**Summer Training Report**

**On**

**Medical Health Expenses Prediction**

**At**

**Academy of Skill Development**

**Submitted in partial fulfillment of the requirement for the award of degree of B.Tech(CSE)**

**Name:- MILIND UDBHAV**

**Roll No.- 1901010044**

**YEAR- 2021-22**



**K.R. Mangalam University**

**School of Engineering & Technology**

**Gurugram**

**CERTIFICATE**

This is to certify that this proposal of minor project entitled **“Medical Health Expense Prediction”** is a record of bonafide work, carried out by **MILIND UDBHAV,** under the guidance at **ACADEMY OF SKILL DEVELOPMENT**. In my opinion, the report in its present form is in partial fulfilment ofthe requirements for the award of the degree of **BACHELOR OF TECHNOLOGY(B.TECH)** and as per regulations of the **ACADEMY OF SKILL DEVELOPMENT*.*** To the best of my knowledge, the results embodied in this report, are original in nature and worthy of incorporation in the present version of the report.

Logo

Description automatically generated

**Guide / Supervisor**

****

------------------------------------------------

**MR. MAHENDRA DATTA**

Project Engineer

ACADEMY OF SKILL DEVELOPMENT

**CANDIDATE’S DECLARATION**

I, **MILIND UDBHAV** bearing roll no. **1901010044**, a student of **B.Tech(CSE)**hereby declare that I own the full responsibility for the information, results etc. provided in this report titled “**Medical Health Expense Prediction**” submitted to **K R Mangalam University**, for the award of **B.Tech(CSE)** degree. I have taken care in all respect to honor the intellectual property right and have acknowledged the contribution of others for using them in academic purpose and further declare that in case of any violation of intellectual property right or copyright I, as a candidate, will be fully responsible for the same. My supervisor should not be held responsible for full or partial violation of copyright or intellectual property right. This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

(Candidate Signature)

**Milind Udbhav**

**1901010044**

Place- Gurugram

Dated:

**ACKNOWLEDGEMENT**

Success of any project depends largely on the encouragement and guidelines of many others. I take this sincere opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project work.

I would like to show our greatest appreciation to **Mr. Mahendra Datta**, Project Engineer at **Academy of skill development, Kolkata.** I always feel motivated and encouraged every time by his valuable advice and constant inspiration; without his encouragement and guidance this project would not have materialized.

Words are inadequate in offering our thanks to the other trainees, project assistants and other members at **Academy of skill development** for their encouragement and cooperation in carrying out this project work. The guidance and support received from all the members and who are contributing to this project, was vital for the success of this project.

**Logo

Description automatically generated**

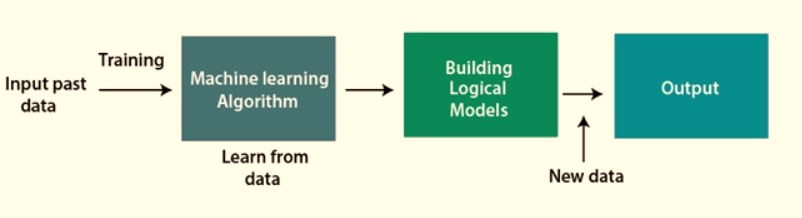
**Name - MILIND UDBHAV**

**Roll No- 1901010044**

# ABSTRACT

**Machine Learning**

* Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. **Machine learning focuses on the development of computer programs** that can access data and use it learn for themselves.
* The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. **The primary aim is to allow the computers learn automatically** without human intervention or assistance and adjust actions accordingly.



**Classification of Machine Learning**

1. **Supervised Learning.**
2. **Unsupervised Learning.**
3. **Reinforcement Learning.**

**1)Supervised Learning**

* Supervised learning, also known as supervised machine learning, is a subcategory of [machine learning](https://www.ibm.com/cloud/learn/machine-learning) and [artificial intelligence](https://www.ibm.com/cloud/learn/what-is-artificial-intelligence).
* It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross-validation process.
* Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox.
* Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.

Supervised learning can be separated into two types of problems when data mining -

classification and regression:

* **Classification** uses an algorithm to accurately assign test data into specific categories.
* It recognizes specific entities within the dataset and attempts to draw some conclusions on how those entities should be labeled or defined.
* Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest, which are described in more detail below.
* **Regression** is used to understand the relationship between dependent and independent variables.
* It is commonly used to make projections, such as for sales revenue for a given business. [Linear regression](https://www.ibm.com/analytics/learn/linear-regression), [logistical regression](https://www.ibm.com/analytics/learn/logistic-regression), and polynomial regression are popular regression algorithms.

**2)Unsupervised learning**

* Unsupervised learning is the training of a machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.
* Here the task of the machine is to group unsorted information according to similarities, patterns, and differences without any prior training of data.

**Unsupervised learning is classified into two categories of algorithms:**

* **Clustering:** A clustering problem is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior.
* **Association:** An association rule learning problem is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y.

**3)Reinforcement learning**

* It is an area of Machine Learning. It is about taking suitable action to maximize reward in a particular situation.
* It is employed by various software and machines to find the best possible behavior or path it should take in a specific situation.
* Reinforcement learning differs from supervised learning in a way that in supervised learning the training data has the answer key with it so the model is trained with the correct answer itself whereas in reinforcement learning, there is no answer but the reinforcement agent decides what to do to perform the given task. In the absence of a training dataset, it is bound to learn from its experience.

**Main points in Reinforcement learning –**

* Input: The input should be an initial state from which the model will start
* Output: There are many possible outputs as there are a variety of solutions to a particular problem
* Training: The training is based upon the input, The model will return a state and the user will decide to reward or punish the model based on its output.
* The model keeps continues to learn.
* The best solution is decided based on the maximum reward.

## **LIST OF COMMON MACHINE LEARNING ALGORITHMS:**

Here is the list of commonly used machine learning algorithms. These algorithms can be applied to almost any data problem:

1. Linear Regression
2. Logistic Regression
3. Decision Tree
4. SVM
5. Naive Bayes
6. KNN (K-Nearest Neighbors)
7. K-Means
8. Random Forest
9. Dimensionality Reduction Algorithms.

There are also Gradient Boosting Algorithms:

1. GBM

2. XGBoost

3. Light GBM

4. CatBoost

## **1. Linear Regression**

It is used to estimate real values (cost of houses, number of calls, total sales etc.) based on continuous variable(s). Here, we establish relationship between independent and dependent variables by fitting a best line. This best fit line is known as regression line and represented by a linear equation Y= a \*X + b.

The best way to understand linear regression is to relive this experience of childhood. Let us say, you ask a child in fifth grade to arrange people in his class by increasing order of weight, without asking them their weights! What do you think the child will do? He / she would likely look (visually analyze) at the height and build of people and arrange them using a combination of these visible parameters. This is linear regression in real life! The child has actually figured out that height and build would be correlated to the weight by a relationship, which looks like the equation above.

In this equation:

* Y – Dependent Variable
* a – Slope
* X – Independent variable
* b – Intercept

These coefficients a and b are derived based on minimizing the sum of squared difference of distance between data points and regression line.

Look at the below example. Here we have identified the best fit line having linear equation **y=0.2811x+13.9**. Now using this equation, we can find the weight, knowing the height of a person.



Linear Regression is mainly of two types: Simple Linear Regression and Multiple Linear Regression.

Simple Linear Regression is characterized by one independent variable and Multiple Linear Regression (as the name suggests) is characterized by multiple (more than 1) independent variables.

While finding the best fit line, you can fit a polynomial or curvilinear regression. And these are known as polynomial or curvilinear regression.

## **2. Logistic Regression**

It is a classification not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variables. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as **logit regression**. Since, it predicts the probability, its output values lie between 0 and 1 (as expected).

odds= p/ (1-p) = probability of event occurrence / probability of not event occurrences

ln(odds) = ln(p/(1-p))

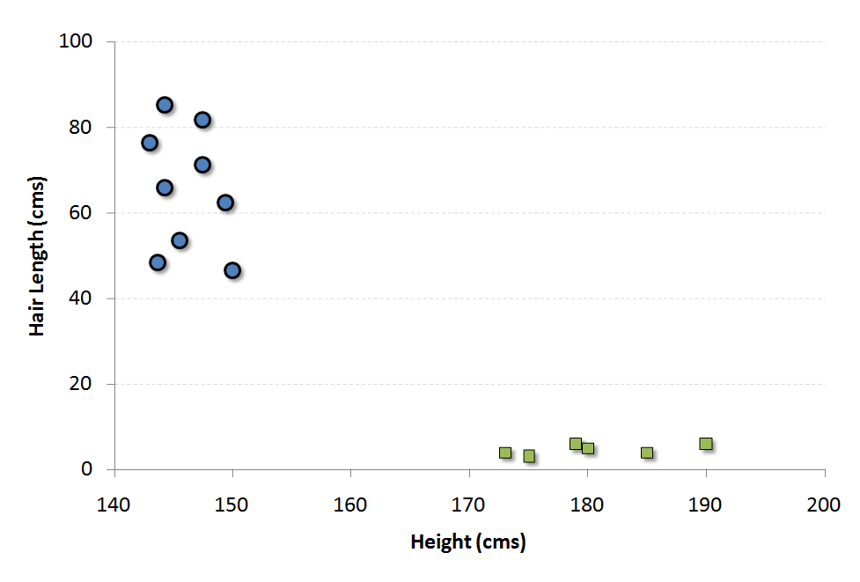
logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3.... +bkXk

Above, p is the probability of presence of the characteristic of interest. It chooses parameters that maximize the likelihood of observing the sample values rather than that minimize the sum of squared errors (like in ordinary regression).

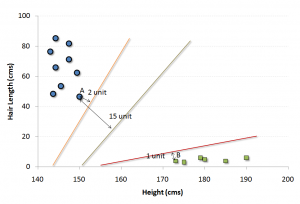
## **3. SVM (Support Vector Machine)**

It is a classification method. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate.

For example, if we only had two features like Height and Hair length of an individual, we’d first plot these two variables in two-dimensional space where each point has two co-ordinates (these co-ordinates are known as **Support Vectors**)



Now, we will find some line that splits the data between the two differently classified groups of data. This will be the line such that the distances from the closest point in each of the two groups will be farthest away.



In the example shown above, the line which splits the data into two differently classified groups is the black line, since the two closest points are the farthest apart from the line. This line is our classifier. Then, depending on where the testing data lands on either side of the line, that’s what class we can classify the new data as.

## **4. kNN (k- Nearest Neighbors)**

It can be used for both classification and regression problems. However, it is more widely used in classification problems in the industry. K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases by a majority vote of its k neighbors. The case being assigned to the class is most common amongst its K nearest neighbors measured by a distance function.

These distance functions can be Euclidean, Manhattan, Minkowski and Hamming distance. First three functions are used for continuous function and fourth one (Hamming) for categorical variables. If K = 1, then the case is simply assigned to the class of its nearest neighbor. At times, choosing K turns out to be a challenge while performing kNN modelling.



KNN can easily be mapped to our real lives. If you want to learn about a person, of whom you have no information, you might like to find out about his close friends and the circles he moves in and gain access to his/her information!

**Things to consider before selecting kNN:**

* KNN is computationally expensive
* Variables should be normalized else higher range variables can bias it
* Works on pre-processing stage more before going for kNN like an outlier, noise removal

## **5. Random Forest**

Random Forest is a trademark term for an ensemble of decision trees. In Random Forest, we’ve collection of decision trees (so known as “Forest”). To classify a new object based on attributes, each tree gives a classification, and we say the tree “votes” for that class.

The forest chooses the classification having the most votes

(Over all the trees in the forest).

Each tree is planted & grown as follows:

1. If the number of cases in the training set is N, then sample of N cases is taken at random but with replacement. This sample will be the training set for growing the tree.
2. If there are M input variables, a number m<<M is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing.
3. Each tree is grown to the largest extent possible. There is no pruning.

**TABLE OF CONTENTS**

**PAGE**

1. **CERTIFICATE .…………………………………………………………………….i**
2. **CANDIDATE’S DECLARATION………………………………………………...ii**
3. **ACKNOWLEDGEMENT…………………………………………………………iii**
4. **ABSTRACT………………………………………………………………………....iv**
5. **INTRODUCTION…………………………………………………………………...1-5**
6. **ABOUT THE COMPANY………………………………………………………….5.1**
7. **PRACTICAL TRAINING………………………………………………………….6-31**
8. **CONCLUSION………………………………………………………………………32**
9. **REFERENCES………………………………………………………………………33**

**Introduction to Medical Health Expense Prediction**

## Problem Statement

Everyone’s life revolves around their health. Good health is essential to all aspects of our lives. Health refers to a person’s ability to cope up with the environment on a physical, emotional, mental, and social level. Because of the quick speed of our lives, we are adopting many habits that are harming our health. One spends a lot of money to be healthy by participating in physical activities or having frequent health check-ups to avoid being unfit and get rid of health disorders. When we become ill, we tend to spend a lot of money, resulting in a lot of medical expenses.

So, an application can be made which can make people understand the factors which are making them unfit, and creating a lot of medical expenses, and it could identify and estimate medical expense if someone has such factors.

## Objective

* Predict the future medical expenses of subjects based on certain features building a robust machine learning model.
* Identifying the factors affecting the medical expenses of the subjects based on the model output.

**DETAILS OF THE PROJECT**

1. **Data Collection.**

* The dataset required for this project was taken from *“kaggle.com”*.
* Kaggle allows users to find and publish datasets to solve challenges or web-based data science projects*.*

1. **About the Dataset.**

* For this project, the data has been imported from the machine learning repository. The dataset contains 1338 rows and 7 columns. The columns present in the dataset are ‘age’,’ sex’,’bmi’, ’children’, smoker’, ’region’, and ‘expenses’. The Expenses column is the target column and the rest others are independent columns. Independent columns are those which will predict the outcome.
* The first column is Age. Age is an important factor for predicting medical expenses because young people are generally more healthy than old ones and the medical expenses for Young People will be quite less as compared to old people.
* The Next column is sex, which has two Categories in this column: Male and Female. The sex of the person can also play a vital role in predicting the medical expenses of a subject.
* After that, you have the ‘bmi’ column, then**BMI is Body Mass Index*.***
* For most adults, an ideal BMI is in the 18.5 to 24.9 range.
* For children and young people aged 2 to 18, the BMI calculation takes into account age and gender as well as height and weight. If your BMI is less than 18.5, you are considered underweight. People with very low or very high ‘bmi’ are more likely to require medical assistance, resulting in higher costs.
* The fourth column is the ‘children’ column, which contains information on how many children your patients have. Persons who have children are under more pressure because of their children’s education, and other needs than people who do not have children.
* The fifth is the ‘smoker’ column. The Smoking factor is also considered to be one of the Most Important factors as the people who smoke are always at risk when their age reaches 50 to 60.
* Next is the ‘region’ column. Some Regions are Hygienic, Clean, Neat, and Prosperous, but some Regions are not, and this information affects health which is related to medical expenses.

1. **ADVANTAGES**

* Understood the importance of Data Analysis and Data Visualization for determining the association between features.
* How to build intuition by building insights from Data Visualization.
* How to deal with categorical variables.
* Learnt about different Assumptions to be satisfied before using a Linear Regression Model.

**More things which we could try: -**

* We can give different labels to each Region. It might give better results.
* Keep 4 and 5 number of children in our analysis instead of capping them and see how the results vary.
* We can try some more predictive models and compare the results.
* Try converting Expense column to a normal distribution using log or square root transformation.

**SOFTWARE REQUIREMENTS**

* os: windows or Linux
* python IDE: python 2.7.x and above
* Jupyter notebook
* setup tools and pin to be installed for 3.6 and above
* language python.

**HARDWARE REQUIREMENTS**

* RAM:4GB and higher
* processor: intel i3 and above
* hard disk:500GB: minimum

**LIBRARY USED IN THIS PROJECT**

* **NumPy:** - It is a general-purpose array processing package.it provides a high-performance multidimensional array object, and tools for working with these arrays. it’s also an efficient multidimensional container of generic data.
* **Pandas: -** It is the most popular python library that is used for data analysis.it provides highly optimized performance with back-end source code is purely written is **C** or **python. subdivided** into two parts it is: - **series and data frame.**
* **Matplotlib: -** It is a python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environment across platforms .it tries to make easy things easy and hard things possible. you can generate plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with just a few lines of code.
* **Seaborn: -** It is a python data visualization library based on matplotlib.it provides a high-level interface for drawing attractive and informative statistical graphics.it is closely integrated with pandas data structures.
* **Scikit-learn: -** it is a free machine-learning library for python programming language.it features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means, DBSCAN and is designed to incorporate with the python numerical and scientific libraries numpy and scypy
* **Plotly**
* The [plotly Python library](https://plotly.com/python/) is an interactive, [open-source](https://plotly.com/python/is-plotly-free) plotting library that supports over 40 unique chart types covering a wide range of statistical, financial, geographic, scientific, and 3-dimensional use-cases.
* Built on top of the Plotly JavaScript library ([plotly.js](https://plotly.com/javascript/)), plotly enables Python users to create beautiful interactive web-based visualizations that can be displayed in Jupyter notebooks, saved to standalone HTML files, or served as part of pure Python-built web applications using Dash.
* The plotly Python library is sometimes referred to as "plotly.py" to differentiate it from the JavaScript library.

**ABOUT THE COMPANY**

* Academy Of Skill Development (ASD) an MSME is a non-profit organization registered with Govt. of West Bengal. ASD has affiliations to conduct global certification of Microsoft Technology Associate (MTA), Autodesk Certified User (ACU), Adobe Certified Associate (ACA), EC-Council Associate Certifications. Mission of ASD is to develop skills of the candidates to meet the dynamics of demand in the job market.
* ASD looks forward to bridge the gap between the industry and academia. Recent survey has revealed the huge gap between demand and supply of candidates. Today industry is looking for ready-made candidates with right set of skills. Hence employability enhancement training and internship is very important. ASD provides the latest and best Skill Development Internships.
* Academy of Skill Development is different because it focusses on practical project based industrial training internships and the programs are conducted with case studies using real life scenarios.
* The content of the training programs is based upon providing insightful, high quality, high value, and innovative inputs from the corporate and from the professors of IITs and NITs. The programs are outcome based that deliver measurable impact and outstanding results.
* It contains skillfully blend proven teaching practices with a consultative, result-centric outlook that emphasizes the practical application of knowledge in line with each students' strategic needs and aims. All the mentors are having industry experience and are certification holder from MNCs like Microsoft, HPE, Oracle, etc.

A picture containing logo

Description automatically generated

**Practical Training Work**

ACTUAL CODES WITH OUTPUT

**Importing necessary libraries and setting parameters for visualization**

Text

Description automatically generated

**Explanation: -**

We import the numpy, pandas, matplotlib, seaborn and plotly (provides high-level interface for attractive and informative statistical graphics). Also, settings parameters for visualization.

**Importing dataset and finding its shape.**

Text

Description automatically generated

Here ,1338 – No. of rows, 7 – columns

**Explanation: -**

The csv file(dataset) is read using the “pd.read\_csv(‘filename’)” command.

**Understanding the Data**

Table

Description automatically generated

**Explanation: -**

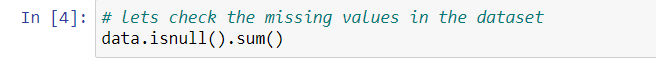
head (), gives us a quick look at our dataset.

**Output: -**

Table

Description automatically generated

**Finding Null Values**



**Output: -**

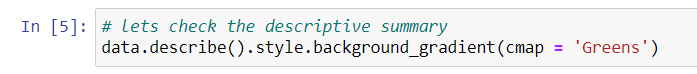
A picture containing table

Description automatically generated

**Explanation: -**

* For finding the number of missing values in a dataset we use data. isnull(). We call data.isnull().sum() to give the output of series containing data about count of NaN in each column.
* We get no null values…thus the data set has no missing values.

**Describing the Data**



**Explanation: -**

Here we check the descriptive summary in which we get parameters like age, bmi, children, expenses.

**Output: -**

Table

Description automatically generated

**Univariate analysis**

Text

Description automatically generated

**Explanation: -**

**Univariate Analysis**

* It involves only one variable.
* It is used to understand the distribution of the variables present in the dataset and derive meaningful insights from them.
* It can be used to check the distribution of both Numerical as well as Categorical variables.

**Output: -**

Chart, bar chart, histogram

Description automatically generated

**Distribution of Smoker, Children, and Regions**

* Here we have used a pie chart to plot the Smoker Column, as the Smoker column has only two values: **Yes and No**.
* Using a Count plot, we have shown the subjects having children ranging from 0 to 5 and it has been computed and observed from the count plot also that those who are having no children are highest in number.
* We have again used a pie chart to plot the number of inhabitants in the region column which consists of four segments: Northeast, Northwest, Southeast, Southwest

**Checking distribution of Age, Bmi and Expenses**

Text

Description automatically generated

**Explanation: -**

* Here we check the distribution of Age, Bmi, expenses which will be represented by different colours.
* The distribution of Age, BMI, and Expenses have been shown using the distribution plot

**Output-**

Chart, line chart, histogram

Description automatically generated

* Here, we have an equal number of people of all ages.
* The BMI of the patients seems to be normally distributed where maximum people have BMI around 30 and very few people have less BMI around 10, similarly very few people have high BMI around 60.
* Expenses column seems to be right-skewed. Here log transformation has been applied to make this distribution normal so it doesn’t create any problem while predicting.

**Check impact of age on Expenses**

Text

Description automatically generated with medium confidence

**Explanation: -**

**Bivariate Analysis:**

* Here we use scatter plot to check the impact of age on medical expenses and Bivariate analysis.
* Bivariate Analysis is one of the simplest forms of quantitative analysis.
* It involves analysis of two variables for determining the empirical relationship between them.
* It can be helpful in testing simple hypothesis of association.

**Output: -**

Chart, scatter chart

Description automatically generated

* With Increasing Age, Expense is expected to increase, but It is not obvious for all the scenarios.

**Check impact of bmi on Expenses**

Graphical user interface, application

Description automatically generated

**Explanation: -**

Again, we use scatter plot to check the impact of bmi on medical expenses and also use Bivariate Analysis.

**Output: -**

Chart, scatter chart

Description automatically generated

**Check the impact of smoking and children in Medical Expenses**

Text

Description automatically generated

**Explanation: -** Here we use boxplot to check the impact of smoking and children in medical expenses

**Output: -**

Chart, box and whisker chart

Description automatically generated

* We plotted the bar plot between the children and expenses, and it has been noticed people with a greater number of children the expense is more as with more children parents need to take care of the health of all of them rather those who have no children or one child, but as there is very less number of people who are having more than 3 children so for them the expense is almost same so we capped them. The people with 3 children have the highest expenses among them.
* We have also seen the relationship between smokers and expenses through box plots and it has been noticed that for smokers the expense is much high than non-smokers as it is obvious because smoking is injurious to health, so smokers are likely to have health issues than non-smokers causing their medical expenses to increase.

**Multivariate Analysis**

A picture containing text

Description automatically generated

Rectangle

Description automatically generated with low confidence

Pair plot between age and expenses for different regions and children

**Output-**

A picture containing Teams

Description automatically generated

* The Expenses of Smokers in all regions ranges from 20 to 60k
* Whereas the Expenses of Non-Smokers in all regions ranges from 10 to 20K
* The Lesser range of Expense is for lesser age people and vice versa.

**EXPLANATION**

* A scatter plot has been created between the age and expenses column, Specifying the smoker column to impart colour in data points. If the patient is a smoker, he or she will be specified with blue colour, and if the patient is a non-smoker, then he or she will be specified with red colour.
* After that, add faceting layers. the “children” column has been added as a facet on the row side, whereas the “Region” column has been added as a facet on the column side. If we look at the Charts Carefully, we can easily figure out that the Expenses of Smokers.

**Again, created a similar kind of scatter plot, where we plotted BMI in the place of age.**

Shape, rectangle

Description automatically generated with medium confidence

**Output-**

A picture containing company name

Description automatically generated

**Pair plot between BMI and expenses for different regions and children**

**Explanation**

* We can clearly see that there is an increasing pattern for BMI as well.
* For smoker with less BMI: Expense is around 20k
* For Smokers with High BMI: Expense is around 50K
* For Non-Smokers BMI is not a Huge Factor, The Expense range from 5k to 10k

**Bubble Chart**

This Chart to Represent the relation of Expense with BMI, Age, smoking only for north west region.

Text

Description automatically generated

**Output- Relation of Expense with BMI, age, and smoking**

Chart, scatter chart, bubble chart

Description automatically generated

**Explanation-**

* This Chart makes it clear that BMI is not powerful indicator Expenses, as people having less BMI also have high Medical Expenses.
* This chart makes it clear that People who smoke have higher Medical Expenses.
* The Size of Bubble, which represents age, shows that people having higher expenses belong to Higher Expenses category

**Bar Polar Chart**

Radar chart

Description automatically generated

**Relation of Expense with Regions**

**Explanation-**

* This Chart clearly depicts that the Southeast region has higher expenses compared to other regions.
* This Chart clearly shows that Males have Higher Expenses in general in all the regions.

**Impact checking and no column removal process**Text

Description automatically generated with medium confidence

**Explanation-**

We check the impact of regions in expenses and no columns are removed as all columns contains important data.

**Data Pre-Processing**Text

Description automatically generated

A picture containing text

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

**Explanation: -**

**Data Pre-processing** refers to the steps applied to make data more suitable for data mining. The steps used for Data Pre-processing usually fall into two categories:

1. selecting data objects and attributes for the analysis.
2. creating/changing the attributes.

Here, Data Pre-Processing used in code is explained

* **Encoding of categorical columns**
* In our dataset, we have three categorical columns: Sex, Smoker, and Region.
* Sex consists of males and females and it has been observed from the data analysis that males have higher medical expenses than females. As a result, we have encoded male as 2 and

female as 1.

* The Smoker column consists of smokers and non-smokers and it has been observed the smoker has more medical expenses than the non-smoker ones so we encoded smokers as 2 and non-smokers as 1.
* The Region column comprises four segments: Southeast, Southwest, Northeast, Northwest.
* It has been noticed from the data analysis the Southeast region has the highest expenses followed by Northeast, Northwest, and Southwest, so we have encoded those regions as 4,3,2,1 respectively.
* **Data Splitting**

We have split the data in the entire dataset in Train and Test with a ratio of 80:20. We have trained the model with trained data and then we applied the model to the test dataset to check the performance of the model.

* **Feature Scaling**

Feature scaling is a method used to normalize the range of independent variables or features of data. Feature Scaling basically helps to normalize the data within the same scale. We have used the fit\_transform function to apply feature scaling on the training data but used the transform function to apply feature scaling on the testing data because the fit\_transform function will first learn all the patterns in the data and then it will apply feature scaling on the data. But we cannot use the fit\_transform function for testing data as you cannot learn the patterns from these data.

**What Is Predictive Modeling?**

* Predictive modeling 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 analyzing current and historical data and projecting what it learns on a model generated to forecast likely outcomes.
* A predictive model is not fixed; it is validated or revised regularly to incorporate changes in the underlying data. In other words, it’s not a one-and-done prediction. Predictive models make assumptions based on what has happened in the past and what is happening now.
* If incoming, new data shows changes in what is happening now, the impact on the likely future outcome must be recalculated, too.

Graphical user interface, text, application

Description automatically generated

**Now comparing all the models and seeing which is best!**

**Random forest model**

* Random Forest is an ensemble learning method for classification and regression by constructing multiple numbers of decision trees at training time.
* Outputting the average prediction of the individual trees in case of regression whereas outputting the class that is the mode of the classes in case of classification.
* It is one of the Most powerful Machine Learning algorithms which works well in most cases.
* First of all, the *RandomForestRegressor* package was imported from sklearn. ensemble library, so that we can use this model to predict the Expenses.
* After that, we specified a Model using this Random Forest Regressor Class. Now, as the Model is Ready, trained our Model using the Training data, for that fit function, was used and used the training data.
* Here, the Training Data refers to *x\_train and y\_train*.  
  Where x\_train is the independent variable, and Y\_train is the dependent variable or target variable.
* After the Model gets trained, we started performing Predictions using this Predictive Model created using the Random Forest Regressor.To do that, used predict function and specified the independent variables inside the function.
* To get the predictions and saved the result produced by the Random Forest in a new variable, so that we can compare the Results later if required. After building the predictive model, evaluated the Model using various Performance Metrics. In this case, we checked the R2 score and RMSE Score as in the case of the linear regressor.

Text

Description automatically generated

**Gradient Boosting Regressor**

* Gradient Boosting is a very popular Boosting technique.
* It works by sequentially adding the previous predictors under fitted  
  predictions to the ensemble ensuring that the errors made previously are corrected.
* Random Forest is also an Ensemble, But in Random Forest the Ensembling happens Parallelly,
* But, In the case of Gradient Boosting, the Ensembling happens Sequentially, which means that the First Model’s Errors will be used to Build the Second Model and the Second Model’s Error will be used to Build the Third Model, and so on.
* The Models will be built until and unless the Errors are optimized in the best way. That means by using the Gradient Boosting Models we can make the least error possible.
* First of all, Using the Gradient Boosting Regressor Model is to be imported from the *sklearn. Ensemble library*. After that, make a Base Gradient Boosting Regressor Model, and then we trained this Model using the fit function on the Training Data that is x\_train and y\_train.
* Where, X\_train is your Independent Variable, and y\_train is your Dependent Variable. After the Model is built, we predicted the Target Variable for our test data using the predict function and save the result in a new variable, so that to compare the results later.
* After that, we performed Model Evaluation using the R2 score and RMSE score Performance Metrics.

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

**Linear Regression**

* Linear regression is probably one of the most important and widely used regression techniques. It’s among the simplest regression methods. One of its main advantages is the ease of interpreting results.
* It is a very powerful technique and can be used to understand the factors that influence profitability. It can be used to forecast sales in the coming months by analyzing the sales data for previous months. It can also be used to gain various insights about customer behavior.
* The objective of a linear regression model is to find a relationship between one or more features (independent variables) and a continuous target variable (dependent variable). When there is only feature it is called Uni-variate Linear Regression and if there are multiple features, it is called Multiple Linear Regression.

Graphical user interface, text

Description automatically generated

* As we can see that the scores are not varying much, so we can say that this model is good.

**Now, final comparison of r2 score of all models**

Text

Description automatically generated

Chart, bar chart

Description automatically generated

**Comparison of R2 Score of three models**

**We have built three models among which the Gradient Boosting Regressor model shows the best result.**

**CONCLUSION**

* We came to know that the Most Important Factor to Predict the Medical Expenses of a subject is Smoking Behaviour and Age, that means, smoking is Bad for Health, as already know that and which inevitably increases medical expenses as due to smoking one is likely to fall ill more than the non-smokers.

We also found that with increasing of age, one needs to take some more care and precautions for your health as with the increase of age health becomes fragile, so they go for frequent medical check-up, likely to fall ill quickly as with the increase of age immunity falls so they adopt measures to stay healthy by taking medicines and engaging in some physical activities like jogging, walking, Yoga which causes an increase of medical expenses.

* Apart from that you also understood that Gender, Number of Children, the Region also have a good impact on determining Medical Expenses.
* We have built three models among which the Gradient Boosting Regressor model shows the best result through which we can say 83.2% variability of expenses can well be explained by predictor variables and which yields comparatively low RMSE value so our predicted expense through this model will not vary too much from the actual expense.

**Takeaways from project**

* Learned different predictive models such as Linear Regression, Random Forest and Gradient Boosting model.
* Learned how to ensemble different models.
* Understood the importance of cross validation on the data.
* Learned the importance of comparing different *predictive models.*

**BIBLIOGRAPHY**

* https://www.youtube.com/c/DataisGood
* www.javatpoint/machine-learning
* [www.kaggle.com](http://www.kaggle.com/)
* <https://scikit-learn.org/>
* https://towardsdatascience.com/polynomial-regression-with-scikit-learn-what-you-should-know-bed9d3296f2