MACHIAN LEARNING PROJECT

The dataset link: https://www.kaggle.com/datasets/chadwambles/supermarket-sales/data

INTRODUCTION :

This project involved building, training, and evaluating machine learning models to address regression and classification problems using the sales dataset.

The models implemented were a Decision Tree Regressor for predicting total price and a Support Vector Machine (SVM) for classifying customer types.

Hyperparameter optimization and cross-validation techniques were employed to enhance model performance.

Methodology:

Key preprocessing steps included:

Label Encoding: Categorical columns and `product_category` were encoded into numerical values using `LabelEncoder`

Feature Selection

Data Splitting: The dataset was split into training and testing subsets using an 80-20 ratio for robust model evaluation

Models and Parameters

Two main machine learning techniques we used:

Decision Tree Regressor (For Regression)
Support Vector Machine (For Classification)

Grid Search for Hyperparameter Tuning

For SVM, grid search was employed to optimize the following parameters:

- Kernel: `['linear', 'rbf']`
- Regularization Parameter (C): `[0.1, 1, 10]`
- Gamma: `['scale', 'auto']`

Cross-Validation

cross-validation approach was applied to:

- Ensure robustness of model evaluation.
- Avoid overfitting.
- Validate performance across multiple data splits.

Accuracy Measures

Regression: Mean Squared Error (MSE) we evaluate the Decision Tree Regressor as it quantifies the average squared difference between predicted and actual values.

classification: Classification Report metrics (precision, recall, F1-score) and overall accuracy were used to assess SVM performance.

Visualizations

1. **Regression Performance**:
    - Scatter plot of actual vs. predicted `total_price` values.
2. **Classification Metrics**:
    - Confusion matrix for customer type predictions.

Results

Decision Tree Regressor

Mean Squared Error: The model achieved an MSE of (`0.00026023691259420063`)

Support Vector Machine

Accuracy : f1_score(0.97) , support(200)

```
SVM Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.94      0.97       108
           1       0.94      1.00      0.97        92

    accuracy                           0.97       200
   macro avg       0.97      0.97      0.97       200
weighted avg       0.97      0.97      0.97       200
```

After Grid Search

```
                  Precision   recall   f1_score    support
           0       1.00      0.95      0.98       108
           1       0.95      1.00      0.97        92

    accuracy                           0.97       200
   macro avg       0.97      0.98      0.97       200

weighted avg       0.98      0.97      0.98       200
```

Discussion

- **Insights**:

  The Decision Tree Regressor demonstrated reasonable predictive capability but could be improved by exploring ensemble methods such as Random Forests.

  SVM provided robust classification results, and grid search significantly enhanced performance by fine-tuning hyperparameters.

- **Challenges**:

  Imbalanced classes in customer types led to lower recall for minority classes. Addressing this imbalance with techniques like SMOTE could improve performance.

  Limited dataset size may have restricted model generalization. Expanding the dataset could yield better results.

Conclusion

he project successfully applied machine learning techniques to predict transaction outcomes and classify customer types.

Key achievements include:

- Building a regression model with reasonable accuracy.
- Classifying customer types with high precision and recall after hyperparameter tuning.