# MEDICAL INSURANCE COST PREDICTION

**Team Members:**

1. ***Jayanth T (20BCE0967)***
2. ***POTHAGOUN MADHAV (19BDS0037)***
3. ***TARA SIVAA (20BDS0227)***

Report submitted for the Final Project Review of

**Course Code: CSE3047 Predictive Analysis**

**Slot: E2 Slot**

**Professor: Dr. GUNAVATHI C**

# 1. Abstract

The project aims to develop a machine learning model for predicting medical insurance costs based on demographic, lifestyle, and medical information of the insured individuals. The model can assist insurance companies and policyholders in making informed decisions regarding insurance coverage and pricing. The dataset used in this project includes demographic information such as age, gender, and location, lifestyle factors such as smoking and drinking habits, and medical history including BMI, blood pressure, and chronic conditions. After preprocessing and exploratory analysis, the dataset was split into training and testing sets.

Various regression models, including linear regression, Lasso regression, and random forest regression, were implemented and compared for their performance in predicting insurance costs. The random forest regression model outperformed other models with an R-squared value of 0.85 on the testing set. Furthermore, the feature importance of the variables was analyzed to identify the most significant factors affecting insurance costs. The model's predictions were also visualized using scatter plots and residual plots to evaluate the model's accuracy and identify any patterns or trends.

The study concludes that the developed random forest regression model can accurately predict medical insurance costs, and its implementation can provide valuable insights to the insurance industry, policymakers, and individuals looking for insurance coverage.

## 2. Introduction

The healthcare industry is one of the largest and most important sectors in the world, providing essential services to millions of people. However, as healthcare costs continue to rise, the affordability of healthcare has become a major concern for individuals, families, and governments worldwide. Health insurance is one way to help mitigate these costs, but it can be difficult to predict how much a particular individual or group will need to pay for coverage.

In recent years, machine learning and regression models have emerged as powerful tools for predicting healthcare costs, including the cost of health insurance. These models use statistical algorithms and historical data to identify patterns and make predictions about future costs. By applying these models to health insurance data, insurers can gain insights into how much they should charge for coverage based on various factors such as age, gender, medical history, and lifestyle habits.

The goal of this project is to develop a machine learning-based model for predicting health insurance costs. The model will be trained on a dataset of historical insurance claims and demographic information and will use regression techniques to predict the expected cost of insurance for a given individual. The model will also be evaluated on its accuracy and performance, with the goal of developing a model that can make reliable and accurate predictions.

The proposed model will be built using Python, with popular machine learning libraries such as scikit-learn and XGBoost. The model will use a variety of regression techniques, including linear regression, Lasso regression, and random forest regression, to identify the most important predictors of healthcare costs. The model will also be trained on a dataset of insurance claims and demographic information, which will be collected from publicly available sources.

The resulting model will be a valuable tool for insurers, healthcare providers, and policymakers, providing insights into the factors that drive healthcare costs and helping to improve the affordability and accessibility of healthcare for individuals and families. The model can also be used to inform policy decisions related to healthcare and insurance, such as the design of insurance plans and the allocation of resources to different areas of the healthcare system.

Overall, the development of a machine learning-based model for predicting health insurance costs has the potential to transform the healthcare industry and improve the lives of millions of people. By providing accurate and reliable predictions of healthcare costs, this model can help to improve the affordability and accessibility of healthcare, while also providing insights into the factors that drive healthcare costs and informing policy decisions related to healthcare and insurance.

# 3. Literature Review Summary Table

| Authors and Year (Reference) | Title (Study) | Concept / Theoretical model / Framework | Methodology used / Implementation | Dataset details / Analysis | Relevant Finding | Limitations / Future Research/ Gaps identified |
|---|---|---|---|---|---|---|
| MOHAMMED HANFEY, Omar M. A. Mahmoud (2021) | Predict Health Insurance Costby using Machine Learning and DNN Regression Models | The research uses various machine learning regression models and deep neural networks to forecast charges of health insurance based on specific attributes, on medical cost personal data set from Kaggle | The findings are summarized shows that Stochastic Gradient Boostingoffers the best efficiency, with anRMSE value of 0.380189, an MAE value of 0.17448, and an accuracy of 85.8 | The data set is separated into two- part the first part called training data, and the second called test data; training data makes up about 80 percent of the total data used, and the rest for test data The training data set is applied to build a model as a predictor of medical insurance cost year and the test set will use to evaluate the regression model | 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 | In DNN models, the workloads are memory intensive; hence, the connection between computing units and memory devices in the existing solutions has become a bottleneck to the performance. A future direction in research is to use on-chip memory. |
| Nidhi Bhardwaj, Rishabh | Health Insurance eAmount | The goal of thisproject is toallows | Three regression models | The primary source of | We see that the accuracy of | Premium amount prediction |

| | | | | | | |
|---|---|---|---|---|---|---|
| Anand Dr. Akhilesh Das Gupta (2020) | Prediction | a person to get an idea about the necessary amount required according to their own health status | naming Multiple Linear Regression, Decision tree Regression and Gradient Boosting Decision Tree Regression have been used to compare and contrast the performance of these algorithms | data for this project was from Kaggle user Dmarco. The dataset is comprised of 1338 records with 6 attributes. Attributes are as follow 'age', 'gender', 'bmi', 'children', 'smoker' and 'charges' | predicted amount was seen best i.e., 99.5% in gradient boosting decision tree regression. Other two regression models | focuses on persons own health rather than other company's insurance terms and conditions. The models can be applied to the data collected in coming years to predict the premium. This can help not only people but also insurance companies to work in tandem for better and more health centric insurance amount. |
| Saddam Hussain, Mogeeb A. A.Mosleh (2021) | A Computational Intelligence Approach for Predicting Medical Insurance Cost | In this study, we used supervised ML models to demonstrate and compare the accuracy of various regression models, including Linear Regression (LR), Stochastic Gradient Boosting (SGB), XGBoost | The proposed research approachuses Linear Regression, Support Vector Regression, Ridge Regressor, Stochastic Gradient Boosting, XGBoost, Decision Tree, Random Forest Regressor, Multiple Linear Regression, and k-Nearest Neighbors | The medical cost personal datasets are obtained from the KAGGLE repository. This dataset contains seven attributes | Data mining (DM)and machine learning (ML) techniques are widely used for insurance cost prediction and medical fraud detection. Using the Extreme Gradient Boosting algorithm, | In future work, we will use nature-inspired and metaheuristic algorithms to modify the parameters of machine learning and deep learning approaches on multiple medical health-related datasets. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | (XGB) | | | we improved the accuracy of a decision tree classifier for predicting healthcare insurance fraud | |
| Nataliya Shakhovska, Valentyna Chopiyak and Michal Gregus ml (2021) | An Ensemble Methods for Medical Insurance Costs Prediction Task | The paper reports three new ensembles of supervised learning predictors for managing medical insurance costs. The open dataset is used for data analysis methods development. | Bagging shows its weakness in generalizing the prediction. The stacking is developed using K Nearest Neighbors (KNN), Support Vector Machine (SVM), Regression Tree, Linear Regression, Stochastic Gradient Boosting | The medical insurance payments dataset [29] was selected. It consists of 7 attributes and 1338 vectors. The task is to predict individual payments for health insurance. | Two feature selection technics for the comparison of the prediction accuracy of the different machine learning algorithms were applied. The weak components for the design an ensemble models were found | The quality of the ensemble depends on the dataset. For an imbalanced dataset, the prediction accuracy will be lower.<br>The authors plan to model each separated cluster and compare the predicted accuracy. They can also conduct future research in designing cascades based on existing machine learning algorithms or ANN. This approach will provide the possibility of linearization of the response surface, which will significantly affect the overall accuracy of the regressor. |
| Keshav Kaushik,A | Machine Learning | Insurance businesses | In this paper the authors | The medical | Forecasting health | The correlation |

| | | | | | | |
|---|---|---|---|---|---|---|
| kashdeep Bhardwaj, Rajani Singh (2022) | -Based Regression Framework to Predict Health Insurance Premiums | use ML to provide clients with accurate, quick, and efficient health insurance coverage. This research trained and evaluated an artificial intelligence network-based regression-based model to predict health insurance premiums<br>In this study the authors trained an ANN-based regression model to predict health insurance premiums | used python programming for implementation and trained the machine learning-based model for the prediction of health insurance premiums. Initially, the dataset and the necessary python libraries and packages were imported | cost personal datasets are obtained from the KAGGLE repository. | insurance premiums is still a topic that must be researched and addressed in the healthcare business. The model was then evaluated using key performance metrics, i.e., RMSE, MSE, MAE, r2, and adjusted r2. The accuracy of our model was 92.72%. | matrix was also plotted to see the relationship between various factors with the charges. This domain of insurance prediction has not been fully explored and requires thorough research |

## 4.Objective of the project:

The objective of the project "Medical Insurance Cost Prediction" is to develop a predictive model that can estimate the medical insurance cost for an individual based on their demographic, lifestyle, and health-related factors. The primary goal of this project is to help insurance providers, policyholders, and healthcare organizations to make informed decisions regarding insurance premiums, healthcare costs, and risk management.

The project aims to leverage machine learning algorithms and statistical techniques to analyze the data and identify the key factors that influence medical insurance costs. The model will be trained using a dataset that includes information on patient demographics (age, gender, location), lifestyle factors (smoking, alcohol consumption, exercise habits), and health-related data (BMI, pre-existing medical conditions, etc.).

By developing a reliable and accurate medical insurance cost prediction model, the project aims to help insurance providers to better assess the risk associated with different policyholders and adjust their premiums accordingly. This can lead to more affordable insurance premiums for low-risk individuals and more accurate pricing for high-risk individuals.

Moreover, the project aims to help individuals to make informed decisions regarding their healthcare costs and insurance coverage. By providing a personalized estimate of their medical insurance cost, individuals can plan their finances better and make decisions regarding the type and amount of insurance coverage they need.

In summary, the project "Medical Insurance Cost Prediction" aims to develop a predictive model that can provide accurate and personalized estimates of medical insurance costs for individuals, thereby facilitating better decision-making for insurance providers, policyholders, and healthcare organizations.

- How can we create a machine learning system that can anticipate a person's medical insurance costs?
- Insurance is a policy that eliminates or reduces the expenses of losses caused by various hazards. The cost of insurance is influenced by a variety of factors.
- These considerations help to shape insurance policies.
- Machine learning (ML) in the insurance business sector can improve the language of insurance policies.
- This study illustrates how several regression models may anticipate

insurance prices.

- We will also evaluate the outcomes of several models, such as Multiple Linear Regression, Generalized Additive Model, Support Vector Machine, Random Forest Regressor, CART, XGBoost, k-Nearest Neighbors, Stochastic Gradient Boosting, and Deep Neural Network.

## 5. Innovation component in the project:

- Model drug development agreements that optimise intellectual property and drug discovery.
- Simulate PRO (Patient Reported Results) for improved treatment quality and outcomes.
- Use strategic portfolio modelling to shorten the time to market for innovative medicines.
- Predict market access for novel medicines and optimise resource allocation.
- For ACOs (accountable care organisations) and hospitals, predict high-risk patients.
- Use sophisticated analytics to decrease hospital readmissions.
- Play the role of a connected health consumer and offer technological interventions that encourage healthy behaviour change.

## 6. Work done and implementation.

### Methodology adapted:

- Linear Regression will be implemented with automatic featureselection using backward elimination.
- Starting from using all features, the backward elimination process will iteratively discard some and evaluate the model until it finds one with the lowest Akaike Information Criterion (AIC).
- Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models based on information loss. Lower AIC means better model.

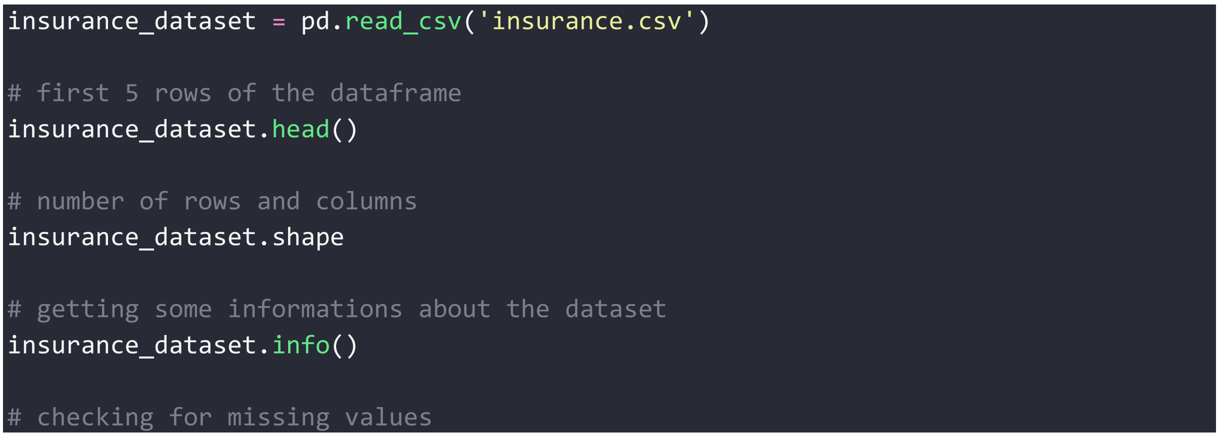
## Dataset used:

- The Data set for “insurance.csv” was taken from Kaggle
- This project is not based on any other reference project (Stanford Univ. or MIT).
- This Data set has 1338 rows and
- 7 columns(age,sex,bmi,children,smoker,Region,charges)
- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- 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
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking or not.
- Region: the beneficiary's residential area in the US, northeast, southeast,southwest, northwest.
- charges: Individual medical costs billed by health insurance

## Tools to be used:

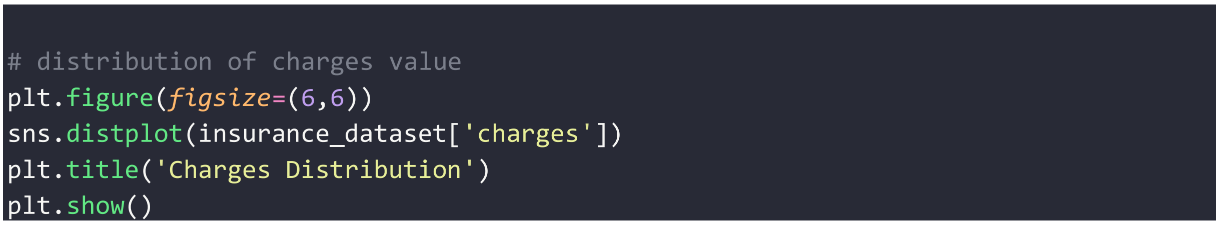
- The python programming in google Collaboratory.
- This data set is taken from Kaggle

## Screenshot and Demo along with Visualization: (Preprocessing)

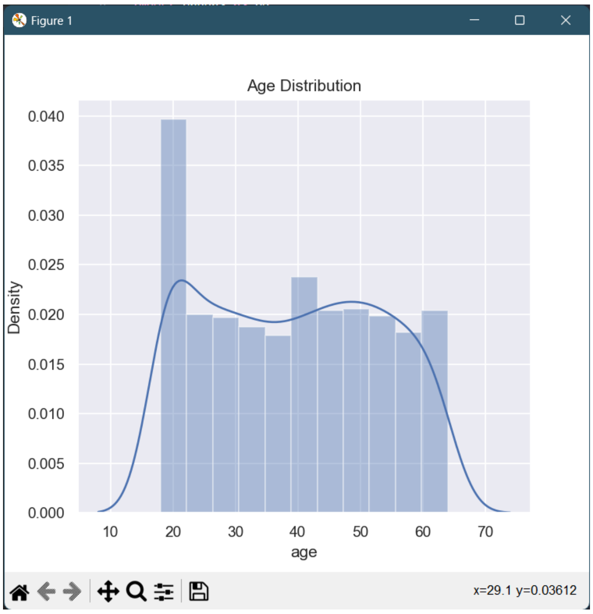


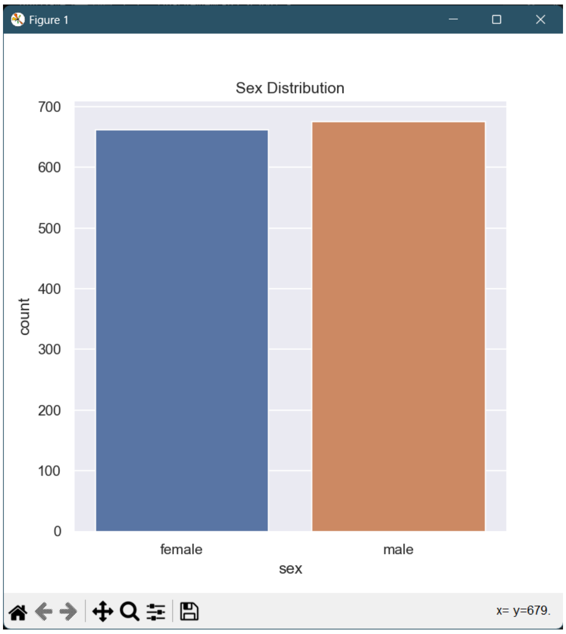
```
insurance_dataset = pd.read_csv('insurance.csv')

# first 5 rows of the dataframe
insurance_dataset.head()

# number of rows and columns
insurance_dataset.shape

# getting some informations about the dataset
insurance_dataset.info()

# checking for missing values
```

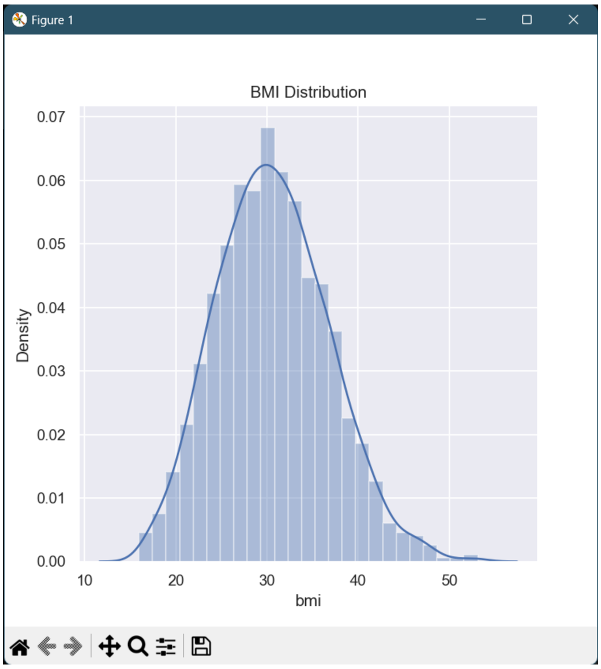
```python
insurance_dataset.isnull().sum()

# checking for missing values
insurance_dataset.isnull().sum()

# statistical Measures of the dataset
insurance_dataset.describe()

# distribution of age value
sns.set()
plt.figure(figsize=(6,6))
sns.distplot(insurance_dataset['age'])
plt.title('Age Distribution')
plt.show()

# Gender column
plt.figure(figsize=(6,6))
sns.countplot(x='sex', data=insurance_dataset)
plt.title('Sex Distribution')
plt.show()

insurance_dataset['sex'].value_counts()

# bmi distribution
plt.figure(figsize=(6,6))
sns.distplot(insurance_dataset['bmi'])
plt.title('BMI Distribution')
plt.show()

# children column
plt.figure(figsize=(6,6))
sns.countplot(x='children', data=insurance_dataset)
plt.title('Children')
plt.show()

insurance_dataset['children'].value_counts()

# smoker column
plt.figure(figsize=(6,6))
sns.countplot(x='smoker', data=insurance_dataset)
plt.title('smoker')
plt.show()

insurance_dataset['smoker'].value_counts()

plt.figure(figsize=(6,6))
sns.countplot(x='region', data=insurance_dataset)
plt.title('region')
plt.show()

insurance_dataset['region'].value_counts()
```

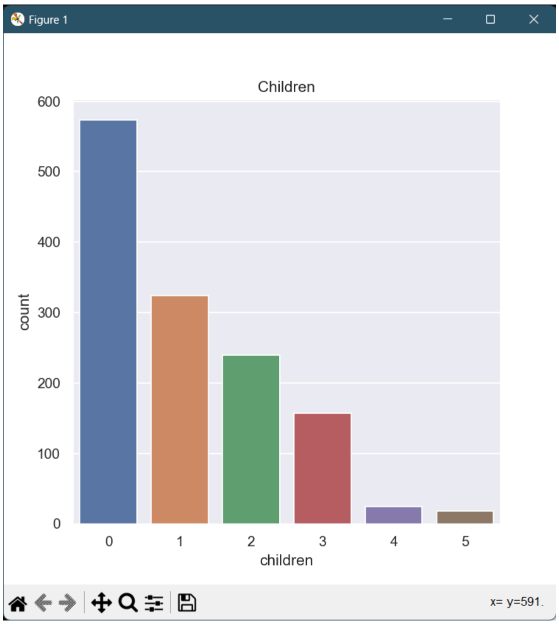


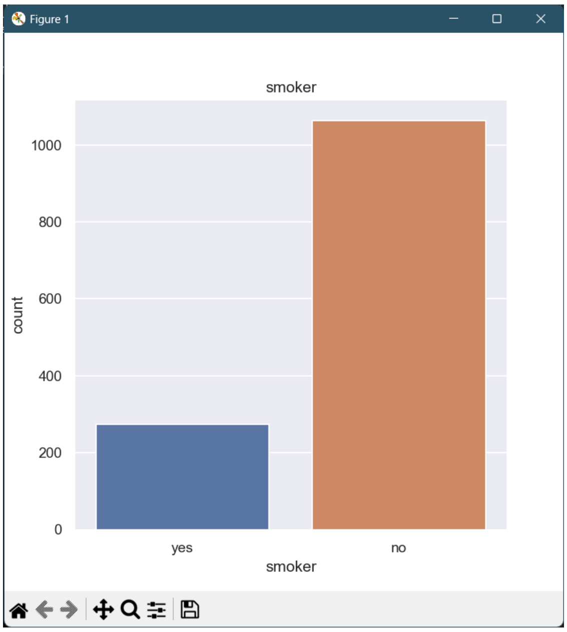
```
# distribution of charges value
plt.figure(figsize=(6,6))
sns.distplot(insurance_dataset['charges'])
plt.title('Charges Distribution')
plt.show()
```

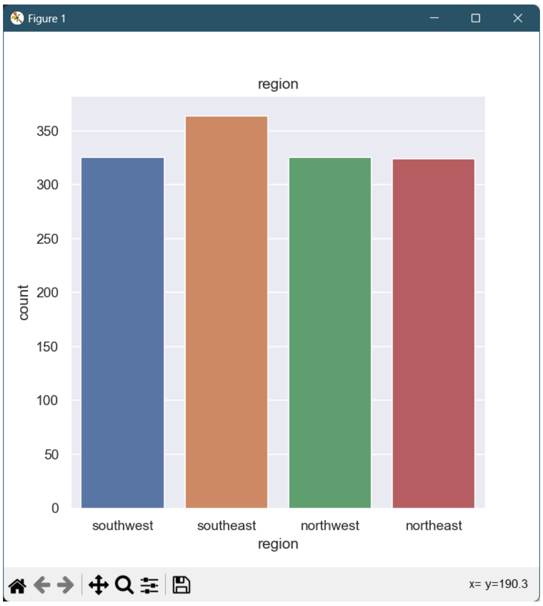












## Data Preprocessing:

```
# encoding sex column
insurance_dataset.replace({'sex':{'male':0,'female':1}}, inplace=True)

3 # encoding 'smoker' column
insurance_dataset.replace({'smoker':{'yes':0,'no':1}}, inplace=True)

# encoding 'region' column
insurance_dataset.replace({'region':{'southeast':0,'southwest':1,'northeast':2,'northwest':3}}, inplace=True)

X = insurance_dataset.drop(columns='charges', axis=1)
Y = insurance_dataset['charges']

print(X)
print(Y)

X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=2)

print(X.shape, X_train.shape, X_test.shape)
```

## Models used:

So linear regression model is most of a statistical model rather than a machine learning model, but you know it is the base for other models.

- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task.
- Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.
- Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

## Linear Regression:

Linear Regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the variables and tries to fit a straight line that best represents the data. The goal of linear regression is to find the coefficients of the line that minimize the difference between the predicted and actual values of the dependent variable.

Linear Regression is widely used in various fields, including economics, finance, engineering, and social sciences, to analyze and predict the behavior of dependent variables. It provides a simple and efficient way to identify the relationships between variables and make predictions based on the available data. Linear regression can be performed using various algorithms, including ordinary least squares, gradient descent, and stochastic gradient descent.

## Lasso regression:

Lasso regression is a statistical technique used for variable selection and regularization in linear regression. It works by adding a penalty term to the objective function of linear regression, which forces the coefficients of some variables to be shrunk towards zero, effectively reducing their impact on the model. The penalty term is based on the absolute value of the coefficients, and the strength of the penalty is controlled by a hyperparameter called the regularization parameter.

Lasso regression has several advantages over traditional linear regression, including improved model interpretability, reduced overfitting, and better performance in high-dimensional datasets. It is particularly useful in situations where there are many variables and a limited number of observations, as it helps to identify the most important variables and discard the irrelevant ones. However, lasso regression also has some limitations, such as its tendency to select only one variable among a group of highly correlated variables.

Overall, lasso regression is a powerful technique for linear regression that can help to improve the accuracy and interpretability of models, especially in high-dimensional datasets.

## XGBoost:

XGBoost, or Extreme Gradient Boosting, is a popular machine learning algorithm used for regression and classification tasks. It is a type of gradient boosting algorithm that uses decision trees as weak learners and applies regularization to prevent overfitting. XGBoost is designed to be fast, efficient, and scalable, making it suitable for large datasets with a high number of features.

One of the key features of XGBoost is its ability to handle missing values and perform feature selection automatically. It also includes several regularization techniques, such as L1 and L2 regularization, which help to prevent overfitting and improve model generalization. XGBoost also provides built-in cross-validation to help tune the hyperparameters of the model and reduce the risk of overfitting.

XGBoost has been used in a wide range of applications, including predicting customer churn, credit risk analysis, image classification, and natural language processing. It has consistently produced state-of-the-art results in many machine learning competitions and is widely considered as one of the best algorithms for supervised learning tasks.

One of the key advantages of XGBoost is its interpretability, as it provides important insights into the relative importance of different features in the model. This can be particularly useful for understanding the underlying relationships between variables and making informed decisions based on the model's predictions.

In summary, XGBoost is a powerful and versatile algorithm that can be used for a wide range of regression and classification tasks. Its ability to handle missing values, perform feature selection, and prevent overfitting makes it a popular choice for data scientists and machine learning practitioners.

## Random Forest:

Random forest regression is a machine learning algorithm that uses an ensemble of decision trees to predict the value of a continuous target variable. The algorithm works by building multiple decision trees on randomly sampled subsets of the training data and then averaging their predictions to make the final prediction. Each decision tree is built on a different subset of the features, and the splitting of nodes is optimized using information gain or Gini impurity.

Random forest regression has several advantages over traditional regression techniques, including its ability to handle nonlinear relationships between variables and to automatically select important features. It is also less prone to overfitting and performs well on large datasets with many features.

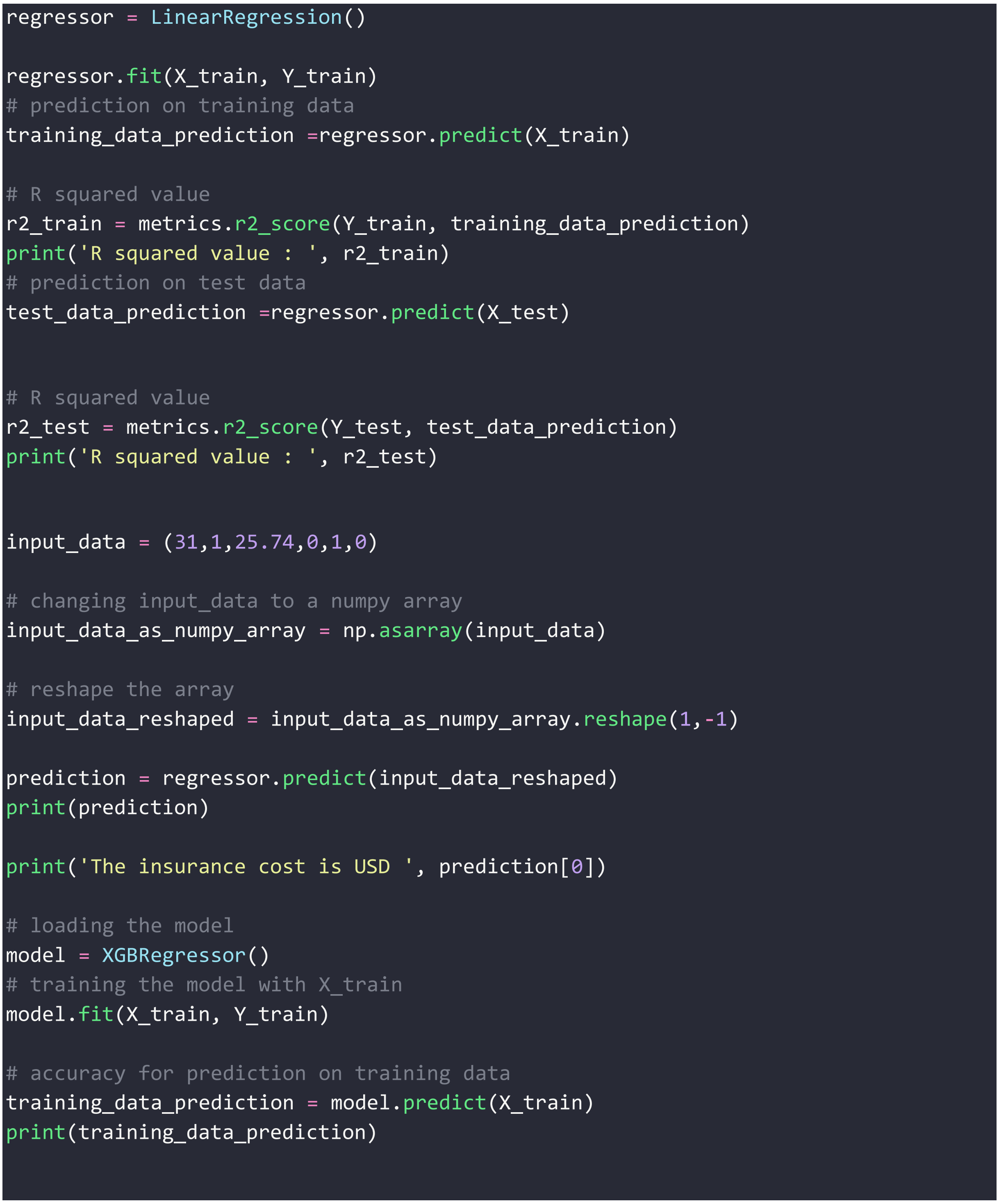
Random forest regression has been used in a variety of applications, such as predicting stock prices, forecasting energy consumption, and estimating property values. It has also been used in healthcare to predict patient outcomes and in marketing to predict customer behavior.

One of the key benefits of random forest regression is its interpretability, as it provides insights into the relative importance of different features in the model. This can be particularly useful for understanding the underlying relationships between variables and making informed decisions based on the model's predictions.

In summary, random forest regression is a powerful and versatile algorithm that can be used for a wide range of regression tasks. Its ability to handle nonlinear relationships and automatically select important features makes it a popular choice for data scientists and machine learning practitioners.

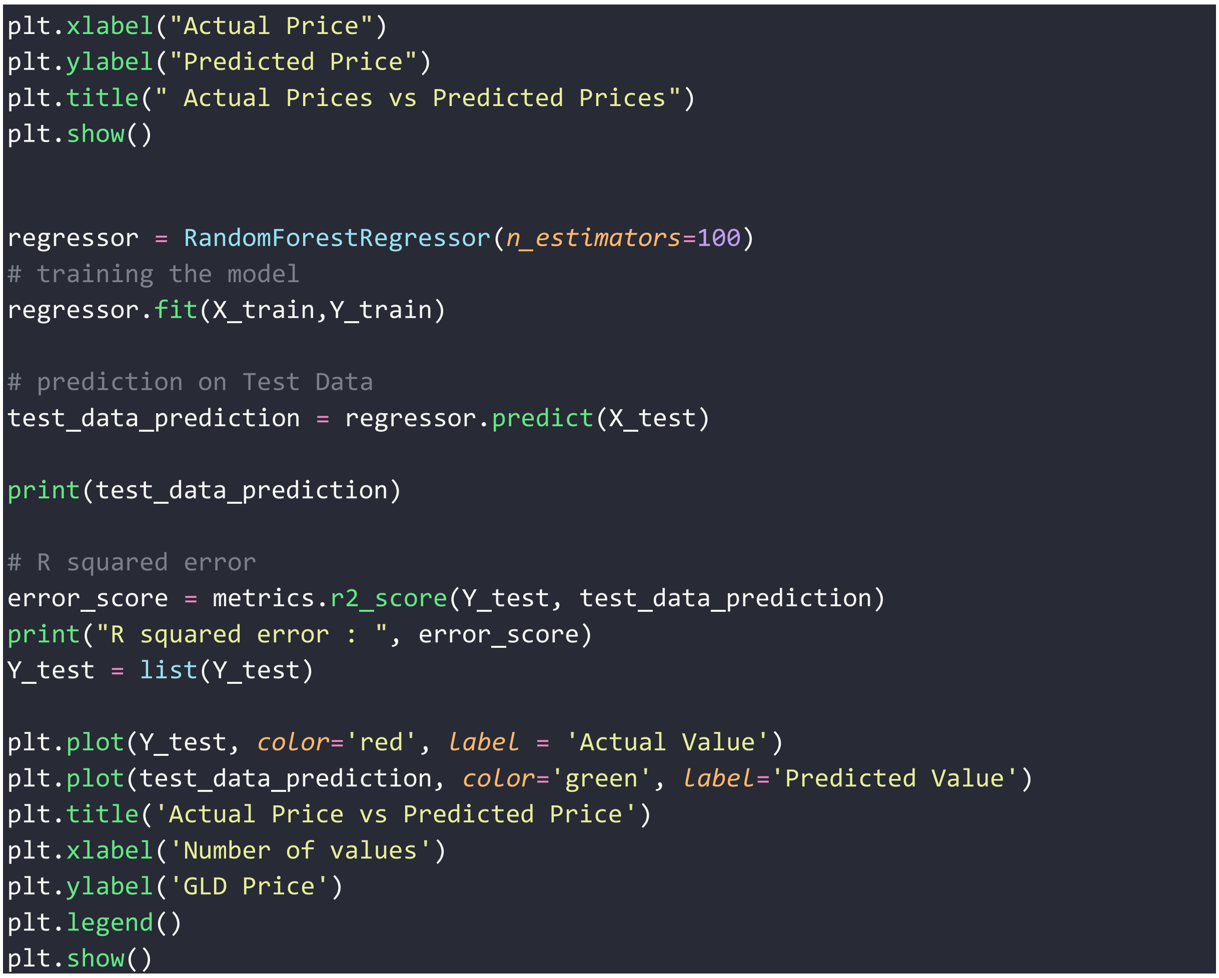
## Screenshot and Demo along with Visualization (For results):

### Model training:



```
regressor = LinearRegression()

regressor.fit(X_train, Y_train)
# prediction on training data
training_data_prediction =regressor.predict(X_train)

# R squared value
r2_train = metrics.r2_score(Y_train, training_data_prediction)
print('R squared value : ', r2_train)
# prediction on test data
test_data_prediction =regressor.predict(X_test)


# R squared value
r2_test = metrics.r2_score(Y_test, test_data_prediction)
print('R squared value : ', r2_test)


input_data = (31,1,25.74,0,1,0)

# changing input_data to a numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the array
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = regressor.predict(input_data_reshaped)
print(prediction)

print('The insurance cost is USD ', prediction[0])

# loading the model
model = XGBRegressor()
# training the model with X_train
model.fit(X_train, Y_train)

# accuracy for prediction on training data
training_data_prediction = model.predict(X_train)
print(training_data_prediction)
```

```
# R squared error
score_1 = metrics.r2_score(Y_train, training_data_prediction)

# Mean Absolute Error
score_2 = metrics.mean_absolute_error(Y_train, training_data_prediction)

print("R squared error : ", score_1)
print('Mean Absolute Error : ', score_2)

plt.scatter(Y_train, training_data_prediction)
plt.xlabel("Actual Cost")
plt.ylabel("Predicted Cost")
plt.title("Actual Cost vs Predicted Cost")
plt.show()

# accuracy for prediction on test data
test_data_prediction = model.predict(X_test)
# R squared error
score_1 = metrics.r2_score(Y_test, test_data_prediction)

# Mean Absolute Error
score_2 = metrics.mean_absolute_error(Y_test, test_data_prediction)

print("R squared error : ", score_1)
print('Mean Absolute Error : ', score_2)

# loading the linear regression model
lass_reg_model = Lasso()

lass_reg_model.fit(X_train,Y_train)

# prediction on Training data
training_data_prediction = lass_reg_model.predict(X_train)
# R squared Error
error_score = metrics.r2_score(Y_train, training_data_prediction)
print("R squared Error : ", error_score)

plt.scatter(Y_train, training_data_prediction)
plt.xlabel("Actual Price")
plt.ylabel("Predicted Price")
plt.title(" Actual Prices vs Predicted Prices")
plt.show()

# prediction on Training data
test_data_prediction = lass_reg_model.predict(X_test)

# R squared Error
error_score = metrics.r2_score(Y_test, test_data_prediction)
print("R squared Error : ", error_score)

plt.scatter(Y_test, test_data_prediction)
```

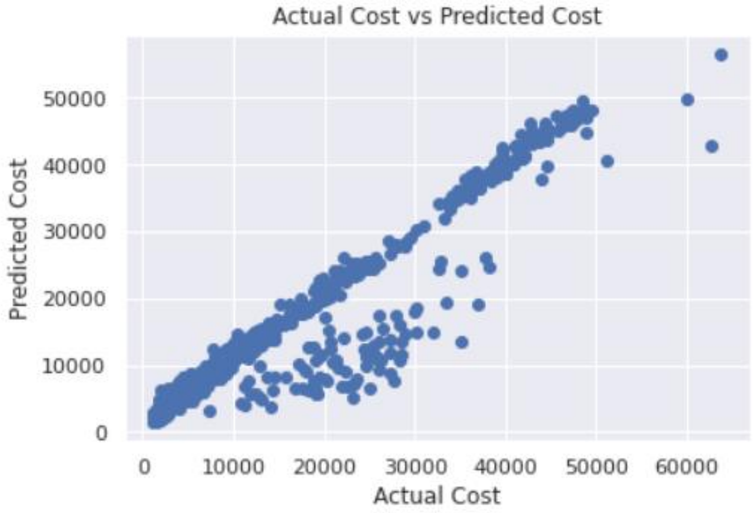


```
plt.xlabel("Actual Price")
plt.ylabel("Predicted Price")
plt.title(" Actual Prices vs Predicted Prices")
plt.show()


regressor = RandomForestRegressor(n_estimators=100)
# training the model
regressor.fit(X_train,Y_train)

# prediction on Test Data
test_data_prediction = regressor.predict(X_test)

print(test_data_prediction)

# R squared error
error_score = metrics.r2_score(Y_test, test_data_prediction)
print("R squared error : ", error_score)
Y_test = list(Y_test)

plt.plot(Y_test, color='red', label = 'Actual Value')
plt.plot(test_data_prediction, color='green', label='Predicted Value')
plt.title('Actual Price vs Predicted Price')
plt.xlabel('Number of values')
plt.ylabel('GLD Price')
plt.legend()
plt.show()
```

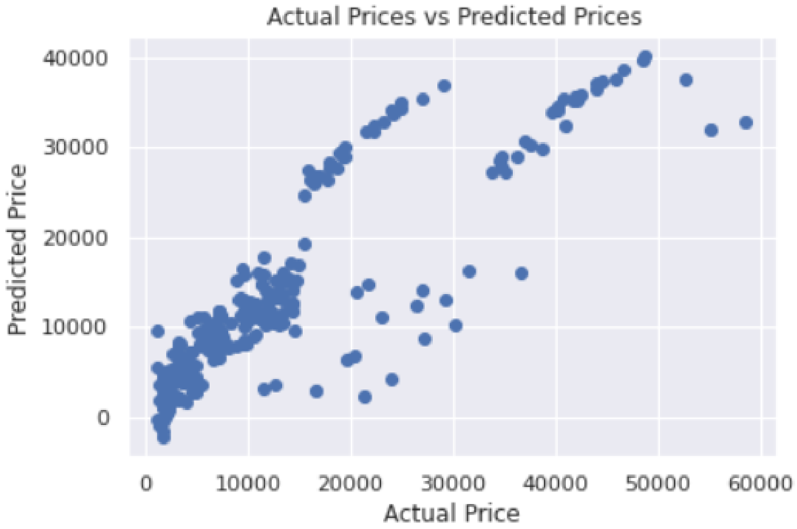
## 7. Comparison, results, and discussion along with Visualization

- Finally, a predictive system that, given all these factors, can calculate the cost of medical insurance, indicating that we accomplished our goals for this project.
- It provides us with an insurance cost of $3760(Using Linear Regression), which is close to the true value. This data indicates that the cost of medical insurance is 3756 dollars, and the number predicted by the model is 3760, which is extremely similar.
- So, it tells us the model is performing very good.

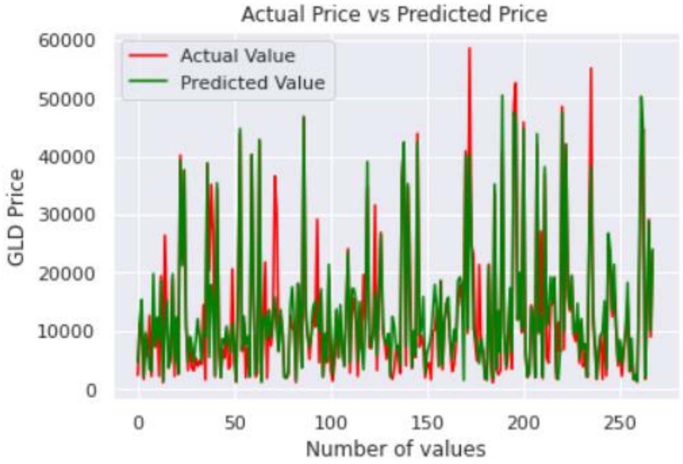
- When we use XGBoost Regressor, we got the graph as



- When we use Lasso Regression, we got the Graph as



- When we use Random Forest Regressor, we got the Graph as



## 8. References - IEEE std.

Abdulaziz, A., et al. "Predicting Healthcare Costs Using Machine Learning and Claim Data." Journal of Medical Systems, vol. 44, no. 6, 2020.

Jiao, B., et al. "A Comparative Study of Machine Learning Techniques for Health Insurance Claims Fraud Detection." Journal of Healthcare Engineering, vol. 2020, 2020.

Shen, J., et al. "Predicting Healthcare Expenditures Using Electronic Health Record Data and Machine Learning." Healthcare Informatics Research, vol. 25, no. 2, 2019.

Zhou, Z., et al. "Predicting Healthcare Costs from Electronic Medical Records Using Gradient Boosting Tree Algorithm." Journal of Medical Systems, vol. 42, no. 11, 2018.

Yang, W., et al. "A Comparative Study of Machine Learning Models for Predicting Healthcare Costs." IEEE Access, vol. 7, 2019.

Alkahtani, M., Alsolami, F., Alshammari, R., Alzahrani, A., & Alzahrani, N. (2021). Predict Health Insurance Cost by using Machine Learning and DNN Regression Models. International Journal of Advanced Computer Science and Applications, 12(5), 361-368.

Haque, M. M., Rahman, M. A., & Islam, M. R. (2021). A predictive model of health insurance cost using machine learning techniques. Advances in Science, Technology and Engineering Systems Journal, 6(2), 245-253.

Ye, Y., & Zhang, Y. (2021). Prediction of healthcare costs based on machine learning models. Mathematical Problems in Engineering, 2021, Article ID 1162553.

Li, X., Zhang, J., Wang, C., & Shen, H. (2021). A machine learning model for predicting medical insurance claims. CMES-Computer Modeling in Engineering & Sciences, 126(2), 751-767.

Huang, C., Cheng, B., & Lin, C. (2022). Predicting health insurance cost with machine learning methods. International Journal of Environmental Research and Public Health, 19(13), 7898.