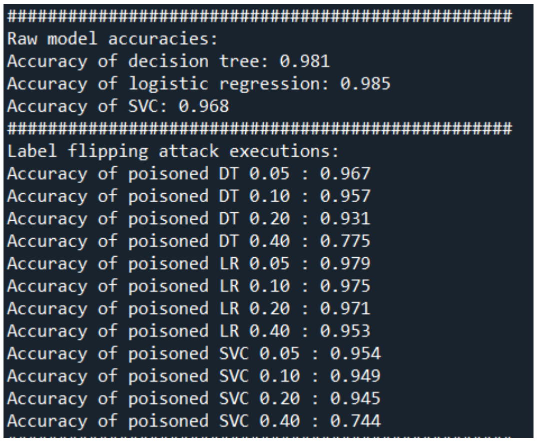
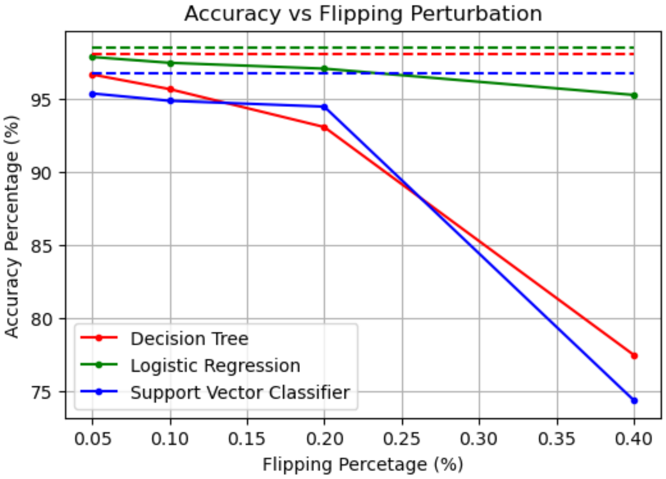
Alp Kural – 71959
Computer Engineering

Comp 430 – Homework 03 Project Report

PART 1 / TASK 1



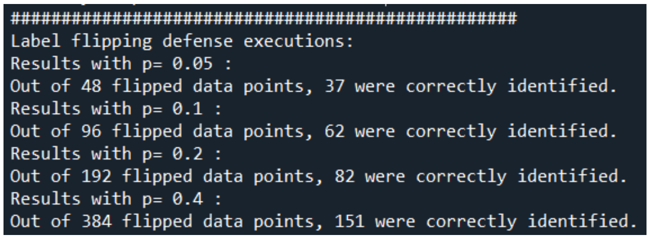
```
####################################################
Raw model accuracies:
Accuracy of decision tree: 0.981
Accuracy of logistic regression: 0.985
Accuracy of SVC: 0.968
####################################################
Label flipping attack executions:
Accuracy of poisoned DT 0.05 : 0.967
Accuracy of poisoned DT 0.10 : 0.957
Accuracy of poisoned DT 0.20 : 0.931
Accuracy of poisoned DT 0.40 : 0.775
Accuracy of poisoned LR 0.05 : 0.979
Accuracy of poisoned LR 0.10 : 0.975
Accuracy of poisoned LR 0.20 : 0.971
Accuracy of poisoned LR 0.40 : 0.953
Accuracy of poisoned SVC 0.05 : 0.954
Accuracy of poisoned SVC 0.10 : 0.949
Accuracy of poisoned SVC 0.20 : 0.945
Accuracy of poisoned SVC 0.40 : 0.744
```



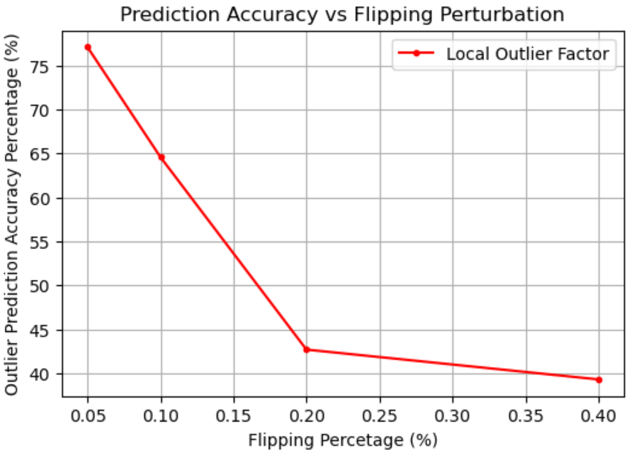
ps. Dotted lines are the unperturbed accuracy percentages of respective models.

Label flipping attack results the accuracy loss of examined supervised machine learning models. As the perturbation percentage of the training data increases, the accuracy of the models on unseen test data decreases. From the experiment, the Logistic Regression model is examined to be more robust against the attack among the other classifiers participated to the experiment listed as Support Vector Classifier and Decision Tree.

PART 1 / TASK 2



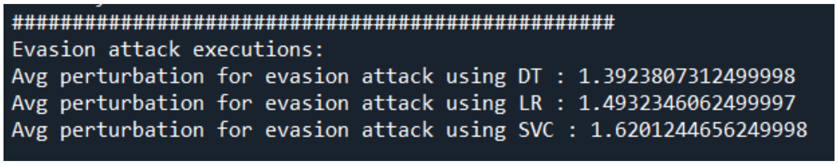
```
####################################################
Label flipping defense executions:
Results with p= 0.05 :
Out of 48 flipped data points, 37 were correctly identified.
Results with p= 0.1 :
Out of 96 flipped data points, 62 were correctly identified.
Results with p= 0.2 :
Out of 192 flipped data points, 82 were correctly identified.
Results with p= 0.4 :
Out of 384 flipped data points, 151 were correctly identified.
```



For the defense of the label flipping attack, Local Outlier Factor model is used to predict the tuples in the dataset that undergo the label flipping attack. The main benefit of Local Outlier Factor model lies in its ability to identify local anomalies, meaning it can detect outliers that are not necessarily global outliers but are anomalous within their local neighborhoods which helps us to follow the heuristic of "if the current data point is surrounded by samples of the opposite class, it was probably flipped".

From the conducted experiment, The Local Outlier Factor model demonstrates high accuracy results in the low perturbation percentages and the overall effectiveness of the defense is examined to be higher than 39% which is reasonably effective. It would be beneficial to train the machine learning model with sainting the dataset from the tuples that are labeled as outliers as the output of the Local Outlier Factor model.

PART 1 / TASK 3

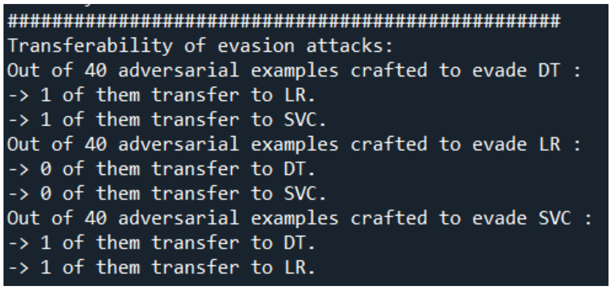


```
##################################################
Evasion attack executions:
Avg perturbation for evasion attack using DT : 1.3923807312499998
Avg perturbation for evasion attack using LR : 1.4932346062499997
Avg perturbation for evasion attack using SVC : 1.6201244656249998
```

The designed heuristic algorithm satisfies the defined success condition which is accomplishing the evasion attack to the ML model with the average perturbation of the original data less than 3.

The algorithm starts with 0.15 step size and exponentially grows at each iteration. At an iteration all the possible combinations of addition and subtraction of a step to the each feature of the original data is applied and checked for whether the prediction label of the ML model is flipped or not.

PART 1 / TASK 4



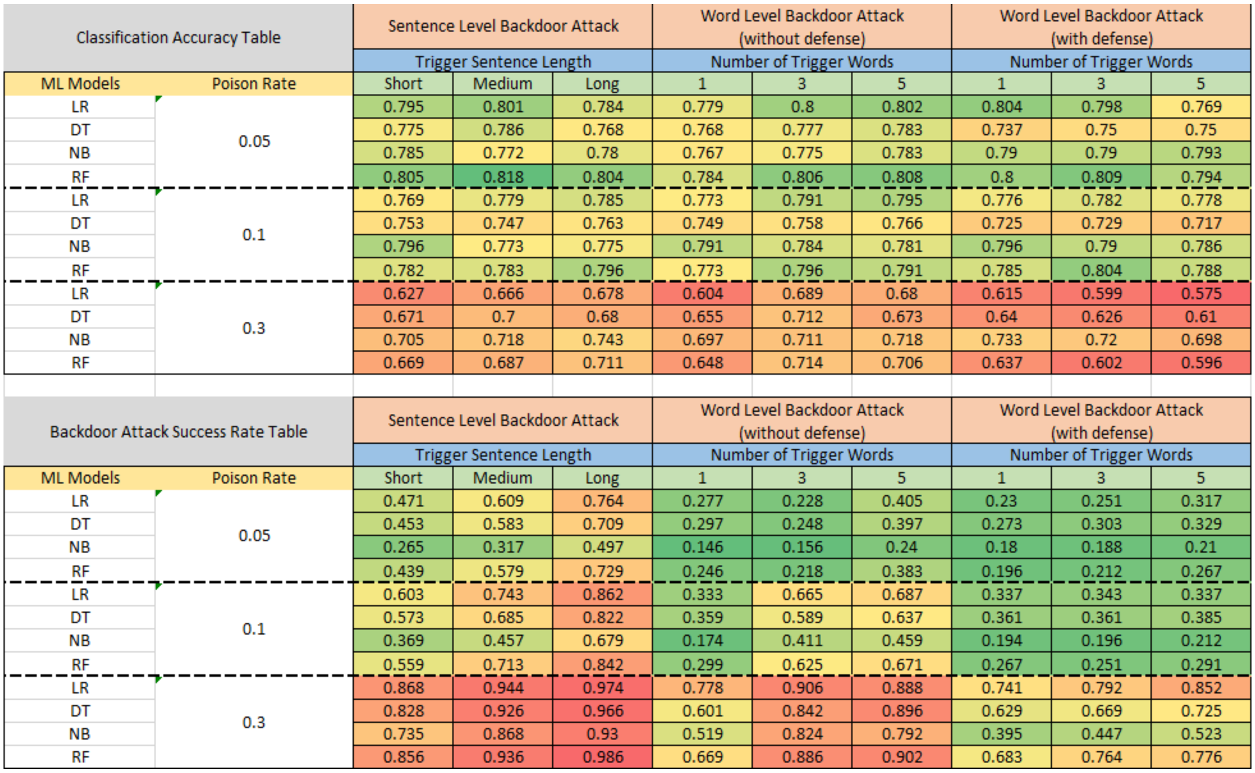
```
##################################################
Transferability of evasion attacks:
Out of 40 adversarial examples crafted to evade DT :
-> 1 of them transfer to LR.
-> 1 of them transfer to SVC.
Out of 40 adversarial examples crafted to evade LR :
-> 0 of them transfer to DT.
-> 0 of them transfer to SVC.
Out of 40 adversarial examples crafted to evade SVC :
-> 1 of them transfer to DT.
-> 1 of them transfer to LR.
```

According to the experiment results, it is hard to tell that the designed heuristic evasion attack algorithm is cross-model transferable.

PART 2 / TASK 3

The defense algorithm employs tokenization with space as a delimiter, therefore, obtaining each token as a word. Algorithm tries to eliminate non-English words from the comments which are treated as suspicious and potential backdoor attack triggers. To determine if a token is in the English vocabulary, it is necessary to remove leading or trailing punctuations in the token first. This is because the comparison with the English vocabulary dataset doesn't account for words with punctuation as a leading or trailing character.

PART 2 / TASK 4



Classification Accuracy Table		Sentence Level Backdoor Attack			Word Level Backdoor Attack (without defense)			Word Level Backdoor Attack (with defense)		
		Trigger Sentence Length			Number of Trigger Words			Number of Trigger Words		
ML Models	Poison Rate	Short	Medium	Long	1	3	5	1	3	5
LR	0.05	0.795	0.801	0.784	0.779	0.8	0.802	0.804	0.798	0.769
DT		0.775	0.786	0.768	0.768	0.777	0.783	0.737	0.75	0.75
NB		0.785	0.772	0.78	0.767	0.775	0.783	0.79	0.79	0.793
RF		0.805	0.818	0.804	0.784	0.806	0.808	0.8	0.809	0.794
LR	0.1	0.769	0.779	0.785	0.773	0.791	0.795	0.776	0.782	0.778
DT		0.753	0.747	0.763	0.749	0.758	0.766	0.725	0.729	0.717
NB		0.796	0.773	0.775	0.791	0.784	0.781	0.796	0.79	0.786
RF		0.782	0.783	0.796	0.773	0.796	0.791	0.785	0.804	0.788
LR	0.3	0.627	0.666	0.678	0.604	0.689	0.68	0.615	0.599	0.575
DT		0.671	0.7	0.68	0.655	0.712	0.673	0.64	0.626	0.61
NB		0.705	0.718	0.743	0.697	0.711	0.718	0.733	0.72	0.698
RF		0.669	0.687	0.711	0.648	0.714	0.706	0.637	0.602	0.596

Backdoor Attack Success Rate Table		Sentence Level Backdoor Attack			Word Level Backdoor Attack (without defense)			Word Level Backdoor Attack (with defense)		
		Trigger Sentence Length			Number of Trigger Words			Number of Trigger Words		
ML Models	Poison Rate	Short	Medium	Long	1	3	5	1	3	5
LR	0.05	0.471	0.609	0.764	0.277	0.228	0.405	0.23	0.251	0.317
DT		0.453	0.583	0.709	0.297	0.248	0.397	0.273	0.303	0.329
NB		0.265	0.317	0.497	0.146	0.156	0.24	0.18	0.188	0.21
RF		0.439	0.579	0.729	0.246	0.218	0.383	0.196	0.212	0.267
LR	0.1	0.603	0.743	0.862	0.333	0.665	0.687	0.337	0.343	0.337
DT		0.573	0.685	0.822	0.359	0.589	0.637	0.361	0.361	0.385
NB		0.369	0.457	0.679	0.174	0.411	0.459	0.194	0.196	0.212
RF		0.559	0.713	0.842	0.299	0.625	0.671	0.267	0.251	0.291
LR	0.3	0.868	0.944	0.974	0.778	0.906	0.888	0.741	0.792	0.852
DT		0.828	0.926	0.966	0.601	0.842	0.896	0.629	0.669	0.725
NB		0.735	0.868	0.93	0.519	0.824	0.792	0.395	0.447	0.523
RF		0.856	0.936	0.986	0.669	0.886	0.902	0.683	0.764	0.776

Sentence Level Backdoor Attack:

- As the poison rate increases, it can be stated that the accuracy of the ML models decreases.
- As the trigger sentence length increases, the success rate of backdoor attack increases.
- As the poison rate increases, the success rate of backdoor attack increases.

Word Level Backdoor Attack:

- As the poison rate increases, it can be stated that the accuracy of the ML models decreases.
- As the number of trigger word increases, the success rate of backdoor attack increases.
- As the poison rate increases, the success rate of backdoor attack increases.

Word Level Backdoor Attack with Defense:

- Defense algorithm slightly decreases the success rate of the word level backdoor attack with slight loss on the accuracy of the ML models.