# Abstract

Artificial Intelligence & Machine Learning: Comparing Logistic Regression & K-NN Accuracy & Measuring Linear Regression Performance

Eoin Fitzsimons | X23151374 | Rural vs Urban Collisions & BMW Car Sales

This project investigates the performance of Logistic Regression and K-Nearest Neighbours (K-NN) for classification tasks and Linear Regression for regression tasks. The study utilises two datasets: a road collision dataset from data.gov.uk for classification and a BMW sales dataset from Kaggle for regression. Results reveal differences between the models, highlighting their strengths and limitations.

# Introduction

Machine learning is a branch of artificial intelligence that enables algorithms to uncover hidden patterns within datasets, allowing them to make predictions on new, similar data without explicit programming for each task.

[1]

Machine learning techniques have a huge impact on the world as we know it today, and their influence is only growing. They allow us to make more efficient decisions in many sectors. I am interested in automobiles so chose two datasets within this field.

This project explores the accuracy of two classification algorithms, Logistic Regression and K-NN, along with Linear Regression for numerical prediction tasks.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Stats** | **price** | **mileage** | **tax** | **mpg** | **engineSize** |
| **mean** | 22733.41 | 25496.99 | 131.70 | 56.40 | 2.17 |

The classification task focuses on predicting whether a location is urban or rural based on road collision data, while the regression task forecasts BMW car prices. The objective is to evaluate the models' performance, understand their behaviour under various configurations, and have the ability to act on the findings.

# Data Statistics

## Road Collision Dataset

* **Description**: Contains 104,258 entries, with features describing environmental and situational factors contributing to road collisions.
* **Key Statistics**:
  + Mean, standard deviation (SD), and quartiles were calculated for numeric columns (e.g., speed limits).
  + Missing values were minimal and addressed through median values.
  + Two target categories, "Data Missing or Out of Range" and "Unallocated," were removed.

## BMW Car Sales Dataset

* **Description**: Contains 10,781 rows and focuses on vehicle specifications and prices.
* **Key Statistics**:
  + Price std = 2.35
  + Features like mileage and engine size showed strong correlations with price.

# Methodology

## Algorithms

1. **Linear Regression**:
   * Applied to predict BMW car prices.
   * Evaluated using R-squared and RMSE.
2. **Logistic Regression**:
   * Utilized for urban/rural classification.
   * Hyperparameters: Max iterations = 1000.
3. **K-NN**:
   * Used for classification, varying k values from 1 to 100, and employing Euclidean and Manhattan distance metrics.
   * Weighting methods: Uniform and distance weighted.

## Preprocessing

* To ensure all features had a mean of zero and unit variance, I applied standardisation.
* Feature selection ensured relevance to the target variables.

## Evaluation Metrics

* Regression: R-squared, RMSE.
* *A metric that tells us the proportion of the variance in the response variable of a regression model that can be explained by the predictor variables. This value ranges from 0 to 1. The higher the R2 value, the better a model fits a dataset.* [2]
* The root mean square error (RMSE) measures the average difference between a statistical model’s predicted values and the actual values. Mathematically, it is the standard deviation of the residuals. Residuals represent the distance between the regression line and the data points. [3]
* Classification: Accuracy, precision, recall, F1-score, and confusion matrix.

Accuracy represents the number of correctly classified data instances over the total number of data instances.

* Precision is true positives/total positives. Recall is true positives/true positives + false negatives.

F1 score is the harmonic mean of precision and recall and is a better measure than accuracy. [4]

# Results and Evaluation

## Linear Regression

* **Performance**:
  + R-squared: 0.86.
  + RMSE: 4257.79.
* **Insights**:
  + Mileage and engine size had the highest predictive influence.

## Logistic Regression

* **Accuracy**: 85.6%.
* **Key Findings**:
  + The model performed consistently across all features.
  + Precision and recall for the rural class were marginally higher, reflecting slight imbalances.

**K-NN**

* **Best Performance**:
  + k = 20, Manhattan metric, distance-weighted: 86.5% accuracy.
* **Insights**:
  + Smaller k values overfit the data, while larger k values underfit.
  + Distance-weighted methods performed better.

## Visualisations:

1. **Linear Regression**:
   * Scatter plot: Actual vs Predicted prices.

A graph with blue dots and a red line

Description automatically generated

1. **Classification**:
   * Accuracy vs. k plot for K-NN.

A graph showing a logistic regression confusion matrix

Description automatically generated

A blue squares with numbers and a blue square

Description automatically generated with medium confidence

* + Confusion matrices for both algorithms.

A graph of numbers and lines

Description automatically generated with medium confidence

# Error Analysis

* **Linear Regression**:
  + Inferior performance in predicting prices for luxury models with limited data points.
  + Remedy: Incorporating additional data fields like brand or demand.
* **Classification**:
  + Misclassified instances mostly occurred at urban-rural boundaries.
  + Remedy: Incorporating geographical data to enhance the feature set.

# Conclusion

This study demonstrates the strengths of Logistic Regression and K-NN for classification, with K-NN slightly outperforming Logistic Regression in specific setups. Linear Regression showed strong predictive capability but struggled with high-value outliers.

# Bibliography

|  |  |
| --- | --- |
| [1] | M. Gupta, “What is Machine Learning?,” Geeksforgeeks, 26 May 2024. [Online]. Available: https://www.geeksforgeeks.org/ml-machine-learning/. [Accessed 1 December 2024]. |
| [2] | Z. Bobbitt, “RMSE vs. R-Squared: Which Metric Should You Use?,” Statology, 22 Jun 2021. [Online]. Available: https://www.statology.org/rmse-vs-r-squared/. [Accessed 1 December 2024]. |
| [3] | J. Frost, “Statistics by Jim,” Statistics by Jim, 8 May 2024. [Online]. Available: https://statisticsbyjim.com/regression/root-mean-square-error-rmse/#:~:text=The%20root%20mean%20square%20error%20%28RMSE%29%20measures%20the,between%20the%20regression%20line%20and%20the%20data%20points.. [Accessed 1 December 2024]. |
| [4] | H. N. B, “Confusion Matrix, Accuracy, Precision, Recall, F1 Score,” Medium, 10 December 2019. [Online]. Available: https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd. [Accessed 1 December 2024]. |