# CAPSTONE PROJECT

## Project Title: Medical Insurance Prediction

PGA-02


## Abstract

The goal of the project is to predict individual's medical insurance charges billed by health insurance companies. The dataset consists of 1338 observations & 7 variables. This data is divided into 2 sets train & test. Several machine learning regression models are trained using train dataset & applied to test dataset in order to evaluate the model performance. These performance measures are then compared to determine which model is best in prediction of insurance charges.



*Submitted by - Begum Zubeda Abbasuddin*

# Table of Contents

# Capstone Project – PGA-02

Medical Insurance Prediction

## 1. Introduction

Insurance dataset comes from the book Machine Learning with R by Brett Lantz. The data contains medical information and costs billed by health insurance companies. The data has been cleaned to match the formatting of the textbook, so it can be assessed as credible and accurate.

The data allows investigation into the factors that affect the amount that an individual spends on health insurance. For insurance companies, the cost of cover usually relates to the risk associated with a payout and the amount they would be liable to pay in the event of a payout. For health insurance specifically, the risk is likely to be higher if an individual is more likely to require medical assistance i.e. they are in worse overall health. In terms of the payout, companies often offer levels of cover based on the quality of care a patient would receive in the event they require medical attention.

Number of observations in the given dataset: 1338

## The Dataset

Categorical Variables:

**1) Sex:** insurance contractor gender, female, male.

**2) Smoker:** smoking status of insurance claimer.

**3) Region:** the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

Numerical Variables:

**1) Age:** age of primary beneficiary.

**2) Children:** no. of children covered by health insurance / no. of dependents.

**3) BMI:** body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9.

**4) Charges:** individual medical costs (in dollars) billed by health insurance.

## 1.1. Summary of the Data:

```
medical_df.describe().T.style.background_gradient(cmap=sns.light_palette("#ea4335", as_cmap=True), axis=1)
```

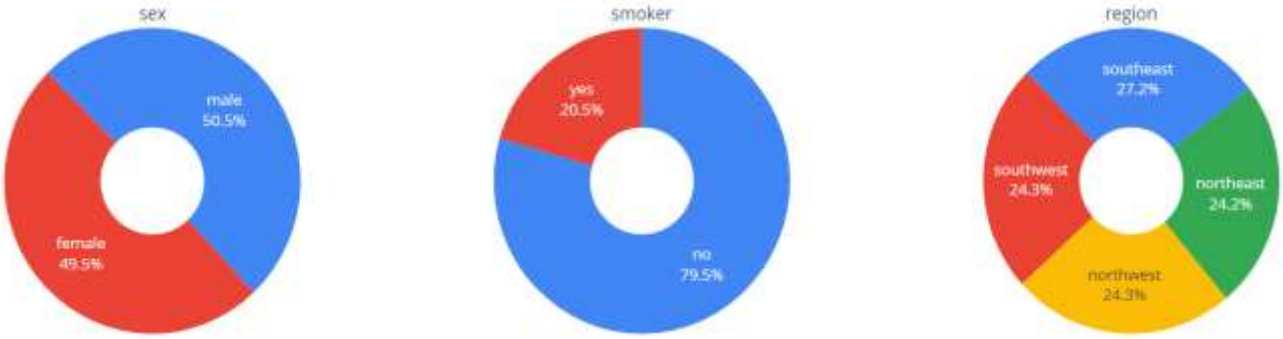
| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| age | 1338.000000 | 39.207025 | 14.049960 | 18.000000 | 27.000000 | 39.000000 | 51.000000 | 64.000000 |
| bmi | 1338.000000 | 30.663397 | 6.098187 | 15.960000 | 26.296250 | 30.400000 | 34.693750 | 53.130000 |
| children | 1338.000000 | 1.094918 | 1.205493 | 0.000000 | 0.000000 | 1.000000 | 2.000000 | 5.000000 |
| charges | 1338.000000 | 13270.422265 | 12110.011237 | 1121.873900 | 4740.287150 | 9382.033000 | 16639.912515 | 63770.428010 |

- We can see that, the age of the primary beneficiary claiming the insurance ranges between 18 and 64 years old.

- The Body Mass Index (ratio of height to weight) of the beneficiaries ranges from around 16 kg/m ^ 2 to around 53 kg/m ^ 2.

- The insurance claimer may fall under the group which do not have any children or dependents or may have up to 5 children or dependents.

- The medical insurance that is billed by health insurance companies ranges from around 1k to around 63.7k dollars.




- Here, we can see out of 1338 there are around 676 male insurance contractors and 662 female insurance contractors.

- Also, most of the insurance claimers are non-smokers and only few insurance claimers (274 out of 1338) are smokers.

- Almost equal number of beneficiaries come from every region of US, the southeast region is taking a slight lead.

# 2. Data Visualization

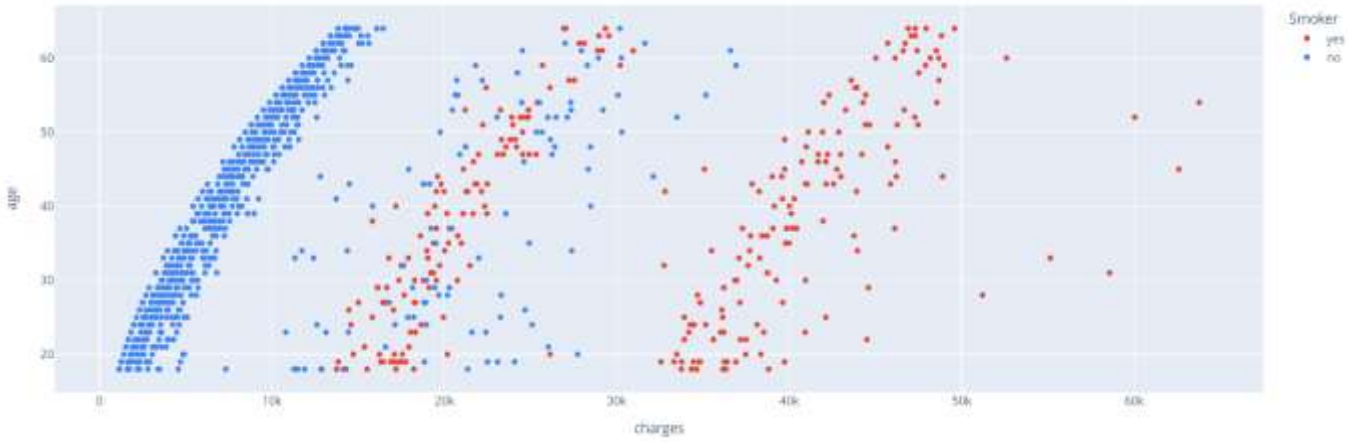



Figure 2.1.

The above graph shows Age-wise smoker distribution, we can see that individuals who smoke have a higher medical bill than individuals who do not smoke.

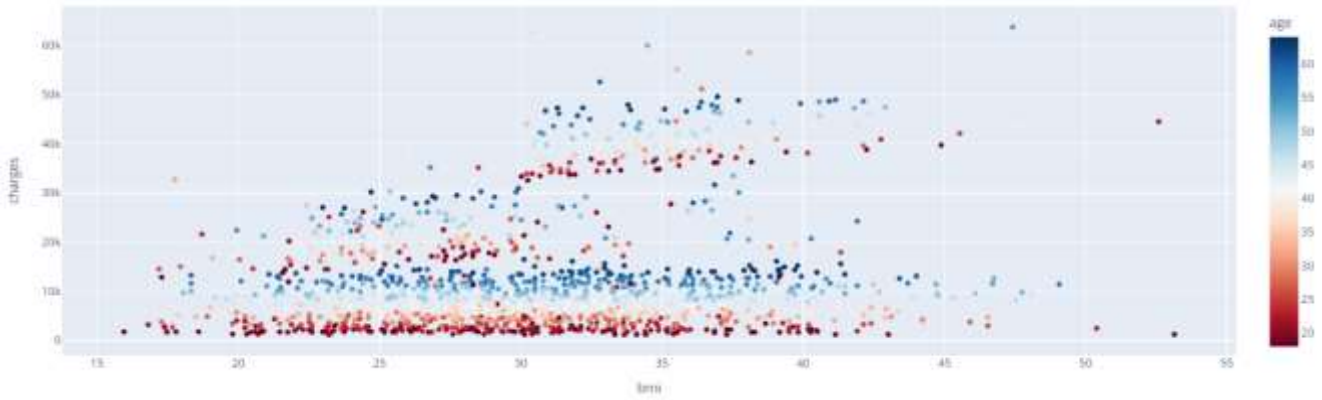



Figure 2.2.

The Age-wise bmi & charges Distribution shows that Charges has a strong correlation with age and this correlation is linear.

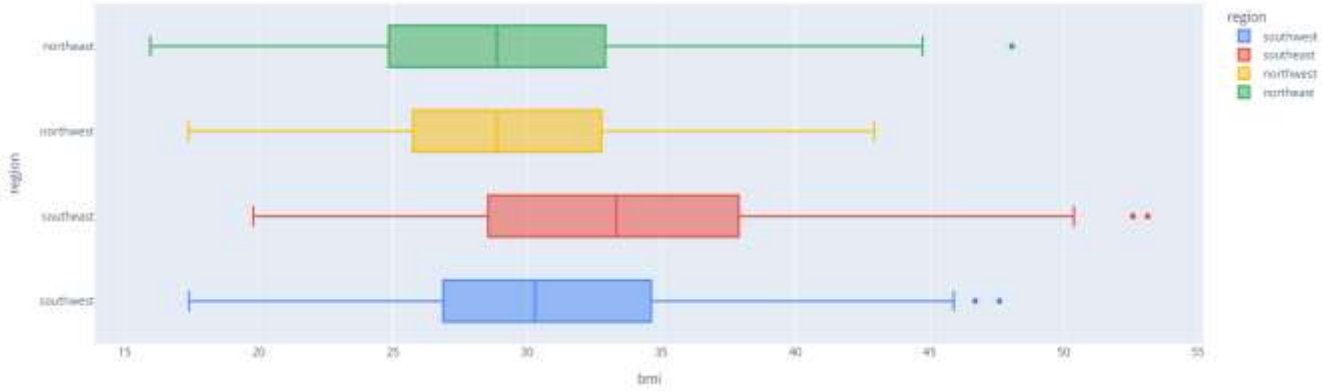



Figure 2.3.

The above graph region vs bmi depicts that the Southeastern people have the highest bmi, a median bmi of 53.13 & the Northeastern people have the lowest bmi, a median bmi of 15.96.

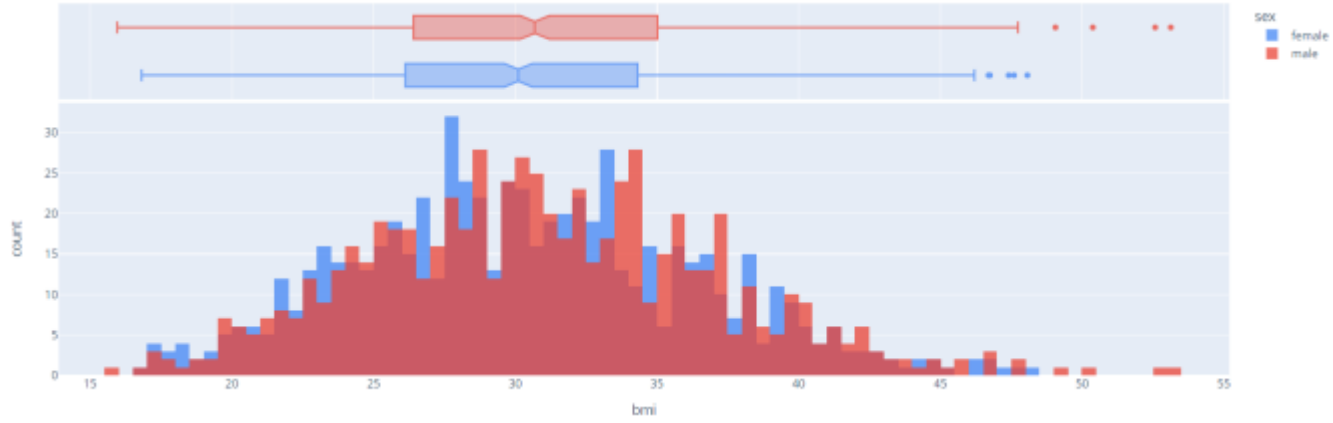



Figure 2.4.

The above figure shows gender-wise BMI distribution. We can conclude that the BMI of a person is independent of their gender.

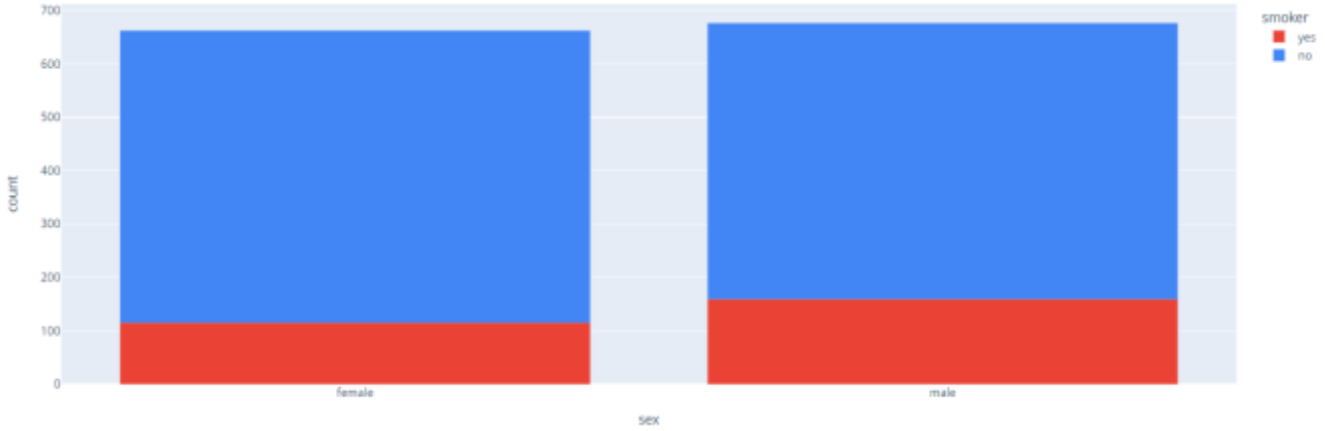



Figure 2.5.

The graph shows the gender-wise smoker distribution, we can conclude that people smoke regardless of their gender. More male beneficiaries seem to be smokers than female beneficiaries, but the difference is minimal, also the number of female beneficiaries tested are slightly less than the number of male beneficiaries.

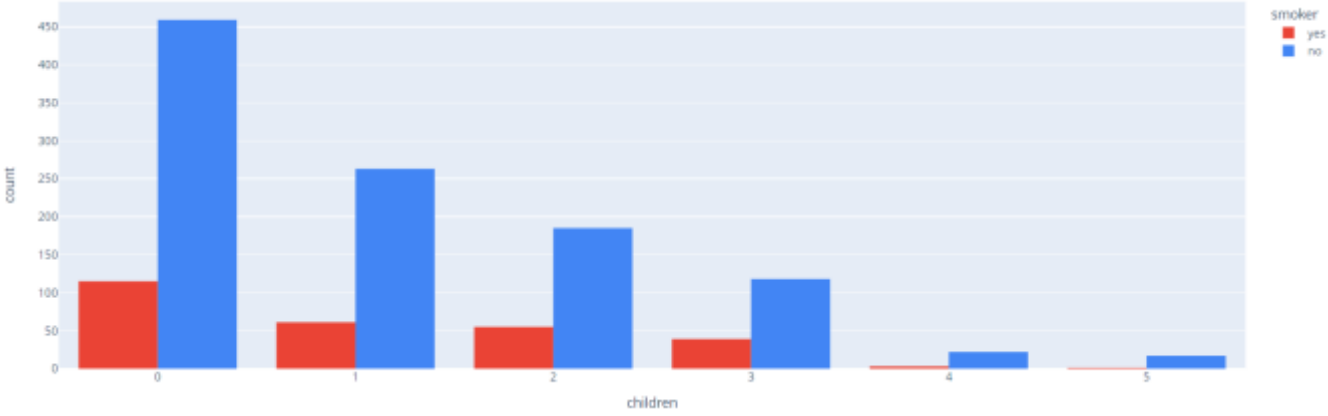



Figure 2.6.

The above bar plot shows that Smokers usually have less children than non-smokers.

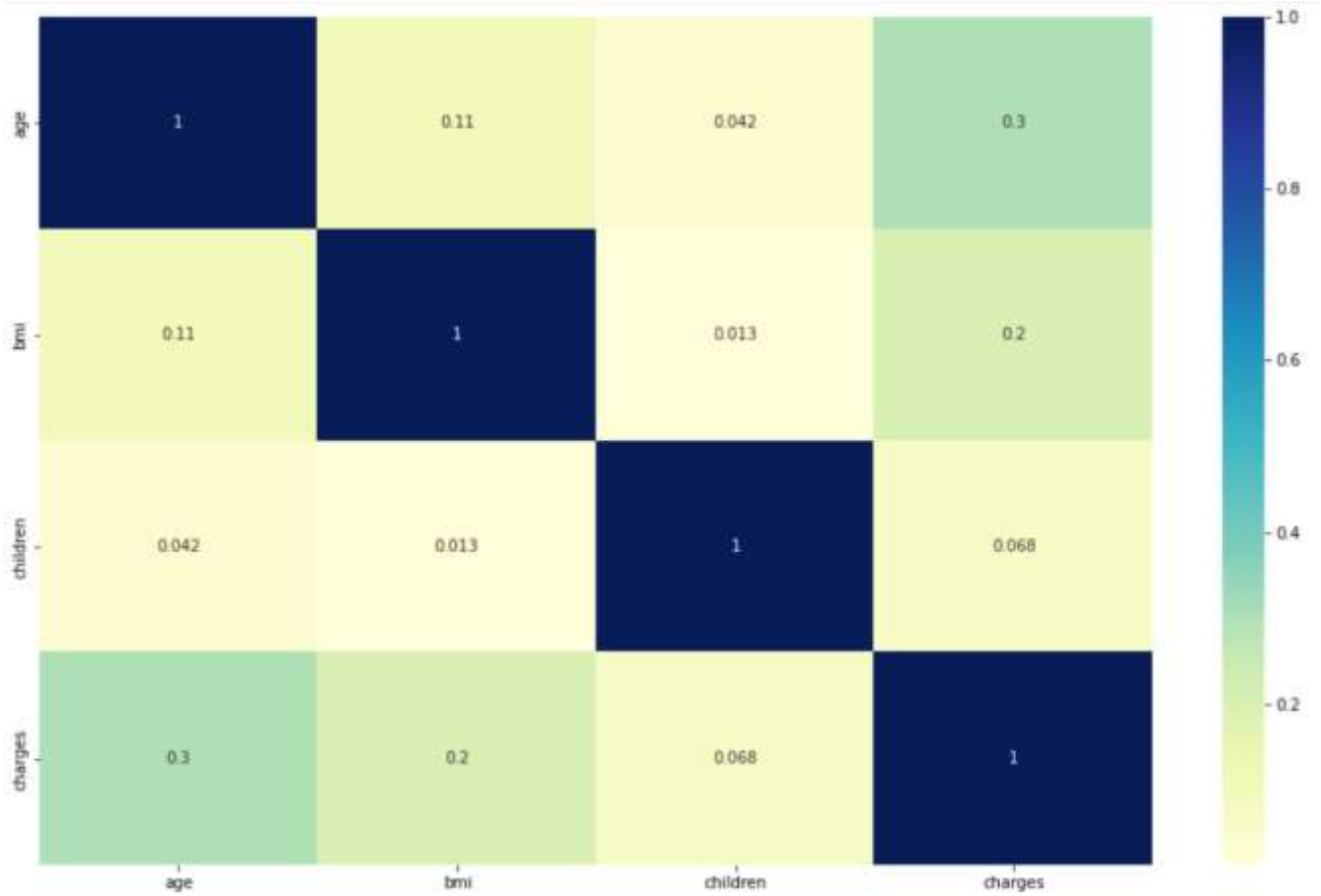



Figure 2.7.

From the above heat-map we can clearly see that there are no correlation between independent variables. Age and the charges show some correlation between them.

# 3. Transforming the Data

## 3.1. Encoding categorical variables:

Sex of the insurance contractor, Smoking status of insurance beneficiary & Region where these beneficiaries reside are the 3 categorical or character variables. In order to train the regression models we need to convert the categorical variables into integer values. The process of converting the categorical variable to integer is called Encoding of categorical variables.

For the insurance dataset, we have used **Label Encoding** which refers to converting the labels into a numeric form so as to convert them into the machine-readable form, for binary categories (Sex, Smoker) converting the values of categorical variables into 0s and 1s. For more than one category we have used **Dummy Encoding** that creates new variable for each category with values of 0s and 1s.

| | age | sex | bmi | children | smoker | charges | region_northeast | region_northwest | region_southeast | region_southwest |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 19 | 0 | 27.900 | 0 | 1 | 16884.92400 | 0 | 0 | 0 | 1 |
| 1 | 18 | 1 | 33.770 | 1 | 0 | 1725.55230 | 0 | 0 | 1 | 0 |
| 2 | 28 | 1 | 33.000 | 3 | 0 | 4449.46200 | 0 | 0 | 1 | 0 |
| 3 | 33 | 1 | 22.705 | 0 | 0 | 21984.47061 | 0 | 1 | 0 | 0 |
| 4 | 32 | 1 | 28.880 | 0 | 0 | 3866.85520 | 0 | 1 | 0 | 0 |

## 3.2. Data Normalization:

The process of standardizing variables present in the data in a fixed range or same scale is referred to as Normalization. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.

In this case, variables Age and BMI are of 2 units but the target variable Charges is in the range of around 1k to 63.7k which highly varying. We used Min-Max scaler to scale our variables, where MinMaxScaler rescales variables into the range [0,1] by default using the formula: y = (x – min) / (max – min).

| | age | sex | bmi | children | smoker | charges | region_northeast | region_northwest | region_southeast | region_southwest |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.021739 | 0 | 0.321227 | 0.0 | 1 | 0.251611 | 0 | 0 | 0 | 1 |
| 1 | 0.000000 | 1 | 0.479150 | 0.2 | 0 | 0.009636 | 0 | 0 | 1 | 0 |
| 2 | 0.217391 | 1 | 0.458434 | 0.6 | 0 | 0.053115 | 0 | 0 | 1 | 0 |
| 3 | 0.326087 | 1 | 0.181464 | 0.0 | 0 | 0.333010 | 0 | 1 | 0 | 0 |
| 4 | 0.304348 | 1 | 0.347592 | 0.0 | 0 | 0.043816 | 0 | 1 | 0 | 0 |

# 4. Hyper-parameter Tuning

A Machine Learning model is defined as a mathematical model with a number of parameters that need to be learned from the data. By training a model with existing data, we are able to fit the model parameters. These parameters are configuration variables that are internal to the model whose value can be estimated by given data (eg. Coefficients in Linear Regression).

There are another kind of parameters, known as **Hyperparameters** which are configuration variables that are external to the model whose value cannot be estimated by given data (eg. Learning rate). The hyperameters cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.

Some examples of model hyperparameters include:

1. The penalty in Logistic Regression Classifier i.e. L1 or L2 regularization
2. The learning rate for training a neural network.
3. The C and sigma hyperparameters for support vector machines.
4. The k in k-nearest neighbors.

Hyperparameter tuning is basically searching the hyperparameter space for set of values that will optimize our model architecture. In this case, we have used GridSearchCV method for hyperparameter tuning of models. In GridSearchCV approach, machine learning model is evaluated for a range of hyperparameter values. This approach is called GridSearchCV, because it searches for best set of hyperparameters from a grid of hyperparameters values. The gridsearch technique will construct many versions of the model with all possible combinations of hyerparameters, and will return the best one.

As in the image, for C = [0.1, 0.2, 0.3, 0.4, 0.5] and Alpha = [0.1, 0.2, 0.3, 0.4].
For a combination ***C=0.3 and Alpha=0.2***, performance score comes out to be **0.726(Highest)**, therefore it is selected.

| C \ Alpha | 0.1 | 0.2 | 0.3 | 0.4 |
|---|---|---|---|---|
| 0.5 | 0.701 | 0.703 | 0.697 | 0.696 |
| 0.4 | 0.699 | 0.702 | 0.698 | 0.702 |
| 0.3 | 0.721 | 0.726 | 0.713 | 0.703 |
| 0.2 | 0.706 | 0.705 | 0.704 | 0.701 |
| 0.1 | 0.698 | 0.692 | 0.688 | 0.675 |

# 5. Steps Performed

**1. Install necessary packages:**

- matplotlib
- seaborn
- sklearn
- numpy
- pandas
- plotly

**2. Import the necessary libraries.**

**3. Load/Read the data from working directory.**

The dataset insurance.csv is a comma separated file that contains 1338 observations & 7 variables.

**4. Developing various plots for Exploratory Data Analysis (EDA).**

**5. Label Encode and Dummy Encode categorical variables.**

The binary categorical variable sex, smoker are label encoded, multi-category variable is dummy encoded.

**6. Normalize numerical variables using Min-Max scaler.**

- age
- bmi
- children
- charges

**7. Splitting the dataset into Training and Test dataset with 80% being the training data and 20% being the test data.**

**8. Building Models.**

In this project, we have used Regression models that are used to predict any quantitative (numerical) data.

- *Linear Regression*: It is a linear approximation of casual relationship between two or more variables. This form of analysis estimates the coefficients of the linear equation, involving one or more independent variables (features/predictors) that best predict the value of the target variable.
- *Random Forest Regressor*: It is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging.
- *Bayesian Ridge*: It allows a natural mechanism to survive insufficient data or poorly distributed data by formulating linear regression using probability distributors rather than point estimates. The output or response is assumed to drawn from a probability distribution rather than estimated as a single value.
- *Support Vector Regression:* It is a supervised learning algorithm that is used to predict discrete values. Support Vector Regression uses the same principle as the

SVMs. The basic idea behind SVR is to find the best fit line. In SVR, the best fit line is the hyperplane that has the maximum number of points.

- *SGD Regressor:* SGD is a simple yet efficient optimization algorithm used to find the values of parameters/coefficients of functions that minimize a cost function. SGD regressor basically implements a plain SGD learning routine supporting various loss functions and penalties on loss function to shrink the model parameters to fit linear regression models.

**9. Predicting Values.**

Trained models are applied on test data in order to evaluate model's performance.

**10. Evaluating model performance based on predictions.**

R-Squared Score and Root Mean Squared Error (RMSE) is used for evaluating models performance.

**11. Interpreting results.**

# 6. Results

| | Mean Squared Error | R-Squared Score |
|---|---|---|
| **SGDRegressor** | 0.101659 | 0.745103 |
| **SupportVectorRegression** | 0.092618 | 0.788427 |
| **BayesianRidge** | 0.090097 | 0.799786 |
| **LinearRegression** | 0.090046 | 0.800015 |
| **RandomForestRegressor** | 0.069377 | 0.881286 |

Figure 6.1.

Plotting the performance measures R-Squared Score – how much better our regression line is than a simple horizontal line through the mean of the data. Closer to 1 better the model, Root Mean Squared Error (RMSE) – gives an absolute number on how much you predicted results deviate from the actual values. The lower the RMSE, the better a given model is able to "fit" a dataset.

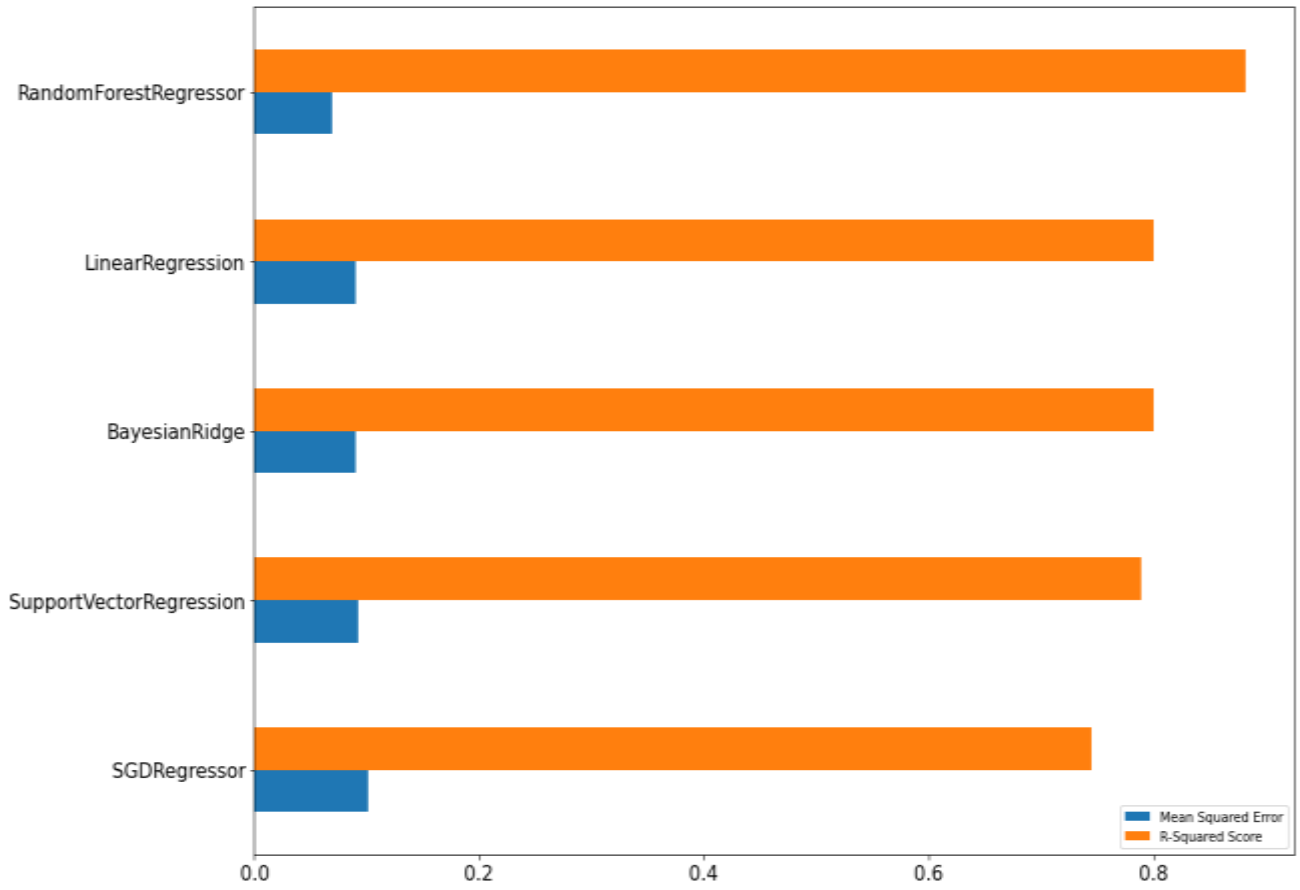



Figure 6.2.

From the above graph we can conclude that, RandomForestRegressor has maximum R-Squared score & least RMSE values as compared to Linear regression, Bayes ridge, Support vector regression & SGD regressor. Hence, RandomForestRegressor is the preferred model for predicting medical insurance.

# 7. Conclusion

People's healthcare cost prediction is now a valuable tool for improving healthcare accountability. The healthcare sector produces a very large amount of data related to patients, diseases, and diagnosis, but since it has not been analyzed properly, it does not provide the significance which it holds along with the patient healthcare cost.

A health insurance policy is a policy that covers or minimizes the expenses of losses caused by a variety of hazards. A variety of factors influence the cost of insurance or healthcare, in our case we have seen in Figure 2.1. that individuals who smoke have a higher medical bill than individuals who do not smoke and in Figure 2.2. Age and insurance charges have correlation the more age the more medical bills.

For a variety of stakeholders and health departments, accurately predicting individual healthcare expenses using prediction models is critical. From our preferred model (RandomForestRegressor) we can predict insurance charges with accuracy of 88%. Accurate cost estimates can help health insurers and, increasingly, healthcare delivery organizations to plan for the future and prioritize the allocation of limited care management resources.

Furthermore, knowing ahead of time what their probable expenses for the future can assist patients to choose insurance plans with appropriate deductibles and premiums. These elements play a role in the development of insurance policies.

# 8. References

## Theory references:


a) https://rpubs.com/tnguyen2921/611065
b) http://rstudio-pubs-static.s3.amazonaws.com/384487_d29d51d197bb429bb661312d2a388da0.html
c) https://machinelearningmastery.com/standardscaler-and-minmaxscaler-transforms-in-python/
d) https://www.geeksforgeeks.org/hyperparameter-tuning/
e) https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/
f) https://www.ibm.com/topics/linear-regression
g) https://www.statology.org/what-is-a-good-rmse/
h) https://www.geeksforgeeks.org/random-forest-regression-in-python/
i) https://www.tutorialspoint.com/scikit_learn/scikit_learn_bayesian_ridge_regression.htm
j) https://towardsdatascience.com/unlocking-the-true-power-of-support-vector-regression-847fd123a4a0
k) https://www.datatechnotes.com/2020/09/regression-example-with-sgdregressor-in-python.html
l) https://learning.imarticus.org/


## Code References:


a) https://plotly.com/python/getting-started/
b) https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html
c) https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html
d) https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html
e) https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
f) https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html
g) https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html
h) https://pandas.pydata.org/docs/reference/api/pandas.get_dummies.html
i) https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html