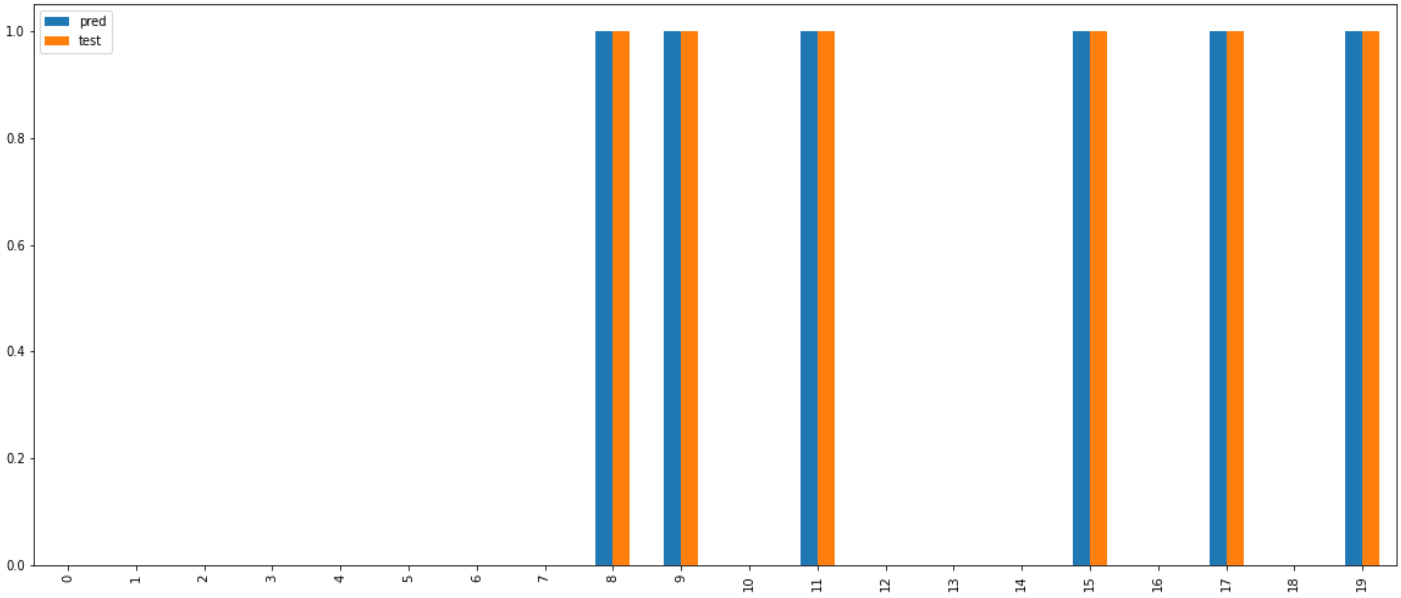
ML_02

MALWARE DETECTION

Team- Knight_2210

- **We have developed a model that can detect malware in a PE file after extracting it into a dataset.**
- **We have used a dataset found in github.**
- **The link for dataset : https://github.com/PacktPublishing/Mastering-Machine-Learning-for-Penetration-Testing/blob/master/Chapter03/MalwareData.csv.gz**
- **We used pandas to convert it to a dataframe.**
- **From the dataframe we don't need the features : Name and md5 because they are objects and they do not have any effect on the legitimacy of the PE file.**
- **Then we made 1 ndarray 'x' by taking all the columns other than 'legitimate' as they are all independent variables and the column legitimate is used for making dependent variable y.**
- **We used square root transformation for managing skewed data.**
- **We imported train_test_split and we split the data into training set and testing set to avoid overfitting of the model.**

- **We trained the model with both managing and without managing the skewed data using RandomForestClassifier. This is imported from ensemble learning module and this is so efficient because it has the collection of 100 decision trees and it gives perfect results.**
- **We found the accuracy for both with and without managing the skewed data.**
- **With managing - 99.53% accuracy**
- **Without managing - 99.52% accuracy**
- **As we can see the accuracy of the models are almost same. We can observe that Decision trees and 'RandomForestClassifer' can give good results even with skewed data**
- **We imported pickle and saved the model in the file 'Malware_classification.pkl' and that is uploaded in the github repository.**



This is the visual representation of comparison between predicted values and test values.

The Dataset and the trained pickle model are uploaded in the github repository I submitted through google form.

----------*----------**