Predicting Insurance Claims Using Demographic and Health Data 2022-2023

Name: M.KARTHIK
Dept. No: 20-UST-021
Name of the **Guide: Mr. M. Syluvai Anthony**

DEPARTMENT OF STATISTICS

LOYOLA COLLEGE (AUTONOMOUS)

CHENNAI – 600 034

Predicting Insurance Claims Using Demographic and Health Data

Submitted to

LOYOLA COLLEGE (AUTONOMOUS)

In

Partial fulfillment of the requirements for the award of the degree of
BACHELOR OF SCIENCE IN STATISTICS

By

M. KARTHIK

Under the Guidance of

Mr. M. Syluvai Anthony

DEPARTMENT OF STATISTICS

LOYOLA COLLEGE (AUTONOMOUS)

CHENNAI-34

DECLARATION

I, KARTHIK M (20-UST-021) from department of statistics declare that the work entitled **"Predicting Insurance Claims Using Demographic and Health Data"** is prepared by me as a partial fulfilment for the requirement of award for Bachelor of Science in statistics, the academic year of this project is during 2022-2023.

Signature

KARTHIK M

ACKNOWLEDGEMENT

For the successful accomplishment of my project, I would like to express my gratitude to my Head of the department **Dr. Edwin Prabakaran**. I would also like to thank the Loyola college management for providing this opportunity and I thank my faculty guide **Mr. M. Syluvai Anthony**. I express my thanks to my friend who helped me in doing this project.

BONAFIDE CERTIFICATE

This is to certify that "**Predicting Insurance Claims Using Demographic and Health Data**" is a bonafide record done by

KARTHIK M (20-UST-021) in the partial fulfillment of the requirement for the degree of B.sc statistics, Loyola College(autonomous), Chennai – 34 under the guidance of Mr. M. Syluvai Anthony, during the year 2022- 2023.

Dr. Edwin Prabakaran.T.

Head of the department

Department of statistics

Loyola College

Chennai - 34

Mr. M. Syluvai Anthony.

Assistant Professor

Department of statistics

Loyola College

Chennai-34

INDEX

ABOUT THE STUDY

This dataset contains insightful information related to insurance claims, giving us an in-depth look into the demographic patterns of those receiving them. The dataset contains information on policyholder's age, gender, BMI (Body Mass Index), steps (average steps walked by a person), medical costs charges, number of children, smoking status and region. By analyzing these key factors across geographical areas and across different demographics such as age or gender we can gain a greater understanding of who is most likely to receive an insurance claim.

This understanding gives us valuable insight that can be used to inform our decision making when considering potential customers for our services. On a broader scale it can inform public policy by allowing for more targeted support for those who are most in need and vulnerable.

These kinds of insights are extremely valuable and this dataset provides us with the tools we need to uncover them!

DATASET

The process of gathering and analyzing accurate data from various sources to find answers to research problems, trends and probabilities, etc., to evaluate possible outcomes is Known as Data Collection. It includes understanding the data to study the hidden patterns and trends which helps to predict and evaluate the results. The methods of data collection can vary depending on the type of data being collected and the purpose for which it is being collected. Here are some common methods of data collection: Surveys, Interviews, Observations, Experiments, Secondary data analysis, Case studies, etc.

The methods I used for this data collection is Secondary data analysis that means it involves analyzing data that has already been collected by others for a different purpose. This method can be used to answer research questions that cannot be answered with new data collection.

The data collected is originally from a person named Sumit Kumar Shukla and the dataset is an insurance dataset available in Kaggle. It several individual's health status, lifestyle habits, and financial situation analyst variables and one binary target variable.

The objective of this dataset is to predict whether an individual will make an insurance claim or not, based on their demographic and lifestyle characteristics. The dataset consist of several independent variables and one dependent variable, i.e., the outcome. Independent variables include policyholder’s age, gender, BMI (Body Mass Index), steps (average steps walked by a person), medical charges bills, number of children, smoking status and region.

Data collection

- The insurance dataset consists of 1338 observation data points, with 9 features each.
- "insuranceclaim" is the feature we are going to predict, value inside this feature are 0 and 1 that means "not claim" and "claim" respectively.
- There is no null values in the dataset.
- The dataset format is stored in CSV format.

Variables description and understanding:

Serial No.	Variables	Description
1	Age	Age of the policyholder
2	Sex	Gender of the policyholder (female=0, male=1)
3	BMI	Body Mass Index, providing an understanding of the 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 25
4	Steps	Average walking steps per day of the policyholder
5	Children	Number of children/dependents of the policyholder
6	Smoker	Smoking state of policyholder (non-smoke=0, smoker=1)
7	Region	The residential area of the policyholder in the US (northeast 0, northwest=1, southeast-2, southwest-3)
8	Charges	Individual medical costs billed by health insurance.
9	Insurance claim	Number of insurance claimed or not (not claimed =0, claimed=1)

INTRODUCTION

Welcome to the project introduction for a data analysis predictive binary project about insurance claims. The purpose of this project is to predict whether an individual will make an insurance claim or not, based on their demographic and lifestyle characteristics.

The independent variables that we will use in this project are age, sex, BMI (Body Mass Index), number of average steps walked per day, number of children, smoking status, region of residence, and medical costs charges. These variables are important as they can provide insights into an individual's health status, lifestyle habits, and financial situation, which may be relevant in predicting their likelihood to make an insurance claim.

The dataset used in this project will be obtained from a health insurance company, which contains information on individuals who have purchased insurance policies from the company. The dataset will be preprocessed, cleaned, and transformed to ensure that it is suitable for analysis.

To build the predictive model, we will use a binary classification algorithm, such as logistic regression or random forest, to train the model using the available data. The model will be evaluated using metrics such as accuracy, precision, recall, and F1-score.

The results of this project can be useful for insurance companies to identify individuals who are more likely to make a claim and adjust their pricing and policies accordingly. Additionally, individuals can also use the insights provided by this project to make informed decisions when choosing insurance policies.

DATA PRE-PROCESSING

Data preprocessing refers to the process of preparing raw data for analysis by cleaning, transforming, and organizing it into a suitable format. It is a critical step in data analysis and machine learning, as the quality of the data directly affects the accuracy and effectiveness of the analysis or model.

Overall, data preprocessing is an essential step in data analysis and machine learning that ensures the accuracy, reliability, and usefulness of the results obtained.

Operations:

I had used **Pyhton** for this purpose as it has the rich collection of machine learning libraries and mathematical operation. I will mostly use common packages as **pandas, numpy, matplotlib, seaborn, scipy, tabulate and sklearn** which help me for mathematical operations and also plotting diagrams, importing and exporting of files.

```
In [1]: import pandas as pd
   ...: import numpy as np
   ...: import matplotlib.pyplot as plt
   ...: import seaborn as sns
   ...: from scipy.stats import chi2_contingency
   ...: from sklearn.model_selection import train_test_split
```

Variables data types are

Binary categorical variables: ‘sex’, ‘smoker’ and 'insuranceclaim’.

Category variable: ‘region’.

Continuous variables: ‘age’,‘bmi’, ‘steps’, ‘children’ and ‘charges’.

Data pre-processing

i. Import and get to know the data

```
In [2]: cd C:\Users\KARTHIK M\Documents\PROJECTS\Final year project
C:\Users\KARTHIK M\Documents\PROJECTS\Final year project

In [3]: data=pd.read_csv("insurance3r2.csv")
   ...: pd.set_option('display.max_columns', None)
   ...: data.head()

Out[3]:
   age  sex     bmi  steps  children  smoker  region      charges  \
0   19    0  27.900   3009         0       1       3  16884.92400
1   18    1  33.770   3008         1       0       2   1725.55230
2   28    1  33.000   3009         3       0       2   4449.46200
3   33    1  22.705  10009         0       0       1  21984.47061
4   32    1  28.880   8010         0       0       1   3866.85520

   insuranceclaim
0               1
1               1
2               0
3               0
4               1

In [4]: data.shape
Out[4]: (1338, 9)
```

Input 2 represent the set current directory where the dataset present.

Input 3 represent importing ‘insurance3r2’ dataset which in CSV format and ‘set_option’ function helps to display full column of the dataset.

Input 4 show us the number of observation (row) and number of variables (column) in the dataset.

Data pre-processing

```
In [5]: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             1338 non-null   int64
 1   sex             1338 non-null   int64
 2   bmi             1338 non-null   float64
 3   steps           1338 non-null   int64
 4   children        1338 non-null   int64
 5   smoker          1338 non-null   int64
 6   region          1338 non-null   int64
 7   charges         1338 non-null   float64
 8   insuranceclaim  1338 non-null   int64
dtypes: float64(2), int64(7)
memory usage: 94.2 KB
```

Input 5 show us the data types of each variables and missing detection of each variables in dataset.

```
In [6]: data['sex'] = pd.Categorical(data['sex'])
   ...: data['smoker'] = pd.Categorical(data['smoker'])
   ...: data['region'] = pd.Categorical(data['region'])
   ...: data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 9 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             1338 non-null   int64
 1   sex             1338 non-null   category
 2   bmi             1338 non-null   float64
 3   steps           1338 non-null   int64
 4   children        1338 non-null   int64
 5   smoker          1338 non-null   category
 6   region          1338 non-null   category
 7   charges         1338 non-null   float64
 8   insuranceclaim  1338 non-null   int64
dtypes: category(3), float64(2), int64(4)
memory usage: 67.2 KB
```

Input 6 changes the respective categorical variables data type into category.

Data pre-processing

```
In [7]: data.describe()
Out[7]:
                age          bmi         steps     children       charges
count  1338.000000  1338.000000   1338.000000  1338.000000   1338.000000
mean     39.207025    30.663397   5328.623318     1.094918  13270.422265
std      14.049960     6.098187   2453.643236     1.205493  12110.011237
min      18.000000    15.960000   3000.000000     0.000000   1121.873900
25%      27.000000    26.296250   3008.000000     0.000000   4740.287150
50%      39.000000    30.400000   4007.000000     1.000000   9382.033000
75%      51.000000    34.693750   8004.000000     2.000000  16639.912515
max      64.000000    53.130000  10010.000000     5.000000  63770.428010
```

Input 7 describe each continuous variables in numerical data properties are Central Tendency and Variation for better understanding of dataset.

i. Cleaning missing observation:

```
In [8]: data=data.dropna()
```

There are no missing data present in this dataset in case if there, the input 8 will remove it out of dataset.

ii. Data transformation:

```
In [9]: data['sex'] = data['sex'].replace({0:'Female', 1: 'Male'})
   ...: data['smoker'] = data['smoker'].replace({0:'non-smoker', 1: 'smoker'})
   ...: data['region'] = data['region'].replace({0:'northeast', 1: 'northwest',2:'southeast',3:'southwest'})
   ...: data.head()
Out[9]:
   age     sex     bmi  steps  children      smoker     region      charges  \
0   19  Female  27.900   3009         0      smoker  southwest  16884.92400
1   18    Male  33.770   3008         1  non-smoker  southeast   1725.55230
2   28    Male  33.000   3009         3  non-smoker  southeast   4449.46200
3   33    Male  22.705  10009         0  non-smoker  northwest  21984.47061
4   32    Male  28.880   8010         0  non-smoker  northwest   3866.85520

   insuranceclaim
0               1
1               1
2               0
3               0
4               1
```

Input 9 help us to give respective category values to respective categorical variables in order to show better understanding in EDA.

Awarding dummy variables:

Data pre-processing

A dummy variable is a binary variable that is used to represent categorical variables in a statistical model. Dummy variables are also known as indicator variables, design variables, or binary variables. Statistical models require that all variables be numerical. Therefore, we need to convert categorical variables into numerical variables. This is where dummy variables come in.

We use dummy variables for several reasons:

To enable us to include categorical variables in regression models or machine learning algorithms, which typically require numerical inputs.

To avoid the assumption that the categories of a categorical variable have an inherent order or magnitude, which may not be appropriate for some variables.

To enable us to estimate the effects of different categories of a categorical variable separately, rather than assuming a single effect for the entire variable.

```
In [10]: # One-hot encoding for 'region'

In [11]: data = pd.get_dummies(data, columns=['region'])

In [12]: data = data.drop('region_southeast', axis=1)

In [13]: # label encoding for 'sex' and 'smoker'

In [14]: data['sex'] = data['sex'].astype('category').cat.codes
    ...: data['smoker'] = data['smoker'].astype('category').cat.codes
    ...: data.head()
Out[14]:
   age  sex     bmi  steps  children  smoker      charges  insuranceclaim  \
0   19    0  27.900   3009         0       1  16884.92400               1
1   18    1  33.770   3008         1       0   1725.55230               1
2   28    1  33.000   3009         3       0   4449.46200               0
3   33    1  22.705  10009         0       0  21984.47061               0
4   32    1  28.880   8010         0       0   3866.85520               1

   region_northeast  region_northwest  region_southwest
0                 0                 0                 1
1                 0                 0                 0
2                 0                 0                 0
3                 0                 1                 0
4                 0                 1                 0
```

Data pre-processing

The above inputs helps to create variables of different kind that are:

One hot encoding:

Creates a new binary feature for each category in a feature. Each binary feature represents whether or not the observation belongs to that category. Each observation would have a value of 1 in the corresponding binary feature and 0 in all other binary features. One hot encoding ensures that there is no inherent order or hierarchy among the categories.

When using these dummy variables in a model, we typically drop one of them to avoid multicollinearity issues. This means that one of the dummy variables serves as the reference category, and the others are compared to it. This reference category is also sometimes called the "baseline" category.

Here we have took ‘region_southeast’ dummy variable as reference category that is "baseline" category, that’s why we removed it you can see in input 12. Because it the large frequency category among the region that is 27.2% from EDA.

Label Encoding:

It involves assigning a unique numerical value to each category in a feature. In label encoding, we would assign the values 0 and 1 to these categories, respectively. The problem with label encoding is that the numerical values may have an inherent order or hierarchy, which may not always be accurate.

We apply dummy variables after doing EDA for better understanding of EDA presentation.

EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is the process of analyzing and understanding a dataset in order to summarize its main characteristics and gain insights into the data. EDA is typically done using visual and quantitative methods, and is often the first step in the data analysis process.

Things that come under EDA include:

- **Data visualization:** This involves creating various types of visualizations, such as histograms, scatter plots, or box plots, to explore the data's distribution, patterns, and relationships. Those Visualization diagrams can be divided into two variants Univariate and Bivariate Analysis.

- **Univariate Analysis:** Univariate analysis involves analyzing a single variable or feature in the data set. It focuses on understanding the distribution of the variable, including its central tendency, dispersion, and shape. Univariate analysis techniques include calculating summary statistics (such as mean, median, mode, and standard deviation), creating histograms and density plots, and calculating probability distributions. Univariate analysis is useful for identifying outliers, understanding the range of values for a given variable, and determining whether the variable is normally distributed

- **Bivariate Analysis:** Bivariate analysis involves analyzing two variables in the data set to understand the relationship between them. It focuses on understanding how changes in one variable affect changes in another variable. Bivariate analysis techniques include scatter plots, correlation analysis, and contingency tables. Bivariate analysis is useful for identifying relationships between variables, determining the strength and direction of those relationships, and identifying potential causal factors.

- **Data summarization:** It is a process of summarizing the main characteristics of a dataset using various statistical measures and methods. The purpose of data summarization is to make it easier to understand the data and to identify any patterns or trends that may exist in the data.

  Data summarization techniques include creating frequency tables, histograms, box plots, and scatter plots. These visualizations can help to identify patterns and trends in the data and to identify any outliers or anomalies.

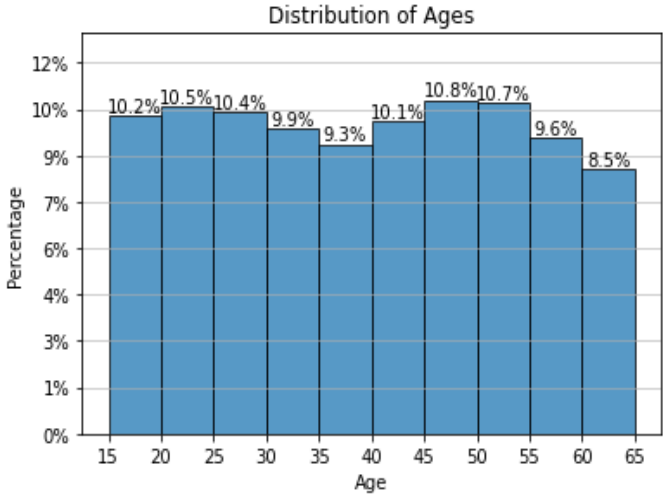
  Data summarization is an important step in exploratory data analysis (EDA) because it provides a quick and easy way to gain insights into the data and to identify any issues or problems that may need further investigation. By summarizing the data, analysts can quickly identify the range, distribution, and patterns in the data, which can inform further analyses and modeling.

i. Univariate Analysis:

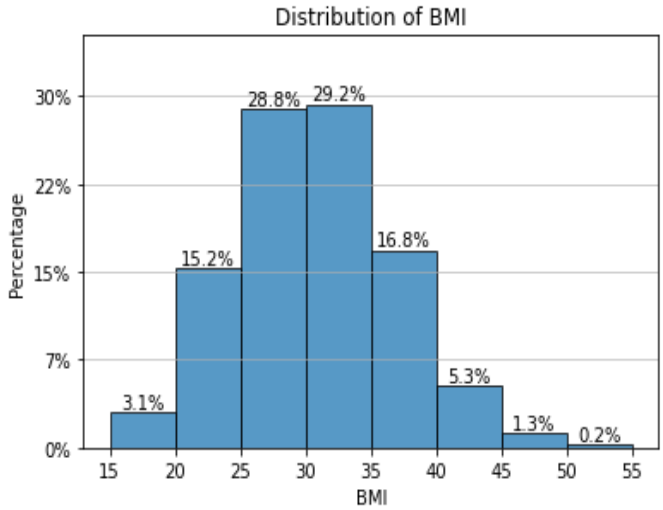
a) Continuous variables:

For continuous variables we can use histogram.

A histogram is a graphical representation of the distribution of numerical data. It consists of a series of rectangles, where the area of each rectangle represents the frequency or proportion of observations within a certain range or bin of values. The x-axis represents the range of values being measured, while the y-axis represents the frequency or count of observations falling within each bin. Histograms are commonly used to visualize the distribution of continuous data.

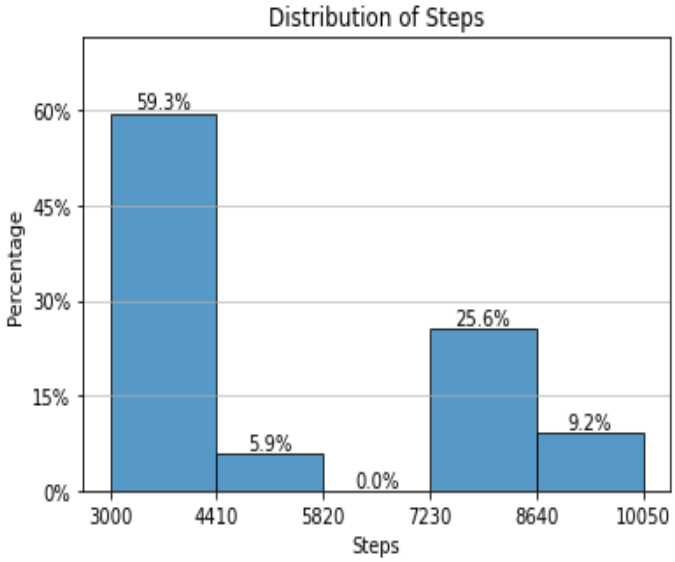


Age distribution is somewhat evenly spread between the 15-19 to 55-60 age groups, with each group representing roughly 9-11% of the population. However, there is a slightly higher proportion of policyholders in the 45-54 age range, which represents over 10% of the population and lower proportion of policyholders in the older age ranges (55-65).

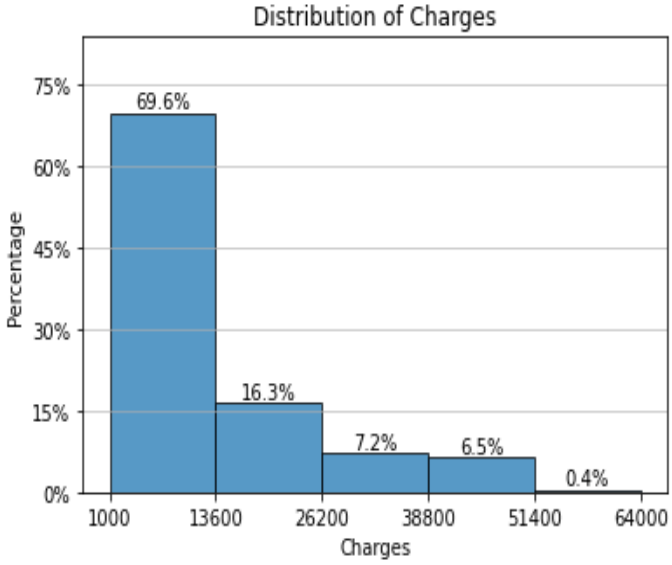


The majority of the policyholders fall within the BMI range of 25-34, which accounts for almost 60% of the population. This indicates that a significant proportion of the population is either overweight or obese, which could have health implications. BMI range of 50-54 is very low, indicating that very few policyholders in the population fall under this category.

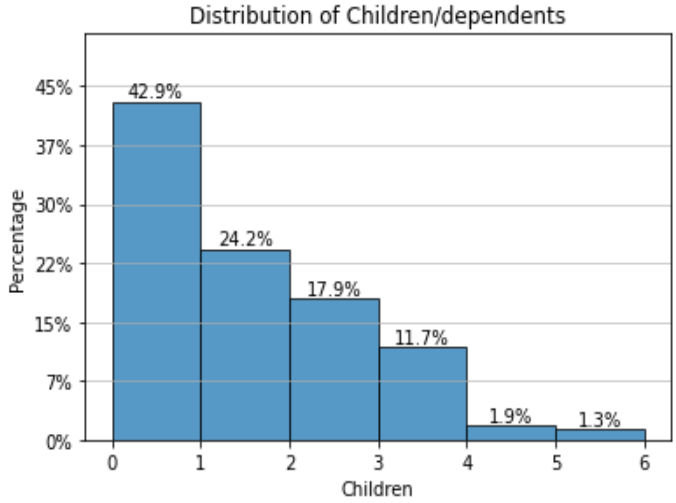
Exploratory Data Analysis



The majority of policyholders (59.3%) take an average of 3000-4409 steps per day. A small percentage of policyholders (5.9%) take an average of 4410-5819 steps per day. No policyholders take an average of 5820-7229 steps per day. Data suggests that a majority of policyholders may need to increase their daily step count to reach recommended levels of physical activity.



The majority of individual medical charges (69.6%) fall within the range of $1000-$13599. That most policyholders are likely to have lower medical expenses. The percentages decrease as we move to higher ranges, with only 0.4% of charges for the range of $51400-$64000. That very few policyholders have extremely high medical expenses. Medical charges is skewed towards lower values. However, there is still a significant portion of policyholders with higher medical expenses.



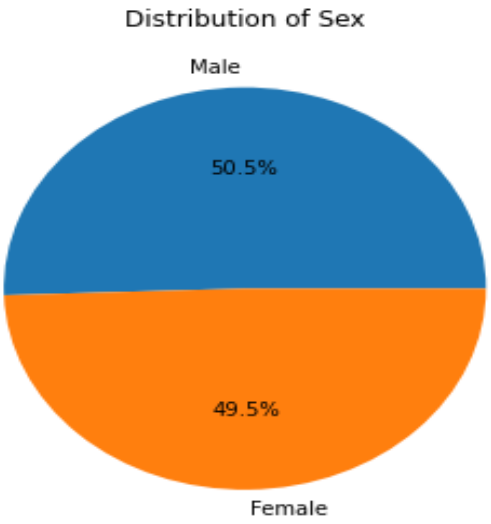
The majority of policyholders (42.9%) do not have any children or dependents. A small percentage of policyholders have four (1.9%) or five (1.3%) children or dependents. policyholders without children may require different types of coverage than those with children or dependents. Similarly, those with more children may require policies with higher coverage limits.

b) Categorical variables:

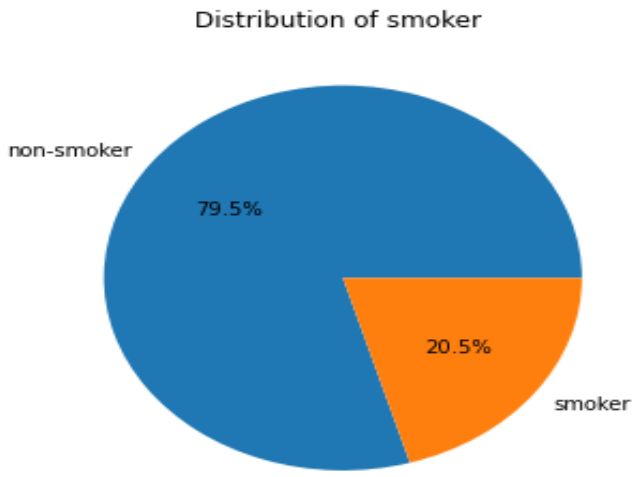
For continuous variables we can use pie chart.

A pie chart is a circular statistical graphic that is commonly used to display proportional data. The chart is divided into slices or sectors, with each slice representing a proportion or percentage of the whole. The size of each slice is proportional to the quantity it represents, and the total area of the chart is equal to 100%.

Pie charts are useful when you want to compare the relative sizes of different categories or show the distribution of a whole into its constituent parts. They are commonly used to represent market shares, demographic data, and other types of categorical data.

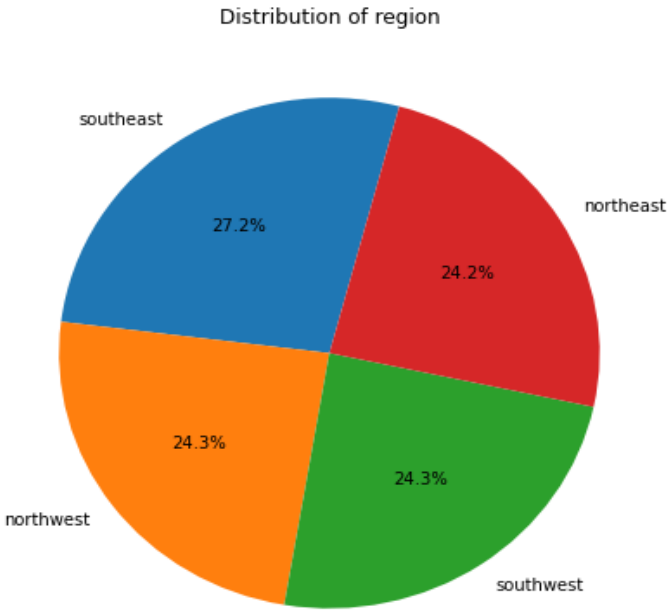


Policyholders' gender distribution reveals a nearly equal distribution between males and females. The pie chart shows that 50.5% of policyholders are male, while 49.5% are female. This information can be useful for insurance companies in developing policies that cater to the diverse needs of both genders.

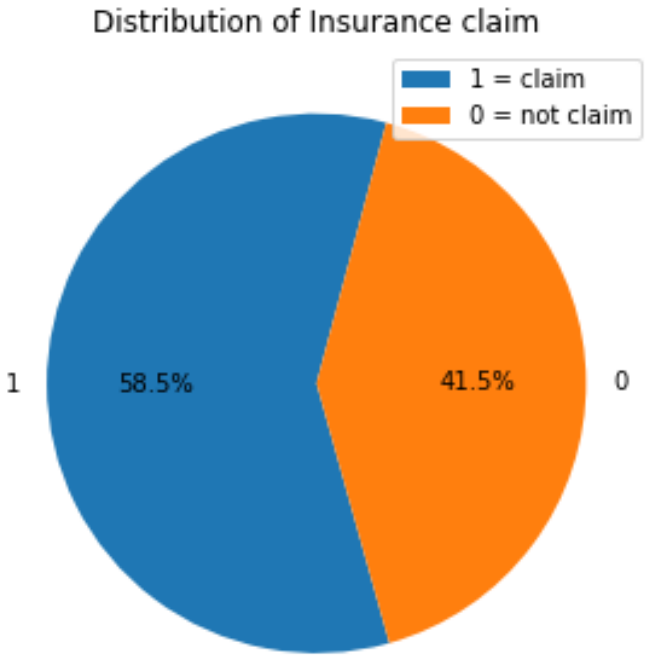


It can be inferred that the majority of policyholders are non-smokers are 79.5% of the total and 20.5% of policyholders are smokers. This data shows importance of promoting healthy habits and discouraging smoking among the population. It is essential for insurance companies to consider the smoker variable while formulating health insurance policies to provide appropriate coverage and benefits to their policyholders.

Exploratory Data Analysis



It shows that the southeast region has the highest percentage of policyholders at 27.2%, followed by the southwest, northwest and northeast regions are almost equally distributed (24.3% & 24.2%). This information can be useful for insurance companies to better understand their customer base and tailor their policies to meet the specific needs of each region. It also highlights the importance of regional demographics and socioeconomic factors in the distribution of health insurance policyholders.



It shows that 58.5% of policyholders have made an insurance claim, while 41.5% have not. This information can provide valuable insights to insurance companies to better understand their customer base and adjust their policies accordingly. It also highlights the importance of having a comprehensive health insurance policy that can provide coverage for unforeseen medical expenses. By analyzing this data, insurance companies can improve their services and ensure that policyholders receive the best possible coverage.

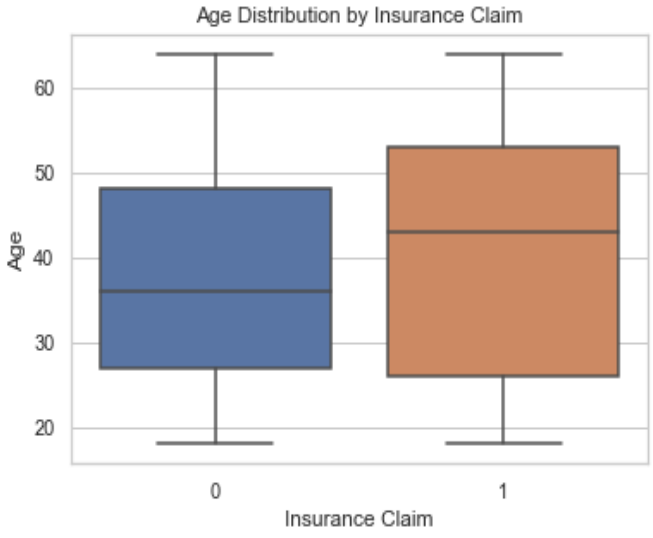
ii. Bivariate Analysis:

a) Continuous variables:

For continuous variables we can use box plot.

A box plot is a graphical representation of a dataset that displays the distribution of the data along with its quartiles, median, and outliers. It consists of a box that represents the interquartile range (IQR) of the dataset, which is the range of values that fall within the 25th to the 75th percentile of the data. The median is represented by a vertical line within the box. The whiskers extend from the edges of the box to the smallest and largest observations that are within 1.5 times the IQR from the box. Any data points outside this range are considered outliers and are plotted as individual points. Box plots are useful for quickly summarizing the distribution of a dataset and identifying any potential outliers.

Common representation for all box plots:

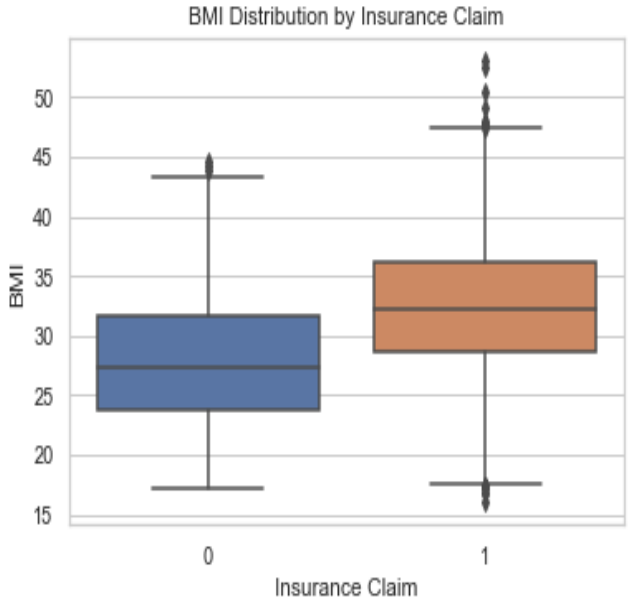


insuranceclaim	min	25%	50%	75%	max
0	18.0	27.0	36.0	48.0	64.0
1	18.0	26.0	43.0	53.0	64.0

The median age for insurance not claim is 36 years, which is lower than the median age for insurance claim, which is 43 years. The interquartile range (IQR) for insurance not claim extends from 27 years to 48 years, while the IQR for insurance claim extends from 26 years to 53 years. This indicates that the age distribution for insurance claim is more spread out than that for insurance not claim.

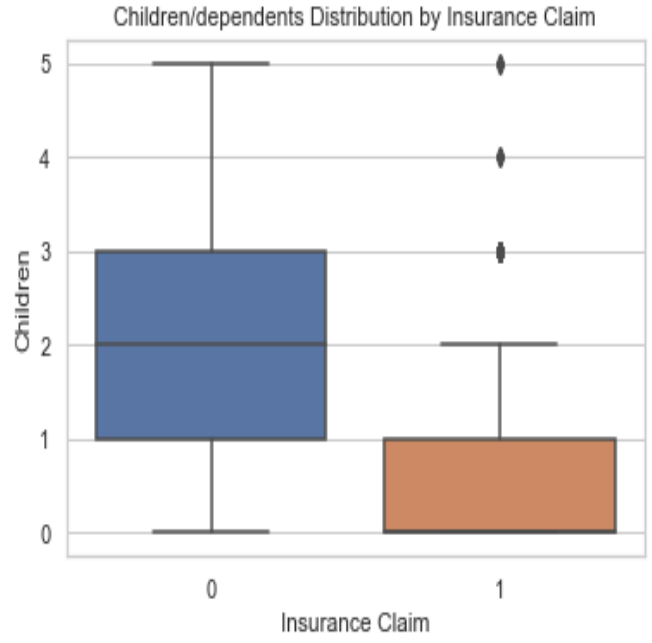
Overall, the box plot suggests that there may be a relationship between age and insurance claim, with older policyholders more likely to make an insurance claim.

Exploratory Data Analysis



insuranceclaim	min	25%	50%	75%	max
0	17.195	23.700	27.265	31.7300	44.70
1	15.960	28.695	32.300	36.1925	53.13

The median BMI is higher for those who made an insurance claim (32.3) compared to those who did not (27.265).The interquartile range (IQR) of BMI is larger for the claimants (from 28.695 to 36.1925) compared to non-claimants (from 23.7 to 31.73). There are some outliers in both groups, with the maximum BMI value being higher for the claimants. I believe that the outliers are genuine data points and not errors, I choose to leave as it is. Policyholders who make insurance claims tend to have higher BMI values and more variability in their BMI compared to those who do not make claims. This information could be useful for insurance companies in assessing the risk and pricing of their policies.

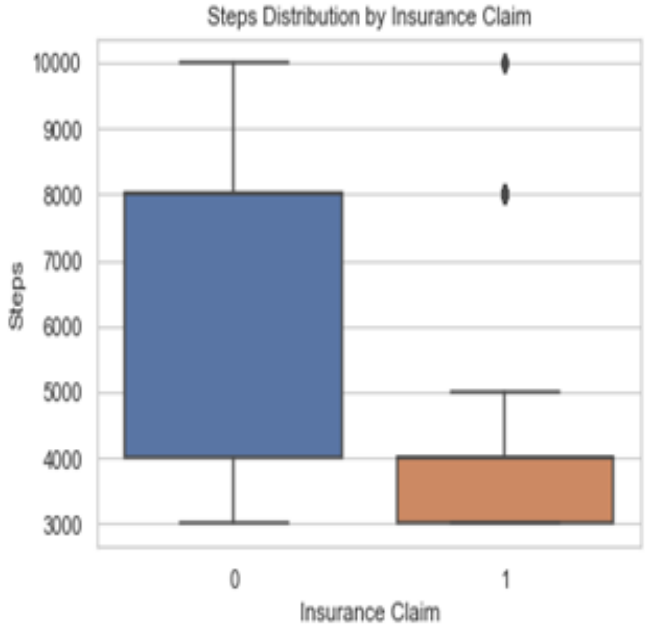


insuranceclaim	min	25%	50%	75%	max
0	0.0	1.0	2.0	3.0	5.0
1	0.0	0.0	0.0	1.0	5.0

For policyholders who not claim insurance (i.e., insuranceclaim=0), the number of children in ranges from 0 to 5, with the median number of children being 2. The interquartile range (IQR), which represents the middle 50% of the data, is from 1 to 3 children. For policyholders who claim (i.e., insuranceclaim=1), the distribution of the number of children in is skewed to the left, with the majority of policyholders having 0 or 1 child. The median number of children is 0, and the IQR is from 0 to 1 child.

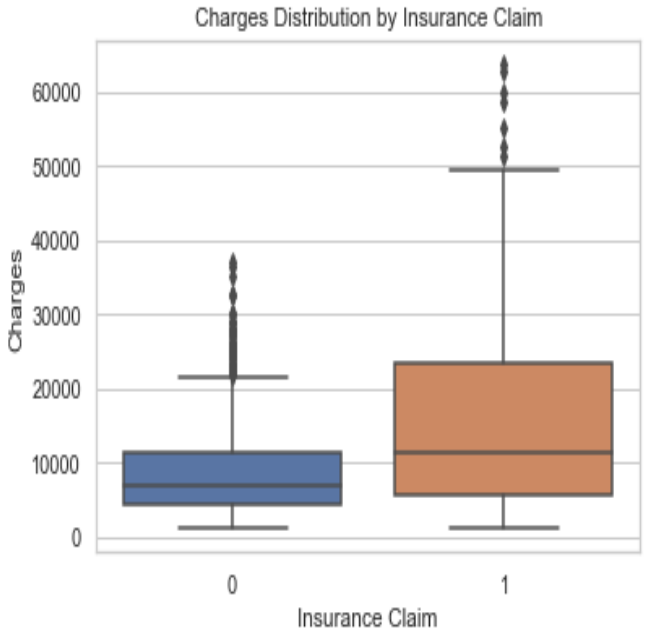
We can conclude this is a relevant variable and that there may be differences in this variable between claim and not claim.

Exploratory Data Analysis



insuranceclaim	min	25%	50%	75%	max
0	3000.0	4003.0	8002.0	8009.0	10010.0
1	3000.0	3007.0	4003.0	4010.0	10010.0

The median Steps is higher for those who made an insurance not claim compared to those who did claim The 25th percentile of group 1 (i.e., policyholders who did not claim insurance) is higher than that of group 2 (i.e., policyholders who claim insurance), indicating that the former group walks more on average. The 75th percentile of group 1 is also higher than that of group 2, indicating a wider range of values for group 1. However, both groups have similar maximum values of around 10,000. We concluded that policyholders who filed an insurance claim tend to walk less on average than those who did not file a claim.



insuranceclaim	min	25%	50%	75%	max
0	1121.87	4445.34	6933.24	11424.21	36910.61
1	1131.51	5733.29	11538.42	23484.79	63770.43

We can see that policyholders who did not make an insurance claim (represented by the blue box plot) generally had lower charges, with an IQR that ranges from about 4,400 to 11,400. Policyholders who made an insurance claim (represented by the orange box plot) generally had higher charges, with an IQR that ranges from about 5,700 to 23,500. The median charges for each group are also displayed as a horizontal line inside each box. There are some outliers in both groups, with the maximum charges value being higher for the claimants. I believe that the outliers are genuine data points and not errors, I choose to leave as it is. We conclude that policyholders who make insurance claims tend to have higher medical charges than those who do not make claims.

b) Categorical variables:

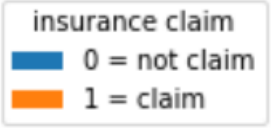
For categorical variables we can use 100% stacked bar chart and contingency table.

A 100% stacked bar chart is a type of bar chart that displays the relative proportion of different categories for two or more variables. Each bar in the chart represents a category, and the height of each segment of the bar corresponds to the proportion of that category for a particular variable. The total height of the bar is always 100%, and the segments are stacked on top of each other to show the total proportion of each variable.

These charts are often used in data visualization to show the distribution of a categorical variable across different groups or to compare the distribution of multiple categorical variables. They are particularly useful for displaying the relative frequency of different categories, highlighting patterns and trends, and showing how different variables contribute to the overall distribution.

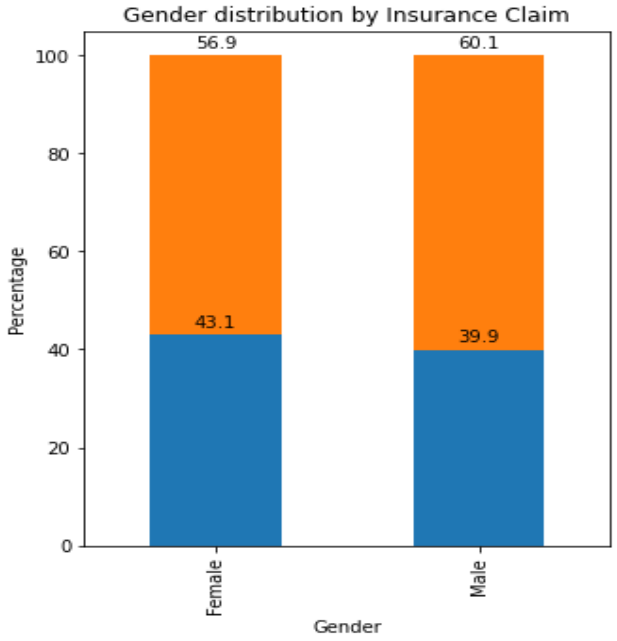
A contingency table is a table that displays the frequency or relative frequency of two or more variables in a dataset. The variables are often categorical, and the table displays the number or proportion of observations that belong to each combination of categories for the variables. Contingency tables are often used to examine the relationship between two variables and to test for independence between them. They can also be used to create 100% stacked bar charts by converting the counts in the table to proportions and plotting them as stacked bars.

Common representation for all 100% Stacked bar charts:



Exploratory Data Analysis

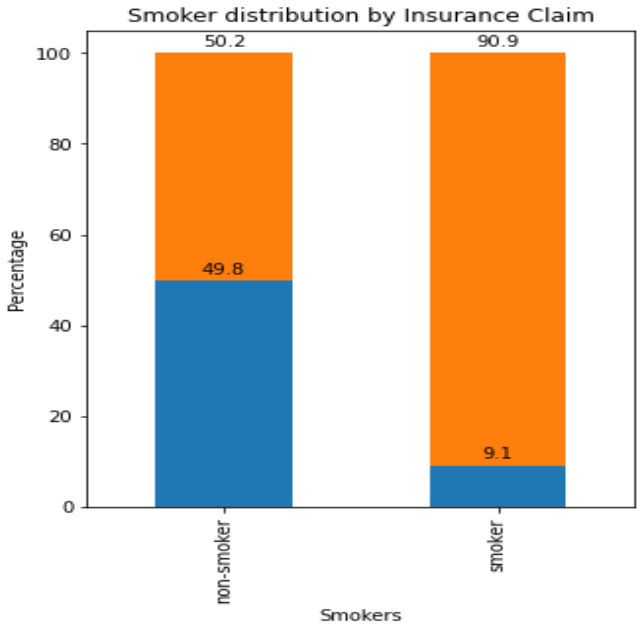
			Sex vs Insurance claim	0	1	Total	0%	1%	Total%
Female	285.00	377.00	662.00	43.05	56.95	100.00	21.30	28.18	49.48
Male	270.00	406.00	676.00	39.94	60.06	100.00	20.18	30.34	50.52
Total	555.00	783.00	1338.00	41.48	58.52	100.00	41.48	58.52	100.00



There are a total of 1338 policyholders in the dataset, out of which 555 (41.48%) did not file an insurance claim and 783 (58.52%) filed an insurance claim.

Among female policyholders, 285 (43.05%) did not file a claim, while 377 (56.95%) did file a claim. Among male policyholders, 270 (39.94%) did not file a claim, while 406 (60.06%) did file a claim. The percentage of female policyholders who filed an insurance claim (56.95%) is slightly higher than that of male policyholders (60.06%).

			Smokers vs Insurance claim	0	1	Total	0%	1%	Total%
non-smoker	530.00	534.00	1064.00	49.81	50.19	100.00	39.61	39.91	79.52
smoker	25.00	249.00	274.00	9.12	90.88	100.00	1.87	18.61	20.48
Total	555.00	783.00	1338.00	41.48	58.52	100.00	41.48	58.52	100.00

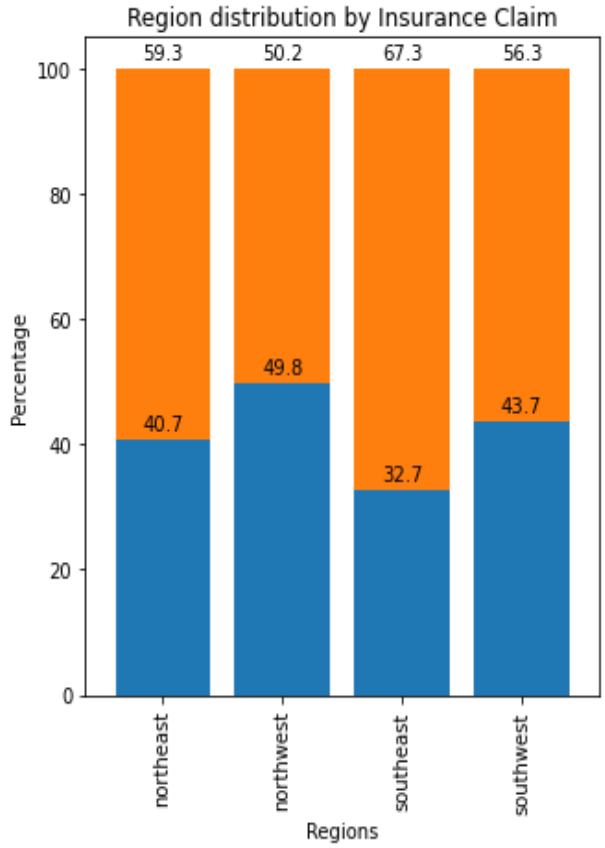


There are a total of 1338 policyholders in the dataset, out of which 555 (41.48%) are non-smokers and 783 (58.52%) are smokers.

Among non-smokers, 530 (49.81%) did not file an insurance claim, while 534 (50.19%) did file an insurance claim. Among smokers, only 25 (9.12%) did not file a claim, while 249 (90.88%) did file a claim. The percentage of smokers who filed an insurance claim (90.88%) is significantly higher than that of non-smokers (50.19%).

Exploratory Data Analysis

			Regions vs Insurance claim	0	1	Total	0%	1%	Total%
northeast	132.00	192.00	324.00	40.74	59.26	100.00	9.87	14.35	24.22
northwest	162.00	163.00	325.00	49.85	50.15	100.00	12.11	12.18	24.29
southeast	119.00	245.00	364.00	32.69	67.31	100.00	8.89	18.31	27.20
southwest	142.00	183.00	325.00	43.69	56.31	100.00	10.61	13.68	24.29
Total	555.00	783.00	1338.00	41.48	58.52	100.00	41.48	58.52	100.00



There are a total of 1338 policyholders in the dataset, out of which 555 (41.48%) did not file an insurance claim and 783 (58.52%) filed an insurance claim. Among policyholders in the northeast region, 132 (40.74%) did not file a claim, while 192 (59.26%) did file a claim. Among policyholders in the northwest region, 162 (49.85%) did not file a claim, while 163 (50.15%) did file a claim. Among policyholders in the southeast region, 119 (32.69%) did not file a claim, while 245 (67.31%) did file a claim. Among policyholders in the southwest region, 142 (43.69%) did not file a claim, while 183 (56.31%) did file a claim.

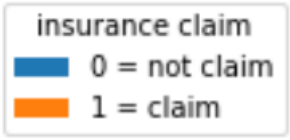
From the 100% stacked bar chart, we can also see the distribution of insurance claims across different regions. We can observe that the percentage of policyholders who filed an insurance claim is highest in the southeast region (67.31%), followed by the northeast region (59.26%), the southwest region (56.31%), and the northwest region (50.15%).

c) Contingency table for Continuous variables:

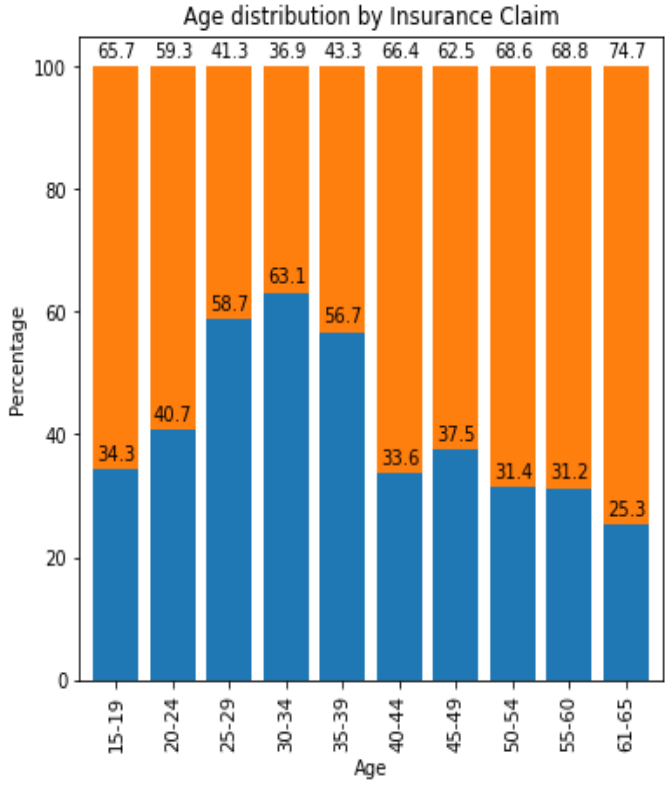
For continuous variables also we can use 100% stacked bar chart and contingency table for better understanding of distribution of data in it.

For creating contingency table and 100% stacked bar chart we need to convert the continuous variable into categorical bins of ranges of the continuous variables in order get to know about distribution of it.

Common representation for all 100% Stacked bar charts:



			Age vs Insurance claim	0	1	Total	0%	1%	Total%
15-19	57.00	109.00	166.00	34.34	65.66	100.00	4.26	8.15	12.41
20-24	57.00	83.00	140.00	40.71	59.29	100.00	4.26	6.20	10.46
25-29	81.00	57.00	138.00	58.70	41.30	100.00	6.05	4.26	10.31
30-34	82.00	48.00	130.00	63.08	36.92	100.00	6.13	3.59	9.72
35-39	72.00	55.00	127.00	56.69	43.31	100.00	5.38	4.11	9.49
40-44	46.00	91.00	137.00	33.58	66.42	100.00	3.44	6.80	10.24
45-49	54.00	90.00	144.00	37.50	62.50	100.00	4.04	6.73	10.76
50-54	44.00	96.00	140.00	31.43	68.57	100.00	3.29	7.17	10.46
55-60	39.00	86.00	125.00	31.20	68.80	100.00	2.91	6.43	9.34
61-65	23.00	68.00	91.00	25.27	74.73	100.00	1.72	5.08	6.80
Total	555.00	783.00	1338.00	41.48	58.52	100.00	41.48	58.52	100.00

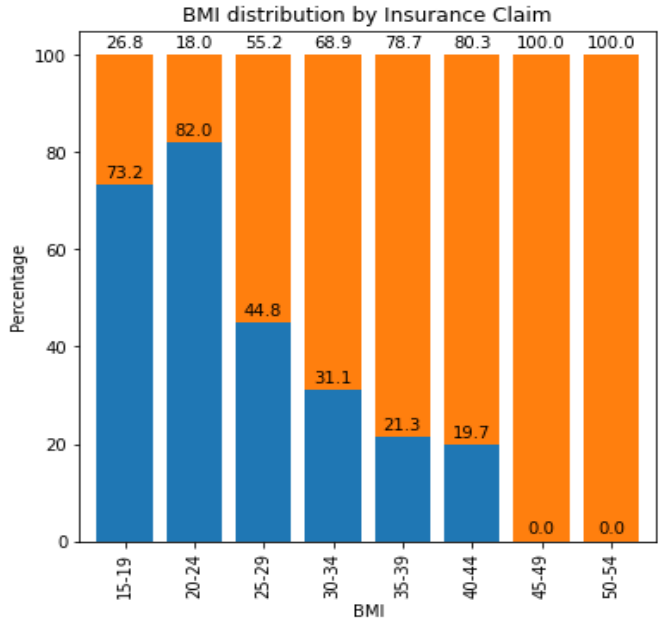


There are a total of 1338 policyholders in the dataset, out of which 555 (41.48%) did not file an insurance claim and 783 (58.52%) filed an insurance claim.

The highest percentage of policyholders who filed an insurance claim is in the age group between 50-65 years (68.6%, 68.8% and 74.7%), next highest by 40-49 years (66.4% and 62.5%), and followed by 15-24 years (65.7 and 59.3%). The lowest percentage of policyholders who filed an insurance claim is in the age group of 26-39 years (41.3%, 36.9% and 43.3%).

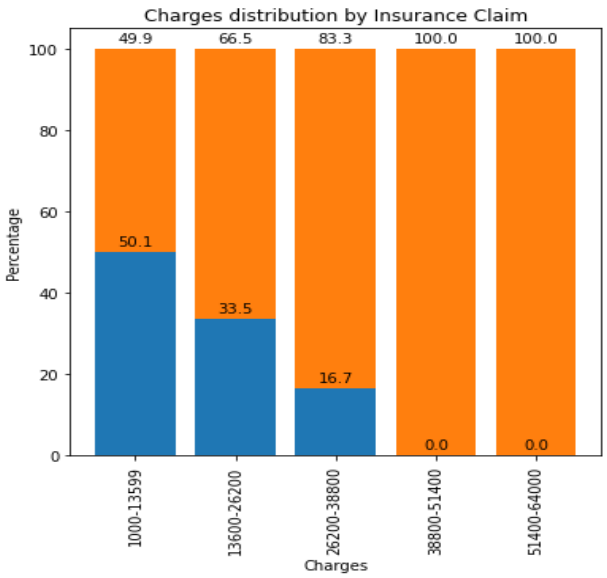
Exploratory Data Analysis

			BMI vs Insurance claim	0	1	Total	0%	1%	Total%
15-19	30.00	11.00	41.00	73.17	26.83	100.00	2.24	0.82	3.06
20-24	169.00	37.00	206.00	82.04	17.96	100.00	12.63	2.77	15.40
25-29	173.00	213.00	386.00	44.82	55.18	100.00	12.93	15.92	28.85
30-34	121.00	268.00	389.00	31.11	68.89	100.00	9.04	20.03	29.07
35-39	48.00	177.00	225.00	21.33	78.67	100.00	3.59	13.23	16.82
40-44	14.00	57.00	71.00	19.72	80.28	100.00	1.05	4.26	5.31
45-49	0.00	17.00	17.00	0.00	100.00	100.00	0.00	1.27	1.27
50-54	0.00	3.00	3.00	0.00	100.00	100.00	0.00	0.22	0.22
Total	555.00	783.00	1338.00	41.48	58.52	100.00	41.48	58.52	100.00



The highest percentage of policyholders who filed an insurance claim is in the BMI group between 35-55 (78.7%, 80.3% and 100%), next highest by 25-34 (55.2% and 68.9%). The lowest percentage of policyholders who filed an insurance claim is in the BMI group of 15-24 (26.8% and 18%).Therefore, the policyholders who have BMI greater than or equal to 30 are more likely to file an insurance claim compared to those with BMI less than 30.

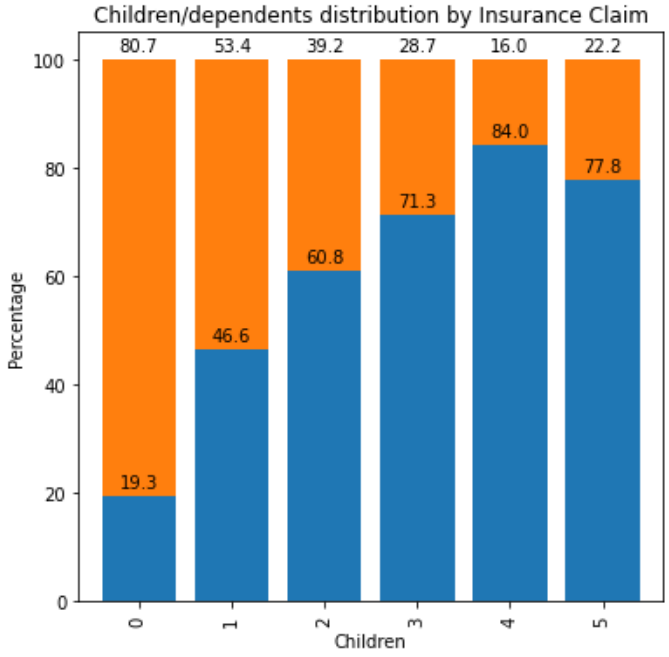
			Charges vs Insurance claim	0	1	Total	0%	1%	Total%
1000-13599	466.00	465.00	931.00	50.05	49.95	100.00	34.83	34.75	69.58
13600-26200	73.00	145.00	218.00	33.49	66.51	100.00	5.46	10.84	16.29
26200-38800	16.00	80.00	96.00	16.67	83.33	100.00	1.20	5.98	7.17
38800-51400	0.00	87.00	87.00	0.00	100.00	100.00	0.00	6.50	6.50
51400-64000	0.00	6.00	6.00	0.00	100.00	100.00	0.00	0.45	0.45
Total	555.00	783.00	1338.00	41.48	58.52	100.00	41.48	58.52	100.00



The highest percentage of policyholders who filed an insurance claim is in the charges group between 26200-64000 medical cost (83.3% and 100%), next highest by 13600-26200 medical cost (66.5%). The lowest percentage of policyholders who filed an insurance claim is in the charge group of 1000-13600 medical cost (49.9%). Therefore, the policyholders who have medical cost greater than or equal to 26200 are more likely to file an insurance claim compared to those with medical cost less than 26200.

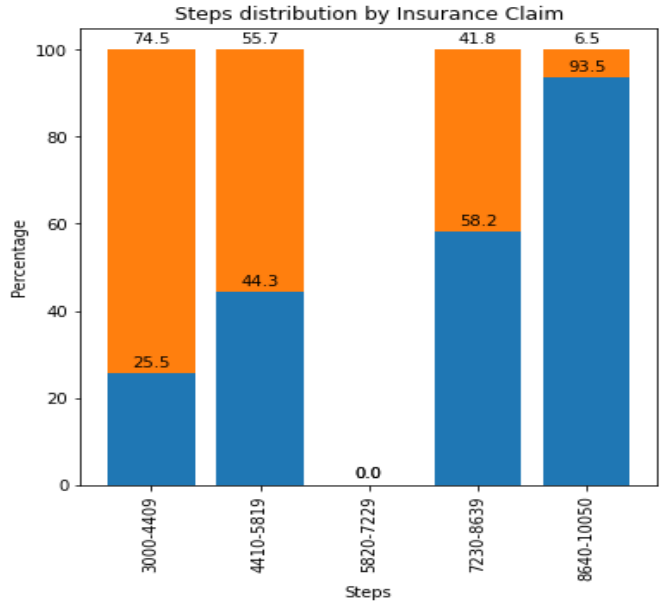
Exploratory Data Analysis

			Children/dependents vs Insurance claim	0	1	Total	0%	1%	Total%
0	111.00	463.00	574.00	19.34	80.66	100.00	8.30	34.60	42.90
1	151.00	173.00	324.00	46.60	53.40	100.00	11.29	12.93	24.22
2	146.00	94.00	240.00	60.83	39.17	100.00	10.91	7.03	17.94
3	112.00	45.00	157.00	71.34	28.66	100.00	8.37	3.36	11.73
4	21.00	4.00	25.00	84.00	16.00	100.00	1.57	0.30	1.87
5	14.00	4.00	18.00	77.78	22.22	100.00	1.05	0.30	1.35
Total	555.00	783.00	1338.00	41.48	58.52	100.00	41.48	58.52	100.00



The highest percentage of policyholders who filed an insurance claim is who policyholders have no and one child (80.7% and 53.4%), next highest by who have 2 and 3 children (39.2% and 28.7%). The lowest percentage of policyholders who filed an insurance claim is who have 4 and 5 children (16% and 22.2%). Therefore, The percentage of policyholders who filed an insurance claim decreases as the number of children/dependents increases.

			Steps vs Insurance claim	0	1	Total	0%	1%	Total%
3000-4409	192.00	560.00	752.00	25.53	74.47	100.00	14.81	43.21	58.02
4410-5819	35.00	44.00	79.00	44.30	55.70	100.00	2.70	3.40	6.10
7230-8639	199.00	143.00	342.00	58.19	41.81	100.00	15.35	11.03	26.39
8640-10050	115.00	8.00	123.00	93.50	6.50	100.00	8.87	0.62	9.49
Total	541.00	755.00	1296.00	41.74	58.26	100.00	41.74	58.26	100.00



The highest percentage of policyholders who filed an insurance claim is in the steps group between 3000-4410 average steps per day (74.5%), next highest by 4410-8640 average steps per day (55.7% and 41.8%). The lowest percentage of policyholders who filed an insurance claim is in the steps group of 8640 -10050 (6.5%).

Therefore, The percentage of policyholders who filed an insurance claim decreases as the number of average steps per day increases.

iii. Hypothesis testing:

Hypothesis testing is a statistical method used to determine whether a hypothesis about a population parameter is supported by the data. In hypothesis testing, we start with a null hypothesis, which represents the status quo or the conventional wisdom, and an alternative hypothesis, which represents a new or competing hypothesis.

The null hypothesis typically assumes that there is no significant difference or relationship between the variables of interest, while the alternative hypothesis assumes that there is a significant difference or relationship. Hypothesis testing involves calculating a test statistic based on the sample data and comparing it to a critical value or p-value, which represents the probability of observing a test statistic as extreme as the one observed, assuming the null hypothesis is true.

If the test statistic falls in the rejection region, which is defined by the critical value or p-value, we reject the null hypothesis and accept the alternative hypothesis, which means that there is evidence to support the new hypothesis. If the test statistic falls in the acceptance region, we fail to reject the null hypothesis, which means that there is insufficient evidence to support the new hypothesis.

a) Continuous variables:

Welch's t-test can be used when one variable is continuous (independent) and another variable is binary categorical (dependent).

Welch's t-test is a statistical hypothesis test used to compare the means of two independent groups. It is similar to the standard two-sample t-test, but it does not assume that the variances of the two groups are equal.

- **Test for "age" and "insurance claim":**
  Null hypothesis: There is no significant difference in the mean age between those who made an insurance claim and those who did not.

  Alternative hypothesis: There is a significant difference in the mean age between those who made an insurance claim and those who did not.

```
Welch's t-test to compare the means of a variable Age between two groups defined
by  variable insuranceclaim results:
t-value = -4.287049792327385
p-value = 1.9460516147410528e-05
There is significant evidence to reject the null hypothesis.
```

- **Test for "bmi" and "insurance claim":**
  Null hypothesis: There is no significant difference in the mean BMI between those who made an insurance claim and those who did not.

  Alternative hypothesis: There is a significant difference in the mean BMI between those who made an insurance claim and those who did not.

```
Welch's t-test to compare the means of a variable BMI between two groups defined
by  variable insuranceclaim results:
t-value: -15.242972959368466
p-value: 4.24712334564955e-48
There is significant evidence to reject the null hypothesis.
```

- **Test for "steps" and "insurance claim":**
  Null hypothesis: There is no significant difference in the mean steps between those who made an insurance claim and those who did not.

  Alternative hypothesis: There is a significant difference in the mean steps between those who made an insurance claim and those who did not.

```
Welch's t-test to compare the means of a variable steps between two groups
defined by  variable insuranceclaim results:
t-value: 15.971419155487517
p-value: 5.737104472076578e-51
There is significant evidence to reject the null hypothesis.
```

- **Test for "children" and "insurance claim":**
  Null hypothesis: There is no significant difference in the mean children between those who made an insurance claim and those who did not.

  Alternative hypothesis: There is a significant difference in the mean children between those who made an insurance claim and those who did not.

```
Welch's t-test to compare the means of a variable children between two groups
defined by  variable insuranceclaim results:
t-value: 15.769758859674198
p-value: 2.9311128338998403e-50
There is significant evidence to reject the null hypothesis.
```

- **Test for "charges" and "insurance claim":**
  Null hypothesis: There is no significant difference in the mean charges between those who made an insurance claim and those who did not.

  Alternative hypothesis: There is a significant difference in the mean charges between those who made an insurance claim and those who did not.

```
Welch's t-test to compare the means of a variable charges between two groups
defined by  variable insuranceclaim results:
t-value: -13.298031957975649
p-value: 1.1105103216309125e-37
There is significant evidence to reject the null hypothesis.
```

b) Categorical variable:
Chi-square test can be used to test the association between two independent categorical variables.

The chi-square test is a statistical test used to determine if there is a significant association or relationship between two categorical variables.

The test involves calculating the difference between the observed frequencies and the expected frequencies. The expected frequencies are the frequencies that would be expected under the assumption that there is no association between the two variables being tested.

The test calculates a test statistic (chi-square statistic) which follows a chi-square distribution. The degrees of freedom for the test depend on the number of categories in the variables being tested.

If the calculated chi-square statistic is larger than the critical value for the given degrees of freedom and level of significance, then we reject the null hypothesis and conclude that there is a significant association between the two variables.

- **Test for "sex" and "insurance claim":**
  Null hypothesis: There is no association between the two variables (i.e., the proportions of insurance claims are the same for males and females).

  Alternative hypothesis: There is an association between the two variables.

```
Chi-square test to compare significant association or relationship between sex
and insuranceclaim results:
Chi-squared statistic: 1.2080752607293883
p-value: 0.2717136468543795
There is not enough evidence to reject the null hypothesis.
```

- **Test for "smoker" and "insurance claim":**
  Null hypothesis: There is no association between the two variables (i.e., the proportions of insurance claims are the same for smokers and non-smokers).

  Alternative hypothesis: There is an association between the two variables.

```
Chi-square test to compare significant association or relationship between
smoker and insuranceclaim results:
Chi-squared statistic: 146.93066085010693
p-value: 8.126230429795898e-34
There is significant evidence to reject the null hypothesis.
```

- **Test for "region" and "insurance claim":**
  Null hypothesis: There is no association between the two variables (i.e., the proportions of insurance claims are the same for all four region).

  Alternative hypothesis: There is an association between the two variables.

```
Chi-square test to compare significant association or relationship between
region and insuranceclaim results:
Chi-squared statistic: 21.67937470273104
p-value: 7.60585329506179e-05
There is significant evidence to reject the null hypothesis.
```

SPLITTING OF DATA

Splitting a dataset is the process of dividing it into two or more parts, typically for training and testing a machine learning model. The goal of splitting a dataset is to use one part of the data to train a model and another part to evaluate the performance of the model.

There are different ways to split a dataset depending on the task at hand. The most common ways are:

Train-Test Split, K-Fold Cross-Validation and Stratified Sampling.

The choice of splitting approach depends on the size of the dataset, the complexity of the model, and the available computing resources. It's essential to split the dataset carefully to avoid overfitting or underfitting the model, which can result in poor performance on new data.

In order to build a model we have to do splitting of dataset. There are different ways to split a dataset the method I choose to use is Train-Test Split.

The train-test split is a technique used to evaluate the performance of a model on an independent dataset. The basic idea is to split the available data into two subsets: one for training the model, and the other for testing its performance.

The training set is used to fit the model and determine the model parameters, while the test set is used to evaluate the model's performance on new, unseen data. This helps to ensure that the model is able to generalize well to new data, rather than just memorizing the training data.

Splitting of data

The train-test split is typically done by randomly selecting a portion of the data to be used for testing, while the remaining data is used for training. The proportion of data allocated to each set can vary depending on the size of the dataset and the complexity of the model being trained. A common practice is to use a 80/20 or 70/30 split for training and testing, respectively.

Once the data has been split into training and testing sets, the model is trained on the training set and its performance is evaluated on the test set. The evaluation metric used will depend on the specific problem being solved and the nature of the data. Common metrics include accuracy, precision, recall, F1 score, and mean squared error.

I have split the data set as 75% training data and 25% testing data.

```
In [193]: from sklearn.model_selection import train_test_split

In [194]: X = data.drop("insuranceclaim", axis=1)
     ...: y = data["insuranceclaim"]

In [195]: X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.25, random_state=42)
```

Input 193 is the necessary library to train – test split.

Input 194 is I am keeping all the independent variables and dropping dependent variable (insuranceclaim) in X and keeping only the dependent variable (insuranceclaim) in y.

Input 195 is the function to split the training and testing data according to X and y. The random_state is the setting the seed of the data.

```
Training set size: 1003 (74.96% of data)
Test set size: 335 (25.04% of data)
```

This the proportion of training and testing data distributed in percentage too.

MODEL BUILDING

Model building is the process of creating a mathematical or computational model that can be used to represent a real-world system or phenomenon. This process typically involves selecting the appropriate mathematical equations or algorithms, defining the model parameters, and specifying the boundary conditions.

Model building is a critical step in many scientific and engineering disciplines, including physics, chemistry, biology, economics, and finance. The resulting model can be used to simulate, predict, or analyze the behavior of the system under different conditions, and can help researchers to gain insights into the underlying mechanisms that govern its behavior.

The process of model building may involve several stages, including:

Selecting the appropriate model type: Selecting the appropriate mathematical or computational model that best represents the system or phenomenon being studied.

Defining the model parameters: Identifying and defining the parameters of the model, which may include physical constants, initial conditions, and boundary conditions.

Validating the model: Comparing the predictions of the model with experimental or observed data, and refining the model as necessary.

Communicating results: Presenting the results of the model development process to stakeholders, including technical and non-technical audiences, in a clear and accessible way.

Model building

Overall, model building is a complex and iterative process that requires careful attention to detail, a deep understanding of the system or phenomenon being studied, and the ability to balance simplicity and complexity to develop a model that is both accurate and useful.

For binary categorical dependent variables (i.e. dependent variables that can take on only two possible values), there are several types of models that can be used, depending on the nature of the data and the research question. The models I choose to do is **Logistic regression, Decision trees** and **Random forest.**

Sensitivity, specificity, positive predictive value, negative predictive value, accuracy, and precision are commonly used measures to evaluate the performance of classification models. Here's a brief description of each measure:

Sensitivity: Sensitivity measures the proportion of true positive cases that are correctly identified by the model. It is calculated as the ratio of true positives to the sum of true positives and false negatives. A high sensitivity indicates that the model is good at identifying positive cases.

Specificity: Specificity measures the proportion of true negative cases that are correctly identified by the model. It is calculated as the ratio of true negatives to the sum of true negatives and false positives. A high specificity indicates that the model is good at identifying negative cases.

Negative Predictive Value (NPV): NPV measures the proportion of true negative cases among all cases predicted as negative by the model. It is calculated as the ratio of true negatives to the sum of true negatives and false negatives. A high NPV indicates that the model is good at correctly identifying negative cases.

Model building

Accuracy: Accuracy measures the proportion of correct predictions made by the model. It is calculated as the ratio of the sum of true positives and true negatives to the total number of cases. A high accuracy indicates that the model is good at making correct predictions.

Precision or Positive Predictive Value (PPV): Precision measures the proportion of true positive cases among all cases predicted as positive by the model. It is calculated as the ratio of true positives to the sum of true positives and false positives. A high precision indicates that the model is good at making precise predictions of positive cases.

i. Logistic Regression:

Logistic regression is a statistical model used for analyzing the relationship between a binary categorical dependent variable and one or more predictor variables. It is widely used in various fields such as social sciences, medical research, and engineering.

The logistic regression model uses a logistic function to model the probability of the dependent variable taking one of two possible values (e.g., success/failure, yes/no, true/false) based on the values of the predictor variables. The logistic function produces an S-shaped curve that ranges from 0 to 1, which represents the probability of the dependent variable being in one of the two categories.

The logistic regression model estimates the coefficients of the predictor variables to determine their effects on the probability of the dependent variable taking on a particular value. These coefficients represent the change in the log-odds (i.e., logit) of the dependent variable for a unit change in the predictor variable, holding other variables constant.

Model building

The logistic regression model can handle both continuous and categorical predictor variables and can be extended to include interactions between predictor variables. The model can also be used to make predictions on new data and evaluate the goodness of fit of the model.

Overall, logistic regression is a powerful and flexible tool for modeling the relationship between a binary categorical dependent variable and one or more predictor variables, and can provide valuable insights into the factors that influence the outcome of interest.

AIC: AIC stands for Akaike Information Criterion, which is a measure of the quality of a statistical model. In logistic regression, AIC is a measure of how well a given model fits the data.

AIC is calculated as follows:

$AIC = -2\log(L) + 2k$

Where:

L is the maximum value of the likelihood function of the logistic regression model.

k is the number of parameters estimated in the model.

The AIC penalizes models for having more parameters, which can help prevent overfitting of the model. The lower the AIC value, the better the model fits the data.

AIC is commonly used to compare the fit of different models with different predictor variables or different levels of complexity. The model with the lowest AIC value is generally considered to be the best model for the given data.

Model building

Odds ratio: In logistic regression, the odds ratio is a measure of the strength of association between a predictor variable and the outcome variable.

The odds ratio represents the odds that an outcome will occur given the presence of a specific predictor, compared to the odds of the outcome occurring in the absence of that predictor, assuming all other predictors remain constant. The odds ratio is calculated from the exponential function of the coefficient estimate based on a unit increase in the predictor.

In logistic regression, the odds ratio is typically used to quantify the effect of a predictor variable on the probability of the outcome variable. An OR greater than 1 indicates that the predictor variable is associated with an increased odds of the outcome, while an OR less than 1 indicates that the predictor variable is associated with a decreased odds of the outcome. An OR of 1 indicates no association between the predictor variable and the outcome.

Model:

Logit Regression Results

Dep. Variable:	insuranceclaim	No. Observations:	1003
Model:	Logit	Df Residuals:	992
Method:	MLE	Df Model:	10
Date:	Thu, 30 Mar 2023	Pseudo R-squ.:	0.4676
Time:	19:43:46	Log-Likelihood:	-363.59
converged:	True	LL-Null:	-682.89
Covariance Type:	nonrobust	LLR p-value:	9.430e-131

	coef	std err	z	P>|z|	[0.025	0.975]	Odds Ratio
const	-9.7717	1.379	-7.086	0.000	-12.475	-7.069	0.000057
age	0.0359	0.009	4.023	0.000	0.018	0.053	1.036559
sex	0.0431	0.186	0.231	0.817	-0.322	0.408	1.044008
bmi	0.3053	0.033	9.196	0.000	0.240	0.370	1.357002
steps	8.275e-05	5.94e-05	1.393	0.164	-3.37e-05	0.000	1.000083
children	-1.5165	0.118	-12.900	0.000	-1.747	-1.286	0.219471
smoker	4.6853	0.534	8.768	0.000	3.638	5.733	108.347731
charges	-1.157e-06	1.79e-05	-0.064	0.949	-3.63e-05	3.4e-05	0.999999
region_northeast	0.4607	0.273	1.688	0.091	-0.074	0.996	1.585201
region_northwest	-0.1052	0.264	-0.398	0.691	-0.623	0.413	0.900150
region_southwest	0.0121	0.268	0.045	0.964	-0.512	0.537	1.012128

AIC: 749.189233073881

Model building

Metric	Value
Sensitivity	0.881773
Specificity	0.833333
Positive Predictive Value	0.890547
Negative Predictive Value	0.820896
Accuracy	0.862687
Precision	0.890547

ii. Decision Trees:

A decision tree is a type of supervised machine learning model that is used for both classification and regression tasks. It is a tree-structured model that makes decisions by splitting the data into smaller and smaller subsets based on a set of decision rules or criteria.

The decision tree consists of nodes and edges. The nodes represent the decision points and the edges represent the possible outcomes. At each node, a decision is made based on a specific feature of the data. The data is split into subsets based on the outcome of that decision, and the process is repeated for each subsequent node until a final decision is made.

The decision rules used to split the data are based on a measure of purity or impurity. In classification tasks, commonly used measures of impurity are Gini impurity and entropy, which quantify the homogeneity of the classes in each subset. The goal is to create splits that result in subsets with the greatest possible homogeneity in terms of class labels. In regression tasks, the mean squared error (MSE) or mean absolute error (MAE) are often used as measures of impurity.

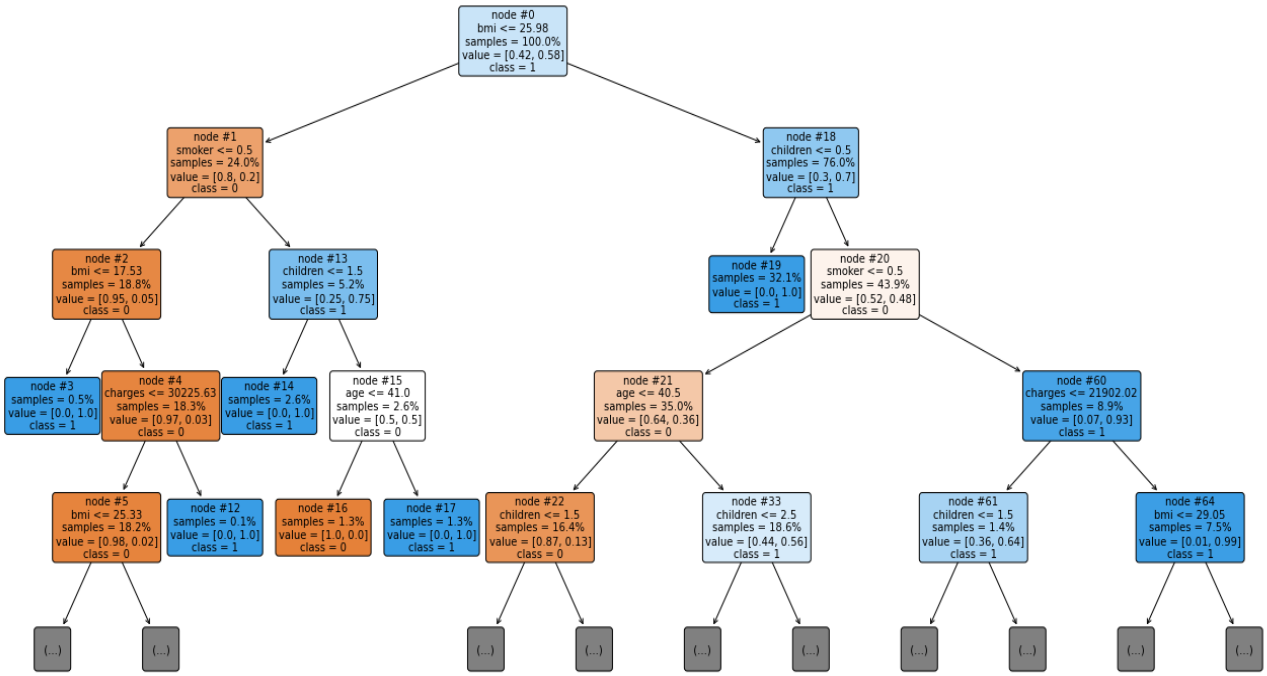
Model building

Once the decision tree is trained on a training dataset, it can be used to make predictions on new data. The tree takes in the features of a new data point and follows the decision rules to classify or predict the target variable.

One advantage of decision trees is that they are easy to interpret and visualize, making them useful for explaining how the model arrived at a particular prediction. However, decision trees can be prone to overfitting if the tree is too complex or if the decision rules are too specific to the training dataset, which can lead to poor performance on new data. To prevent overfitting, techniques such as pruning, limiting the depth of the tree, or using ensemble methods like random forests or gradient boosting can be used.

Model:

Metric	Training	Testing
Tree size	71	71
Accuracy	1	0.985075
Precision	1	0.982685
Recall	1	0.98636
ROC AUC score	1	0.98636



Model building

Metric	Value
Sensitivity	0.980296
Specificity	0.992424
Positive Predictive Value	0.995
Negative Predictive Value	0.97037
Accuracy	0.985075
Precision	0.995

iii. Random Forest:

Random Forest is a popular ensemble learning algorithm that combines multiple decision trees to improve the accuracy and stability of predictions. The basic idea behind Random Forest is to build a large number of decision trees (hence the term "forest"), each using a random subset of the features and a random subset of the training data. The individual trees are trained using a process called bagging (bootstrap aggregating), where each tree is trained on a bootstrapped sample of the training data (i.e., a random sample with replacement).

During the prediction phase, each tree in the forest predicts the class of the input data, and the final prediction is made by aggregating the predictions of all the trees, either through majority voting (for classification problems) or averaging (for regression problems). This aggregation of predictions helps to reduce the variance and overfitting that can occur in individual decision trees.

Model building

Random Forest has several advantages over single decision trees:

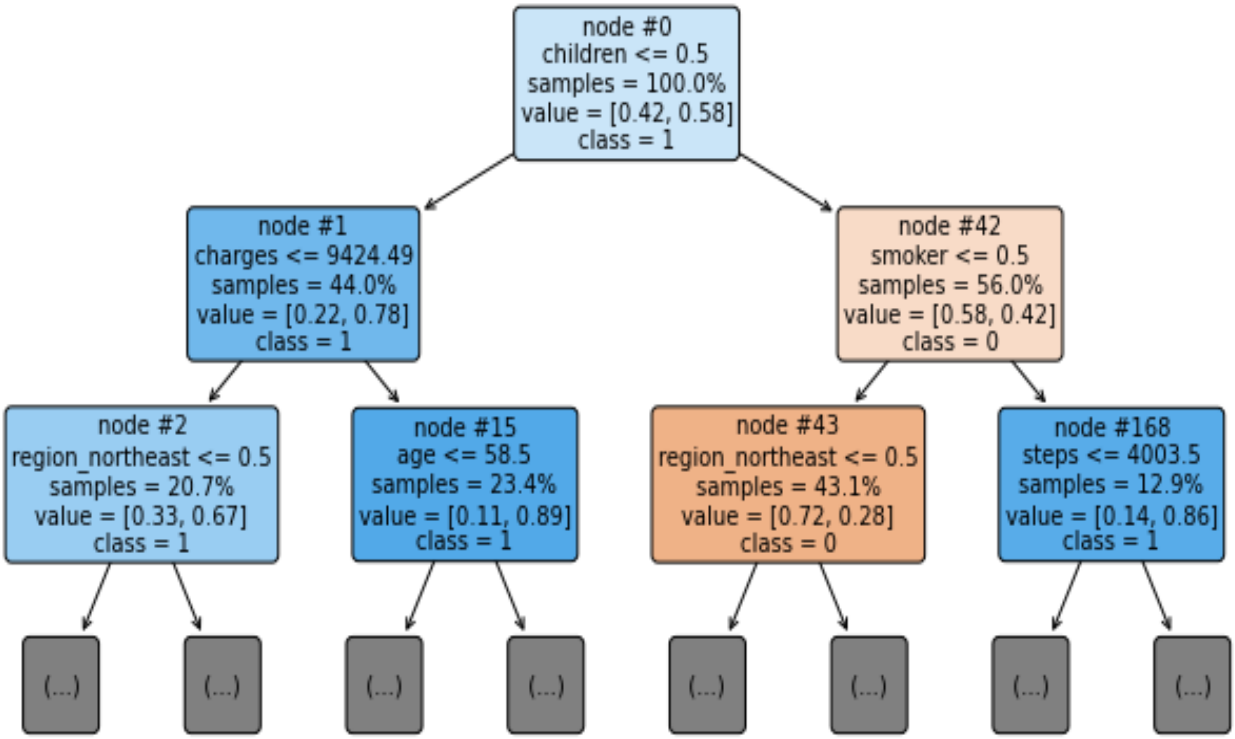
1. It can handle both classification and regression problems.
2. It can handle high-dimensional data with many features.
3. It is less prone to overfitting than single decision trees.
4. It can provide estimates of feature importance, which can be used for feature selection and interpretation.

Random Forest is widely used in machine learning applications due to its high accuracy, robustness, and scalability. However, it may require more computational resources than single decision trees, especially when dealing with large datasets or a large number of trees.

Model:

Decision tree1:

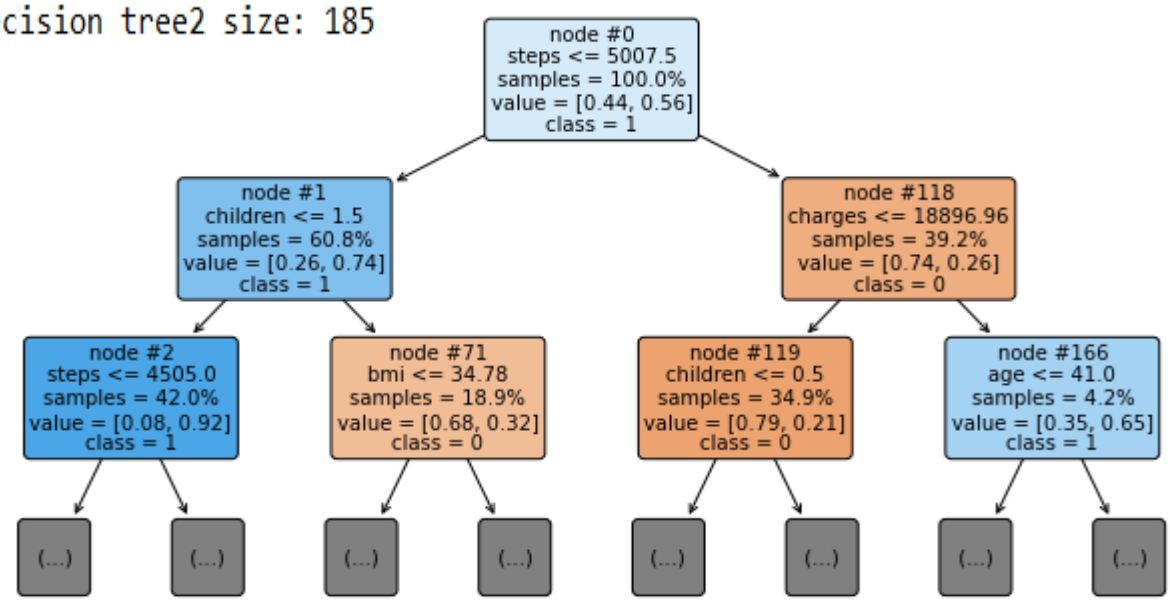
```
Decision tree1 size: 185
```



Model building

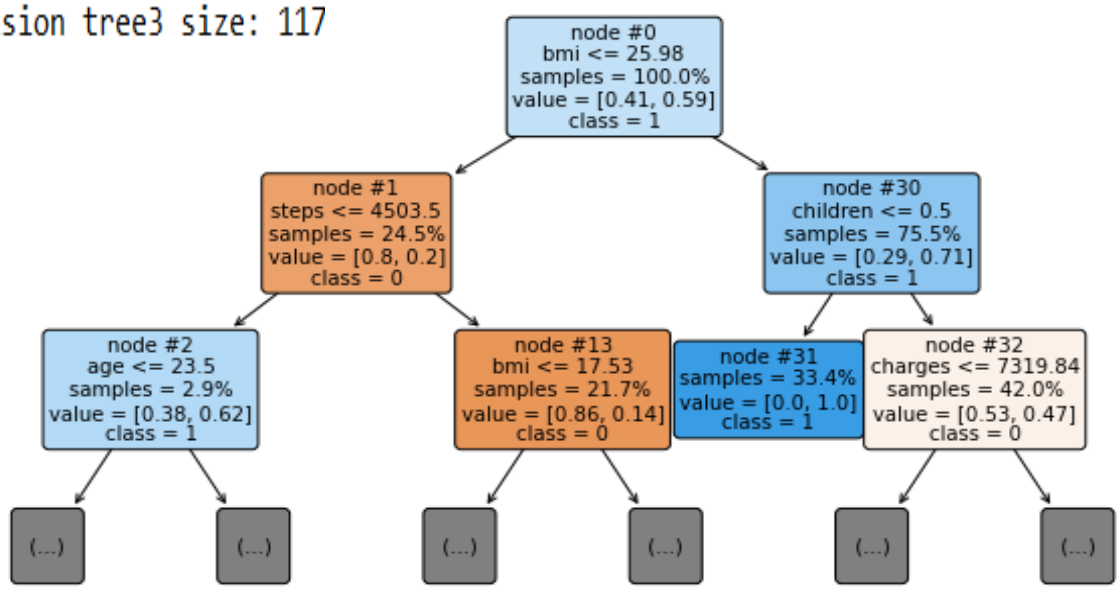
Decision tree2:

Decision tree2 size: 185



Decision tree3:

Decision tree3 size: 117



Metric	Value
Sensitivity	0.940887
Specificity	0.969697
Positive Predictive Value	0.979487
Negative Predictive Value	0.914286
Accuracy	0.952239
Precision	0.979487

Model building

Confusion Matrix:

A confusion matrix is a table that is used to evaluate the performance of a classification model by comparing the actual and predicted class labels of a set of data points. It is a tool to help visualize the number of true positive, true negative, false positive, and false negative predictions made by a classification model.

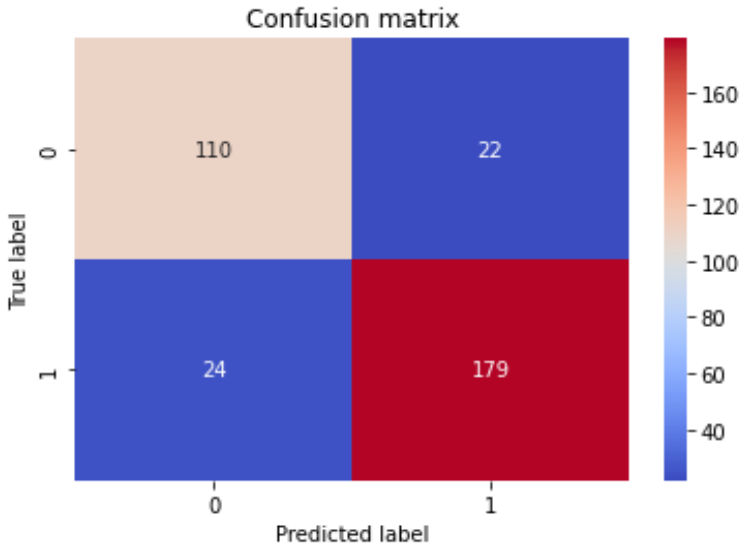
Here's an example of a confusion matrix:

	Predicted Positive	Predicted Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

The rows represent the actual class labels of the data points, while the columns represent the predicted class labels. The true positives (TP) are the number of data points that were correctly classified as positive by the model, while the true negatives (TN) are the number of data points that were correctly classified as negative. False positives (FP) are the number of data points that were incorrectly classified as positive, while false negatives (FN) are the number of data points that were incorrectly classified as negative.

Now I have compared the confusion matrices of all three models Logistic Regression, Decision Tree and Random Forest visually below:

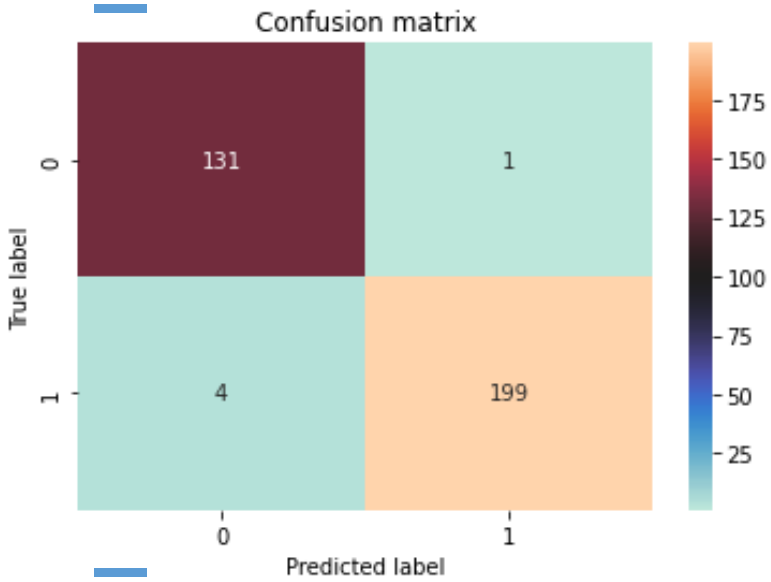
Model building



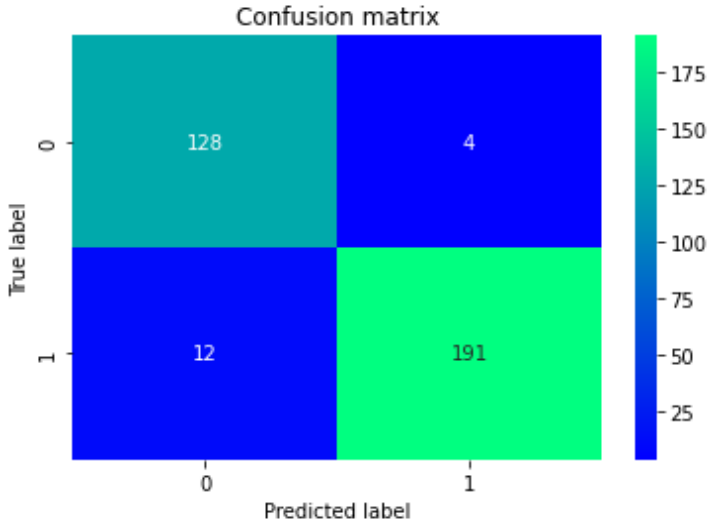
LOGISTIC REGRSSION

VS

DECISION TREE



VS



RANDOM FOREST

Model building

ROC Curve:

A receiver operating characteristic (ROC) curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. It is a plot of the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings.

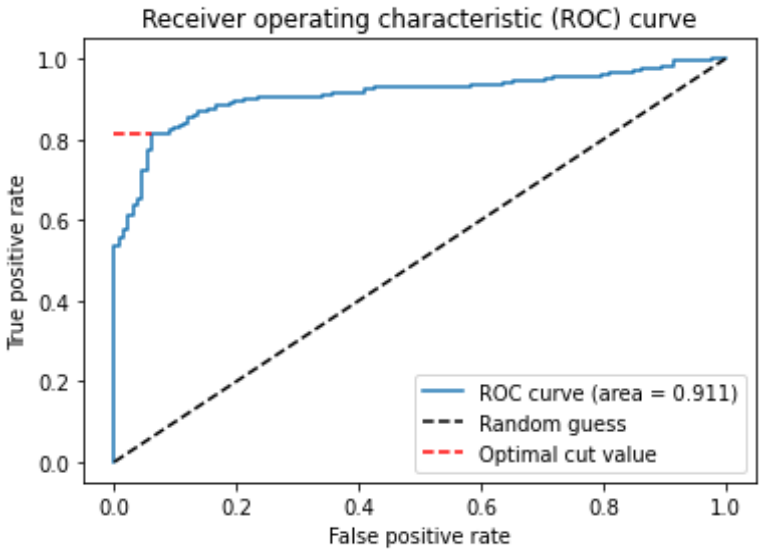
To create a ROC curve, the model's predicted probabilities for the positive class are sorted in descending order, and a threshold is set to convert the predicted probabilities into binary class predictions. As the threshold varies, the true positive rate and false positive rate are calculated for each threshold value, resulting in a set of points that can be plotted on an ROC curve.

The ROC curve is useful for evaluating the performance of a binary classification model, especially when the classes are imbalanced or when the cost of false positives and false negatives is different. The area under the ROC curve (AUC) is often used as a summary statistic to measure the overall performance of the model. An AUC of 0.5 indicates that the model performs no better than random guessing, while an AUC of 1.0 indicates perfect classification performance.

In general, a good ROC curve is one that lies closer to the top left corner of the plot, indicating high true positive rates and low false positive rates across a range of threshold values.

Now I have compared the receiver operating characteristic (ROC) curve of all three models Logistic Regression, Decision Tree and Random Forest visually below:

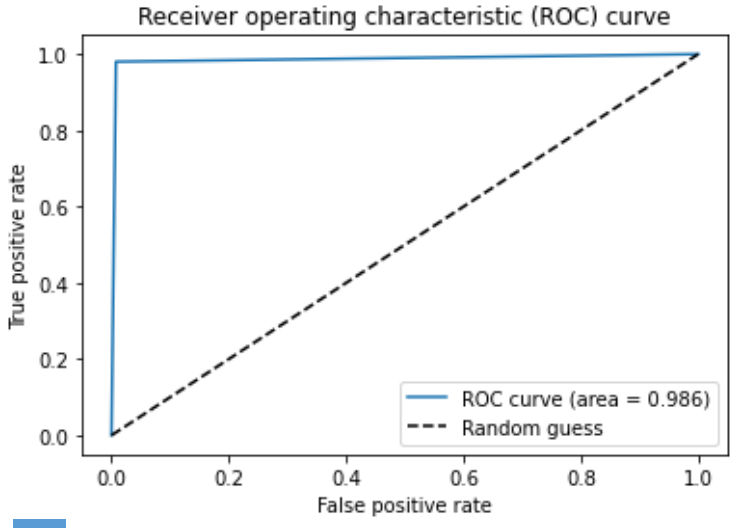
Model building



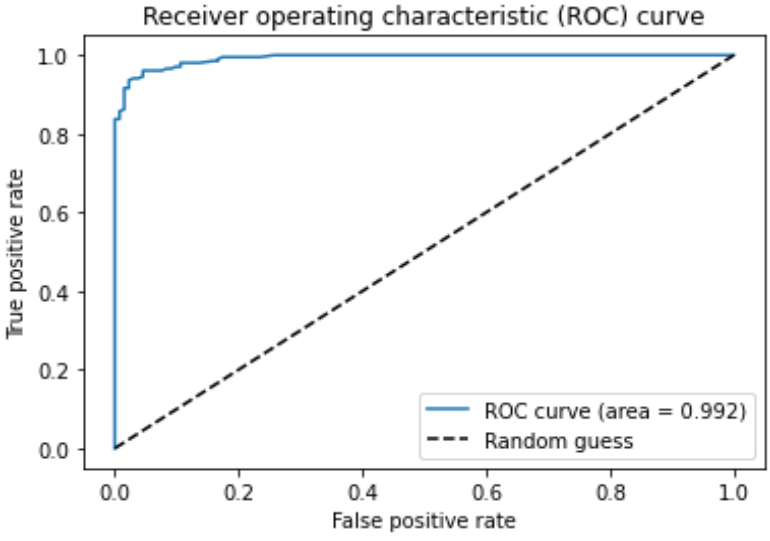
LOGISTIC REGRSSION

VS

DECISION TREE



VS



RANDOM FOREST

CONCLUSION

In this project, we compared the performance of Logistics Regression, Decision Tree and Random Forest Based on the results obtained from the three models, it can be concluded that all three models have performed well in predicting the insurance claim status for the given dataset.

The logistic regression model has an accuracy of 86.26%, while the decision tree model has an accuracy of 98.51%, and the random forest model has an accuracy of 95.22%.

In terms of sensitivity and specificity, the decision tree model performed the best with a sensitivity of 98.03% and a specificity of 99.24%. However, the random forest model also performed well in terms of sensitivity and specificity with a sensitivity of 94.09% and a specificity of 96.97%.

Regarding the ROC-AUC value, the random forest model had the highest ROC-AUC value of 0.992, followed by the decision tree model with a value of 0.986, and the logistic regression model with a value of 0.911.

Overall, the random forest model has the highest ROC-AUC value of 0.992, indicating that it has the best overall performance among the three models.

In conclusion, the random forest model is the best model for predicting the insurance claim status for the given dataset. However, it is important to note that the choice of model ultimately depends on the specific requirements and constraints of the project.