NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY,
BELGAUM, APPROVED BY AICTE & GOVT. OF KARNATAKA

PROJECT REPORT

on

MEDICAL COST PREDICTION

Submitted in partial fulfilment of the requirement for the award of Degree of

Bachelor of Engineering

in

Computer Science and Engineering

Submitted by:

ANURAG NEPAL	1NT19CS036
AVISHEK RIJAL	1NT19CS045
BAIBHAV DHAKAL	1NT19CS048
NABIN KUMAR K.C.	1NT1NCS116

Under the Guidance of
Dr. Vani V
Professor , Dept. of CS&E, NMIT

Department of Computer Science and Engineering
(Accredited by NBA Tier-1)

2021-2022

NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY, BELGAUM
, APPROVED BY AICTE & GOVT.OF KARNATAKA)

Department of Computer Science and Engineering
(Accredited by NBA Tier-1)

CERTIFICATE

This is to certify that the ***Medical Cost Prediction*** is an authentic work carried out by **Anurag Nepal(1NT19CS036), Avishek Rijal(1NT19CS045), Baibhav Dhakal(1NT19CS045), and Nabin Kumar K.C(1NT19CS116)** bonafide students of **Nitte Meenakshi Institute of Technology**, Bangalore in partial fulfilment for the award of the degree of ***Bachelor of Engineering*** in COMPUTER SCIENCE AND ENGINEERING of Visvesvaraya Technological University, Belgavi during the academic year ***2021-2022.*** It is certified that all corrections and suggestions indicated during the internal assessment have been incorporated in the report. This project has been approved as it satisfies the academic requirement in respect of project work presented for the said degree.

Internal Guide	**Signature of the HOD**	**Signature of Principal**
Vani V Professor, Dept. CSE, NMIT Bangalore	Dr. Sarojadevi H. Professor, Head, Dept. CSE, NMIT Bangalore	Dr. H. C. Nagaraj Principal, NMIT, Bangalore

DECLARATION

We hereby declare that

(i) The project work is our original work
(ii) This Project work has not been submitted for the award of any degree or examination at any other university/college/Institute.
(iii) This Project Work does not contain other persons' data, pictures, graphs, or other information unless specifically acknowledged as being sourced from other persons.
(iv) This Project Work does not contain other persons' writing unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
a) their words have been re-written but the general information attributed to them has been referenced;
b) where their exact words have been used, their writing has been placed inside quotation marks and referenced.
(v) This Project Work does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source is detailed in the thesis and the References sections.

NAME	**USN**	**Signature**
ANURAG NEPAL	1NT19CS036	
AVISHEK RIJAL	1NT19CS045	
BAIBHAV DHAKAL	1NT19CS048	
NABIN KUMAR K.C	1NT1NCS116	

Date:

ACKNOWLEDGEMENT

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crowned our effort with success. I express my sincere gratitude to our Principal **Dr. H. C. Nagaraj**, Nitte Meenakshi Institute of Technology for providing facilities.

We wish to thank our HoD**, Dr. Sarojadevi H.** for the excellent environment created to further educational growth in our college. We also thank him for the invaluable guidance provided which has helped in the creation of a better project.

I hereby like to thank our guide **Vani V, Professor**, Department of Computer Science & Engineering for her periodic inspection, time to time evaluation of the project, and help to bring the project to the present form.

Thanks to our Departmental Project coordinators. We also thank all our friends, teaching and non-teaching staff at NMIT, Bangalore, for all the direct and indirect help provided in the completion of the project.

NAME	**USN**	**Signature**
ANURAG NEPAL	1NT19CS036	
AVISHEK RIJAL	1NT19CS045	
BAIBHAV DHAKAL	1NT19CS048	
NABIN KUMAR K.C	1NT1NCS116	

Date:

ABSTRACT

There are a variety of automated prediction techniques and models using numerous supervised learning models but most of these models don’t take independent attributes like BMI and if the person smokes or not as a factor. Taking that into account, we have chosen to develop a supervised machine learning project and predict the medical cost of patients based on numerous factors. In this project, we have used various data mining tasks to build a model that predicts the insurance amount based on the factors like age, gender, region, BMI, etc. We chose a standard Dataset from Kaggle and used the same in order to train the models and that training helps to come up with efficient and accurate predictions. This project does not give the exact amount required for any health insurance company but our project aims to give sufficient information about the amount associated with it for the health insurance of an individual.

TABLE OF CONTENTS

INTRODUCTION

The goal of this project is to allow a person to get an idea about the necessary amount required according to their health status. Later they can comply with any health insurance company and their schemes & benefits keeping in mind the predicted amount from our project. This can help a person in focusing more on the health aspect of insurance rather than the futile part.

Our project does not give the exact amount required for any health insurance company but gives enough idea about the amount associated with an individual for his/her health insurance.

We are thinking of applying the data mining tools that we have been learning this semester to try and tackle this problem to predict the insurance cost of a person based on age, sex, BMI, number of children, whether he/she is a smoker, and the region they reside in.

1.1 MOTIVATION

As we progress further to the digital age, access to mobile phones and inexpensive internet has taken the world by storm mostly for the better. One of the advantages we can notice is the rise of online services like insurance, consultations etc. Furthermore, We've seen a huge rise of online insurance companies that claim to provide cheap insurance to people but most of them lack to explain and/or lack to take factors like BMI, region and fail to take the fact if the user is a smoker or a non-smoker which increases the cost of insurance for the rest of us. So, in order to tackle this problem and detect the price of insurance more effectively and efficiently, for this project we have taken factors like age, gender, region, BMI, etc.

1.2 PROBLEM DOMAIN

The domain we use for this project are Data Prediction, Predictive Modelling Technique, and Machine Learning. Predictive data mining is data mining that is done for the purpose of using business intelligence or other data to forecast or predict trends. This type of data mining can help business leaders make better decisions and can add value to the efforts of the analytics team. Hence, because of its vast use case scenario and its ability to predict trends we used Data Prediction. Also, predictive modelling is a statistical technique using machine learning and data mining to predict and forecast likely future outcomes with the aid of historical and existing data. It works by analysing current and historical data and projecting what it learns on a model generated to forecast likely outcomes. Predictive modelling can be used to predict just about anything, from TV ratings and a customer's next purchase to credit risks and corporate earnings. Furthermore, Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

1.3 AIM AND OBJECTIVES

- Taking various factors into account that is often left behind by the insurance companies.
- Implement domains like Data Prediction, Predictive Modelling Technique, and Machine Learning Algorithms.
- Use various data mining tasks to build an efficient model that predicts the insurance amount based on various factors
- Help the average insurance payee pay less by providing an accurate amount by taking count of the various factors

The main aim of this project was to take various factors into account that most of the insurance companies fail to consider and hence ultimately decrease the price of insurance for the average insurance payee.

DATA SOURCE AND DATA QUALITY

2.1 DATASET USED

The data set we chose from Kaggle has the columns as

- age: age of the primary beneficiary

- sex: insurance contractor gender, female, male

- BMI: Body mass index, providing an understanding of the body, weights that are relatively high or low relative to height,

an objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9

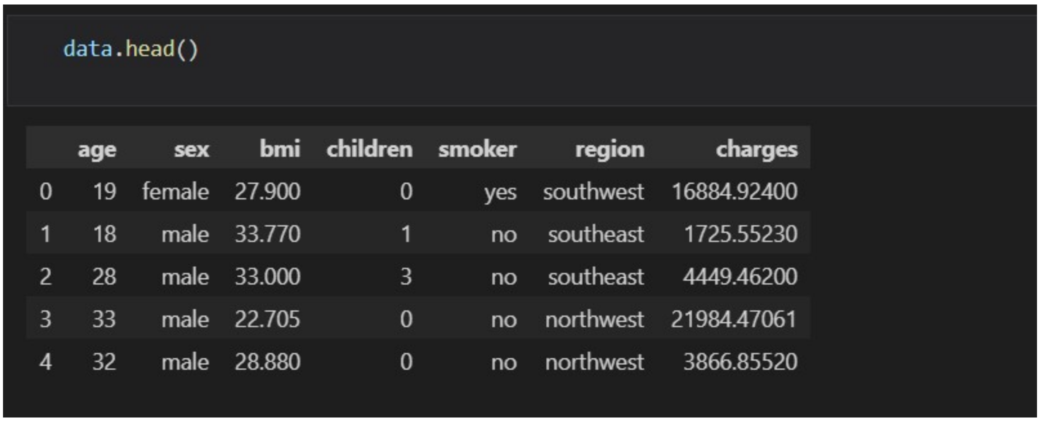
- children: Number of children covered by health insurance / Number of dependents

- smoker: Smoking

- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

- charges: Individual medical costs billed by health insurance

We can see some of the values in the dataset



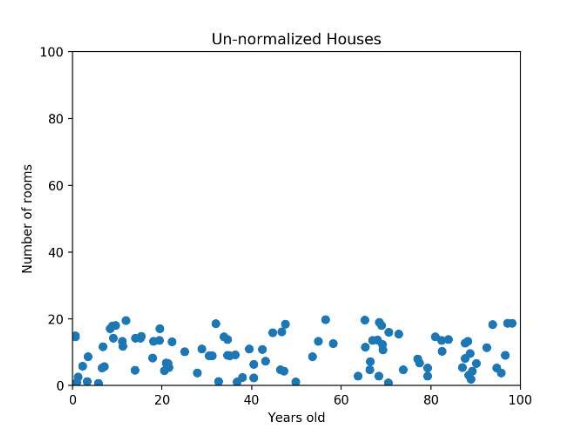
```
data.head()
```

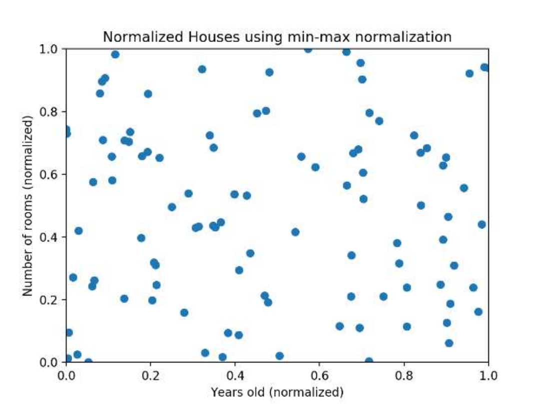
	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

2.2 DATA PREPROCESSING

The dataset is not suited for the regression to take place directly. So cleaning of dataset becomes important for using the data under various regression algorithms. Many machine learning algorithms attempt to find trends in the data by comparing features of data points. However, there is an issue when the features are on drastically different scales.

For example, consider a dataset of houses. Two potential features might be the number of rooms in the house, and the total age of the house in years. A machine-learning algorithm could try to predict which house would be best for you. However, when the algorithm compares data points, the feature with the larger scale will completely dominate the other.





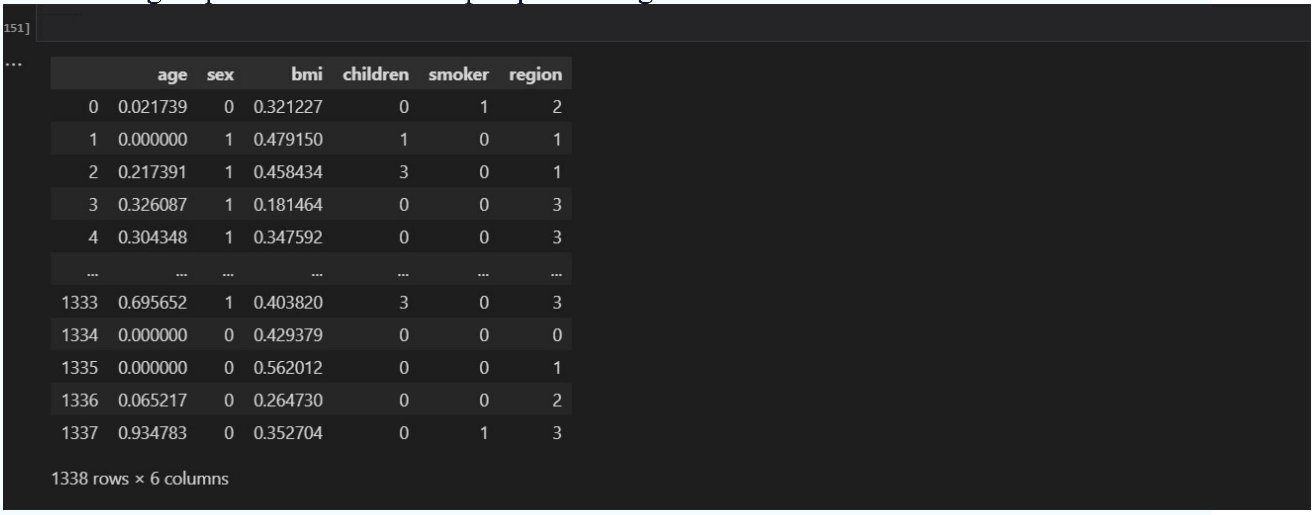
When the data looks squished like that, we know we have a problem. The machine learning algorithm should realize that there is a huge difference between a house with 2 rooms and a house with 20 rooms. But right now, because two houses can be 100 years apart, the difference in the number of rooms contributes less to the overall difference.

The goal of normalization is to make every datapoint have the same scale so each feature is equally important. The image below shows the same house data normalized using min-max normalization.

Min-max normalization is one of the most common ways to normalize data. For every feature, the minimum value of that feature gets transformed into a 0, the maximum value gets transformed into a 1, and every other value gets transformed into a decimal between 0 and 1. For example, if the minimum value of a feature was 20, and the maximum value was 40, then 30 would be transformed to about 0.5 since it is halfway between 20 and 40.

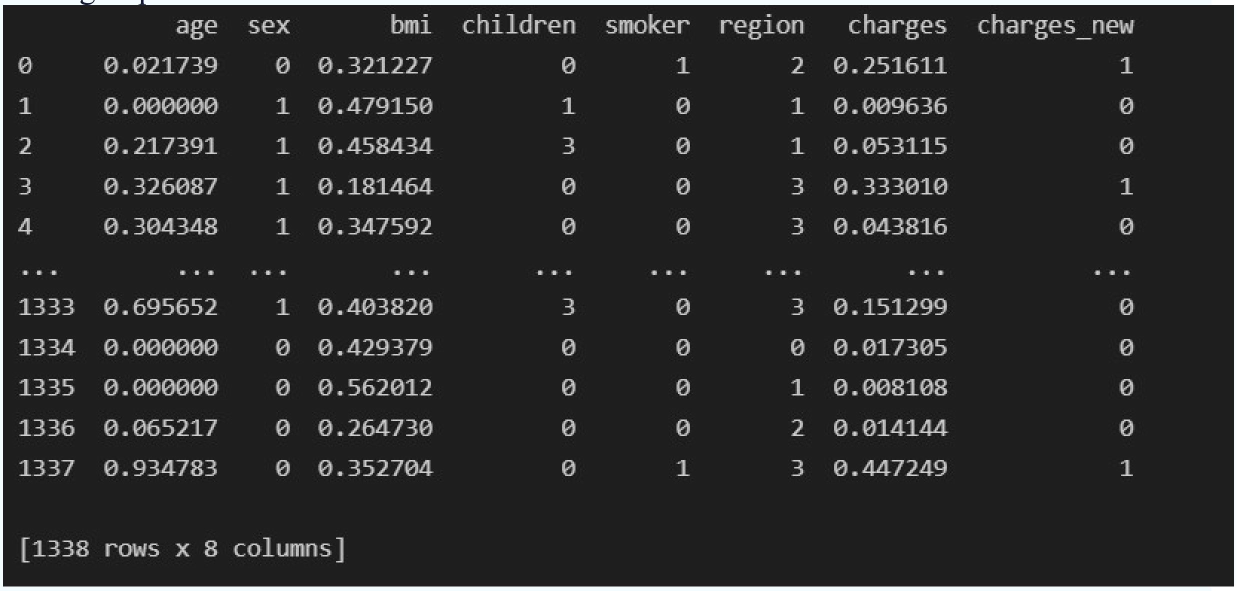
So we used a min-max scaler to normalize the data of age, BMI, and charges into values between 0 and 1. The values of smoker and region being discrete are changed into values 0 or 1.

Below is a glimpse of the data after pre-processing is done:



```
           age  sex       bmi  children  smoker  region
0     0.021739    0  0.321227         0       1       2
1     0.000000    1  0.479150         1       0       1
2     0.217391    1  0.458434         3       0       1
3     0.326087    1  0.181464         0       0       3
4     0.304348    1  0.347592         0       0       3
...        ...  ...       ...       ...     ...     ...
1333  0.695652    1  0.403820         3       0       3
1334  0.000000    0  0.429379         0       0       0
1335  0.000000    0  0.562012         0       0       1
1336  0.065217    0  0.264730         0       0       2
1337  0.934783    0  0.352704         0       1       3

1338 rows × 6 columns
```

In order to perform classification tasks we needed to change the continuous output attribute into categorical one. So taking mean, we classified the data as high or low. If the charges are greater than the mean we group it under high (1) and if the charges are lower or equal to the mean we group it under low (0). We saved it as a new attribute under charges new and below is the glimpse of the data.

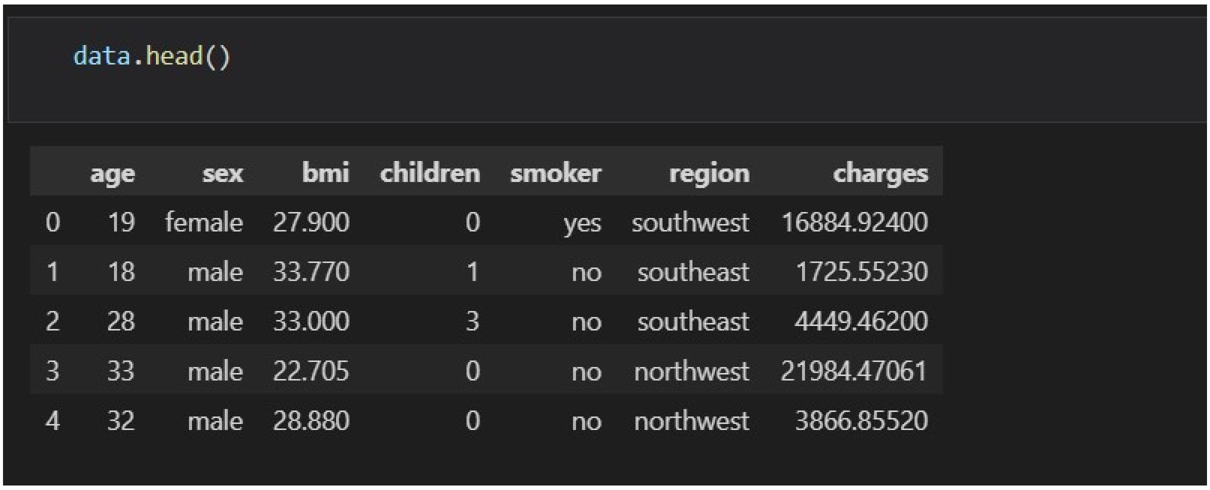


```
           age  sex       bmi  children  smoker  region   charges  charges_new
0     0.021739    0  0.321227         0       1       2  0.251611            1
1     0.000000    1  0.479150         1       0       1  0.009636            0
2     0.217391    1  0.458434         3       0       1  0.053115            0
3     0.326087    1  0.181464         0       0       3  0.333010            1
4     0.304348    1  0.347592         0       0       3  0.043816            0
...        ...  ...       ...       ...     ...     ...       ...          ...
1333  0.695652    1  0.403820         3       0       3  0.151299            0
1334  0.000000    0  0.429379         0       0       0  0.017305            0
1335  0.000000    0  0.562012         0       0       1  0.008108            0
1336  0.065217    0  0.264730         0       0       2  0.014144            0
1337  0.934783    0  0.352704         0       1       3  0.447249            1

[1338 rows x 8 columns]
```

METHODS AND MODELS

3.1 DATA MINING QUESTIONS

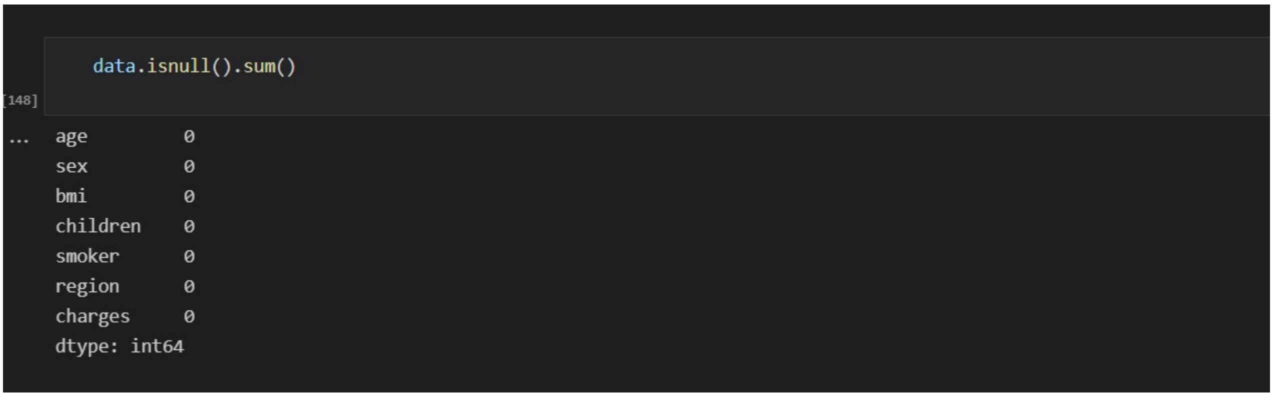
- How does the data look like?



```
data.head()
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520

- Does the data have missing values?



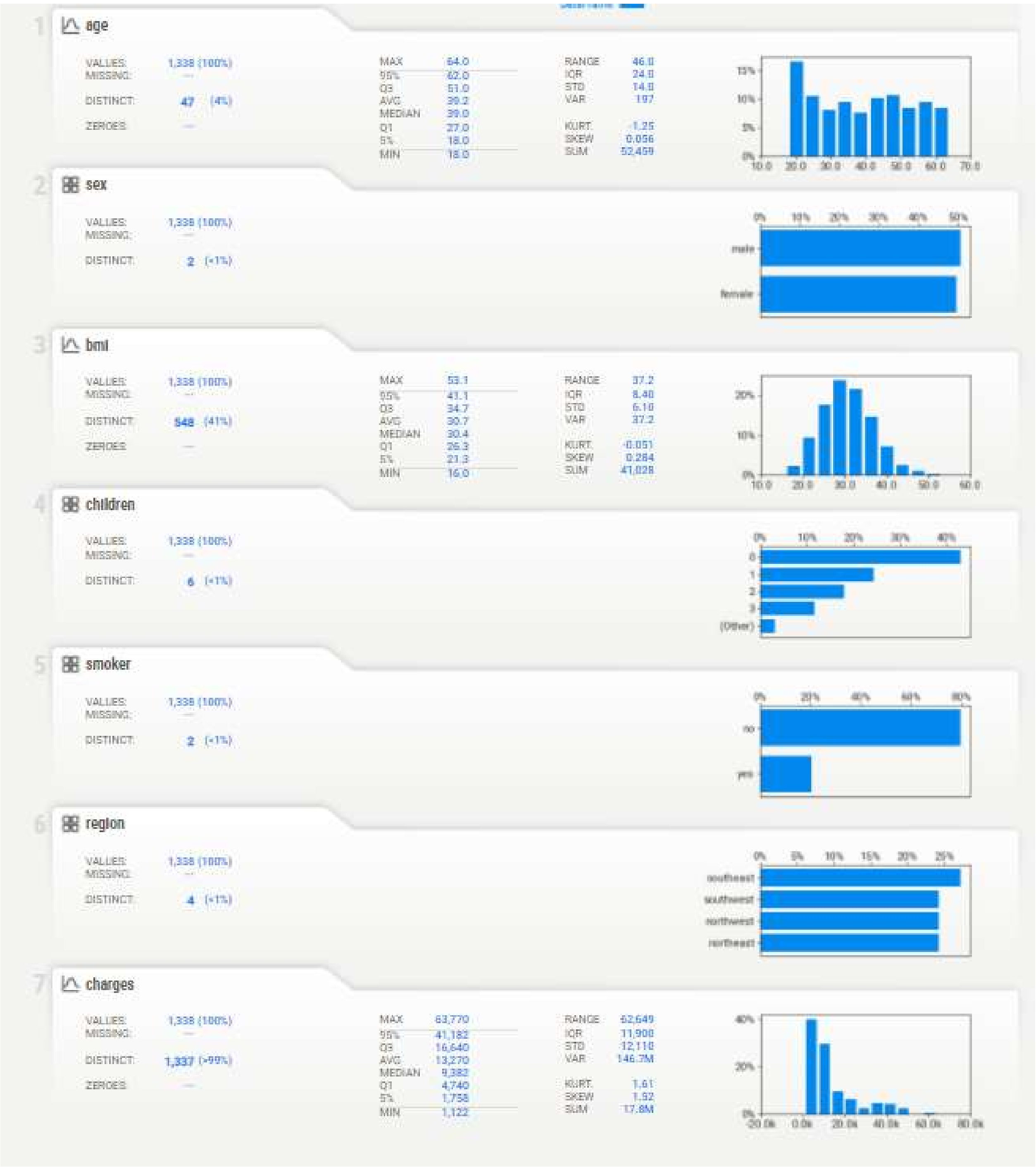
```
data.isnull().sum()
```

```
[148]
...  age         0
     sex         0
     bmi         0
     children    0
     smoker      0
     region      0
     charges     0
     dtype: int64
```

As we can see, in this case, since the problem is for beginners, we do not have missing values, however, this is not always the case. In the following articles, we will have data with missing values and have the possibility to understand how to cope with that.

- What is the distribution of the numerical features?

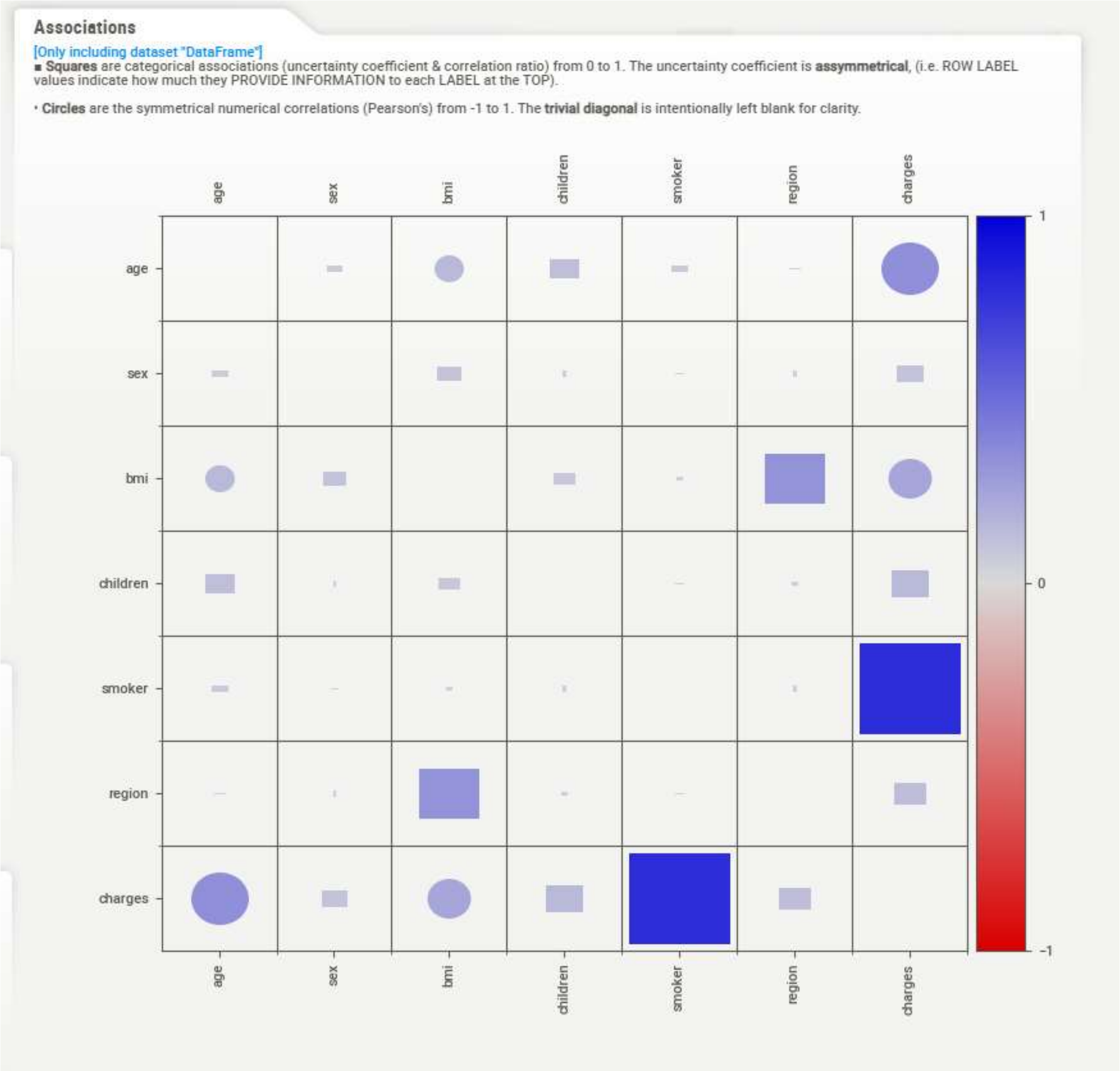
  Below we have shown the distribution of the data from the dataset as bar graphs. There we have the information such as the minimum, maximum values along with the statistical values like mean,median ,interquartile vales, variance, standard deviation and skewness.



```
1  age
   VALUES:    1,338 (100%)     MAX     64.0      RANGE  46.0
   MISSING:   —                95%     62.0      IQR    24.0
                               Q3      51.0      STD    14.0
   DISTINCT:  47  (4%)         AVG     39.2      VAR    197
                               MEDIAN  39.0
   ZEROES:    —                Q1      27.0      KURT.  -1.25
                               5%      18.0      SKEW   0.056
                               MIN     18.0      SUM    52,459

2  sex
   VALUES:    1,338 (100%)
   MISSING:   —
   DISTINCT:  2  (<1%)

3  bmi
   VALUES:    1,338 (100%)     MAX     53.1      RANGE  37.2
   MISSING:   —                95%     41.1      IQR    8.40
                               Q3      34.7      STD    6.10
   DISTINCT:  548  (41%)       AVG     30.7      VAR    37.2
                               MEDIAN  30.4
   ZEROES:    —                Q1      26.3      KURT.  -0.051
                               5%      21.3      SKEW   0.284
                               MIN     16.0      SUM    41,028

4  children
   VALUES:    1,338 (100%)
   MISSING:   —
   DISTINCT:  6  (<1%)

5  smoker
   VALUES:    1,338 (100%)
   MISSING:   —
   DISTINCT:  2  (<1%)

6  region
   VALUES:    1,338 (100%)
   MISSING:   —
   DISTINCT:  4  (<1%)

7  charges
   VALUES:    1,338 (100%)     MAX     63,770    RANGE  62,649
   MISSING:   —                95%     41,182    IQR    11,900
                               Q3      16,640    STD    12,110
   DISTINCT:  1,337 (>99%)     AVG     13,270    VAR    146.7M
                               MEDIAN  9,382
   ZEROES:    —                Q1      4,740     KURT.  1.61
                               5%      1,758     SKEW   1.52
                               MIN     1,122     SUM    17.8M
```

As we can see the distributions of the data look to be unbalanced because our output attribute is continuous and continuous target variables that need to be predicted often have many rare and extreme values. In our case too our output attribute (charges) resembles similar pattern and results in unbalanced dataset.

- What is a correlation between the target variable and features?



Associations

[Only including dataset "DataFrame"]

■ **Squares** are categorical associations (uncertainty coefficient & correlation ratio) from 0 to 1. The uncertainty coefficient is **assymmetrical**, (i.e. ROW LABEL values indicate how much they PROVIDE INFORMATION to each LABEL at the TOP).

• **Circles** are the symmetrical numerical correlations (Pearson's) from -1 to 1. The **trivial diagonal** is intentionally left blank for clarity.

Squares are categorical associations (uncertainty coefficient & correlation ratio) from 0 to 1. The uncertainty coefficient is asymmetrical, (i.e. ROW LABEL values indicate how much they PROVIDE INFORMATION to each LABEL at the TOP). Circles are the symmetrical numerical correlations (Pearson's) from -1 to 1. The trivial diagonal is intentionally left blank for clarity.

3.2 DATA MINING ALGORITHMS

Here we have compared six different Data Mining Algorithms to see how they fare in the prediction of the insurance cost. they are briefly discussed below:

- **Linear Regression**: Regression refers to a type of supervised machine learning technique that is used to predict any continuous-valued attribute. In regression, the nature of the predicted data is ordered. The regression can be further divided into linear regression and non-linear regression.

Linear regression is the type of regression that forms a relationship between the target variable and one or more independent variables utilizing a straight line. The given equation represents the equation of linear regression

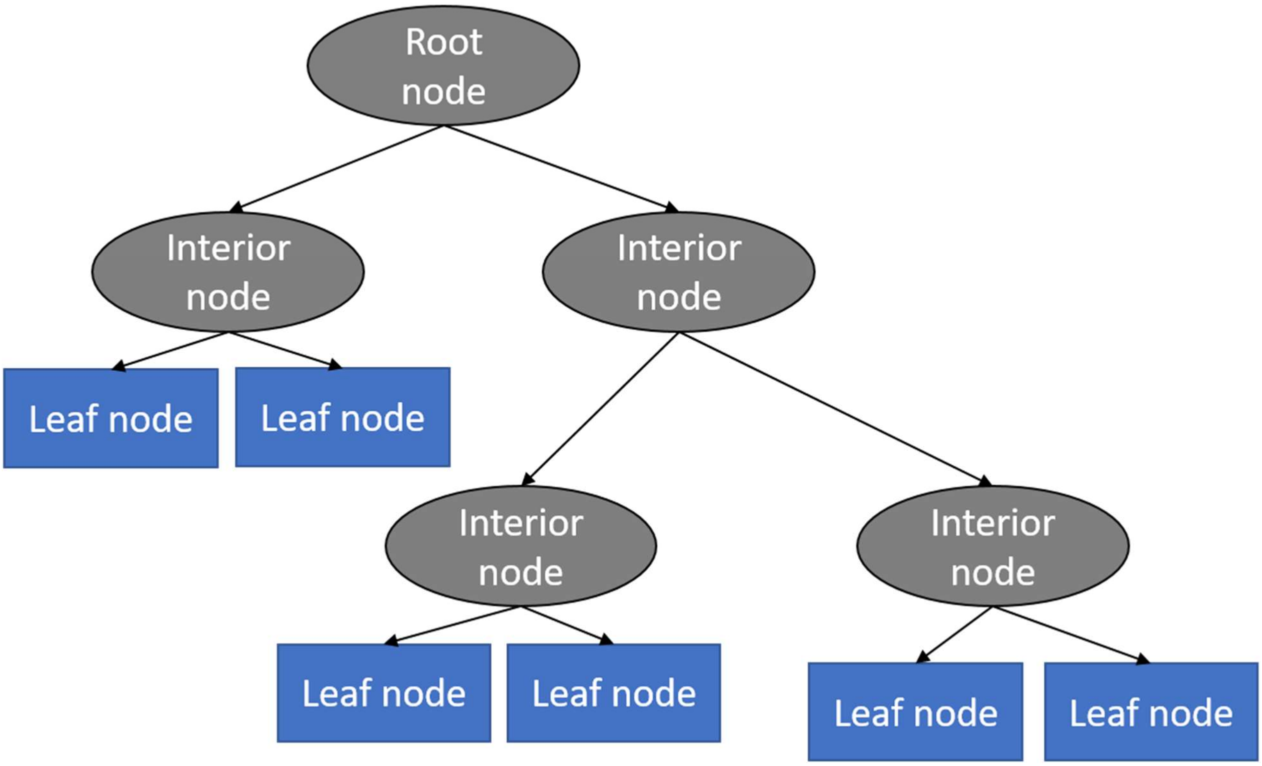
Y = a + b*X + e, where a represents the intercept represents the slope of the regression line, e represents the error and Y represent the predictor and target variables, respectively.

If X is made up of more than one variable, termed as multiple linear equations.

In linear regression, the best fit line is achieved utilizing the least squared method, and it minimizes the total sum of the squares of the deviations from each data point to the line of regression. Here, the positive and negative deviations do not get cancelled as all the deviations are squared.

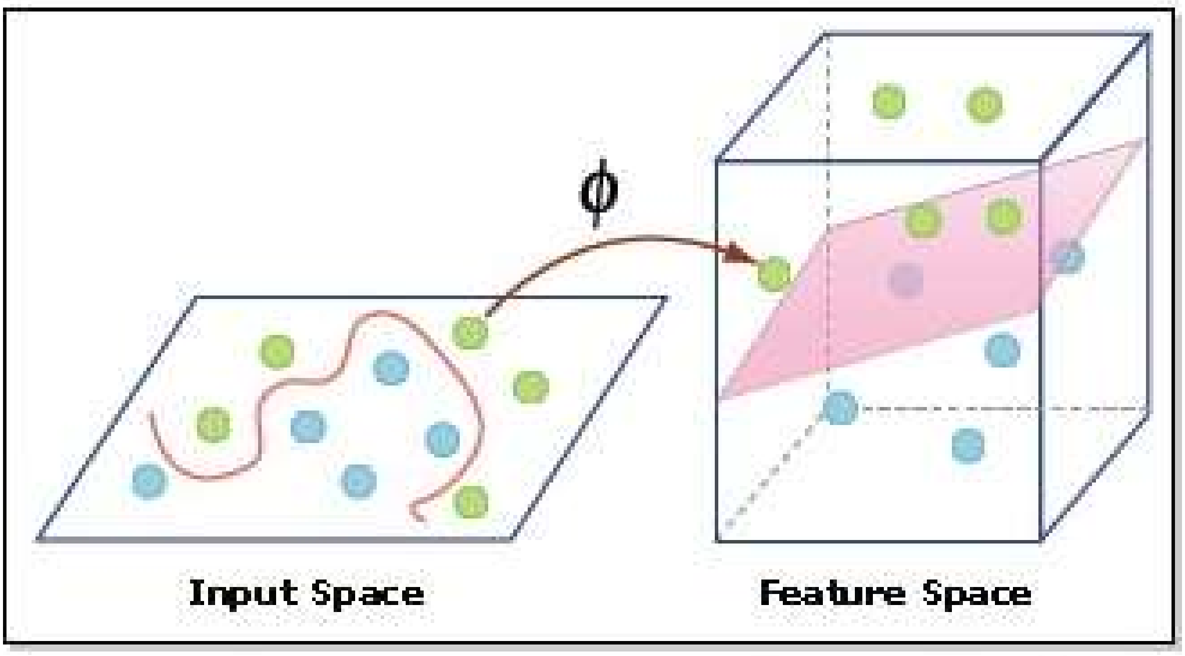
- **Decision Tree Regressor**: Decision Tree is one of the most commonly used, practical approaches for supervised learning. It can be used to solve both Regression and Classification tasks with the latter being put more into practical application.

It is a tree-structured classifier with three types of nodes. The Root Node is the initial node which represents the entire sample and may get split further into further nodes. The Interior Nodes represent the features of a data set and the branches represent the decision rules. Finally, the Leaf Nodes represent the outcome. This algorithm is very useful for solving decision-related problems.



With a particular data point, it is run completely through the entirely tree by answering True/False questions till it reaches the leaf node. The final prediction is the average of the value of the dependent variable in that particular leaf node. Through multiple iterations, the Tree is able to predict a proper value for the data point.

- **SVR Linear** : Support Vector Machines (SVMs) are well known in classification problems. The use of SVMs in regression is not as well documented, however. These types of models are known as Support Vector Regression (SVR).



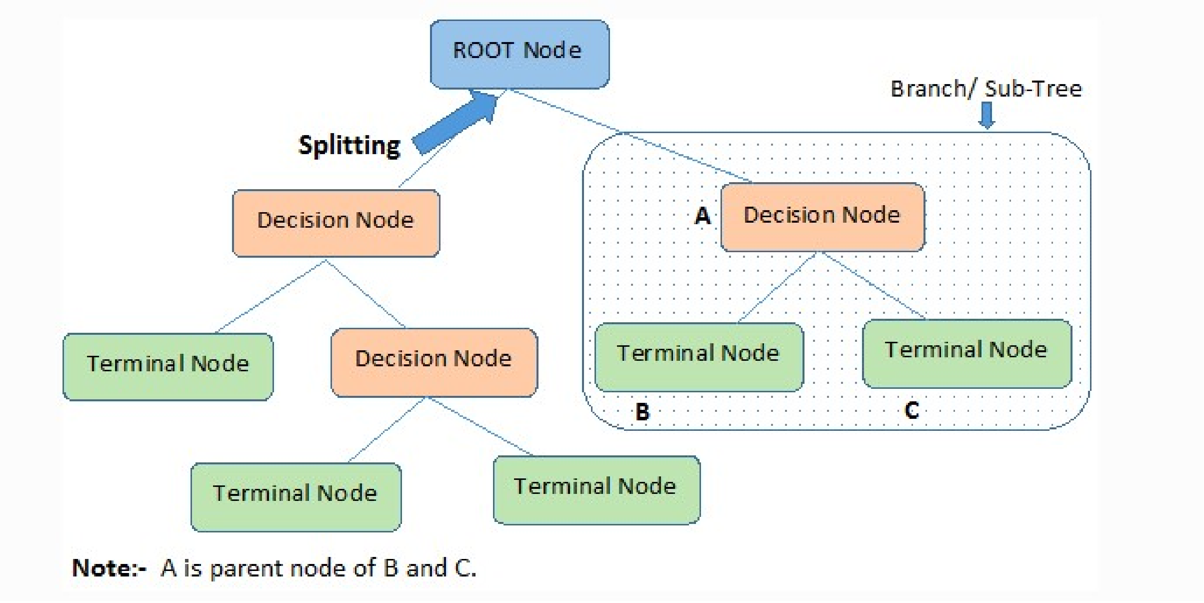
There are a few important parameters of SVM that we should be aware of before proceeding further:

○ Kernel: A kernel helps us find a hyperplane in the higher dimensional space without increasing the computational cost. Usually, the computational cost will increase if the dimension of the data increases. This increase in dimension is required when we are unable to find a separating hyperplane in a given dimension and are required to move in a higher dimension:

○ Hyperplane: This is basically a separating line between two data classes in SVM. But in Support Vector Regression, this is the line that will be used to predict the continuous output

○ Decision Boundary: A decision boundary can be thought of as a demarcation line (for simplification) on one side of which lie positive examples and on the other side lie the negative examples. On this very line, the examples may be classified as either positive or negative. This same concept of SVM will be applied in Support Vector Regression as well

- **Decision Tree Classifier**: Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving regression and classification problems too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by learning simple decision rules inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the root of the tree. We compare the values of the root attribute with the record's attribute. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

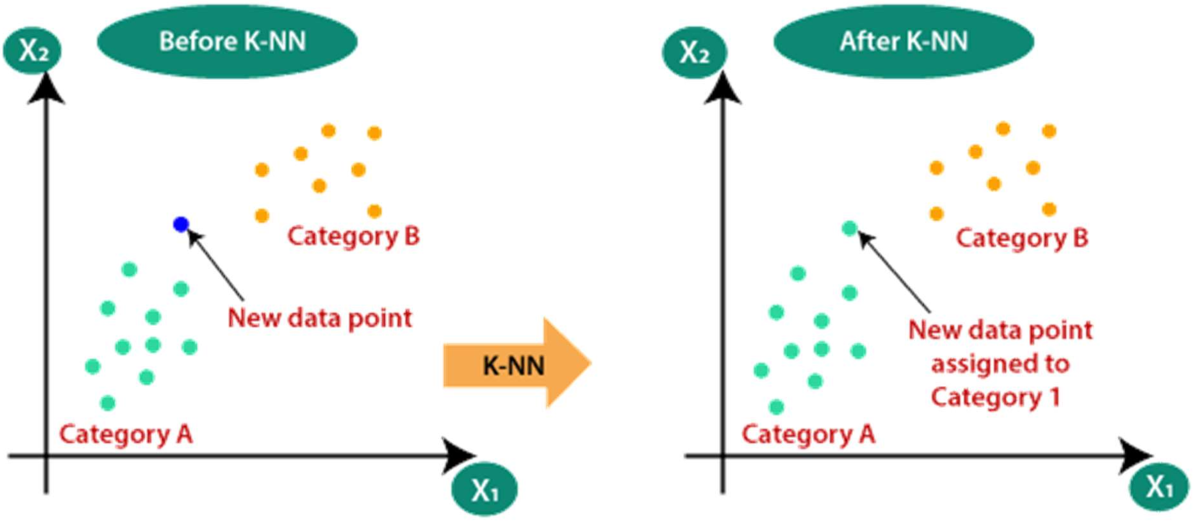


- **KNN Classifier**:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

The K-NN working can be explained on the basis of the below algorithm:

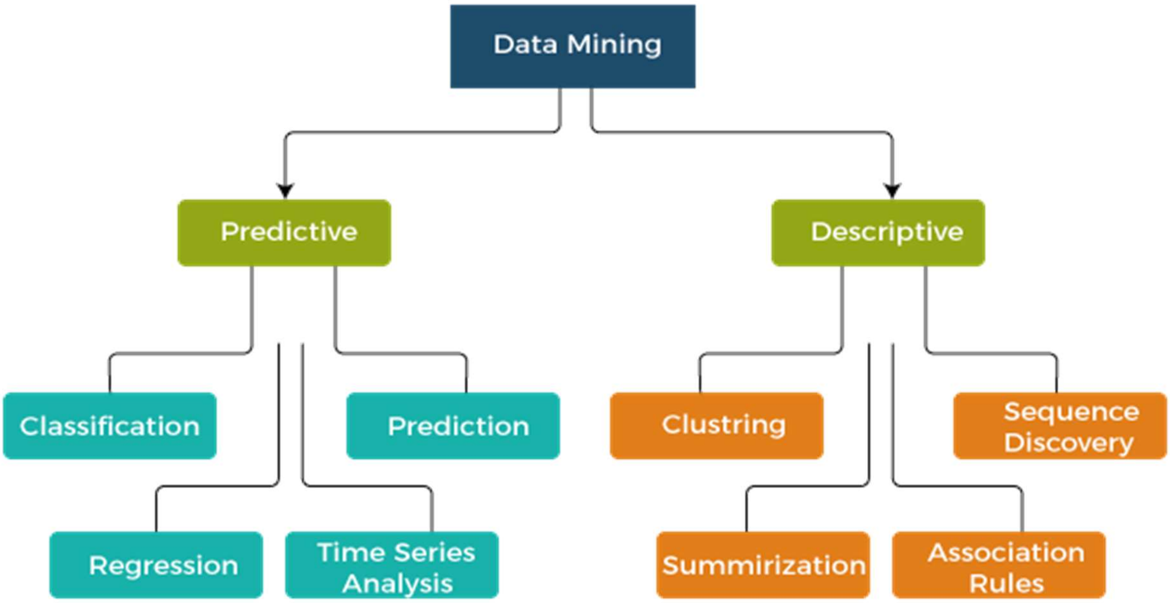
- **Step-1:** Select the number K of the neighbours
- **Step-2:** Calculate the Euclidean distance of **K number of neighbours**
- **Step-3:** Take the K nearest neighbours as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbours, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbour is maximum.
- **Step-6:** Our model is ready.



All of these algorithms that we have discussed in this section have their own advantages and disadvantages. So we are comparing and contrasting between these to see which one would do a better job at predication task at hand.

3.3 DATA MINING MODELS

A Data mining model refers to a method that is usually used to present the information and various ways in which they can apply information to specific questions and problems. Below are the models that we have in Data Mining.



As our project is about predicting the cost of insurance, we have used the predictive Data mining model.

A predictive data mining model predicts the values of data using known results gathered from the different data sets. Predictive modelling cannot be classified as a separate discipline; it occurs in all organizations or industries across all disciplines. The main objective of predictive data mining models is to predict the future based on the past data, generally but not always on the statistical modelling. Predictive modelling is used in healthcare industries to identify high-risk patients with congestive heart failures, high blood pressure, diabetes, infection, cancer, etc. It is also used in the vehicle insurance company to assign the risk of accidents to the policyholder. A predictive model of a data mining task comprises classification, regression, prediction, and time series analysis. The predictive model of data mining is also called statistical regression. It refers to a monitoring learning technique that includes an explication of the dependency of a few attribute's values upon the other attribute's value in the same product and the growth of a model that can predict these attribute's values in previous cases.

- **Classification:**

In data mining, classification refers to a form of data analysis where a machine learning model assigns a specific category to a new observation. It is based on what the model has learned from the data sets. In other words, classification is the act of assigning objects to many predefined categories. One example of classification in the banking and financial services industry is identifying whether transactions are fraudulent or not. In the same way, machine learning can also be used to predict whether a loan application would be approved or not.

- **Regression:**

Regression refers to a method that verifies the value of data for a function. Generally, it is used for appropriate data. A linear regression model in the context of machine learning or statistics is basically a linear approach for modelling the relationships between the dependent variable known as the result and your independent variable is known as features. If your model has only one independent variable, it is called simple linear regression, and else it is called multiple linear regression.

MODEL EVALUATION & DISCUSSION

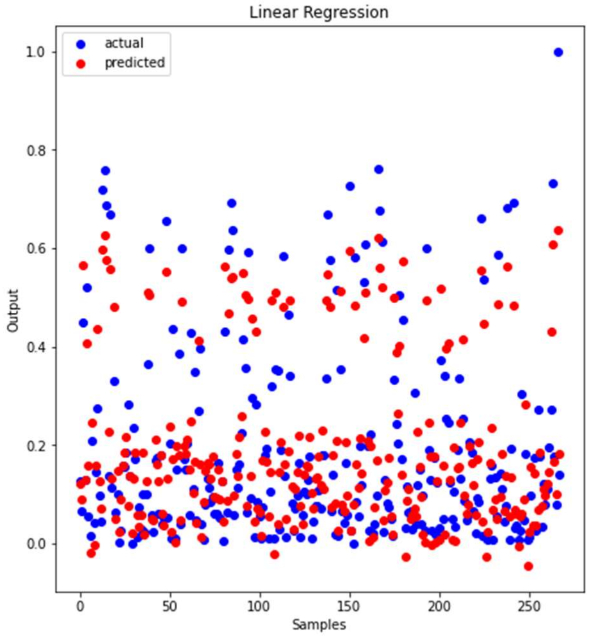
We used the test data instance to analyse and measure the accuracy of different algorithms. There are 1338 instances of data and 80% of the data was used for training the model whereas the rest 20% for testing the model.

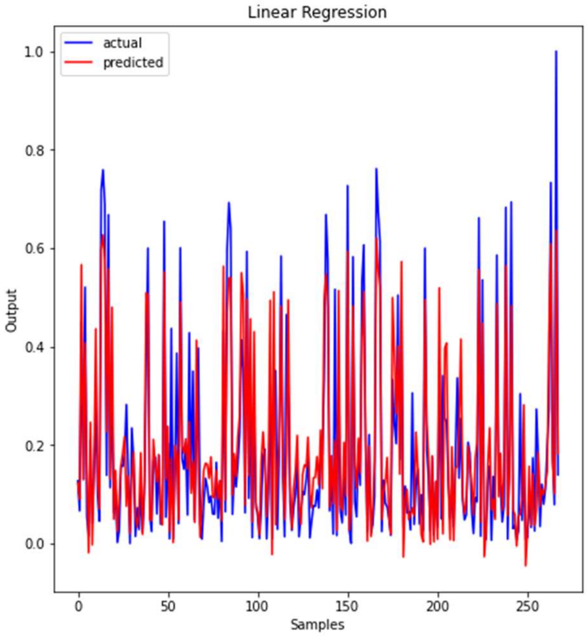
We used three different types of regression algorithms and they are Linear Regression, Decision Tree Regressor and Support Vector Regression with linear kernel.

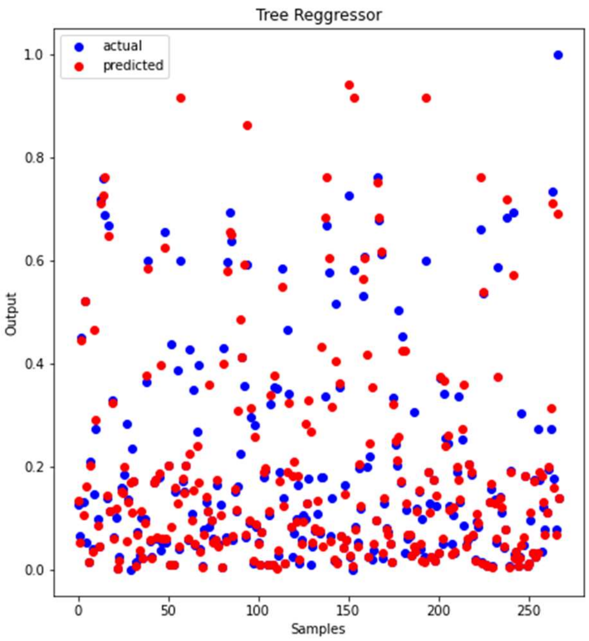
The metrics used to analyse the performance of the regression algorithm is R2 score.

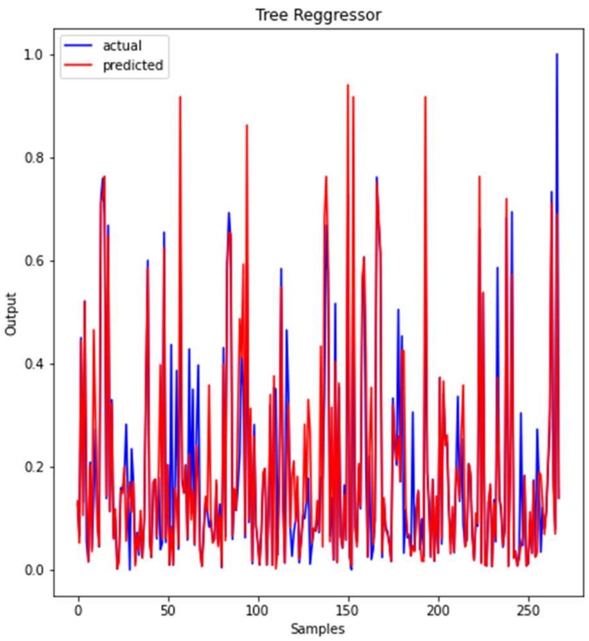
Coefficient of determination also called as R2 score is used to evaluate the performance of a linear regression model. It is the amount of the variation in the output dependent attribute which is predictable from the input independent variable(s).

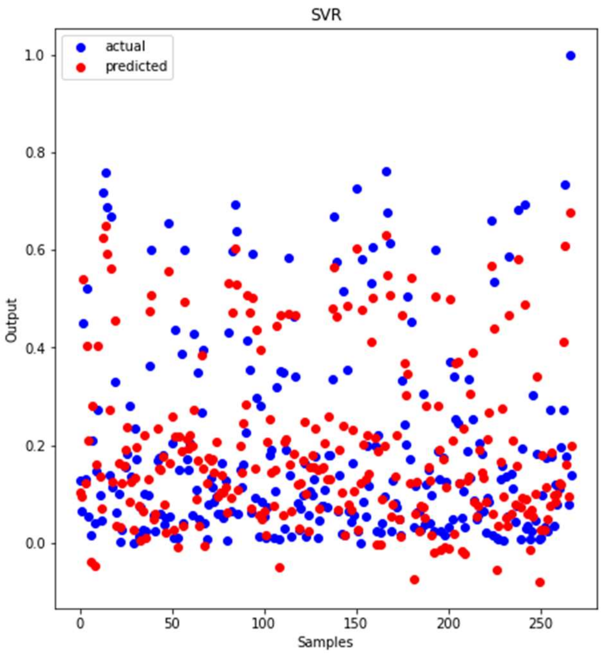
Linear Regression model performed the best with a R2 score of 0.780 whereas Decision Tree Regressor had the score of 0.733 and SVR had the score of 0.762.

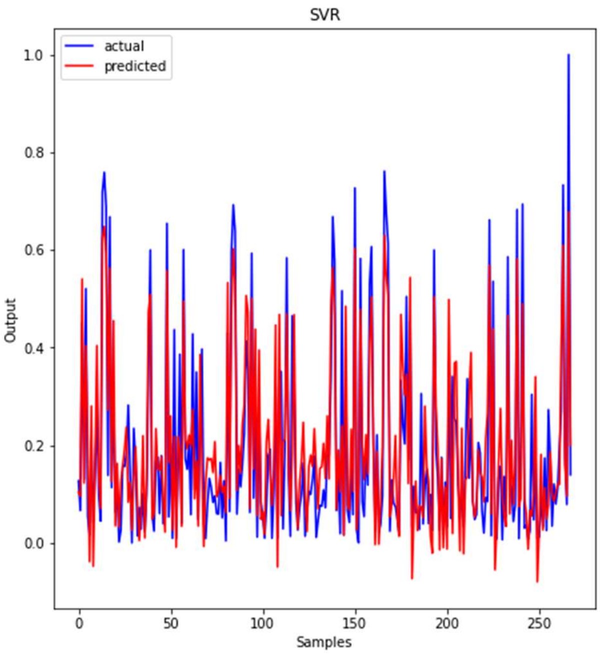






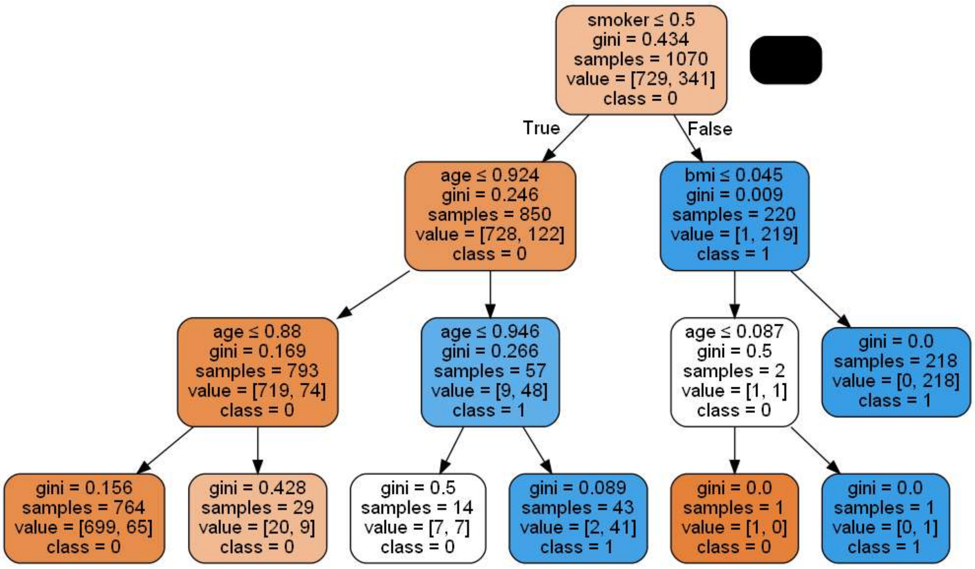






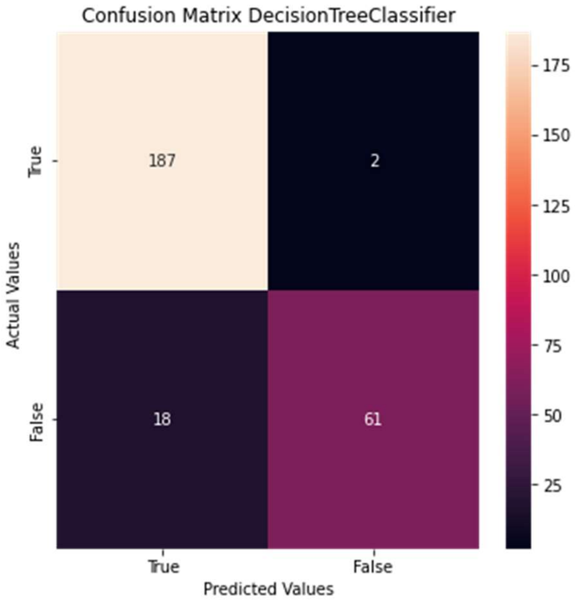
We also used 2 different classification algorithms as we normalized continuous attributes to categorical(binary). We used Decision Tree Classifier and KNN classifier algorithms. We used the sklearn.metrics.accuracy_score function to compute the accuracy and also created a confusion matrix to see how it performed with the test instance.

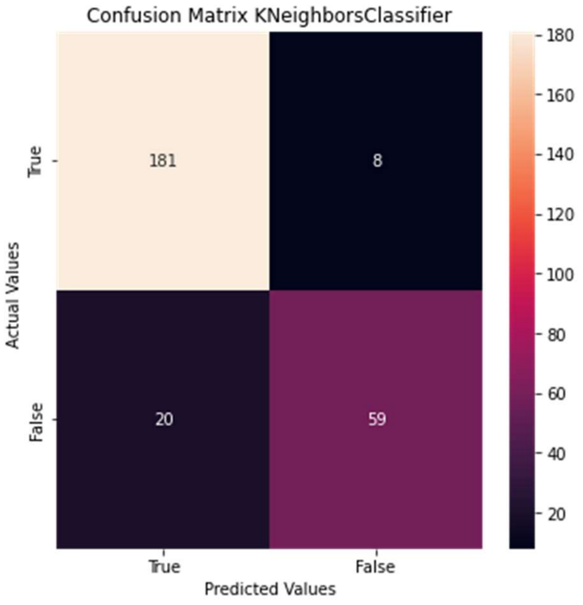
For the Decision Tree Classifier we used Gini index measure with max depth of tree as 3. We achieved 0.925 accuracy.



For the KNN classifier we used k or N neighbours as 3 and achieved an accuracy of 0.8955. Furthermore classification algorithms can be analysed by the confusion matrix.

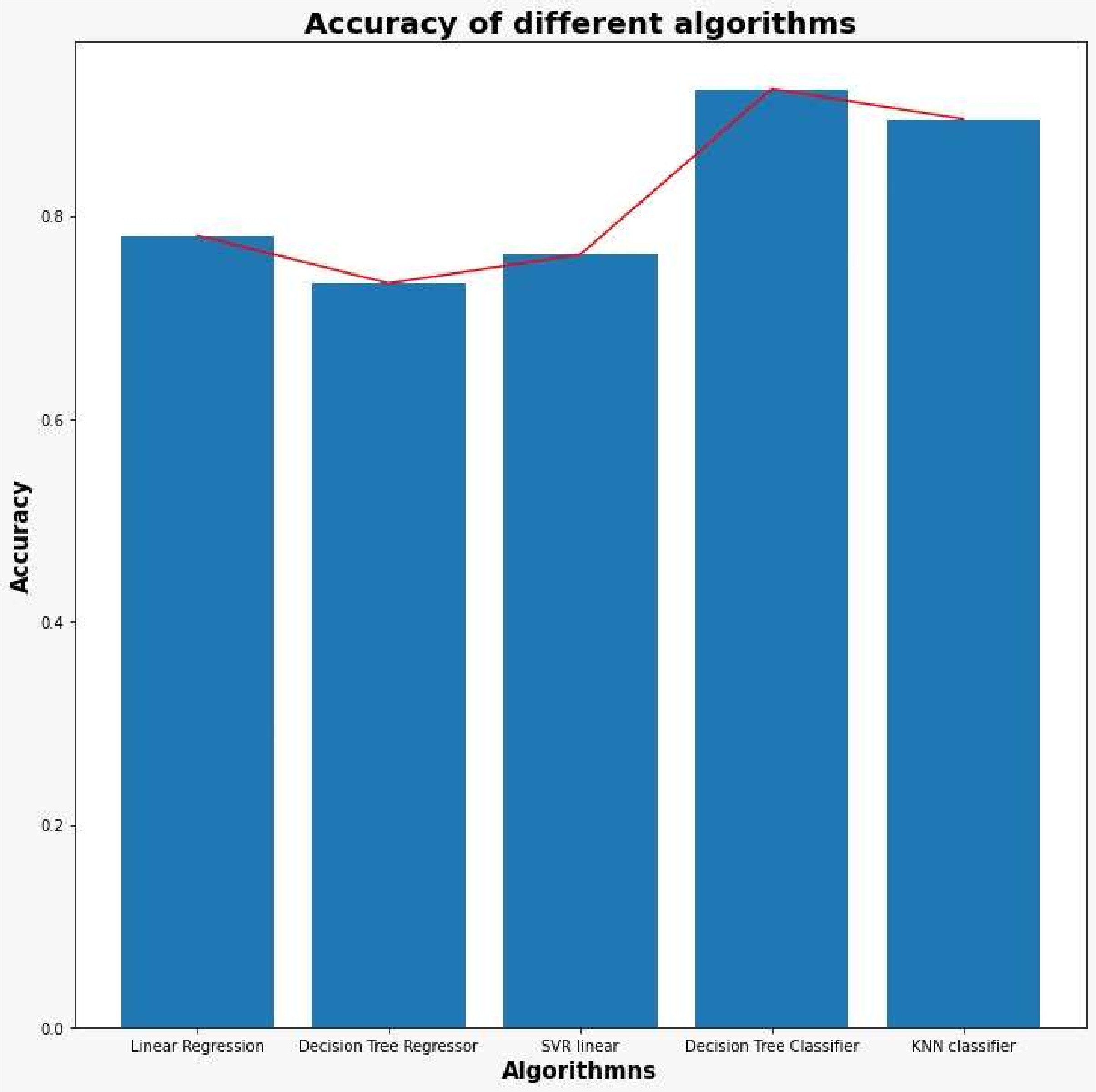
A confusion matrix is a summary of prediction results on a classification problem. A confusion matrix prints the correct and also incorrect values in number count . It gives us insight not only into the errors being made by a classifier but more importantly the types of errors that are being made. There are 4 different classes in the matrix and they are true positive, true negative, false positive and false negative. A true positive is an outcome where the model correctly predicts the positive class. Similarly, a true negative is an outcome where the model correctly predicts the negative class. A false positive is an outcome where the model incorrectly predicts the positive class. And a false negative is an outcome where the model incorrectly predicts the negative class.





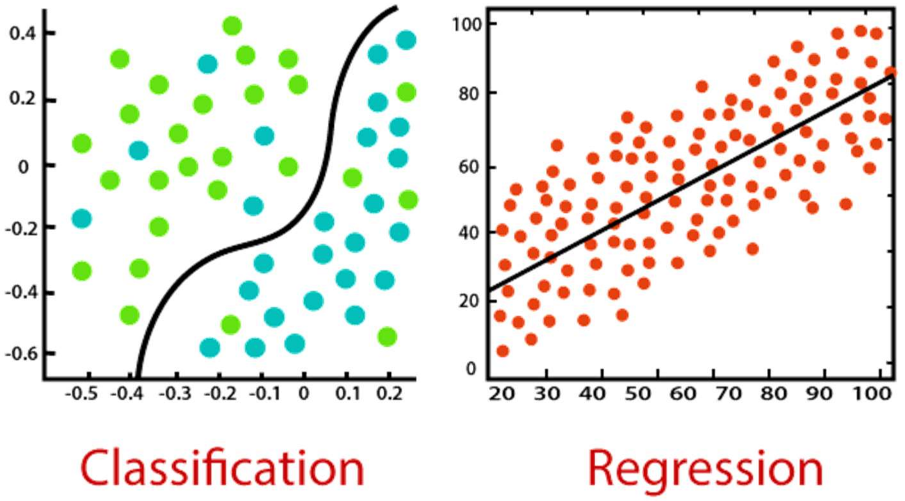
CONCLUSION & FUTURE DIRECTION

While we tested different algorithms and different techniques, we observed that a single problem statement can be solved using different data mining approaches.



In terms of accuracy the decision tree classifier performed the best with the highest accuracy whereas all the regression algorithms were slightly poorer.

In Regression, we try to find the best fit line, which can predict the output more accurately. In Classification, we try to find the decision boundary, which can divide the dataset into different classes.



Classification

Regression

For the given problem statement a regression model could predict the medical charges more efficiently as the output attribute is continuous however the lack in accuracy made us try a different approach. We normalized the output attribute from continuous to categorical (binary) i.e. 0 and 1 and applied classification techniques. We observed that the classification techniques observed higher accuracy however they can only predict the class of the output whereas the regression model could predict the value itself.

To solve this, in the future we can create and build more complex regression models to try to capture the remaining variance. We could add polynomial terms to model the nonlinear relationship between an independent variable and the target variable. We can also do different kernel and hyperparameter testing to find and select an optimal regression model.

REFLECTION PORTFOLIO

Looking back at the work we had to do to get this project done we can definitely say that we have learnt about various data mining algorithm as well as the models that are applicable for a certain type of data. We went from zero knowledge in implementing these algorithms to somewhat fulfilling the task at hand and we can say that this has greatly increasef our understanding of the Data Mining topics.

Here are the things that we learnt during the preparation of this project as this report. We have learnt that we can do anything that we set out to do with a full heart and stop from doubting ourselves. This report shows that all we’ve researched and how we implemented the algorithms and compared their accuracies. We also learnt how to effectively deal with the pressure to complete the assignment and overcame it through team work.

REFERENCES

Chauhan, N. S. (2021). *Decision Tree Algorithm, Explained*. KDnuggets. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

Codecademy. (n.d.). *Normalization*. https://www.codecademy.com/article/normalization

Data Mining Models - Javatpoint. (2021). Www.Javatpoint.Com. https://www.javatpoint.com/data-mining-models

K, G. M. (2021, December 15). *Machine Learning Basics: Decision Tree Regression - Towards Data Science*. Medium. https://towardsdatascience.com/machine-learning-basics-decision-tree-regression-1d73ea003fda

Sethi, A. (2020, April 1). *Support Vector Regression In Machine Learning*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2020/03/support-vector-regression-tutorial-for-machine-learning/

APPENDIX

Dataset from Kaggle:

https://www.kaggle.com/mirichoi0218/insurance

Pyhton Code Implemented:

https://github.com/avishek-r/DataMining-LA2

Setup to execute the code:

- Python 3.7.1 and up
- Pandas
- Scikit learn
- Matplotlib
- Seaborn
- Graphviz