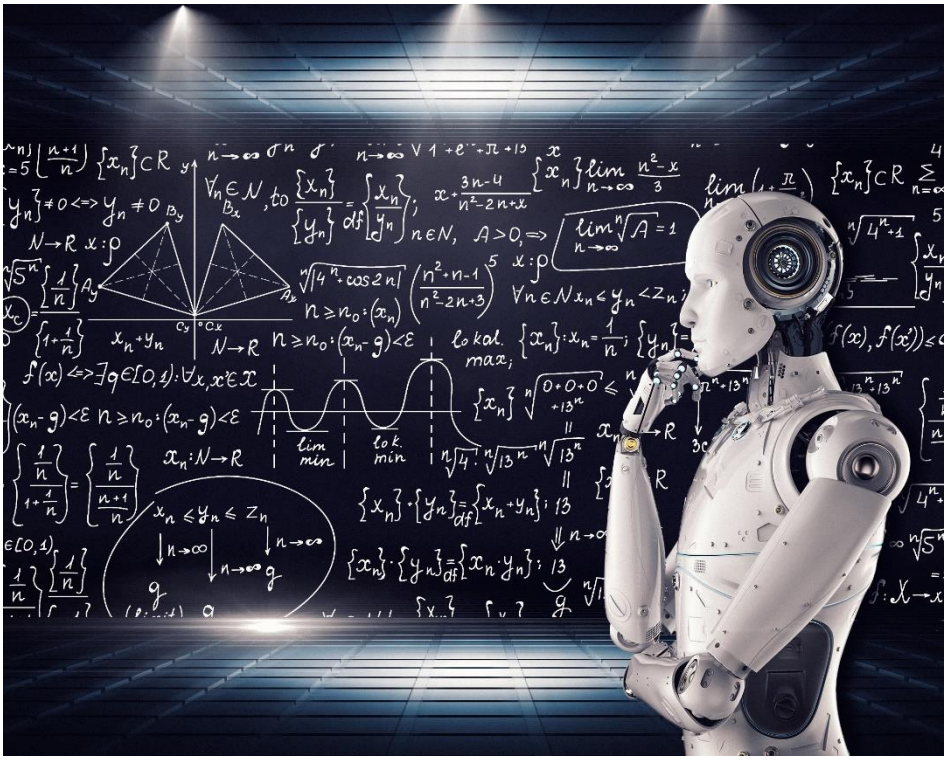
THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

UNIVERSITY OF LONDON



MACHINE LEARNING WITH PYTHON

MODULE : MACHINE LEARNING - ST3189

UOL STUDENT NUMBER : 210458262

NUMBER OF PAGES : 10

(EXCLUDING COVER PAGE, TABLE OF CONTENTS AND BIBLIOGRAPHY)

Contents

1.0 Unsupervised Learning

“Unsupervised learning uses machine learning (ML) algorithms to analyze and cluster unlabeled data sets. These algorithms discover hidden patterns or data groupings without the need for human intervention” (IBM, n.d.). Unsupervised learning involves three main tasks: dimensionality reduction, clustering, and association.

The dataset used for the analysis, named "Penguin Species," comprises penguin features; culmen length, culmen depth, flipper length, body mass, and sex, obtained from the Kaggle platform.

1.1 Existing Literature

In a study in sexing Chinstrap Penguins it was found that “The degree of sexual size dimorphism the Chinstrap Penguin is similar to that exhibited by other penguin species in flipper length, but is smaller in bill length and body mass (Marchant and Higgins 1990). Trivelpiece and Trivelpiece (1990) found sexual differences in body mass for Chinastraps during the breeding season. Croll et al. (1991) stated, without stating data, that Chinastrap Penguins do not exhibit sexual dimorphism.” (Juan A. Amat)

1.2 Data Preprocessing

Null values, duplicated rows, and negative values, since they are unrealistic, were dropped from the dataset. Additionally, outliers were removed to ensure data quality and reliability for analysis. Furthermore, the numerical data was standardized, and the categorical variables were encoded using one-hot encoding to obtain indicator variables.

1.3 Research Questions

- Do male and female penguins exhibit significant differences in culmen length, depth, flipper length, or body mass?
- Is there a correlation between different physical characteristics (culmen length, culmen depth, flipper length, body mass)?
- Can we identify natural groupings or clusters of penguins based on their physical characteristics?

1.4 Exploratory Data Analysis(EDA)

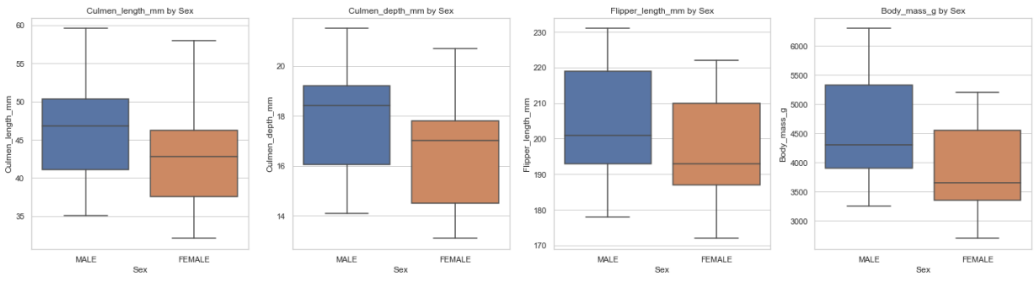


Figure 1

In Figure 1, the box plots illustrate that the lower (Q1) and upper (Q3) quartiles, as well as the minimum and maximum values, are consistently lower for females compared to males across all four features: culmen length, culmen depth, flipper length, and body mass. This suggests that, on average, females tend to have smaller measurements for these penguin characteristics compared to males. Further investigation is conducted through an independent two-sample t-Test. The t-statistics for culmen length/depth, flipper length, and body mass indicate a significant difference between male and female penguins ($p < 0.05$), suggesting statistically significant differences in the features between the two groups. With all p-values indicating a

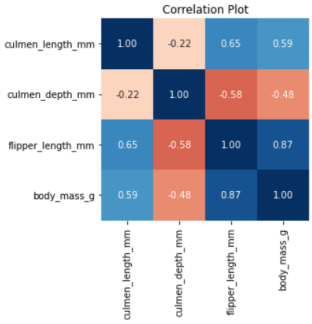


Figure 2

significance level below 0.05, we can confidently reject the null hypothesis and infer that there are notable variances in the measured characteristics between male and female penguins.

Figure 2 shows that culmen length and depth have a weak correlation of 0.22, whereas flipper length and body mass are strongly correlated with 0.87. The others are moderately correlated, with some being positive and some negative.

1.5 Unsupervised Models

1.5.1 Principal Component Analysis (PCA)

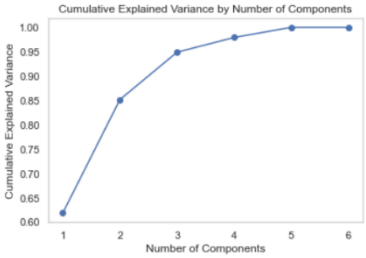


Figure 3

PCA reduces the dimensionality of datasets by transforming numerous variables into smaller ones while retaining most of the original information. After preprocessing (removing missing values and outliers), the dataset comprised 6 columns and 332 rows. Prior to PCA, standardization ensured unbiased results. Initially, 6 principal components were generated. To determine the optimal number of components, a cumulative explained variance graph was plotted. Figure 3 demonstrates that variance explained by each component decreases as their count rises, with marginal change becoming minimal after the first 2 components. Aiming to retain 85% variance, 2 principal components were chosen, capturing 85% of the dataset's variation. Thus, 2 components were selected for dimensionality reduction.

1.5.2 KMeans Clustering

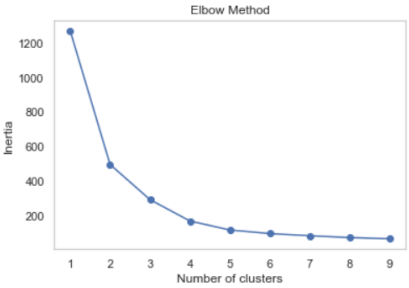


Figure 4

The Elbow method plot helps find the best number of clusters. As we increase the clusters, the data's within-cluster sum of squares generally decreases. In Figure 4, our focus lies on identifying the 'elbow' point, where the rate of decrease starts to slow down, suggesting that further addition of clusters does not notably enhance the clustering. Subsequently, following the Elbow method to ascertain the optimal number of clusters (in this instance, 4), KMeans clustering was implemented. Each point in the plot represents a data point, colored according to its cluster. Clusters are groups of data points close to each other in the data space, aiming to identify meaningful patterns. Overall, KMeans partitions the data into groups based on similarity. This helps reveal underlying patterns and relationships, though careful interpretation within the context of the problem is essential.

1.5.3 Dendogram (Hierachical Clustering)

Hierarchical clustering groups similar data points into clusters based on their pairwise distances. Three linkage methods—Ward, Complete, and Average—calculate cluster distances differently, resulting in varied cluster structures. The dendrogram visualizes the hierarchical relationship between data points and clusters. In the Ward method plot, the x-axis represents cluster sizes, while the y-axis indicates distances between clusters. Moving upward on the y-axis merges clusters, starting from individual points and ending with one cluster. The height of vertical lines indicates merge distances: longer lines signify dissimilar clusters, while shorter ones indicate similarity. By analyzing the dendrogram, we determine optimal cluster numbers by identifying significant changes in merge distances, suggesting natural breakpoints or "jumps." Dendrograms provide insights into data hierarchy, aiding in the understanding of cluster formation and relationships between data points.

1.5.4 Gaussian Mixture Model (GMM)

GMM clustering categorizes data points into groups based on their statistical distribution. In our analysis, we aimed to find clear clusters within the dataset. The GMM algorithm identified 4 clusters, each representing data points with similar traits. These clusters are shown in a scatter plot, with data points colored by their cluster label. Overall, GMM clustering helps us grasp natural groupings in the data, offering insights into its characteristics. It enables us to spot distinct clusters and understand relationships between data points, aiding in further analysis and decision-making.

1.5.5 Density Based Spatial Clustering(DBSCAN)

DBSCAN stands as a clustering technique designed to organize data points by their closeness, thereby facilitating the identification of natural groupings within the dataset. It identifies dense regions as clusters and outliers as noise. Unlike K-means, it doesn't require the number of clusters to be predefined. The algorithm's parameters, 'eps' and 'min_samples', control cluster density. In our analysis, DBSCAN identified clusters in the dataset (2), shown in a scatter plot. Each point is colored by its cluster. This helps reveal data grouping and outliers. DBSCAN's flexibility makes it valuable for exploratory data analysis without needing prior knowledge of the number of clusters.

1.5.6 Comparison of clusters

In comparison to the others, GMM clusters are well-separated, comprising four distinct groups. K-means also results in four clusters, while the dendrogram reveals two distinct clusters. However, DBSCAN identifies two clusters in the dataset.

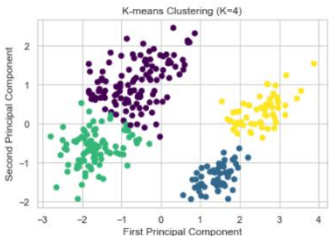


Figure 5

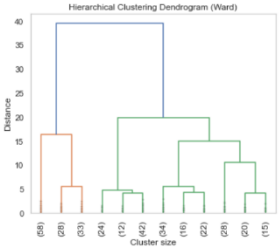


Figure 6

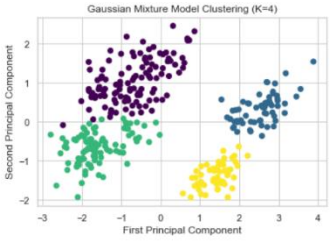


Figure 7

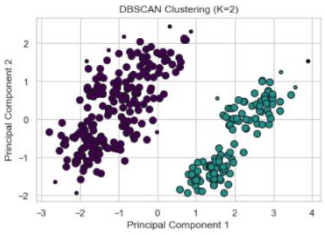


Figure 8

2.0 Supervised Learning

Supervised learning is when well-labelled data are used to instruct or train a machine. “ Which means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples(data) so that the supervised learning algorithm analyses the training data (set of training examples) and produces a correct outcome from labeled data” (GeeksforGeeks, 2024).

2.1 Regression

“Regression is a technique used to capture the relationships between independent and dependent variables, with the main purpose of predicting an outcome” (Tech Target, n.d.).

The dataset used for this task is the ‘Car Price Prediction’ dataset obtained from the Kaggle platform. The data contains information about car features and its prices. The aim is to predict the car price using regression in this task.

2.1.1 Existing Literature

Supply and demand dynamics influence pricing. For instance, an abundance of similar models entering the used car market can depress prices due to increased supply relative to demand. “Some vehicles

are more likely to sell than others, there are some automobile models that also have a better reputation for durability and other aspects, making them ideal options for those wanting to buy a used car, for example, Japanese models tend to have a better reputation for resale than German models, but of course, it is not in all cases" (What are the factors that influence a used car price?, 2021). Kilometers traveled significantly influence a car's value, typically favoring vehicles with lower mileage. While low mileage generally boosts a car's value, usage and mechanical condition will be pivotal factors, determinable through expert inspection and test drives.

2.1.2 Data Preprocessing

Data cleaning involved removing duplicates and filtering out low-price data to prevent unrealistic predictions. New columns such as 'Car Age' and 'Turbocharged' were added, focusing on specific car categories: 'Jeep', 'Hatchback', 'Sedan', and 'Minivan'. Outliers were addressed using the 5th and 95th percentiles method. Transformations like square root and logarithmic were applied to 'Price', 'Mileage', and 'Car Age' to stabilize variance, reduce skewness, and improve symmetry, enhancing modeling and interpretability, as observed in Figures 9 and 10. Numerical variables were standardized, and categorical variables were encoded using one-hot encoding to obtain indicator variables.

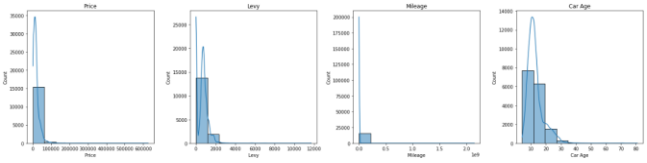


Figure 9

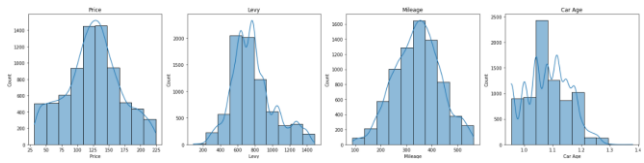


Figure 10

2.1.3 Research Questions

- Which manufacturers produce the most expensive cars on average?
- How has the average price of cars changed over its age?
- Is there a correlation between various car features with car price?
- What is the best regression model to predict car prices?

2.1.4 Exploratory Data Analysis(EDA)

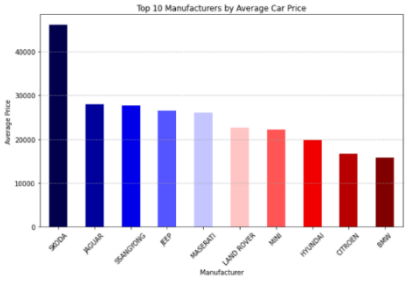


Figure 11

The bar graph in Figure 11 displays the average prices for each manufacturer, with manufacturers sorted based on their average prices. The top 10 manufacturers are selected, revealing brand perception and market positioning. For instance, 'Skoda' emerges as a luxury brand with the highest average prices.

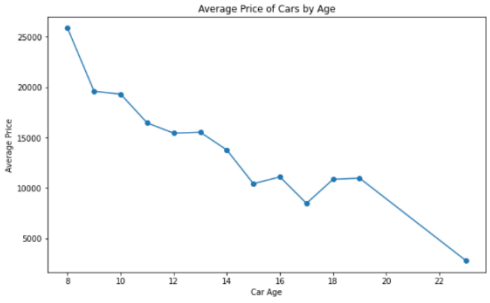


Figure 12

Figure 12 depicts pricing trends over time, illustrating how car age affects average prices. It helps understand depreciation patterns, showing that older cars tend to have

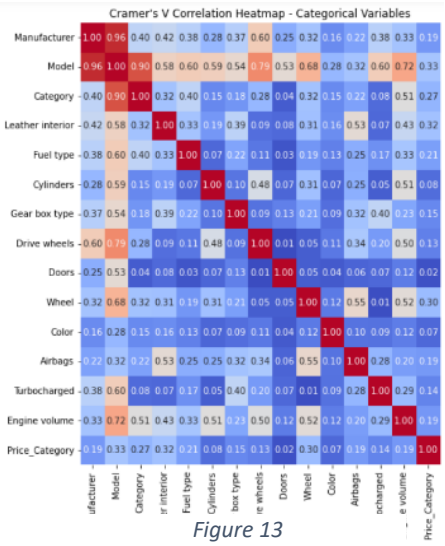


Figure 13

lower prices. In Figure 13, contingency tables are computed, and chi-square tests are conducted to determine the strength of association between categorical variables by deriving Cramer's V statistic. Notably, variables such as 'Model', 'Category', 'Leather Interior', 'Fuel Type', and 'Wheel' show correlations to have a moderate relationship.

2.1.5 Regression Models

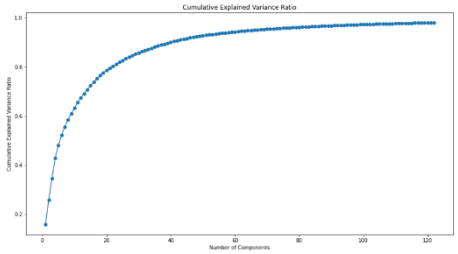


Figure 14

The dataset was split into 80% training and 20% test sets for model training and evaluation. To ensure alignment with assumptions, dimensional reduction via PCA was employed for multiple linear regression. Figure 14 depicts diminishing variance explained as component numbers increase, with marginal changes becoming negligible thereafter. Higher marginal changes indicate greater retained information in principal components.

2.1.5.1 Multiple Linear Regression(MLR)

“To build a MLR model, five key assumptions must be met: linearity, homoskedasticity, independence of errors, normality, and independence of independent variables” (Complete Dissertation, n.d.) Figures 15, 16, and 17 assess for linearity violations. In Figure 15, the alignment of the blue line around zero indicates minimal bias between errors, supporting homoskedasticity. Figure 17's residual plot suggests errors are somewhat normally distributed, affirming this assumption. Additionally, Figure 13's correlation plot reveals no perfect multicollinearity. The scatter plot in Figure 16 demonstrates a significant linear relationship between predictor and actual values, supporting the linearity assumption. Autocorrelation was tested using the Durbin-Watson statistic, yielding a value of 1.906. “The null and alternative Hypothesis are as follows; Null Hypothesis (H0): There is no autocorrelation present in the residuals of the model. Alternative Hypothesis (H1): Autocorrelation exists in the residuals of the model. Since this falls between 1.5 and 2.5, we fail to reject the null hypothesis, indicating no autocorrelation in the residuals” (BOBBITT, 2020). With all linear assumptions satisfied, the model can effectively predict car prices.

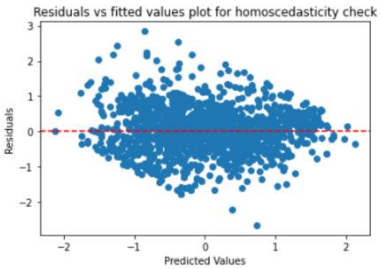


Figure 15

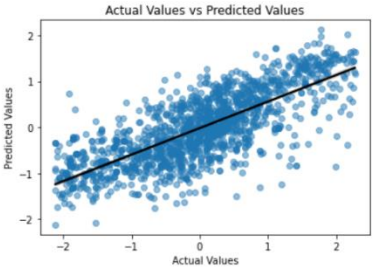


Figure 16

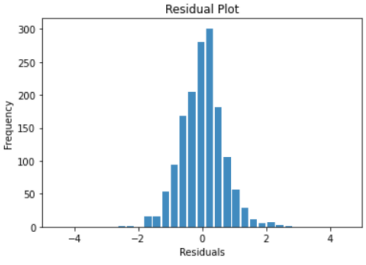


Figure 17

2.1.5.2 Ridge Regression

Ridge regression is effective for addressing multicollinearity and overfitting, making it well-suited for our predictive task. Ridge, a cross-validated variant was used, to determine the optimal regularization parameter (alpha), which balances model complexity to avoid overfitting. The Ridge model was trained using a five-fold cross-validation strategy on the training dataset to ensure robustness. Following model fitting, the optimal alpha value was determined to be 1.0, indicating that moderate regularization is beneficial. We subsequently used the trained model to make predictions on the test dataset.

2.1.5.3 Lasso Regression

Lasso is a linear regression technique that combines variable selection and regularization to enhance prediction accuracy and interpretability. It assigns weights to each feature, indicating their importance in making predictions. However, many coefficients are close to zero, suggesting minimal contribution to predictions. Lasso encourages sparsity by penalizing the absolute size of coefficients, leading to automatic feature selection. This is beneficial for datasets with numerous features, simplifying the model and potentially improving generalization. The best alpha value, found through cross-validation, indicates the level of regularization applied. In this case, the best alpha value is approximately 0.0002978, implying the optimal balance between bias and variance in the model.

2.1.5.4 Random Forest Regressor

“Random Forest Regressor is an ensemble algorithm that combines predictions from multiple models to enhance accuracy” (Geeks for Geeks, n.d.) . At its core, it comprises decision trees, which break down problems based on data features. Each tree is trained on random subsets of data and features to prevent overfitting. After training, predictions from each tree are aggregated through a voting process, often resulting in an average for regression tasks. Random Forest is flexible, handling various data types with minimal preprocessing, and robust to outliers and noisy data. Eventhough it is complex, it offers some interpretability through feature importance analysis. In summary, Random Forest Regressor is versatile, robust, and capable of accurate predictions even with complex datasets.

2.1.5.5 Decision Tree Regressor

It operates by dividing the dataset into smaller subsets, constructing a tree-like structure in the process. Each "node" in the tree represents a question based on a feature of the data. By asking the most relevant questions first, it splits the data into subsets with distinct outcomes. Moving through the tree, it reaches a "leaf node" where it provides the predicted value. This method is straightforward to understand, even for those without a technical background in machine learning. Nevertheless, there is a potential for overfitting, wherein the model becomes overly tailored to the training data, resulting in diminished performance when applied to unseen data. To address this, it's crucial to fine-tune the model parameters and use techniques like cross-validation to ensure reliable performance on unseen data.

2.1.5.6 K-nearest neighbors (KNN) Regression

KNN Regression identifies the nearest data points within the training set to the one targeted for prediction. The parameter "K" denotes the number of closest neighbors considered for prediction. In the Car Price Prediction model, K is configured to 3, signifying that the algorithm identifies the three nearest points and computes their average values to make predictions. KNN Regression offers simplicity in comprehension and implementation, rendering it well-suited for novices in machine learning. It adeptly manages data featuring intricate patterns and refrains from assuming any specific data distribution.

2.1.6 Comparison of Model Results and Conclusion

The following metrics will be used to evaluate performance: R-squared indicates the proportion of variance explained by the regression model. MSE (Mean Squared Error) measures the average squared difference between actual and predicted values in the dataset. RMSE (Root Mean Squared Error) is the square root of MSE, representing the standard deviation of residuals. Lastly, MAE (Mean Absolute Error) calculates the average absolute difference between actual and predicted values in the model.

The R-squared value ranges from 0 to 1, with a strong fit indicated by a value greater than 0.80. In this case, the Random Forest Regressor shows the best fit among all the models assessed. Lower MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error) values also indicate better model fit to the dataset, with RMSE being more effective at capturing larger errors. (Chugh, 2020)

	R Squared	MSE	RMSE	MAE
Multiple Linear Regression	0.58	0.42	0.65	0.50
Ridge Regression	0.60	0.39	0.63	0.48
Lasso Regression	0.60	0.40	0.63	0.49
Random Forest Regressor	0.82	0.18	0.18	0.27
Decision Tree Regressor	0.65	0.35	0.59	0.35
KNN Regression	0.73	0.26	0.51	0.32

When comparing model results, the Random Forest Regressor stands out due to its high R-squared, low MAE, and low RMSE. The Reason for MLR to perform low and Random Forest Regressor to be the best can be due to;

- MLR assumes a linear relationship between features and the target variable, which may not capture complex data patterns as effectively as Random Forest, capable of handling non-linear relationships.
- Random Forest can effectively manage feature interactions and identify complex patterns within the data, resulting in a better fit.
- MLR might overfit if there are too many irrelevant features, while Random Forest, with its ensemble of decision trees, is less prone to overfitting as it averages predictions across multiple trees.

In conclusion, the Random Forest Regressor likely outperforms MLR due to its ability to capture complex relationships and interactions in the data, along with its resilience to overfitting.

2.2 Classification

“Classification is a fundamental task in machine learning that involves categorizing data into different classes or categories based on their features or characteristics where algorithms are trained on labeled datasets, with input variables and corresponding class labels” (AYADATA, 2023) . By identifying patterns in the data, the classifier can assign new instances to the appropriate class. The output is discrete, meaning it falls into distinct categories or classes.

The dataset used for this task is the "Airline Passenger Satisfaction" dataset obtained from the Kaggle platform. This dataset contains a US airline passenger satisfaction survey. The final goal of this task is to classify whether the passenger is satisfied or not.

2.2.1 Existing Literature

“Tsafarakis et al. (2017) suggested that enhancing onboard entertainment and Wi-Fi services could enhance airline passenger satisfaction, as per a multi-standard satisfaction analysis method. Hess (2018) studied these factors individually across various market segments and concluded that visit times, flight durations, and airfares are crucial for both business and leisure travelers. Lucini et al. (2019) utilized text mining to analyze online customer feedback, predicting passenger attitudes. They recommended tailored customer service for first-class and premium economy passengers, and

highlighted aspects like checked baggage, wait times, cabin staff, in-flight service, and cost performance as key dimensions in predicting airlines recommended by passengers. Brochado (2020), through quantitative content analysis of airline passenger web reviews using Leximancer, found that on-board service, airport operations, ground service, and other factors significantly impact service quality assessment" (Scientific Reports, 2022).

2.2.2 Data Preprocessing

The data preprocessing steps involved dropping duplicated rows and null values, as well as removing ratings with a value of 0 (indicating "not applicable"). Outliers were also identified and eliminated using the interquartile range (IQR) method. Additionally, numerical variables underwent Box-Cox transformation to achieve a more symmetrical distribution and stabilize variance, which is beneficial for improving the performance of statistical models. These steps collectively aim to ensure data cleanliness, address anomalies, and enhance the suitability of the dataset for subsequent analyses. Furthermore, numerical values were standardized, and categorical variables were encoded using one-hot encoding to obtain indicator variables.

2.2.3 Research Questions

- How does passenger satisfaction vary with flight distance and age group, stratified by Class of travel?
- Which aspects of the flight experience (e.g., inflight entertainment, cleanliness) have the most significant impact on passenger satisfaction?
- What is the best classification model to predict satisfaction?

2.2.4 Exploratory Data Analysis(EDA)

Figure 18 illustrates that children and teenagers in the Eco class shows satisfaction, followed by middle-aged individuals up to their 60s who prefer mostly the Business class. Older individuals tend to find satisfaction mostly in the Eco class.

Figure 18

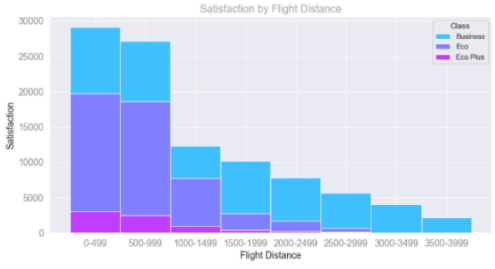


Figure 19

Figure 19 demonstrates that satisfaction levels, particularly in the Business class, exists with longer flight distances.

In Figure 20, contingency tables are computed, and chi-square tests are conducted to determine the strength of association between categorical variables by Cramer's V statistic. We can observe that 'Inflight wifi service,' 'Seat Comfort,' 'Online Boarding,' and 'Inflight entertainment' have a correlation of more than 0.4 showing a moderate to strong relationship

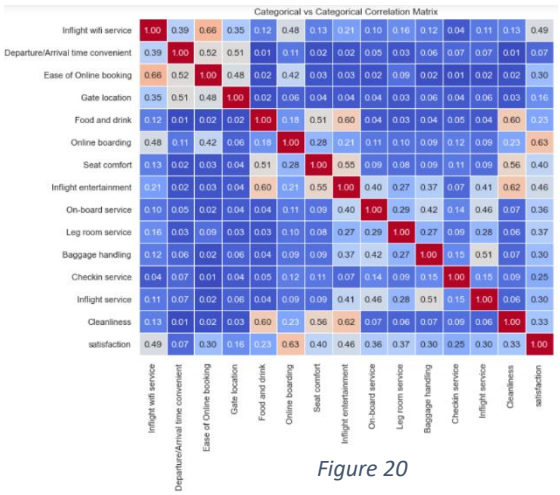


Figure 20

2.2.5 Classification Models

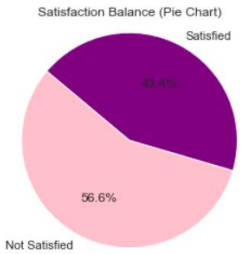


Figure 21

The dataset was split into 80% training and 20% test sets for model training and evaluation. The predictor variable 'satisfaction' indicates a distribution of 56.6% and 43.4% for the two classes, respectively, suggesting that the classes are approximately balanced. Therefore, resampling is not needed.

2.2.5.1 Decision Tree Classifier

Decision Trees ask if-else questions about the data, splitting it into groups to predict outcomes. The top (root) node starts with the most important feature, leading to branches based on features. Terminal nodes (leaves) give predictions. Entropy measures data disorder, indicating how well a feature divides data. Information Gain assesses a feature's effectiveness in reducing uncertainty about class labels, helping to improve prediction accuracy.

2.2.5.2 Logistic Regression

Similar to linear regression, predicts binary outcomes like passenger satisfaction. Using a logistic function, it calculates probabilities ranging from 0 to 1. The model estimates coefficients for predictor variables, showing their impact on satisfaction. These coefficients indicate how much the log-odds of satisfaction change with a one-unit increase in predictors. To predict outcomes, the model calculates probabilities. If the probability exceeds a threshold (typically 0.5), it predicts satisfaction; otherwise, it predicts dissatisfaction.

2.2.5.3 Random Forest Classifier

Random Forest operates like an ensemble of decision trees, with each tree being trained on a distinct subset of the data. This approach aids in mitigating overfitting. During the construction of each tree, only a limited set of random features is assessed at each step, contributing to the diversity and robustness of the model.Random Forest also tells us which features are most important for making predictions and gives more reliable results by taking a vote or average from all the trees. So, to make a prediction, the data goes through each tree, and the final result is based on what most of the trees predict.

2.2.5.4 Naïve Bayes Classifier

The Multinomial Naive Bayes (MNB) classifier assumes feature independence given the class label, simplifying analysis for real-world datasets. It performs well with count or frequency-based features, such as term frequencies. During training, MNB estimates feature probabilities for each class and uses them to calculate the likelihood of observing input features for each class, incorporating prior class probabilities. In terms of interpretation, MNB provides insights into feature importance for class differentiation. By examining learned probabilities, significant features can be identified. For predictions, MNB calculates the likelihoods of observing input features for each class based on training-learned probabilities. It then applies Bayes' theorem to determine the posterior probability of each class given the input features, predicting the class with the highest posterior probability.

2.2.5.5 Support Vector Machines (SVM)

SVM emerge as a formidable model, particularly effective when dealing with data that doesn't exhibit clear separability along a linear boundary. Rather than attempting to delineate classes with a straight line or plane, SVM seeks out the optimal hyperplane that maximizes the margin between them. Support vectors, representing data points closest to this boundary, play a pivotal role in delineating it and facilitating predictions. SVM can transform data into higher-dimensional spaces using different kernel functions. In this analysis, a polynomial kernel of degree 5 is used, allowing the SVM to capture

complex relationships. After fitting the SVM to the training data, it learns the optimal hyperplane, which is then used to predict new data points. The random state parameter ensures result reproducibility.

2.2.5.6 Extreme Gradient Boosting (XG Boost)

XG Boost belongs to the boosting algorithms, a type of ensemble method that improves accuracy by combining predictions from multiple base estimators. Boosting sequentially trains weak learners, like decision trees, with each one focusing on the errors of its predecessors. XG Boost enhances traditional gradient boosting by using a more regularized model to control overfitting and handle sparse data better. The objective of XG Boost is to minimize a loss function while penalizing model complexity to prevent overfitting. It achieves this by adding new weak learners iteratively, each aiming to reduce the overall model loss. A notable feature of XG Boost is its ability to handle missing values internally, making it robust to datasets with incomplete information. Additionally, XG Boost supports parallel processing, making it efficient for large datasets.

2.2.6 Comparison of Model Results and Conclusion

The performance will be compared using four score parameters;

- Accuracy score: This measures the proportion of correct predictions made by the model, considering both true positives and true negatives. It indicates the overall correctness of the model's predictions.
- Precision score: This evaluates the proportion of positively predicted instances that are actually correct. It helps assess the reliability of the model's positive predictions.
- Recall score: Also known as true positive rate or sensitivity, this metric measures the model's ability to correctly identify positive instances out of all actual positive instances.
- F1 score: This score is calculated as the harmonic mean of precision and recall, providing a balanced measure of a model's performance in terms of both precision and recall. (II, n.d.)

	Accuracy	Precision	Recall	F1 Score
Decision Tree Classifier	94.76%	0.95	0.95	0.95
Logistic Regression	93.16%	0.93	0.93	0.93
Random Forest Classifier	96.06%	0.96	0.96	0.96
Naïve Bayes Classifier	88.61%	0.88	0.88	0.88
SVM	95.93%	0.96	0.96	0.96
XG Boost	96.16%	0.96	0.96	0.96

All models exhibit over 80% accuracy, highlighting their strong performance. Notably, XG Boost leads with an impressive 96.16% accuracy, emphasizing its effectiveness in predicting outcomes.

A ROC curve illustrates the trade-off between a binary classifier's true positive rate (sensitivity) and false positive rate across various thresholds. A curve near the top-left corner signifies superior performance. The area under the curve summarizes overall model performance, with higher values indicating better discrimination ability. In this case, XG Boost outperforms other models.

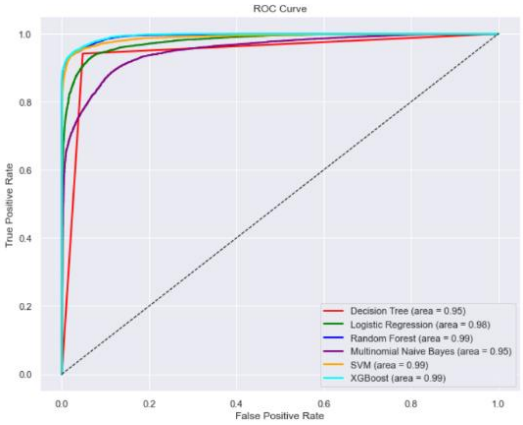


Figure 22

3.0 Bibliography

(n.d.). Retrieved 04 02, 2024, from IBM: https://www.ibm.com/topics/unsupervised-learning

(n.d.). Retrieved 04 03, 2024, from Tech Target: https://www.techtarget.com/searchenterpriseai/feature/What-is-regression-in-machine-learning

(n.d.). Retrieved 04 03, 2024, from Complete Dissertation: https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-multiple-linear-regression/

(n.d.). Retrieved 04 03, 2024 , from Geeks for Geeks: https://www.geeksforgeeks.org/random-forest-regression-in-python/

(2022, 07 01). Retrieved 04 02, 2024, from Scientific Reports: https://www.nature.com/articles/s41598-022-14566-3

(2023, 11 1). Retrieved 04 03, 2024, from AYADATA: https://www.ayadata.ai/blog-posts/data-classification-in-machine-learning/

(2024, 3 13). Retrieved 3 29, 2024, from GeeksforGeeks: https://www.geeksforgeeks.org/supervised-unsupervised-learning/

BOBBITT, Z. (2020, 07 21). Retrieved 04 03, 2024, from STATOLOGY: https://www.statology.org/durbin-watson-test-python/

CarSellZone. (2021, 12 05). Retrieved 04 01, 2024, from https://carsellzone.com/blog/detail/factors-affect-car-price

Chugh, A. (2020, 12 8). Retrieved 04 03, 2024, from Medium: https://medium.com/analytics-vidhya/mae-mse-rmse-coefficient-of-determination-adjusted-r-squared-which-metric-is-better-cd0326a5697e

II, S. M. (n.d.). Retrieved 04 03, 2024, from KLU: https://klu.ai/glossary/accuracy-precision-recall-f1

Juan A. Amat, J. V. (n.d.). Sexing Chinstrap Penguins (Pygoscelis antarctica) by Morphological Measurements. *Colonial Waterbirds*. Retrieved 04 02, 2024

Dataset Links;

KLEIN, T. (n.d.). Retrieved 04 02, 2024, from Kaggle: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

ABOELWAFA, Y. (n.d.). Retrieved 04 02, 2024, from Kaggle: https://www.kaggle.com/datasets/youssefaboelwafa/clustering-penguins-species

CONTRACTOR, D. (n.d.). Retrieved 04 02, 2024, from Kaggle: https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge/data