PROJECT NAME:- MEDICAL INSURANCE COST PREDICTION

Submitted to Globesyn Finishing School

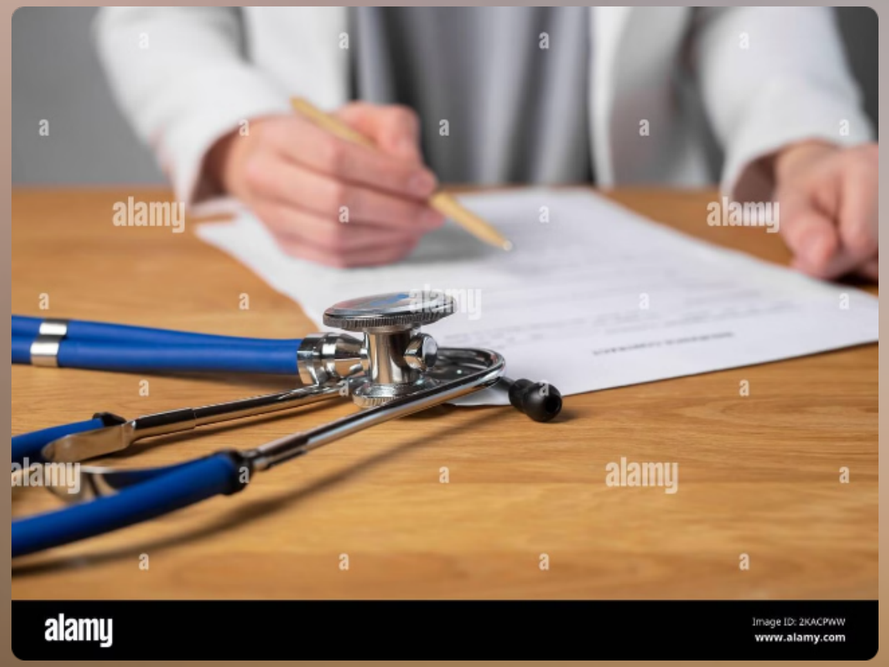
PRESENTED BY

AHANA MANDAL

BAJRANG KUMAR

Medical Insurance Cost Prediction

Predicting medical insurance costs is vital for individuals and insurance companies. By analysing various factors, we can develop models that accurately estimate insurance premiums, leading to fairer pricing and better healthcare management.



Understanding the Problem Statement

1 Accurate Prediction

The goal is to build a model that can accurately predict the cost of medical insurance for individuals based on their characteristics.

2 Fair Pricing

The model should ensure fair pricing by taking into account relevant factors that influence medical expenses.

3 Data-Driven Insights

Understanding the key drivers of insurance costs can provide valuable insights for healthcare providers and policy makers.

Libraries used to make this machine learning model

1. Numpy
2. Pandas
3. Matplotlib.pyplt
4. Linear regression from sklearn.linear_model
5. Seaborn
6. train_test_split from sklearn.mosel_selection

Data Collection and Preprocessing

1

Data Sources

Collect data from kaggle to do the further tasks.

2

Data Preprocessing

Preprocess the data by analyse it , describe it , defining the shape of the data, analysing the categorical values etc

3

Data Transformation

Transform the data into a format suitable for machine learning algorithms. This might involve scaling, normalisation, or encoding categorical variables.

Exploratory Data Analysis

1. Defining the head values
2. Defining the shape:- (1338,7)

It means it has 1338 rows and 7 columns

3. By using the info() key we get some information about the dataset
4. Defining the categorical values

   Here the categorical values are

i) sex

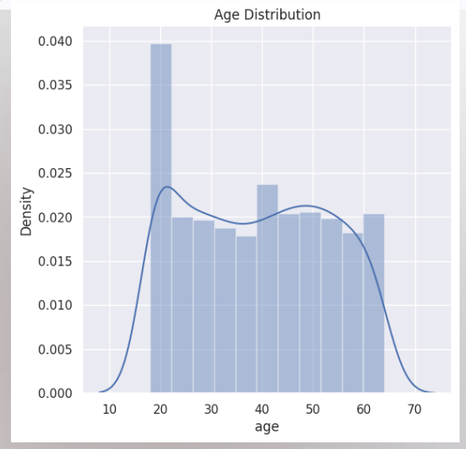
ii) Smoker

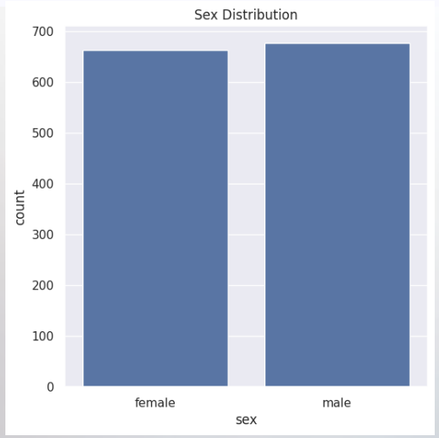
iii) Region

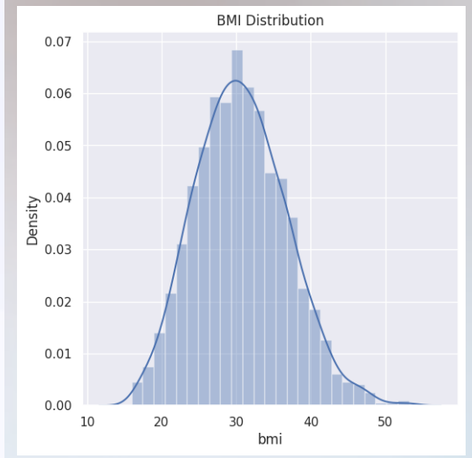
5. Checking the dataset if there is any null value

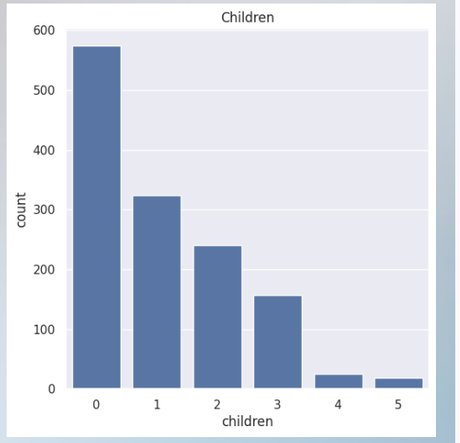
   By using the insull().sum() key we can check whether there is any null value or not

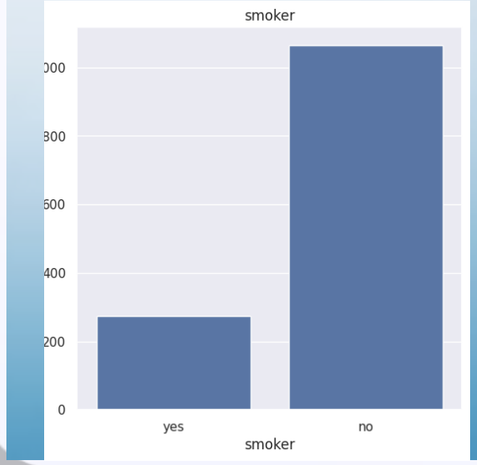
   >>In the dataset there is o null value.

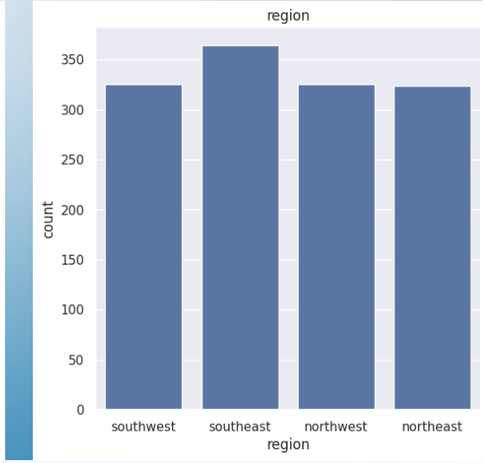












DATA ANALYSIS

Describing the Data

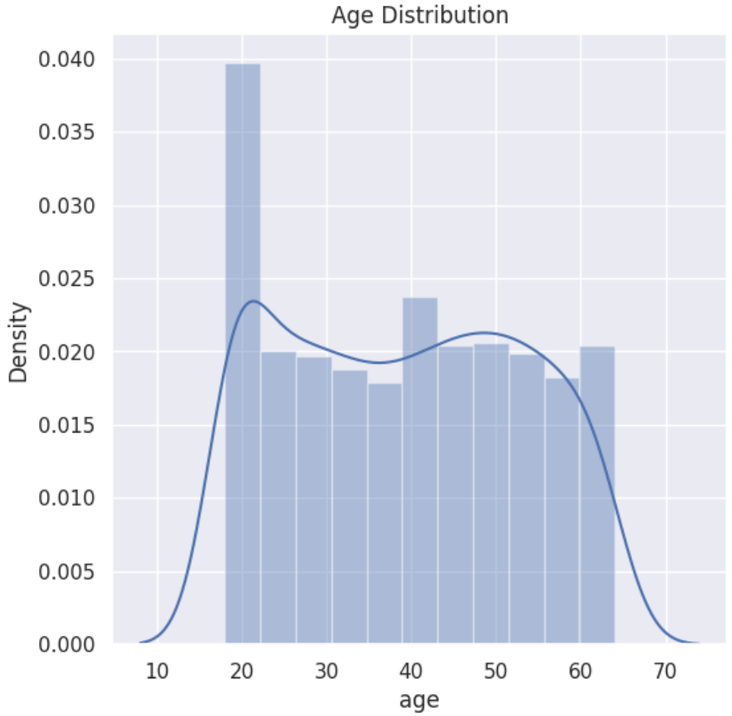
By using describe() key we can see the mean value, standard deviation and

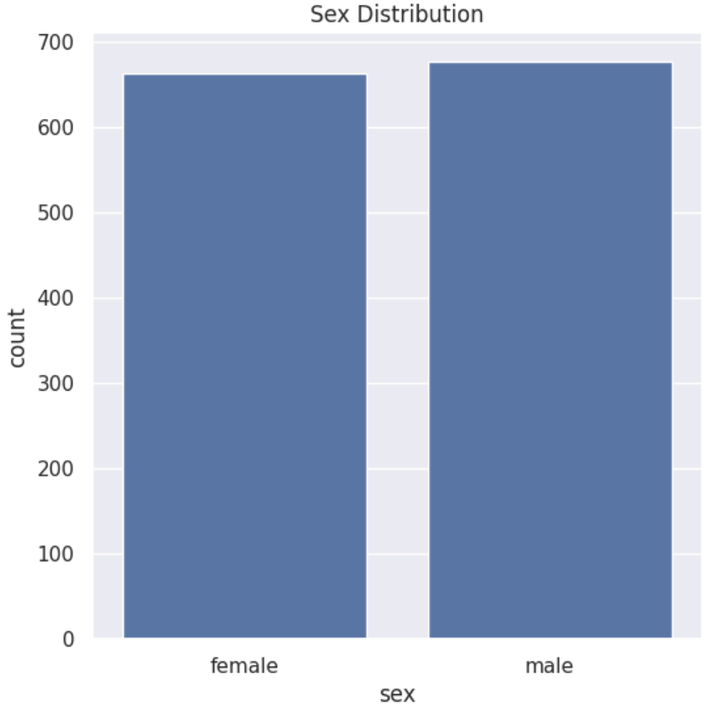
Feature Selection

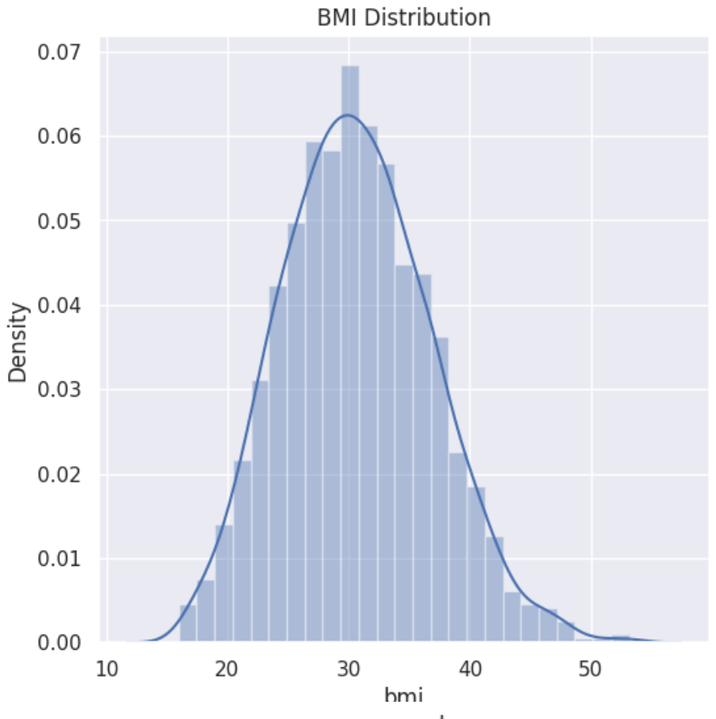
The most relevant features for this dataset are ‘age’,’sex’,’bmi’,’children’,’smoker’, ,’region’.

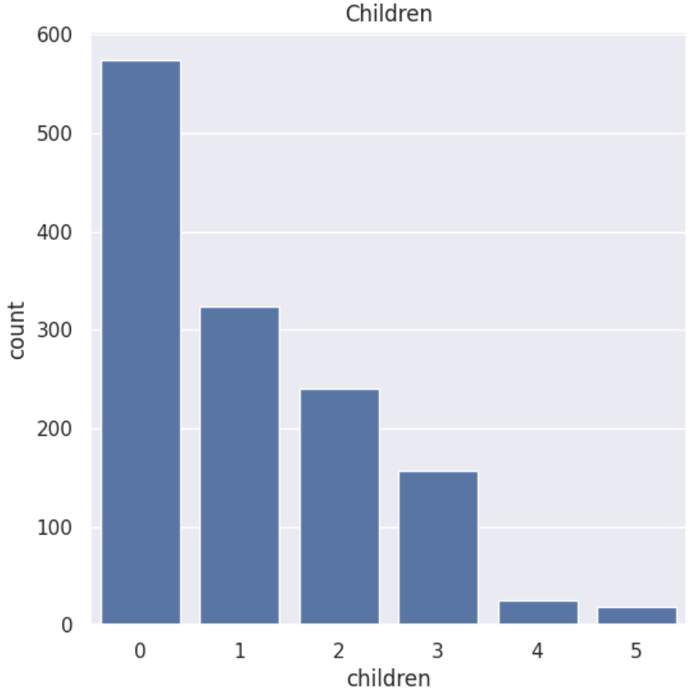
Feature Scaling

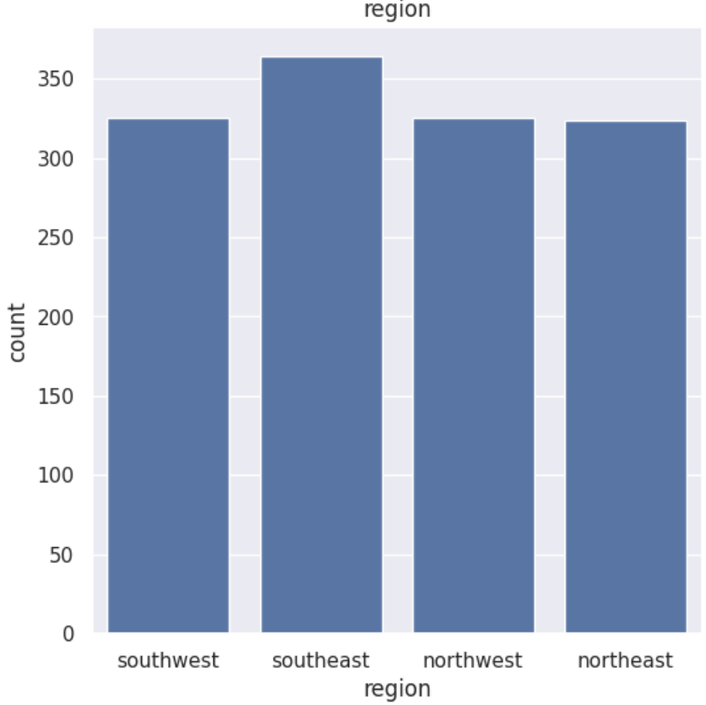
Scale the selected features to a common range to improve model performance. This helps ensure that different features contribute equally to the model.

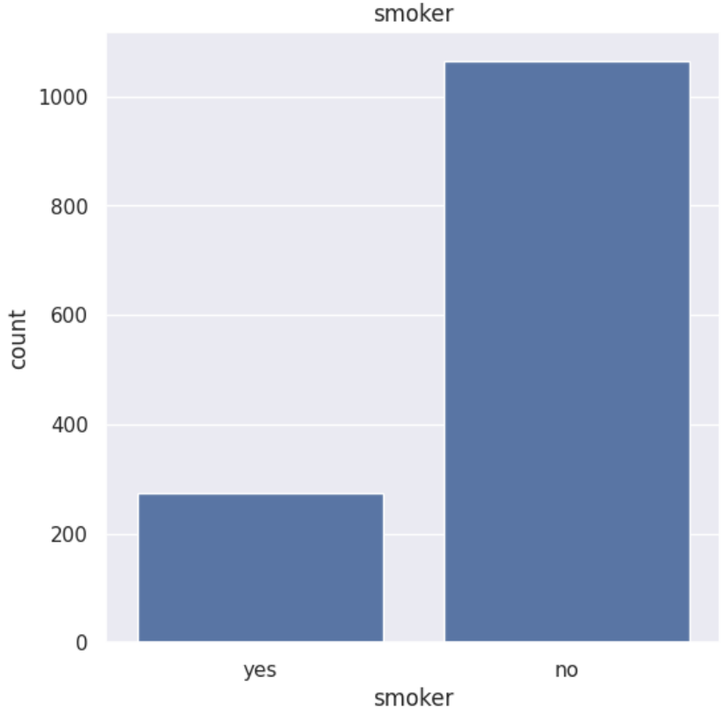

Age Distribution
Density
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000
10
20
30
40
50
60
70
age

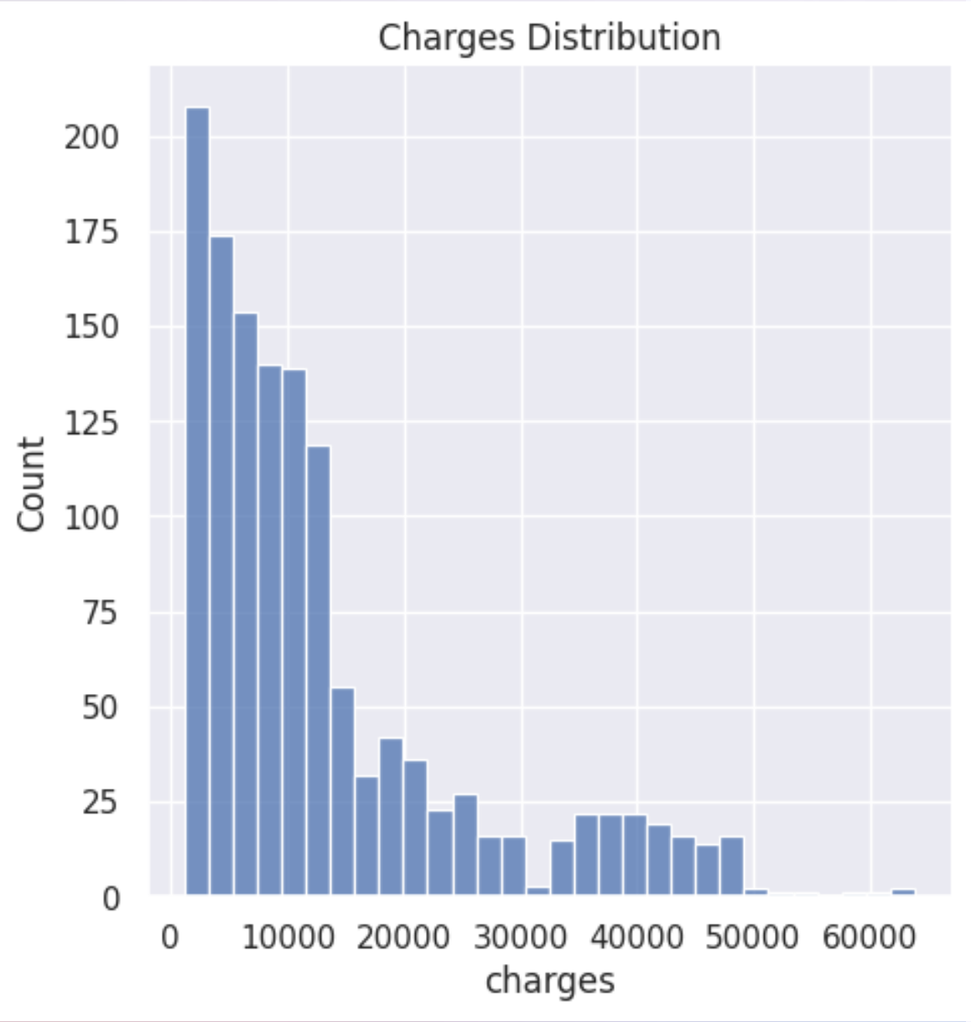

Sex Distribution
count
700
600
500
400
300
200
100
0
female
male
sex


BMI Distribution
Density
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
10
20
30
40
50
bmi


Children
count
600
500
400
300
200
100
0
0
1
2
3
4
5
children


region
count
350
300
250
200
150
100
50
0
southwest
southeast
northwest
northeast
region

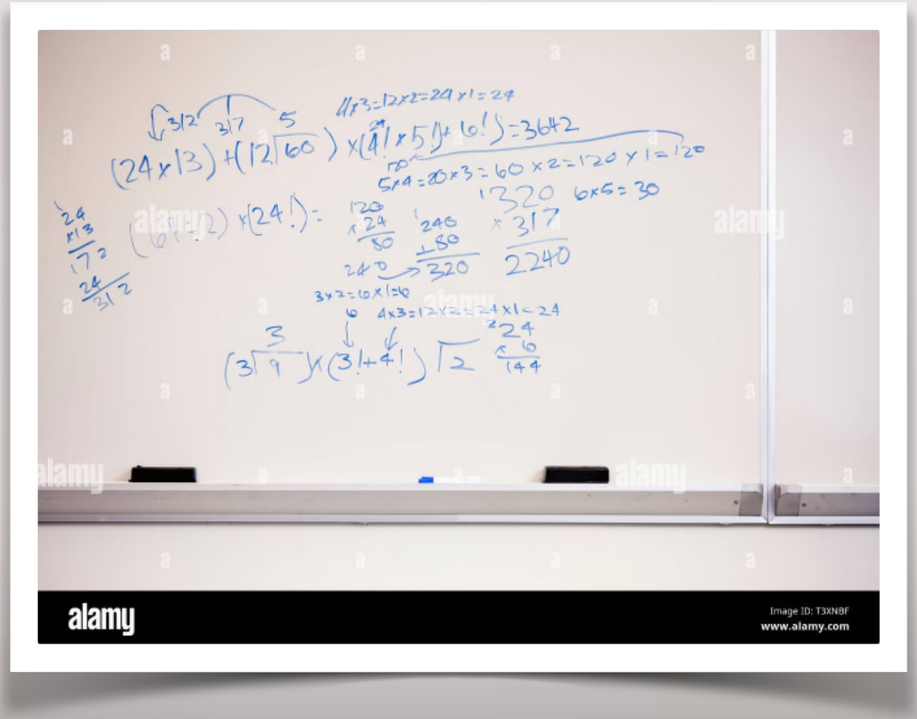

smoker
count
1000
800
600
400
200
0
yes
no
smoker



Data Preprocessing

Data Encoding

1. encoding ‘sex’ column
 male : 0 , female: 1
2. Encoding ‘smoker’ column
 Yes : 0, No:1
3. Encoding ‘region’ column
Southeast:0, southwest:1,
Northwest:2, Northeast:3



Linear Regression Model Development

1

Splitting the dataset

In this level we split the features and target,we split this in x and y

2

Model Training

Importing the linear regression model from sklearn and then dividing the x and y into trains and testing data.

3

Model Evaluation

Evaluate the model's performance on unseen data to assess its accuracy and generalizability.

Hyperparameter Tuning and Optimisation

Regularisation

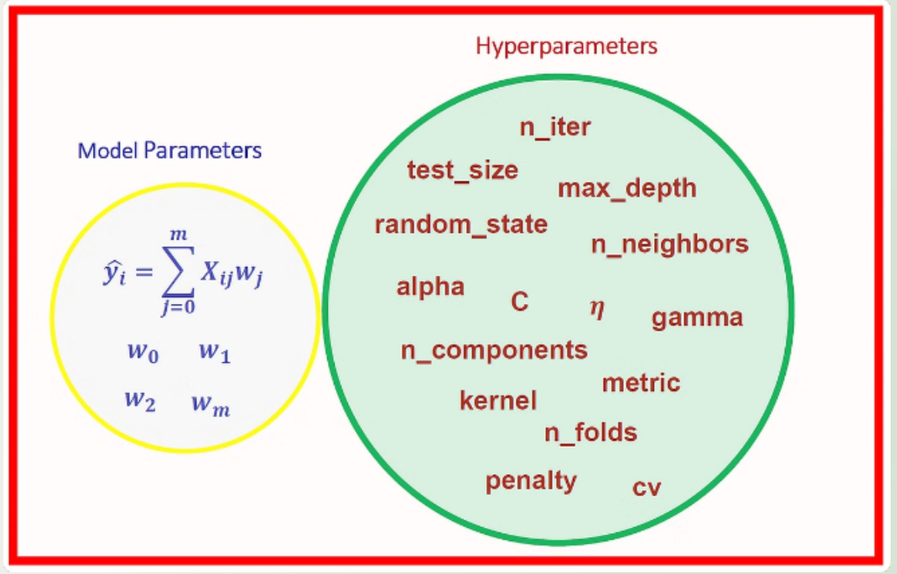
Apply techniques like L1 or L2 regularisation to prevent overfitting and improve model generalisation.

Feature Importance

Analyse the importance of different features in the model to identify key drivers of insurance costs.

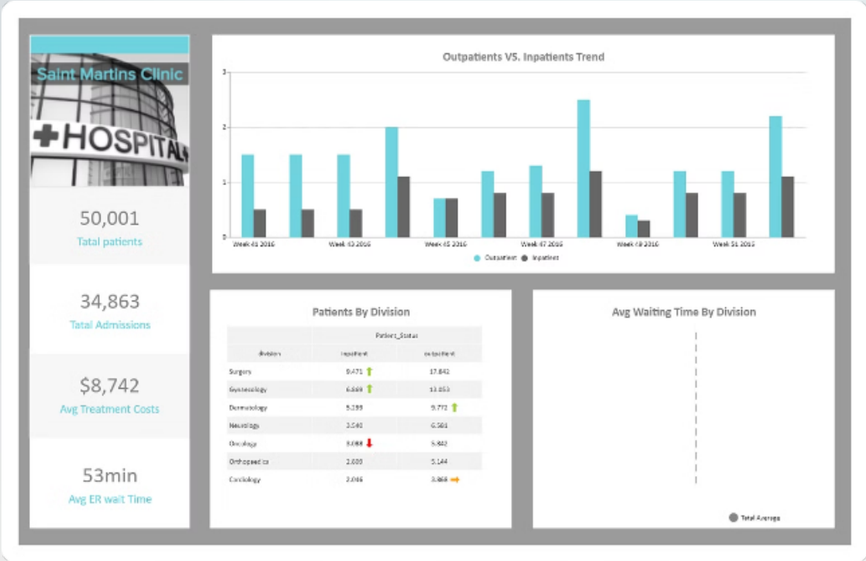
Cross-Validation

Use cross-validation techniques to evaluate the model's performance on different subsets of the data.



Conclusion and Future Considerations

By leveraging linear regression, we can develop accurate medical insurance cost prediction models. This leads to fair pricing, improved healthcare management, and valuable insights. Future work can focus on incorporating more complex machine learning models, exploring new data sources, and addressing the ethical implications of using predictive models in healthcare.



Thank You