Machine Learning

2228 - CSE 6363 - SEC 002

HW1 - REPORT

Names of Group members

1. Pratik Antoni Patekar (1001937948)
2. Amrita Singh (1001937490)
3. Ruthvik Kumar Myadam (1002026231)
4. Prathibha Lakkidi (1001962876)
5. Harshini Kandimalla (1001960046)

Table of contents

1.0. Results

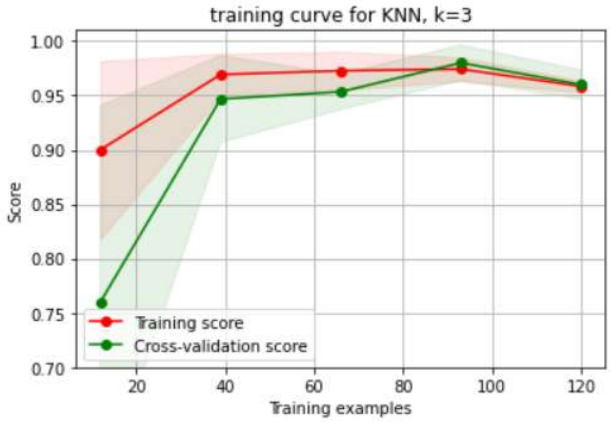
1.1. Iris dataset:

1.1.1. Model 1: KNN model:

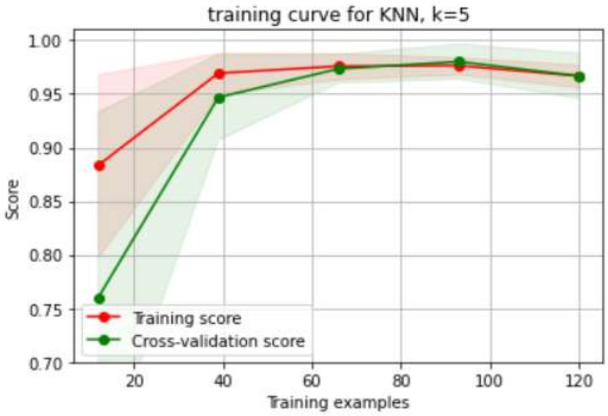
Values of N	Accuracy	Precision	Recall	F1 score
N = 3	0.973	0.973	0.971	0.97
N = 5	0.947	0.95	0.942	0.942
N = 7	0.947	0.95	0.942	0.942
N = 9	0.96	0.962	0.956	0.956

Learning Curve

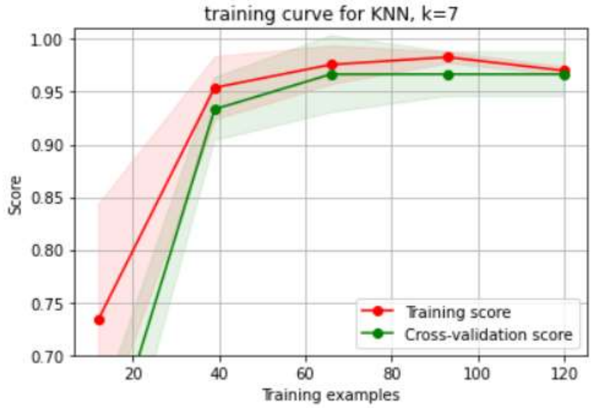
k=3



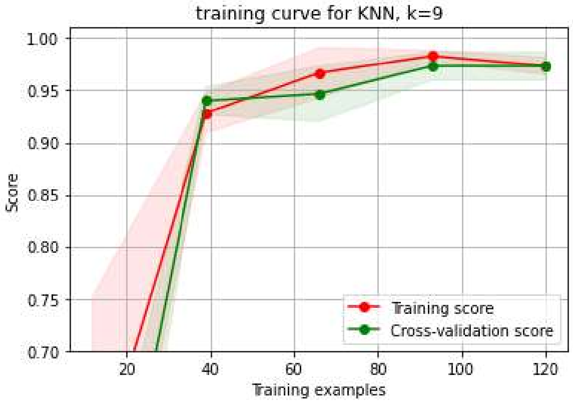
k=5



k=7



k=9

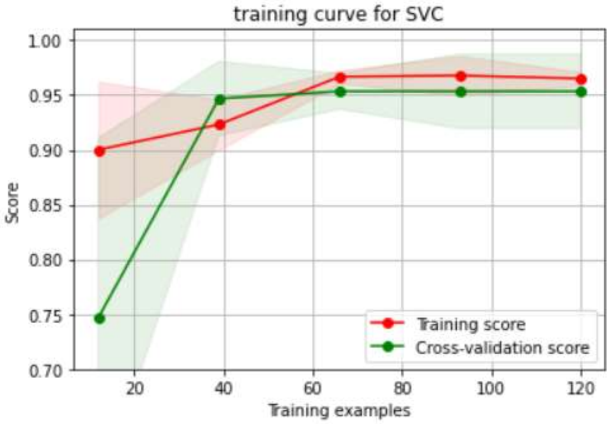


1.1.2. <u>SVC model:</u>

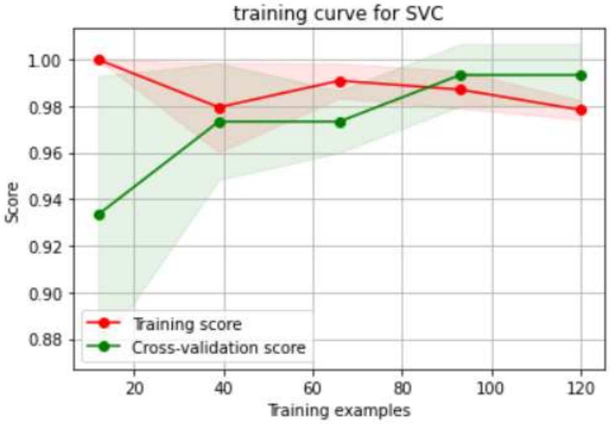
Parameter kernel	Accuracy	Precision	Recall	F1 score
rbf	1	1	1	1
linear	1	1	1	1
poly	0.96	0.962	0.957	0.956
sigmoid	0.173	0.12	0.188	0.147

Learning Curve

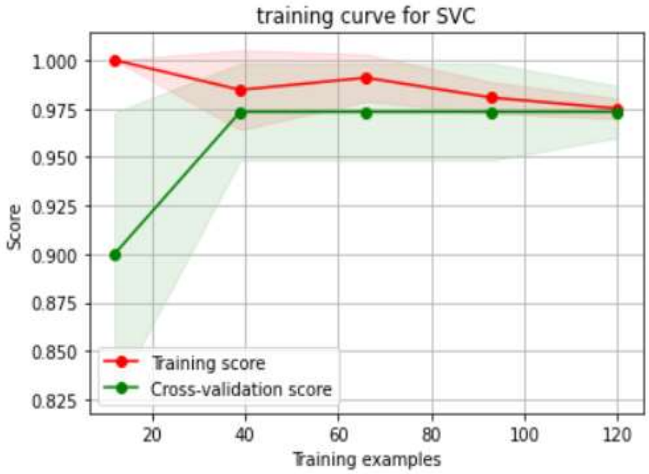
Kernel: rbf



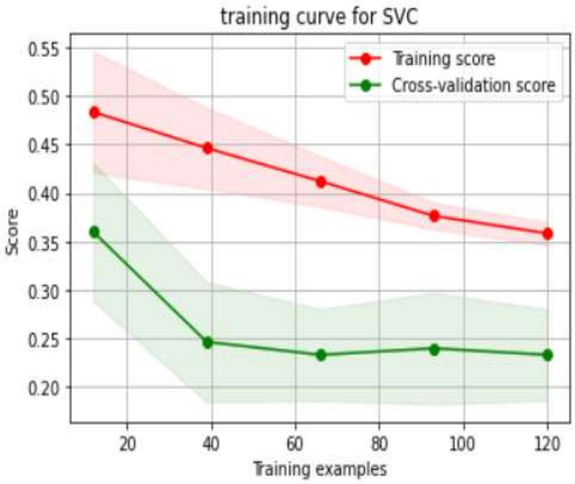
Kernel: linear



Kernel: poly



Kernel: sigmoid

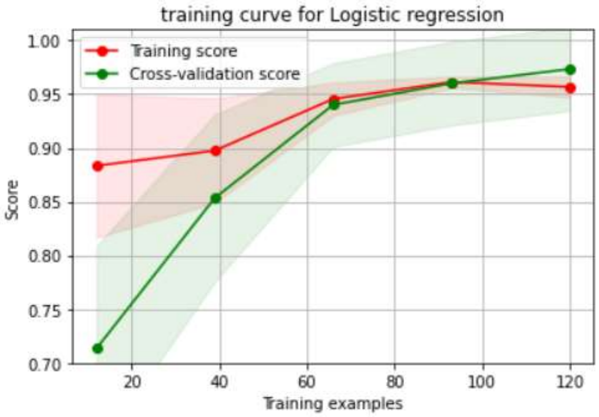


1.1.3. Logistic regression:

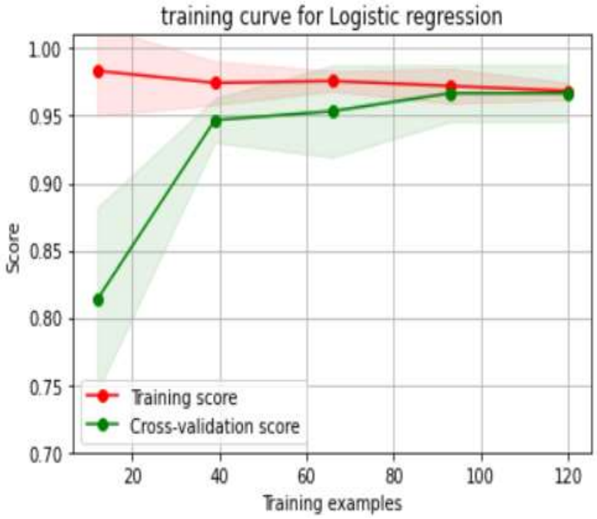
Penalty	Solver	Accuracy	Precision	Recall	F1 score
l1	liblinear	0.987	0.986	0.986	0.986
l2	newton-cg	1	1	1	1
none	lbfgs	0.96	0.962	0.957	0.956
l2	sag	1	1	1	1
l2	saga	1	1	1	1
none	saga	0.987	0.986	0.986	0.986

Learning curve

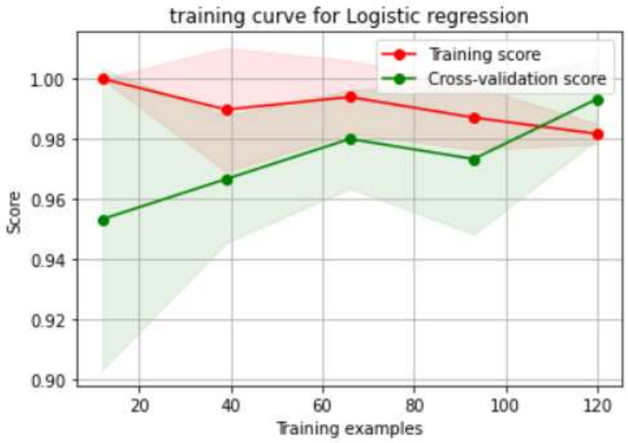
Penalty: l1 **Solver: liblinear**



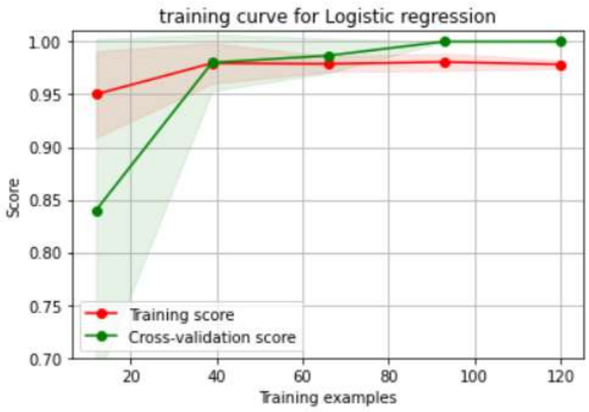
Penalty: l2 **Solver: newton-cg**



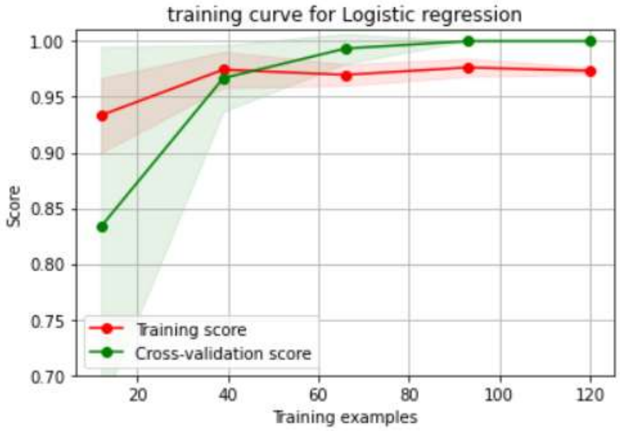
Penalty: none Solver: lbfgs



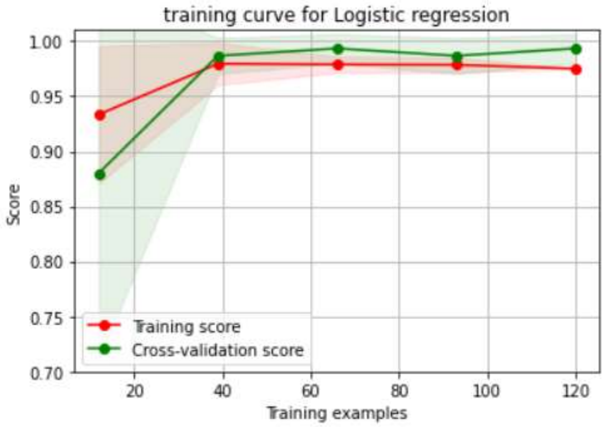
Penalty: l2 Solver: sag



Penalty: l2 Solver: saga



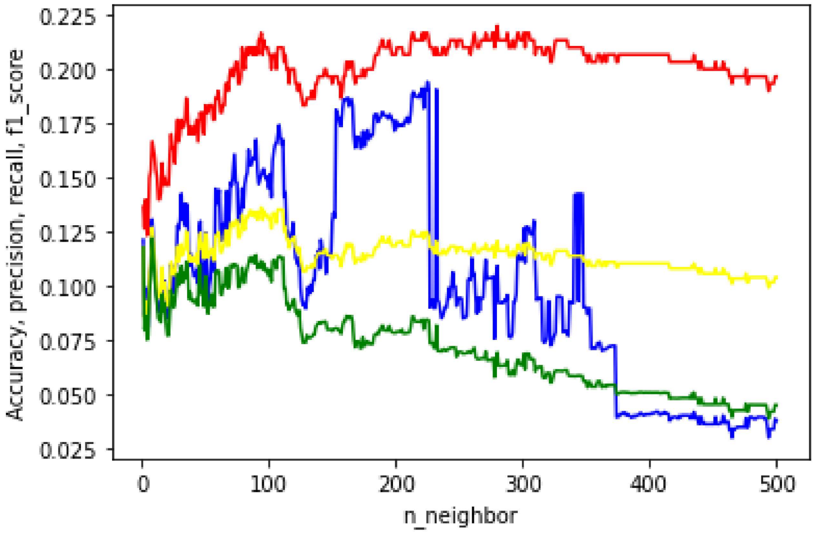
Penalty: none Solver: saga



1.2. SVHN Dataset:

1.2.1. KNN model:

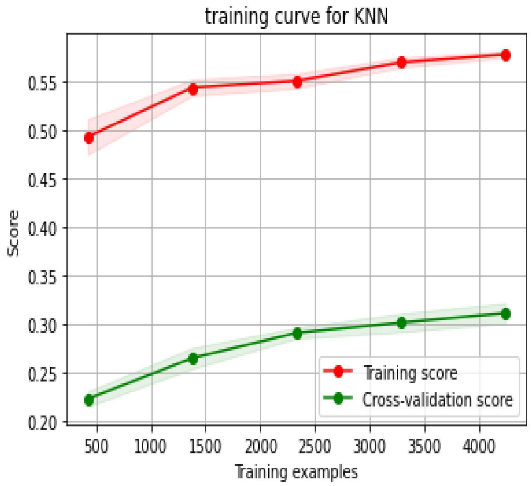
Values of N	Accuracy	Precision	Recall	F1 score
N = 3	0.14	0.10	0.09	0.08
N = 5	0.13	0.08	0.08	0.07
N = 7	0.15	0.11	0.11	0.10
N = 9	0.16	0.12	0.12	0.11
N=100	0.20	0.14	0.13	0.10
N=200	0.20	0.17	0.11	0.07
N=300	0.21	0.11	0.11	0.06
N=400	0.20	0.04	0.11	0.05
N=500	0.19	0.03	0.10	0.04



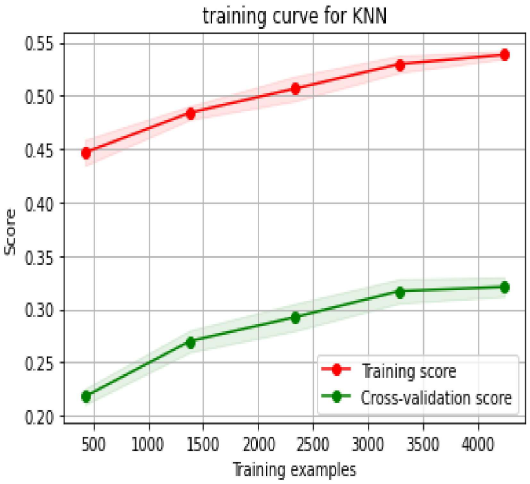
As we were not getting better accuracy for smaller values of n-neighbors, we plotted a graph showing the trends of all the metrics for different n values. The above plot shows the accuracy, precision, recall and f1 score for different values of n-neighbors (from 1 to 501). We can see that all the parameters have approximate maximum values at n-neighbors = 100. Following are the metrics calculated for the n-neighbors = 210.

Learning Curve:

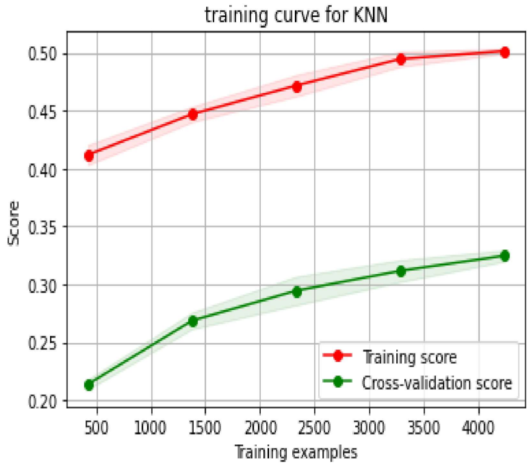
N = 3



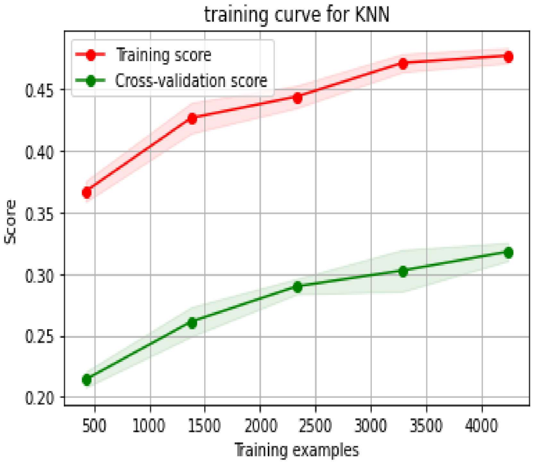
N = 5



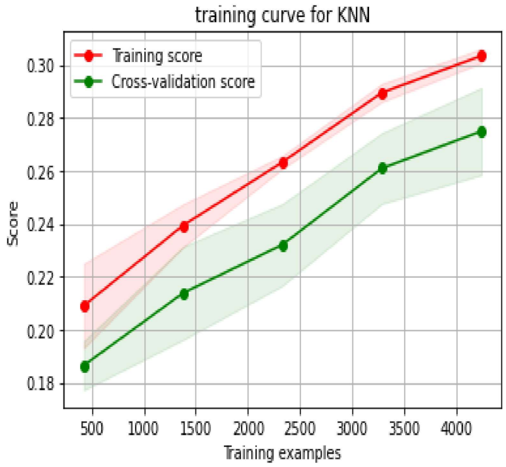
N = 7



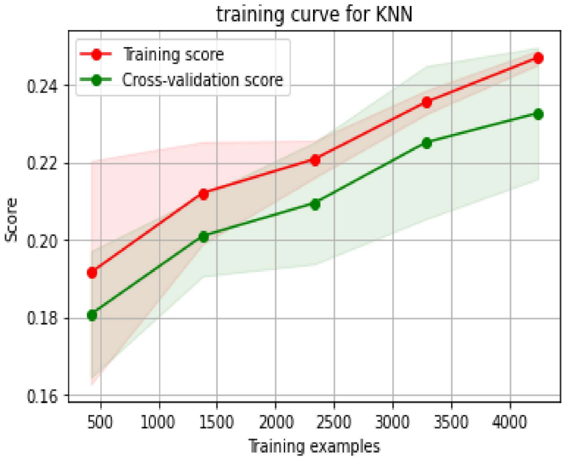
N = 9



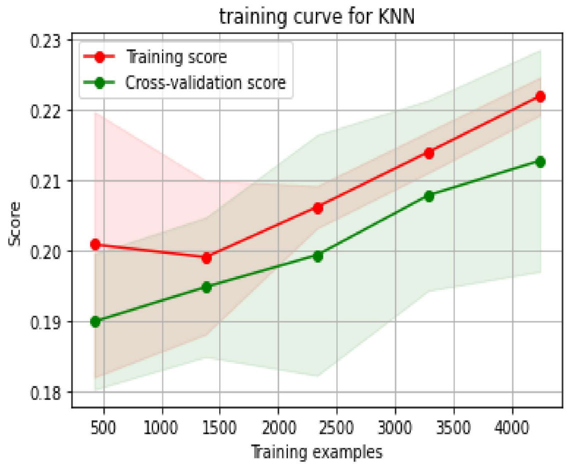
N = 100



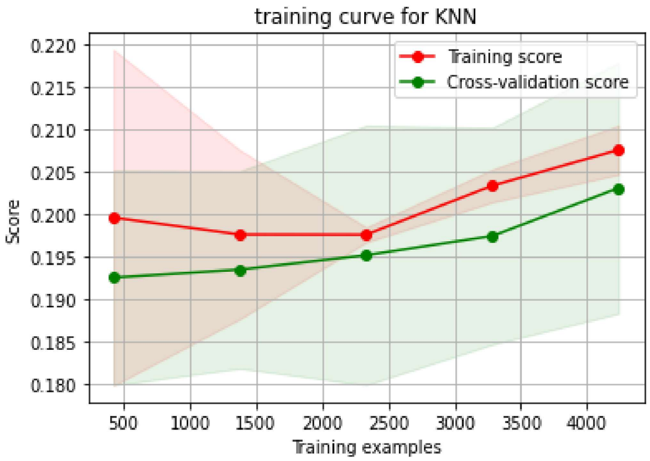
N = 200



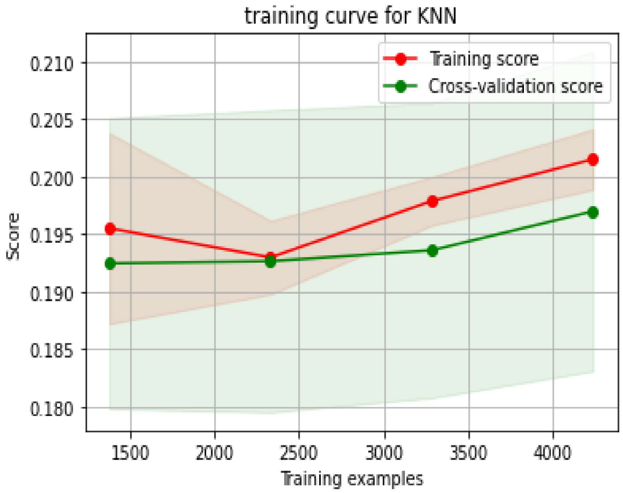
N = 300



N = 400



N = 500

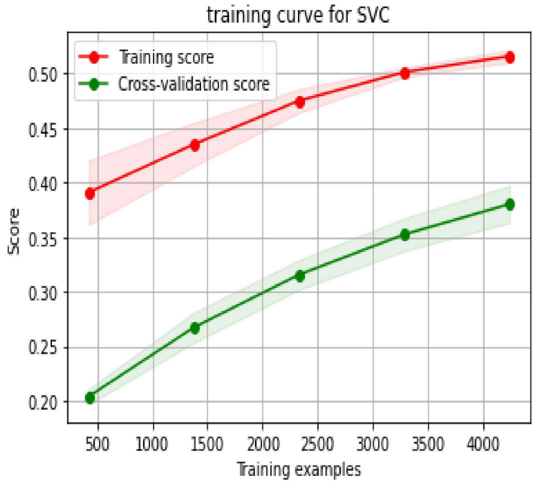


1.2.2. SVC model:

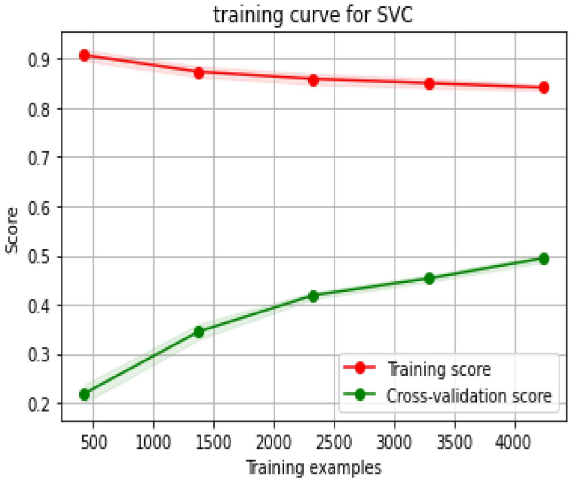
Parameter kernel	Accuracy	Precision	Recall	F1 score
rbf	0.18	0.10	0.11	0.09
poly	0.11	0.08	0.08	0.08
sigmoid	0.15	0.04	0.12	0.06

Learning Curve

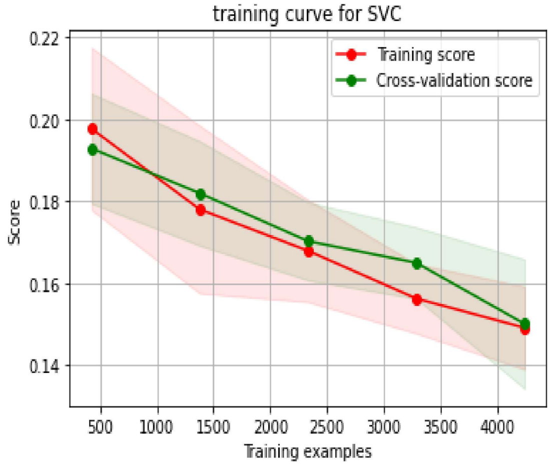
Kernel: rbf



Kernel: poly



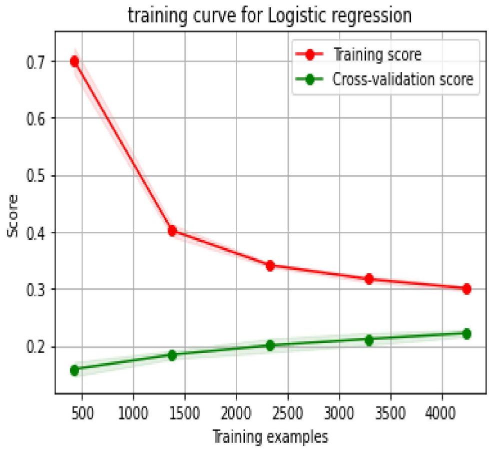
Kernel: sigmoid



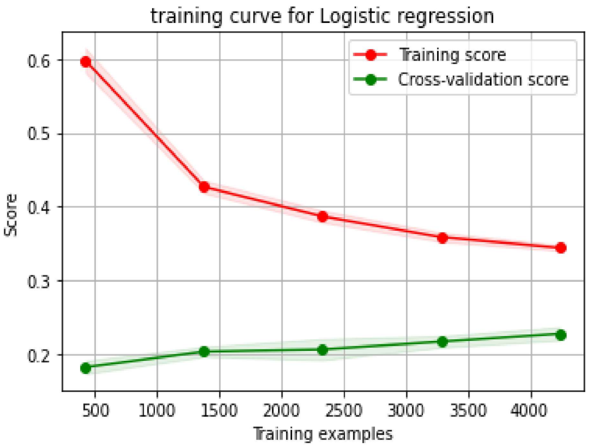
1.2.3. Logistic regression:

Penalty	Solver	Accuracy	Precision	Recall	F1 score
none	lbfgs	0.14	0.09	0.10	0.09
l2	sag	0.16	0.11	0.11	0.10
l2	saga	0.17	0.12	0.12	0.10
none	saga	0.17	0.12	0.12	0.11

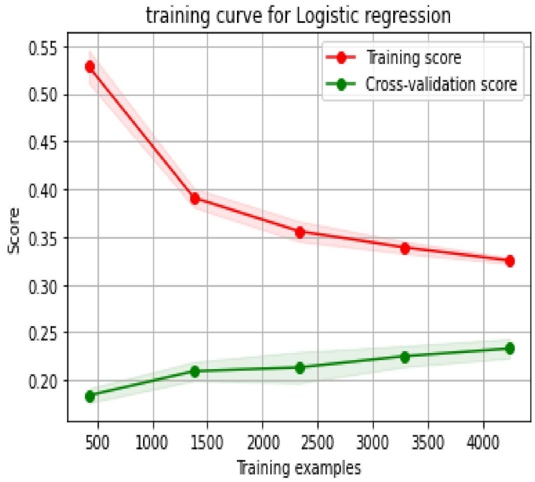
Penalty: none **Solver: lbfgs**



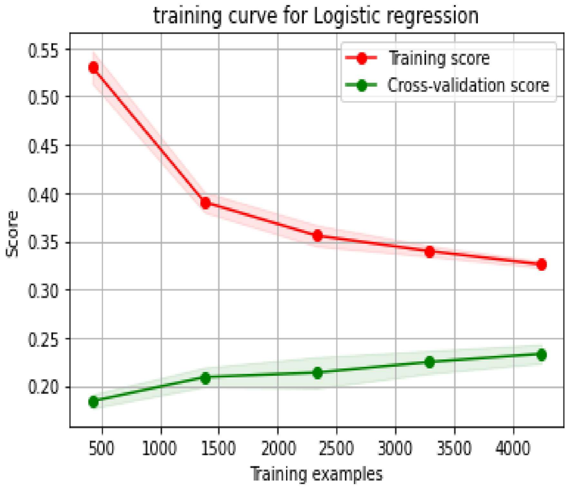
Penalty: l2 **Solver: sag**



Penalty: l2 **Solver: sag**



Penalty: none **Solver: saga**



2.0. Discussions:

2.1. Dataset 1: Iris

For the ***KNN model***, the n-neighbors number of 3 provides the best accuracy. Since the dataset is balanced (the number of data points in each class is the same), we can use accuracy to discover the ideal parameter value. Also, we can observe that the accuracy improves with larger n-neighbor numbers (9 and beyond), although it takes more processing to do so. As a result, with an n-neighbor number of 3, we have the best of both worlds: less computation and more accuracy.

For the ***SVC model***, we gain greater accuracy for kernel parameters of 'rbf' and 'linear'. The time required to train the SVC model with the rbf kernel is not linear in terms of computation.Therefore, for the parameter kernel in SVC model, linear is best value.

On similar lines, for ***Logistic Regression***, we can see that the L2 penalty has the best accuracy, precision, recall, and f1 score. And for the solver parameter, in combination with the L2 penalty, ‘newton-cg’, ‘sag’ and ‘saga’ provide the best metric results.

2.2. Dataset 2: SVHN

In comparison to the first dataset(iris), the SVHN dataset is more imbalanced as the target class has an uneven number of observations. Thus we cannot depend solely on accuracy to select the optimum parameter values because it might be misleading. We need to take precision and recall into consideration as well.
Note: *The values of the parameters in this dataset are low as the training has been done with a small dataset (5000).*

For the ***KNN model***, based on the four parameters i.e. accuracy, precision, recall and the f1 score, we can see that when N is set to 100, the values are more ideal. Despite the fact that the greatest f1 score is 0.11 for N = 9, we can see that the accuracy is better between values 100 and 200. As a result, we can say that both f1 score and accuracy are better for N = 100 which is the best case.

For the ***SVC model***, we can observe that the f1 score and the accuracy are higher for the rbf kernel, hence it is the best case.

For ***Logistic Regression***, as the parameter values are higher with accuracy 0.17 and f1 score 0.11 we can conclude that the solver option ‘saga’ with penalty none is the best case scenario.