**MEETING MINUTES: 27/11/2023**

**Milestone 2: Work done**

* Collecting dataset 100+ entries
* Basic arithemetics, sorting, searching, and dynamic programming
* Trained and tested a model
* prismtestcase.py, testp1.py
* First code uses random forest classifier 46%
* Size: 100+
* Function, function description, unit test case and labelling
* Second code will give higher accuracy 83%

**Feedback:**

* Features must be present
* Combinations must be available
* Good to have more no of features to give more accuracy
* More samples will give more accuracy
* More samples with more features: problem is execution time
* Filters should be used to remove redundancy
* Last column is the result
* Dataset creation is manual
* They asked us to classify whether it is positive and negative
* Features must be available
* Feature analysis is required
* For any dataset, equation based, text based or anything, we can make use of the features to get the accuracy
* Random forest gives 90% above accuracy
* Classification as positive and negative doesn't depend upon imbalance of +ve and -ve
* Whether the test case is rightly classified as +ve and -ve is what matters
* Make use of standard benchmark dataset (700-1000 entries)
* First note the execution time to run the dataset
* Time doesn't matter now, the major focus is learning
* a+b-c\*d/e
* a,b, c,d,e is features
* Some features are redundant. We come to this stage only at advanced stage
* Understand each feature and how it contributes to the result
* You can add the same dataset to the ppt
* Clarify the doublt with samsung mentors
* Our job now is not to tell which algo is performing well
* Some datasets will work better on some algo. Algo depends upon the type of datset
* If an algo can only understand btwn black and white, other colours will not give the right output. In this case, random result is chosen
* To enrich the dataset, add more specifications
* Features are correlated functionally
* Find the relationship
* Let there be 2 sets of datasets and two algo. Make a tabular column for the execution
* The next ppt is to make this classification
* Generally, only 70% is used for training. Entire dataset is not used for testing
* I show 10 different types of pen. You must be abler to identify if it comes under the category of pen or if it is a pencil
* If more sample is used for training and less is used for testing, obviously, the accuracy will be more
* Task: Look into the features