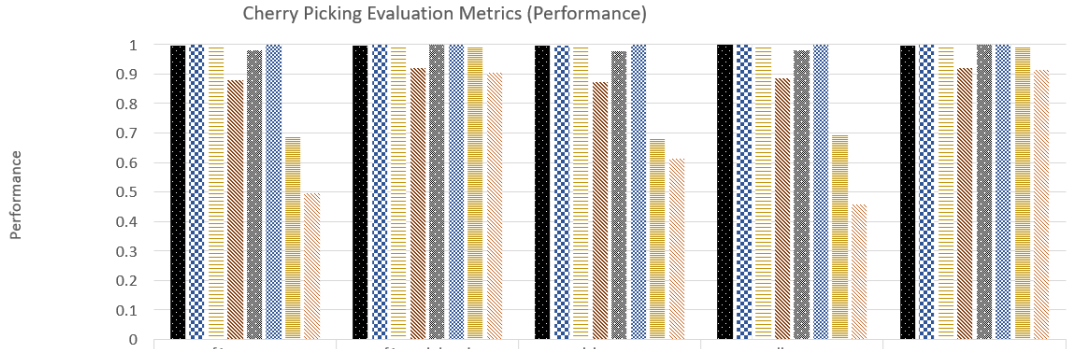
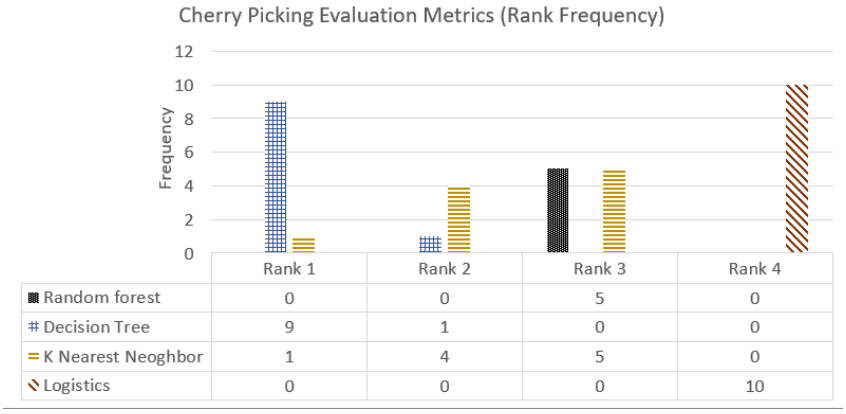
Cherry-Picking Evaluation Metrics: Report

Team: team_49
Student IDs: 17316735, 17311349, 17311213
Total Time Required (in hours): 25



	f1_macro	f1_weighted	precision_macro	recall_macro	accuracy
Random forest (skinnonskin)	0.996234871	0.997523278	0.994780047	0.997711803	0.997519199
Decision Tree(skinnonskin)	0.997739522	0.998515515	0.997296993	0.998188141	0.998514661
K Nearest Neoghbor (skinnonskin)	0.997580165	0.998402966	0.996243081	0.998933997	0.998400389
Logistics (skinnonskin)	0.879125776	0.919999219	0.874002637	0.884624715	0.919431332
Random forest (Sum Without Noise)	0.979029312	0.999959009	0.978315428	0.979816514	0.999966948
Decision Tree(Sum Without Noise)	0.999729535	0.999995886	0.999519231	0.999944568	0.999995868
K Nearest Neoghbor (Sum Without Noise)	0.686607113	0.991665468	0.680685308	0.692853123	0.99262913
Logistics (Sum Without Noise)	0.496963062	0.90363491	0.611527955	0.457325253	0.914925183



	Rank 1	Rank 2	Rank 3	Rank 4
Random forest	0	0	5	0
Decision Tree	9	1	0	0
K Nearest Neoghbor	1	4	5	0
Logistics	0	0	0	10

Findings/Answer (200-300 words)

Our observation showed that Decision Tree algorithm performed better than our Novel Algorithm Random Forest overall metrics.

- According to our results on skin-no-skin datasets- Random Forest, Decision Tree and K-Nearest Neighbor have performed consistently on all evaluation metrics. However Logistic regression failed on the consistent performance on all metrics. On Sum-without-noise database, we observed different performances of the K-Nearest Algorithm and Random forest algorithms on

different metrics. Where we can observe performance for K-Nearest Algorithm at 0.9916 and 0.9926 on f1_weighted and accuracy metrics respectively, we make completely opposite observations on f1_macro, precison_macro, and recall_macro evaluating at 0.686, 0.68 and 0.69 respectively.

- The same was true for **Logistic regression**, with f1_macro evaluating to 0.49 vs 0.87, precision_macro evaluating to 0.61 vs 0.87 and recall_macro evaluating to 0.45 vs 0.88 on the sum_without_noise and skin-no-skin datasets respectively, while all other algorithms performed consistently well in both the datasets on all metrics.

Therefore, we can conclude that algorithm performance cannot be judged using a single evaluation metric. Each metric evaluates algorithm differently focusing on different parts.

Additional Information

- Taking up the whole records from the dataset for execution had taken long execution time so we have considered 500,000 of rows for creating our dataset.
- Random Forest was taken as Novel algorithm and remaining 3 are the baseline
- Sum dataset was not suitable Logistics Regression Algorithm
- Unfortunately, we could not come up with any made-up metrics

Data, Algorithms, etc.

Novel Algorithm	Random forest
Baseline Algorithm 1	Decision Tree
Baseline Algorithm 2	K Nearest Neighbor - 6
Baseline Algorithm 3	Logistics
Dataset 1	Skin-noskin
Dataset 2	Sum dataset(Without Noise)
Common Metric 1	f1_macro
Common Metric 2	f1_weighted
Common Metric 3	precision_macro
Common Metric 4	recall_macro
Common Metric 5	accuracy
Made Up Metric 1	None
Made Up Metric 2	None

Contributions (max. 200 words)

17311213 implemented Decision Tree Algorithm

17311349 implemented Random forest and Logistic Regression

17616735 implemented K-Nearest Algorithm with 6 neighbors.

Everyone contributed equally for report generation.