الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Project Report

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Table Of Content

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Problem Statement

I am given a dataset that represents the people who attend to take loans from the bank.

My objective is to use the available information in the loan data to try to predict whether to give a loan or not. The dataset will come in a form of CSV file and will have 32581 observations and 12 fields including the target variable "Loan_status".

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Machine Learning Techniques

- **Machine Learning Techniques Used In This Project:**

  - **Decision Tree:**

    A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

  - **Support Vector Machine (SVM):**

    A Support Vector Machine (SVM) classifier is a supervised machine learning algorithm used for classification tasks. It's particularly effective for both linear and non-linear classification tasks and is widely used in various fields, including image recognition, text classification, and bioinformatics.

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Dataset Description

The dataset I'm using is given by the supervisor and it's 1.7MB size. Here is more information about the dataset:

- Number of rows and columns: (32581 , 21).
- Missing or Null values: 4011.
- Duplicate rows: 165.
- Duplicate columns: 0.

- **Attributes Description:**

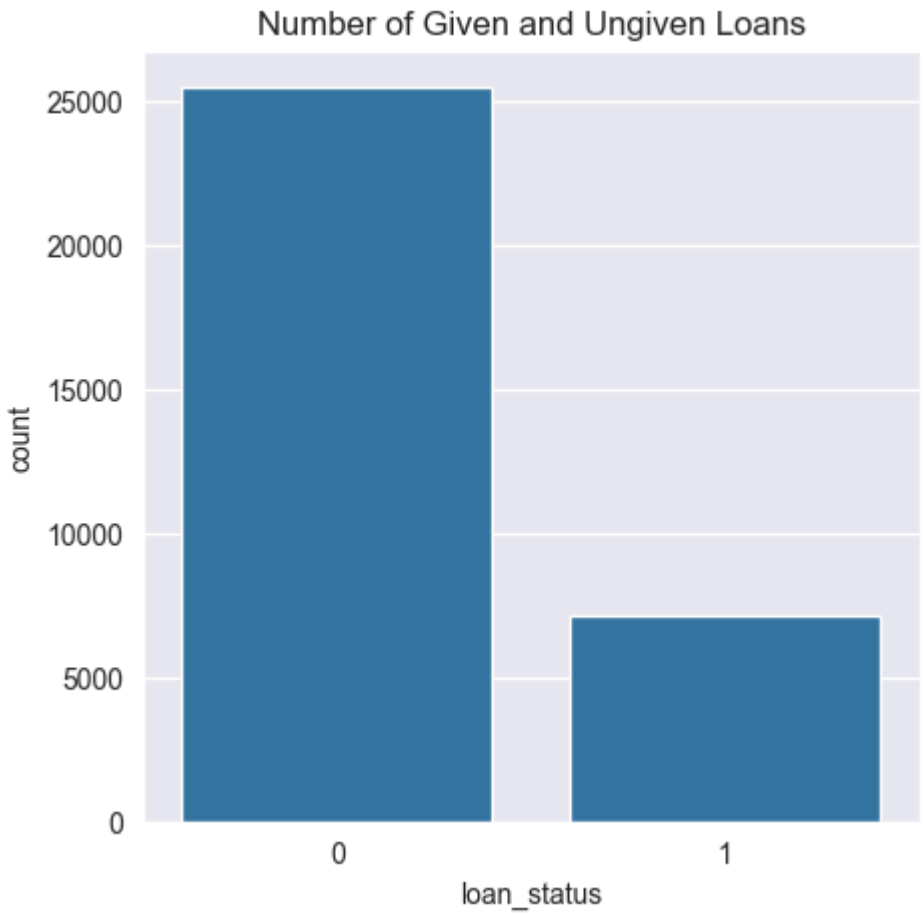
  The dataset includes 12 columns, one of these columns is the label column, and it is called "Loan_status". Its values are 0 or 1. 1 means that the bank would give a loan, 0 means the opposite, which is don't give the loan.

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

- **Number of Given and Ungiven Loans Plot:**



Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Data Preprocessing

Preprocessing data is an important step for data analysis. The following are some benefits of preprocessing data:

- It improves accuracy and reliability. Preprocessing data removes missing or inconsistent data values resulting from human or computer error, which can improve the accuracy and quality of a dataset, making it more reliable.

- It makes data consistent. When collecting data, it's possible to have data duplicates, and discarding them during preprocessing can ensure the data values for analysis are consistent, which helps produce accurate results.

- It increases the data's algorithm readability. Preprocessing enhances the data's quality and makes it easier for machine learning algorithms to read, use, and interpret it.

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

- **Dealing With Null Values:**

  Missing data is a common issue encountered in data analysis and machine learning tasks. Whether due to errors in data collection, incomplete records, or other reasons, missing values can significantly impact the accuracy and reliability of analytical results. In this guide, we explore one popular technique for handling missing data: median imputation.

  Median imputation is a method for replacing missing values with the median of the available data. Unlike mean imputation, which uses the average of the values, median imputation is less sensitive to outliers and skewed distributions, making it particularly useful in datasets with non-normally distributed variables.

  I used the following function, and applied it to the columns that contain missing/null values, which are “loan_int_rate” and “person_emp_length”. The function fills the null values using the median operation.

```
columns_to_impute = ["loan_int_rate", "person_emp_length"]
data[columns_to_impute] = data[columns_to_impute].fillna(data[columns_to_impute].median())
```

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

- **Encoding Object/Categorical Variables to Numerical:**

  Categorical variables are commonplace in real-world datasets, representing qualitative attributes such as gender, color, or country. While many machine learning algorithms require numerical input, dealing with categorical variables poses a challenge. In this article, we delve into the process of encoding categorical variables into a numerical representation, exploring popular encoding techniques and their applications.

  Categorical encoding is the process of converting categorical variables into numerical form, allowing them to be incorporated into machine learning models. This transformation is necessary because most algorithms operate on numerical data and cannot directly handle categorical variables.

  I used the “LabelEncoder” function from the “sklearn.preprocessing” library, then applied the function on every column that included object values.

```
label_encoders = {}
for col in x.select_dtypes(include=['object']).columns:
    label_encoders[col] = LabelEncoder()
    x[col] = label_encoders[col].fit_transform(x[col])
```

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

- **Feature Selection:**

  What is Feature selection? What features are important for your problem statement?
  Choosing the key features for the model is known as feature selection. A feature is a trait that affects or helps solve an issue. While developing the machine learning model, only a few variables in the dataset are useful for building the model, and the rest features are either redundant or irrelevant. If we input the dataset with all these redundant and irrelevant features, it may negatively impact and reduce the overall performance and accuracy of the model. Hence it is very important to identify and select the most appropriate features from the data and remove the irrelevant or less important features, which is done with the help of feature selection in machine learning.

  I applied a feature selection function that selects 9 columns instead of 12, the function uses “chi2” criteria as a hyperparameter. Obviously, this function improved my model performance.

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

- **What is “chi2”?**

  In scikit-learn's SelectKBest function, the "chi2" hyperparameter refers to the chi-squared statistical test for non-negative features to select the best features. Chi-squared (χ^2) test is a statistical method used to determine whether there is a significant association between two categorical variables.

  The chi-squared test measures the difference between the observed and expected frequencies of categorical variables to determine whether there is a significant relationship between them. It is commonly used in feature selection to assess the importance of categorical features in classification tasks. Features with higher chi-squared test scores are considered more informative and relevant for prediction.

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

- **The Dataset After Applying The Feature Selection Function:**

```
new_x=featureSelect_dataframe(x,y,chi2,9)
new_x
```

✓ 0.0s

	person_income	person_home_ownership	person_emp_length	loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_percent_income	cb_person_default_on_file
0	59000	3	123.0	4	3	35000	16.02	0.59	1
1	9600	2	5.0	1	1	1000	11.14	0.10	0
2	9600	0	1.0	3	2	5500	12.87	0.57	0
3	65500	3	4.0	3	2	35000	15.23	0.53	0
4	54400	3	8.0	3	2	35000	14.27	0.55	1
...	...	...	...	...	...	...	...	...	...
32576	53000	0	1.0	4	2	5800	13.16	0.11	0
32577	120000	0	4.0	4	0	17625	7.49	0.15	0
32578	76000	3	3.0	2	1	35000	10.99	0.46	0
32579	150000	0	5.0	4	1	15000	11.48	0.10	0
32580	42000	3	2.0	3	1	6475	9.99	0.15	0

32581 rows × 9 columns

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

- **Train-Test Split:**

  The train-test split is a technique for evaluating the performance of a machine learning algorithm.

  It can be used for classification or regression problems and can be used for any supervised learning algorithm.

  The procedure involves taking a dataset and dividing it into two subsets. The first subset is used to fit the model and is referred to as the training dataset. The second subset is not used to train the model; instead, the input element of the dataset is provided to the model, then predictions are made and compared to the expected values. This second dataset is referred to as the test dataset.

  I used the “train_test_split” function from the “sklearn.model_selection” library, the parameters of this function are “new_x” the selected features, “y” the label, with 0.3 test size and random state with value of 42.

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

- **Dataset Scaling:**

  In machine learning, feature scaling is a preprocessing technique used to standardize or normalize the range of independent features or variables in the dataset. It aims to bring all features to a similar scale to prevent features with larger magnitudes from dominating those with smaller magnitudes. Feature scaling is essential for many machine learning algorithms that are sensitive to the scale of input features, such as gradient descent-based optimization algorithms, k-nearest neighbors (KNN), support vector machines (SVM), and neural networks.

  There are two common methods of feature scaling:

  - Standardization (Z-score normalization): In standardization, each feature is rescaled so that it has a mean of 0 and a standard deviation of 1. This transformation ensures that the feature values are centered around the mean, with a standard deviation that accounts for the variability of the data.

  - Normalization (Min-Max scaling): In normalization, each feature is scaled to a fixed range, usually between 0 and 1. This transformation preserves the relative relationships between feature values and is particularly useful when the features have different units or scales.

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Building Models

Why do we need to train and test different models?

We need to train and test different models in data mining because it allows us to evaluate the performance of the models and select the one that best suits our needs.

When we train a model, we are teaching it how to make predictions based on the data we provide. We use a portion of the available data (called the training set) to train the model, and we adjust its parameters until it can make accurate predictions.

However, the ultimate goal of building a model is to use it to make predictions on new, unseen data. Therefore, it's important to test the model on a separate set of data (called the testing set) to see how well it performs on new data. This is known as model evaluation.

By testing different models on the same testing set, we can compare their performance and choose the one that performs best. This helps us to avoid overfitting (where the model is too closely tailored to the training data and performs poorly on new data) and to ensure that our model is robust and reliable.

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

- **Model1 (Decision Tree):**

  A decision tree model is a predictive model used in data mining. It uses a tree-like structure to represent a sequence of decisions and their possible outcomes. The tree is built using a dataset of labeled examples, and the algorithm selects the best attribute to split the dataset at each node. Once the tree is constructed, it can be used to classify or predict new examples by following the sequence of decisions. Decision tree models are easy to interpret and visualize, but they can overfit the training data.

  One of the most important thing to consider while building any model is Hyperparameter Tuning. So, after tuning the hyperparameters of this model, I found out that the best ones are “criterion=’gini’ ” which is the default value, “max_depth=9” and “max_features=5”.

  Now I can build the model with the previously mentioned hyperparameters.

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Here is the results I got:

```
Precision :  0.958148383005707

Recall :  0.6988899167437558

Accuracy Score :  0.9266496163682865

F1 Score :  0.8082374966568602

 [[7547   66]
 [ 651 1511]]
```

The Confusion Matrix Plot:

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

Machine Learning Project

- **Model2 (SVM):**

The fundamental concept behind SVM is to find the hyperplane that best separates the data points of different classes in the feature space. In two dimensions, this hyperplane is a line, while in higher dimensions, it's a hyperplane.

SVM aims to find the hyperplane that maximizes the margin, which is the distance between the hyperplane and the nearest data points (support vectors) from each class. Maximizing the margin helps improve the generalization performance of the model by maximizing the separation between classes.

The optimal hyperplane is determined by solving an optimization problem that involves minimizing the classification error while maximizing the margin. This optimization problem can be formulated as a quadratic programming problem.Once the optimal hyperplane is determined, SVM assigns new data points to one of the classes based on their position relative to the hyperplane. Data points on one side of the hyperplane are classified as one class, while data points on the other side are classified as the other class.

As I mentioned before, I tuned the hyperparameters of this model and found that the best ones are “C=100”, “kernel=”rbf’ ” and “degree=3”.

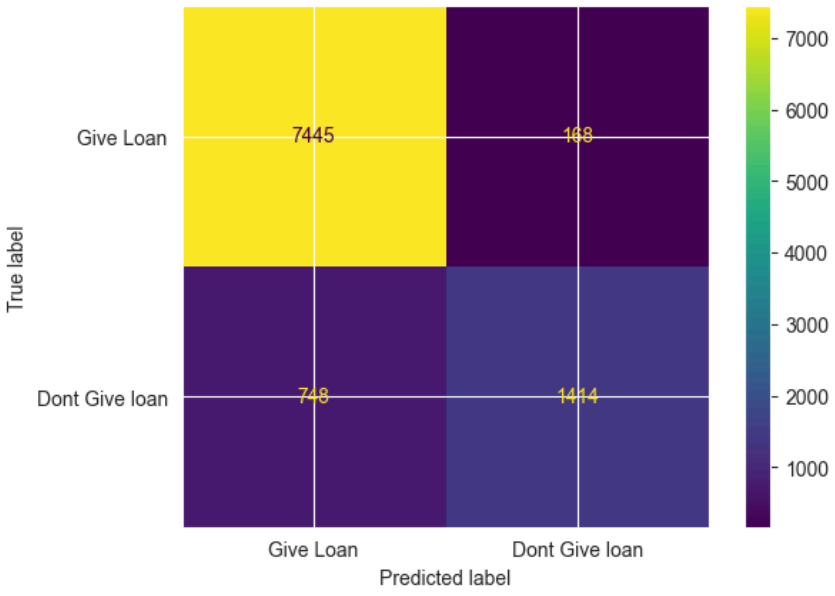
الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Here is the results I got:

```
Precision :  0.8938053097345132

Recall :  0.6540240518038853

Accuracy Score :  0.9062915601023018

F1 Score :  0.7553418803418803

 [[7445  168]
 [ 748 1414]]
```

The Confusion Matrix Plot:



Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Model Evaluation & Results

How do we compare models? What are the various metrics used? Comparing models is an important step in machine learning to determine the performance of different models on a given task. There are various metrics used to evaluate the performance of a model, depending on the type of task and the nature of the data. Some commonly used metrics are:

Accuracy: Accuracy is a common evaluation metric used in classification tasks to measure the proportion of correctly classified instances among all instances in the dataset. It is calculated as the ratio of the number of correct predictions to the total number of predictions made by the model.

Mathematically, accuracy can be expressed as:

Accuracy = Number of correct predictions/Total number of prediction x 100%

In simpler terms, accuracy answers the question: "What proportion of the predictions made by the model are correct?"

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Precision: Precision is a metric used in binary classification tasks to measure the proportion of correctly predicted positive instances (true positives) among all instances predicted as positive by the model. It focuses on the accuracy of the positive predictions made by the model.

Precision is calculated as:

Precision = True Positives / True Positives + False Positives

In simpler terms, precision answers the question: "Of all the instances predicted as positive by the model, how many were actually positive?"

Recall: Recall, also known as sensitivity or true positive rate, is a metric used in binary classification tasks to measure the proportion of correctly predicted positive instances (true positives) out of all actual positive instances in the dataset. It focuses on the model's ability to correctly identify all positive instances, regardless of any false positives.

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Recall is calculated as:

Precision = True Positives / True Positives + False Positives

In simpler terms, recall answers the question: "Of all the actual positive instances in the dataset, how many did the model correctly identify?"

F1-score: The F1-score is a metric used to evaluate the performance of a binary classification model, taking into account both precision and recall. It is the harmonic mean of precision and recall, providing a single score that balances the trade-off between these two metrics.

The formula for calculating the F1-score is:

F1_score = 2 x Precision x Recall / Precision + Recall

In simpler terms, the F1-score represents the balance between precision and recall, with values ranging from 0 to 1. A high F1-score indicates that the model has both high precision and high recall, meaning it has a good balance between correctly identifying positive instances and avoiding false positives.

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Confusion Matrix: A confusion matrix is a table that is used to evaluate the performance of a classification model by comparing the predicted and actual values of the target variable. It is also known as an error matrix.

A confusion matrix consists of four components:

1. True Positives (TP): The number of instances that are correctly classified as positive.
2. False Positives (FP): The number of instances that are incorrectly classified as positive.
3. True Negatives (TN): The number of instances that are correctly classified as negative.
4. False Negatives (FN): The number of instances that are incorrectly classified as negative.

The confusion matrix is typically displayed in a table format with the predicted values along the top row and the actual values along the first column. The entries in the table represent the number of instances in each combination of predicted and actual values.

Using the values in the confusion matrix, several metrics can be calculated to evaluate the performance of a classification model, including accuracy, precision, recall, and F1-score. For example, accuracy is calculated as (TP + TN) / (TP + TN + FP + FN), and precision is calculated as TP / (TP + FP).

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

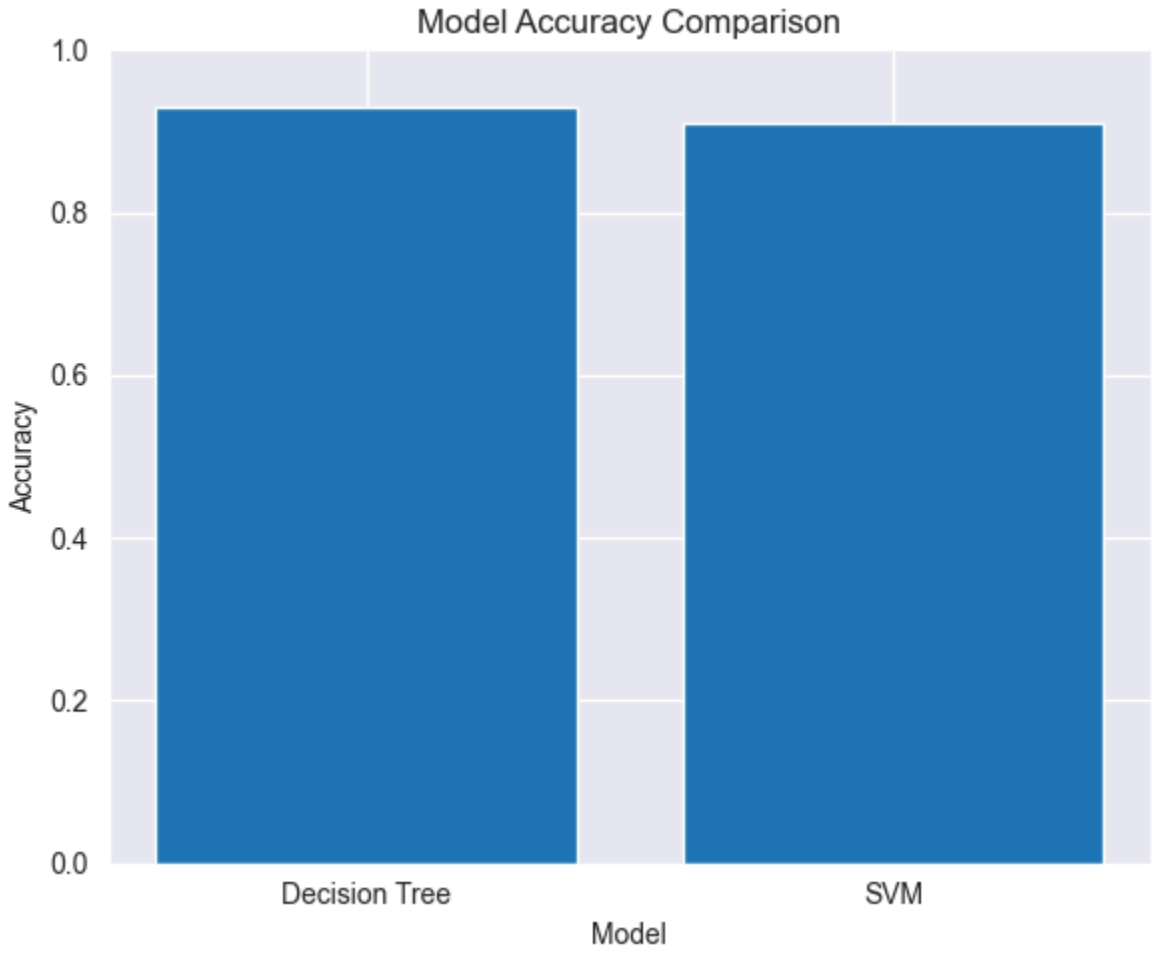
الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Inferences & conclusion

Objective of this project is to find a model that classifies an email as spam or not spam with better accuracy. The given dataset is trained using two models DecisionTree, and SVM.

Models Accuracy Comparison:

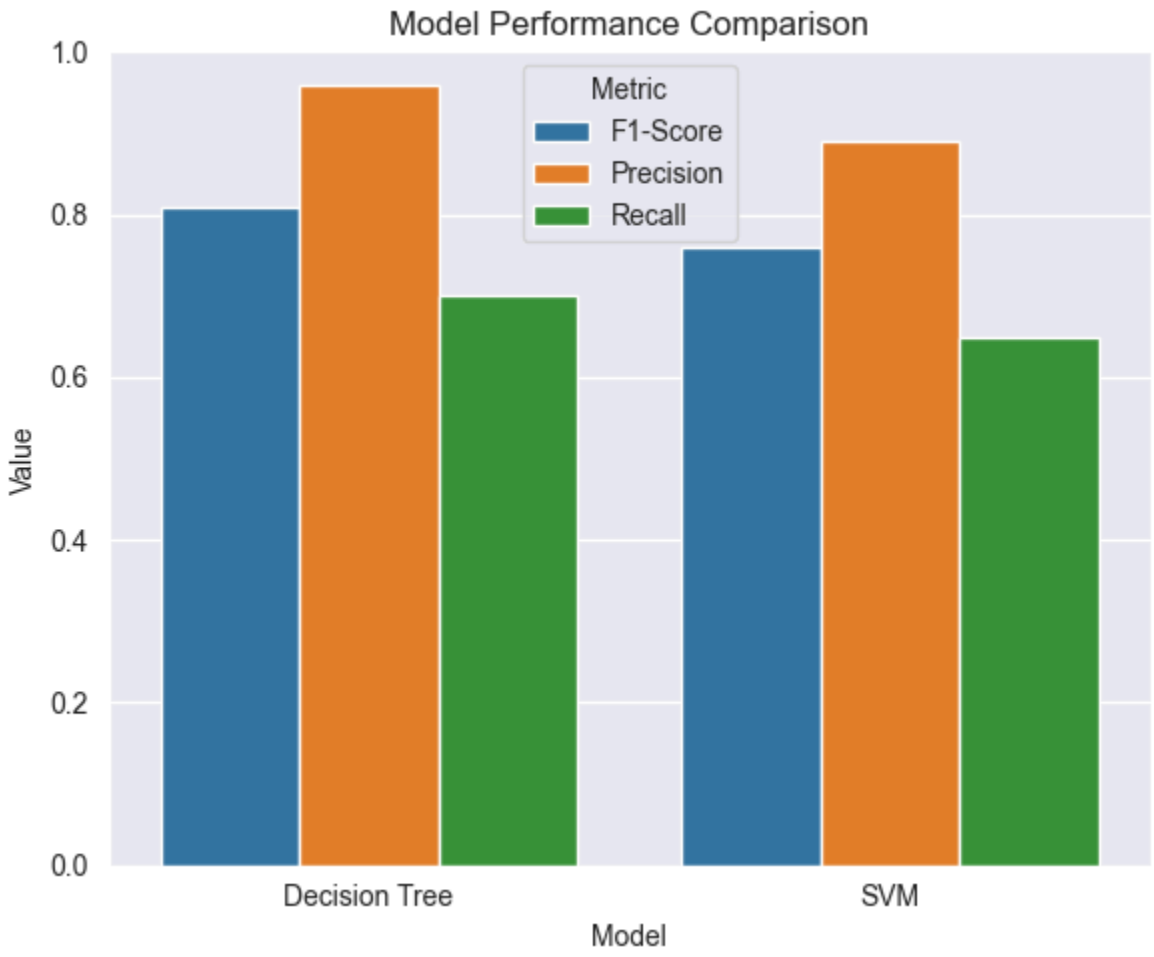


Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Models Performance Comparison:



Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**

الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Machine Learning Project

Both models resulted in high accuracy , which is quite good. The following table is a comparison table that compares the two models to find the better one.

Model	Decision Tree	SVM
Precision	0.96	0.90
Recall	0.70	0.65
Accuracy	0.93	0.91
F1_Score	0.80	0.76

After comparing all the attributes, it is clear that the Decision Tree model is better than the SVM model with better precision, recall, accuracy and F1 score.

Supervisor: Eng. Anas Abdulaziz **Submitted By: Owais Hilal 4200051**