

**Machine Learning - Project Report Document**

|  |  |
| --- | --- |
| **Student Name** | Tella Krishna Prasad |
| **Batch** | AI Elite 18 |
| **Project Name** | Medical Prediction |
| **Project Domain** | Healthcare Domain |
| **Type of Machine Learning** | Supervised |
| **Type of Problem** | Regression |
| **Project Methodology** | CRISP-DM |
| **Stages Involved** | * Business understanding. * Data understanding. * Data preparation. * Modeling. * Evaluation. |

**Stage 1: Business Understanding:**

Insurance is like a safely net for life’s uncertainties. People pay a little money (premiums) regularly to an insurance company, which promises to cover big expenses if something bad happens, like a car crash or a house fire. It helps insurance companies figure out who’s likely to need help and how much they should charge to make sure everyone’s protected fairly.

**Business constraints:**

Insurance must navigate the delicate balance between risk mitigation and customer satisfaction, ensuring that policies adequately protect clients without overburdening them with excessive premiums.

**Stage 2: Data Collection and Understanding:**

1. **Data Collection:** Data collected from the client.

**Methods involved in data collection:**

Age & Gender: Demographic details aiding in risk assessment and customer segmentation.

BMI & Smoker Status: Health-related factors influencing insurance premiums and coverage.

Children & Region: Family structure and geographic location impacting insurance needs and costs.

Charges: Financial data reflecting premium payments and revenue generation for the insurer.

This representation emphasizes the interplay between demographic, health, family, geographic, and financial factors in the insurance context.

1. **Data Understanding:**

In data set 3 types of data are is observed i.e., int64, float64 and object.

It is observed that age and charges have more correlation when compared with other features in dataset.

|  |  |  |
| --- | --- | --- |
| **S No** | **Feature Name** | **Data Type** |
| 1 | Age | Int64 |
| 2 | Sex | object |
| 3 | bmi | Float64 |
| 4 | children | Int64 |
| 5 | smoker | object |
| 6 | region | object |
| 7 | charges | Float64 |

**Stage 3: Data Preparation**

**a) Exploratory Data Analysis:** Data provided was good need there was no required of exploratory data analysis.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S no** | **Type of Cleaning** | **Technique** | **Feature Name** | **Reason** |
| 1 | Missing value | Imputing with mean |  |  |
| 2 | Encoding | One hot | Sex  Smoker  Region | Categorical column applied OHE that help to convert categorical to numerical |
| 3 | Scaling | Standard Scaling | RobustScaler | scaled features are less affected by outliers |

**Stage 4: Model Building:**

**Algorithm selected for the problem statement:**

* Linear Regression: A linear modeling technique that predicts a continuous outcome by fitting a linear equation to the observed data.
* KNeighbors Regressor: A non-parametric regression algorithm that predicts the target variable by averaging the values of its k-nearest neighbors.
* DecisionTree Regressor: A tree-based regression algorithm that predicts the target variable by recursively partitioning the feature space into regions and assigning a constant value to each region.
* RandomForest Regressor: An ensemble learning method that constructs multiple decision trees during training and outputs the average prediction of the individual trees for regression tasks.
* SVR (Support Vector Regressor): A regression algorithm that finds the hyperplane with the maximum margin between data points and predicts continuous outcomes by fitting a linear function with support vectors.
* AdaBoost Regressor: An ensemble learning technique that combines multiple weak regressors sequentially, focusing on the errors of the previous models to improve overall prediction accuracy.
* GradientBoosting Regressor: An ensemble learning method that builds predictive models sequentially, each correcting errors of the previous models, often achieving high predictive accuracy for regression tasks.

|  |  |  |
| --- | --- | --- |
| **S No** | **Type of Problem** | **Algorithm Name** |
| 1 | Regression | Linear Regression |
| 2 | Regression | KNeighbors Regressor |
| 3 | Regression | DecisionTree Regressor |
| 4 | Regression | RandomForest Regressor |
| 5 | Regression | SVR |
| 6 | Regression | AdaBoost Regressor |
| 7 | Regression | GradientBoosting Regressor |

**Stage 5: Model Training:**

**Details of metrics used for evaluation:**

|  |  |  |
| --- | --- | --- |
| **S No** | **Algorithm Name** | **Metric used for Evaluation** |
| 1 | Linear Regression | Mean Absolute Error (MAE) |
| 2 | KNeighbors Regressor | Mean Absolute Error (MAE) |
| 3 | DecisionTree Regressor | Mean Absolute Error (MAE) |
| 4 | RandomForest Regressor | Mean Absolute Error (MAE) |
| 5 | SVR | Mean Absolute Error (MAE) |
| 6 | AdaBoost Regressor | Mean Absolute Error (MAE) |
| 7 | GradientBoosting Regressor | Mean Absolute Error (MAE) |

**Stage 5: Model Evaluation:**

|  |  |  |
| --- | --- | --- |
| **S No** | **Algorithm Name** | **Metric Score** |
| 1 | Linear Regression | 4015.599661 |
| 2 | KNeighbors Regressor | 4080.069988 |
| 3 | DecisionTree Regresso | 4466.368943 |
| 4 | RandomForest Regressor | 4015.652004 |
| 5 | SVR | 8476.749886 |
| 6 | AdaBoost Regressor | 5706.685523 |
| 7 | GradientBoosting Regressor | 3977.770820 |

**Challenges Faced:**

Analysing the dataset poses challenges in understanding how demographics like age and gender, lifestyle factors such as smoking status, and regional differences influence insurance charges. Addressing biases, managing outliers, and ensuring compliance with regulations are key considerations in data analysis.

**Conclusion:**

In conclusion, the models were evaluated based on Mean Absolute Error (MAE), which measures the average absolute difference between predicted and actual values. Among the models tested, GradientBoosting Regressor performed the best with the lowest MAE of 3977.77, followed closely by RandomForest Regressor with an MAE of 4015.65. Linear Regression and SVR showed higher MAE values, indicating less accurate predictions compared to other models. DecisionTree Regressor and AdaBoost Regressor also exhibited relatively higher MAE values, suggesting room for improvement in their predictive performance.