THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

UNIVERSITY OF LONDON

MACHINE LEARNING WITH PYTHON

MODULE : MACHINE LEARNING - ST3189

UOL STUDENT NUMBER : 200643838

NUMBER OF PAGES : 10

(EXCLUDING COVER PAGE, TABLE OF CONTENTS AND REFERENCES)

Table Of Contents

TASK 1: UNSUPERVISED LEARNING

1.1 INTRODUCTION

Unsupervised Learning is a technique used in machine learning for clustering and identifying patterns within unlabeled data. This will help us better understand whether our dataset has any homogeneous groups without the need of human interference. Unsupervised learning is also used for dimensionality reduction to reduce unnecessary data and bring down the number of features to a few relevant ones (Ross, 2022).

The dataset used for this task is the 'Wholesale Customers dataset' obtained from the UCI Machine Learning Repository. This comprises the amounts that clients of a wholesaler spent on products in different regions and channels in Portugal. We aim to perform a dimensionality reduction technique and then identify homogeneous groups of clients (clusters) so that relevant marketing campaigns can be carried out by the wholesaler.

1.2 EXISTING LITERATURE

In a study conducted on the success of different channels in selling wholesale food items in Portugal, it was found that the retail channel has been declining and that the Horeca channel has been doing really well by covering 44% of the total sales (Clemente, 2013). In a research done to gather information as to why Horeca channel is successul, it was identified that customer's spending for Fresh food items is higher in Horeca channels than in Retail channels (Hubeni et al., 2020). A research conducted on the 'Wholesale Customers dataset' Identified eight different homogeneous groups using the K-Means clustering technique (Pokharel et al., 2021).

1.3 RESEARCH QUESTIONS

1. Is the Horeca channel doing better than the Retail channel in terms of sales?
2. Do customers spend more on Fresh food in the Horeca channel than in the retail channel?
3. How many clusters (homogeneous groups) can be identified using this dataset?

1.4 Exploratory Data Analysis (EDA)

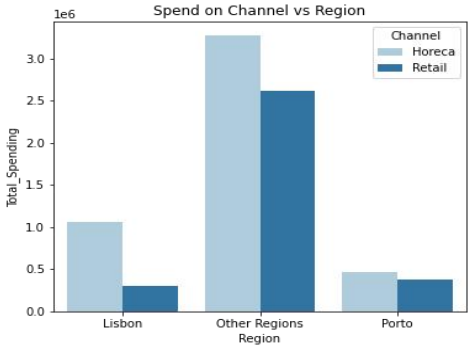


Figure 1. Bar plot

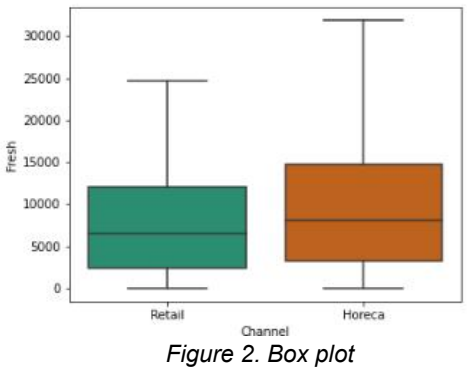


Figure 2. Box plot

From Figure 1, we can see that the bars for the Horeca channel (light blue) is bigger than the bars for the retail channel (dark blue) in all the regions in Portugal. Hence we can say that the amount spent by customers through the Horeca channel is higher than through the retail channel. Let's assume that Total customer spending on items is a proxy for the total sales. This indeed tells us that Horeca channel is doing better than the retail channel in terms of customer sales. It can also be observed that customers haven't spent much for buying wholesale items in Porto but they have spent a very high amount through both Horeca and Retail channels in regions other than Lisbon and Porto in Portugal. As seen from the previous study, we too saw that the Horeca channel is doing better than the Retail Channel in selling more wholesale items.

In Figure 2, the median value of each box is given by the horizontal black line inside each box. We use the median values of amount spent on Fresh food items in both Retail and Horeca channels to observe the differences in amounts spent on Fresh food items in both the channels. It can be seen that the median amount spent on Fresh food is higher in Horeca than in Retail channel and the maximum amount spent on Fresh food is higher in Horeca than in Retail channels. The minimum amount spent is the same for both the channels and takes a value of zero. Hence, we can say that the Horeca channel does sell more Fresh food products than the Retail channel. As seen from the previous study, from the analysis done in our study we too have observed that customers do spend more on fresh food in the Horeca channel than in the retail channel.

1.5 Principal Component Analysis (PCA)

"Principal Component Analysis is a dimensionality reduction method that is used to reduce the dimensionality of large datasets, by converting a large set of variables into smaller variables that still contains most of the information in the large set" (Jaadi, 2022).

The dataset contained 8 columns and 332 rows after removing missing values and potential outliers. Before computing principal components, the variables in the dataset were standardized to ensure unbiasedness of results. Since the dataset contained 8 columns, 8 principal components were constructed.
A cumulative explained variance graph was plotted to find the optimal number of principal components that should be used for the dataset.

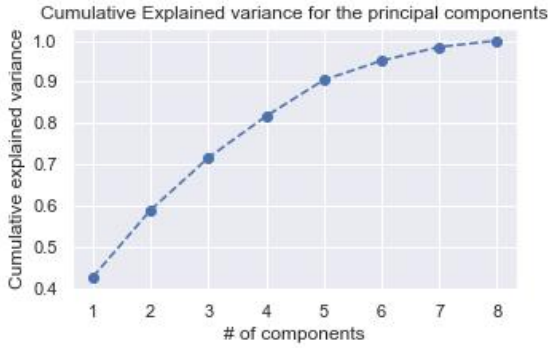


Figure 3. Cumulative explained variance graph

From Figure 3, we can see that the proportion of variance explained by each component decreases as the number of components increases. The marginal change in cumulative explained variance becomes negligibly small after first 6 principal components. Higher the marginal change in the cumulative explained variance, greater the information retained by that principal component. The first six principal components cumulatively explain 95% of the total variation in the dataset which is close to 100%, hence we would choose to use 6 principal components.

It should be noted that each principal component is a linear combination of the original features.
The correlation plot in Figure 4 was plotted to show the level of impact that each original feature has on the six principal components.
Let's assume that variables with an absolute correlation value greater than or equal to 0.45 are the most important contributors,

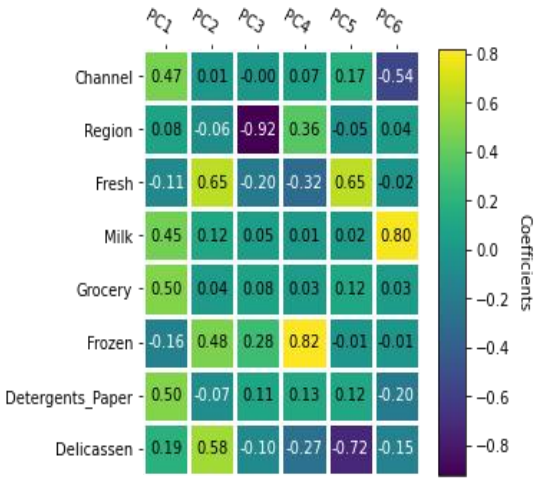


Figure 4. Correlation plot

PC1: It is seen that all the variables except Frozen and Fresh are positively correlated with PC1. Region is the least correlated with PC1 whereas Grocery, Channel, Milk and Detergents_Paper had a moderate positive influence on PC1. Grocery and Detergents_Paper contribute the most to PC1.
PC2: Channel, Region, Grocery and Detergents_Paper have a weak correlation with PC2 whereas Fresh, Frozen and Delicassen has moderate positive influence on PC2. It is seen that Fresh contributes the most to PC2.
PC3: Region has a very strong negative influence on PC3.
PC4: Frozen has a very strong positive influence on PC4.
PC5: Fresh has a moderate positive impact and Delicassen has a strong negative correlation with PC5.
PC6: Channel has a moderate negative impact on PC6 and Milk has a very strong positive influence on PC6.

For the sake of making the interpretation of the principal components easy and simple, we visualise using a 2-dimensional figure.
The Biplot in Figure 5 is a 2 dimensional figure plotted for 2 different principal components and it shows us the impacts of the original variables on each principal component.

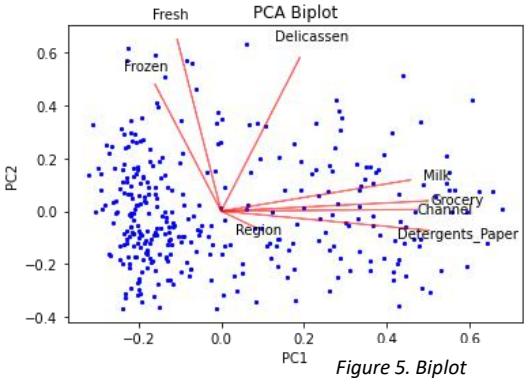


Figure 5. Biplot

The Biplot was plotted against principal components 1 and 2 (PC1 and PC2). It is seen that variability of all the features except Region is highly represented by PC1 and PC2 (long red lines). Features with red lines close to each other (small angles between them) are said to be highly positively correlated with each other. Hence it is observed that Milk, Grocery, Channel and Detergents Paper are highly positively correlated to each other. It also can be seen that Fresh and Frozen are highly positively correlated to each other. Any features with red lines having an angle of intersection equal to 90 degrees are said to be uncorrelated. Therefore, it can be seen that both Fresh and Milk are uncorrelated whereas Delicassen and Region are also seen to be uncorrelated to each other.

1.6 K-MEANS CLUSTERING

"K-means is a centroid-based algorithm, where we calculate the distances to assign a point to a cluster and it's main objective is to minimize the sum of distances between the points and their respective cluster centroid " (Sharma, 2023).

In order to find the optimal number of clusters that can be formed using K-Means , the elbow graph was plotted. The elbow graph makes use of the Within Cluster Sum of Squares (WCSS) to choose the optimal number of clusters that we should go for, based on the place where there is an elbow point in the graph (gupta, 2021).

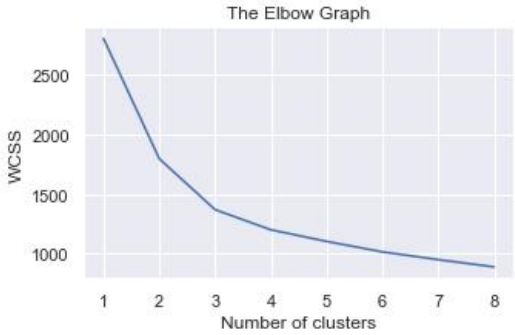


Figure 6. Elbow graph

As seen from the elbow graph in figure 6, it is clearly visible that there is an elbow point at number of clusters being equal to 2. Hence we will perform K-Means using this optimal number of 2 clusters.

K-Means clusters were plotted in a scatter diagram and it was seen that there were clear clusters formed in scatter diagrams between PC1 and the other Principal components (that is PC1 and PC2, PC1 and PC3, PC1 and PC4, PC1 and PC5 , PC1 and PC6).

The scatter plot with clusters between PC1 and PC2 is given below to show why 2 clusters is the optimal number.

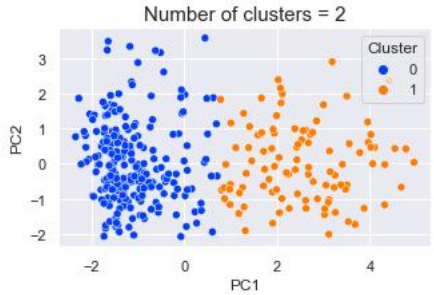


Figure 7. Scatter plot for 2 clusters

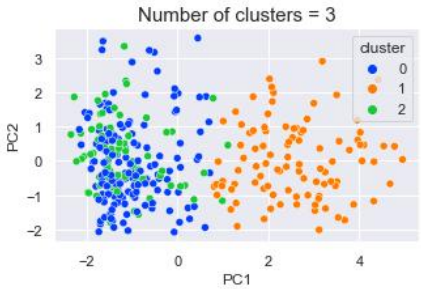


Figure 8. Scatter plot for 3 clusters

As seen from Figure 7, the clusters are linearly separable when 2 clusters are used in K-Means. But when we used visualised using 3 clusters in Figure 8, we could see that the clusters started overlapping. Hence the optimal number of clusters is two with 231 observations in Cluster '0' (green) and 101 observations in Cluster '1'. Therefore 2 clusters is enough for identifying 2 homogeneous groups for the Wholesaler's clients. Hence our results seem to different compared to existing literature since in the previous study, eight clusters were found without using PCA. Since PCA was used in our study, we were able to find only 2 homogeneous clusters.

TASK 2: REGRESSION

2.1 INTRODUCTION

Supervised Learning is the method of training of algorithms using input data that is labelled for a specific output and this trained algorithm is then used to predict continuous values or classify data points (Petersson, 2021).
"Regression is a supervised learning method that identifies whether some independent variables and a dependent variable are related ,and if so what the relationship is" (Castillo, 2023).
The dataset used for this task is the 'Boston housing prices' dataset obtained from the Kaggle platform. The data was collected in 1978 and it contains information about house features and the median house prices. We aim to predict the median value of house prices using regression in this task.

2.2 EXISTING LITERATURE

In a research done to find the impact of house features on the final house price found that the number of rooms had a significant impact on the final house price (Owusu-Manu et al., 2019). In a research done previously on the Boston dataset, it was found that the house prices were most notably affected by percentage of people in the population who have a lower socio-economic status (Khosravi et al., 2022). In an earlier study done on Boston dataset, it was found that random forest was a better model than linear regression and decision tree in predicting house prices, with a root mean squared error of 2.901 (Begum et al., 2022).

2.3 RESEARCH QUESTIONS

1. Does the number of rooms have an impact on housing prices?
2. Do the lower status of the population have the biggest influence on housing prices?
3. What is the best regression model to predict house prices?

2.4 Exploratory Data Analysis (EDA)

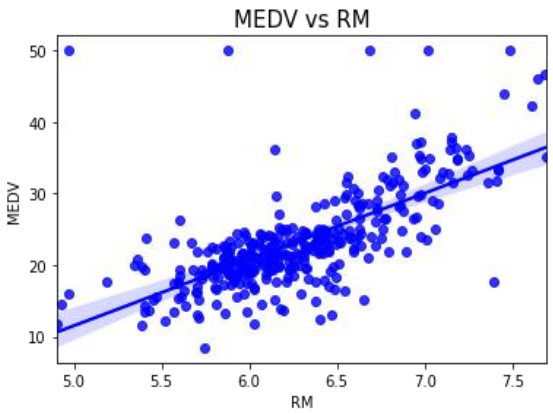
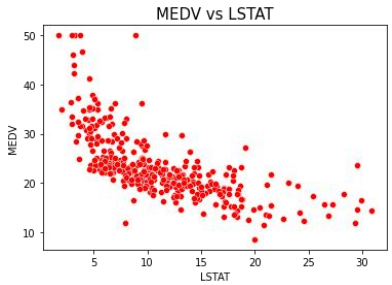


Figure 9. Scatter plot between MEDV and RM

A scatter plot was plotted to answer the first research question. The scatter plot depicted in Figure 9 shows us that there is kind of a linear relationship between average number of rooms per dwelling and the house price. As the number of rooms increased, the house prices also increased in a linear manner.The correlation coefficient between these variables was found to be +0.68 which indicates that there is a moderate positive linear relationship between average number of rooms and house price.

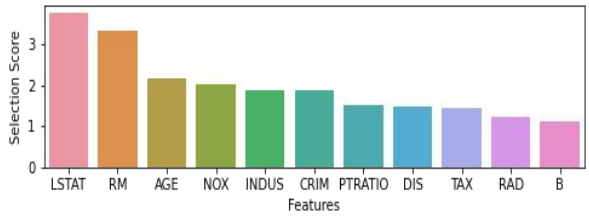


From Figure 10, we can see that as LSTAT (% lower status of population) increases, MEDV (median house price) decreases gradually in not a very linear way. This tells us that the lower status of population has a negative relationship with the house prices. The correlation coefficient between MEDV and LSTAT was found to be -0.68 which indicates that there is a moderate negative linear relationship between lower status of population and house price.

Figure 10. Scatter plot between MEDV and LSTAT

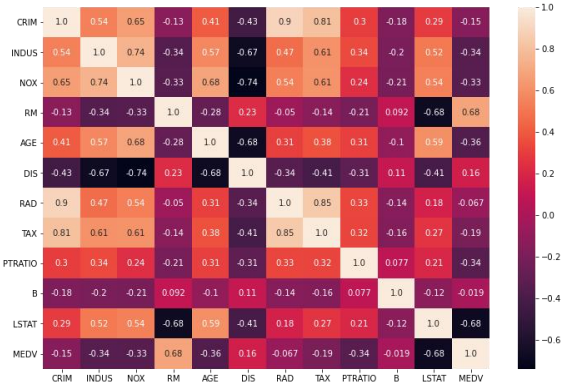
2.5 FEATURE SELECTION

The two variables named CHAS and ZN were dropped in the data cleaning process since they took ambiguous values. The best features that impact the target variable 'MEDV' (housing price) from the remaining features was found using selectKbest method in python.



The bar plot in Figure 11 shows us the selection score for each feature ,where each selection score tells you the impact that each feature has on the target variable 'MEDV'. Hence it is seen that LSTAT has the highest impact on MEDV and B has the least impact on MEDV.

Figure 11. Feature selection bar plot



A correlation of greater than or equal to 0.8 was assumed to be a high correlation. It was found from the correlation plot in Figure 12 that CRIM and RAD, RAD and TAX , TAX and RAD were the variables that were highly correlated. Since we should try to avoid multicollinearity (high correlation between predictor variables), the least important features 'TAX', 'RAD' and 'B' were dropped before modelling.

Figure 12. Correlation plot

2.6 REGRESSION MODELS

The dataset was divided into train and test sets with a split ratio of 80%-20% respectively. The purpose of this was to train the models using the train set and evaluate the model performance on the test set.

MULTIPLE LINEAR REGRESSION

It is essential that we check whether our dataset follows the five basic assumptions of the linear regression model before fitting the model.

The assumptions for a linear model are:

- There is a linear association between target and predictor variables.
- Error terms have a constant variance (homoscedastic).
- Error terms are independent (No Autocorrelation).
- Error terms are normally distributed.
- No perfect multicollinearity between predictors (No correlation of +1 or -1 between predictors)

The below graphs and statistics are used to check for any linearity assumption violations.

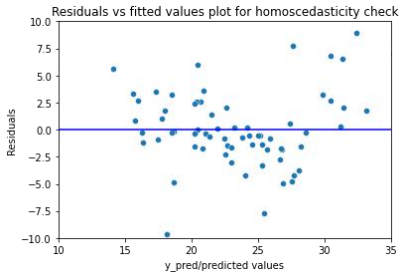


Figure 13. Residual v fitted plot

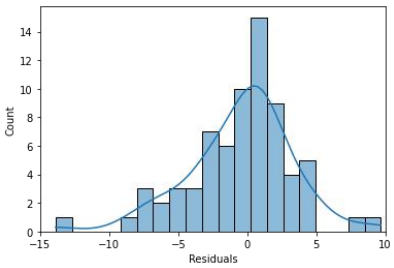


Figure 14. Histogram of the error terms

Figure 13 plots the residuals against the predicted values. Since the blue line is centered at zero, this indicates that there is approximately zero bias between the errors and hence the homoscedasticity assumption is satisfied. From the histogram in Figure 14, it is visible that the errors seem to be somewhat normally distributed hence this assumption too holds.

Scatter plots were plotted to visualise the impact of the predictor variables on the target variable 'MEDV' and the predictors were seen to have a significant linear relationship with 'MEDV', hence the linearity assumption holds. From the correlation plot in Figure 12, it was found that there was no correlation between predictors with a value of +1 or -1, hence 'no perfect multicollinearity' assumption is valid.

To check whether the autocorrelation assumption is valid, we conduct a Durbin Watson hypothesis test. The null and alternative Hypothesis are as follows,

H0: Model does not suffer from autocorrelation
H1: Model suffers from autocorrelation

The python code durbin_watson() was used to obtain the durbin watson test statistic value of **1.701.** Since this value is between 1.5 and 2.5, we fail to reject H0 at 5% significance level and conclude that the statistical results suggest that the model does not suffer from autocorrelation (*How to perform a Durbin-Watson Test in python* 2021).

Therefore all the linear assumptions are satisified by the multiple linear regression model and hence the multipl linear regression model can be used to predict house prices.

Other Regression Models

Regression models such as ***Polynomial Regressor of degree 2***, ***Random Forest Regressor***, ***Decision Tree Regressor*** and ***XGBoost Regressor*** were also used so that we compare and evaluate all the regression models to come up with the best regression model that will predict the house prices.

The following table shows the evaluation metrics such as R squared, MSE, RMSE and MAE of the models with their best hyper parameters. The **R squared** value tells us the % of variation in the target variables that is explained by the predictor variables hence we will ideally want the R squared to be as high as possible. The Mean Squared Error (**MSE**) , Root Mean Squared Error (**RMSE**) and Mean Absolute Error (**MAE**) are loss functions and we would ideally want it to be as low as possible.

	R squared	MSE	RMSE	MAE
Polynomial Regression(degree 2)	0.8298	8.404	2.899	2.040
Random Forest	0.8121	9.281	3.047	2.235
XGBoost	0.7918	10.28	3.207	2.341
Decision Tree	0.7291	13.38	3.658	2.475
Multiple Linear Regression	0.7034	14.65	3.827	2.747

The Polynomial Regression model (degree 2) performed the best in predicting the house prices when compared to the other four regression models since it has the highest R squared of 0.8298 and lowest values for all loss functions. It is also identified that multiple linear regression performed the least well out of all the models, with a R squared value of 0.7034.

2.7 CONCLUSION

- As seen from the previous study, we too saw in this study that the number of rooms per dwelling was found to have an impact with the house prices. In fact, the number of rooms per dwelling was seen to have a positive linear relationship with the house prices.

- As seen from the previous study, we too saw in this study that percentage of people in the population with a lower socio-economic status had the highest influence on the house prices with the help of the feature selection bar plot.

- The best regression model for predicting house prices was found to be the polynomial regression model (degree 2) with a RMSE value of 2.899. This RMSE value was lower than that of the best performing random forest model that was found in a previous study by (Begum et al., 2022) using the same Boston dataset.

TASK 3: CLASSIFICATION

3.1 INTRODUCTION

"Classification is a supervised learning method that is used to categorise a given set of input data into classes based on one or more variables" (Ramakrishnan, 2023).

The dataset used for this task is the 'Pima Indians Diabetes' dataset obtained from the Kaggle platform. This dataset is about the North American Indians and contains certain measurements of them along with whether they are diabetic or not. The final goal of this task is to classify whether a person has diabetes or not.

3.2 EXISTING LITERATURE

A research done on the likelihood of hypertension in diabetic and non-diabetic people found that diabetic individuals are twice as probable to develop hypertension when compared to those without diabetes (Petrie et al., 2018). Petrie et al. also stated that patients with hypertension are more likely to have diabetes as they often are insulin resistant. A previous study conducted on the risk of BMI on diabetes found that people with higher BMI values are more prone to be diabetic (Gupta & Bansal, 2020). A research done on the Pima Indians dataset found that out of 5 classification machine learning models created, Deep Neural Network (DNN) model was the best performing model with an accuracy of 77.86% (Wei et al., 2018).

3.3 RESEARCH QUESTIONS

1. Does hypertension have an effect on diabetic and non-diabetic people?

2. Does a person's BMI have an effect on the person having diabetes?

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

3.4 Exploratory Data Analysis (EDA)

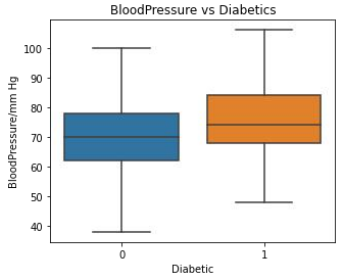


Figure 15

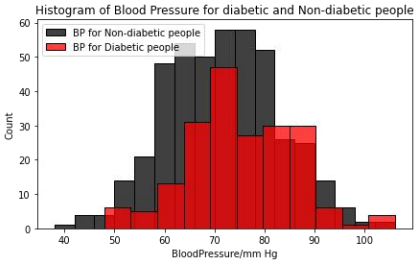


Figure 16

From the side by side box plot in Figure 15, it can be seen that the median (the line in the middle of the boxes) value for blood pressure is higher for a person with diabetics (Diabetic=1) than a person who is non-diabetic (Diabetic=0). If the blood pressure is greater than 100 mm Hg, it is highly likely that the person has diabetes since the maximum of blood pressure for non-diabetic people ends at 100 mm Hg but the maximum for diabetic people is greater than that. The histogram in Figure 16 too shows that the distribution of blood pressure for Diabetic people is skewed to the right when compared to that for non-diabetic people. Hence it is quite evident that people with diabetes are more likely to have hypertension (high blood pressure).

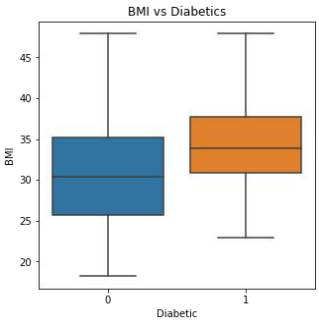


Figure 17

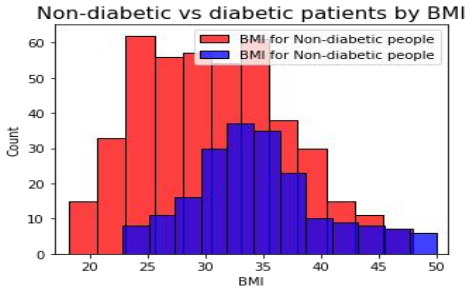


Figure 18

The above side by side box plot in Figure 17 shows that the median BMI value for diabetic people is greater than that for non-diabetic people and that people with BMI values of less than 23 are more likely to be non-diabetic. The above histogram in Figure 18 shows that the distribution of BMI for diabetic people takes higher values for BMI when compared with that for non-diabetic people. Hence it is quite evident from the graphs that people with higher BMI have a greater chance of being diabetic.

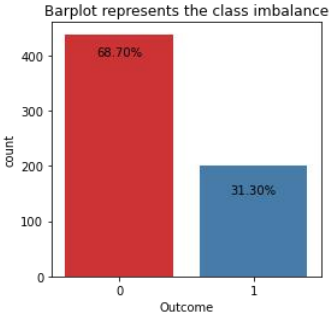


Figure 19

Figure 19 shows the counts of the target variable 'Outcome'. 'Outcome' has been classified into two - "0" if the person is non-diabetic and "1" if the person has diabetes. Through this bar plot, we can see that there is a huge mismatch between the number of people who had diabetes and did not have diabetes. The number of non-diabetic people(439) are greater than the number of diabetic people(200) in the dataset. Hence it is evident that the dataset is imbalanced since the dataset contains 68.7% non-diabetic people and only 31.30% diabetic people.

3.5 FEATURE SELECTION

All features except for SkinThickness were taken for as predictor variables for the models since 'SkinThickness' had a very low impact on the target variable 'Outcome'. The impact was found through the selectkbest feature selection method in Python. It was also found that 'Glucose' had the greatest impact on the target variable 'Outcome'. It was observed that none of the predictor variables were very strongly linearly related using the correlation plot. The highest correlation of +0.57 was between 'Age' and 'Pregnancies' but since this isn't very close to 1, these variables were not dropped.

3.6 CLASSIFICATION MODELS

The following models were used for this task,

- ***Decision Tree :-*** A decision tree classifier ,also known as classification tree, is used when the target variable is categorical in nature. It predicts the class of the target variables using simple decision rules that are obtained from the features.

- ***Logistic Regression:-*** Logistic Regression is a classification model that uses multiple predictor variables to predict a binary response. The binary response in this dataset are the values 0 and 1 that the target variable "Outcome" takes. This technique helps to directly get the probabilities of classifying a patient being diabetic and non-diabetic.

- ***Random Forest:-*** Random Forest is made up of various decision trees and it trains the sub decision trees in such a way that the resulting predictions are less correlated (Brownlee, 2021).

- ***XGBoost:-*** XGBoost is an efficient gradient boosting tree algorithm that trains various decision trees and tries to predict the output variable by getting estimates of weak models (Mishra, 2019).

The evaluation metrics used for the model comparison are stated below,

- **Accuracy** - A measure that tells you how close the model's predictions are to the true values.

- **Precision**- A measure that tells how well the model predicts a particular category (Dang, 2022)

- **Recall**- A measure that indicates the ability of a model to detect the cases that are predicted positive and are actually positive.

- **F1 score**- It involves both precision and recall combined to judge the model's performance.

- **Confusion Matrix**- A matrix that is used to visualise the actual and predicted classes.

- **ROC curve**- "A probability curve that is plotted for binary classification problems and the area under the curve tells us how well the model is at differentiating between the positive and negative classes" (Bhandari, 2023).

From the following equations, the accuracy , precision, recall and F1 scores can be calculated directly from the confusion matrix,

$$\textbf{Accuracy} = \frac{\text{Total number of correctly predicted cases}}{\text{Total number of cases}} \qquad \textbf{Precision} = \frac{\text{Number of correctly predicted positive cases}}{\text{Total number of predicted positive cases}}$$

$$\textbf{Recall} = \frac{\text{Number of correctly predicted positive cases}}{\text{Total number of actual positive cases}} \qquad \text{" } \textbf{F1 score} = \frac{2 \times \textbf{Precision} \times \textbf{Recall}}{\textbf{Precision} + \textbf{Recall}} \text{ " (Brownlee, 2019)}$$

We would ideally want all these measures to be as high as possible.

Before fitting the models, the dataset was divided into train and test datasets. Since we found that the data was imbalanced, observations in the training set were sampled using a randomization process to get a balanced training set (equal number of 0s and 1s of target variable). This was done to ensure to ensure that the model will have a 50% chance of predicting each class and to avoid class bias.

The following table shows the evaluation metrics for the models. It should be noted that the evaluation metrics shown in the table are for models after they were tuned for the best hyperparameters.
The confusion matrix in the table assumes that negative means the target variable "Outcome" being equal to 0 and positive means "Outcome" being equal to 1.

The total number of cases considered in the confusion matrices is 128.

MODEL	ACCURACY	PRECISION	RECALL	F1 SCORE	CONFUSION MATRIX
Decision Tree	74.2%	65.1%	60.9%	62.9%	Predicted Negative / Predicted Positive Actual Negative: 67 / 15 Actual Positive: 18 / 28
Logistic Regression	78.1%	67.3%	76.1%	71.4%	Predicted Negative / Predicted Positive Actual Negative: 65 / 17 Actual Positive: 11 / 35
Random Forest	79.7%	75%	65.2%	69.8%	Predicted Negative / Predicted Positive Actual Negative: 72 / 10 Actual Positive: 16 / 30
XGBoost	80.5%	78.4%	63.0%	69.9%	Predicted Negative / Predicted Positive Actual Negative: 74 / 8 Actual Positive: 17 / 29

Since the dataset is imbalanced, there is no point in comparing the models in terms of their accuracies since accuracies will be misleading for imbalanced datasets. Hence we will compare the models in terms of F1 score and the ROC curves to determine the best model for classifying diabetes. Since F1 score gives you a balance between recall and precision, F1 score will be used to compare the models (and not the precision and recall).
As observed from the above table, F1 score is the highest for the Logistic Regression Model and has a value of 71.4%. The next best was XGBoost, followed by Random Forest and Decision Tree.

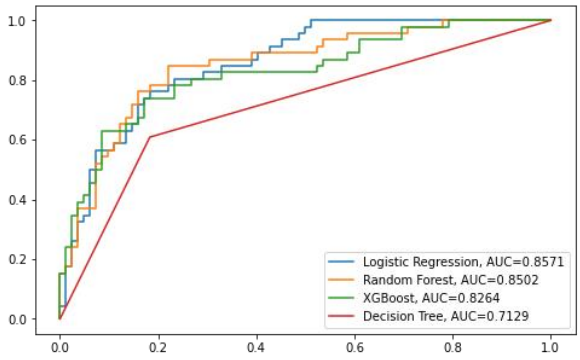


Figure 20

The ROC curves for the models are shown in Figure 20. The model with a ROC curve that is more closer to the top left corner is said to be the best model in differentiating between diabetic and non-diabetic people (Nahm, 2022). The AUC (Area Under the Curve) can be used to check this and we see from the legend that the AUC of the ROC curve of the Logistic Regression model is 0.8571 and is the highest. Therefore since the Logistic Regression model has the highest F1 score and highest AUC in the ROC curve, Logistic Regression is the best model in classifying diabetes.

3.7 CONCLUSION

- As seen from previous study, we too saw in this study that hypertension had an impact on the diabetic and non-diabetic people. It was evident that people with diabetes are more likely to have hypertension than non-diabetic people.
- As seen in the previous study, we too saw in this study that a person's BMI had an influence on the person having diabetes. People with a higher BMI were found to be more likely to have diabetes.
- The best classification model for classifying people with diabetes and non-diabetic people was found to be the Logistic Regression model with an accuracy of 78.1%. This accuracy was much higher than that of the best performing Deep Neural Network model that was found in the previous study by (Wei et al., 2018) using the same Pima Indians Diabetes dataset.

4.0 References

Begum, A., Kheya, N.J. and Rahman, M.Z. (2022) “Housing price prediction with machine learning,” *International Journal of Innovative Technology and Exploring Engineering*, 11(3), pp. 42–46. Available at: https://doi.org/10.35940/ijitee.c9741.0111322.

Bhandari, A. (2023) *Guide to AUC ROC curve in machine learning : What is specificity?*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/#What_are_Sensitivity_and_Specificity? (Accessed: March 24, 2023).

Brownlee, J. (2019) *Classification accuracy is not enough: More performance measures you can use*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use/#:~:text=F1%20Score,the%20precision%20and%20the%20recall. (Accessed: March 27, 2023).

Brownlee, J. (2021) *How to develop a random forest ensemble in Python*, *MachineLearningMastery.com*. Available at: https://machinelearningmastery.com/random-forest-ensemble-in-python/ (Accessed: March 24, 2023).

Castillo, D. (2023) *Machine learning regression explained*, *Seldon*. Available at: https://www.seldon.io/machine-learning-regression-explained#:~:text=Machine%20Learning%20Regression%20is%20a,used%20to%20predict%20continuous%20outcomes. (Accessed: March 21, 2023).

Clemente, J.P.de A.A.F. (2013) *Recheio: Growing through a declining channel*. Available at: https://repositorio.ucp.pt/bitstream/10400.14/15674/1/Complete%20thesis%20_Jo%C3%A3o%20Clemente_16-09_final.pdf (Accessed: March 25, 2023).

Dang, T. (2022) *Guide to accuracy, precision, and recall*, *Mage*. Available at: https://www.mage.ai/blog/definitive-guide-to-accuracy-precision-recall-for-product-developers (Accessed: March 24, 2023).

gupta, R.K. (2021) *K -mean clustering*, *Medium*. Medium. Available at: https://10012000rahulgupta.medium.com/k-mean-clustering-a87b2d4dbabd (Accessed: March 23, 2023).

Gupta, S. and Bansal, S. (2020) “Does a rise in BMI cause an increased risk of diabetes?: Evidence from India,” *PLOS ONE*, 15(4). Available at: https://doi.org/10.1371/journal.pone.0229716.

Hubeni, Y. *et al.* (2020) “Globalization and local determinants of Horeca customers market behavior in the Wholesale Food Market,” *Zeszyty Naukowe SGGW w Warszawie - Problemy Rolnictwa Światowego*, 20(1), pp. 25–39. Available at: https://doi.org/10.22630/prs.2020.20.1.3.

Jaadi, Z. (2022) *A step-by-step explanation of principal component analysis (PCA)*, *Built In*. Available at: https://builtin.com/data-science/step-step-explanation-principal-component-analysis (Accessed: March 25, 2023).

Khosravi, M. *et al.* (2022) “Performance evaluation of machine learning regressors for estimating real estate house prices.” Available at: https://doi.org/10.20944/preprints202209.0341.v1.

Mishra, A. (2019) *Machine learning in the AWS cloud: Add intellegence to applications with Amazon Sagemaker and Amazon Rekognition*, *Amazon*. Sybex, a Wiley brand. Available at: https://docs.aws.amazon.com/sagemaker/latest/dg/xgboost-HowItWorks.html (Accessed: March 24, 2023).

Nahm, F.S. (2022) “Receiver operating characteristic curve: Overview and practical use for clinicians,” *Korean Journal of Anesthesiology*, 75(1), pp. 25–36. Available at: https://doi.org/10.4097/kja.21209.

Owusu-Manu, D.-G. *et al.* (2019) “Housing attributes and relative house prices in Ghana,” *International Journal of Building Pathology and Adaptation*, 37(5), pp. 733–746. Available at: https://doi.org/10.1108/ijbpa-01-2019-0003.

Petersson, D. (2021) *What is supervised learning?*, *Enterprise AI*. TechTarget. Available at: https://www.techtarget.com/searchenterpriseai/definition/supervised-learning (Accessed: March 21, 2023).

Petrie, J.R., Guzik, T.J. and Touyz, R.M. (2018) “Diabetes, hypertension, and cardiovascular disease: Clinical insights and vascular mechanisms,” *Canadian Journal of Cardiology*, 34(5), pp. 575–584. Available at: https://doi.org/10.1016/j.cjca.2017.12.005.

Pokharel, M., Bhatta, J. and Paudel, N. (2021) “Comparative analysis of K-means and enhanced K-means algorithms for clustering,” *NUTA Journal*, 8(1-2), pp. 79–87. Available at: https://doi.org/10.3126/nutaj.v8i1-2.44044.

Ramakrishnan, M. (2023) *What is classification in Machine Learning and why is it important?*, *Emeritus Online Courses*. Available at: https://emeritus.org/blog/artificial-intelligence-and-machine-learning-classification-in-machine-learning/#:~:text=In%20machine%20learning%2C%20classification%20is,datasets%20of%20input%20and%20output. (Accessed: March 23, 2023).

Ross, A. (2022) *What is unsupervised learning?*, *Unsupervised*. Available at: https://unsupervised.com/resources/blogs/what-is-unsupervised-learning/ (Accessed: March 22, 2023).

Sharma, P. (2023) *The Ultimate Guide to K-means clustering: Definition, methods and applications*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/#:~:text=K%2Dmeans%20clustering%20is%20a,assigned%20cluster%20mean%20is%20minimized. (Accessed: March 26, 2023).

Wei, S., Zhao, X. and Miao, C. (2018) “A comprehensive exploration to the machine learning techniques for diabetes identification,” *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)* [Preprint]. Available at: https://doi.org/10.1109/wf-iot.2018.8355130.

Zach (2021) *How to perform a Durbin-Watson Test in python*, *Statology*. Available at: https://www.statology.org/durbin-watson-test-python/ (Accessed: March 21, 2023).