Graphical user interface, text, application, website

Description automatically generated

Text

Description automatically generated with low confidence

Text

Description automatically generated

**Team members :**

Yenimetla Venkata Krishna Chaitanya

Rahul R

Yandrapu Andrea Chamili Varsha

**Submitted to :**

Gurvansh Singh

M.Tech

Knowledge Solutions India

TABLE OF CONTENT

|  |  |
| --- | --- |
| Table of figures/graphs | 2 |
| Abstract | 3 |
| Chapters(Introduction,Dataset, Software-libraries used, Algorithm) | 3-4, 10-12 |
| Data Preprocessing | 4-9 |
| Conclusion | 13 |

TABLE OF FIGURES/GRAPHS

|  |
| --- |
| EDA(displot of charges) Graph 1 , 2 |
| Barplot Graph 3,4,5,6 |
| Lmplot Graph 7,8,9 |
| Violinplot Graph 10 |
| Heatmap Graph 11 |

**Abstract:**

Insurance is a policy that eliminates or decreases losscosts occurred by various risks. Various factors influence the cost of insurance. These considerations contribute to the insurance policy formulation. Machine learning (ML) for the insurance industry sector can make the wording of insurance policies more efficient. This study demonstrates how different models of regression can forecast insurance costs. And we will compare the results of models, for example, Multiple Linear Regression, Random Forest Regressor. This paper offers the best approach to the Random Forest Regressormodel with an MAE value of 3930.33, RMSE value of 5643.21and R -squared value of 0.8.

**Keywords:** regression, machine learning,forecast, insurance

**INTRODUCTION**

We are on a planet full of threats & uncertainty, people, households, companies, properties, and property are exposed to different risk forms and the risk levels can vary. These dangers contain the risk of death, health, and property loss or assets. Life and wellbeing are the greatest parts of people's lives but, risks cannot usually be avoided. So the world offinance has developed numerous products to shield individuals and organizations from these risks by using financial capital to reimburse them.

Insurance is, therefore, a policy that decreases or removes loss costs incurred by various risks. Concerning the value of insurance in the lives of individuals, it becomes important for the companies of insurance to be sufficiently precise to measure or quantify the amount covered by this policy and the insurance charges which must be paid for it various variables estimates these charges. Each factor of these is important. If any factor is omitted when the amounts are computed, the policy changes overall. It is therefore critical that these tasks are performed with high accuracy. As human mistakes are could occur, insurers use people with experience in this area. They also use different tools to calculate the insurance premium. ML is beneficial here. ML may generalize the effort or method to formulate the policy. These ml models can be learned by themselves. The model is trained on insurance data from the

past. The requisite factors to measure the payments can then be defined as the model inputs, then the model can correctly anticipate insurance policy costs. This decreases human effort and resources and improves the company's profitability. Thus the accuracies can be improved with ml. our objective is to

forecast insurance charges in this article. The value of insurance fees is based on different variables. as a result, insurance fees are continuous values. the regression is the best choice available to fulfill our needs. We use multiple linear regression in this analysis since there are many independent variables used to calculate the dependent(target) variable. For this study, the dataset for cost of health insurance is used. Pre-processing of the dataset done first. then we trained regression models with training data and finally evaluated these models based on testing data. in this article, we used several models of regression, for example, multiple linear regression, Random Forest Regressor. It is found that the Random Forest Regressor boosting provides the highest accuracy with an r-squared value of 0.8 . The key reason for this study is to include a new way of estimating insurance costs.

**DATASET**

Insurance cost dataset

Name Description

age Age of the client

BMI body mass index

The Number of Kids number of children the client have

gender Male / Female

smoker whether a client is a smoker or not

region where the client lives southwest, southeast,

northwest or northeast

Charges(target variable) Medical Cost the client pay

**DATA PREPROCESSING**

The dataset includes seven variables every one of these attributes has some contribution to estimate the cost of the insurance, which is our dependent variable. In this stage, the data is scrutinized and updated properly to efficiently apply the data to the ML algorithms. First of all, all unknown values are cleaned. The unknown numerical values are replaced with the mean.

The target variable (charges) would then be examined.

Graphical user interface, text

Description automatically generated

**EDA**

We can seehow the charges are distributed

Chart

Description automatically generated

Graph 1

This distribution is right-skewed. To make it closer to normal we can apply natural log

Chart, histogram

Description automatically generated

Graph 2

charges vs region

Graphical user interface

Description automatically generated with medium confidence

Graph 3

So overall the highest medical charges are in the Southeast and the lowest are in the Southwest. Taking into account certain factors (sex, smoking, having children) let's see how it changes by region

Graphical user interface, application

Description automatically generated

Graph 4

Chart

Description automatically generated with low confidence

Graph 5

Graphical user interface

Description automatically generated

Graph 6

As we can see from these barplots the highest charges due to smoking are still in the Southeast but the lowest are in the Northeast. People in the Southwest generally smoke more than people in the Northeast, but people in the Northeast have higher charges by gender than in the Southwest and Northwest overall. And people with children tend to have higher medical costs overall as well..

Now let'sanalyze the medical charges by age, bmi and children according to the smoking factor

Graphical user interface, text, application

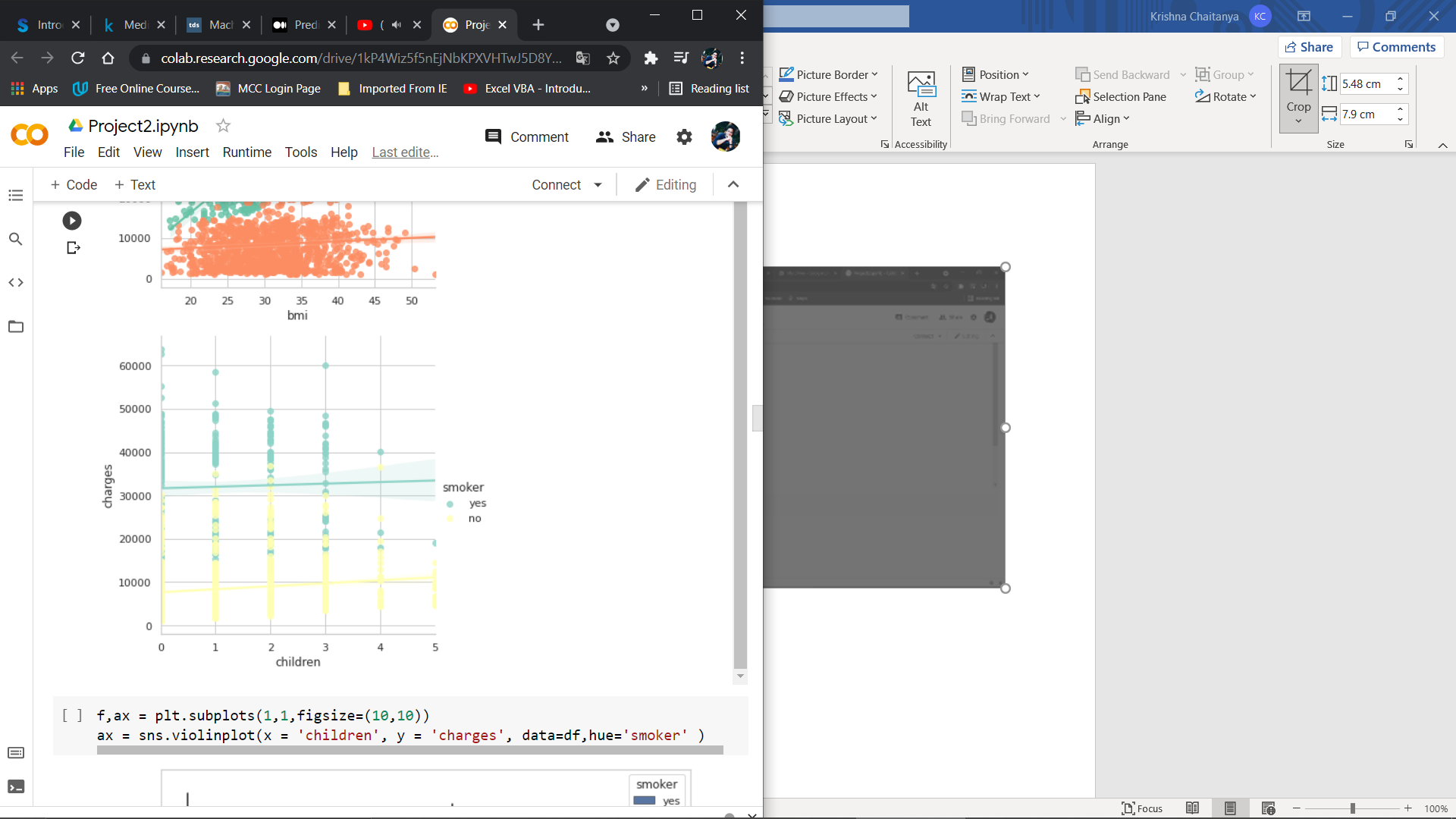
Description automatically generated

Graph 7

Graphical user interface

Description automatically generated with medium confidence

Graph 8



Graph 9

Smoking has the highest impact on medical costs, even though the costs are growing with age, bmi and children. Also people who have children generally smoke less, which the following violinplots shows too…

Graphical user interface

Description automatically generated

Graph 10

No correlation, except with the smoking

Graphical user interface, website

Description automatically generated

Graph 11

**Software-libraries used**

**Pandas**- Pandas is a Python library.Pandas is used to analyze data.

**Numpy** – Numpy is a Python library.Numpy is used for working with arrays.

**Matplotlib**– low level graph plotting library in python that serves as a visualization utility.

**Seaborn** – high-level interface for drawing attractive and informative statistical graphics.

**Scikit-learn** – efficient tools for predictive data analysis.

**Sklearn.model\_selection.train\_test\_split**–Split arrays or matrices into random train and test subsets.

**Sklearn.preprocessing.LabelEncoder** – Encode target labels with value between 0 and n\_classes – 1 .

**Sklearn.linear\_model.LinearRegression** – Linear Regression fits a linear model with coefficients w = (w1,…..,wp) to minimize the residual sum of squares between the observed targets in the dataset and the target predicted by the linear approximation.

**Sklearn.metrics.mean\_squared\_error** – Mean squared error regression loss.

**Sklearn.metrics.r2\_score** - R2 (coefficient of determination) regression score function.

Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse). A constant model that always predicts the expected value of y, disregarding the input features, would get a R2 score of 0.0.

**Sklearn.metrics** – it implements functions assessing prediction error for specific purposes.

**Sklearn.preprocessing.StandardScaler** – By removing the mean and scaling to unit variance.

**Sklearn.decomposition.PCA** – Linear dimensionality using Singular Value Decomposition of the data to project it to a lower dimensional space.

**Sklearn.ensemble.RandomForestRegressor** and **Sklearn.ensemble.RandomForestClassifier**– A random forest is a meta estimator that fits a number of classifying decision tree on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over -fitting.

**Algorithm**

Multiple Linear Regression

In practice, we often have more than one predictor. For example, with the

data set used in this study, we may wish to understand if independent variables (6 independent variables), (linearly) related to the dependent variable (charges). this is referred to as the multiple linear regression (MLR) model . An MLR model with t independent features X1, X2, ..., XT , and Y results can be calculated as in the following equation

Graphical user interface, text, application

Description automatically generated

In the above equation, u is the residual regression while a is the weight of each independent variable or parameter assigned.

Random Forest

Random forests reflect a shift to the bagged decision trees that create a broad number of decorrelated trees so that predictive efficiency can be improved further. They are a very popular 'off-the-box' or off-the-shelf' learning algorithm, with good predictive performance and relatively few hyperparameters.Random forests create an average predictive value as a result throughout the regression of individual trees. Random forests resolve to overfit . As in the following equation , a random model for forest regressors

can be expressed.

Graphical user interface, text, application

Description automatically generated

where g is the final model that is the sum of all models. Each model f(x) is a decision tree.

Principal Component Analysis

PCA reduce the number of variables of a data set, while preserving as much information as possible

Principal Component Analysis Standardization

Mean Absolute Error(MAE)

The Mean Absolute Error (MAE) is the difference between the original and forecast values obtained by averaging the absolute difference over the data set.

MAE= Graphical user interface, text, application, Word

Description automatically generated

Root Mean Squared Error(RMSE)

The RMSE of the disparity between the expected values and the real values is determined as the square root. For an accurate forecast, the RMSE must be low so there would be less variance among the expected values and the real values.

RMSE = Graphical user interface, text, application, Word

Description automatically generated

Where N = Number of overall observations, Yv= expected insurance fee values, Y = real insurance fee values.

R-squared

The R-squared is often called the coefficient of decision. The proportion of variance is estimated from the independent variables in the dependent variable.

R-squared = Explained variance**/**Total variance

The more R-squared, the better the model output and indicates that the model deviates less from real values. A R-squared score of 1 indicates that it suits perfectly.

**CONCLUSION**

The research uses various machine learning regression models to forecast charges of health insurance based on specific attributes, on medical cost

personal data set . Random Forest Regressor can therefore be used in the estimation of insurance costs with better performance than Multiple linear regression. Forecasting insurance costs based on certain factors help insurance policy providers to attract consumers and save time in formulating plans for every individual. Machine learning can significantly minimize these

individual efforts in policymaking, as ML models can do cost calculation in a short time, while a human being would be taking a long time to perform the same task. This will help businesses improve their profitability. The ML models can also manage enormous amounts of data. **Smoking** is the greatest factor that affects medical cost charges, then it's **bmi** and **age.**