
- 2. DecisionTreeClassifier
- 3. GradientBoostingClassifier
- 4.KNeighborsClassifier
- 5. XGBClassifier
- 6. VotingClassifier

Prediction:

The Evaluation metrics that are used to assess theperformance of our machine learning model:

Accuracy Score: Indicates the proportion of correctly classified instances:

Accuracy Score is : 0.9021304243750641

F1-score: The harmonic mean of precision and recall, which provides a balanced measure of model performance:

F1 Score is : 0.9037853829086263

Recall Score: Represents the ability of the model to correctly identify positive instances:

Recall Score is : 0.9171670305676856

Precision Score: Reflects the model's ability to correctly classify instances as positive:

Precision Score is : 0.8907886017229953

Classification Report: Provides precision, recall, F1-score, and support for each class, as well as macro-averaged and weighted-averaged metrics:

```
Classification Report is :          precision    recall  f1-score   support
0         0.91       0.89       0.90      14587
1         0.89       0.92       0.90      14656
accuracy                             0.90      29243
macro avg       0.90       0.90       0.90      29243
weighted avg       0.90       0.90       0.90      29243
```

Model Deployment and Use:

Once the machine learning model for predicting heart disease has been trained and evaluated, it can be deployed and utilized in real-world scenarios to assist in early detection and prevention efforts. Here are some key points to consider:

- Deployment of the Model:

Web or Mobile Applications: The trained model can be integrated into web or mobile applications that allow healthcare professionals or individuals to input relevant data and receive predictions on the likelihood of heart disease.

Electronic Health Records (EHR) Systems: The model can be integrated into existing EHR systems used by healthcare providers, enabling automated prediction and risk assessment for patients.

Remote Health Monitoring: In remote health monitoring systems, the model can analyze real-time patient data (e.g., vital signs, activity levels) and provide alerts or recommendations to healthcare providers or individuals.

- Utilization in Real-World Scenarios:

Early Detection and Risk Assessment: The model can help identify individuals who are at a higher risk of developing heart disease, allowing for targeted interventions and preventive measures.

Personalized Treatment Planning: By considering the predicted risk of heart disease, healthcare professionals can create personalized treatment plans, lifestyle recommendations, and interventions to manage and reduce the risk.

Public Health Planning: Aggregated predictions from the model can provide insights for public health planning, such as identifying high-risk populations or areas that require targeted interventions and awareness campaigns.

By deploying the trained model in real-world scenarios and leveraging its predictive capabilities, healthcare providers, policymakers, and individuals can benefit from early detection, personalized interventions, and improved planning for heart disease prevention and management. This can ultimately lead to better patient outcomes, reduced healthcare costs, and a healthier population.