Applied Machine Learning and Data Mining

**2020/21 SECOND SUBMISSION – UNSUPERVISED LEARNING**

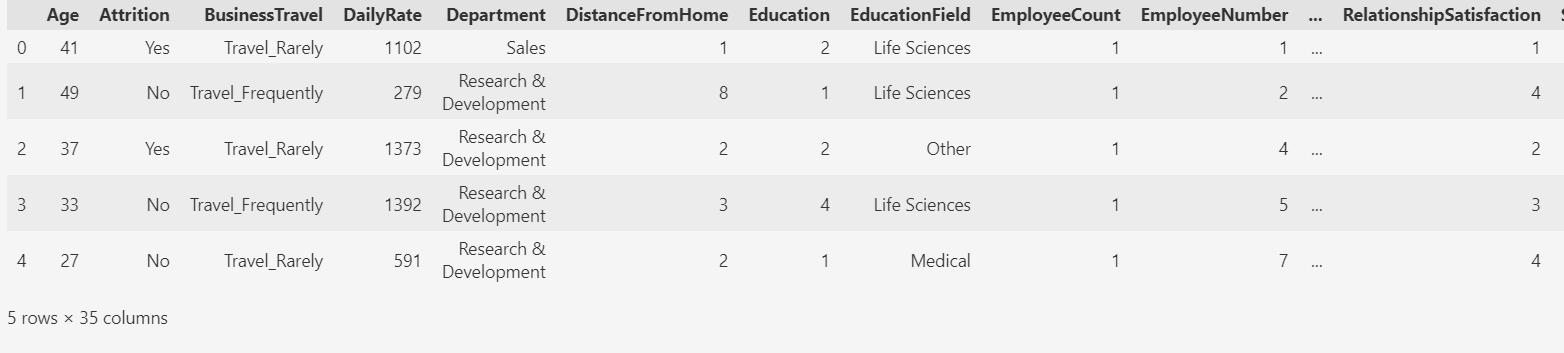
**UP2166428**

# Part 1: IBM HR Attrition Case Study

## 1.1 Business Problem Definition

Employee attrition is a significant concern for organisations due to its impact on costs, productivity, and profit. The objective of this project is to identify factors influencing employee attrition, suggest measures to retain employees, and predict potential attrition. The steps involved are data extraction, preprocessing, analysis, model development, and result interpretation.

## 1.2 IBM HR Attrition

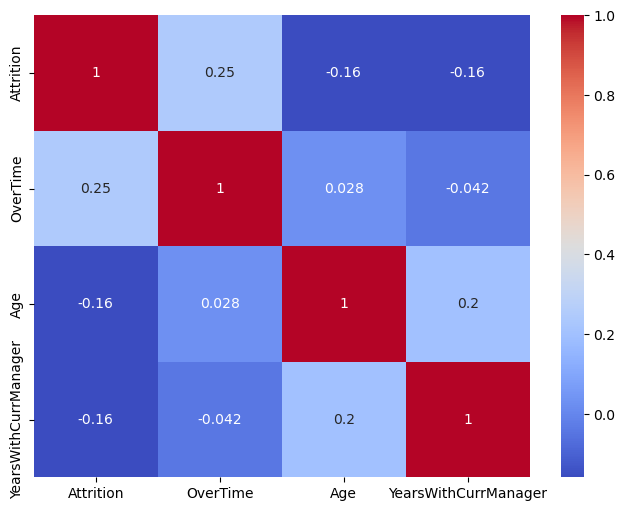
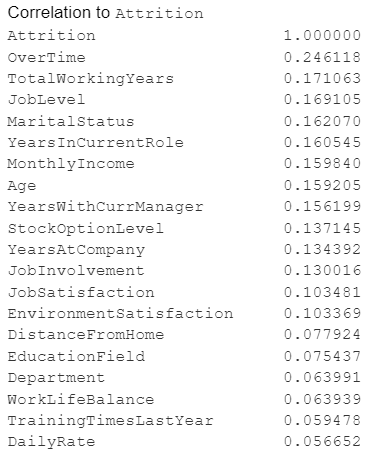
****

*Table 1*

## 1.3 Data preprocessing

This dataset has many columns, and many are in the ‘object’ format. Not all these features will be important for our target, ‘Attrition’. Before we use a correlation matrix, we need to change the remaining columns into numbers. This is important for our machine learning model.

The correlation matrix shows that features with a correlation less than 0.137145 corr might not be very helpful. So, we might need to remove them from our machine learning model.certain feature such as OverTime has positive correlation while age has a negative

*Figure 1*

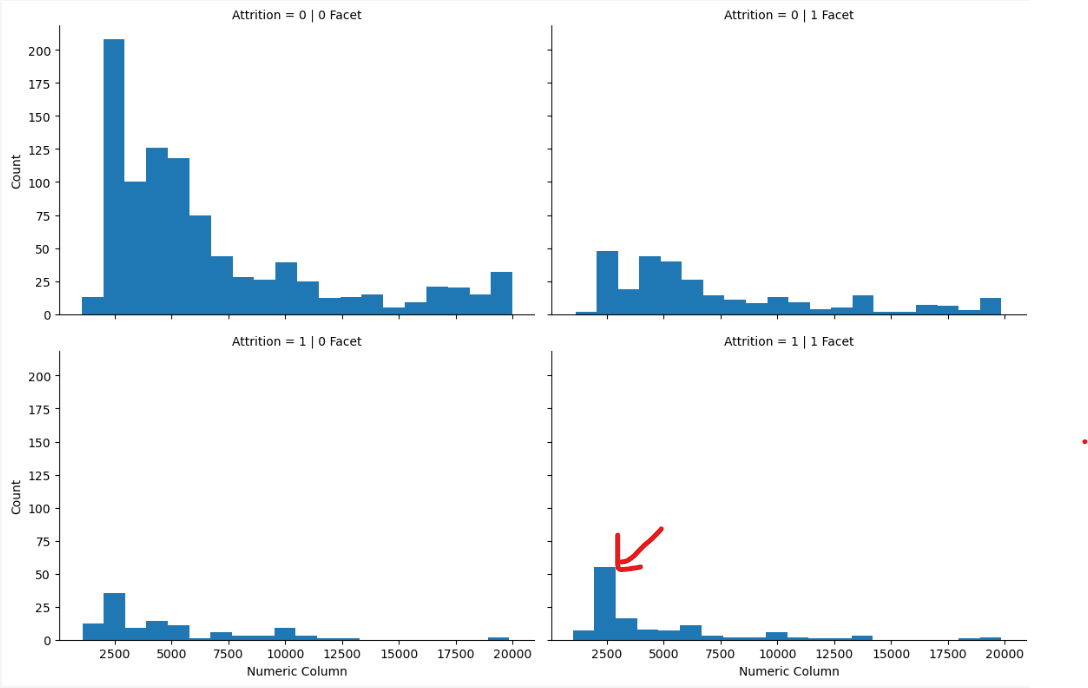


*Table 2*

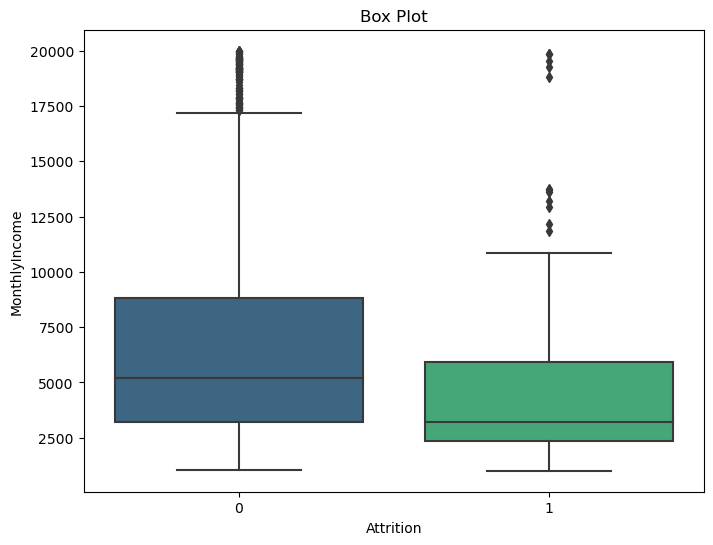
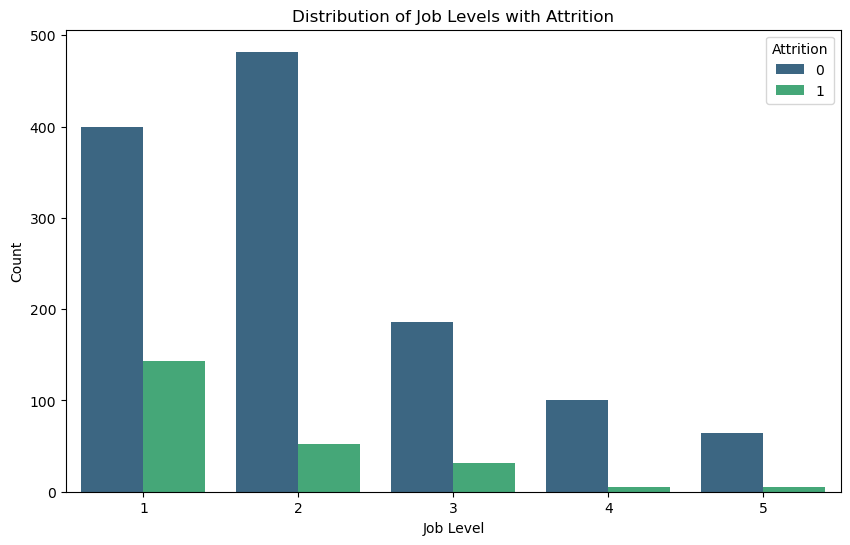
The data analysis in Table 2 reveals the following key points:

* The average age of employees in our dataset is 36.9 years. With a standard deviation of 9, this suggests that the majority of employees are between 29 and 46 years old. Furthermore, 75% of employees are 43 years or older, which is considered the upper quartile.
* In terms of work experience, the average employee has worked for about 11.28 years. However, this varies widely, with some employees having worked for as little as 0 years and others as much as 40 years.
* The job level data indicates that the majority of employees are at Level 1, which corresponds to the position of Research Scientist.

## 1.3 Data visualisation



*Figure 3*

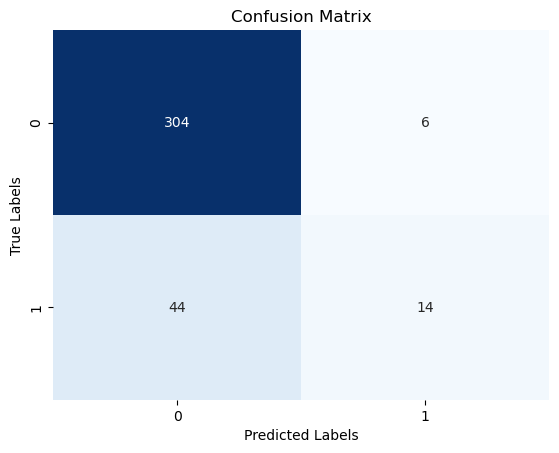
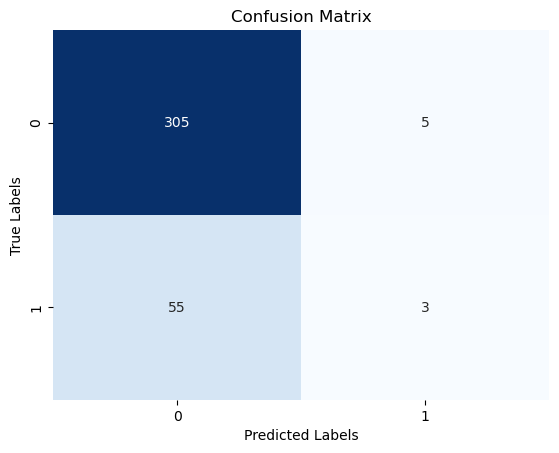
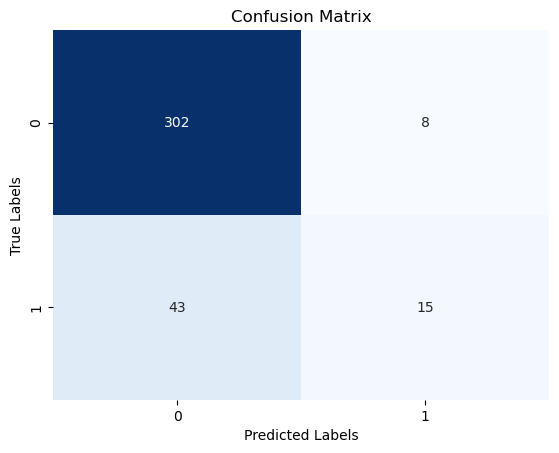


*Figure 3 Figure 4*

Data visualisation techniques such as Box plots, Bar plots, and FacetGrids were utilised in our analysis. It was observed that employees who are required to work overtime tend to leave the job more frequently. The highest attrition was seen among employees who work overtime and have a lower salary. Another pattern that emerged was that attrition is higher among employees at beginner level positions who work overtime, compared to those who have reached higher positions and also do overtime.

In conclusion, it was determined that employee attrition is associated with overtime, particularly among beginner level positions with lower salaries.

## 1.4 Model Training Overview

*Figure 4 Figure 5* 

*Figure 6*

I applied three different classification algorithms to predict if an employee is likely to quit based on a given set of features. The algorithms I used are random forest, k-nearest neighbors (KNN), and support vector machine (SVM). I compared the performance of these algorithms using accuracy as the evaluation metric.

Random forest is an ensemble learning method that builds multiple decision trees and combines their predictions. It can handle both categorical and numerical features, and it is robust to noise and overfitting. I used GridSearch to find the best parameters for the random forest model, such as the number of trees, the maximum depth, and the minimum samples per leaf. The confusion matrix for the random forest model is shown in Figure 4. The accuracy of the random forest model on the test set is 86.14%.

Some of the advantages of random forest are:

* It can handle both regression and classification problems
* It can handle a large number of features and missing values

Some of the disadvantages of random forest are:

* It can be computationally expensive and slow to train and test
* It can overfit if the number of trees is too high or the depth is too large

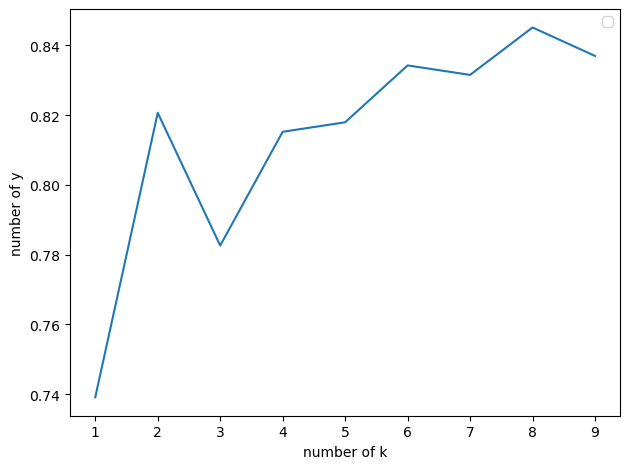
KNN is a lazy learning method that assigns the label of the majority of the k nearest neighbours to a new instance. It is simple and intuitive, but it can be sensitive to the choice of k and the distance metric. I used the elbow method to find the optimal value of k, which is the value that minimises the error rate. The confusion matrix for the KNN model is shown in Figure 5. The optimal value of k is 8, as shown in Figure 7. The accuracy of the KNN model on the test set is 84.51%.

Some of the advantages of KNN are:

* It can handle both regression and classification problems
* It can handle non-linear and complex relationships between the features and target variable

Some of the disadvantages of KNN are::

* It can be computationally expensive and slow to test
* It can suffer from the curse of dimensionality and require feature scaling



*Figure 7*

SVM is a kernel-based learning method that finds the optimal hyperplane that separates the data into different classes. It can handle nonlinear and high-dimensional data, but it can be computationally expensive and sensitive to the choice of the kernel and the regularization parameter. I used the radial basis function (RBF) kernel and the default value of the regularization parameter for the SVM model. The confusion matrix for the SVM model is shown in Figure 6. The accuracy of the SVM model on the test set is 86%.

Some of the advantages of SVM are :

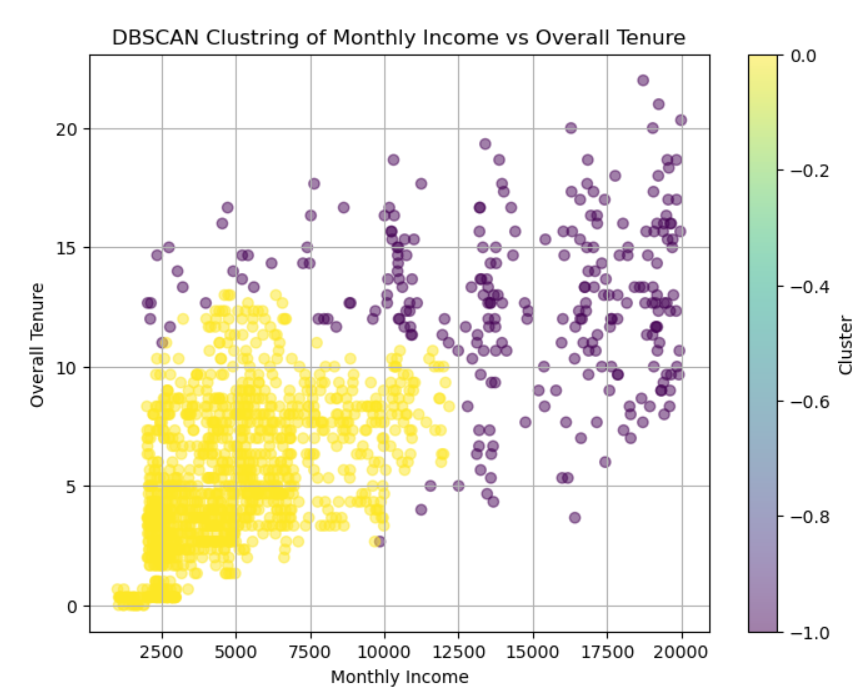
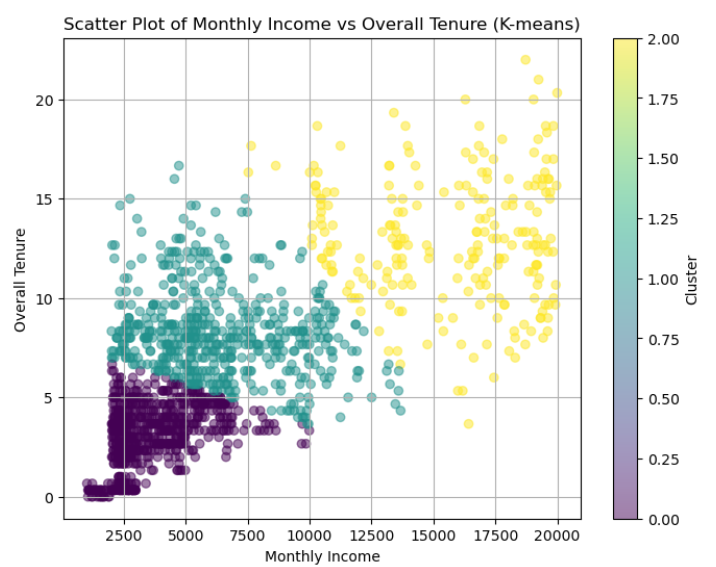
* It can handle both linear and non-linear classification problems
* It can handle high-dimensional data and reduce the feature space

Some of the disadvantages of SVM are :

* It can be computationally expensive and slow to train and test
* It can require feature scaling and one-hot encoding for categorical features

Based on the results, I can conclude that the three algorithms have similar performance in predicting employee attrition, with SVM and random forest slightly outperforming KNN. However, the differences are not statistically significant, and the accuracy scores are not very high, indicating that there is room for improvement.

## 1.5 Clustering Model



*Figure 8 Figure 9*

## **K-Means Clustering (Figure 8)**

The k-means clustering in Figure 8 suggests that employees can be segmented into three distinct groups based on their monthly income and overall tenure. This segmentation could imply that employees with lower incomes and shorter tenures are more likely to leave (attrition), while those with higher incomes and longer tenures might be more inclined to stay.

## **DBSCAN Clustering (Figure 9)**

The DBSCAN clustering in Figure 9 presents two groups but does not show as clear segmentation as k-means; however, it still indicates that employees with longer tenures tend to have higher incomes.

## **Suggestions**

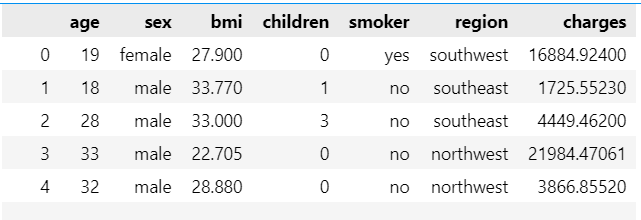
Based on these observations from Figures 8 and 9, it can be suggested that there may be a correlation between an employee’s likelihood to stay with an organisation (lower attrition) and their monthly income combined with overall tenure. Employees who have been with the company longer or who earn a higher salary are potentially less likely to experience attrition.

# Part 2 : Medical Cost Dataset

## 2.1 Business Problem Definition

Understanding why health insurance costs what it does is crucial for making it fair and accessible. By using computer models, we're trying to figure out how different things like age, lifestyle, and where you live affect those costs. This helps insurance companies and governments make better decisions about how much to charge and who needs help the most.

## 2.2 Medical cost dataset

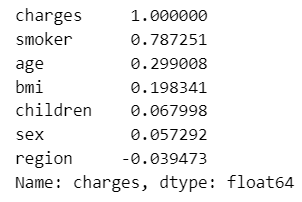
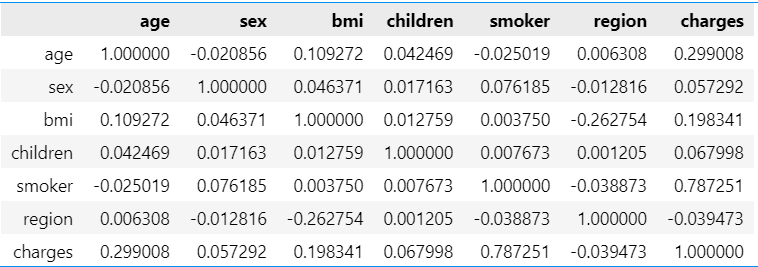


*Figure 10*

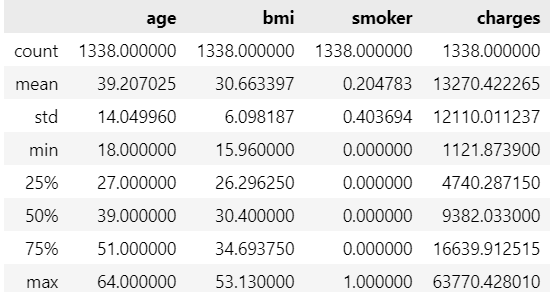
## 2.3 Data preprocessing

This dataset has many columns, and many are in the ‘object’ format. Not all these features will be important for our target, ‘charges’. Before we use a correlation matrix, we need to change the remaining columns into numbers. This is important for our machine learning model.

The correlation matrix shows features with a correlation less than 0.00…. corr might not be very helpful. So, we might need to remove ‘region’, ‘sex, ‘children’’ from our machine learning model.certain feature such as Smoker has positive correlation



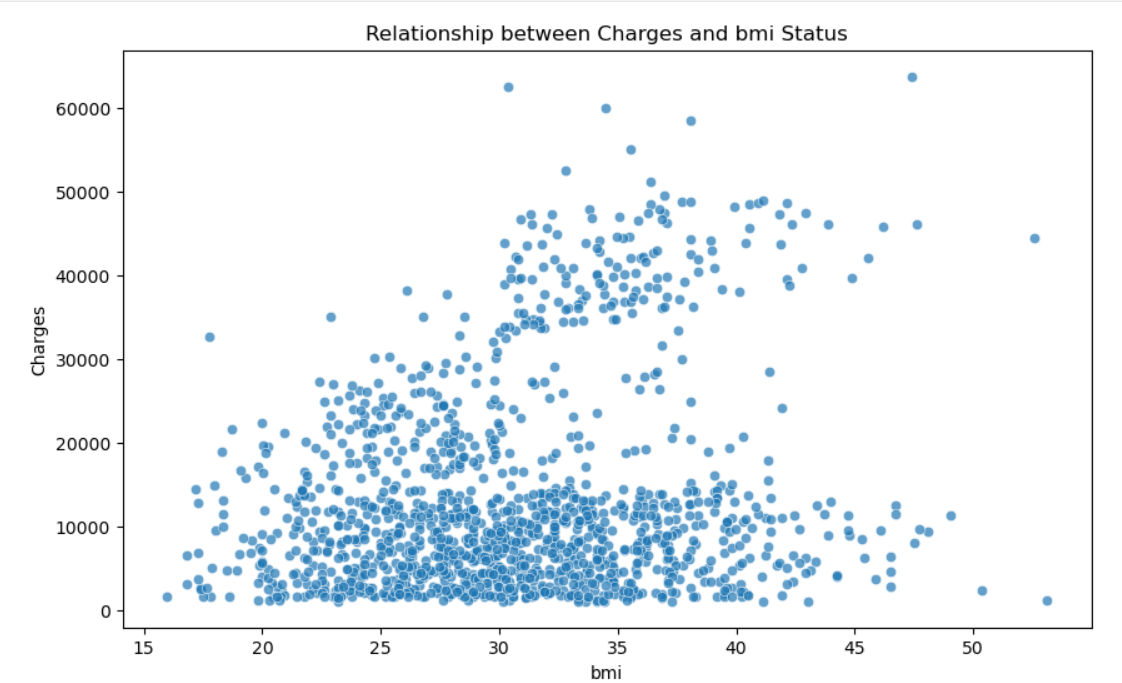
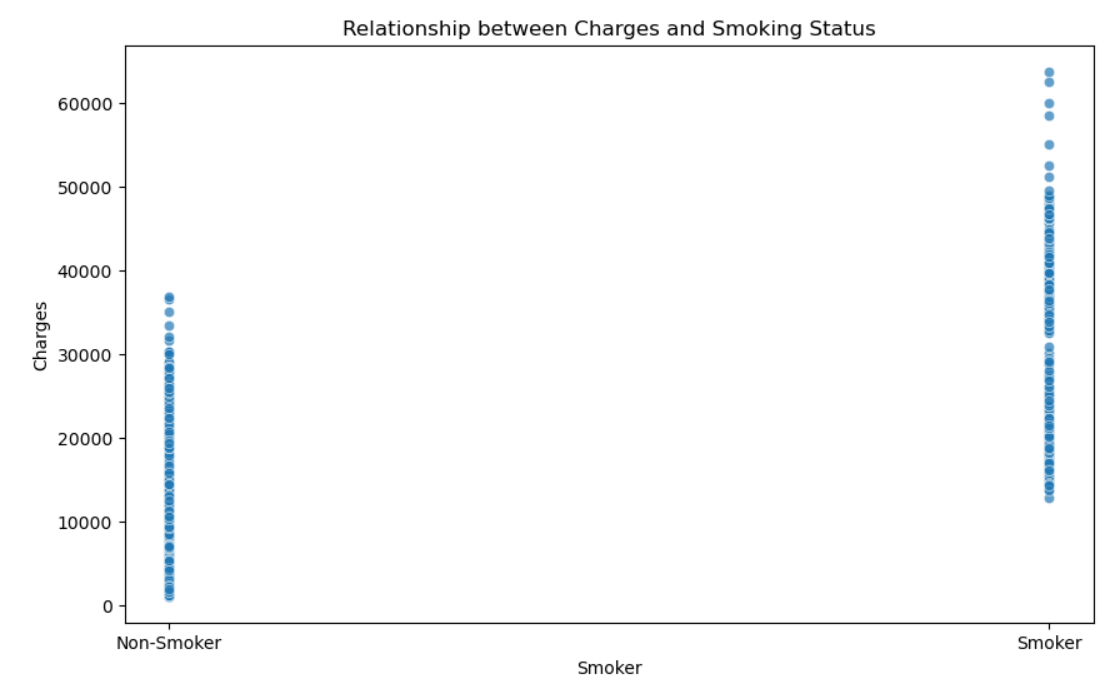
*Figure 11 Figure 12*

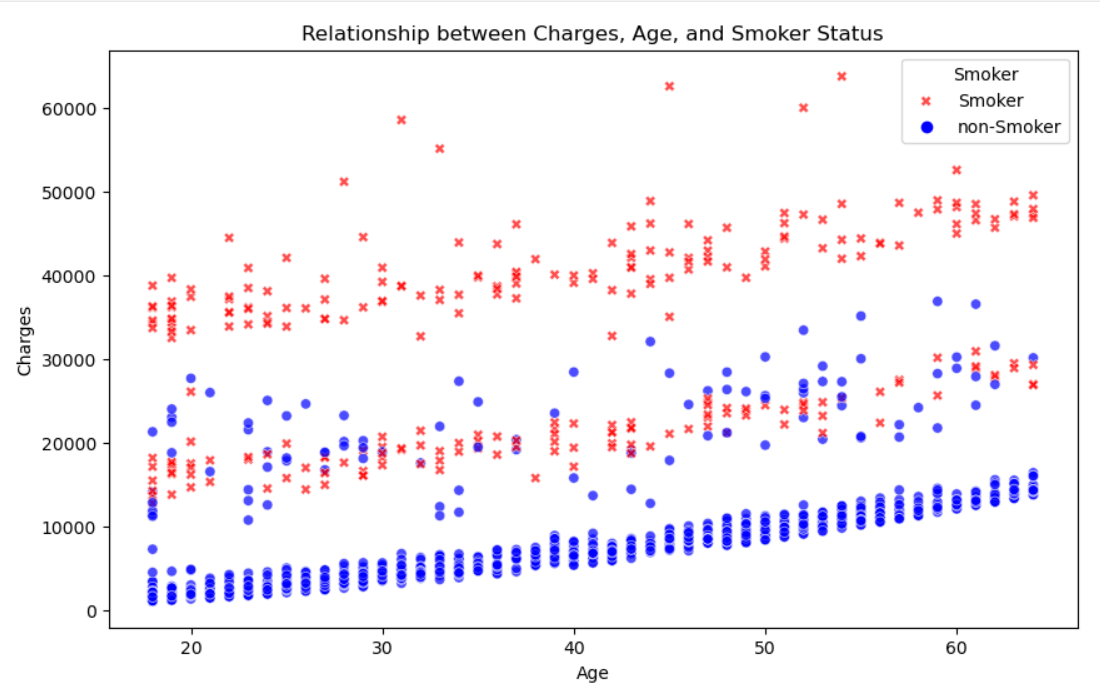
*Figure 13*

The data analysis in Figure11 reveals the following key points:

* The average age of policyholders in our dataset is 39.20 years. With a standard deviation of 14, this suggests that the majority of employees are between 25 and 53 years old. Furthermore, 75% of employees are 51 years or older, which is considered the upper quartile.
* In terms of bmi, the average policyholder has a BMI of 0,20s..

## 2.4 Data Visualization

 *Figure 14 Figure 15*



*Figure 16*

1. Smoking Status and Charges: The first scatter plot showed a clear distinction between the medical charges incurred by smokers and non-smokers. Smokers generally have higher medical charges compared to non-smokers. This suggests that smoking status is a significant factor associated with higher medical expenses.
2. BMI and Charges: The second scatter plot showed the relationship between a person’s Body Mass Index (BMI) and their medical charges. While the data points were widely spread out, there was a slight positive trend. This indicates that higher BMI might be associated with higher charges, although the relationship was not very strong.
3. Age, Smoking Status, and Charges: The third scatter plot showed the relationship between a person’s age, smoking status, and their medical charges. As age increases, there is a general trend of increasing charges for both smokers and non-smokers. Additionally, across all ages, smokers have higher charges compared to non-smokers. This reaffirms the observation from the first scatter plot and adds the dimension of age to the analysis.

In conclusion, the visualisations suggest that a person’s smoking status, BMI, and age are all factors associated with their medical expenses. Among these, smoking status appears to have the most significant impact

## 2.5 Model Training Overview

#### **Algorithm Selection:**

For the regression task of predicting health insurance charges, I opted to utilise the Multiple Linear Regression algorithm. This choice was made due to its simplicity, interpretability, and ability to capture linear relationships between multiple independent variables (features) and the target variable (health insurance charges). Multiple Linear Regression is well-suited for this task as it allows us to analyse the combined effect of various demographic and lifestyle factors on insurance charges.

#### **Performance evaluation:**

The Multiple Linear Regression model achieved an R2 score of approximately 0.778, indicating that around 77.8% of the variance in health insurance charges is explained by the model. The adjusted R2 score of approximately 0.775 provides a more conservative estimate, considering the number of predictors. Additionally, the mean squared error (MSE) of approximately 34512843.88 reflects the model's average prediction accuracy. Overall, these metrics suggest the model's capability in capturing relationships between demographic and lifestyle factors and health insurance charges, albeit with room for further refinement.

# Part 3 : association rule

# 

The Apriori algorithm has been used to mine frequent itemsets and learn association rules in a dataset. The results are presented as a list of association rules, each with three key metrics: support, confidence, and lift. These metrics are crucial for evaluating the strength and relevance of the association rules discovered.

Here are some of the key findings:

1. Rule: Ozark -> Demon Slayer
   * Support: 0.005573330581071318
   * Confidence: 0.6300000000000001
   * Lift: 3.097176459275536

This rule suggests that if a viewer watches ‘Ozark’, they are likely to watch ‘Demon Slayer’ as well. The confidence of 0.63 indicates that 63% of the customers who watch ‘Ozark’ also watch ‘Demon Slayer’. The lift of 3.09 suggests that viewers of ‘Ozark’ are about three times more likely to watch ‘Demon Slayer’ than randomly chosen viewers.

1. Rule: Mr. Robot -> Ozark
   * Support: 0.0059750322222107
   * Confidence: 0.6021739130434805
   * Lift: 2.34837771556454

This rule suggests that viewers of ‘Mr. Robot’ are likely to watch ‘Ozark’. The confidence of 0.60 indicates that 60% of the customers who watch ‘Mr. Robot’ also watch ‘Ozark’. The lift of 2.35 suggests that viewers of ‘Mr. Robot’ are about 2.35 times more likely to watch ‘Ozark’ than randomly chosen viewers.

1. Rule: Mr. Robot -> Stranger Things
   * Support: 0.006521854315934827
   * Confidence: 0.6578947368421052
   * Lift: 3.161780855242083

This rule suggests that viewers of ‘Mr. Robot’ are likely to watch ‘Stranger Things’. The confidence of 0.66 indicates that 66% of the customers who watch ‘Mr. Robot’ also watch ‘Stranger Things’. The lift of 3.16 suggests that viewers of ‘Mr. Robot’ are about 3.16 times more likely to watch ‘Stranger Things’ than randomly chosen viewers.

These findings provide valuable insights into the viewing patterns of customers. They can be used by netflix to make recommendations to viewers, thereby enhancing their viewing experience and potentially increasing viewer engagement.