UNIVERSITY OF LONDON

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Module – ST3189

Student ID- 210473416

Number of pages – 10 (Excluding the table of contents and bibliography

Table of Contents

TASK 1 – UNSUPERVISED LEARNING

1.1 - Introduction

Unsupervised learning uses machine learning algorithms to analyze and cluster unlabeled datasets. Unsupervised learning models are utilized mainly for clustering, association, and dimensionality reduction. (IBM, n.d.)

The dataset used for this task is the iris dataset from Kaggle. The iris dataset contains measurements on four different attributes from three different species. There are 150 data points in the dataset containing equal number of unique values for each species. The task below will focus on using clustering to group data into meaningful clusters.

1.2 – Literature Review

(Enthought, n.d.) approached a similar problem and used K-means clustering algorithm to segment the flowers and used elbow method to determine the number of clusters. The research was conducted, and it was identified that there were three homogenous clusters.

1.3 – Research Questions

1. Is there a relationship between sepal length, sepal width and the species?

2. Is there a relationship between petal length, petal width and the species?

3. How many clusters can be identified using this dataset?

1.4 – Exploratory Data Analysis

The heatmap indicates that petal length and petal width have high correlations, petal length and sepal length have good correlations, and petal width and sepal length have high correlations .

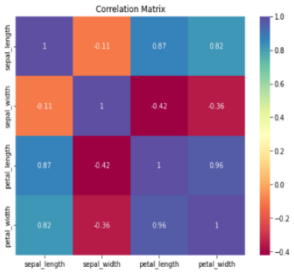


Figure 1

Figure2-Figure5 show the histograms with the distribution plot for the petal length, petal width, sepal length and sepal width. It's noteworthy that in the case of petal length and petal width there's hardly any overlapping but there's significant overlapping in the sepal length and sepal width. Hence, petal length and petal width can be used as features for clustering.

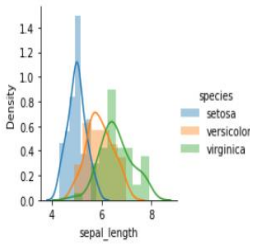


Figure 2

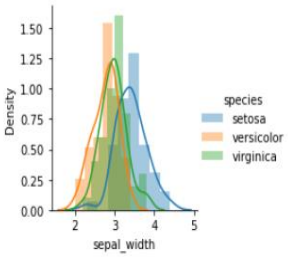


Figure 3

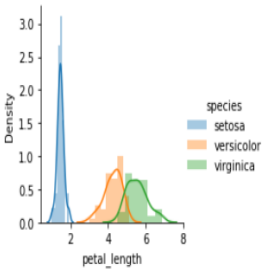


Figure 4

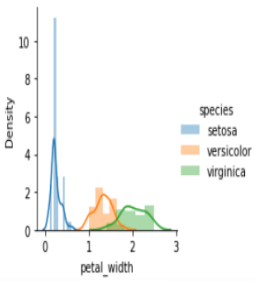


Figure 5

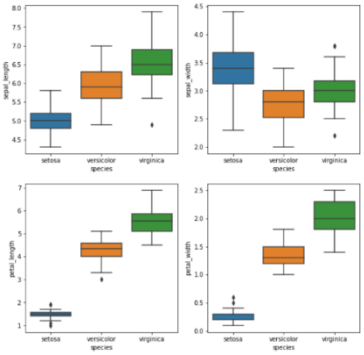


Figure 6

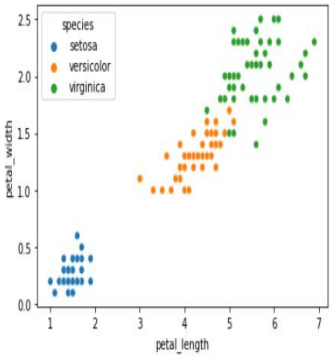


Figure 7

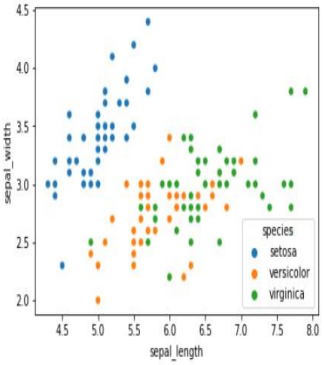


Figure 8

Figure-6 indicates that Setosa species shows the smallest features with relatively low distribution, Virginica illustrates the largest features and Versicolor displays average features. Figure-7 indicates that Setosa has smaller petal lengths and petal widths, Virginica has the largest of petal widths and petal lengths and Versicolor lies in the middle of the two features in terms of petal length and petal width. While figure-8 shows that Setosa has smaller sepal lengths but larger sepal widths, Virginica has larger sepal lengths but smaller sepal widths, Versicolor lies in the middle of the two species in terms of sepal length and sepal width.

1.5 – K-Means clustering

K-means clustering algorithm divides a set of n observations into k clusters. It forms groups in a manner that minimizes the variance between the data points and the clusters centroid. (statistics, n.d.)

The elbow method is a graphical representation of finding the optimal K in a K-means clustering. It works by finding the within-cluster sum of square i.e. the sum of the square distance between points in a cluster and the cluster centroid. (analytics, n.d.)

The elbow graph indicates that there is an elbow point at number of clusters being equal to 3. Therefore, K-Means clustering will be performed using this optimal number 3.

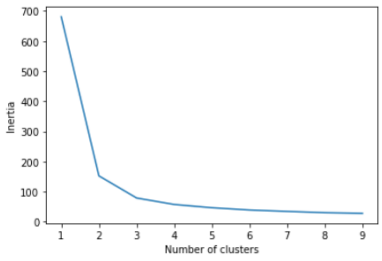


Figure 9- Elbow graph

K-Means clustering was plotted using the scatter plot and 3 clusters were formed as it's the optimal number of clusters as shown in the elbow graph. Petal width and petal length were taken as a pair of features to form clusters and sepal length and sepal were taken as the second pair of features to form clusters.

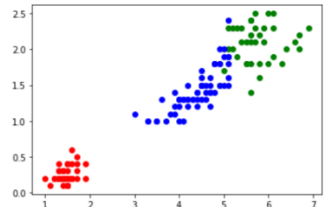


Figure 10- clustering with respect to petal length and petal width

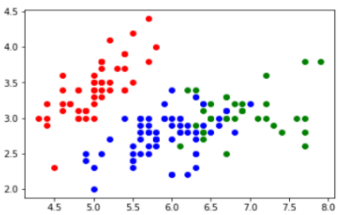


Figure 11- clustering with respect to sepal length and sepal width

Its noteworthy from the graphs above that petal width and petal length is a good feature pair than sepal length and sepal width when used in clustering.

1.6 – Principal Component Analysis (PCA)

PCA is a dimensionality reduction method that is often used to reduce the dimensionality of large datasets by changing a large set of variables into a simplified one that still consists most of the information in the large set. (builtin, n.d.)

Although the number of features used in this dataset is small, PCA was still conducted to analyse the most significant features.

From figure-12 its noteworthy that as the number of components increase, the proportion of variance explained by each component decreases. After the first three principal components the marginal change in cumulative explained variance becomes negligibly small. The first three components explain around 98% of the total variation in the dataset which is close to 100% hence we would use 3 principal components for the dataset.

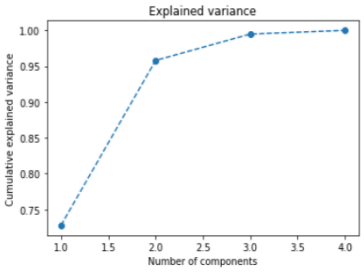


Figure 12

Each principal component is a linear combination of the original features.

From the explained variance ratio, we can see that PC1 contributes to 72.8% of the variance, and PC2 contributes to 23% of the variance, PC3 contributes to 3.6% which sums up to 99.4% of the variance in the data which is explained by these three Principal Components.
The greater the value, the more it contributes to the variance of the Principal Components. From the explained variance its noteworthy that PC1 contributes to 2.93 units of variance in the original dataset, PC2 contributes to 0.93 units and PC3 contributes to 0.14 units.

	Explained Variance	Explained Variance Ratio
PC1	2.930354	0.727705
PC2	0.927404	0.230305
PC3	0.148342	0.036838

Figure 13

Biplot is a 2-dimensional figure plotted for two principal components and it shows us the impacts of the original variables on each principal component.
Figure-14 indicates that petal length, petal width and sepal length are in the same direction as PC1, hence they are positively correlated with PC1. Sepal length is in the opposite direction with PC1, which indicates a negative correlation with PC1. In simpler terms, PC1 tends to signify smaller sepal widths alongside larger sepal lengths. Additionally, it often corresponds to greater petal measurements.

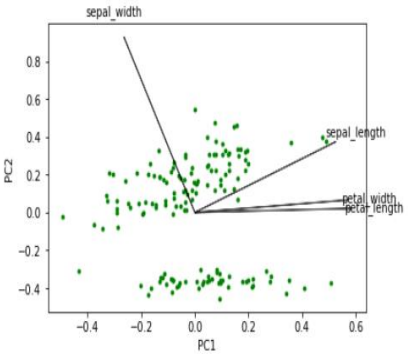


Figure 14

TASK 2 – CLASSIFICATION

2.1 - Introduction

Supervised learning is a subcategory of machine learning and artificial intelligence. It relies on labelled datasets to train algorithms effectively, enabling them to accurately classify data or make predictions about outcomes (IBM, n.d.).

Classification entails the identification, comprehension, and categorization of objects and concepts into predefined groups, also known as “sub-populations” (simplilearn, n.d.). The dataset used for the classification task is the diabetes dataset obtained from Kaggle. The objective of this dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset.

2.2 – Literature Review

Research was done to check the interaction between age group and risk factors of diabetes. Among the 2776 individuals analysed, diabetes was found to affect 15.1% of the population, while 52.3% were identified as having prediabetes. Notably, both diabetes and prediabetes were more prevalent among the elderly compared to the middle-aged group (NIH, n.d.). Researchers from the university of Cambridge studied that those in the highest BMI group had a 11- fold increased risk of diabetes compared to participants in the lowest BMI group (british heart foundation, n.d.). Information from (health, n.d.) indicates that high blood pressure is twice as likely to strike a person with diabetes than without diabetes. Moreover, (analytics vidhya, n.d.) did similar research and found that out of four models, random forest is the best model for this prediction with an accuracy score of 0.76.

2.3 – Research Questions

1. Does a person’s age have an impact on the person having diabetes?

2. Does a person’s BMI and blood pressure have an impact on the person having diabetes?

3. what is the best classification model to classify non-diabetic people and diabetic people?

2.4 – Exploratory Data Analysis

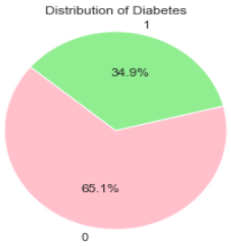


Figure 15

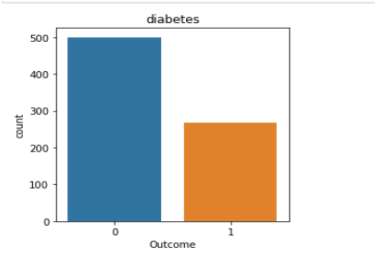


Figure 16

Its noteworthy, that if the outcome is **1** it indicates that the person has **diabetes** and **0** if the person **doesn’t have diabetes**. Figure-15 and figure-16 shows that there is a significant disparity between the number of individuals diagnosed with diabetes and those without diabetes. Hence, it’s clear that the dataset is imbalanced (where one class significantly outweighs the other).

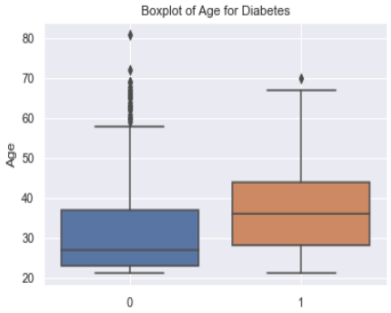


Figure 17

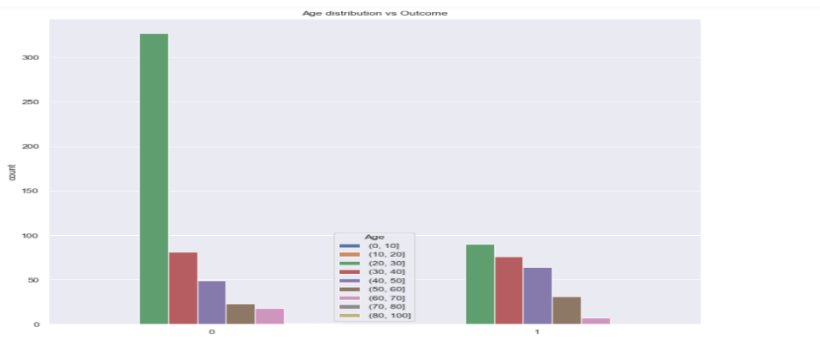


Figure 18

From figure-17 its noted that the median age for a person with diabetes is higher than for a person not having diabetes. Figure-18 shows that people having diabetes starts from age 20 and above and most of the people between the age 20-30 doesn't have diabetes. Moreover, the number of people having diabetes above the age of 40 is higher compared to people who don't have diabetes. Hence, the older the person is, there is more risk the person will have diabetes.

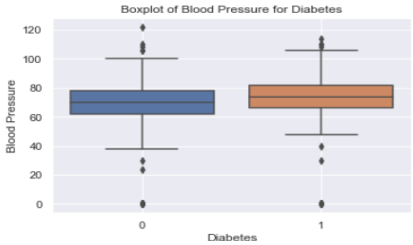


Figure 19

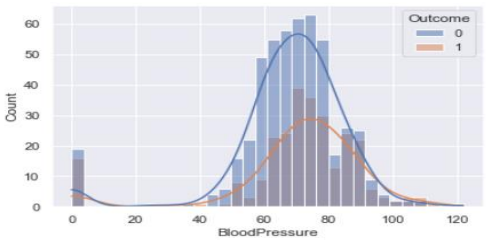


Figure 20

From figure-19 it shows that the median blood pressure for a person with diabetes is higher than for a person without diabetes. Excluding the outliers in the dataset, its noteworthy that if the blood pressure is more than 100 the person has diabetes. From figure-20 it depicts that the graph for the people with diabetes is skewed to the right. Hence, people with high blood pressure are more likely to have diabetes.

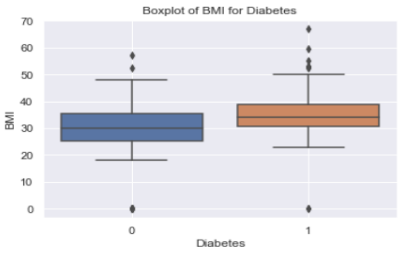


Figure 21

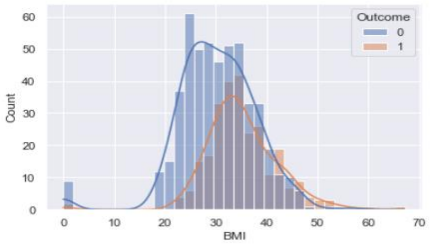


Figure 22

Figure-21 indicates that the median BMI for the people with diabetes is higher than a person without diabetes. The histogram in figure-22 reveals that diabetic individuals tend to exhibit higher BMI values compared to non-diabetic individuals. Consequently, its apparent from the graphs that individuals with elevated BMI are more likely to have diabetes.

2.5 – Feature Selection

The target variable for the model is "Outcome". All features were taken as the predictor variables. It was found that glucose exerted the most significant influence on the target variable. From examining the correlation plot, it was noted that none of the predictor variables exhibited a highly robust linear relationship with the variable "Outcome". The highest

correlation was between age and pregnancies with a value of +0.54. Since the value is not close to 1 the variables were not removed.

2.6 – Classification Models

Four classification models were tested, which are as follows:

1.Support Vector Classifier- This model classifies data by finding an optimal hyperplane or line that maximises the distance between each class in an N-dimensional space. (IBM, n.d.)

2.Decision Tree classifier- It is a non-parametric supervised learning model which has a hierarchical, tree structure which consists of branches, leaf nodes, internal nodes, and root nodes. (IBM, n.d.)

3.Random Forest Classifier- It is a model that combines the output of multiple decision trees to reach a single result. (IBM, n.d.)

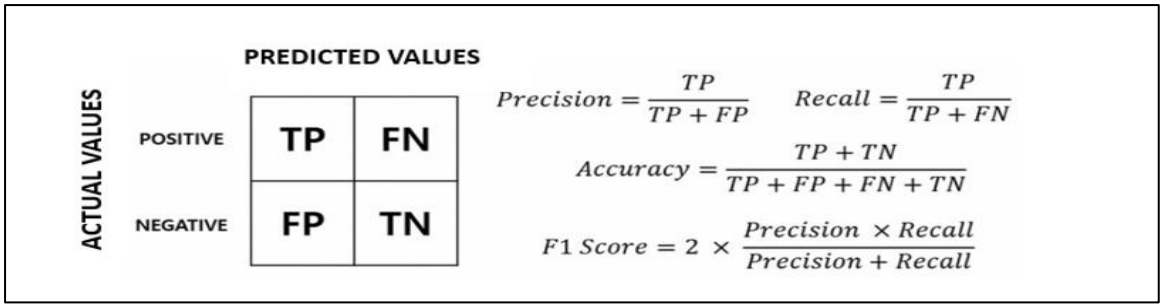
4.Logistic Regression- It is a model which estimates the probability of an event occurring based on a given set of independent variables. (IBM, n.d.)

Below are the evaluation metrics utilized for evaluating the model.

Confusion matrix- It is a means of displaying the number of accurate and inaccurate instances based on the models' predictions.

ROC curve- It is a graphical plot that illustrates the performance of a binary classifier model at varying threshold values.

The accuracy, precision, recall and F1 scores can be calculated using the following equations from the confusion matrix. Where TP refers to true positive values, TN refers to true negative values, FN refers to false negative values and FN refers to false negative values.



Accuracy – Out of all the predictions which were made how many were true?

Precision – Out of all the positive predictions which were made how many were true?

Recall – Out of all the data points that should be predicted as true, how many were correctly predicted as true?

F1 Score- It is a measure that combines both precision and recall.

Before fitting the model, the data was divided into training and testing data in a ratio of 80%-20% and the models were then tested. The following table shows the evaluation metrics obtained from the confusion matrix from each model. Where the total number of cases considered in the confusion matrix is 154.

Model	Accuracy	Precision	Recall	F1-SCore	ROC-AUC score
Decision Tree	71.4%	66.7%	53.3%	59.2%	68.2%
Logistic Regression	80.5%	82.6%	63.3%	71.7%	77.4%
Random Forest Classifier	78.6%	78.7%	61.7%	69.2%	75.5%
Support Vector Machine	81.8%	90%	60%	72%	77.9%

Accuracy is a common metric used to evaluate model performance but if the dataset is imbalanced, accuracy can be misleading. Since F1-Score is a combination of the recall and precision, F1-Score will be used instead of recall and precision. Hence, to determine which model is the best using the evaluation metrics F1-Score and the ROC-AUC score will be used. As observed from the above table its shown that the F1-score is the highest for support vector machine with a value of 72%. The next highest was Logistic regression, followed by Random Forest Classifier.

Greater the ROC-AUC score, better is the model in differentiating between diabetic and non-diabetic people. It's noted from the above table that Support vector machine has the highest ROC-AUC score. Since Support vector machine has the highest ROC-AUC score and the highest F1-score it's the best model in classifying the diabetes dataset.

TASK 3 – REGRESSION

3.1 – Introduction

Regression is a supervised learning technique where a set of statistical methods are used for the estimation of relationships between a dependent variable and one or more independent variables (CFI, n.d.). The dataset used for this task is the medical cost insurance dataset consisting of seven variables of which the charges (individual medical costs billed by health insurance) is the continuous target variable.

3.2 – Literature Review

Research was done and found that smokers must pay more money in policy premiums than a non-smoker (Outlook, n.d.). Research, which was done, found that women usually pay higher medical insurance than men (future generali, n.d.). Similar research was done on the medical cost insurance dataset, and it was found that the random forest regression model performs better as compared to the remaining models considering the RMSE and R2 (IJRTI).

3.3 – Research Questions

1. Does smoking impact the level of medical insurance charged for an individual?

2. Does gender and the number of children impact the level of medical insurance charged for an individual?

3. What is the best regression model to predict the insurance charged?

3.4 – Exploratory Data Analysis

Figure-23 illustrates the relationship between BMI and charges, distinguishing between smokers and non-smokers. It reveals that, regardless of BMI level, smokers tend to incur higher insurance charges compared to non-smokers.

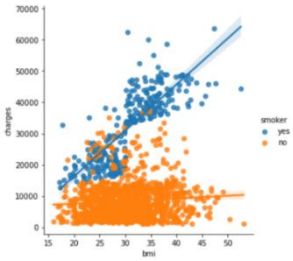


Figure 23

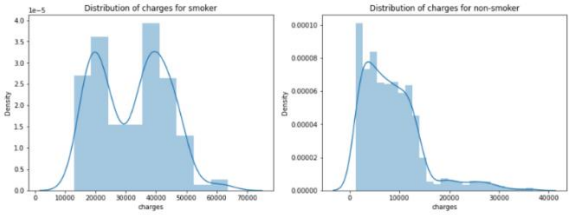


Figure 24

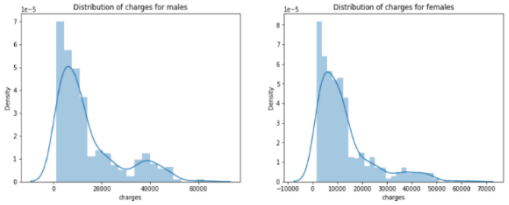


Figure 25

Figure-24 indicates that the graph for smokers is bimodal, and it shows that smokers face higher insurance charges than non-smokers. Additionally, there is a notable prevalence of smokers with elevated charges, indicating a greater proportion of smokers paying higher premium compared to those paying lower premiums. Moreover, figure-25 indicates that the overall insurance charged for males is higher compared to females and most of the insurance charged for males is less than 20,000 while for females its less than 15,000. It's also noteworthy that the graphs in figure-25 are positively skewed.

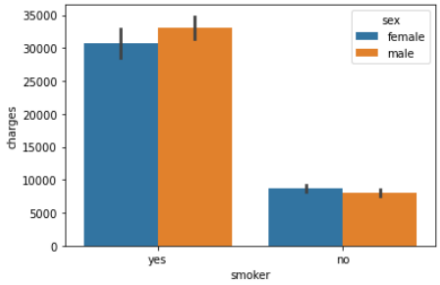


Figure 26

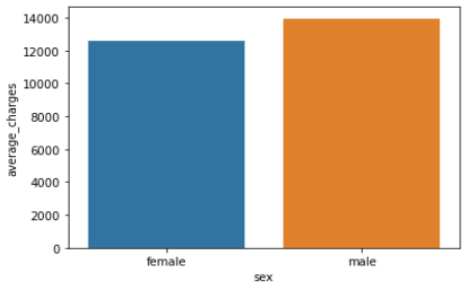


Figure 27

Figure-26 shows that the insurance charged for male is higher than for a female if the individual is a smoker and it is the opposite if the individual is a non-smoker. It's also noteworthy that in figure-27 it indicates that the overall average charge for a male is higher than a female. Which goes against the information in the literature review.

Figure-28 indicates that out of all the regions, southeast is the region which is charged the most and it contains more smokers compared to the other regions.

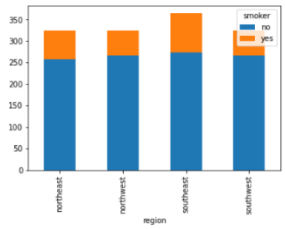


Figure 28

It's shown from figure-29 that as the number of children increase the insurance charged increases until the number of children is 3. If the number of children is more than 3 the insurance charged reduces as the number of children increases.

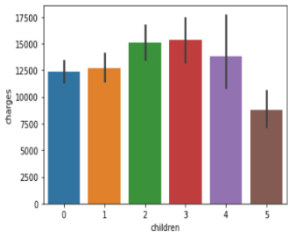


Figure 29

3.5 – Feature Selection

Figure-30 is a correlation heatmap which displays the correlation between multiple variables as a colour coded matrix. The regions were converted into numerical value by using dummy variables. Its noteworthy that the most relevant features to charges are smoker, age, and BMI, which brings it down to only three features. Hence, if the correlation value with charges was more than 0.01 the feature was considered as the predictor variable to predict the insurance charged. Therefore, the only feature which was removed was the region Northeast.

Figure 30

3.6 – Regression Models

The data was divided into training and testing data in a ratio of 80%-20% and then the models were then tested. The objective was to utilize the training dataset to train the models and then assess their performance using this test dataset. The dataset was fitted using multiple different regression models to come up with the best regression model which can be used to predict the insurance charges.

Below are the evaluation metrics utilized for evaluating the model.

R-Square- It is a statistical measure that determines the proportion of variance in the dependent variable that can be explained by the independent variable. Hence, we will want the R-square to be high as possible.

Mean Squared Error (MSE)- It is a statistical measure that assesses the average squared difference between the observed and predicted values.

Root Mean Squared Error (RMSE)- It is the standard deviation of the residuals.

Mean Absolute Error (MAE)- It is a measure of errors between paired observations expressing the same phenomenon.

Explained Variance – This measures the variance between the model and the actual data.

Since MSE, RMSE, and MAE are loss functions we would want it to be low as possible, but we would want the R-Squared and explained variance values to be high as possible.

The below table shows the models used for this task and the evaluation metrics values for each model.

MODEL	R-SQUARED	MSE	RMSE	MAE	EXPLAINED VARIANCE
Random Forest	0.86	4956.56	70.40	2783.04	0.86
Decision Tree	0.70	7093.62	84.22	3438.60	0.70
Linear Regression	0.76	6252.75	79.07	4321.00	0.76
XGB Regressor	0.83	5214.59	72.21	3005.90	0.83

Its noteworthy from the above table that Random Forest is the best model in predicting the insurance charges as it has the highest R-square, explained variance and the lowest RMSE, MSE, and MAE. Therefore, this is the best model for predicting insurance charges, with 86% of variance being explained.

Hyperparameter tuning was done to the XGB regressor to improve the accuracy, however the original model was more accurate.

Moreover, it's noteworthy that to do the linear regression model the following assumptions should be satisfied.

i) No multicollinearity in the data
ii) Homoscedasticity of residuals
iii) No autocorrelation in residuals
iv) Predictors are normally distributed.

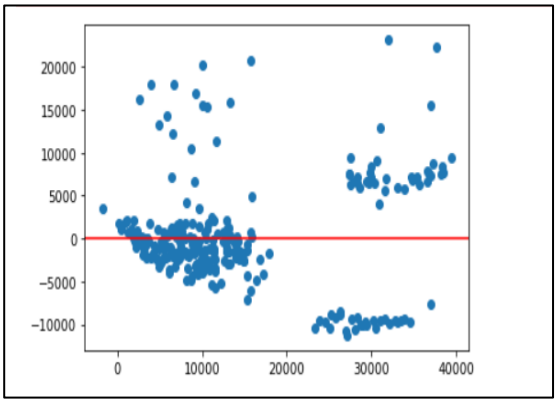
OLS Regression Results

Dep. Variable:	charges	R-squared:	0.751
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	500.8
Date:	Mon, 25 Mar 2024	Prob (F-statistic):	0.00
Time:	13:22:10	Log-Likelihood:	-13548.
No. Observations:	1338	AIC:	2.711e+04
Df Residuals:	1329	BIC:	2.716e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.194e+04	987.819	-12.086	0.000	-1.39e+04	-1e+04
age	256.8564	11.899	21.587	0.000	233.514	280.199
sex	-131.3144	332.945	-0.394	0.693	-784.470	521.842
bmi	339.1935	28.599	11.860	0.000	283.088	395.298
children	475.5005	137.804	3.451	0.001	205.163	745.838
smoker	2.385e+04	413.153	57.723	0.000	2.3e+04	2.47e+04
southwest	-960.0510	477.933	-2.009	0.045	-1897.636	-22.466
northwest	-352.9639	476.276	-0.741	0.459	-1287.298	581.370
southeast	-1035.0220	478.692	-2.162	0.031	-1974.097	-95.947

Omnibus:	300.366	Durbin-Watson:	2.088
Prob(Omnibus):	0.000	Jarque-Bera (JB):	718.887
Skew:	1.211	Prob(JB):	7.86e-157
Kurtosis:	5.651	Cond. No.	311.

Ols results



Residual plot

Its noteworthy from the Ols regression results and the residual plot that the error terms are homoscedastic, and the error terms are uncorrelated. Hence, linear regression can be used.

Bibliography

(n.d.). Retrieved 3 18, 2024, from IBM: https://www.ibm.com/topics/unsupervised-learning

(n.d.). Retrieved 3 18, 2024, from statistics: https://statisticsbyjim.com/basics/k-means-clustering/

(n.d.). Retrieved from analytics: https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/

(n.d.). Retrieved 3 19, 2024, from builtin: https://builtin.com/data-science/step-step-explanation-principal-component-analysis

(n.d.). Retrieved 3 22, 2024, from health: https://www.hopkinsmedicine.org/health/conditions-and-diseases/diabetes/diabetes-and-high-blood-pressure

(n.d.). Retrieved 3 22, 2024, from analytics vidhya: https://www.analyticsvidhya.com/blog/2022/01/diabetes-prediction-using-machine-learning/#:~:text=Conclusion,in%20the%20dataset%20have%20diabetes.

(n.d.). Retrieved from IBM: https://www.ibm.com/topics/support-vector-machine

(n.d.). Retrieved from IBM: https://www.ibm.com/topics/decision-trees

(n.d.). Retrieved from IBM: https://www.ibm.com/topics/random-forest

(n.d.). Retrieved from IBM: https://www.ibm.com/topics/logistic-regression

(n.d.). Retrieved 3 22, 2024, from CFI: https://corporatefinanceinstitute.com/resources/data-science/regression-analysis/

(n.d.). Retrieved 3 22, 2024, from Outlook: https://business.outlookindia.com/insurance/how-does-smoking-affect-your-health-insurance-premiums#:~:text=Although%20there%20is%20a%20belief,etc.%2C%20as%20a%20smoker.

(n.d.). Retrieved 3 22, 2024, from future generali: https://life.futuregenerali.in/life-insurance-made-simple/life-insurance/how-health-insurance-premium-varies-by-gender

british heart foundation. (n.d.). Retrieved 3 22, 2024, from https://www.bhf.org.uk/what-we-do/news-from-the-bhf/news-archive/2020/august/body-mass-index-is-a-more-powerful-risk-factor-for-diabetes-than-genetics

Enthought. (n.d.). Retrieved 3 18, 2024, from https://www.enthought.com/blog/number-of-clusters/

IBM. (n.d.). Retrieved 3 22, 2024, from https://www.ibm.com/topics/supervised-learning#:~:text=Supervised%20learning%2C%20also%20known%20as%20supervised%20machine%20learning%2C,that%20to%20classify%20data%20or%20predict%20outcomes%20accurately.

IJRTI. (n.d.). Retrieved from https://www.ijrti.org/papers/IJRTI2304248.pdf

NIH. (n.d.). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9843502/

simplilearn. (n.d.). Retrieved from https://www.simplilearn.com/tutorials/machine-learning-tutorial/classification-in-machine-learning