OSTİM TEKNİK ÜNİVERSİTESİ
ANKARA

OSTIM TECHNICAL UNIVERSITY

CENG 441 DATA MINING:

Classification of "Aparments" Dataset using Data Mining techniques

instructor: Hakan Bekir Aksebzeci

Introduction

Objective:

- To analyze and classify data using various machine learning classifiers.
- To compare the performance of classifiers using evaluation metrics.

Problem Statement:

- Large datasets often require efficient methods for analysis and prediction.
- Identifying the most effective classification technique is crucial for accurate results.

Importance:

- Data classification helps in decision-making and pattern recognition.
- This project demonstrates practical implementation of Data Mining techniques.

Dataset Overview

Dataset

- Consists of real estate data with features like price, location, rooms, area (square meters), floor, and building conditions.
- Contains 39302 entries.

Key Features

- Price: Target variable for prediction or analysis.
- Rooms, Square, Floor: Descriptive attributes used for classification.
- Location: Categorical variable indicating neighborhoods.

Preprocessing

- Removed missing values.
- Cleaned and normalized numerical data.
- Encoded categorical variables if necessary.

Methodology

Tools Used:

- Python (libraries:Pandas, NumPy, Matplotlib, Scikit-learn)
- Visual Studio Code as IDE

Steps of KDD Process:

1. Data Selection: Chose relevant attributes like price, rooms, square, and location.
2. Preprocessing:
    - Handled missing values.
    - Encoded categorical variables.
    - Normalized numerical data.
3. Transformation:
    - Feature scaling (Min-Max normalization).
    - Converted dataset for machine learning models.
4. Data Mining:
    - Applied 5 classifiers:
        - Logistic Regression
        - Decision Tree
        - Random Forest
        - Support Vector Machine (SVM)
        - k-Nearest Neighbors (k-NN)
5. Evaluation: Used cross-validation and performance metrics like accuracy, precision, recall, and F1-score.

Results Overview

Key Points:

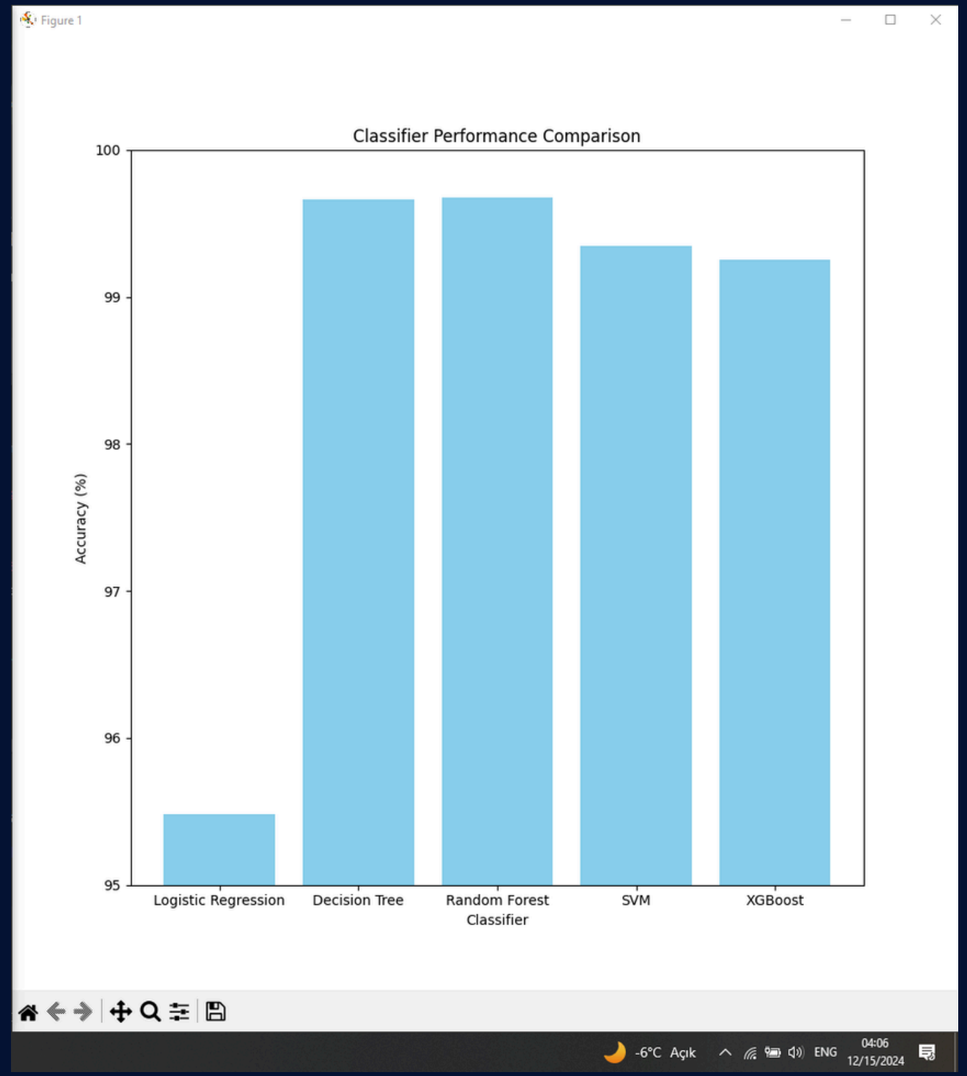
- Comparison of classifiers using evaluation metrics: accuracy, precision, recall, and F1-score.
- Random Forest emerged as the best-performing model, achieving the highest accuracy (99.68%).
- SVM and Decision Tree followed closely with high accuracies.
- Visualizations are provided in the following slides.

Results Overview: Classifier Accuracy Bar Chart

```
import matplotlib.pyplot as plt

classifiers = ['Logistic Regression',
'Decision Tree', 'Random Forest', 'SVM',
'XGBoost']
accuracies = [95.48, 99.66, 99.68, 99.35,
99.25] # Accuracy values from results

plt.figure(figsize=(10, 6))
plt.bar(classifiers, [acc * 100 for acc in
accuracies], color='skyblue')
plt.xlabel('Classifier')
plt.ylabel('Accuracy (%)')
plt.title('Classifier Performance Comparison')
plt.ylim(95, 100)
plt.show()
```



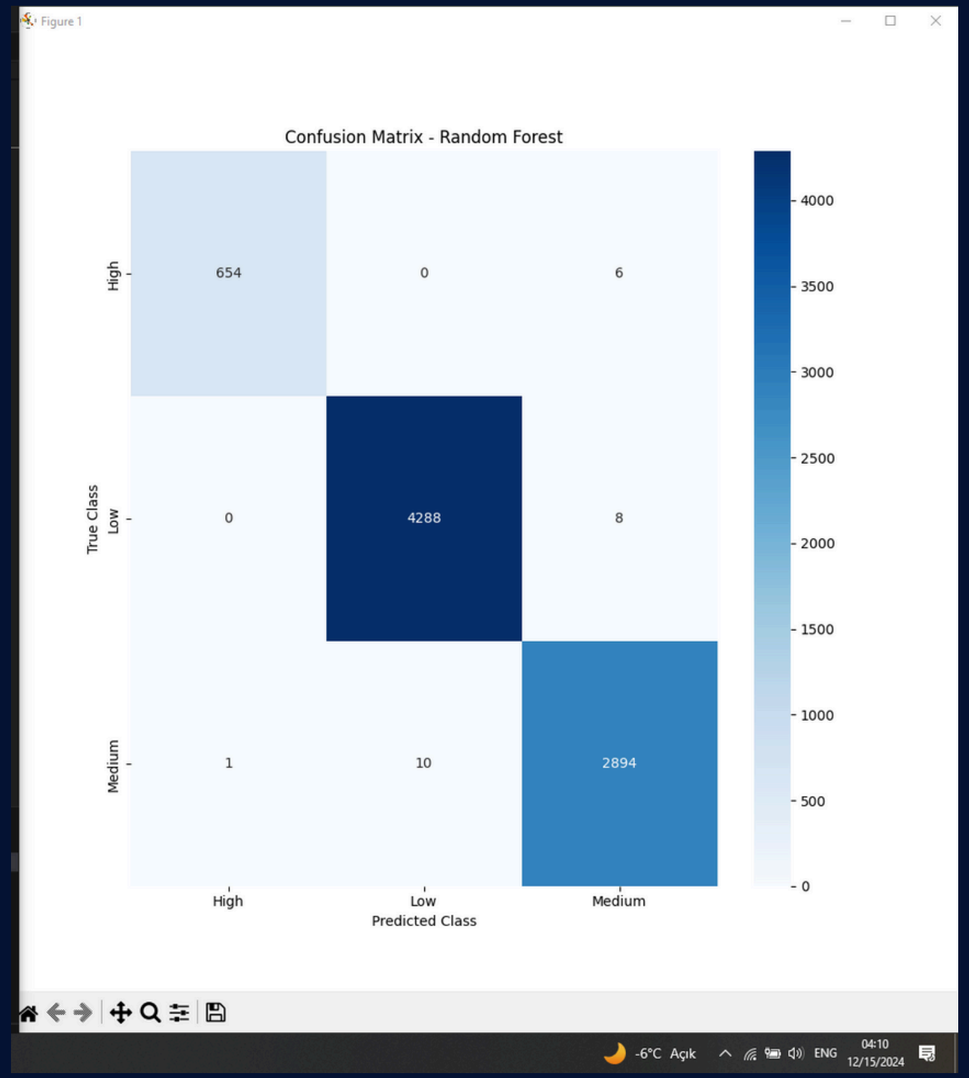
Results Overview:
Confusion Matrix – Random Forest

```
import seaborn as sns
import matplotlib.pyplot as plt

cm_rf = [
    [654, 0, 6],  # True High
    [0, 4288, 8],  # True Low
    [1, 10, 2894]  # True Medium
]

labels = ['High', 'Low', 'Medium']

plt.figure(figsize=(8, 6))
sns.heatmap(cm_rf, annot=True, fmt='d',
cmap='Blues', xticklabels=labels,
yticklabels=labels)
plt.xlabel('Predicted Class')
plt.ylabel('True Class')
plt.title('Confusion Matrix - Random Forest')
plt.show()
```

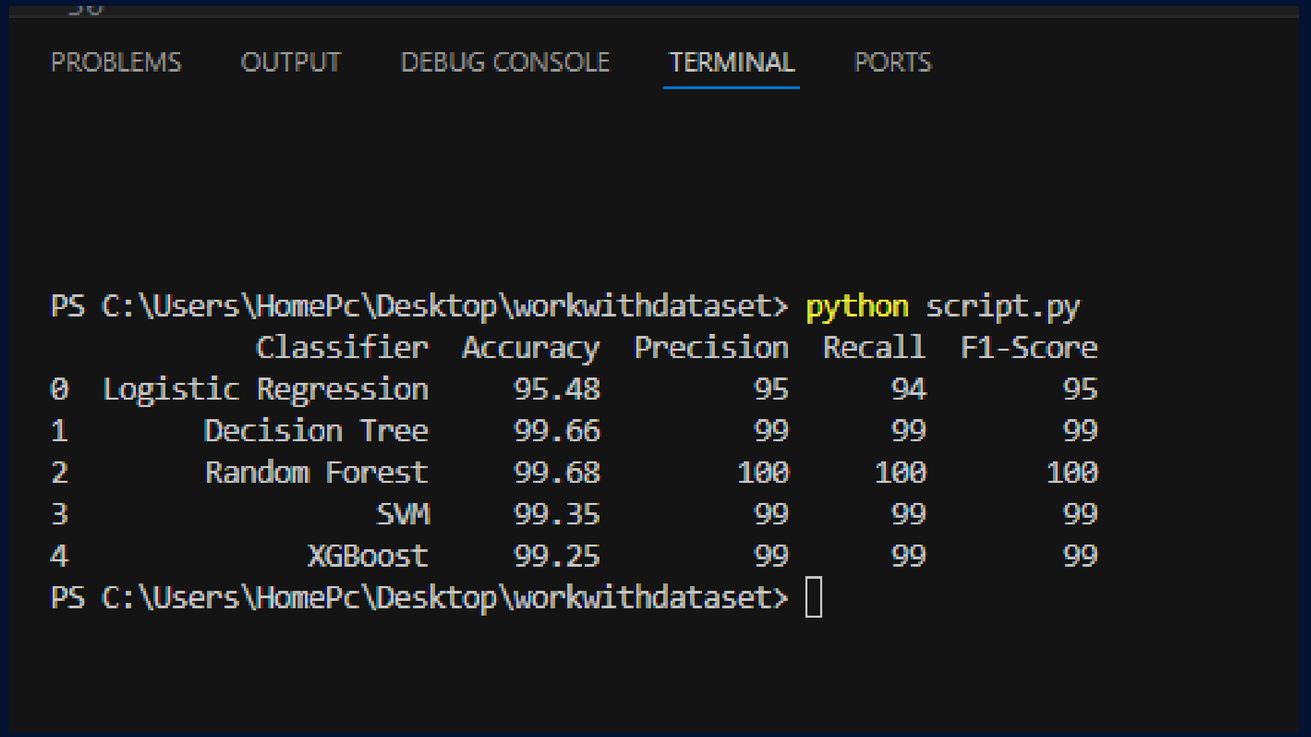


Results Overview: Performance Comparison Table

```
import pandas as pd

data = {
    'Classifier': ['Logistic Regression',
'Decision Tree', 'Random Forest', 'SVM',
'XGBoost'],
    'Accuracy (%)': [95.48, 99.66, 99.68,
99.35, 99.25],
    'Precision (%)': [95, 99, 100, 99, 99],
    'Recall (%)': [94, 99, 100, 99, 99],
    'F1-Score (%)': [95, 99, 100, 99, 99]
}

df_metrics = pd.DataFrame(data)
print(df_metrics)
```



Conclusion

Key Points:

- The Random Forest classifier outperformed all other models, achieving the highest accuracy (99.68%) and excellent precision, recall, and F1-score.
- Ensemble-based methods proved effective for classifying complex datasets.
- The dataset's key features, such as square, rooms, and location, significantly influenced classification accuracy.
- Insights gained from this study can guide future work on price prediction or similar classification problems.
- Limitations and areas for improvement:
  - Incorporating additional features like demographics or economic indicators.
  - Exploring deep learning models for enhanced accuracy.

OSTİM TEKNİK ÜNİVERSİTESİ
ANKARA

CENG 441 DATA MINING:
Classification of "Aparments" Dataset using Data Mining techniques

OSTIM TECHNICAL UNIVERSITY

Thank you for your attention!

instructor:
Hakan Bekir Aksebzeci

members:
Arifali Baghirli
Yunis Guliyev
Sadik Emre Duzgun
Mohammedghazi A M hattab