MTH 410 Data Mining for Cybersecurity

MIDTERM REPORT

1999 DARPA Intrusion Detection Evaluation Dataset

Spring , 2024

Group no: 8

Cennet Sude Arıkan 120200057

Sude Nur Ertürk 120200039

Doğa Bice 120200048

Table of Contents

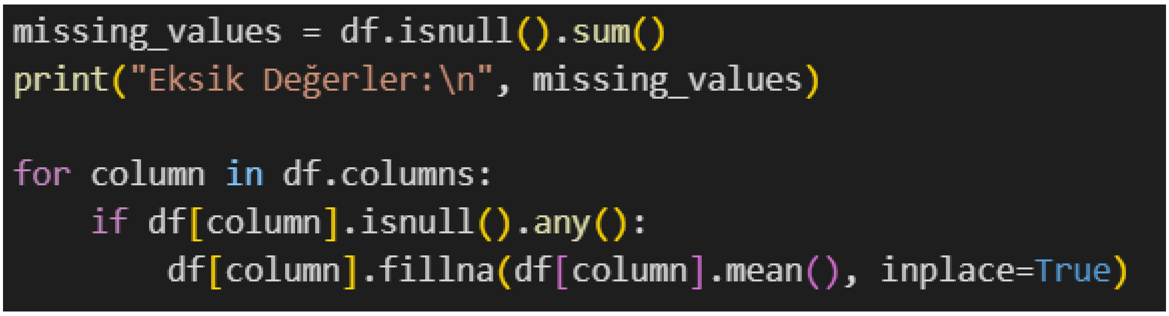
a. Introduction

Cybersecurity is a critical concern in today's digital age, where detecting and preventing network intrusions is paramount. This report focuses on the preprocessing, analysis, and application of data mining techniques to the 1999 DARPA Intrusion Detection Evaluation Dataset, a benchmark dataset for evaluating intrusion detection systems. The goal is to preprocess the data, perform exploratory data analysis (EDA), and apply data mining techniques to detect and classify potential security threats.

b. Data Preprocessing

Data preprocessing is essential to prepare the dataset for analysis. This involves handling missing values, selecting relevant features, and transforming data appropriately.

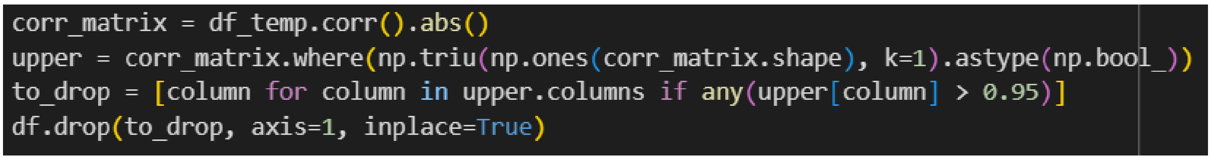
1. Handling Missing Values:



```python
missing_values = df.isnull().sum()
print("Eksik Değerler:\n", missing_values)

for column in df.columns:
    if df[column].isnull().any():
        df[column].fillna(df[column].mean(), inplace=True)
```

- Missing values were filled with the mean of the respective columns. In this section, missing values in the data set are checked and missing values are filled with average values.

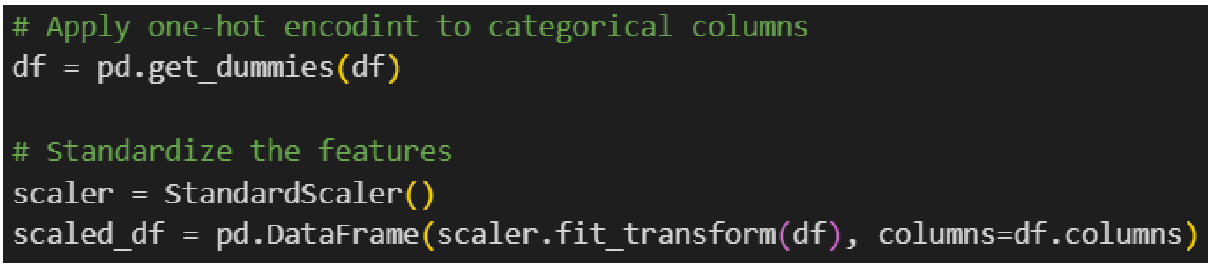
2. Feature Selection:



```python
corr_matrix = df_temp.corr().abs()
upper = corr_matrix.where(np.triu(np.ones(corr_matrix.shape), k=1).astype(np.bool_))
to_drop = [column for column in upper.columns if any(upper[column] > 0.95)]
df.drop(to_drop, axis=1, inplace=True)
```

- The correlation matrix is calculated and highly correlated columns are identified and removed from the data set.

3. One-Hot Encoding



```
# Apply one-hot encodint to categorical columns
df = pd.get_dummies(df)

# Standardize the features
scaler = StandardScaler()
scaled_df = pd.DataFrame(scaler.fit_transform(df), columns=df.columns)
```

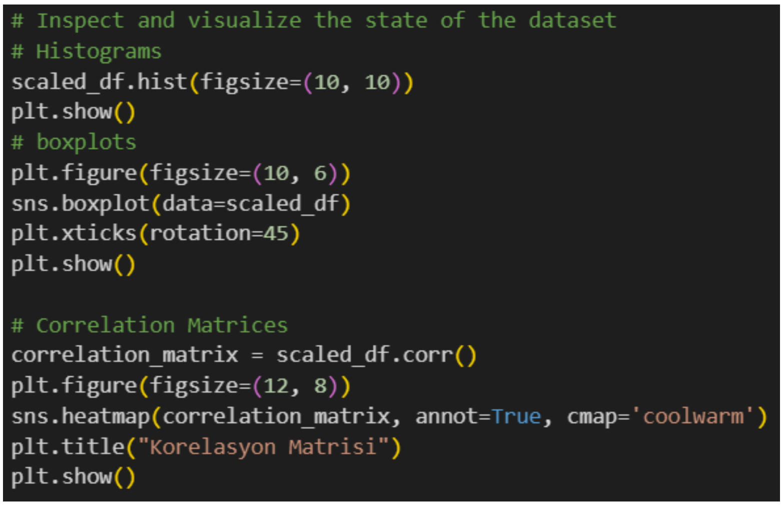
- Categorical columns are converted to numeric data using the one-hot encoding method. Features are standardized, that is, features are scaled so that their mean is zero and their standard deviation is one.

c. Exploratory Data Analysis (EDA)

EDA helps in understanding the data distribution, detecting patterns, and gaining insights into feature characteristics.

Visualizations:
We visualized the distribution of the target variable and other important features.

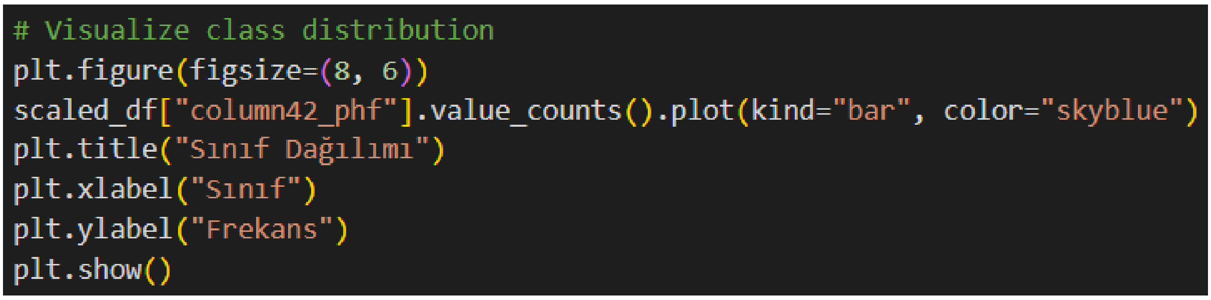
Correlation Matrix:
We visualized the correlation matrix to understand feature interdependencies.



```
# Inspect and visualize the state of the dataset
# Histograms
scaled_df.hist(figsize=(10, 10))
plt.show()
# boxplots
plt.figure(figsize=(10, 6))
sns.boxplot(data=scaled_df)
plt.xticks(rotation=45)
plt.show()

# Correlation Matrices
correlation_matrix = scaled_df.corr()
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title("Korelasyon Matrisi")
plt.show()
```

d. Data Mining Technique Section and Application

We selected the Random Forest classification technique to analyze the dataset, as it is well-suited for handling a mixture of numerical and categorical features and can provide insights into feature importance.



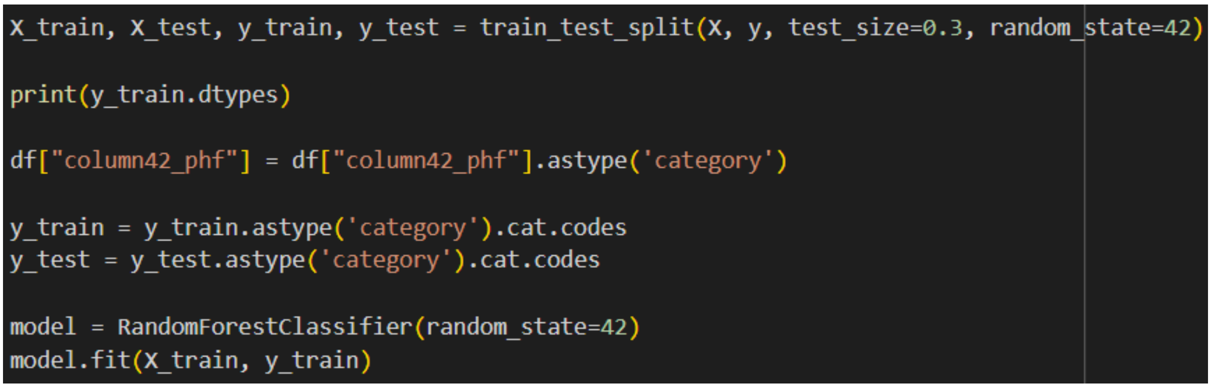
```
# Visualize class distribution
plt.figure(figsize=(8, 6))
scaled_df["column42_phf"].value_counts().plot(kind="bar", color="skyblue")
plt.title("Sınıf Dağılımı")
plt.xlabel("Sınıf")
plt.ylabel("Frekans")
plt.show()
```

1. Preparing the Data:

We split the dataset into training and testing sets.

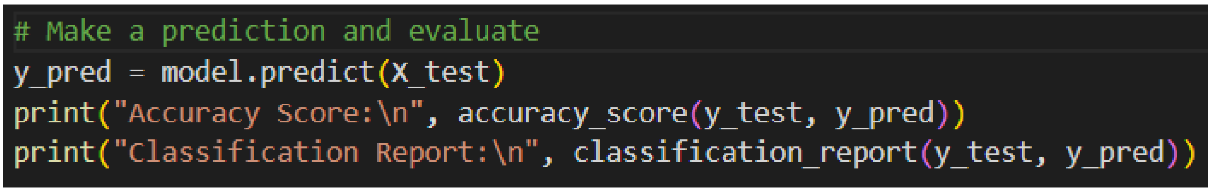
2. Training the Model:

We trained a Random Forest classifier on the training set.



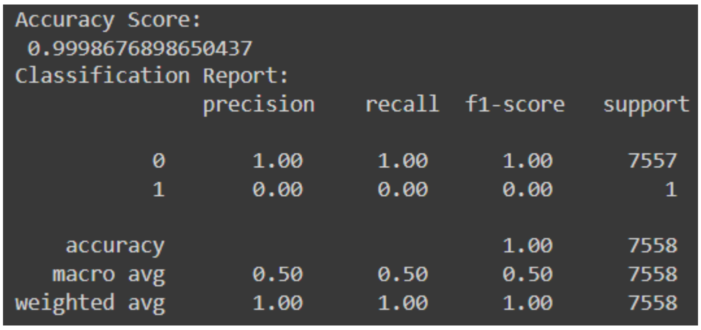
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

print(y_train.dtypes)

df["column42_phf"] = df["column42_phf"].astype('category')

y_train = y_train.astype('category').cat.codes
y_test = y_test.astype('category').cat.codes

model = RandomForestClassifier(random_state=42)
model.fit(X_train, y_train)
```

3. Prediction and Evaluation



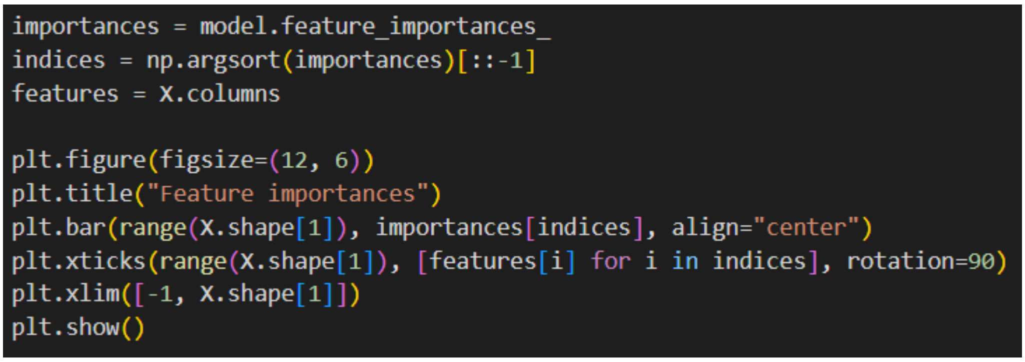
```
# Make a prediction and evaluate
y_pred = model.predict(X_test)
print("Accuracy Score:\n", accuracy_score(y_test, y_pred))
print("Classification Report:\n", classification_report(y_test, y_pred))
```

OUTPUT :



```
Accuracy Score:
 0.9998676898650437
Classification Report:
               precision    recall  f1-score   support

           0       1.00      1.00      1.00      7557
           1       0.00      0.00      0.00         1

    accuracy                           1.00      7558
   macro avg       0.50      0.50      0.50      7558
weighted avg       1.00      1.00      1.00      7558
```

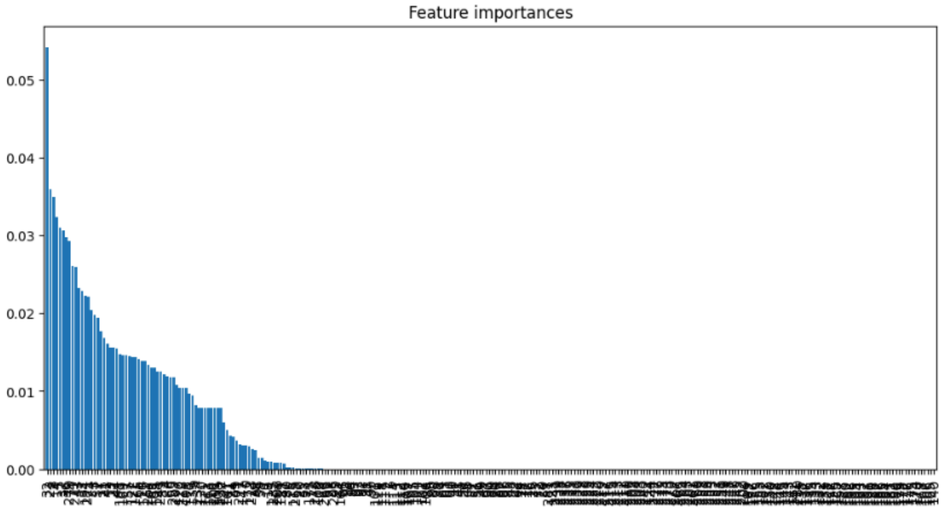
4. Feature Importance



```
importances = model.feature_importances_
indices = np.argsort(importances)[::-1]
features = X.columns

plt.figure(figsize=(12, 6))
plt.title("Feature importances")
plt.bar(range(X.shape[1]), importances[indices], align="center")
plt.xticks(range(X.shape[1]), [features[i] for i in indices], rotation=90)
plt.xlim([-1, X.shape[1]])
plt.show()
```

We identified the most important features using the trained model.

OUTPUT:



e. Conclusion

In this report, we demonstrated the complete process of data preprocessing, exploratory data analysis, and application of data mining techniques on a cybersecurity dataset. The Random Forest classifier proved effective in identifying important features and detecting security threats. The EDA provided valuable insights into the data distribution and feature characteristics.

This report provides a comprehensive overview of the steps involved in analyzing a cybersecurity dataset, from data preprocessing to the application of a Random Forest classifier, and illustrates the importance of each step in ensuring accurate and effective results.

REFERENCES:

- https://archive.ll.mit.edu/ideval/files/1999_DARPA_EvaulationSumPlans.pdf
- https://www.ll.mit.edu/r-d/datasets/1999-darpa-intrusion-detection-evaluation-dataset
- https://mesopotamian.press/journals/index.php/CyberSecurity/article/view/17/37
- https://ak-tyagi.com/static/pdf/56.pdf