Machine Learning Project

by

Syarafana Begum

3 April 2025

Page Count: 10 excluding cover page, table of contents and references

Table of Contents

1. Introduction

Machine Learning is a collection of techniques that can automatically identify patterns in data and make predictions based on them (Murphy, 2013). It includes supervised learning—where labeled datasets train algorithms for data classification and accurate outcome prediction—and unsupervised learning—where unlabeled datasets use algorithms for detection of hidden data patterns without human intervention (IBM, 2021).

2. Unsupervised Learning

2.1. Substantive Issue

Global well-being is a growing area of interest for policy makers, international organizations, and researchers. Understanding the factors that differentiate happier nations from less happy ones can help guide resource allocation and inform policy decisions aimed at improving citizens' quality of life. The World Happiness Report data provides a snapshot of various happiness-related indicators—such as GDP per capita, healthy life expectancy, social support, and more—across different countries. By using unsupervised learning, we can discover hidden groupings of countries that share similar socio-economic and well-being characteristics, without relying on any predefined labels.

2.2. Research Questions

The analysis addresses the following research questions (RQ):

- **RQ1:** How many natural clusters of countries emerge when considering the happiness-related indicators?
- **RQ2**: What distinct characteristics define each cluster in terms of economic, social, and health factors?
- **RQ3**: Which groups of countries appear to have higher well-being metrics, and what insights can be drawn to potentially replicate their situations?

2.3. Dataset & Variables

The original happiness dataset comprises 156 rows and 9 columns, with no missing or duplicate data. Descriptive statistics generally indicate a positively skewed distribution, with mean values exceeding medians, suggesting a tail towards higher values. The correlation matrix plotted shows a significant positive correlation between GDP per capita and Healthy life expectancy, along with a slight negative correlation between GDP per capita and Generosity.

Variable Name	**Variable Description**
Overall rank	Rank of the country/region based on the Happiness Score
Country or region	Name of the country/region
Score	A metric measured in 2019 by asking the sampled people the question: "How would you rate your happiness on a scale of 0 to 10 where 10 is the happiest."
GDP per capita	The extent to which GDP contributed to the calculation of the Happiness Score
Social support	The extent to which social support contributed to the calculation of the Happiness Score
Healthy life expectancy	The extent to which healthy life expectancy contributed to the calculation of the Happiness Score

Freedom to make life choices	The extent to which freedom to make life choices contributed to the calculation of the Happiness Score
Generosity	The extent to which generosity contributed to the calculation of the Happiness Score
Perceptions of corruption	The extent to which perceptions of corruption contributed to the calculation of the Happiness Score

Table 1: Happiness Dataset Variables Description

2.4. Methodology

The task objective is to group countries in the happiness dataset based on similarities and differences in well-being indicators, thereby revealing underlying patterns or subgroups. Unsupervised learning techniques are applied, including principal component analysis (PCA) for dimensionality reduction and clustering algorithms such as K-means and hierarchical clustering for data organisation and visualisation. This approach facilitates the classification of countries into distinct clusters that reflect varying levels of socio-economic development and overall happiness. The optimal number of clusters is determined using the silhouette and elbow methods.

2.5. Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms correlated variables into a smaller set of uncorrelated components, retaining as much information as possible. After data scaling, four principal components were chosen, collectively explaining 93% of the overall variance and thus reducing dimensionality, as illustrated in Figure 1.1. According to the PCA presented in Figure 1.2, PC1 represents “quality of life” as it is heavily loaded to GDP per capita, Social support and Healthy life expectancy. PC2 represents “social capital” as it is heavily loaded onto Freedom to make life choices, Generosity and Perceptions of corruption. PC3 represents “trust and freedom” since Perceptions of corruption, Generosity and Social support load heavily on PC3. PC4 represents “personal freedom” as Freedom to make life choices is most heavily loaded on PC4. Among all principal components, PC1 accounts for the largest share of variability in the dataset, as depicted in Figure 1.3.

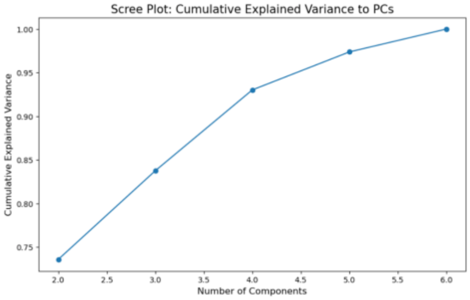
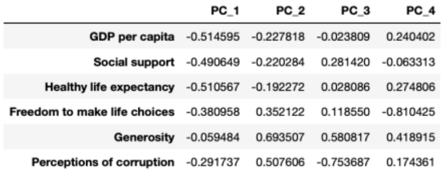
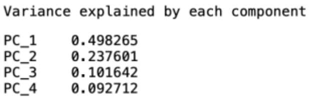


Figure 1.1: Scree Plot



	PC_1	PC_2	PC_3	PC_4
GDP per capita	-0.514595	-0.227818	-0.023809	0.240402
Social support	-0.490649	-0.220284	0.281420	-0.063313
Healthy life expectancy	-0.510567	-0.192272	0.028086	0.274806
Freedom to make life choices	-0.380958	0.352122	0.118550	-0.810425
Generosity	-0.059484	0.693507	0.580817	0.418915
Perceptions of corruption	-0.291737	0.507606	-0.753687	0.174361

Figure 1.2: Principal Component Analysis

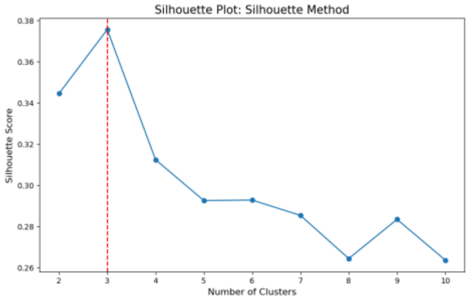


```
Variance explained by each component

PC_1     0.498265
PC_2     0.237601
PC_3     0.101642
PC_4     0.092712
```

Figure 1.3: Variances

K-means clustering partitions data into groups by minimising the squared distance between each data point and its assigned cluster centroid. The algorithm works iteratively by assigning points to the nearest centroid and recalculating the centroid positions until they stabilize. It uses Euclidean distance as a similarity measure. To determine the most suitable number of clusters, both the Elbow Method and Silhouette Analysis were applied, as shown in Figure 1.4 and Figure 1.5. The Elbow Method evaluates how the within-cluster sum of squared errors (inertia) changes with different values of *k*. The optimal *k* is identified where the reduction in inertia starts to level off—this “elbow point” appears at k = 3. Similarly, the highest silhouette score is also observed at k = 3, indicating well-separated and cohesive clusters. Based on these findings, three clusters were chosen for further analysis.



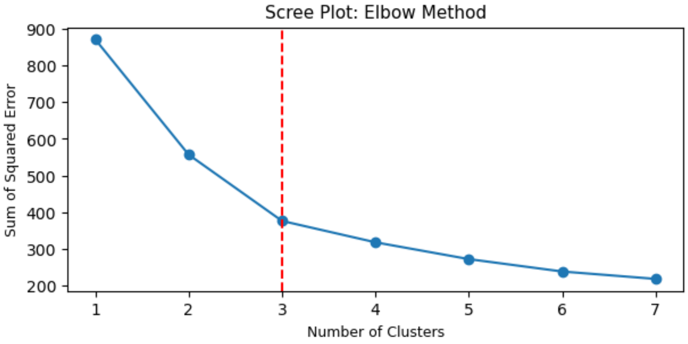
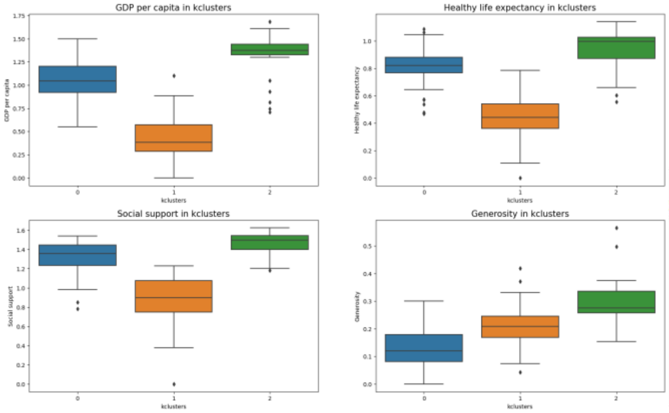


Figure 1.4: Silhouette Plot

Figure 1.5: Elbow Method

To identify the cluster representing the highest overall well-being, we referred to PC1 scores. PC1 captures key dimensions such as GDP per capita, income, life expectancy, and child mortality—indicators closely tied to national happiness. Cluster 2 has the highest GDP per capita, healthy life expectancy, social support and lowest generosity, as depicted in Figure 1.6. This cluster comprises developed countries, primarily located in North America, Oceania, with a minor presence in Europe, as illustrated in Figure 1.7. Meanwhile, Cluster 1 has the lowest GDP per capita, healthy life expectancy, social support and moderate generosity, representing underdeveloped countries mostly across Africa and Asia. Cluster 0, depicting developing countries, lies across North America, South America, Europe, Asia and Africa.



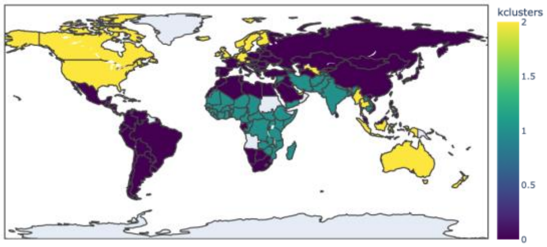


Figure 1.6: Variables in Clusters

Figure 1.7: World Map in Clusters

Hierarchical clustering is an unsupervised learning method that builds nested clusters by progressively merging the most similar groups, as visualised through the dendrogram in Figure 1.8. It uses a bottom-up approach, starting by measuring the distance between individual data points and gradually combining the closest ones into larger clusters. This process continues until all data points are grouped into a single cluster. In this analysis, we applied hierarchical clustering using Euclidean distance and Ward's linkage method, which minimises within-cluster variance at each merging step to form similarity-based groupings.

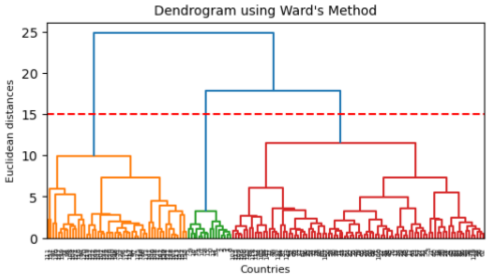
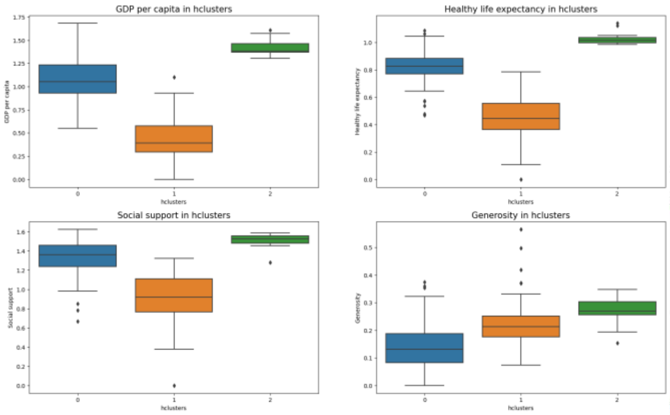


Figure 1.8: Dendrogram

Cluster 2 demonstrated the strongest outcomes, characterised by the highest levels of GDP per capita, healthy life expectancy, social support and generosity, indicating it represents developed countries. As shown in Figure 1.9 and Figure 1.10, these countries are mainly concentrated in North America and Oceania, with some representation in Europe.

In contrast, Cluster 1 had the least favourable results, including the lowest GDP per capita, healthy life expectancy, social support and generosity, suggesting it comprises underdeveloped countries. This cluster is primarily found in regions of Africa and Asia.

Cluster 2 displayed moderate values across all indicators, representing developing countries. It includes countries from South America, parts of Europe and Asia, and North Africa, reflecting transitional economies with mid-range well-being scores.



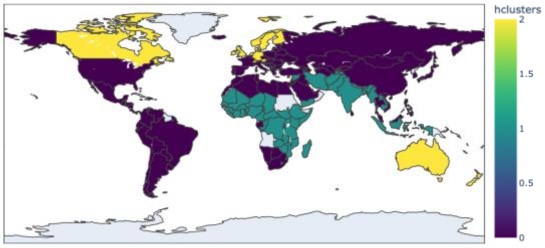


Figure 1.9: Variables in Clusters

Figure 1.10: World Map in Clusters

2.6. Results & Recommendations

RQ1: How many natural clusters of countries emerge when considering the happiness-related indicators?

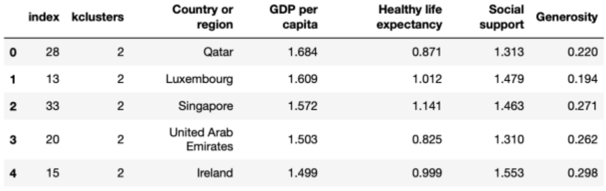
- The Silhouette and Elbow methods indicated that the optimal number of clusters is three, which correspond to groups of developed, developing, and less developed countries.

RQ2: What distinct characteristics define each cluster in terms of economic, social, and health factors?

- Based on the PCA analysis, each principal component captures a distinct dimension of well-being and socio-economic development. PC1 represents “quality of life”, as it is strongly influenced by GDP per capita, social support, and healthy life expectancy. PC2 reflects “social capital”, with high loadings on freedom to make life choices, generosity, and perceptions of corruption. PC3 highlights aspects of “trust and freedom”, being heavily associated with perceptions of corruption, generosity, and social support. Lastly, PC4 represents “personal freedom”, as it is most strongly linked to freedom to make life choices.

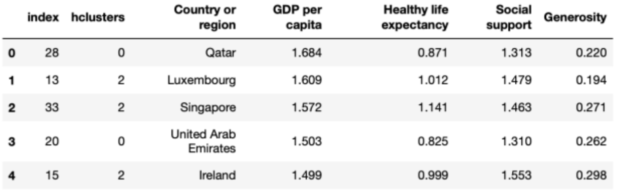
RQ3: Which groups of countries appear to have higher well-being metrics, and what insights can be drawn to potentially replicate their situations?

- Figures 1.11 and 1.12 illustrate that both K-Means and Hierarchical clustering produced comparable results, consistently grouping countries such as Qatar, Luxembourg, Singapore, United Arab Emirates and Ireland together. These nations share characteristics commonly associated with development, including high GDP per capita, healthy life expectancy, social support and low generosity.



	index	kclusters	Country or region	GDP per capita	Healthy life expectancy	Social support	Generosity
0	28	2	Qatar	1.684	0.871	1.313	0.220
1	13	2	Luxembourg	1.609	1.012	1.479	0.194
2	33	2	Singapore	1.572	1.141	1.463	0.271
3	20	2	United Arab Emirates	1.503	0.825	1.310	0.262
4	15	2	Ireland	1.499	0.999	1.553	0.298

Figure 1.11: K-Means Recommendations



	index	hclusters	Country or region	GDP per capita	Healthy life expectancy	Social support	Generosity
0	28	0	Qatar	1.684	0.871	1.313	0.220
1	13	2	Luxembourg	1.609	1.012	1.479	0.194
2	33	2	Singapore	1.572	1.141	1.463	0.271
3	20	0	United Arab Emirates	1.503	0.825	1.310	0.262
4	15	2	Ireland	1.499	0.999	1.553	0.298

Figure 1.12: Hierarchical Recommendations

3. Regression

3.1. Substantive Issue

The resale value of a second-hand car is influenced by multiple factors, including brand reputation, model specifications, mileage, fuel type, and age. Understanding these relationships is crucial for buyers and sellers to make informed decisions. Car dealerships, online marketplaces, and individual sellers can benefit from predictive models that estimate the fair market value of a used vehicle based on its specifications. Given the numerous factors influencing a car's resale value, regression analysis is a powerful tool for deriving insights from historical sales data and predicting future prices.

3.2. Research Questions

The analysis addresses the following research questions (RQ):

- **RQ1:** Which regression model provides the most accurate prediction of second-hand car prices?
- **RQ2:** What are the key factors influencing the selling price of a used car?

3.3. Dataset & Variables

The original used car dataset comprises 100 rows and 13 columns, with no missing or duplicate data. Descriptive statistics indicate a positively skewed distribution, with mean exceeding the median.

Variable Name	**Variable Description**
Car_ID	A unique identifier for each car listing
Brand	The brand or manufacturer of the car (e.g., Toyota, Honda, Ford, etc.)
Model	The model of the car (e.g., Camry, Civic, Mustang, etc.)
Year	The manufacturing year of the car
Kilometers_Driven	The total kilometers driven by the car
Fuel_Type	The type of fuel used by the car (e.g., Petrol, Diesel, Electric, etc.)
Transmission	The transmission type of the car (e.g., Manual, Automatic)
Owner_Type	The number of previous owners of the car (e.g., First, Second, Third)
Mileage	The fuel efficiency of the car in kilometers per liter
Engine	The engine capacity of the car in CC (Cubic Centimeters)
Power	The maximum power output of the car in bhp (Brake Horsepower)
Seats	The number of seats available in the car
Price	The selling price of the car in INR (Indian Rupees), which is the target variable to predict

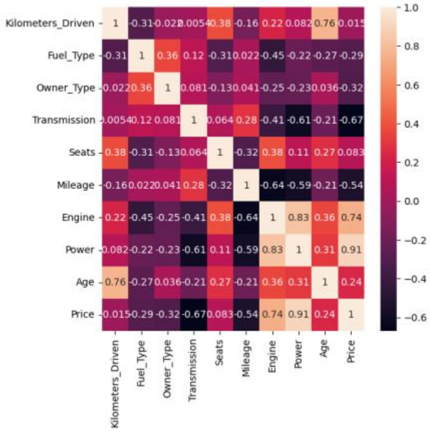
Table 2: Used Car Dataset Variables Description

3.4. Methodology

Exploratory data analysis (EDA) will be conducted on the vehicle dataset to explore the relationships between input features and the target variable. Regression models including linear regression, gradient boosting, and random forest will be used to predict selling prices. These models will be assessed and compared to determine the most effective method for price prediction. Additionally, feature importance analysis will be carried out to identify the variables that most significantly influence selling price and should be prioritized in the final model. This approach will support the selection of the best-performing machine learning model while uncovering key drivers of price.

3.5. Analysis

A correlation matrix was first generated to explore the relationships among variables. It revealed that power has strong positive correlation with price, suggesting that greater power generally leads to higher prices, as illustrated in Figure 2.1. To reduce the risk of skewing the model, data points with engine exceeding 4,000 or power over 350 were considered outliers and subsequently removed, as shown in Figure 2.2. Prior to splitting the training and test sets 80:20, data preparation was conducted by removing irrelevant ‘Car_ID’, ‘Brand’ and ‘Model’ fields, reorganising columns to a given order, encoding categorical variables and excluding ‘Price’ and ‘Kilometers_Driven’ from features.



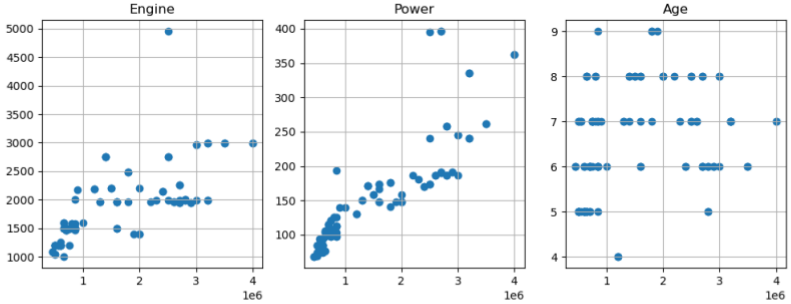
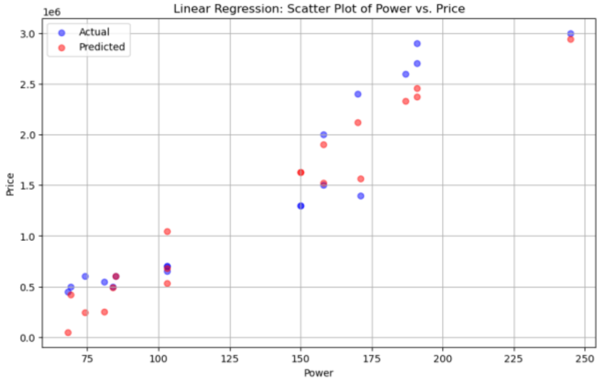


Figure 2.1: Correlation Matrix

Figure 2.2 Scatter Plot of Engine, Power and Age

Linear regression is a widely used algorithm in machine learning that models a continuous target variable based on a linear relationship with several input features. It uses the least squares method to calculate coefficients that minimise error between actual and predicted values. The model assumes normally distributed errors and linear dependencies between predictors and the target. However, as shown in Figures 2.3 and 2.4, the predicted values significantly deviate from the actual ones, and the data points are not tightly aligned with the regression line. This suggests that linear regression may not be the most suitable approach for this dataset.



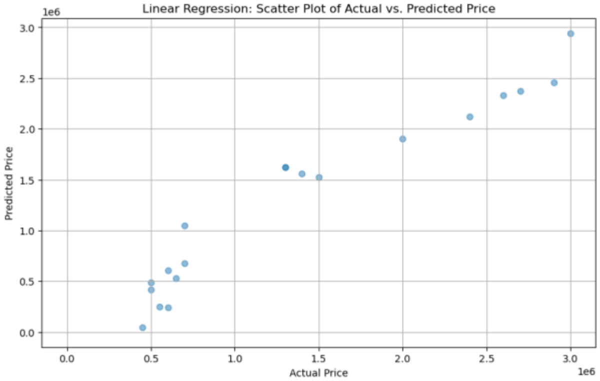


Figure 2.3: Scatter Plot of Power vs Price

Figure 2.4: Scatter Plot of Actual vs Predicted Price

Gradient boosting is a machine learning method that builds an ensemble of weak learners—typically decision trees—by sequentially minimising a loss function such as mean squared error. Each new model is trained on the residuals of the previous ones, allowing the algorithm to progressively correct errors while using techniques like shrinkage to prevent overfitting. This step-by-step learning process results in a robust model with lower bias and variance than individual models. As shown in Figures 2.5 and 2.6, the predicted values from gradient boosting align more closely with the actual values, and the data points are more tightly clustered around the regression line compared to linear regression. This indicates that gradient boosting offers greater predictive accuracy.

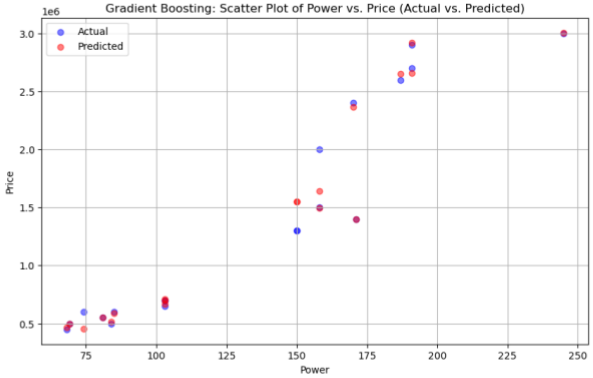


Figure 2.5: Scatter Plot of Power vs Price

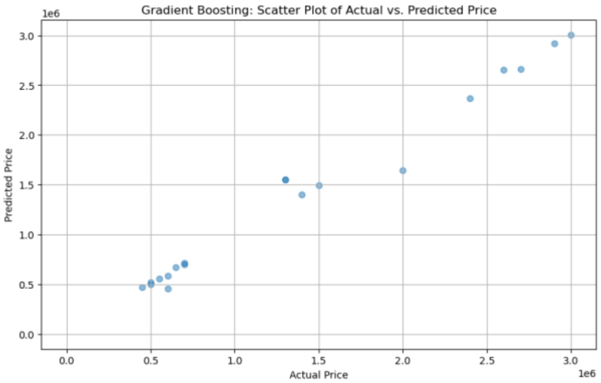


Figure 2.6: Scatter Plot of Actual vs Predicted Price

Random forest is an ensemble learning technique that builds multiple decision trees using bootstrapped subsets of the data and randomly selected features for each split, which helps reduce overfitting seen in single decision trees. It combines the outputs of all trees—averaging in regression tasks or majority voting in classification—to produce final predictions. A validation curve, as shown in Figure 2.7, was used to assess training and cross-validation scores across different tree counts, leading to the selection of 500 trees for optimal performance.

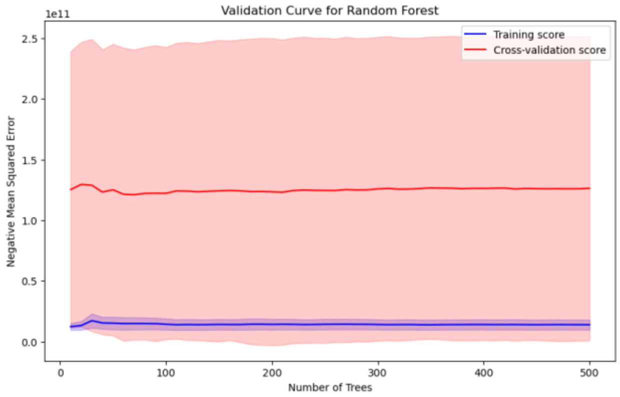


Figure 2.7: Validation Curve for Random Forest

Compared to linear regression, the predicted values from the random forest model were more closely aligned with actual values, with points clustering nearer to the regression line (Figure 2.8 and Figure 2.9). However, to determine whether random forest outperforms gradient boosting, performance metrics must be further analysed.

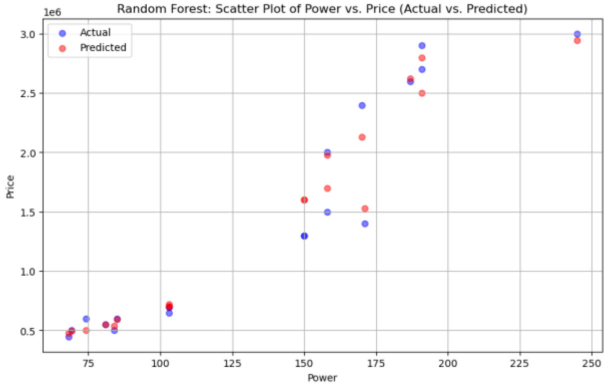


Figure 2.8: Scatter Plot of Power vs Price

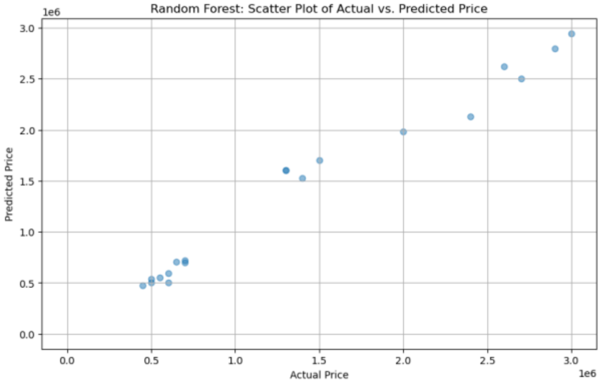
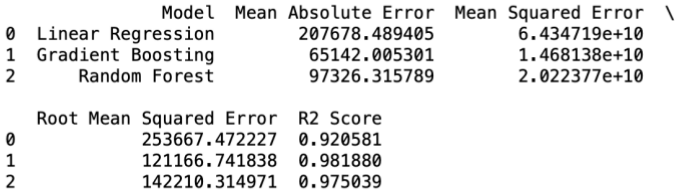


Figure 2.9: Scatter Plot of Actual vs Predicted Price

Linear Regression was used as a benchmark model, showing a high error with a Mean Absolute Error (MAE) of 207,678.49 and a Mean Squared Error (MSE) of 6.4347×10^{10}, accounting for about 92.06% of the variance (Figure 2.10). Gradient Boosting performed better, achieving much lower MAE, MSE, and RMSE values, along with an R² score of 0.981880—indicating it explained roughly 98.2% of the variance. Although Random Forest produced slightly higher error metrics compared to Gradient Boosting, it still achieved a strong R² score of 0.975039, capturing around 97.5% of the variance. Overall, Gradient Boosting emerged as the most effective model for predicting used car prices, with the highest accuracy and lowest error rates.



```
               Model  Mean Absolute Error  Mean Squared Error  \
0  Linear Regression        207678.489405        6.434719e+10
1  Gradient Boosting         65142.005301        1.468138e+10
2      Random Forest         97326.315789        2.022377e+10

   Root Mean Squared Error  R2 Score
0            253667.472227  0.920581
1            121166.741838  0.981880
2            142210.314971  0.975039
```

Figure 2.10: Evaluation of Models

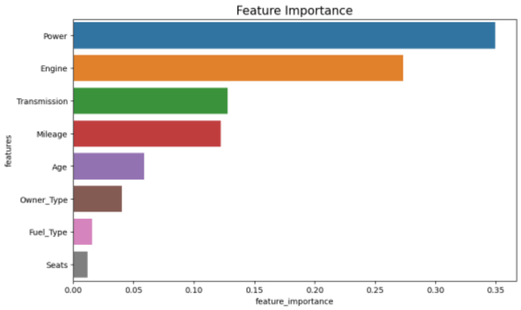


Figure 2.11: Feature Importance

The Random Forest model was also analyzed for feature importance to assess how much each input variable contributed to predicting car prices. Prior to this, hyperparameter tuning was conducted using maximum depth and grid search cross-validation. The resulting feature importance scores revealed that *Power* (0.349566) and *Engine* (0.273362) were the most influential variables in generating accurate predictions, while *Seats* (0.011661) had the least impact on the model's performance.

3.6. Results & Recommendations

RQ1: Which regression model provides the most accurate prediction of second-hand car prices?

- Gradient Boosting delivered the strongest performance among all evaluated models, achieving the lowest error values and the highest R² score, highlighting its accuracy and reliability in forecasting vehicle selling prices.

RQ2: What are the key factors influencing the selling price of a used car?

- The most influential factors in determining a car's selling price are Power, followed by Engine and Transmission.

4. Classification

4.1. Substantive Issue

Employee retention and workforce stability are crucial for organisational success. High employee turnover can result in increased hiring costs, loss of institutional knowledge, and disruptions to business operations. Identifying factors that contribute to employee attrition can help organizations develop better retention strategies, improve job satisfaction, and optimize workforce planning.

Machine learning classification techniques can be applied to predict whether an employee is likely to leave a company based on factors such as education, experience, salary tier, and job history. This study aims to build a classification model that assists in proactive employee retention efforts.

4.2. Research Questions

The analysis addresses the following research questions (RQ):

- **RQ1:** Which classification model provides the highest accuracy in predicting employee attrition?
- **RQ2:** What are the key factors influencing an employee’s decision to leave the company?

4.3. Dataset & Variables

The original employee dataset comprises 4653 rows and 9 columns, with no missing values. There is duplicate data but has been removed.

Variable Name	**Variable Description**
Education	The educational qualifications of employees, including degree, institution, and field of study
JoiningYear	The year each employee joined the company, indicating their length of service
City	The location or city where each employee is based or works
PaymentTier	Categorisation of employees into different salary tiers
Age	The age of each employee, providing demographic insights
Gender	Gender identity of employees, promoting diversity analysis
EverBenched	Indicates if an employee has ever been temporarily without assigned work
ExperienceInCurrentDomain	The number of years of experience employees have in their current field
LeaveOrNot	a target column

Table 3: Employee Dataset Variables Description

4.4. Methodology

Exploratory data analysis (EDA) will be performed on the employee dataset to examine the relationships between input features and the target variable. Following this, machine learning models including logistic regression, random forest, and K-nearest neighbors will be applied to predict the likelihood of heart disease. The performance of these models will be evaluated and compared to determine the most effective method for predicting attrition. Additionally, key features contributing to prediction accuracy will be identified to enhance the final model. This overall process will help in selecting the best machine learning approach for forecasting attrition and understanding its main influencing factors. A comparison with existing literature review (Analytics Vidhya, 2021) can provide helpful context by comparing the performance metrics of machine learning models applied to employee attrition, particularly in evaluating how well models such as random forest and logistic regression performed in terms of accuracy and feature importance.

4.5. Analysis

A correlation heatmap was first generated to identify strong positive or negative relationships among the variables. It revealed a positive correlation between the LeaveOrNot target variable and JoiningYear, as illustrated in Figure 3.1. Further analysis of attrition trends based on age and gender indicated that most employees who left were between the ages of 24 and 31 (Figure 3.2), and that females were more likely to leave compared to males (Figure 3.3). The dataset was then preprocessed by separating numerical and categorical features, and examining the distribution of the target variable, which showed that the majority of employees stayed. Standardization was performed using StandardScaler, and categorical variables like Age and ExperienceInCurrentDomain were encoded using one-hot encoding to prepare the data for modeling.

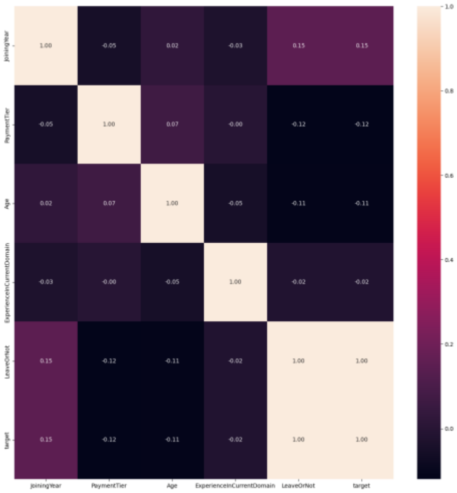
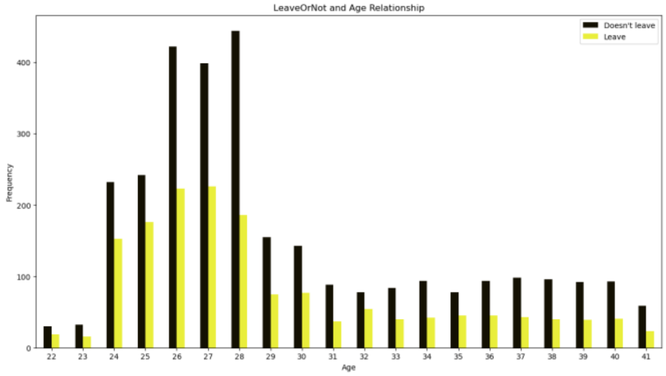


Figure 3.1: Correlation Heatmap



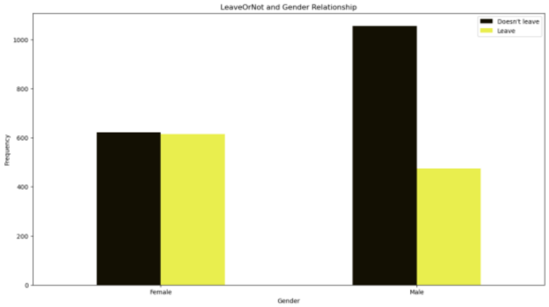
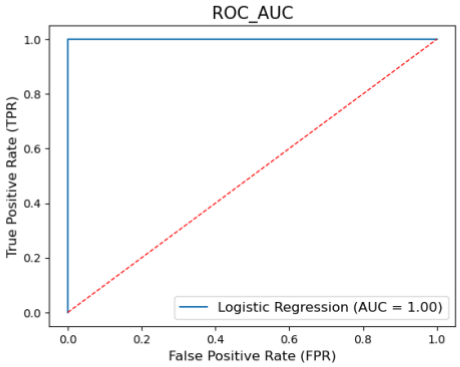


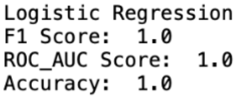
Figure 3.2: LeaveOrNot and Age Relationship

Figure 3.3: LeaveOrNot and Gender Relationship

Categorical columns such as education, joining year, payment tier, and gender were transformed into dummy variables using get_dummies, while any miscategorized columns from the process were removed. The dataset was then split into training and testing sets with a 70:30 ratio. A model evaluation function was created to handle predictions, compute evaluation metrics, plot the ROC and precision-recall curves, and display key performance scores including F1 score, ROC AUC, and accuracy. It also generated a classification report and visualized the confusion matrix. This function was applied to the dummy-encoded data to ensure it performed as expected. All models incorporated GridSearchCV for hyperparameter tuning, which adjusted regularization strength and improved model performance.

Logistic Regression is a statistical classification method that estimates the probability of a data point belonging to a particular class by fitting a logistic function. The model learns its parameters through maximum likelihood estimation and produces binary predictions by converting probabilities into class labels. It performed strongly, achieving an AUC of 1, F1 score of 1, and accuracy of 1 (Figure 3.4). It produced 100% true positives and 100% true negatives (Figure 3.5).





```
Logistic Regression
F1 Score:  1.0
ROC_AUC Score:  1.0
Accuracy:  1.0
```

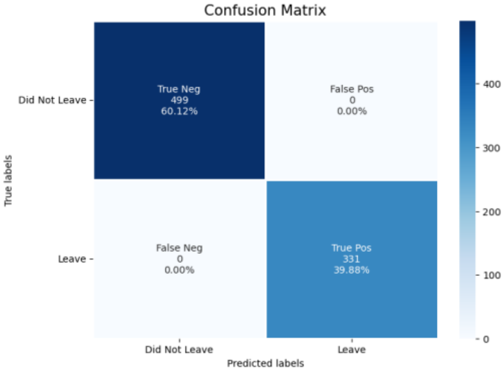
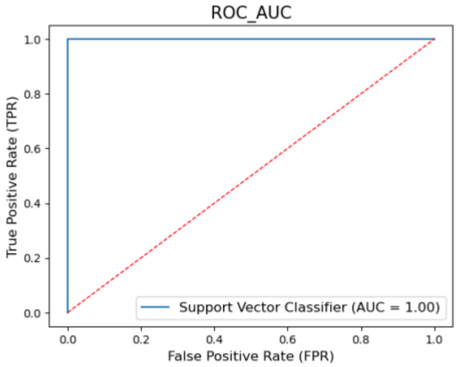
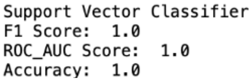


Figure 3.4: Logistic Regression ROC Curve

Figure 3.5: Logistic Regression Confusion Matrix

Support Vector Machine (SVM) is a supervised learning algorithm that identifies the optimal hyperplane to separate classes by maximizing the margin between data points. SVM achieved the same AUC score of 1 as logistic regression, but its performance was slightly weaker, with an F1 score of 1 and accuracy of 1 (Figure 3.6).





```
Support Vector Classifier
F1 Score:  1.0
ROC_AUC Score:  1.0
Accuracy:  1.0
```

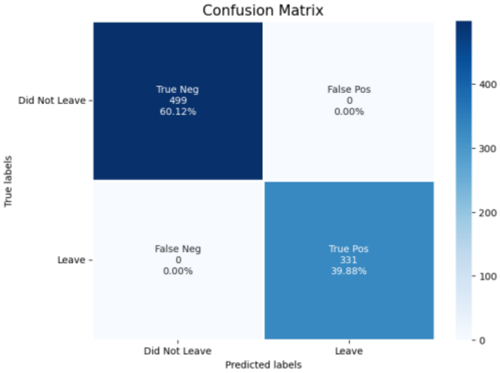
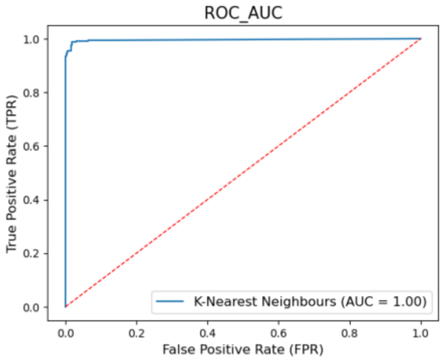
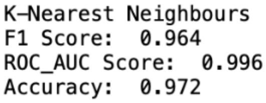


Figure 3.6: SVM ROC Curve

Figure 3.7: SVM Confusion Matrix

K-Nearest Neighbors (KNN) is a non-parametric algorithm that classifies new instances based on the most common class among their *k* closest neighbors, using a distance metric such as Euclidean distance. KNN achieved decent results with an AUC of 0.996, F1 score of 0.964, and accuracy of 0.972 (Figure 3.8). While it did not outperform logistic regression or SVM overall, it showed a higher false negative rate than SVM (Figure 3.9).





```
K–Nearest Neighbours
F1 Score:  0.964
ROC_AUC Score:  0.996
Accuracy:  0.972
```

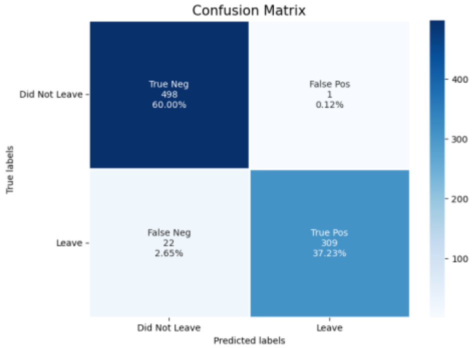
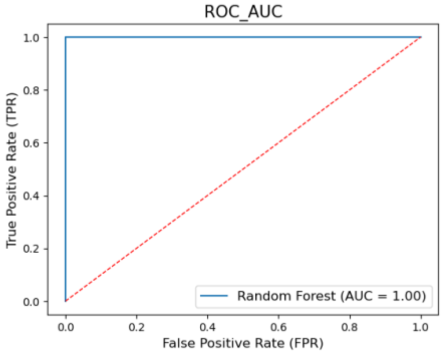


Figure 3.8: K-nearest ROC Curve

Figure 3.9: K-nearest neighbours Confusion Matrix

Random Forest is an ensemble learning technique that constructs multiple decision trees and outputs the class with the majority vote. The model was fine-tuned by plotting out-of-bag errors across different tree counts, settling on 150 trees as optimal. It achieved an AUC score of 1, an F1 score of 1, and an accuracy of 1 (Figure 3.10), with the same false positive rate as KNN, but a lower false negative rate (Figure 3.11).



```
Random Forest
F1 Score:  1.0
ROC_AUC Score:  1.0
Accuracy:  1.0
```

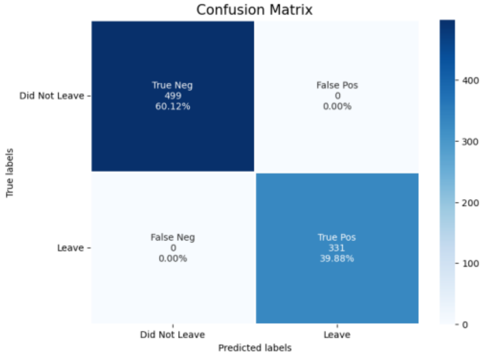


Figure 3.10: Random Forest ROC Curve

Figure 3.11: Random Forest Confusion Matrix

Overall, Logistic Regression emerged as the best-performing model based on its highest F1 score (1) and shared top accuracy (1), offering the best balance between precision and recall. However, Random Forest achieved the highest AUC (1), indicating superior ability to distinguish between classes. To better understand model performance, feature importance analysis was conducted, revealing that the most influential predictors of employee attitrion are the joining year, payment tier and gender (Figure 3.12).

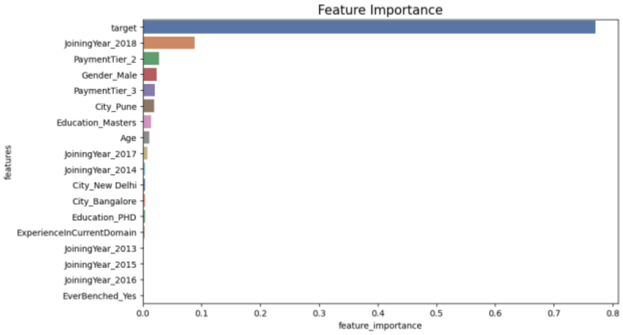


Figure 3.12: Random Forest ROC Curve

4.6. Results & Recommendations

RQ1: Which classification model provides the highest accuracy in predicting employee attrition?

- Logistic Regression is the best-performing model based on F1 score, ROC AUC, and accuracy.

RQ2: What are the key factors influencing an employee's decision to leave the company?

- The most influential factors are joining year, payment tier and gender.

5. Reference

IBM, 2013
https://www.ibm.com/think/topics/supervised-vs-unsupervised-learning

Abmatic AI, 2023
https://abmatic.ai/blog/importance-of-customer-segmentation-for-businesses

Analytics Vidhya, 2021
https://www.analyticsvidhya.com/blog/2021/11/employee-attrition-prediction-a-comprehensive-guide/

6. Dataset Links

6.1. Unsupervised Learning

World Happiness Report Dataset:
https://www.kaggle.com/datasets/unsdsn/world-happiness?select=2019.csv

6.2. Regression

Secondhand Car Price Dataset:
https://www.kaggle.com/datasets/sujithmandala/second-hand-car-price-prediction

6.3. Classification

Employee Dataset:
https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset