Name: Ninaad Akella

Group: M

Reflective Essay

The main objective of our project was to train Machine Learning models with existing statistical data on tumors from the breast region with labels defining whether it is cancerous (Malignant) or non-cancerous (Benign) and find out which model is best for new data. We decided to go with UCI dataset from Wisconsin USA because this dataset was produced by a very reputable organization and we could be sure of the quality of the data, i.e, low human errors. In hindsight, a bigger dataset could have been used for this project.

We decided to use Z-score normalization over min-max normalization because it handles outliers better though all the features are not in the exact same scale. This was because datasets which consists mainly of statistical features (like area_mean) can consist of a lot of outliers as one large value can skew the data. We

We had some issues with a former member of our team Nidhiben. Any work that was assigned to her was not completed and eventually other members had to do it. Eventually we reported this issue to the faculty and on their advice, she was moved to another team. However due to this some delays were caused in our project. For coding part other members and me divided the work with me working on pre-processing of data as well as two models. Haomin was responsible for Data Analysis and two models and finally Boyu was responsible for research and two model. Both of them did their works well. As indicated by the faculty, our project was lacking in EDA especially Visual EDA. This is not any one person's fault but an issue because of all the members.

For tuning of the Hyperparameters of the various models used in our project (Decision Tree, K- Nearest Neighbours, Logistic regression, Naïve Bayes, SVM, Random Forest) we used Grid Search Cross Validation. It forms a grid of all the mentioned values of hyperparameters and trains the model with all combinations of these values for k folds each time and gives you back the best combination. This means for two hyperparameters a and b, a has 3 values, b has 5 values and we set k fold to 10, then the number of fits for this model will be 3*5*10=150 fits. This seems like the best way to tune a model (given our current level knowledge).

For scoring the models, we used primarily f1 (as we used it for tuning) along with jaccard-similarity score and ROC curve. From all these metrics, we found that K-Nearest Neighbours to be best and SVM to be second best overall. However, I don't think there is any significant statistical difference between their performance.

Looking back, we could have done a few things differently like acquiring a bigger dataset, performing a better EDA, trying some sort of dimensionality reduction techniques. In future, we want to build on this project and create a more sophisticated model than can be used directly on the image data.