**Machine Learning Algorithms with Python: Analyzing Medical Insurance Costs**

University Canada West

BUSI 652 (Section- 12): Predictive Analysis: What Works?

Prepared by: Aditya Arte

Student ID: 2305505

Date: June 8, 2024

**Table of Contents**

[**Abstract 3**](#_2t9rkro8ena4)

[**Introduction 4**](#_u9u8o7e1h1y)

[**Data Description and Exploratory Analysis 5**](#_z88p4cn79cez)

[**Methodology 7**](#_hppnsxu8z9rq)

[Data Preprocessing 7](#_28rlnsu14axy)

[Dummy Variable Creation 7](#_mwdhq9tw6d8b)

[Data Normalization 7](#_4nuk633uxhuz)

[**Model Training and Evaluation 8**](#_x7iovjieuepq)

[Linear Regression and Piecewise Linear Regression 8](#_s3gxe0t17hco)

[Evaluation Metrics 8](#_bqacx3hwsdym)

[Evaluation of Prediction Models 9](#_r1iu3owsh32r)

[Linear Regression 9](#_e66xz9st3830)

[Decision Tree 9](#_q41q58lz4usi)

[K-Nearest Neighbours 9](#_rw6rarkrpfl5)

[Piecewise Linear Regression 10](#_f0rntjkdonei)

[**Visualization 11**](#_o70ph0831lu9)

[**Conclusion 16**](#_q3uy36gdpoe9)

# **Abstract**

In this paper, the application of different machine learning algorithms for predicting medical cost insurance based on continuous features such as age, BMI, number of children, and categorical parameters like sex, smoking status, and region is discussed, and which prediction model is better is evaluated. It includes data preprocessing, outlier removal, dummy variable creation, data normalization, and model evaluation, which includes Linear Regression, Decision Tree, KNN, and piecewise Linear Regression. Performance metrics include evaluating Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared.

# **Introduction**

Professionals in the field have developed machine learning as an important technique for making predictions and inferences from large data sets. In the corporate world, whenever a situation analysis is done to estimate and predict the trend or feasibility of a project or product, the employees engaged in such projects make use of machine learning to transform the large set of input data and make predictions. In this paper, the author has based the predictions on a medical insurance dataset with the target of making the prediction of insurance charges based on several input features. For both the insurance provider and the policyholder, accurate forecasting of the costs is beneficial in and of itself. The author tries to develop machine learning models with maximum prediction accuracy for themselves.

# **Data Description and Exploratory Analysis**

The dataset includes 1,338 observations with attributes: Age, BMI, Number of Children, Charges, Sex, Smoking, and Region. These variables are going to be used as independent variables to predict the insurance charges.

Age: The age of the individual. The given parameter has no missing values; the percentage of missing values is null. The minimum and maximum observed ages are 18 and 64 years, respectively. The average and median ages are 39 and 25, and 75 percent of the individuals are below 27 and 51 years, respectively.

BMI: Body Mass Index is the body fat obtained from height and weight by applying a person's mass formula. This parameter does not have any missing values; the percentage of missing values is zero. The minimum and maximum observed BMIs are 15.96 and 53.13, respectively. The average and median BMIs are 30.66 and 30.4, respectively, which means that most people are not fit. 25 percent of the people have a BMI of below 26.29, and 75 percent are below 34.69.

Children: The number of children or dependents covered by the insurance. This parameter has no missing values; the percentage of missing values is zero. The minimum and maximum recorded number of children is 0 and 5, respectively. The average and median number of children is 1, and this means that most people have one kid. 25 percent of the individuals have no kids, and 75 percent have 2 kids.

Charges: Medical insurance cost. This parameter doesn't contain missing values; the missing percentage is zero. The minimum and the maximum observed charges are 1121.87 and 63770.43, respectively. The average and median charges are 13270.42 and 9382.03, respectively. 25 percent of individuals pay below 4740.28, and 75 percent below 16639.91.

Sex: Gender of the individual (male/female). There are 676 men and 662 women.

Smoker: Is the individual a smoker or not (yes/no)? There are 274 smokers and 1064 who do not smoke.

Region: Where the individual comes from (northeast, northwest, southeast, southwest).

324 persons live in the Northeast region, the same number as in the Northwest, 325 persons live in the Southwest, and 364 persons live in the Southeast, dividing the population evenly across the regions.

# **Methodology**

## **Data Preprocessing**

Data preprocessing includes handling missing values, although the dataset in this study contains no missing entries. Outliers are identified and removed using the Interquartile Range (IQR) method, which ensures data points fall within acceptable limits, thus enhancing the model's effectiveness.

## **Dummy Variable Creation**

Categorical variables such as sex, smoker status, and region are converted into dummy variables. This transformation allows these categorical variables to be used effectively in machine learning models by converting them into a binary format.

## **Data Normalization**

Continuous data is scaled to fit within a range so that the K-Nearest Neighbour model does not disregard the effect of any parameter.

# **Model Training and Evaluation**

## **Linear Regression and Piecewise Linear Regression**

Linear regression models are trained on various parameters and also a combination of categorical data to improve prediction accuracy. By segmenting the data based on categories such as gender, smoking status, and region, the author created a piecewise linear regression model tailored to each segment.

## **Evaluation Metrics**

The models are evaluated using the following metrics:

Mean Squared Error (MSE): Measures the average squared difference between actual and predicted values. Lower MSE suggests a better fit.

Root Mean Squared Error (RMSE): The square root of MSE gives error metrics in the same unit as the target variable.

Mean Absolute Error (MAE): Measures the average absolute difference between actual and predicted values. A lower MAE indicates a good fit.

R-squared (R²): Reflects the proportion of variance in the dependent variable explained by the independent variables; the closer to 1, the better the fit.

## **Evaluation of Prediction Models**

### Linear Regression

MSE: 23437328.29

RMSE: 4841.21

MAE: 3316.73

R square: 70.96%

### Decision Tree

MSE: 28993970.45

RMSE: 5384.60

MAE: 2300.56

R square: 73.4%

### K-Nearest Neighbours

MSE: 27989408.02

RMSE: 5290.50

MAE: 3294.13

R square: 63.77%

### Piecewise Linear Regression

For the test Data:

MSE: 23,296,700.81

RMSE: 4,826.67

MAE: 2,876.09

R square: 84.11%

For the training data:

MSE: 22,728,811.60

RMSE: 4,767.47

MAE: 2,957.51

R square: 84.48%

These results suggest that the piecewise linear regression model has relatively high predictive accuracy, with R² values indicating that the model can explain a substantial portion of the variance in insurance charges.

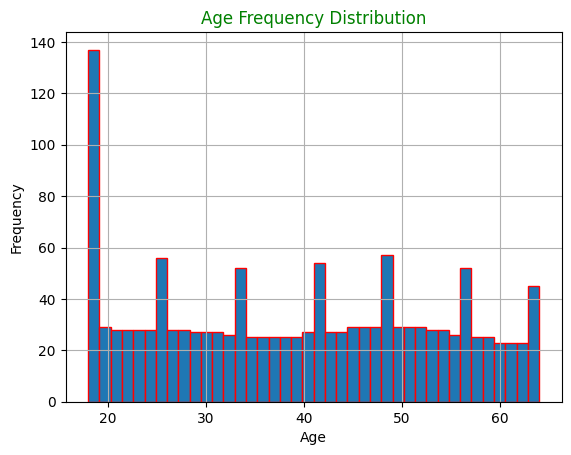
# **Visualization**

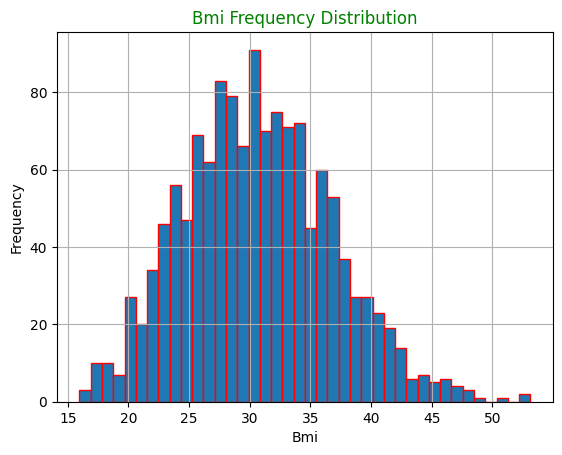
Histograms for the frequency distribution of the age and BMI attributes are generated to visualize the data distribution. These visualizations help in understanding the spread and central tendency of the data. For instance, the age distribution shows a higher frequency of younger individuals, while the BMI distribution appears to be approximately normal, with a peak around 30.

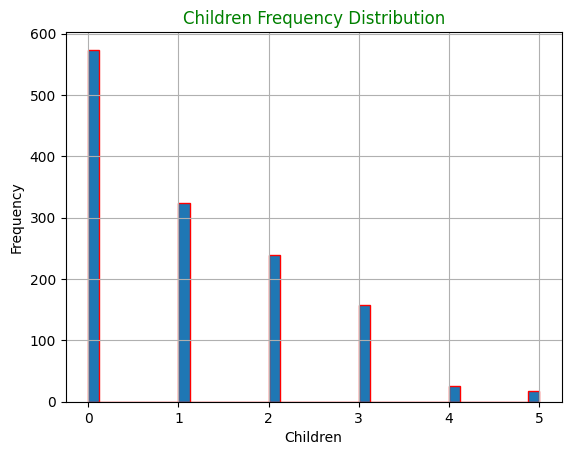
The boxplot shows the outliers and the data about the quartiles. This can be used to check for minimum and maximum acceptable values.

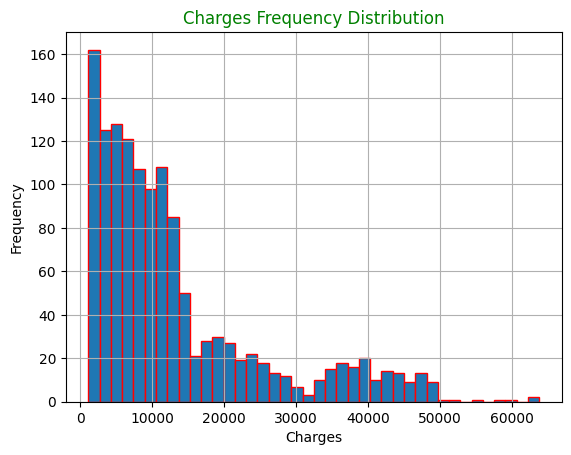
**Figure 1**

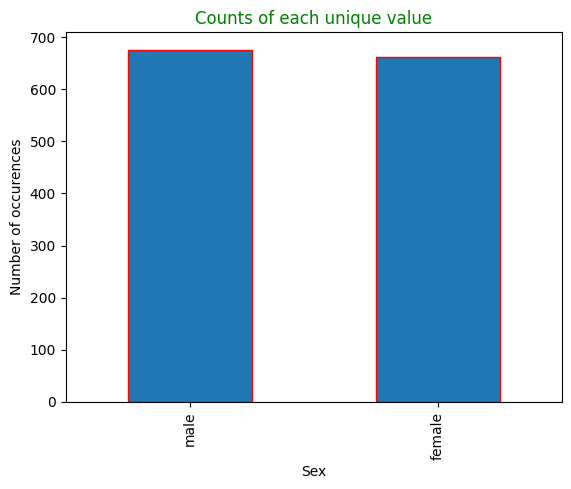
Frequency Distribution of all the parameters

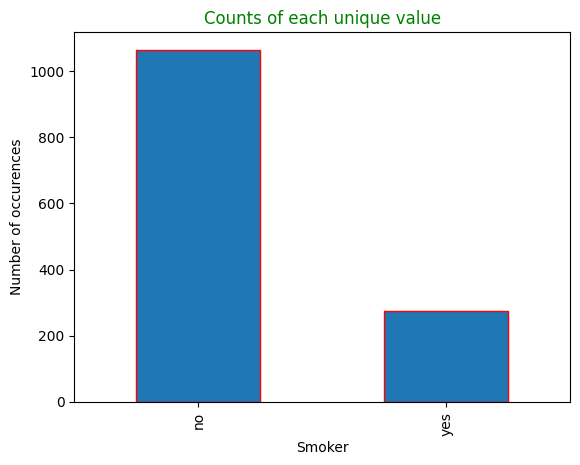
****

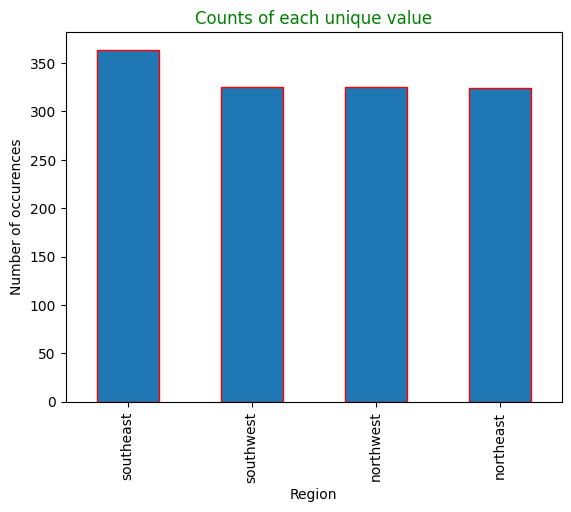
****

****

****

****

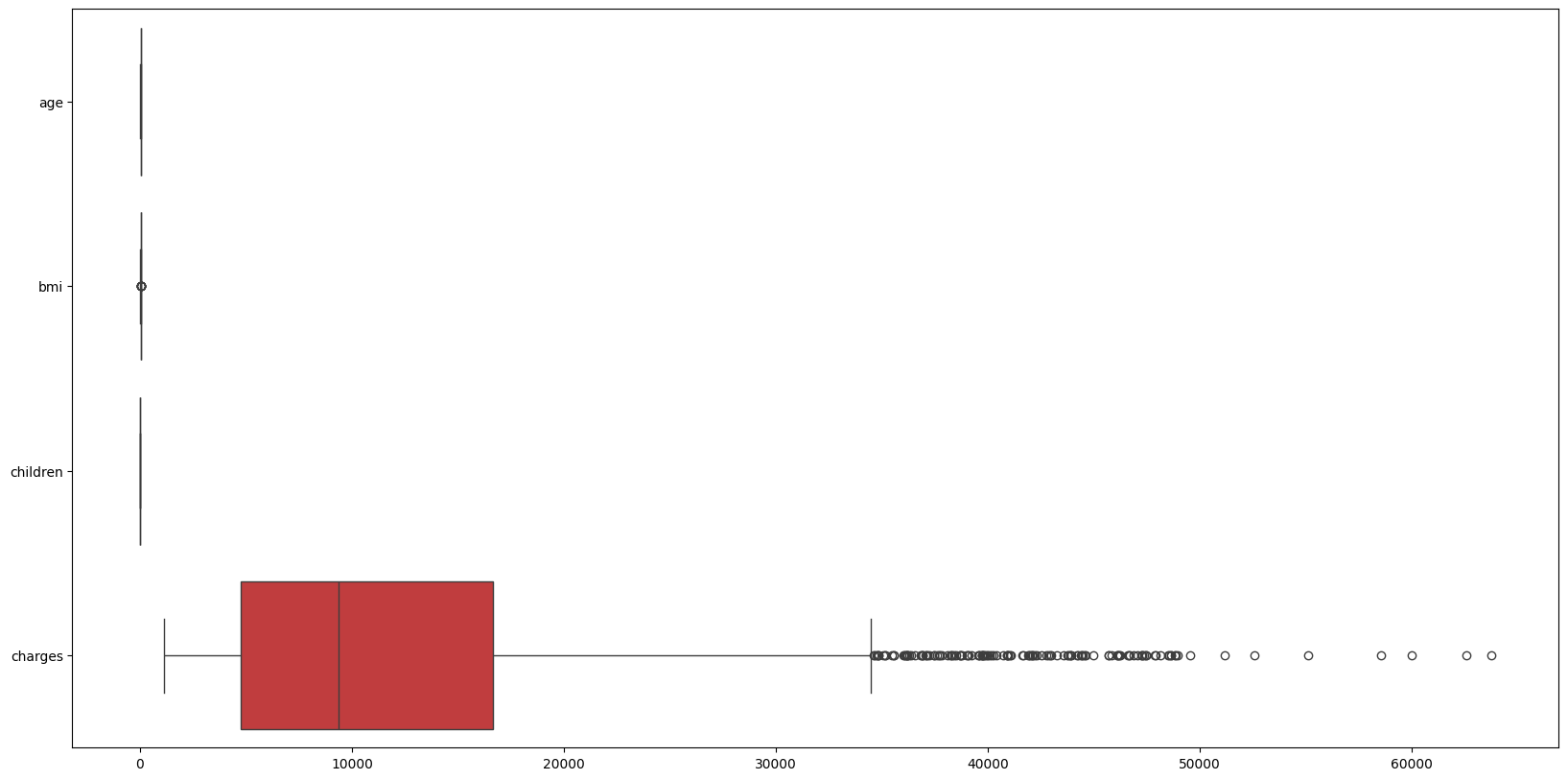
****

****

*Note.* From MLOutput

**Figure 2**

Boxplots of all parameters



*Note.* From MLOutput

# **Conclusion**

The study was conducted to predict medical insurance costs through several existing machine-learning techniques. Categorical variables are used, and the data is segmented based on that categorical variable. The performance of all these models in predicting medical insurance costs is studied using standard metrics. The study shows extremely promising results, especially for linear regression models.

It is a good practice and finding to include categorical variables for data segmentation according to the categories. Future works should implement more features and other advanced machine learning algorithms in order to increase the prediction accuracy. The best optimization of the models could be achieved by using techniques such as cross-validation and hyperparameter tuning.