BİM309 Artificial Intelligence Homework

Task 1 : Odometer Type Classification:

1) **Introduction:**
   The goal of this task is to classify odometer types into analog and digital categories. This classification is crucial for understanding the diversity within the TRODO dataset and has potential applications in various domains.
2) **Dataset Overview:**

The TRODO dataset consists of odometer images with corresponding annotations. The distribution of analog and digital odometers within the dataset was analyzed, revealing insights into the dataset's composition.

Annotations in the Pascal VOC 1.1 format were converted into a format suitable for odometer type classification. This was essential for training classification algorithms. These were given in a JSON file, and I converted them into a form that the program could understand with the help of a loop on Python.

3) **Data Preprocessing**

Images were resized to a consistent 64x64 pixel dimension to facilitate uniformity in model training. Additionally, normalization techniques were applied to enhance model convergence. The images are converted to a NumPy array and its flattened version is obtained. This converts the image to a flat vector. With these preprocessing steps, I made our data set ready for machine learning algorithms.

4) **Model Selection and Training**

Three classification algorithms, namely K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Random Forest Classifier (RFC), were selected for model training.

A) KNN (K-Nearest Neighbors): I choose this algorithm because it's Easy to implement, Adapts easily and have Few and simple hyperparameters. To simplify my project, I first wanted to choose a simple classification algorithm. Since KNN is a lazy algorithm, it takes up more memory and data storage compared to other classifiers. But of course, no algorithm is perfect.

[1] The k value in the k-NN algorithm defines how many neighbors will be checked to determine the classification of a specific query point. I choose the k value in mu project 3. It strikes a balance between capturing local patterns and generalizing well to broader trends. k=3 is often considered a balanced choice that avoids the extremes of being too sensitive to local variations (small k) or overly smoothed (large k).

B) SVM (Support Vector Machine) :

[2]SVM performs well in high-dimensional spaces, making it suitable for problems with a large number of features. This is particularly beneficial in applications such as image recognition, text classification, and genomics.

I choose for the parameter kernel as RBF. The choice of the Radial Basis Function (RBF) kernel in Support Vector Machine (SVM) models is often motivated by its flexibility and ability to capture complex relationships in the data.

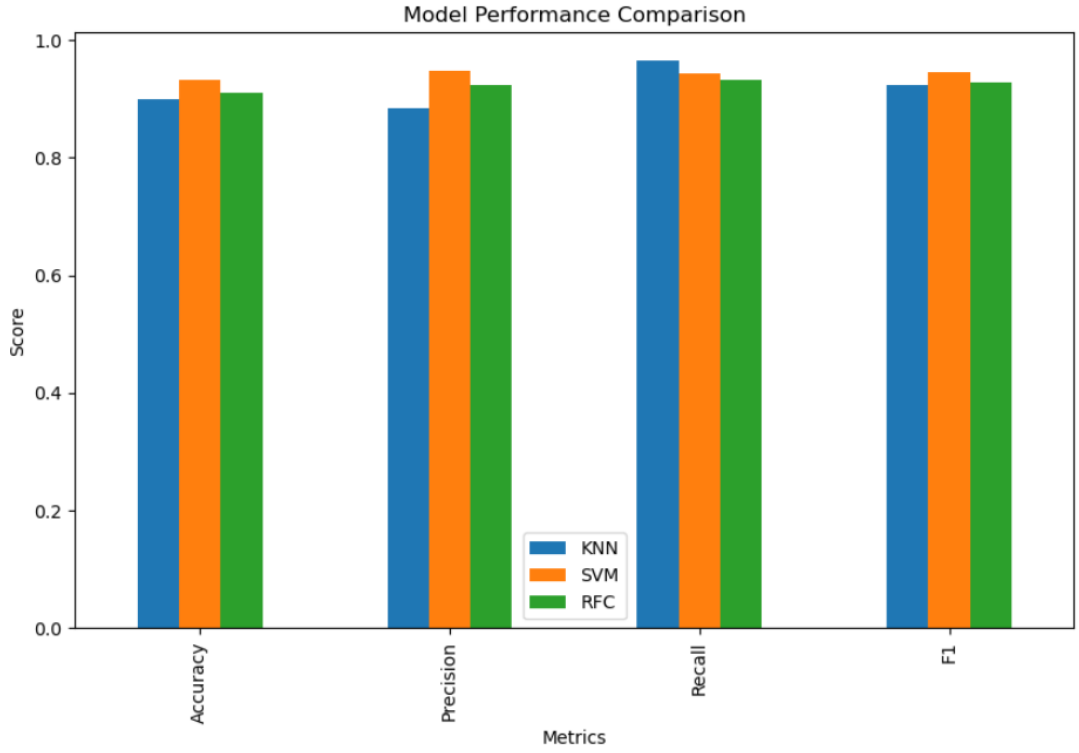
In summary, the RBF kernel is chosen in SVM models for its ability to handle non-linear relationships, expressiveness, versatility, and empirical success in capturing complex patterns in diverse datasets. However, the choice of kernel should be based on the specific characteristics of the data and the problem at hand.

C) Random Forest :

Random Forest's ensemble learning approach and many beneficial features make it a popular choice for a variety of machine learning tasks.

-Ensemble Learning:

During training, several decision trees are constructed using the Random Forest ensemble learning technique, which then combines them to produce a prediction that is more reliable and accurate. The ensemble of trees improves the model's overall performance and lessens overfitting.



5) Conclusion

At the evaluation and analysis stage, criteria like accuracy, precision, recall, and F1 score were used to evaluate each method. The outcomes included information on the advantages and disadvantages of each model, assisting in the decision-making process of which one to use for the purpose of classifying odometer types. It's crucial to remember that the features of the dataset and the particular demands of the task determine how effective these methods are. It might take more investigation, testing, and tweaking to get the models just right for practical uses.

Task 2: Mileage Extraction:

1. Data Preprocessing:

Firstly, this function reads XML files containing information about images (filename, object labels) from a Pascal VOC dataset, constructs the full image paths, and stores the information in a pandas DataFrame for further analysis or use in machine learning tasks.

```
# Read XML files and add relevant information to the data structure
for xml_filename in os.listdir(xml_folder):
    if xml_filename.endswith(".xml"):
        xml_path = os.path.join(xml_folder, xml_filename)
        tree = ET.parse(xml_path)
        root = tree.getroot()

        image_filename = root.find("filename").text
        image_path = os.path.join(image_folder, image_filename)

      # Extract the tag value as text
        label = root.find("object/name").text

        data['image_path'].append(image_path)
        data['label'].append(label)
```

The function load_and_preprocess_images is designed to load a list of images from given file paths, resize them, and normalize their pixel values to a range between 0 and 1. Here's a breakdown of the code:

```
def load_and_preprocess_images(image_paths):
    images = []
    for path in image_paths:
        img = Image.open(path)
        img = img.resize(image_size)
        img = np.array(img) / 255.0  # Normalize
        images.append(img)
    return np.array(images)
```

2. Model Selection and Training:

!! I choose two machine learning algorithms:

1. ***Support Vector Classifier (SVC):***

Support Vector Classifier (SVC) is a versatile algorithm for classification tasks. It effectively handles high-dimensional data and offers flexibility with various kernel functions, enabling it to capture complex decision boundaries. By maximizing the margin between classes, SVC builds robust models that generalize well to new data.

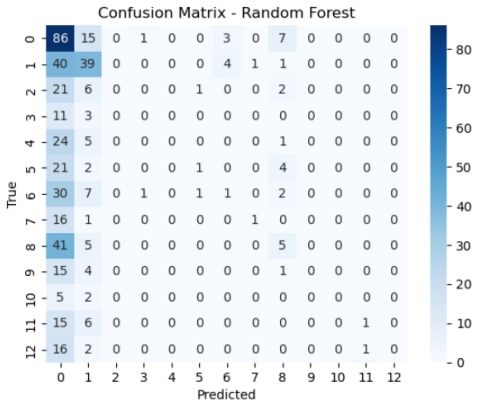
2. ***Random Forest:***

Random Forest is an ensemble learning algorithm that constructs multiple decision trees during training. By aggregating predictions from these trees, it improves accuracy and mitigates the risk of overfitting. Random Forest provides valuable insights into feature importance, aiding in feature selection. Its ability to handle missing values without imputation makes it versatile, making it suitable for diverse datasets.

Model performances:

1. Random Forest:

```
Random Forest Train set Accuracy: 0.9874411302982732
Random Forest Test set Accuracy: 0.2803347280334728
Random Forest Precision: 0.22613364413782774
```



2. SVC:

```
Test set accuracy for SVC: 0.26778242677824265
Precision for SVC: 0.15570319827136178
```

!! The observed low accuracy of approximately 28% in the model could be attributed to several factors. Firstly, the dataset might lack diversity and representation, hindering the model's ability to generalize effectively. Overfitting could also be a concern, as the model may be overly tailored to the training data, limiting its performance on new examples. Additionally, an imbalanced distribution of classes within the dataset might lead to biased learning. Furthermore, the features in the dataset may not be effectively capturing the underlying patterns, necessitating improvements in feature engineering. Lastly, issues with data preprocessing, such as normalization, could impact the model's ability to learn effectively.

3. Conclusion:

Mileage extraction from images faces several challenges, including variability in image quality, text recognition errors, diverse odometer designs, reflections, and varying numeric formats. Inconsistent data presentation across different vehicles adds complexity.

To address these challenges, potential improvements include data augmentation for model robustness, advanced OCR techniques for accurate text recognition, and adaptive algorithms that can handle different odometer designs and numeric formats. Image preprocessing, such as glare removal, and integration with vehicle information can enhance accuracy.

References:

[1]https://www.ibm.com/topics/knn#:~:text=Next%20steps-,K%2DNearest%20Neighbors%20Algorithm,of%20an%20individual%20data%20point.

[2]https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/#:~:text=SVM%20is%20a%20powerful%20supervised,work%20best%20in%20classification%20problems.