***Project***

***Report***

***Topic :***

*Used Car selling prizes Pridiction*

*Data Analysis and Machine learning Project*

*Dataset from Kaggle*

*Module 1*

***Prepared by : Sabah Shiekh***

***Submitted to: Miss Iqra***

***Used Car Prediction Analysis***

Abstract

This project explores the application of machine learning techniques for predicting the selling price of used cars. The dataset utilized contains various features relevant to car sales, including car name, year, kilometers driven, fuel type, seller type, transmission, and owner. The project involves a comprehensive data preprocessing pipeline encompassing data loading, exploratory data analysis, handling missing values, removing duplicates, encoding categorical features, outlier detection, data splitting, and feature scaling using standardization and Principal Component Analysis (PCA). Several regression models are trained and evaluated, including Linear Regression, Decision Tree Regression, Random Forest Regression, K-Nearest Neighbors Regression, and Support Vector Regression. Model performance is assessed using Mean Squared Error (MSE) and R-squared metrics. A comparative analysis reveals that the K-Nearest Neighbors (KNN) model exhibits the highest accuracy and reliability in predicting car prices based on the given dataset and features. The project demonstrates the effectiveness of machine learning in addressing the task of used car price prediction and provides insights into the selection of suitable models for this application.

Introduction

Predicting used car prices accurately is crucial for both buyers and sellers in the automotive market. Traditional valuation methods can be subjective and lack precision. This research leverages machine learning to develop a more robust and accurate car price prediction model. Using a dataset of used cars with features like car name, year, mileage, and fuel type, we train and evaluate various regression models, including Linear Regression, Decision Tree, Random Forest, K-Nearest Neighbors, and Support Vector Regression. By comparing their performance, we identify the most suitable model for this task, aiming to provide a valuable tool for informed decision-making in the used car market

Dataset Description

The dataset used in this study comprises information on various attributes of used cars and their corresponding selling prices. It was obtained from a publicly available source and contains a collection of real-world car sale records. The dataset includes the following features:

**Car\_Name:** The name or model of the car.

**Year:** The year in which the car was manufactured.

**Selling\_Price:** The selling price of the car in the market.

**Present\_Price:** The current showroom price of the car.

**Kms\_Driven:** The total distance the car has been driven in kilometers.

**Fuel\_Type:** The type of fuel used by the car (e.g., petrol, diesel, CNG).

**Seller\_Type:** The type of seller (e.g., individual, dealer).

**Transmission:** The type of transmission (e.g., manual, automatic).

**Owner:** The number of previous owners of the car.

**Data Characteristics:**

* The dataset consists of approximately 301 entries (rows) representing individual car sales records.
* It includes a mix of numerical and categorical features.
* Numerical features like "Year," "Selling\_Price," "Present\_Price," and "Kms\_Driven" provide quantitative information about the car.
* Categorical features like "Car\_Name," "Fuel\_Type," "Seller\_Type," "Transmission," and "Owner" represent qualitative characteristics of the car.
* The dataset may contain some missing values or outliers, which were addressed during the data preprocessing stage.

**Data Source:**

The dataset was downloaded from Kaggle, a popular platform for data science competitions and datasets. It is widely used for car price prediction tasks and serves as a valuable resource for research and analysis in the automotive domain

Methodology

This project employs a data-driven methodology to predict used car prices leveraging machine learning techniques. The workflow involves the following key stages:

**1. Data Loading and Initial Exploration:**

* The dataset "car\_data.csv" is loaded using the Pandas library.
* Preliminary exploration is conducted using df.info() and df.describe() to understand the data's structure, data types, and basic statistical properties.
* Duplicate entries are identified and removed using df.duplicated() and df.drop\_duplicates().
* The distribution of key categorical features like car name, transmission, fuel type, and seller type is examined using value\_counts().

**2. Data Preprocessing and Feature Engineering:**

* **Categorical Feature Encoding:** Label Encoding is applied to transform categorical features (Car\_Name, Fuel\_Type, Transmission, Seller\_Type) into numerical representations using LabelEncoder from scikit-learn.
* **Outlier Detection:** Outliers in numerical features are identified using the Interquartile Range (IQR) method. However, outliers are not removed, considering their potential significance in car price fluctuations.
* **Data Splitting:** The dataset is split into training (70%) and testing (30%) sets using train\_test\_split from scikit-learn with a random\_state for reproducibility.
* **Feature Scaling:** Standardization is performed on the feature matrix (X) using StandardScaler from scikit-learn to ensure features have zero mean and unit variance.
* **Dimensionality Reduction (PCA):** PCA is applied to reduce the number of features while preserving 95% of the variance using PCA from scikit-learn. The reduced feature set is used for model training.

**3. Model Training and Evaluation:**

* **Model Selection:** Five regression models are chosen: Linear Regression, Decision Tree Regression, Random Forest Regression, K-Nearest Neighbors Regression, and Support Vector Regression (SVR).
* **Model Training:** Each model is trained using the training data (X\_train, y\_train) with appropriate hyperparameters.
* **Model Evaluation:** Model performance is assessed on the testing data (X\_test, y\_test) using Mean Squared Error (MSE) and R-squared metrics.
* **Visualization:** Scatter plots are generated to visualize the relationship between actual and predicted selling prices for each model.

**4. Model Comparison and Selection:**

* A comparison table is created, summarizing the MSE and R-squared scores for all models.
* A bar chart is plotted to visually compare the R-squared values of the models.
* The model with the highest R-squared value is identified as the best-performing model. In your case, K-Nearest Neighbors (KNN) emerged as the most suitable model for this particular dataset and task.

**Tools and Libraries:**

The project utilizes the following libraries within the Google Colab environment:

* Pandas: For data loading, manipulation, and analysis.
* NumPy: For numerical operations and array handling.
* Scikit-learn: For machine learning algorithms, preprocessing, model evaluation, and dimensionality reduction.
* Matplotlib and Seaborn: For data visualization and creating plots

**Model Performance Summary:**

The table below summarizes the performance of each model:

| **Model** | **MSE** | **R-squared** |
| --- | --- | --- |
| Linear Regression | 5.598941 | 0.759190 |
| Decision Tree | 9.379437 | 0.596590 |
| Random Forest | 12.229710 | 0.474000 |
| KNN | 4.329124 | 0.813804 |
| SVR | 5.993469 | 0.742221 |

**Analysis of Results:**

* **K-Nearest Neighbors (KNN):** The KNN model demonstrated the best performance among the tested models, achieving the lowest MSE (4.329124) and the highest R-squared (0.813804). This indicates that KNN effectively captured the underlying patterns in the dataset and provided the most accurate predictions for car prices.
* **Linear Regression:** Linear Regression also performed reasonably well, with an R-squared of 0.759190 and an MSE of 5.598941. While not as accurate as KNN, it still provided a decent level of predictive power.
* **Support Vector Regression (SVR):** SVR exhibited performance comparable to Linear Regression, with an R-squared of 0.742221 and an MSE of 5.993469.
* **Decision Tree and Random Forest:** Both Decision Tree and Random Forest models showed relatively lower performance compared to KNN, Linear Regression, and SVR. Decision Tree obtained an R-squared of 0.596590 and an MSE of 9.379437, while Random Forest had the lowest R-squared (0.474000) and the highest MSE (12.229710) among all models. This suggests that these models might not be as suitable for this specific dataset or require further hyperparameter tuning.

Conclusion :

This project explored various machine learning models for predicting used car prices. K-Nearest Neighbors (KNN) emerged as the most accurate model, outperforming Linear Regression, Decision Tree, Random Forest, and Support Vector Regression. These findings demonstrate the potential of machine learning, specifically KNN, as a valuable tool for car price prediction, empowering buyers, sellers, and businesses in the automotive market. Future work could involve exploring other algorithms, feature engineering, and testing with larger datasets for further performance improvement.

.