DATA MINING PROJECT

TITLE: SANITY TEST

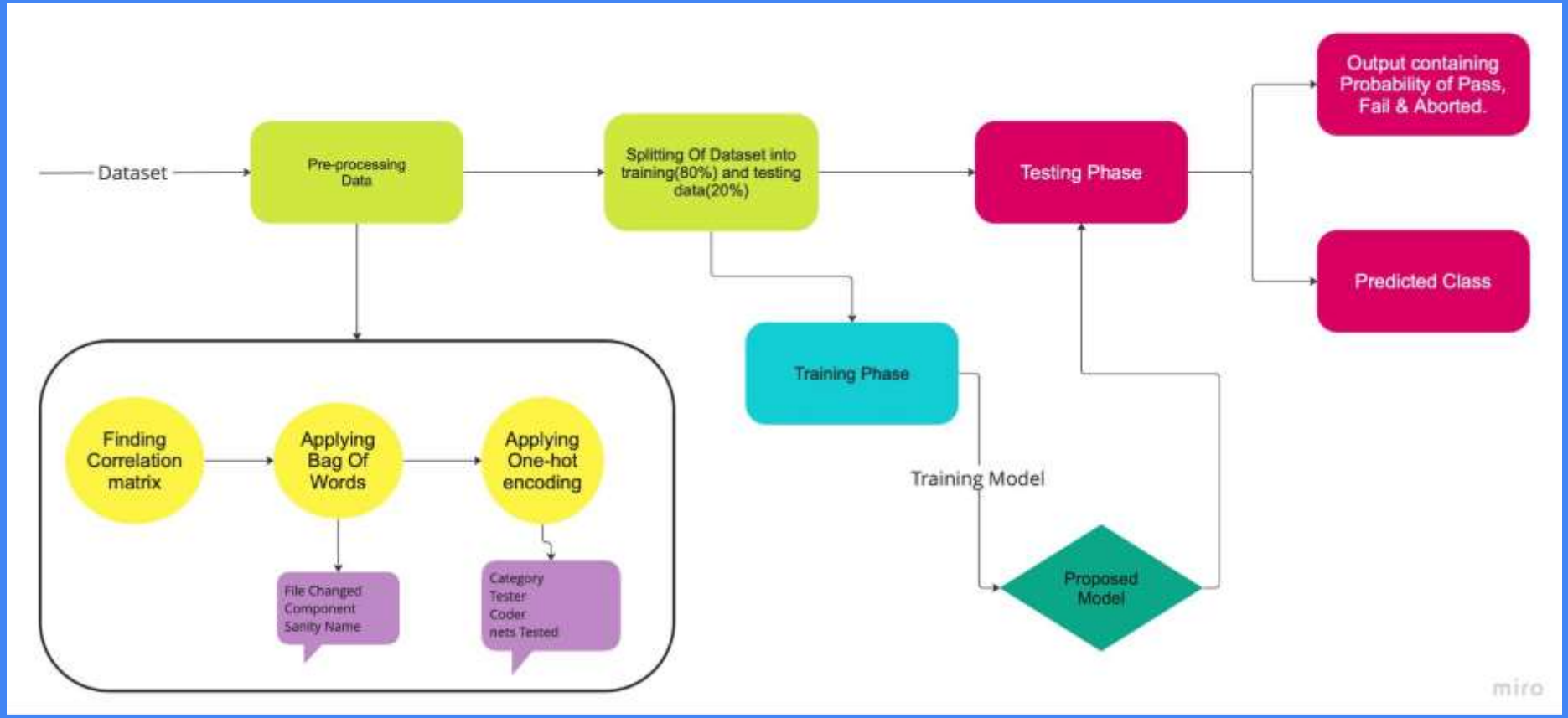
Made By: Nidhi Dixit

Roll No.-21ec3012

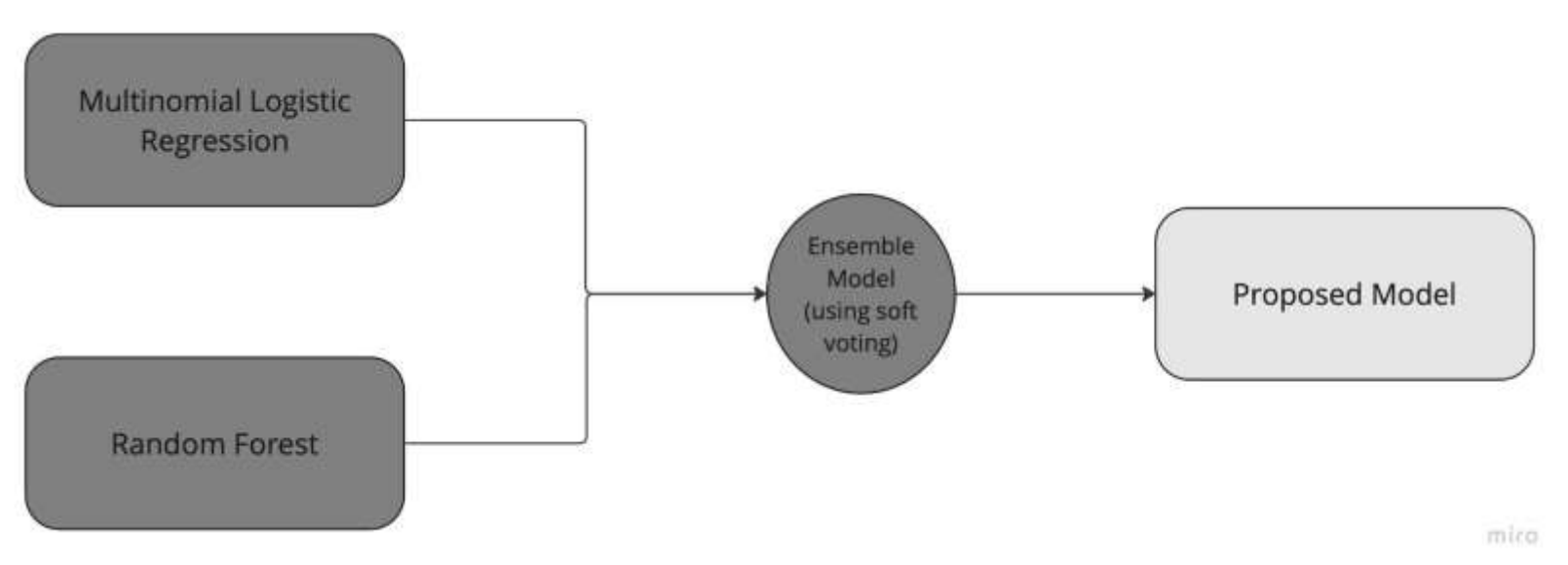
PROBLEM STATEMENT

A dataset containing various attributes that impact the result of the sanity test was given and when a particular file in the given directory changes and is committed to GitHub. A classification model was built to predict if a sanity test case will fail on the platform for a given combination of attributes when a particular file is changed. The model is enhanced to identify the scenarios which have a very low probability of failure for the change in a particular set of files.

Flowchart


Dataset
Pre-processing Data
Splitting Of Dataset into training(80%) and testing data(20%)
Testing Phase
Output containing Probability of Pass, Fail & Aborted.
Predicted Class
Training Phase
Training Model
Proposed Model
Finding Correlation matrix
Applying Bag Of Words
Applying One-hot encoding
File Changed
Component
Sanity Name
Category
Tester
Coder
nets Tested
miro

Proposed Model



Analysis done on the problem statement

- This was a multiclass classification problem.
- I've found the correlation matrix in this problem statement and concluded that output was dependent on all the attributes.
- Some attributes in the dataset were given as english texts so, in order to train the machine learning model I vectorized all of the attributes. Here I used NLP for the process of vectorization.

var2 var1	Category	Coder	Component	File Changed	Sanity Result	Sanity name	Tester	nets Tested
Category	NaN	7.804164e-42	4.826306e-183	2.708323e-120	7.636152e-14	1.330431e-301	2.032295e-184	2.928507e-282
Coder	7.804164e-42	NaN	2.517941e-217	2.342799e-52	1.064056e-07	3.929845e-118	3.710492e-184	5.548302e-65
Component	4.826306e-183	2.517941e-217	NaN	0.000000e+00	1.252796e-26	0.000000e+00	0.000000e+00	5.347049e-199
File Changed	2.708323e-120	2.342799e-52	0.000000e+00	NaN	5.131404e-20	2.964803e-285	1.303508e-175	1.126545e-115
Sanity Result	7.636152e-14	1.064056e-07	1.252796e-26	5.131404e-20	NaN	4.973096e-14	3.724433e-33	1.798112e-10
Sanity name	1.330431e-301	3.929845e-118	0.000000e+00	2.964803e-285	4.973096e-14	NaN	0.000000e+00	2.104017e-298
Tester	2.032295e-184	3.710492e-184	0.000000e+00	1.303508e-175	3.724433e-33	0.000000e+00	NaN	1.023524e-164
nets Tested	2.928507e-282	5.548302e-65	5.347049e-199	1.126545e-115	1.798112e-10	2.104017e-298	1.023524e-164	NaN

Analysis done on the problem statement (Conti…)

- After researching deeply I choose to apply “Bag Of Words” on the attributes whose count of unique value (categories) > 20.

  We applies “Bag of words” on three attributes which were:

a) File changed
b) Component
c) Sanity name

- For attributes whose count of unique value (categories) < 20, we applies “One Hot Encoding”.

	File Changed	Component	Coder	Tester	Category	Sanity name	nets Tested	Sanity Result
count	351	351	351	337	351	351	351	351
unique	176	45	7	18	11	25	14	3
top	cpp/csyu/category/fnf	asr1001-forge-cgn	patrick	manishakang	PD-asr1001	vistra-cxr	asr1001-PX,asr1001-X64	Pass
freq	2	70	130	109	146	122	105	210

Analysis done on the problem statement (Conti…)

- After preprocessing the data, finally the dataset looked like this

	1	2	3	4	5	6	7	8	9	10	...	193	194	195	196	197	Coder	Tester	Category	nets Tested	Sanity Result
0	0	0	0	0	0	0	0	0	0	0	...	1	0	0	0	0	1	12	4	1	2
1	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	7	9	1	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	7	4	1	1
3	0	1	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	7	4	1	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	7	4	1	0

5 rows x 202 columns

Model Building

After applying different classification models like XGBoost, Naive Bayes, Decision Tree, Multinomial logistic regression, the accuracy of our proposed model is listed below.

Models Applied	Accuracy
XGBoost	74.6478
Naive Bayes	80.2816
Decision Tree	91.5492
Multinomial Logistic Regression	87.3239
Random Forest	94.3661
Proposed Model	97.1830

Output

	1	2	3	4	5	6	7	8	9	10	...	196	197	Coder	Tester	Category	nets Tested	Prob_Aborted	Prob_Fail	Prob_Pass	Predicted
157	0	0	0	0	0	0	0	0	0	0	...	0	0	6	7	4	10	0.68	0.05	0.27	Aborted
342	0	0	0	0	0	0	0	0	0	0	...	0	0	0	7	4	10	0.17	0.29	0.54	Pass
316	0	0	0	0	0	0	0	0	0	0	...	0	0	6	7	4	10	0.04	0.54	0.42	Fail
234	0	0	0	0	0	0	0	0	0	0	...	0	0	3	1	6	2	0.01	0.04	0.95	Pass
155	0	0	0	0	0	0	0	0	0	0	...	0	0	6	7	4	12	0.07	0.25	0.68	Pass
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
181	0	0	0	0	0	0	0	0	0	0	...	0	0	6	7	10	1	0.39	0.06	0.55	Aborted
179	0	0	0	0	0	0	0	0	0	0	...	0	0	6	7	4	11	0.50	0.17	0.33	Aborted
199	0	0	0	0	0	0	0	0	0	0	...	1	0	2	6	2	10	0.03	0.11	0.86	Pass
327	0	0	0	0	0	0	0	0	0	0	...	0	0	6	7	4	10	0.03	0.02	0.95	Pass
228	0	0	0	0	0	0	0	0	0	0	...	0	0	2	13	6	2	0.00	0.00	1.00	Pass

71 rows × 205 columns

Thanks!