Machine Learning (ST3189)

Coursework Project

Student Number : 200640778

Module Code : ST3189

Number of Pages : 10

Table of Contents

1.0 Unsupervised Learning

Unsupervised learning is, using machine learning algorithms to cluster unlabeled data and identify any patterns or groups to make data analysis without human involvement. Unsupervised learning is made up of 3 chores. Which is dimensionality reduction, clustering, and association. In this part of the coursework, 1 model from each dimensionality reduction and clustering will be analyzed further. And research questions will be analyzed using exploratory data analysis to understand any relationships in-between the factors. The dataset used for the analysis will be a dataset named “customer personality analysis data”. The dataset includes customers’ age, educational levels, the amount they spent on products, what they spent money on, and how many kids they have. etc. link for the dataset is available in the bibliography.

1.1 Research questions

- What are the factors affecting the spending decision of the customers?

1. Does the amount they spend on buying products from the company has a relationship with the customer’s educational level?

2. Is there a relationship between the amount customers spent on products like wine, meat, and gold when the number of kids in the house changes?

1.2 Exploratory data analysis (EDA)

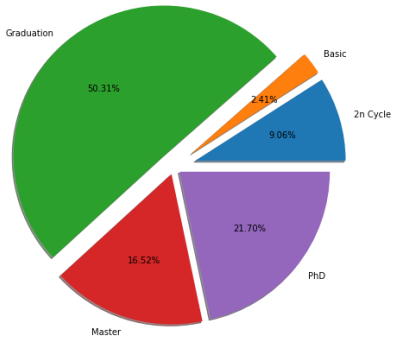
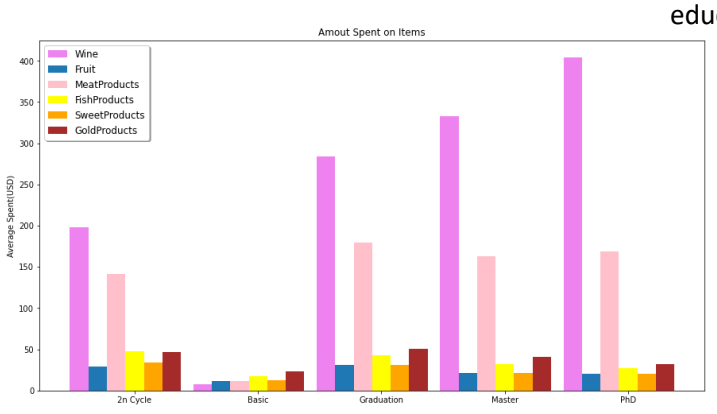


Figure1: education level of customers

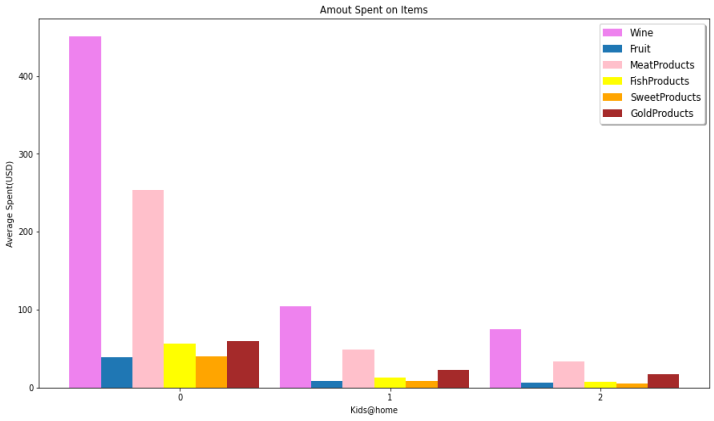
by observing the pie chart, it is obvious that around half of the customers are graduated. And about 88% of the customers at least have a graduate degree. And only 11.46% of the customers are below the graduate level.

When analyzed further, it is possible to see that more educated people have a higher intention to spend more on buying wine. Compared to other education levels, customers with a basic education level spend a small amount overall. A large gap exists between 2n Cycle and basic salary level in overall spending. Customers have mostly spent money on wine and next is meat products.



Therefor from the above analysis, it is possible to say that, there are evidence on more educated customers tends to spend more in every product type.

Figure2: education level vs spendings on different products



When comparing the spending habits of the customers with respect to how many kids they have at home, it is apparent that the overall spendings drop drastically from no kids to 2 kids at home. Spendings on wine are almost 4 times when there are no kids at home compared to one kid. Spending on sweets, fruits, and fish drop to below 15$ when at least one kid is at home.

Figure3: Number of kids at home vs average spendings on different products

1.2 Literature review

A study done by Chris butter has found that wine consumption and spending money on wine are correlated with demographic factors such as income, age, and education. It has divided customers into 4 groups. It states that higher-income customers consume more than 3 times the amount middle-income customers consume and more than 8 times the amount consumed by the customers in the lower income levels. (Bitter, 2020)

It also states that when comparing the education level of each income group customers in the higher income level with a graduate degree consume more than 15% of what a consumer in the same income level consumes without a graduate degree. Consumers in the lower 75% income level only account for 18% of the total wine sales. In that income group also, graduated consumer accounts for 36.5% of the consumption while non-graduates only consume 23.5% of the consumption. But they spend money on other alcohol products. (Bitter, 2020)

1.3 Clustering analysis

Clustering is used to process row and unclassified data. Patterns or groups will be identified from the dataset when using clustering. There are a few cluster types exclusive, overlapping, and hierarchical.

1.3.1 K means clustering

K means clustering is an exclusive clustering method. Data will be assigned into k number of groups. Grouping will be based on the distance to each group centroid. Points closest to a centroid will be in the same cluster. When analyzing the customer personality analysis dataset using k means clustering, we use the elbow method to the dataset, and it's clear that the optimum number is 3 clusters. The elbow method is a method used to identify the optimum value for k.

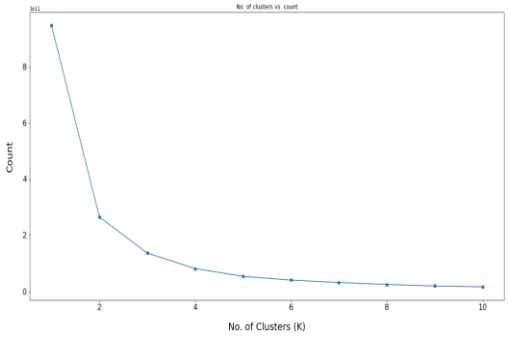
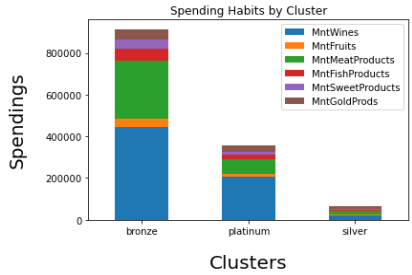


Figure3: Elbow method graph to choose optimal k value

Figure4: graph of variations in the clusters



In this graph, it is clear how the 3 clusters are being divided. And it's clear how each cluster spends money on different products. The 3 clusters will be named bronze, platinum, and silver for identification purposes.

By analyzing the clusters, it's clear that bronze is the cluster that spends the most.

The highest amount is spent on wine. And the least spending is coming from the silver cluster. comparing the income of each cluster, the bronze cluster has the highest income as well. But the silver cluster has some outliers which spend way above the mean spending of the cluster.

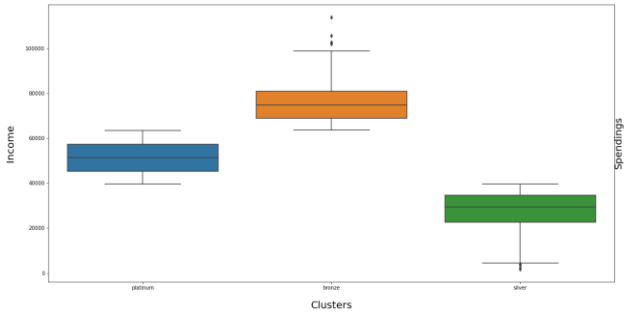


Figure5: Boxplot of incomes of each cluster

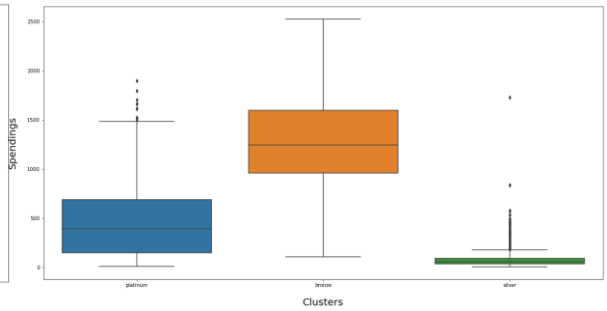


Figure6: Boxplot of spendings of each cluster

1.4Dimensionality reduction method

Dimensionality reduction is used when a dataset has too many attributes, by using this method it can reduce the number of attributes and keep the variation of the dataset as much as it can. There are several dimensionality reduction methods, one is principal component analysis (PCA).

1.4.1 Principal component analysis

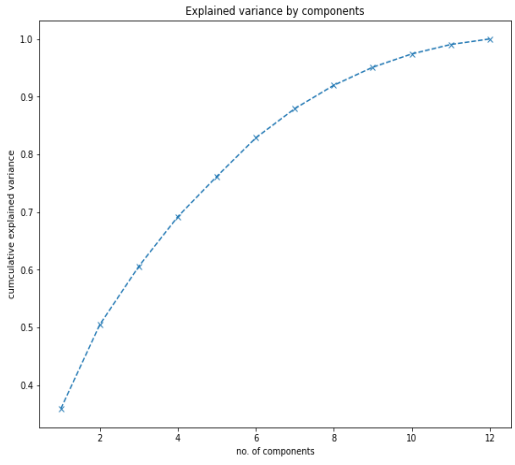


Figure7: PCA graph to get the optimal component value

PCA can transfer a larger number of correlated variables to a smaller number of uncorrelated variables. Which are known as principal components. By analyzing the dataset using PCA the optimum number for Principal components is 7 components. The value is selected by understanding the explained variance graph for the dataset. Since the graph gets flatter as it goes, the optimal value will be to use 7 as the number of uncorrelated variables. When using 7 components, the explained variance ratio will be 0.88. which suggests that using 7 variables will be giving 88% of the variance in the original dataset.

2.0 Regression

Regression analysis is used to understand relationships between dependent and independent variables. Using regression to analyze data is the most fundamental task in the industry. The dataset used in this part of the coursework will be “Co2 emission by vehicles”. Where Co2 emission will be analyzed depending on features such as vehicle class, fuel type, engine size, cylinder in the engine, and fuel consumption. link for the dataset is available in the bibliography.

The dataset will be split into training and test datasets with a ratio of 0.75 and 0.25 of the main datasets respectively.

2.1 Research question

- What factors affect the high Co2 emission in vehicles?

1. Is there a correlation between Co2 emission(g/km) and fuel consumption (L/100 km) of vehicles?

2. Does fuel type have the influence to differentiate fuel consumption of different vehicles?

3. does Co2 emission Differs from the brand of a vehicle?

2.2Exploratory data analysis (EDA)

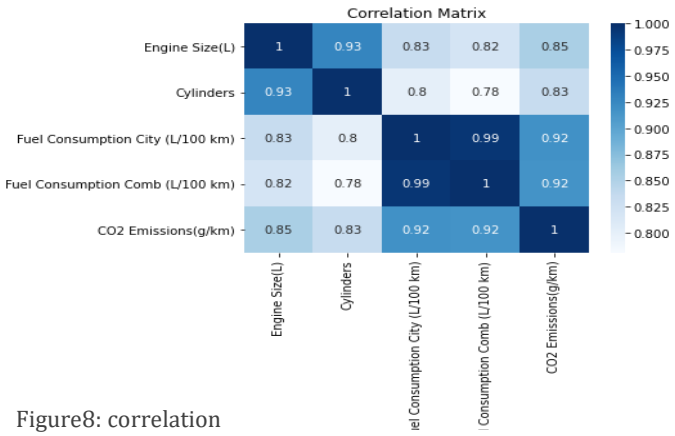


Figure8: correlation matrix of each attribute

By observing the correlation matrix, it is distinct that engine size, cylinders, and fuel consumption have a positive correlation with the Co2 emission of a vehicle. But fuel consumption has a very strong relationship with a correlation score of 0.93. looking at the factors which affect low fuel consumption will give a better image of what factors affect high Co2 emissions.

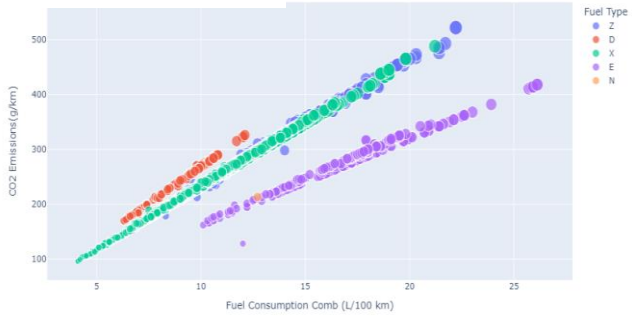


Figure9: Fuel type vs. fuel consumption and Co2 emission

Analyzing further it is obvious that fuel type also has a positive relationship with the Co2 emission of vehicles. Whereas diesel vehicles are very inefficient in fuel consumption and produce more Co2. But ethanol(E85) vehicles emit less Co2 and require less fuel per 100km.

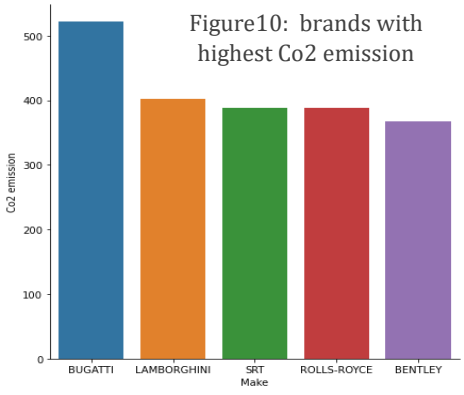


Figure10: brands with highest Co2 emission

When referring to the average Co2 produced per kilometer from a vehicle, the graph shows how much Co2 a vehicle from each brand emits. Only the top 5 brands are plotted in the graph. Therefore, it is possible to see that all the brands are high-end luxury vehicle brands with larger engine displacements.

Bugatti leads the table by producing 0.522kgs of Co2 per kilometer. Therefore, it's obvious that the larger the engine displacements are, the higher the Co2 emission from each vehicle.

2.3 Literature review

A study done by Valentinas mickunaitis, Alvidas pikunas, and Igor mackoit has found that there is a relationship between the emission of Co2 and the fuel consumption of a vehicle. It also states that the average fuel consumption of a vehicle using diesel is 26% less than that of a vehicle with the same features which use petrol. When burning the fuel, 1L of diesel produces 2.7kgs of Co2, but 1L of petrol only produces 2.4kgs of Co2. The study also states that the large engine displacements lead to a higher Co2 emission because of very less fuel consumption rates.

It has also been found that the mass of the vehicle acts a crucial part in increasing the Co2 emission. When the mass increases by 100kg, it states that Co2 emissions grow by 6.5% for petrol engines and 7.1% for diesel engines. Transmission of the vehicle also affects the Co2 emission directly. Vehicles using automatic transmission increase Co2 emissions by 16% in petrol vehicles, compared to 33% in diesel vehicles. (Mickūnaitis et al., 2007)

Research done by Uswitch(UK-based company) has compared more than 50 car brands using an international testing standard to calculate fuel efficiency and emission levels. And the results of the research show that luxury car brands and sports car brands have the highest emission levels and least fuel efficiency levels. (Wise, 2021)

2.4 Regression models and results

To recognize which model is the best to use with the dataset, 5 regression models will be compared with each other. 4 measures will be checked from each model.

- R squared value is the proportional value of variance which was explained by each regression model.
- MSE (mean squared error) gives the average value of the squared difference between original and predicted values in the dataset used.
- RMSE is the square root value of MSE. It represents the standard deviation of residuals.
- MAE (mean absolute error) represents the average value of the absolute difference between actual and predicted values in the model used.

Values obtained from using each model will be as followed,

Score / Model type	R squared value	Mean squared error value (MSE)	Root mean squared error value (RMSE)	Mean Absolute error value (MAE)
Linear regression model	**0.8785**	**441.32**	**21.008**	**13.96**
KNN regression model	**0.9727**	**99.25**	**9.962**	**4.74**
Support vector regression model (SVR)	**0.8063**	**703.85**	**26.530**	**9.92**
Decision tree regression model	**0.9566**	**157.72**	**12.558**	**3.97**
Random forest regression model	**0.9739**	**94.99**	**9.746**	**3.75**

2.5 Conclusion

- Value of R squared ranges from 0 to 1. A strong fit is a model with an r-squared value greater than 0.80. This implements that all our models are strongly fitting the dataset, but the random forest regressor fits the most out of all the models.
- MAE is the average error and RMSE indicates the average squared error, therefor lower the values of MAE and RMSE more the model fits the dataset.
- RMSE is good at capturing larger errors because it gets squared, but MAE only gives the average error.
- Therefore looking at the results obtained from the models used, it's accurate to use a random forest regressor as a model to make predictions on this dataset because, it has the best values in r squared, RMSE, and MAE.

3.0 Classification

Classification is a supervised machine learning process of predicting classes or categories of data based on predefined classes of data which was labeled.

The dataset used for classification modeling will be “airplane passenger satisfaction”. The ratio of the training and test data will be 0.8 and 0.2 respectively. link for the dataset is available in the bibliography.

3.1 Research questions

- What factors affect the higher satisfaction level in airplane passengers?

1. Does inflight entertainment have a relationship with the satisfaction level of airplane passengers?

2. Does a better online boarding process have an impact on the overall satisfaction of the passengers?

3.2 Exploratory data analysis (EDA)

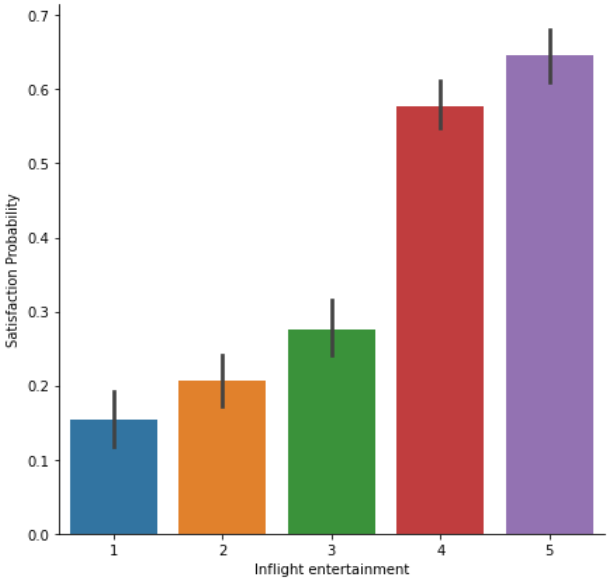


Figure11: Rating for inflight entertainment with passenger satisfaction probability

- Exploring the dataset, it is obvious that inflight entertainment has a positive relationship with the satisfaction level of the passengers. 66% and 61% of the passengers who have graded inflight entertainment with a rate of 4 and 5 out of 5 respectively are satisfied with their overall experience.
- More than 70% of the passengers who have graded inflight entertainment with a rate of 1,2 or 3 out of 5 are dissatisfied or neutral about the overall experience.

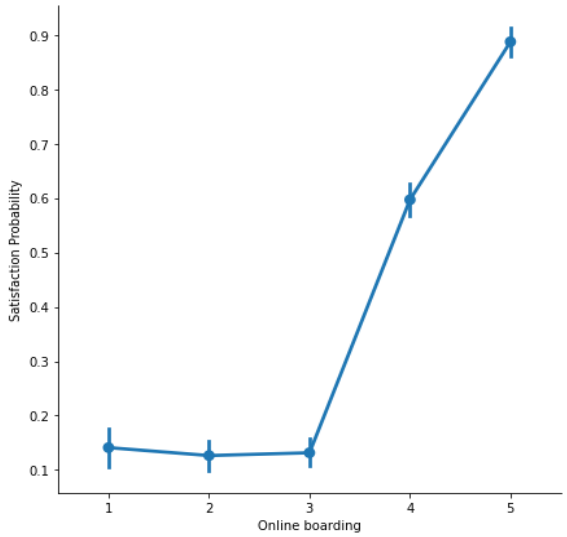


Figure12: Rating for online boarding with passenger satisfaction probability

- Observing the point plot, it's clear that online boarding has a direct impact on the overall satisfaction of airplane passengers. Passengers who are not satisfied with the online boarding process (passengers who rated with 1,2 or 3 out of 5) are not satisfied with the overall experience.
- When the ratings for online boarding are 4 and 5 out of 5, there's a rapid growth in the overall satisfaction level of the passengers. As a percentage around 67% of the passengers who rated 4 out of 5 for online boarding and 89% of the passengers who rated 5 out of 5 are satisfied with the overall experience.

3.3 Literature review

A study done by Sahar tahanisaz has found that inflight entertainment act as a deciding factor for improving the overall satisfaction level of airplane passengers. The study mentions how important it is to use an inflight VR entertainment system for the passengers, where they can access it in 2D or 3D format. Passengers are more concerned about easily downloading a free video player app where they can choose and watch a movie or a television episode. It also states that kids-only mode in the entertainment system should allow kids to be entertained throughout the journey without parents being worried. A customized magazine for each passenger with interesting content should help increase inflight entertainment satisfaction by a huge margin. The study also states that the variety and quality of inflight entertainment system in business class is not provided in economy class, which leads to dissatisfaction among the passengers. (Tahanisaz & shokuhyar, 2020)

Data mining research done by Tri Noviantoro and Jen-Peng Huang using feature selection has found that online boarding features are the most critical service to increase passenger satisfaction levels. It states that online boarding using a mobile application or a website will help passengers to avoid waiting in line to check-in. It also states that the time spend on personal contact with an agent in a standard check-in is saved for the passengers and they have more time to explore terminal stores. It's also an advantage to the airport as well because it states that airports receive a significant part of their revenue from in-terminal stores. Therefore, using online boarding makes passengers feel that they are in control of their trip, Which leads to overall satisfaction. (Noviantoro & Huang, 2022)

3.4 Classification models and results

- classification models will be used to analyze the dataset and a comparison of each model will be done by comparing 4 score parameters.

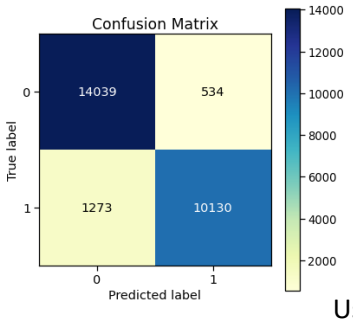
Accuracy scores ➡ The ratio of true positive and true negative to all positive and negative observations. This gives the percentage of how often the model will predict correctly. (Kumar, 2023)

Precision score ➡ This measure the proportion of positively predicted which are actually correct. Precision is also known as the positive predictive value. (Kumar, 2023)

Recall score ➡ This gives the model ability to correctly predict the positives out o the actual positives. It's also known as the true positive rate or sensitivity. (Kumar, 2023)

F1 score ➡ This is a model score given as a function of recall and precision. It's also known as the harmonic mean of precision and recall. (Kumar, 2023)

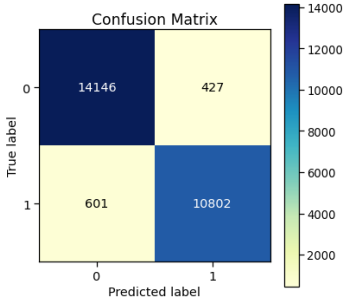
3.4.1. K-nearest neighbor classifier



K-nearest neighbor uses proximity to make predictions about the dataset. K value in the K-NN algorithm refers to the number of neighbors which will be used to identify the classification of a specific query point. K-NN is simple to use and good in accuracy but, it also has weaknesses such as it doesn't work with high dimensional data and a lower value of K can easily overfit the model predictions.

Using KNN the accuracy of the model is 93.04%, precision is 94.99%, recall is 88.84% and F1 is 91.81%.

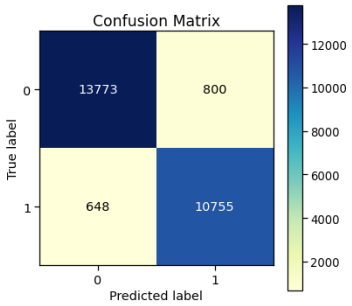
3.4.2. XGB classifier



Extreme gradient boosting is an optimized gradient boosting library. XGB is very efficient in handling missing values without any pre-processing.

Using XGB the accuracy of the model is 96.04%, precision is 92.20%, recall is 94.73% and f1 is 95.46%.

3.4.3 Decision tree classifier

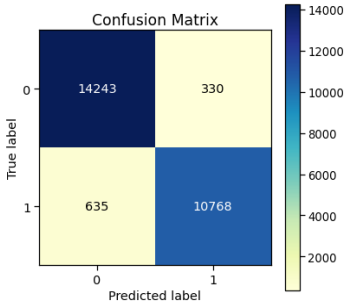


A decision tree classifier uses previously labeled data to train an algorithm that can be used to make predictions about the dataset.

In the decision tree the top node is known as the root node, each decision point is a decision node and the final decision point is called the leaf node.

Using the decision tree classifier as the model, accuracy is 94.39%, precision is 93.10%, recall is 94.21% and f1 is 93.65%.

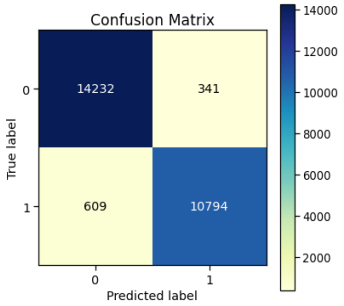
3.4.4 Random Forest classifier



Random forest is made up of multiple decision trees to give an output of a single result. It makes an uncorrelated forest of random forests to give that output. Feature randomness is used to assure the least correlation between decision trees.

Using a random forest classifier as the model, accuracy is 96.29%, precision is 97.07%, recall is 94.40% and f1 is 95.72%.

3.4.5 LGBM classifier



LGBM is a fast and high-performance gradient-boosting machine learning framework. It's also supported by decision tree algorithms. The difference is it splits tree leafs with the simplest fit possible. Therefore, it leads to better accuracy compared to other boosting algorithms.

Using the LGBM classifier as the model, accuracy is 96.34%, precision is 96.94%, recall is 94.66% and f1 is 95.78%.

3.5 Conclusion

The objective of the analysis is to identify the model with the highest accuracy. Therefore, it is noticeable that LGBM and Random Forest classifiers have the highest accuracies. Out of these two, the highest observed predictive performance(accuracy) is from the LGBM model. Therefore, it is the best model for the prediction purposes of this dataset.

4.0 Bibliography

Tahanisaz, S., & shokuhyar, S. (2020). Evaluation of passenger satisfaction with service quality: A consecutive method applied to the airline industry. *Journal of Air Transport Management*, *83*, 101764. https://doi.org/10.1016/j.jairtraman.2020.101764

Noviantoro, T., & Huang, J.-P. (2022). Investigating airline passenger satisfaction: Data Mining Method. *Research in Transportation Business & Management*, *43*, 100726. https://doi.org/10.1016/j.rtbm.2021.100726

Kumar, A. (2023, March 17). *Accuracy, precision, Recall & F1-Score - Python examples*. Data Analytics. Retrieved March 31, 2023, from https://vitalflux.com/accuracy-precision-recall-f1-score-python-example/

Mickūnaitis, V., Pikūnas, A., & Mackoit, I. (2007). Reducing fuel consumption and CO2 emission in Motor Cars. *TRANSPORT*, *22*(3), 160–163. https://doi.org/10.3846/16484142.2007.9638119

Wise, M. (2021, June 30). *Which car brands emit the most carbon dioxide?* Earth911. Retrieved March 31, 2023, from https://earth911.com/business-policy/which-car-brands-emit-the-most-carbon-dioxide/

Bitter, C. (2020, February 28). *Demographics ...* Demographics and Wine: The Class Divide. Retrieved April 1, 2023, from https://www.vineconomics.com/blog/demographics-and-wine-the-class-divide#:~:text=The%20spending%20differential%20is%20partly%20related%20to%20the,alcohol%20budgets%20to%20wine%20at%20all%20income%20levels.

Airplane passenger satisfaction dataset:

Klein, T. J. (2020, February 20). *Airline passenger satisfaction*. Kaggle. Retrieved April 3, 2023, from https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

Customer personality analysis dataset:

Patel, A. (2021, August 22). *Customer personality analysis*. Kaggle. Retrieved April 3, 2023, from https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

Co2 emission by vehicles dataset:

Podder, D. (2020, August 5). *CO2 emission by vehicles*. Kaggle. Retrieved April 3, 2023, from https://www.kaggle.com/datasets/debajyotipodder/co2-emission-by-vehicles