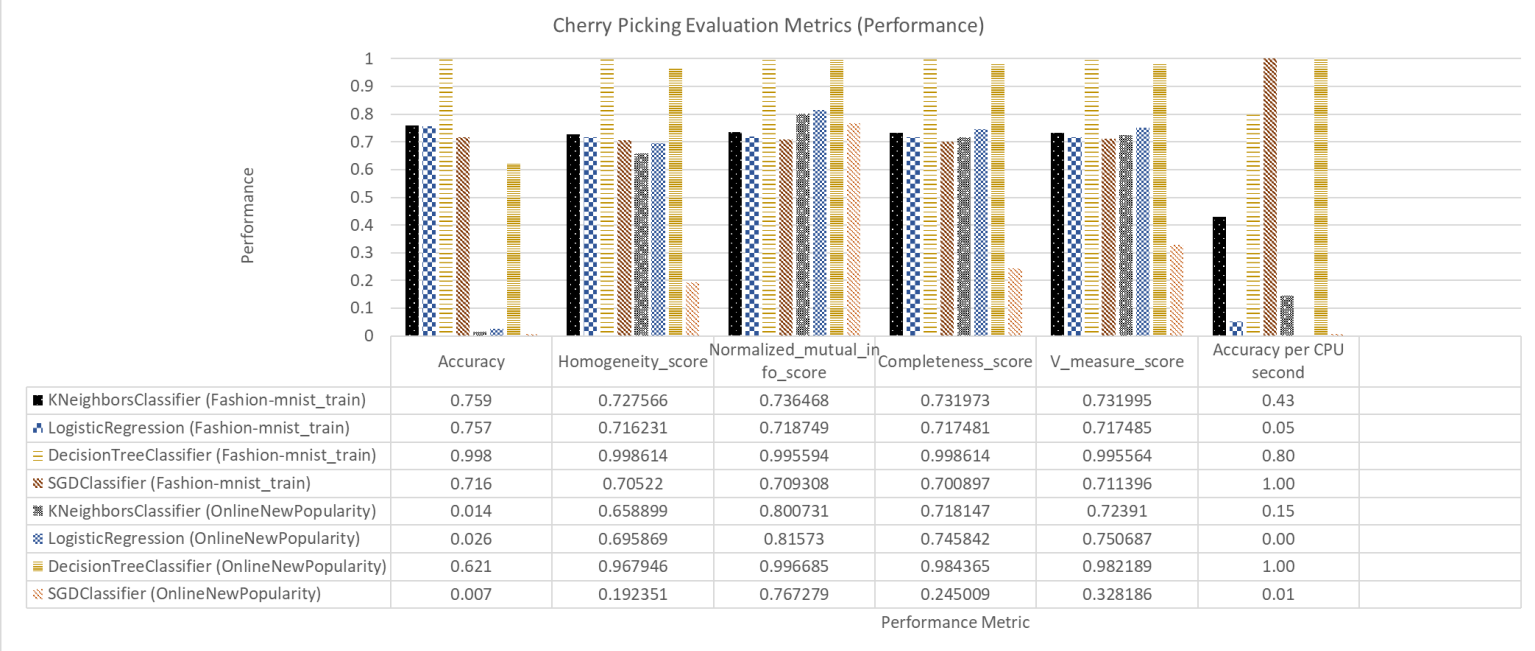
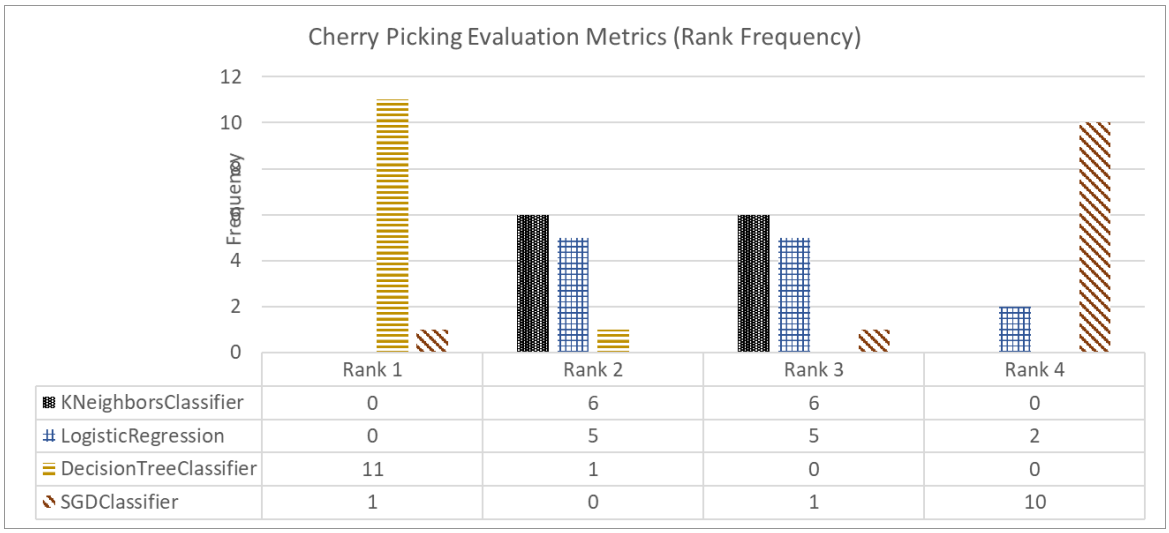
Cherry-Picking Evaluation Metrics: Report

Team: 22
Student IDs: Kevin Morris – 14315027
Joseph Fitzpatrick- 14312993
Callum Duffy - 14315135
Total Time Required (in hours): 15



	Accuracy	Homogeneity_score	Normalized_mutual_info_score	Completeness_score	V_measure_score	Accuracy per CPU second
KNeighborsClassifier (Fashion-mnist_train)	0.759	0.727566	0.736468	0.731973	0.731995	0.43
LogisticRegression (Fashion-mnist_train)	0.757	0.716231	0.718749	0.717481	0.717485	0.05
DecisionTreeClassifier (Fashion-mnist_train)	0.998	0.998614	0.995594	0.998614	0.995564	0.80
SGDClassifier (Fashion-mnist_train)	0.716	0.70522	0.709308	0.700897	0.711396	1.00
KNeighborsClassifier (OnlineNewPopularity)	0.014	0.658899	0.800731	0.718147	0.72391	0.15
LogisticRegression (OnlineNewPopularity)	0.026	0.695869	0.81573	0.745842	0.750687	0.00
DecisionTreeClassifier (OnlineNewPopularity)	0.621	0.967946	0.996685	0.984365	0.982189	1.00
SGDClassifier (OnlineNewPopularity)	0.007	0.192351	0.767279	0.245009	0.328186	0.01



	Rank 1	Rank 2	Rank 3	Rank 4
KNeighborsClassifier	0	6	6	0
LogisticRegression	0	5	5	2
DecisionTreeClassifier	11	1	0	0
SGDClassifier	1	0	1	10

Findings/Answer (200-300 words)

How meaningful are evaluation metrics ("meaningful" in terms of how consistent are they in assessing the performance of machine-learning algorithms)?

Evaluation metrics are an extremely useful tool when it comes to machine learning. They influence the comparison of different machine learning algorithms and allow them to be measured well, to add meaning to the values and data you receive from the algorithm. The most common way to demonstrate a metric is to use a 10-fold cross-validation test harness, which is what we have done here. This is done because it is the most likely scenario where evaluation metrics will be employed.

At first reading, the results of the test harness are difficult to understand, and therefore it is hard to come to a conclusion as to whether the machine has learnt something from the data, and this is where evaluation metrics come in. It is by running metrics with the test harness that we finally add meaning to this resultant data. Each metric scores the algorithm that ran over the data to allow us to consistently see across all of the algorithms, how ell each has performed in comparison to each other.

The majority of metrics give their results in ascending order, however there are some which give results in descending order, for example mean squared error, we tried to avoid these as we decided in order to make our results as meaningful as they could be, we wanted to make it easy to plot all of our results on the same scale.

This decision was what we saw as the best way for our metric results to maintain consistency across all of our machine learning algorithms, as it was then very easy to read across the graphs and charts to see how each algorithm performed.

Additional Information

We as a team only managed to come up with one 'made up metric'. And this was the performance score of accuracy per cpu second. To do this we recorded the time taken for the accuracy score for each algorithm, and then divided the accuracy by time taken. However these values were a bit difficult to read across the rest of out charts, so to add consistency to these values we decided to divide all results by the highest score. Therefore giving a result from 0 to 1, where the results could be easily compared to each other.

Data, Algorithms, etc.

Novel Algorithm	KNeighborsClassifier
Baseline Algorithm 1	LogisticRegression
Baseline Algorithm 2	DecisionTreeClassifier
Baseline Algorithm 3	SGDClassifier

Dataset 1	Fashion-mnist_train
Dataset 2	OnlineNewPopularity
Common Metric 1	Accuracy
Common Metric 2	Homogeneity_score
Common Metric 3	Normalized_mutual_info_score
Common Metric 4	Completeness_score
Common Metric 5	V_measure_score
Made Up Metric 1	Accuracy per CPU second
Made Up Metric 2	$made_up_metric_2

Contributions

For this task3 we all met up and picked whether use four algorithms of our choice for either a regression/prediction or a classification task. We then as a team chose the 5-evaluation metrics to use and spoke about the 2 new evaluation metrics to choose. Callum Duffy and Kevin Morris picked a dataset each and worked on the structure of the data to ensure they were compatible with the algorithms and wrote the code to implementing the metrics on the algorithms with the help of Joseph Fitzpatrick. For the two-new evaluation metrics we all worked together to implement our one new evaluation metric.