UNIVERSITY OF LONDON

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

MACHINE LEARNING – ST3189

UOL STUDENT NUMBER: 210468113

TABLE OF CONTENTS

1. UNSUPERVISED LEARNING

a) INTRODUCTION

Unsupervised learning is a powerful tool for extracting valuable insights from unlabelled data through discovering patterns, structures or groups in the data automatically without the need for human interference. By using techniques such as clustering, dimensionality reduction or association analysis, businesses can obtain new insights, reveal hidden information and improve decision making (Patrick, 2023).

The “Wholesale customers” dataset from the UCI Machine Learning Repository was used for this task. It includes the amounts that clients in Portugal spend on various items in different regions through different channels. The objective is to conduct a dimensionality reduction technique and then identify clusters where elements within the clusters are homogeneous (characteristics of each element are similar) and elements between clusters are heterogeneous (different between each cluster).

b) EXISTING LITERATURE

According to a study on the success of various channels for selling wholesale food items in Portugal, the Horeca channel has been doing well, accounting for 44% of total sales, whereas the retail channel has been declining (Clemente, 2013). Another study discovered that a large proportion of milk products are sold through retail channels (Popovic, 2009). Research conducted on this dataset revealed that the ideal number of clusters needed was four using the K-means clustering technique (Pokharel, Bhatta, & Paudel, 2021).

c) RESEARCH QUESTIONS

1) In terms of total spendings, which channel (Retail or Horeca) is doing better?
2) How does spending on milk vary between the two channels?
3) Using this dataset, how many homogeneous clusters can be identified?

d) EXPLORATORY DATA ANALYSIS

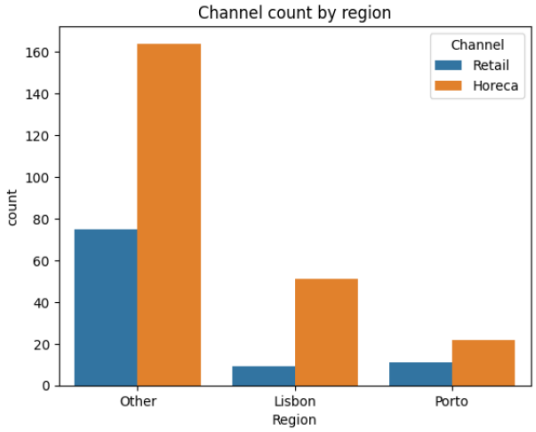

Figure 1.1

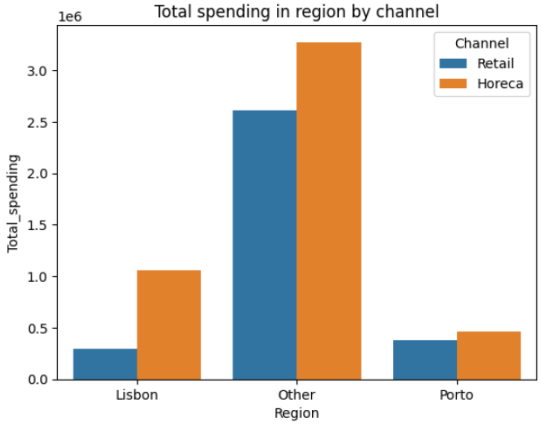

Figure 1.2

From figure 1.1, we can see that Horeca is the most frequently used channel in all the regions. In addition to that, Figure 1.2 shows that total spending through Horeca channel is the highest among all the regions. It can also be noted that people tend to spend less in Porto and the Retail channel is not preferred in Lisbon when compared to Horeca. As seen from the previous study too, Horeca is the better performing channel than Retail in terms of total spendings on wholesale items.

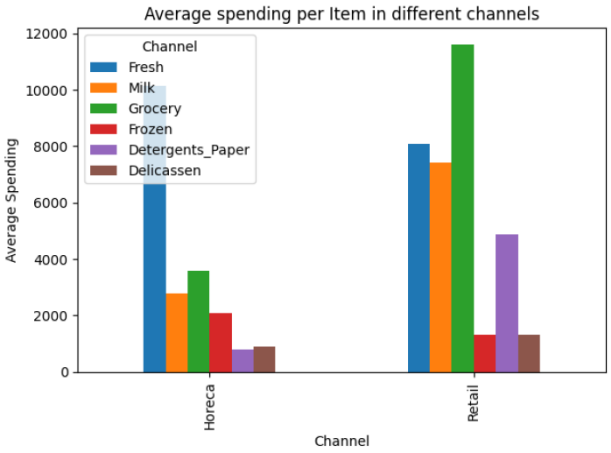

Figure 1.3

It can be easily observed that the average spending on Milk is significantly higher in the Retail channel than the Horeca channel. Therefore, as seen from the previous study, clients spend more on Milk through Retail channel.

e) PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is a dimensionality reduction technique used to reduce the number of dimensions in a dataset by extracting the principal components containing the most information and preserving all the essential details (BasuMallick, 2023).

After cleaning, the dataset contained 332 entries and then the variables were standardised to eliminate bias in the results. As there are 8 columns, 8 different principal components was computed. A cumulative explained variance ratio graph was plotted to find the optimal number of components needed for this analysis.

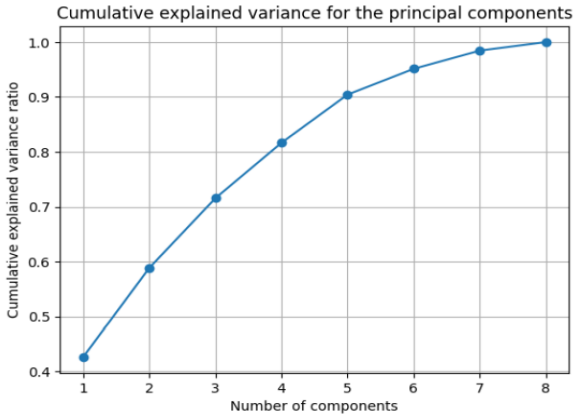

Figure 1.4

From figure 1.4, we can observe that as the number of components increases, the proportion of variance explained by each component reduces. The slope of the curve rises sharply but slowly falls and after the 6th component, the curve begins to flatten. The first 6 components explain about 95% of the total variation in the dataset and hence we choose 6 components for further analysis.

A correlation plot (figure 1.5) was formed to show the association of each feature with every principal component. Absolute value of the coefficients greater than 0.5 are considered significant.

PC1: All the features except Fresh and Frozen are positively correlated with PC1. Region has negligible correlation whereas Channel, Milk, Grocery and Detergents_Paper have a moderate positive correlation with PC1.

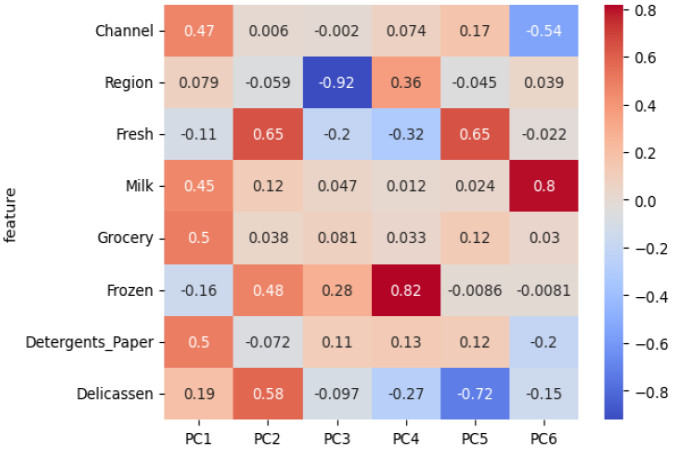

Figure 1.5

PC2: Fresh and Delicassen has a reasonable positive influence on PC2 whereas the rest of the features do not have a significant correlation.
PC3: Region is very strongly negative correlated with PC3.
PC4: Frozen is very strongly positive correlated with PC4.
PC5: Fresh has a moderately positive correlation with PC5 whereas Delicassen has a moderately negative correlation with PC5.
PC6: Milk is strongly positive correlated with PC6 although Channel has a reasonably negative impact on PC6.

f) K-MEANS CLUSTERING

"K-means clustering is a method used for clustering analysis, especially in data mining and statistics. It aims to partition a set of observations into a number of clusters (k). It can be considered a method of finding out which group a certain object really belongs to." (Sharma, 2023)

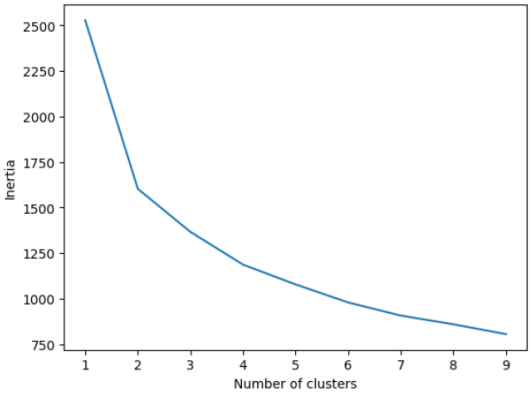

Figure 1.6: Elbow curve

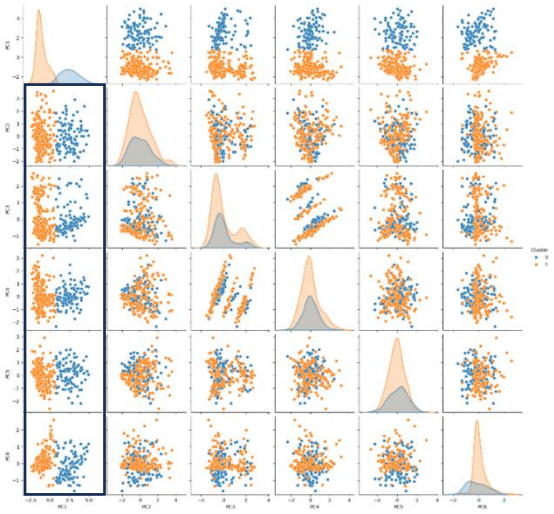

Figure 1.7: Scatter plot

In order to find the optimal number of clusters, an elbow graph is plotted where we select the point at which there is a significant deviation in the direction of the graph. From figure 1.6, we can observe that at 2 clusters, the curve bends significantly like an elbow. Hence, we go ahead with 2 clusters. This contradicts the existing literature in the previous study as 4 clusters was needed. Clusters were formed and plotted in a scatter graph and it was quite evident from figure 1.7 that there were clear clusters formed between PC1 and the other principal components (highlighted).

g) HIERACHICAL CLUSTERING

"In this technique, each data point is considered as a cluster initially. At each level, similar clusters combine with other clusters until a single cluster or K clusters are formed." (Patlolla, 2018). A dendrogram is plotted to show how the clusters are formed in the best possible way. We can observe that the Euclidean distance to form the final single cluster is significantly larger than at 2 clusters. Therefore, we choose a cutoff point at distance 30.

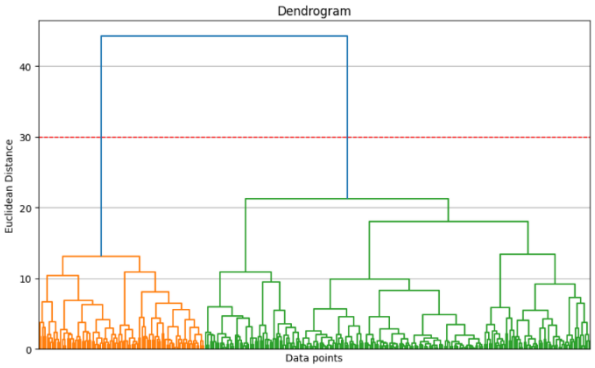

Figure 1.8

2. REGRESSION

a) INTRODUCTION

Regression is a supervised learning technique used to find a relationship between a metric dependent (target) variable and several independent (predictor) variables.

The CO_2 emissions dataset from Kaggle was chosen for this task. The dataset contains information of various types of vehicles in Canada over a period of 7 years. There are 11 features and 7385 entries and our goal is to predict the level of CO_2 emitted from a vehicle through regression. It is measured in grams per kilometre travelled (g/km).

b) EXISTING LITERATURE

A study in 2007 found out that heavy duty diesel vehicles produced the highest level of CO_2 followed by light duty diesel vehicles. In addition to that, light duty gasoline vehicles released a lower amount of CO_2 (Nilrit, Sampanpanish, & Bualert, 2017). In research conducted to find a relationship between vehicle fuel consumption and CO_2 emissions, it was found that gasoline consumption and diesel consumption has a positive linear relationship with CO_2 emissions (Pinto & Oliver-Hoyo, 2008).

c) RESEARCH QUESTIONS

1) Does diesel vehicles produce the highest level of CO_2 when compared to gasoline vehicles?
2) On average, which vehicle class produces the highest level of CO_2?
3) Does gasoline and diesel have a positive linear relationship with CO_2 emissions?
4) What is the best regression model to predict CO_2 emissions?

d) EXPLORATORY DATA ANALYSIS

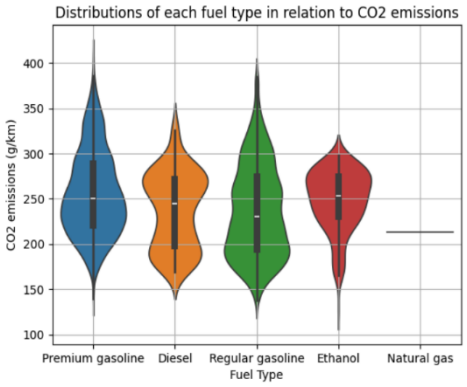

Figure 2.1

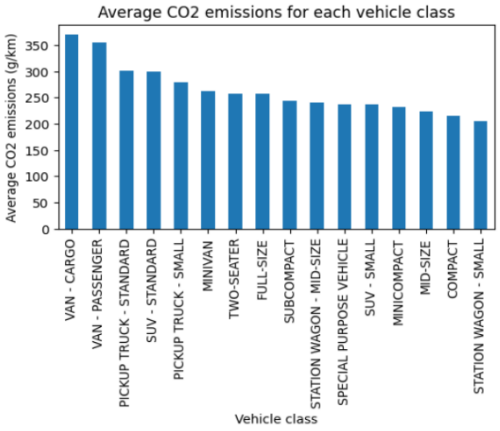

Figure 2.2

Figure 2.1 shows violin plots for each fuel type in relation to the level of CO_2 emissions. We can see that in terms of median CO_2 emissions:

Ethanol > Premium gasoline > Diesel > Regular gasoline > Natural gas

This contradicts the existing study partially which found out that gasoline vehicles released less CO_2 than diesel vehicles.

Figure 2.2 illustrates the average levels of CO_2 emitted for each vehicle class. We can clearly see that Cargo vans and Passenger vans produce the highest levels (greater than 350g/km). Smaller vehicles tend to produce less CO_2 as seen in the graph.

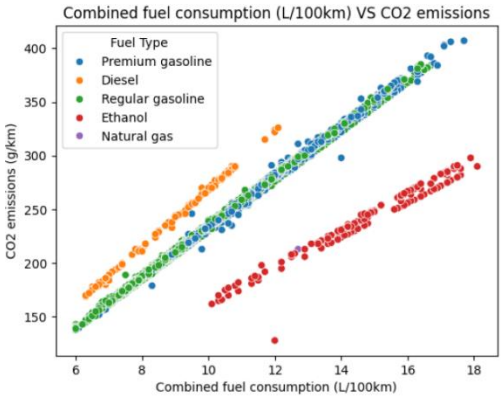

Figure 2.3

Figure 2.3 is a scatterplot showing the relationship between combined fuel consumption and CO_2 emissions for each fuel type. We can observe that for both Premium and Regular gasoline, there is a very similar positive correlation between fuel consumption and CO_2 emissions. When compared to gasoline engines, Diesel engines release more CO_2 since the slope for Diesel engines is steeper. As seen from the previous study, both gasoline and diesel consumption has a positive linear relationship with CO_2 emissions.

e) FEATURE SELECTION

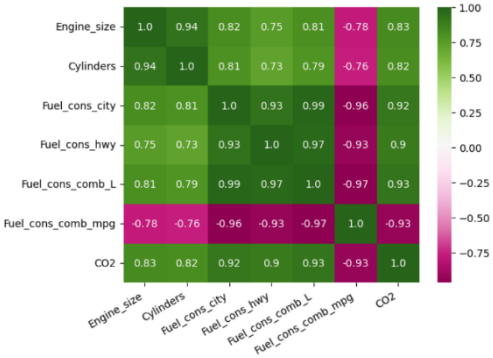

Figure 2.4: Correlation heatmap

The correlation heatmap shows that multicollinearity exists among all the features therefore it is not possible to select the best features using the correlations. To select features that influence the target variable the most, we use the K-best feature selection method. In this case, all the features have been used.

f) REGRESSION MODELS

The variables were split into train and test sets with a split ratio of 75%-25% respectively and then standardised. Various regression models were trained and tested to see which model performed the best.

- MULTIPLE LINEAR REGRESSION

It is crucial that we check whether our dataset holds the basic assumptions of linear regression. The assumptions are:

1) Error terms have a constant variance (homoscedastic)
2) Error terms are normally distributed with mean zero and constant variance

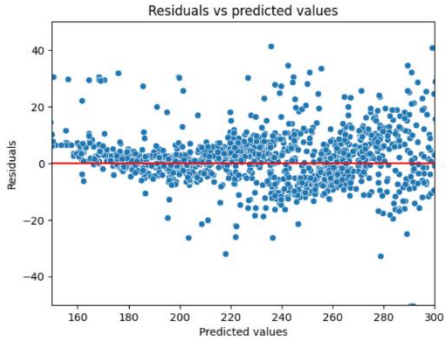

Figure 2.5

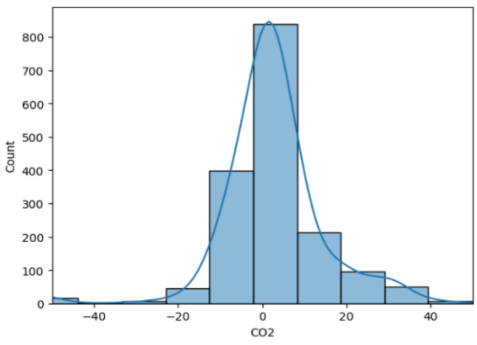

Figure 2.6

Assumption 1: Figure 2.5 plots the residuals against the predicted values. On either side of the red line at zero, there seems to be no bias between the errors therefore the homoscedastic assumption holds.

Assumption 2: Figure 2.6 plots the distribution of the error terms and it is somewhat normally distributed with zero mean therefore this assumption holds.

Since both the assumptions are satisfied by the multiple linear regression model, regression can be used to predict the level of CO_2 emissions from vehicles.

- OTHER REGRESSION MODELS

Several regression models have been implemented and evaluated. Models were tuned using hyperparameters and then results were compared to identify the best model for this dataset. In regression, the **R-Squared** value is considered to be the most important measure as it determines the % of the variance in the dependent (target) variable that is explained by the independent (predictor) variables. The Mean Squared Error (**MSE**), Mean Absolute Error (**MAE**) and Root Mean Squared Error (**RMSE**) are loss functions hence the values should be as low as possible. The values obtained for each model is shown in the table below.

Model	R-Squared	MSE	MAE	RMSE
Multiple linear regression	0.9043	232.0008	9.2514	15.2316
Decision tree regressor	0.9616	93.0875	4.0603	9.6482
Random forest regressor	0.9698	73.1886	3.7043	8.5550
Polynomial regression with degree 2	0.9401	145.1122	6.0885	12.0463
Gradient boosting regressor	0.9639	87.4298	3.9563	9.3504

The Random Forest regressor achieved the best scores out of all the models when predicting the level of CO_2 released from vehicles. The R-Squared value was 0.9698 which is the highest and the values of the loss functions were the lowest among all models. Random Forest regressor is the best model. Multiple linear regression was the worst performing model with a comparatively low R-Squared value and a very high MSE value.

g) CONCLUSION

- It was found that although regular gasoline engines released less CO_2 than diesel engines, premium gasoline engines produced more CO_2 than diesel engines which contradicted the existing literature.
- It was found that on average, cargo vans and passenger vans released the highest level of CO_2.
- It was found that for both diesel and gasoline powered vehicles, there is a positive linear relationship between fuel consumption and CO_2 emissions thus supporting the previous study conducted by Pinto & Oliver-Hoyo.
- The best regression model to predict CO_2 emissions in vehicles was found to be the Random forest regressor.

3. CLASSIFICATION

a) INTRODUCTION

Classification is a supervised learning method that predicts observations to distinct classes based on their features. Specific categories are identified based on one or more independent variables (Ramakrishnan, 2022).

For this task, the “Pima Indians Diabetes” dataset obtained from Kaggle was used. This dataset contains several diagnostic measurements of 440 female patients who are North American Indians aged 21 and above. The objective is to classify whether a patient has diabetes or not.

b) EXISTING LITERATURE

A study found out that when glucose level is more than 150, the patient is more likely to have diabetes (Zhan, 2022). A previous study on the effect of BMI on diabetic patients found out that people are more likely to have diabetes if their BMI is high (Gupta & Bansal, 2020). Research found out that the best performing model out of 7 different models on this dataset is the logistic regression model with an accuracy of 92.26% (Patil & Ingle, 2021).

c) RESEARCH QUESTIONS

1) What is the level of glucose at which the patient is more likely to have diabetes?
2) Does BMI have an effect on the patient having diabetes?
3) What is the most suitable model to predict whether a patient has diabetes or not?

d) EXPLORATORY DATA ANALYSIS

In each of the graphs below the outcome 0 denotes that the patient does not have diabetes and 1 denotes that the patient has diabetes.

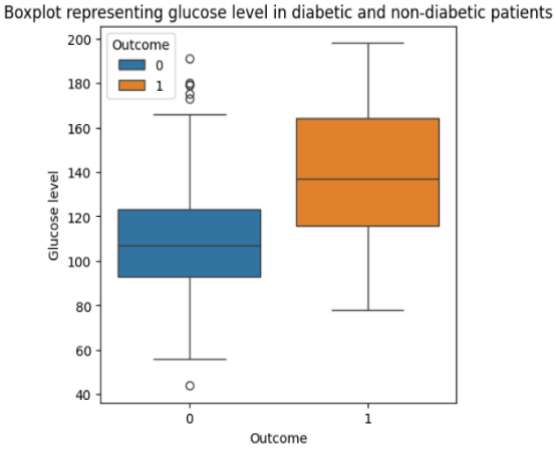

Figure 3.1

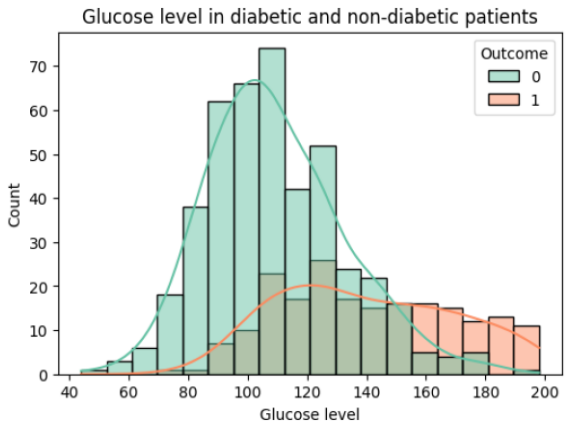

Figure 3.2

When comparing blood glucose levels for diabetic and non-diabetic patients, we can observe that from figure 3.1, the side-by-side boxplots show that median glucose level is 108 for non-diabetic patients whereas for diabetic patients it is at 138. Additionally, we can deduce that around 25% of diabetic patients have a glucose level of more than 160. Patients are highly likely to not have diabetes if the glucose is less than 80. The histogram in figure 3.2 shows that the distribution of glucose level for diabetic patients is skewed to the left when compared to non-diabetic patients. The patient’s risk of developing diabetes increases as the glucose level rises above 150 which was also seen from the previous study.

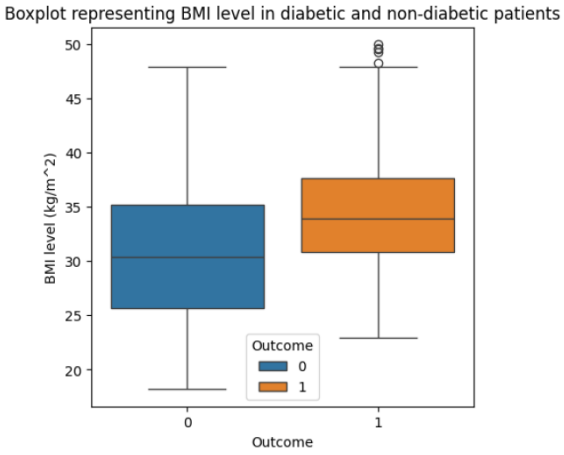

Figure 3.3

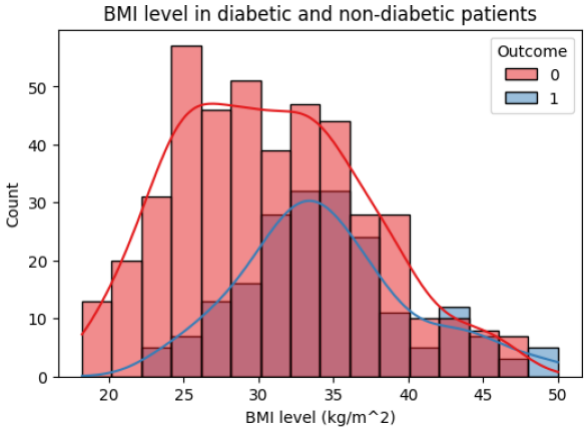

Figure 3.4

Figure 3.3 displays side-by-side boxplots comparing BMI for diabetic and non-diabetic patients. For diabetic patients, the median BMI is 34 kg/m^2 which is higher than for patients without diabetes (30 kg/m^2). Although there is a difference in the minimum BMI for both types of patients, the same cannot be said for the maximum BMI level which is similar at 48 kg/m^2. In figure 3.4, the histograms show that the distribution of BMI for diabetic patients is almost symmetric with its peak at about 33 kg/m^2. It is also quite evident that for almost all BMI levels, the patient is more likely to not have diabetes which is a contradiction to the existing literature. Therefore, we conclude that BMI does not have a clear effect on a patient having diabetes.

e) FEATURE SELECTION

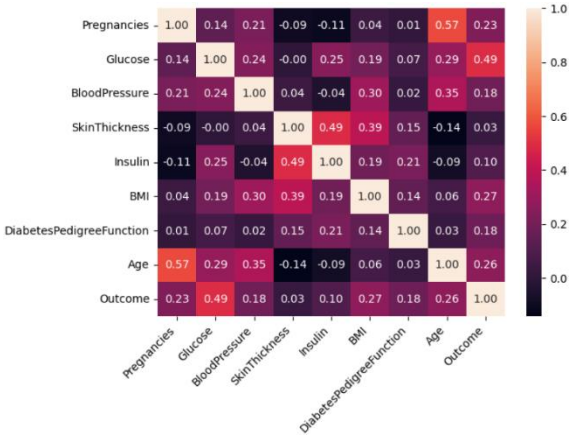

Figure 3.5: Correlation heatmap

From the correlation heatmap, we can see that SkinThickness and Insulin variables are moderately correlated with each other (+0.49) and both these have variables have almost zero correlation with the target variable (Outcome). Hence, we remove SkinThickness and Insulin columns from the dataset.

From the remaining features, the 4 best variables were selected using the K-Best feature selection method. In this case, the columns selected were Pregnancies, Glucose, BMI and Age.

f) CLASSIFICATION MODELS

The variables were split into train and test sets with a split ratio of 80%-20% respectively and then standardised. It was also noted that the training data was imbalanced therefore the data was balanced through a resample technique. This ensures that the model has a 50% chance of predicting each class. To determine which classification model performed the best, a number of models were trained and tested. The following models were used for this task:

1) **Decision tree classifier:** “Decision Tree is a Supervised Machine Learning Algorithm that uses a set of rules to make decisions, similarly to how humans make decisions.” (Bento, 2021).

2) **Random forest classifier:** Random forests uses many decision trees on various subsets of the given dataset and gets the average to improve the accuracy of the predictions (Random Forest Algorithm, n.d.).

3) **Logistic regression:** It is a type of classification where the dependent variable is binary and the independent variables are continuous or categorical. Using the relationships found between the predictor variables, the model calculates the probability of the target variable being in one of the 2 classes (Shah, 2023).

4) **CatBoost classifier:** Cat (Categorical data) Boost (Gradient boosting) uses many decision trees for classification (How CatBoost algorithm works, n.d.).

5) **Gradient boosting classifier:** In gradient boosting, many weak learning models are combined to create a powerful predicting model (Gradient Boosting Algorithm in Python with Scikit-Learn, 2023).

The metrics used to evaluate the models are:

1) **Accuracy:** How close the predicted values are to the true values.

2) **Precision:** "The proportion of positive class predictions that were actually correct." (Kundu, Precision vs. Recall: Differences, Use Cases & Evaluation, 2022).

3) **Recall: "**The proportion of actual positive class samples that were identified by the model." (Kundu, Precision vs. Recall: Differences, Use Cases & Evaluation, 2022).

4) **F1 score:** This measure combines precision and recall to measure the model's accuracy. (Kundu, F1 Score in Machine Learning: Intro & Calculation, 2022).

5) **Confusion matrix:** A matrix that compares the number of predicted and actual outcomes.

6) **ROC curve:** It is a probability curve that shows the model's capability to distinguish between the different classes (Narkhade, 2018).

Models were tuned using hyperparameters and the values obtained for each model is shown in the table below.

Model	Accuracy	Precision	Recall	F1 score	Confusion matrix
Decision tree classifier	74.21%	62.50%	58.14%	60.24%	Predicted Negative / Predicted Positive Actual Negative: 70 / 15 Actual Positive: 18 / 25
Random tree classifier	71.09%	58.82%	46.51%	51.95%	Predicted Negative / Predicted Positive Actual Negative: 71 / 14 Actual Positive: 23 / 20
Logistic regression	65.63%	49.06%	60.47%	54.17%	Predicted Negative / Predicted Positive Actual Negative: 58 / 27 Actual Positive: 17 / 26
CatBoost classifier	74.22%	66.67%	46.51%	54.79%	Predicted Negative / Predicted Positive Actual Negative: 75 / 10 Actual Positive: 23 / 20
Gradient boosting classifier	71.09%	60.71%	39.53%	47.89%	Predicted Negative / Predicted Positive Actual Negative: 74 / 11 Actual Positive: 26 / 17

As the dataset was imbalanced, it is inappropriate to compare the models using the accuracy metric as this metric can be misleading. Therefore, the F1 score and the ROC curves are considered to determine the best performing model.

We can see from the table in the previous page that the Decision tree classifier model has the largest F1 score with 60.24% while CatBoost classifier scored 54.79%. To get a more detailed view on the performance of each model, ROC curves for all the models are plotted below.

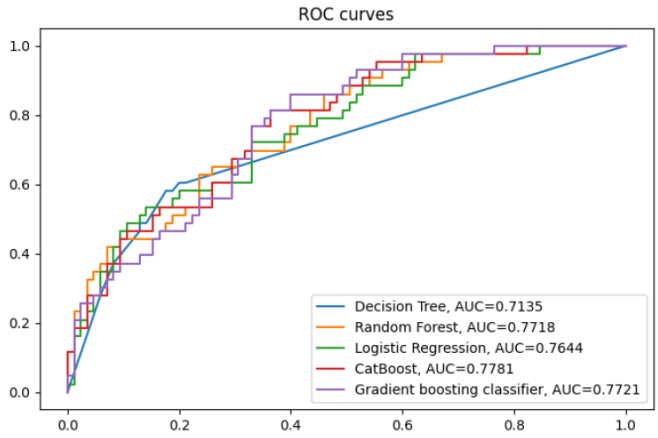

Figure 3.6

If the curve for a model is closer to the top left corner, then that model is more precise in predicting whether a patient has diabetes or not. The Area Under the Curve (AUC) can be used to determine the performance of the model at different classification thresholds. From the curves in figure 3.6, we can see that CatBoost classifier has the highest AUC with 0.7781 closely followed by Gradient boosting classifier (0.7721) and Random Forest (0.7718). Although decision tree provided the highest F1 score, the AUC is the lowest among all models with a value of 0.7135.

Since the CatBoost classifier gave the highest AUC and the 2nd highest F1 score, CatBoost classifier is the most suitable model to predict diabetic and non-diabetic patients.

g) CONCLUSION

- As seen in the prior study, the patient's risk of getting diabetes increases as the blood glucose level increases past 150.
- BMI does not have an influence on a patient being diabetic or non-diabetic. This contradicted the previous research conducted.
- It was found that the CatBoost classifier was the most appropriate model to implement on this dataset to predict whether a patient has diabetes or not.

4. BIBLIOGRAPHY

BasuMallick, C. (2023, September 25). *What Is Principal Component Analysis (PCA)? Meaning, Working, and Applications*. Retrieved March 27, 2024, from Spiceworks: https://www.spiceworks.com/tech/big-data/articles/what-is-principal-component-analysis/

Bento, C. (2021, June 28). *Decision Tree Classifier explained in real-life: picking a vacation destination*. Retrieved April 2, 2024, from Towards Data Science: https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b2b60575

Clemente, J. P. (2013). *Recheio : growing through a declining channel : a case on wholesale food private labels and its competitiveness.* Retrieved March 27, 2024, from http://hdl.handle.net/10400.14/15674

Gradient Boosting Algorithm in Python with Scikit-Learn. (2023, August 16). Retrieved April 2, 2024, from Simplilearn: https://www.simplilearn.com/gradient-boosting-algorithm-in-python-article#:~:text=Gradient%20Boosting%20is%20a%20functional,produce%20a%20powerful%20predicting%20model.

Gupta, S., & Bansal, S. (2020, April 1). Does a rise in BMI cause an increased risk of. *Plos One*. Retrieved March 31, 2024, from https://doi.org/10.1371/journal.pone.0229716

How CatBoost algorithm works. (n.d.). Retrieved April 2, 2024, from ArcGIS Pro: https://pro.arcgis.com/en/pro-app/3.1/tool-reference/geoai/how-catboost-works.htm#:~:text=CatBoost%20is%20a%20supervised%20machine,gradient%20boosting%20(the%20Boost).

Kundu, R. (2022, December 16). *F1 Score in Machine Learning: Intro & Calculation*. Retrieved April 2, 2024, from V7Labs: https://www.v7labs.com/blog/f1-score-guide

Kundu, R. (2022, September 19). *Precision vs. Recall: Differences, Use Cases & Evaluation*. Retrieved April 2, 2024, from V7Labs: https://www.v7labs.com/blog/precision-vs-recall-guide

Narkhade, S. (2018, June 26). *Understanding AUC - ROC Curve*. Retrieved April 2, 2024, from Towards Data Science: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

Nilrit, S., Sampanpanish, P., & Bualert, S. (2017). *Comparison of CO2 Emissions from Vehicles in Thailand.* Applied Environmental Research. Retrieved March 30, 2024, from https://d1wqtxts1xzle7.cloudfront.net/105386078/67059-libre.pdf?1693389323=&response-content-disposition=inline%3B+filename%3DComparison_of_CO2_Emissions_from_Vehicle.pdf&Expires=1711784597&Signature=IZ9KpbcH714rqdleV9PACxb-kmkINzEso2-sJjQT3bT3tkEa1EV1Px3

Patil, V., & Ingle, D. R. (2021). Comparative Analysis of Different ML Classification Algorithms with Diabetes Prediction through Pima Indian Diabetics Dataset. *2021 International Conference on Intelligent Technologies (CONIT)*. doi:https://doi.org/10.1109/CONIT51480.2021.9498361

Patlolla, C. R. (2018, December 10). *Understanding the concept of Hierarchical clustering Technique*. Retrieved April 2, 2024, from Towards Data Science: https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec

Patrick. (2023, May 12). *Unsupervised Learning: Clearly Explained*. Retrieved March 25, 2024, from Alexanderthamm: https://www.alexanderthamm.com/en/blog/this-is-how-unsupervised-machine-learning-works/

Pinto, G., & Oliver-Hoyo, M. T. (2008). Using the Relationship between Vehicle Fuel Consumption and CO2 Emissions To Illustrate Chemical Principles. *Journal of Chemical Education*, 218. doi:https://doi.org/10.1021/ed085p218

Pokharel, M., Bhatta, J., & Paudel, N. (2021, December 31). Comparative Analysis of K-Means and Enhanced K-Means Algorithms for Clustering. *NUTA Journal*, 79-87. doi:https://doi.org/10.3126/nutaj.v8i1-2.44044

Popovic, R. (2009). *Effects of market structure changes on dairy supply chain in Serbia.* Retrieved March 28, 2024, from https://www.ifama.org/resources/files/2009-Symposium/1003_paper.pdf

Ramakrishnan, M. (2022, November 23). *What is Classification in Machine Learning and Why is it Important?* Retrieved March 31, 2024, from Emeritus: https://emeritus.org/blog/artificial-intelligence-and-machine-learning-classification-in-machine-learning/

Random Forest Algorithm. (n.d.). Retrieved April 2, 2024, from JavatPoint: https://www.javatpoint.com/machine-learning-random-forest-algorithm

Shah, D. (2023, March 31). *Logistic Regression: Definition, Use Cases, Implementation*. Retrieved April 2, 2024, from V7labs: https://www.v7labs.com/blog/logistic-regression

Sharma, N. (2023, August 8). *K-Means Clustering Explained*. Retrieved March 29, 2024, from Neptune.ai: https://neptune.ai/blog/k-means-clustering

Zhan, W. (2022, December 30). A Comparative Study on Machine Learning Based Type 2 Diabetes Mellitus Prediction. *Proceedings of the 2022 International Conference on Computer Science, Information Engineering and Digital Economy (CSIEDE 2022)*. Retrieved March 31, 2024, from https://doi.org/10.2991/978-94-6463-108-1_95