Digit Recognition: A Case Study

Dipayan Banik, Scout Doran, Shataydrian Marshall, Tony Tai

Introduction

Handwritten character recognition is one of the practically important issues in pattern recognition. The applications of digit recognition includes in bank check processing, data entry, etc. Digit Recognition is a computer vision technique to predict the correct digits from pixel values of images. The objective of this project is to explore the field of image processing using K Nearest Neighbor (KNN) algorithm to predict digits and find the accuracy of the model.

Methodology

The aim of the project is to build an efficient model with an accuracy of at least 90% through testing and training on a Kaggle dataset. Through Exploratory Data Analysis and evaluation metrics the datasets were explored in detail to find the correctness of the image label.

Evaluation Metric

The F1, P-Value, and Kappa score is used as the evaluation metric for the model. A score of 1 for F1 and Kappa is considered as a perfect model and p values less than 0.05 are reliable.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

.

Techniques

KNN is an effective algorithm to solve the digit recognition problem. KNN calculates the distance between the new instance and all the instances in the entire training set. The algorithm looks for the top nearest K instances and outputs the class with the highest frequency as the prediction. Cross validation can be used to choose the best value for K that results in the highest accuracy.

Results

Exploratory Data Analysis

After investing the Digit Recognizer Kaggle dataset, we have done some exploratory data analysis on the training and test data set. The data set contains gray-scale images of handwritten digits. The training dataset an extra column which has the label of the digit and its pixels.

Modelling

Cross validation has been used to choose the best value for K and we have choose K to be 5 because it has the highest accuracy. A confusion matrix has been printed to see the predicted and actual outcome of the digits and the evaluation metrics. An F1 score of approximately 0.9 was predicted for the digits, Kappa value for 0.9 and p-value less than 0.05.

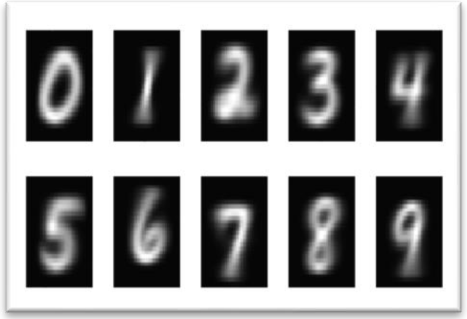


Figure 1: Digit visualization

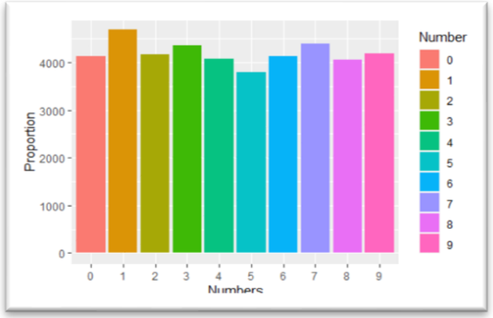
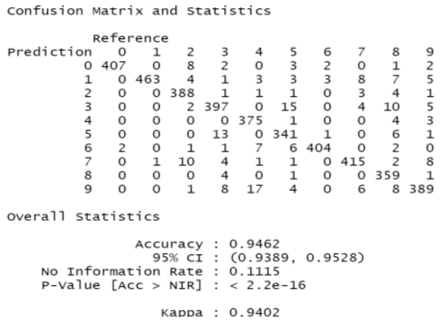


Figure 2: Digit proportion



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1   2   3   4   5   6   7   8   9
         0 407   0   8   2   0   3   2   0   1   2
         1   0 463   4   1   3   3   3   8   7   5
         2   0   0 388   1   1   1   0   3   4   1
         3   0   0   2 397   0  15   0   4  10   5
         4   0   0   0   0 375   1   0   0   4   3
         5   0   0   0  13   0 341   1   0   6   1
         6   2   0   1   1   7   6 404   0   2   0
         7   0   1  10   4   1   1   0 415   2   8
         8   0   0   0   4   0   1   0   0 359   1
         9   0   0   1   8  17   4   0   6   8 389

Overall Statistics

               Accuracy : 0.9462
                 95% CI : (0.9389, 0.9528)
    No Information Rate : 0.1115
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9402
```

Figure 3: Confusion Matrix

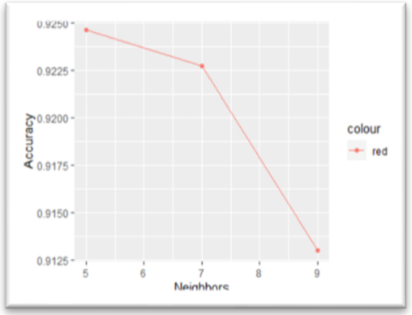


Figure 4: KNN Cross Validation

References

"Digit Recognizer," Kaggle. [Online]. Available: https://www.kaggle.com/c/digit-recognizer/overview. [Accessed: 11-Apr-2021].

A. Marjani, "Digit Recognition with KNN, Neural Network, and SVM," 09-Oct-2017. [Online]. Available: https://rstudio-pubs-static.s3.amazonaws.com/326787_431decc80c1a4ec797c3ef1e0b095e41.html. [Accessed: 11-Apr-2021].

O. Harrison, "Machine Learning Basics with the K-Nearest Neighbors Algorithm," Medium, 14-Jul-2019. [Online]. Available: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761. [Accessed: 11-Apr-2021].

R. Agarwal, "The 5 Classification Evaluation metrics every Data Scientist must know," Medium, 11-Sep-2020. [Online]. Available: https://towardsdatascience.com/the-5-classification-evaluation-metrics-you-must-know-aa97784ff226. [Accessed: 11-Apr-2021].

Acknowledgements

We would like to thank Dr. William Eberle for giving us the opportunity to work on this project and the Office of Research for showcasing our project.

Conclusion

In conclusion we have discovered that the percentage of images that were accurately matched with the correct digit was 94.62%. By comparing Kappa value, p value and F1 score we concluded that KNN is a good model for predicting the digits. The digit labels are printed into a csv file with its associated image ID.

Future Work

In the future, we plan to continue to our analysis through using different techniques such as Neural Networks and Support Vector Machines(SVM) to check if the accuracy improves.