**Machine Learning**

**Documentation**

**1.Sales prediction**

**Problem Statement:**

The goal is to predict sales for a product based on the customer price index (CPI), discounts offered, and number of reward offers. A linear regression model is used to make these predictions.

**Data:**

The data used to train the model is a table with the following columns:

* CPI: Customer price index
* discounts: Percentage discount offered
* offers: Number of reward offers
* Sales: Actual sales

The table contains 6 data points.

**Model Development:**

A linear regression model was created using the (sklearn.linear\_model.LinearRegression)class. The model was trained on the data in the DataFrame. The features used for training were the CPI, discounts, and offers columns. The target variable was the Sales column.

**Model Evaluation:**

The performance of the model can be evaluated using metrics such as Mean Squared Error (MSE), R-squared (R²) score, and others. These metrics provide insights into how well the model is performing in terms of predicting sales based on the input features.

R squared: 0.95

Mean of Residuals : -8.73

**Predictions:**

The model was used to predict sales for two new data points:

* CPI = 5000, discounts = 3, offers = 20
* CPI = 4000, discounts = 8, offers = 19

The predicted sales for these data points are:

* 5000 CPI, 3% discounts, 20 offers: Sales = 826645.35
* 4000 CPI, 8% discounts, 19 offers: Sales = 732680.36

**2.Loan prediction:**

**Problem Statement:**

Develop a model to predict whether a customer qualifies for a loan offer based on their:

* Number of credit cards
* Debit card ownership
* Insurance ownership
* Age
* CIBIL score

**Data:**

The provided table contains information for 1340 customers with their loan offers (0 - no offer, 1 - offer).

**Model Development:**

A logistic regression model was created using the (sklearn.linear\_model.LogisticRegression)class. The model was trained on the data in the DataFrame. The features used for training were the Cards, Debit cards, Insurance, Age and CIBIL score columns. The target variable was the Loan Offer column.

**Evaluation:**

The model's performance should be evaluated using metrics like accuracy Score, confusion matrix and ROC curve. Splitting the data into training and testing sets is crucial for unbiased evaluation.

Accuracy Score: 0.70

**3.income prediction:**

**Problem Statement:**

This project aims to classify customer data into meaningful groups based on their geographic and personal details. We will explore different classification models like Decision Tree, Random Forest, KNN, K-Means, and SVM to identify the most effective approach for this task.

**Data:**

The provided dataset contains information about customers like age, work class, education, income, and other personal details. We will use this data to train and evaluate the classification models.

**Model Development:**

* **Decision Tree:** This model creates a tree-like structure where each node represents a feature and branches represent possible values. It's interpretable but prone to overfitting.
* **Random Forest:** An ensemble method combining multiple decision trees for improved accuracy and robustness. Less interpretable compared to a single decision tree.
* **KNN:** Classifies data points based on the majority vote of their k nearest neighbors in the training data. Simple and efficient for small datasets, but sensitive to noise and dimensionality.
* **K-Means:** This unsupervised clustering algorithm groups data points into k predefined clusters based on their similarity. Not strictly a classification model, but useful for exploratory data analysis and potentially identifying natural groupings.
* **SVM:** This model aims to find a hyperplane separating different classes with the largest margin. Effective for high-dimensional data but requires careful parameter tuning.

**Evaluation:**

Each model will be evaluated based on metrics like accuracy, precision, recall, and F1 score. We will also consider interpretability and computational efficiency when selecting the best model.

|  |  |
| --- | --- |
| **Model** | **Accuracy Score** |
| Decision Tree | 0.84 |
| Random forest | 0.85 |
| KNN | 0.77 |
| SVM | 0.75 |