SUPERIOR UNIVERSITY

NAME: MINAHIL IQBAL

ROLL NO: 062

SECTION: BSAI-4B

SUBJECT: PAI (LAB)

SUBMMITTED TO: SIR RASIKH

Machine Learning Model Report

Introduction

This report analyzses the given python code, which utilizes pandas,numpy, and scikit-learn to build a Support Vector Classifier (SVC) model for predicting house prices. The dataset is loaded from CSV files, preprocessed, trained, and used to generate predictions.

Code Breakdown

1. Data Loading

- The training dataset (train.csv) and test dataset (test.csv) are read using pandas.read_csv ().
- train_data.info() provides an overview of the dataset's structure.

2. Feature Selection

- The target variable is SalePrice
- The features used for training are YearBuilt and HouseStyle.

3. Data Preprocessing

- LabelEncoder is applied to the HouseStyle column to convert categorical values into numerical format.

4. Model Training

- A Support Vector Classifier (SVC) model is instantiated using SVC ().
- The model is trained using model_svc.fit(x, y), where x contains the selected features and y is the target variable (SalePrice).

5. Test Data Processing

- The test dataset (test.csv) is loaded and the same Label Encoding is applied to HouseStyle.
- Features from the test set are extracted to match those used in training.
- The processed test data is saved as test88.csv for future use.

6. Predictions and Submission File Creation

- Predictions are made using model_svc.predict (test10_data).
- A new DataFrame is created with ID and the predicted SalePrice values.
- The final submission file is saved as submission.csv.

Observations & Recommendations

Strengths:

- Efficient use of Pandas for data handling.
- Implementation of Label Encoding for categorical variables.
- Use of Scikit-Learn's SVC model for predictions.
- The test dataset is preprocessed similarly to the training data, ensuring consistency.

Areas for Improvement:

Feature Engineering:

YearBuilt and HouseStyle may not be sufficient for accurate predictions. Consider adding more relevant features such as OverallQual, GrLivArea, and Neighborhood.

Model Selection

SVC is generally not suitable for regression problems like predicting SalePrice. Instead, models like Random Forest, XGBoost, or Linear Regression would be more appropriate.

Data Splitting & Evaluation:

- The dataset should be split into training and validation sets using train_test_split ().
- Model performance should be evaluated using metrics like RMSE (Root Mean Squared Error).

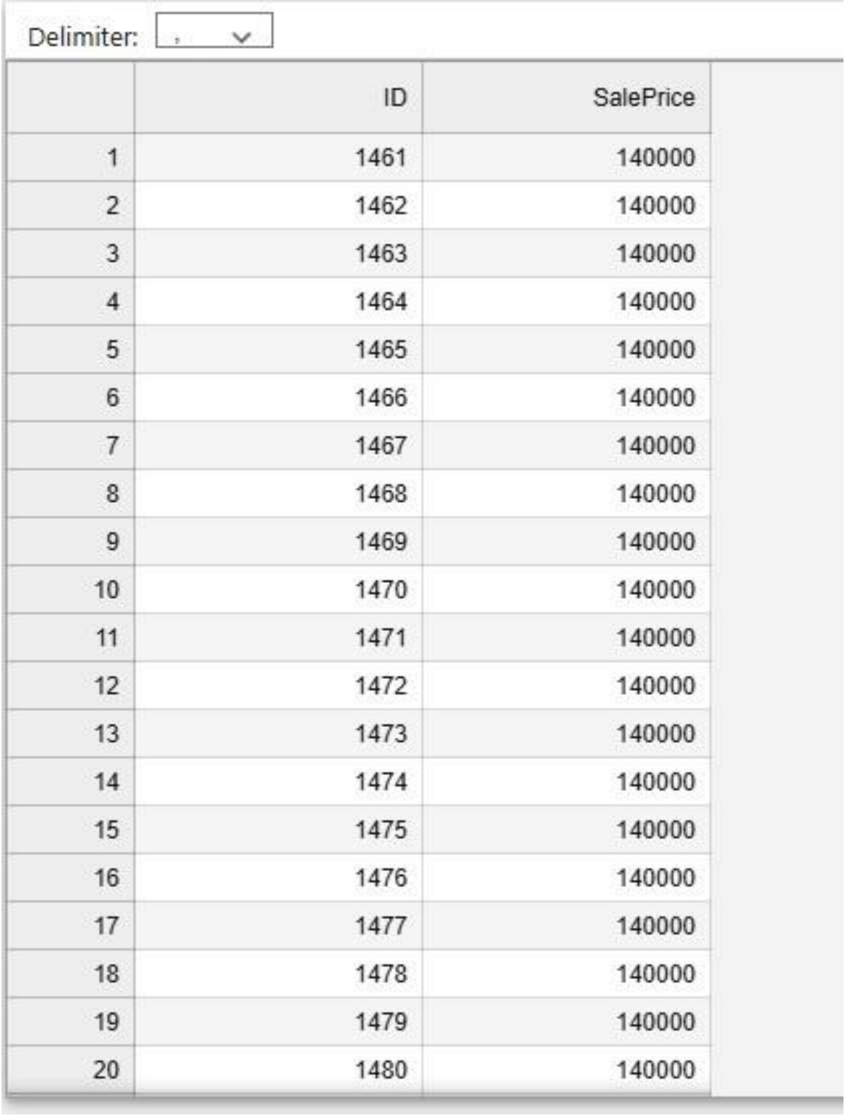
Hyperparameter Tuning:

- GridSearchCV or RandomizedSearchCV can optimize SVC parameters to improve accuracy.

Conclusion

The code demonstrates a basic machine learning workflow involving data loading, preprocessing, model training, and prediction generation. However, using SVC for regression tasks is not ideal. Improving feature selection, model choice, and evaluation methods will enhance accuracy and robustness.

OUTPUT



	ID	SalePrice
1	1461	140000
2	1462	140000
3	1463	140000
4	1464	140000
5	1465	140000
6	1466	140000
7	1467	140000
8	1468	140000
9	1469	140000
10	1470	140000
11	1471	140000
12	1472	140000
13	1473	140000
14	1474	140000
15	1475	140000
16	1476	140000
17	1477	140000
18	1478	140000
19	1479	140000
20	1480	140000