**Assignment 3**

**INSERT YOUR NAME HERE**

**Data and Data Processing:**

In this section explain your pre-processing, feature engineering, feature selection, and any data-augmentation you performed. Also, show a visualization of the data.

You must perform some data preprocessing, and you must explain what you did. Below is a list of things to consider. You must do some of these, but you don’t need to do all of them. Use your best judgement.

* What features will you use?
* How will you deal with categorical features?
* Will you eliminate uninformative features? If so, how will you determine which feature are uninformative?
* Will you perform any binning/bucketing?
* Will you perform any normalization or standardization?
* What is the class balance? Will you perform any under-sampling or oversampling?

You must state the number of samples, how you encoded the labels, and the number of true samples and false samples.

You must visualize the data and insert a plot using PCA and another plot using TSNE. You must explain how much variability is explained in the PCA plot (per each dimension and combined). Your plots won’t have clear class boundaries. Explain what this tells you about the data.

sklearn.decomposition.PCA

sklearn.manifold.TSNE

**Results:**

You must fill in the following table

|  |  |  |
| --- | --- | --- |
| Model | Hyperparameters | Average 5-fold CV accuracy |
| Majority Class Baseline |  |  |
| KNN | n\_neighbors=5 distance=euclidean |  |
| Decision Tree |  |  |
| Logistic Regression |  |  |
| Polynomial Logistic Regression |  |  |
| Neural Network |  |  |
| Random Forest |  |  |
| AdaBoost |  |  |
| Linear SVM |  |  |
| RBF SVM |  |  |

Using the following implementations:

Majority Class Baseline = code yourself

K Nearest Neighbors (KNN) = sklearn.neighbors.KNeighborsClassifier

Decision Tree = sklearn.tree. DecisionTreeClassifier

Logistic Regression = sklearn.linear\_model.SGDClassifier

Logistic Polynomial Regression = sklearn.linear\_model.SGDClassifier with sklearn.preprocessing.PolynomialFeatures

Neural Network = sklearn.neural\_network.MLPClassifier

Random Forest = sklearn.ensemble.RandomForestClassifier

AdaBoost = sklearn.ensemble.AdaBoostClassifier

Linear SVM = sklearn.linear\_model.SGDClassifier

RBF SVM = sklearn.svm.SVC

List any hyperparameters and the values that you specified in the hyperparameters column. List one hyperparameter=value pair per line, using the ones I listed for KNN as a template

Warning: Polynomial Logistic regression can be extremely slow for high degree polynomials, and RBF SVM can also be slow

**Best Results:**

In this section simply state the best performing classifier, it’s performance, and the hyperparameters you used.