Supervised Learning Comparison Report

By:-
Pritam Borate
ID:800936897

Overview

This report is mainly directed towards the comparison between different classifier models on digit recognition dataset and amazon dataset. The models which are taken into consideration are Naive Bayes, Decision tree, KNN, Neural Networks and SVM. In addition to that, their comparison with boosting is also done. Let's start with each of the algorithms overview:

Digital Recognition Dataset

Neural Networks

Neural networks has produced one among the best results on the dataset. Its prediction accuracy reaches till 97%. It is producing very good results and can be termed as the best model for this dataset.

Precision and Recall values of the labels are also one of the highest compared to other models. For the number 0 the precision and recall value is 1. It means that this model is correctly classifying every number 0.

Accuracy	**97**
Complexity	**0.811**
Precision	**[1. 0.95212766 0.99431818 0.98305085 0.97802198 0.94210526 1. 0.98809524 0.95238095 0.91145833]**
Recall	**[1. 0.98351648 0.98870056 0.95081967 0.98342541 0.98351648 0.98342541 0.9273743 0.91954023 0.97222222]**

Confusion Matrix	**[[178 0 0 0 0 0 0 0 0 0] [0 179 1 0 1 0 0 0 7 0] [0 0 175 1 0 0 0 0 0 0] [0 0 1 174 0 2 0 0 0 0] [0 0 0 0 178 0 2 1 0 1] [0 0 0 1 0 179 0 5 2 3] [0 0 0 0 0 0 178 0 0 0] [0 0 0 2 0 0 0 166 0 0] [0 2 0 3 0 0 1 1 160 1] [0 1 0 2 2 1 0 6 5 175]]**

Decision Tree

As far as prediction accuracy is concerned, the decision tree model produces an accuracy of almost 85% which is slightly lower than the Neural network model for this dataset. The better part of this model is that it has the less time complexity than the neural networks.

Considering the precision of this model, we have 10 numbers from 0 through 10 which acts as labels in the dataset. The highest precision which I got is of the number 0 which seems valid. The calculated precision percentage for the number 0 is almost 94%. However, this mitigates when the number 1 and number 7 comes into the picture. This maybe due to the fact that they both are quite similar in their array structure. So, the less accuracy in this case seems viable. This is the precisions of the different labels. The detail can be seen in the table.

Precision and Recall come hand to hand with each other where the accuracy is a big concern. On a similar note as precision, the accuracy of the different labels varies. However, the highest recall percentage is more than 95% for the number 0 which is higher than the precision percentage of the same number which is a good sign for this model in this dataset.

Accuracy	**85**
Complexity	**0.043**
Precision	**[0.94444444 0.84126984 0.81005587 0.86627907 0.84431138 0.88268156 0.93478261 0.90566038 0.74489796 0.8125]**
Recall	**[0.95505618 0.87362637 0.81920904 0.81420765 0.77900552 0.86813187 0.95027624 0.80446927 0.83908046 0.86666667]**

Confusion Matrix	**[[170 0 1 0 6 1 0 0 1 1] [0 159 4 1 10 3 2 1 5 4] [1 4 145 7 2 6 1 6 6 1] [0 7 4 149 0 2 0 1 3 6] [4 0 3 0 141 1 3 6 4 5] [2 0 0 5 4 158 0 2 3 5] [0 1 1 0 3 7 172 0 0 0] [0 1 6 4 2 0 0 144 2 0] [1 3 12 8 9 1 3 11 146 2] [0 7 1 9 4 3 0 8 4 156]]**

KNN

KNN or K-Nearest Neighbor can also be considered as one of the efficient classifier for this dataset. For this dataset, it has the highest accuracy percentage which is around 98% which makes it one of the best and efficient classifier for this dataset. On a similar note, the Precision and Recall of this dataset for each label is very good. Like neural network it has also the highest precision and recall value as 1 for a particular label.

Accuracy	**98**
Complexity	**0.819**
Precision	**[1. 0.92857143 1. 0.98360656 0.99441341 0.99444444 1. 0.98870056 0.9702381 0.96685083]**
Recall	**[1. 1. 0.98305085 0.98360656 0.98342541 0.98351648 1. 0.97765363 0.93678161 0.97222222]**
Confusion Matrix	**[[178 0 0 0 0 0 0 0 0 0] [0 182 3 0 2 0 0 0 9 0] [0 0 174 0 0 0 0 0 0 0] [0 0 0 180 0 0 0 0 1 2] [0 0 0 0 178 1 0 0 0 0] [0 0 0 0 0 179 0 0 0 1] [0 0 0 0 0 0 181 0 0 0] [0 0 0 2 0 0 0 175 0 0] [0 0 0 1 1 0 0 1 163 2] [0 0 0 0 0 2 0 3 1 175]]**

Naive Bayes

Naive Bayes model is mostly based on the probability distribution methodology. There are three different approaches of Naive Bayes taken into the consideration i.e. Multinomial Naive Bayes, Gaussian Naive Bayes and Bernoulli Naive Bayes. Unlike the previous models, it is not working very efficiently in this dataset. It has prediction accuracy of 89%. The best part of it that its time complexity is very low to almost a factor of ten when compared to other models.

Precision and Recall of the labels of this dataset is showing variations. For some digits, the Precision value is more than 0.94 but some it is pretty less. Almost the same case is there with the Recall values.

Accuracy	**89**
Complexity	**0.01**
Precision	**[0.99428571 0.79069767 0.9 0.93902439 0.92021277 0.97633136 0.95027624 0.93513514 0.7679558 0.75471698]**
Recall	**[0.97752809 0.74725275 0.86440678 0.84153005 0.9558011 0.90659341 0.95027624 0.96648045 0.79885057 0.88888889]**

Confusion Matrix	**[[174 0 0 1 0 0 0 0 0 0] [0 136 7 0 1 0 4 1 22 1] [0 15 153 1 0 0 0 0 1 0] [0 1 3 154 0 0 0 0 0 6] [4 0 0 0 173 1 4 1 1 4] [0 0 0 2 0 165 1 0 0 1] [0 7 0 0 0 1 172 0 1 0] [0 0 1 6 3 0 0 173 1 1] [0 9 9 9 3 2 0 3 139 7] [0 14 4 10 1 13 0 1 9 160]]**

SVM(Support Vector Machines)

SVM has pretty promising prediction accuracy percentage. It is about 96% for the digital recognition dataset. Its time complexity is lower than neural networks and KNN but higher than the other methodologies which are taken into consideration. The precision and recall values for the labels are also quite better.

Accuracy	**96**
Complexity	**0.314**
Precision	**[0.98882682 0.91794872 0.96045198 0.96045198 0.98360656 0.91836735 0.99444444 0.98802395 0.9695122 0.94413408]**

Recall	**[0.99438202 0.98351648 0.96045198 0.92896175 0.99447514 0.98901099 0.98895028 0.92178771 0.9137931 0.93888889]**
Confusion Matrix	**[[177 0 0 1 0 0 0 0 0 1] [0 179 7 0 0 0 0 0 8 1] [0 0 170 5 0 1 0 0 1 0] [0 0 0 170 0 0 0 0 3 4] [0 0 0 0 180 0 1 1 0 1] [1 0 0 3 0 180 0 7 2 3] [0 1 0 0 0 0 179 0 0 0] [0 0 0 2 0 0 0 165 0 0] [0 1 0 1 1 0 1 0 159 1] [0 1 0 1 0 1 0 6 1 169]]**

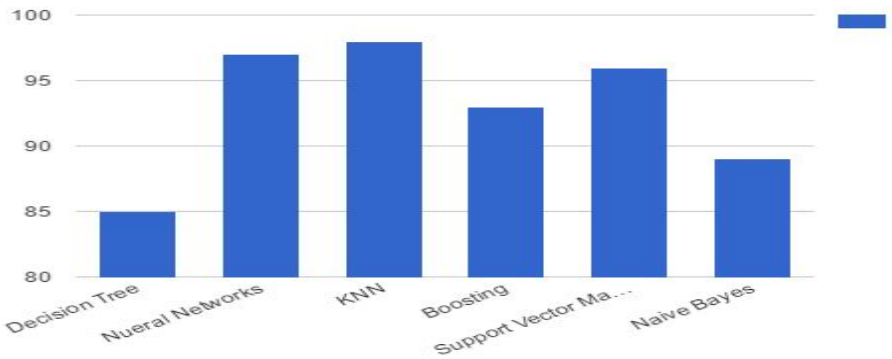
Boosting

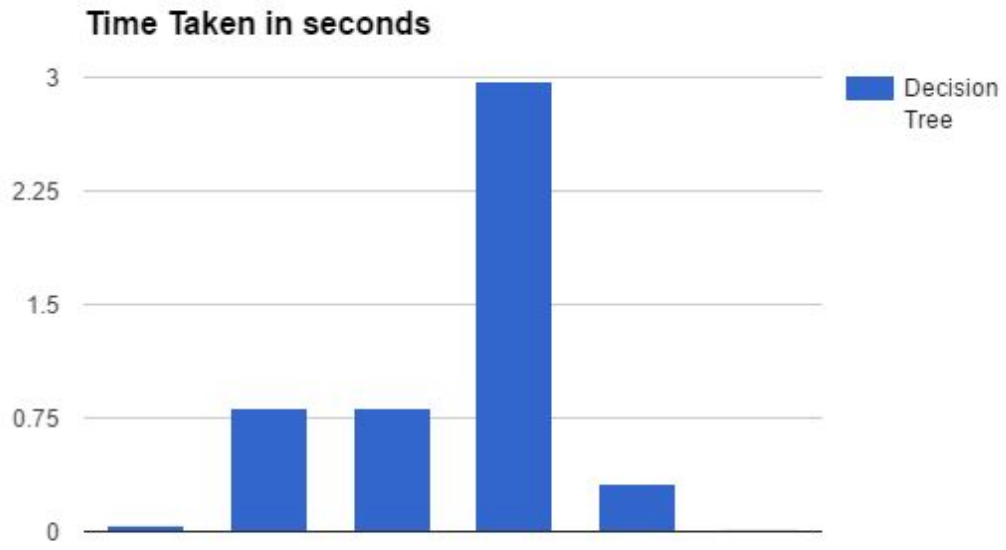
Boosting does not seems effective and efficient for this dataset. In terms of time complexity which is very high compared to other models and the accuracy percentage is lower than the most of the models other than Naive Bayes and Decision Trees.

Accuracy	**93**
Complexity	**2.975**
Precision	**[0.98882682 0.95512821 0.97126437 0.94827586 0.95135135 0.97765363 0.98870056 0.98076923 0.75348837 0.82178218]**

Recall	**[0.99438202 0.81868132 0.95480226 0.90163934 0.97237569 0.96153846 0.96685083 0.8547486 0.93103448 0.92222222]**
Confusion Matrix	**[[177 0 0 0 0 1 1 0 0 0] [0 149 0 1 3 0 2 0 1 0] [0 0 169 1 0 0 0 2 2 0] [0 1 0 165 0 0 0 0 2 6] [0 0 0 1 176 1 1 4 0 2] [0 0 0 1 0 175 0 0 0 3] [0 0 0 0 0 2 175 0 0 0] [0 0 2 1 0 0 0 153 0 0] [1 26 6 8 1 1 2 5 162 3] [0 6 0 5 1 2 0 15 7 166]]**

Following are the graphical representation of the models and their accuracy following the models and their time complexity:

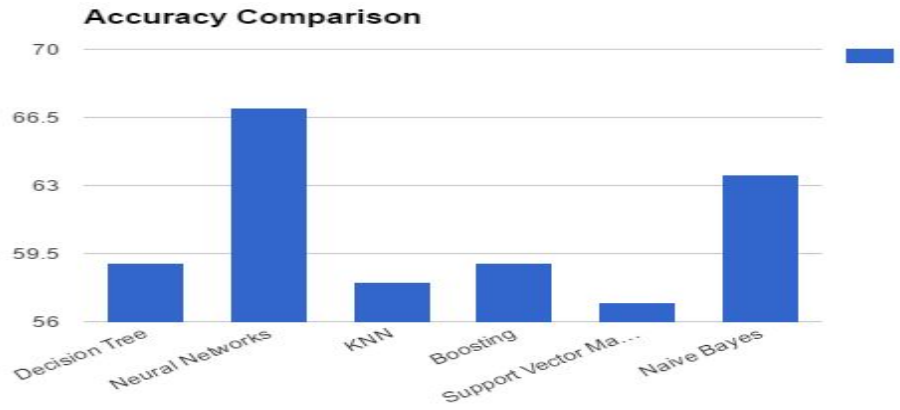


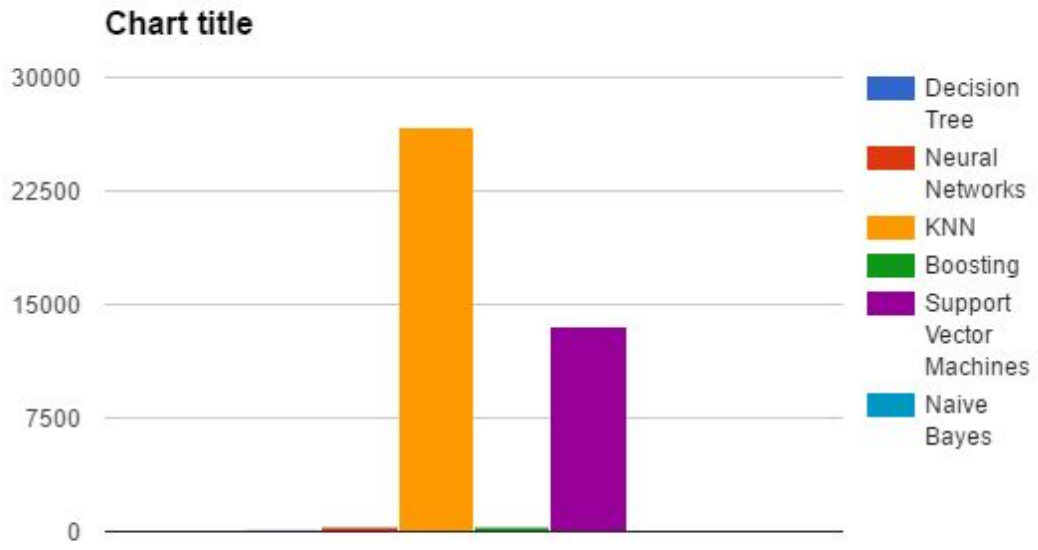


Best Pick : KNN for accuracy and Naive Bayes overall

Amazon Dataset

Let us first look into the pictorial data representation of accuracy percentage comparison and the time complexity comparison of different models which are taken into consideration:





Neural Networks

For this dataset the accuracy percentage by the neural network model is approximately 67% which is not anywhere near to the accuracy produced by the same model on the digit recognition dataset.

Following are the few parameters such as complexity,precision, recall and confusion matrix in the tabular form for this model:

Accuracy	67
Complexity	325.55
Precision	[0.51127633 0. 0.32166218 0.4545829 0.77074315]
Recall	[0.68904476 0. 0.24763033 0.33378705 0.90780007]
Confusion Matrix	[[2063 933 525 186 328] [0 0 0 0 0] [388 574 836 524 277] [87 254 977 2207 1330] [456 478 1038 3695 19052]]

Decision Tree

The accuracy percentage when Decision Tree model is used is lower than the neural networks which is 59%. Time complexity is pretty much better than the Neural network model. Following are the further details:

Accuracy	59
Complexity	112
Precision	[0.62931034 0. 0. 0.32240099 0.60258439]
Recall	[0.09752839 0. 0. 0.07879613 0.97989231]
Confusion Matrix	[[292 58 29 22 63] [0 0 0 0 0] [0 0 0 0 0] [155 194 387 521 359] [2547 1987 2960 6069 20565]]

Boosting

Boosting produces almost the similar results as the Decision Tree model but has worse complexity than the Decision Tree model. It is consuming almost thrice the time taken by the Decision Tree model. Following are the further details:

Accuracy	59
Complexity	323.39
Precision	[0.62931034 0. 0. 0.32198142 0.60258439]
Recall	[0.09752839 0. 0. 0.07864489 0.97989231]
Confusion Matrix	[[292 58 29 22 63] [0 0 0 1 0] [0 0 0 0 0] [155 194 387 520 359] [2547 1987 2960 6069 20565]]

Naive Bayes

Naive Bayes produces better results than Decision Tree and Boosting but not better than the Neural Networks. It has almost 63.55% of accurate predictions. Time complexity is however promising for this model in this dataset.

Accuracy	63.55
Complexity	25.92
Precision	[0.48099261 0.25031766 0.32773438 0.40935132 0.77560656]
Recall	[0.60855043 0.17597142 0.24851896 0.33499698 0.84538047]
Confusion Matrix	[[1822 649 389 300 628] [364 394 338 227 251] [286 469 839 600 366] [131 267 798 2215 2000] [391 460 1012 3270 17742]]

KNN

K-Nearest Neighbor or KNN produces results almost similar to Boosting and Decision Tree. However, the time taken to compute it is very high compared to the other models.

Accuracy	58
Complexity	26650
Precision	[0.62931034 0. 0. 0.32198142 0.60258439]
Recall	[0.09752839 0. 0. 0.07864489 0.97989231]

Confusion Matrix	[[292 58 29 22 63] [0 0 0 1 0] [0 0 0 0 0] [155 194 387 520 359] [2547 1987 2960 6069 20565]]

SVM(Support Vector Machines)

SVM does not seem to be a better model for this type of dataset. It has worse time complexity as well as the less prediction percentage. Following are the details:

Accuracy	57
Complexity	13600
Precision	[0.48099261 0.25031766 0.32773438 0.40935132 0.77560656]
Recall	[0.60855043 0.17597142 0.24851896 0.33499698 0.84538047]
Confusion Matrix	[[1822 649 389 300 628] [364 394 338 227 251] [286 469 839 600 366] [131 267 798 2215 2000] [391 460 1012 3270 17742]]

Best Pick : Neural Network for accuracy and Naive Bayes overall