Insurance Premium Prediction

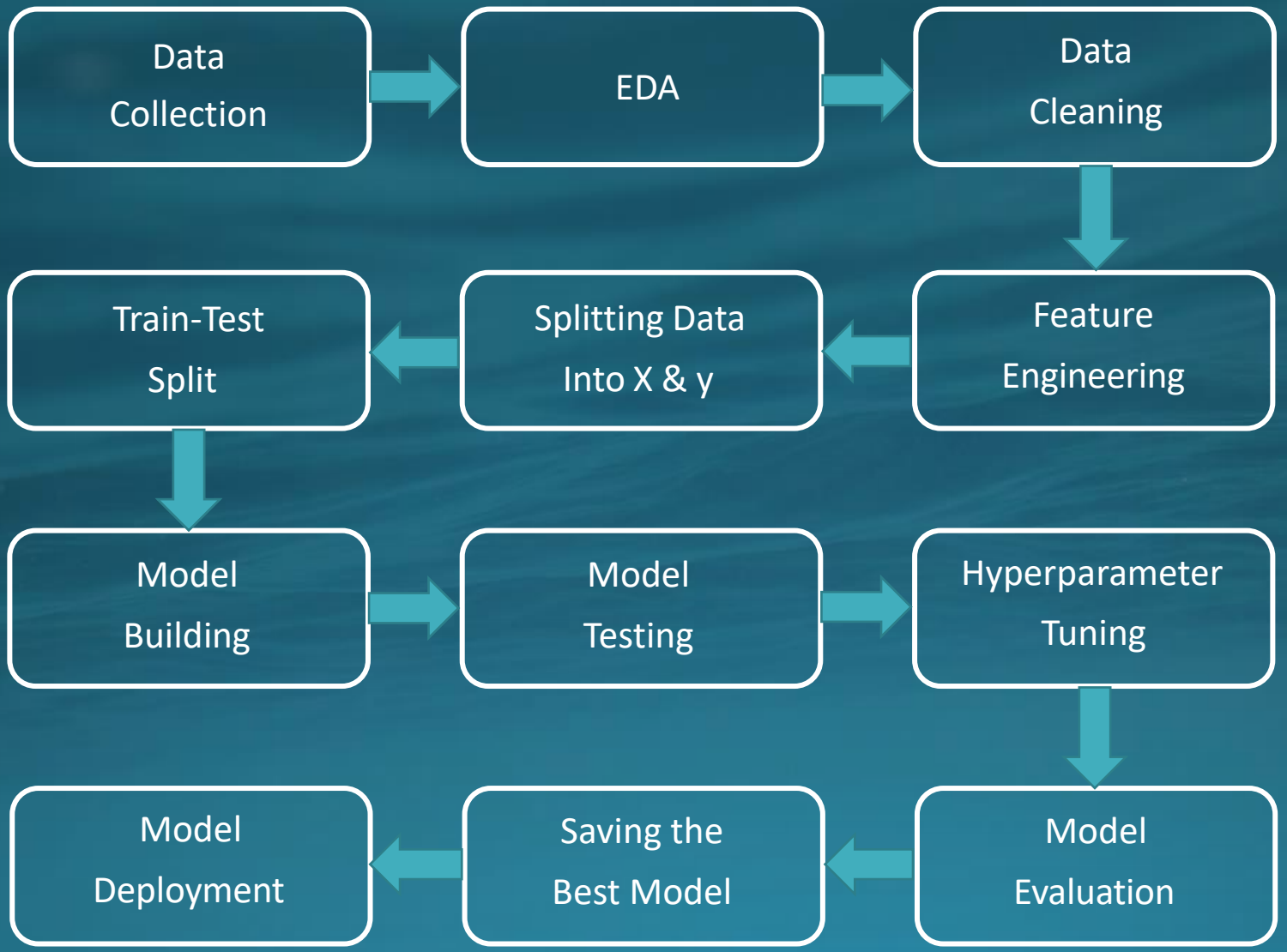
Objective:

The objective of this project is to estimate the insurance premium charges of a person and identify those patients with any health issues or not and build a solution that should be able to predict the premium of the personal for health insurance.

Benefits:

- Gives better insight about how much premium is required annually according to their health condition.
- Will help the person for making better health driven decisions by estimating the premium.
- Helps in giving premium of health insurance.

Architecture



Data Collection and Data Validation

- The dataset is in .csv format and is collected from Kaggle platform.
- Number of Columns : Validation of number of columns present in the file, and if it doesn’t match then it is corrected.
- Name of Columns : The name of the columns is validated and should be the same as given in the file.
- Data type of columns : Validating the data type of the columns. If wrong then it was corrected.
- Null values in columns : If any of the columns have NULL or missing values, we discard such file.

Model Training

➢ Data Pre-processing :

- Performing EDA to get insights of the data like identifying distribution, outliers, trend among data etc.
- Check for any null values in the dataset. If present then impute those null values.
- Encode the categorical values with numeric values.
- Perform Standard Scalar to scale down the values.

Model Selection

After pre-processing step we do model training, we find the best model for predicting the insurance premium. By using various regression algorithms like Linear Regression, Support Vector, Random Forest, Gradient Boosting we trained multiple models. We perform hyperparameter tuning to get the best model parameters. We calculate r2_score for each model and select the model with the best score. Metrics like MAE (Mean Absolute Error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error) are used for evaluation.

Predictions

- The trained models are then used for validating on test dataset.
- We perform pre-processing techniques on it.
- The model with lowest MAE and best r2_score is saved for predictions.

Q & A

Q1) What's the source of data?

The data is collected from Kaggle platform. The data is in the form of 'csv' file.

Q2) What was the type of data?

The data was the combination of categorical and numerical values.

Q3) What's the complete flow you followed in this project?

Refer the 3rd slide for better understanding.

Q4) What are the techniques used for data pre-processing?

- Visualizing relations between independent variables and dependent variable.
- Checking distribution of numerical variables.
- Cleaning data and imputing if null values are present.
- Removing outliers.
- Checking correlation using heatmap.
- Converting categorical features into numerical.

Q5) How training was done or what models were used?

- Before training the model we split the dataset into training set and testing/validation set.
- The model is fitted on training set and predictions are done using testing/validation set.

- Algorithms like Linear Regression, SVR, Random Forest, Gradient Boosting were used for model training and based on MAE and r2_score best model is saved for validation.

Q6) How prediction was done?

The model with best score is used to perform predictions. An API interface is created for estimation of premium cost, were the user can get predictions on a single click.

Q7) What are the different stages of deployment?

When the model is ready we deploy it in Heroku platform.