SKILL VERTEX

REPORT

Topic: **Mnist Digit Recognition with csv data**

Name: **Sri Krishna Bellam**

October 2022

Index

Introduction

Problem Statement:

- To find the best algorithm for digit recognition using csv mnist dataset

Objective:

- To Train the machine using machine learning algorithms
- To find the best algorithm for best result
- To find highest possible accuracy

Applications:

- Enables scanning and processing of handwritten data directly
- Creation of OCR (optical character recognition)

Data Description

The Given Dataset is a csv form pixels where each column represents the values of each pixel in 28x28 grid for forming each handwritten digit.

It also contains a column called **label**, which is the actual digit representation of that specific row, helpful to train and test the data

Approach

My approach to the problem would be using all the classification algorithms, since it is a classification based dataset, and find the best algorithm in terms of mean absolute error, mean squared error, r2 score and accuracy, along with creating a confusion matrix to enable better visualization of the prediction

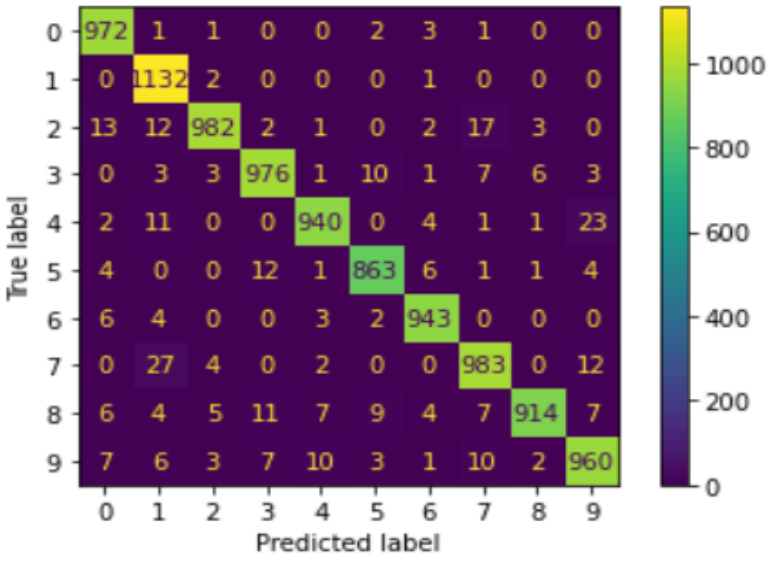
Visualization

There is no visualization required for this dataset since it is the csv form of a the pixels of handwritten digits.

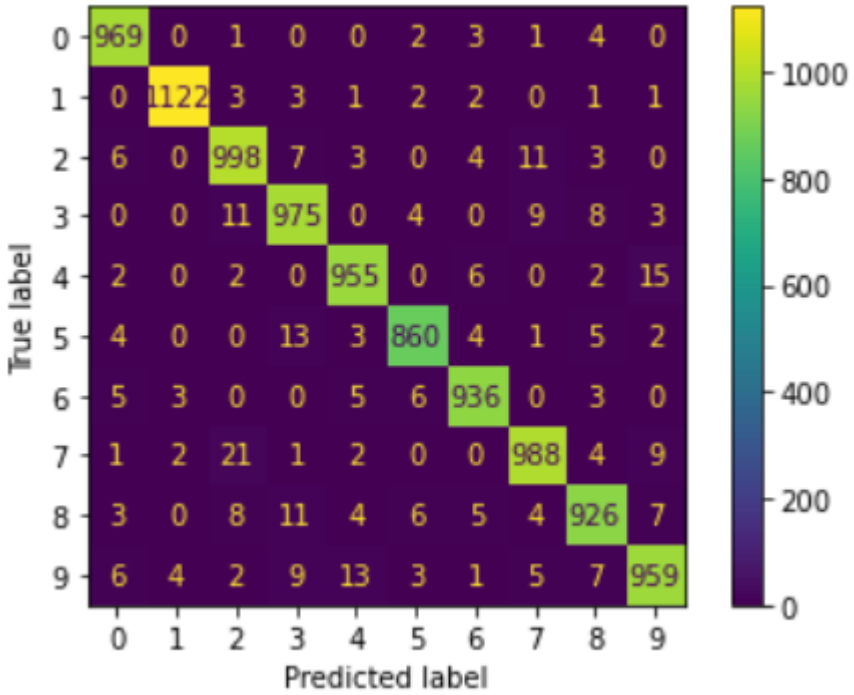
But visualization is possible for the evaluation, by creating a confusion matrix.

The confusion Matrix of KNN, Random Forest Classifier and Decision Tree Classifier is given in the next page

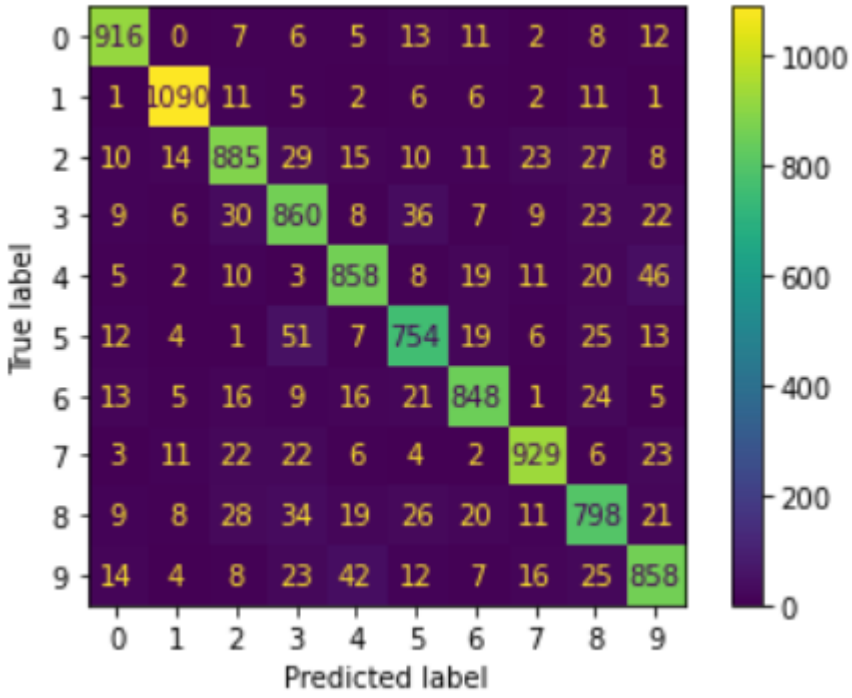
KNN:



Random Forest Classifier:



Decision Tress Classifier:



Algorithms

As I mentioned, I have used various classification models on this dataset and they have different accuracy and other performance measures. I have used the following machine learning algorithms on our dataset.

1. **KNN:**

   KNN algorithm used for both classification and regression problems. KNN algorithm based on feature similarity approach.

2. **Random Forest Classifier:**

   Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time.

3. **Decision Tress Classifier:**

   Supervised Machine Learning Algorithm that uses a set of rules to make decisions, similarly to how humans make decisions.

Evaluation:

1) Mean_absolute_error:

KNN	0.1325
Random Forest Classifier	0.1182
Decision Tree Classifier	0.4341

2) Mean_squared_error:

KNN	0.6713
Random Forest Classifier	0.5856
Decision Tree Classifier	2.0751

3) R2_score:

KNN	0.9199422
Random Forest Classifier	0.9301626
Decision Tree Classifier	0.7525281

4) Accuracy_score:

KNN	96.65
Random Forest Classifier	96.88
Decision Tree Classifier	87.96

Result and discussion

We have applied KNN, Random Forest Classifier and Decision Tree Classifier Algorithm to determine the best algorithm with the highest accuracy, we can also observe here that the KNN and Random Forest Classifier's accuracy score is more or less equal.

Conclusion

After applying these visualizations and algorithms we can conclude that the Random forest Classifier is the most suited algorithm to perform this project.

Future Work

In the future using OpenCV module we can create a program that can dynamically read handwritten datasets.

References

https://scikit-learn.org/stable/supervised_learning.html