CLASSIFYING POKER HANDS: NOT A GAMBLER'S TASK

JULIE LENZER

MSML651 – FINAL PROJECT

AGENDA

- Project Overview
- Data Representation and Challenges
- Model Evaluation and Selection
- Model Training: Spark MLlib process
- Performance Metrics
- Conclusions

PROJECT OVERVIEW

- Given a draw of 5 cards from a standard deck of 52 cards, what is the associated poker hand?
- Supervised multivariate machine learning (*each card in hand represents a feature*)
- Data from https://archive.ics.uci.edu/ml/datasets/Poker+Hand
- Training data: 25,010 possible hands, testing data: 1,000,000 possible hands
- Evaluated several models to see which performed *best*
- Used Spark Mllib for training and prediction with selected model

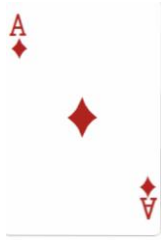
DATA REPRESENTATION – ALL CATEGORICAL

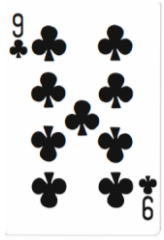
Suits

Hearts=2, Spades=2, Diamonds=3, Clubs=4

Rank

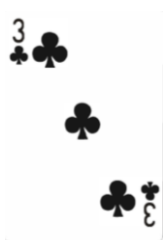
12 

1 

9 

7 

5 

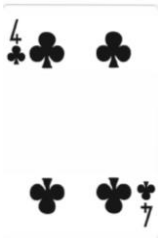
3 

11 

10 

8 

6 

4 

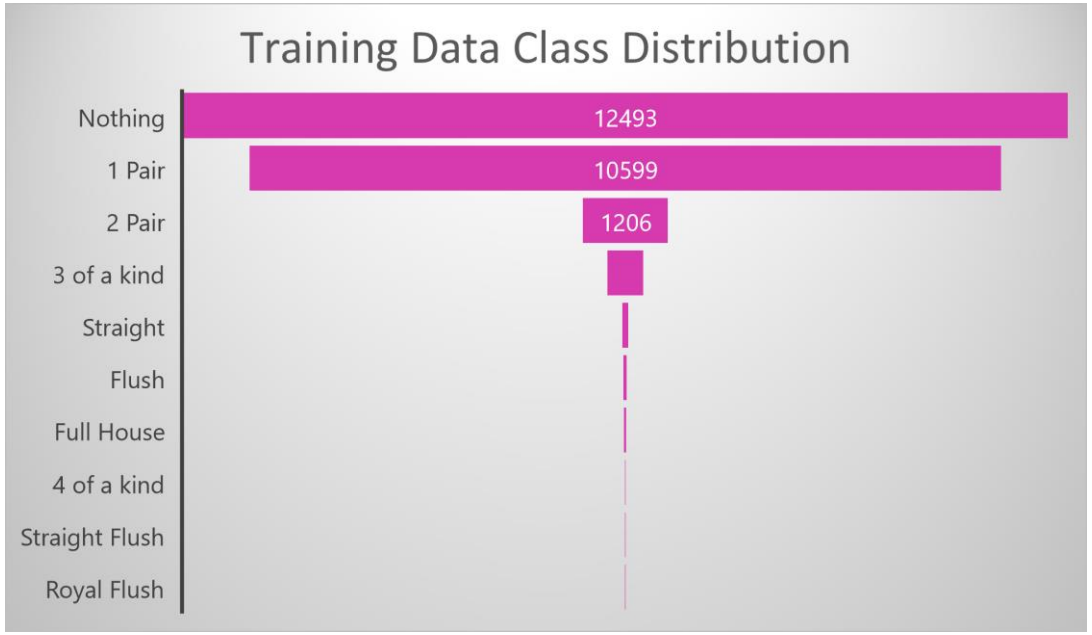
2 

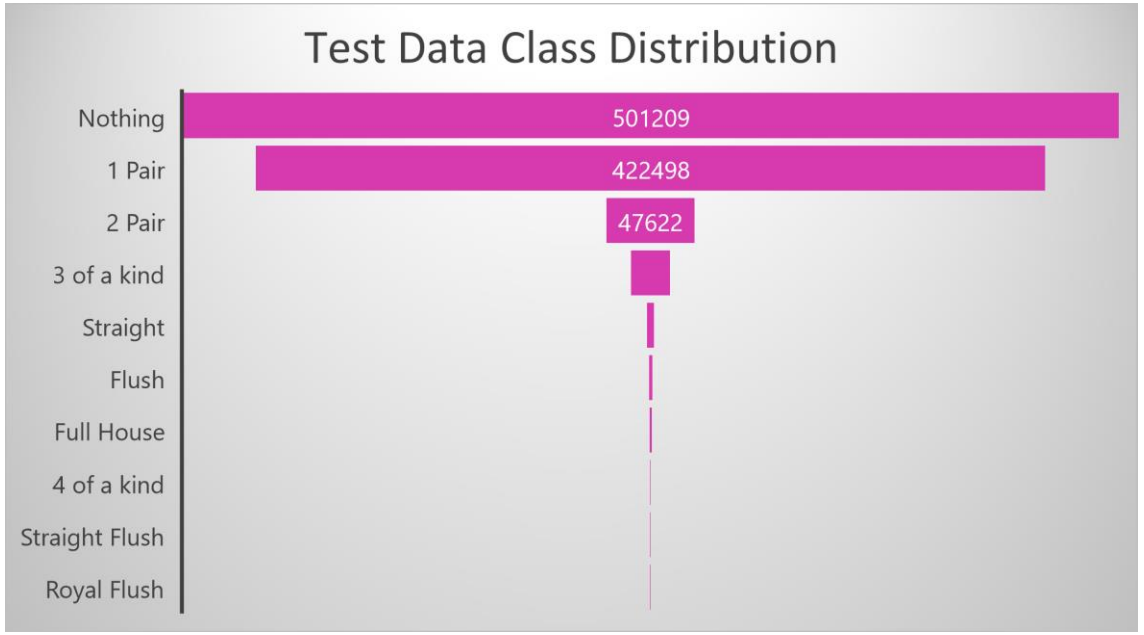
13

POSSIBLE POKER HANDS – CLASS TO PREDICT

Label	Hand	Description
0	Nothing in the hand	Not a recognized poker hand
1	One pair	One pair of equal rank (different suit)
2	Two pairs	Two pairs of equal rank
3	Three of a kind	Three equal ranks
4	Straight	Five cards sequentially ranked with no gaps
5	Flush	Five cards of the same suit
6	Full House	One pair and three of a kind in a different rank
7	Four of a kind	Four equal ranks
8	Straight flush	Straight in same suit (flush)
9	Royal flush	Ace, king, queen, jack and ten in same suit

DATA CHALLENGE: HUGE IMBALANCE

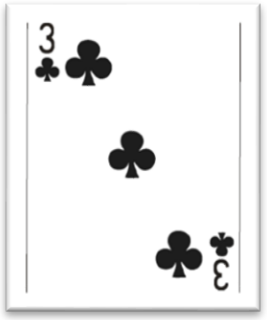




Intended to mirror real probability distribution

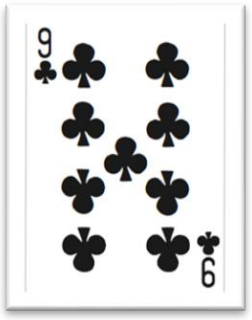
DATA CHALLENGE: REPRESENTATION OF SEQUENCE

Class: 1
One Pair

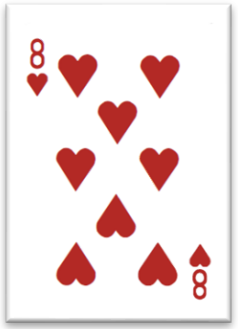


S1 = 4
C1 = 3

S2 = 1
C2 = 13



S3 = 4
C3 = 9



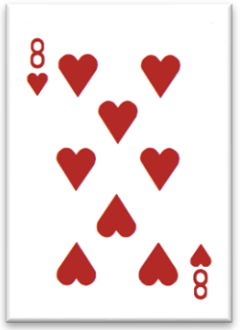
S4 = 1
C4 = 8

S5 = 4
C5 = 13

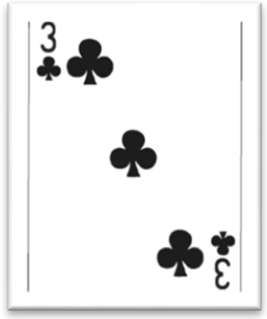
Class: 1
One Pair

S1 = 1
C1 = 13

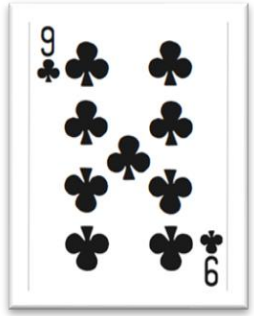
S2 = 4
C2 = 13



S3 = 1
C3 = 8

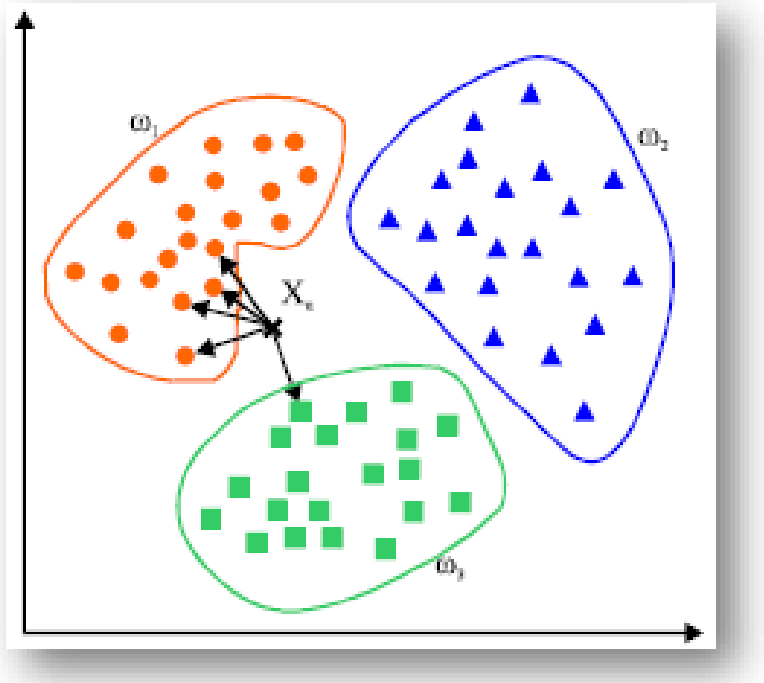


S4 = 4
C4 = 3



S5 = 4
C5 = 9

MODELS CONSIDERED



K-Nearest Neighbors

https://www.mathworks.com/matlabcentral/fileexchange/63621-knn-classifier

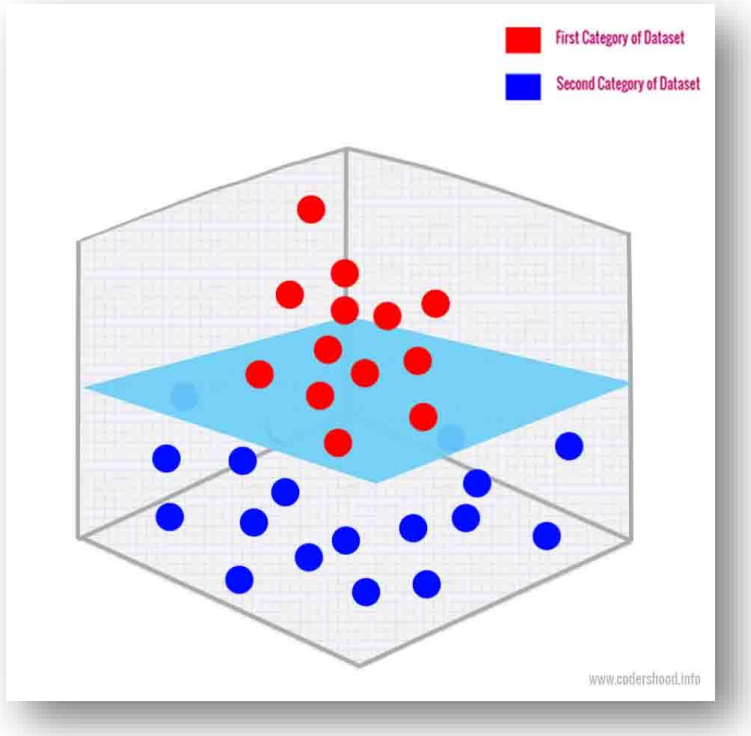
Accuracy Score of KNeighborsClassifier : 0.532853525256564

Predicted Result	0	1	2	3	4
Actual Result					
0	2585	1181	3	0	1
1	1756	1408	12	7	2
2	148	197	4	3	0
3	39	95	3	1	0
4	9	16	0	1	0
5	6	4	0	0	0
6	6	12	0	0	0
7	0	1	0	0	0
8	0	2	0	0	0
9	0	1	0	0	0

	precision	recall	f1-score	support
0	0.57	0.69	0.62	3770
1	0.48	0.44	0.46	3185
2	0.18	0.01	0.02	352
3	0.08	0.01	0.01	138
4	0.00	0.00	0.00	26
5	0.00	0.00	0.00	10
6	0.00	0.00	0.00	18
7	0.00	0.00	0.00	1
8	0.00	0.00	0.00	2
9	0.00	0.00	0.00	1
accuracy			0.53	7503
macro avg	0.13	0.11	0.11	7503
weighted avg	0.50	0.53	0.51	7503

F1 Score: 0.511597732805661

MODELS CONSIDERED



Support Vector Machine

https://www.codershood.info/2019/01/10/support-vector-machine-machine-learning-algorithm-with-example-and-code/support-vector-machine-machine-learning-algorithm-with-example-and-code-higher-dimension/

Accuracy Score of SVM : 0.5024656803945089

Predicted Result 0

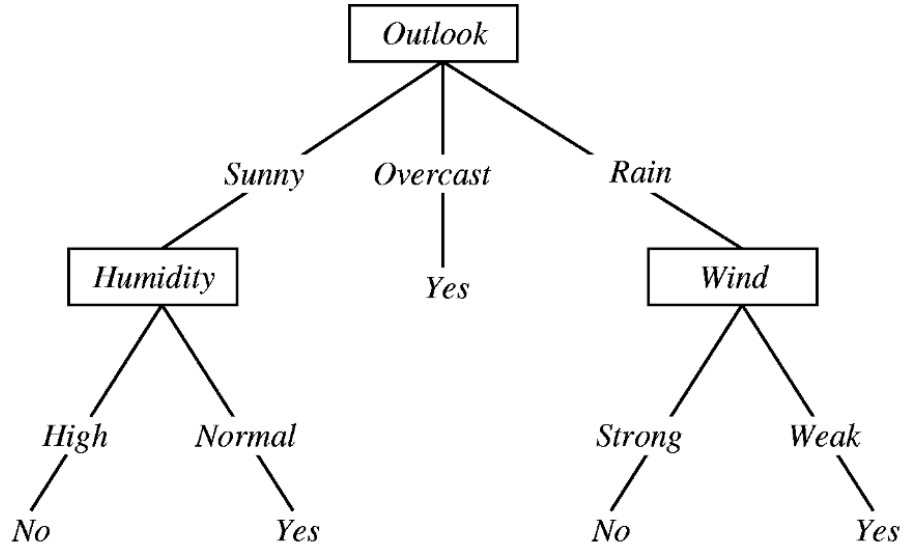
Actual Result

0	3770
1	3185
2	352
3	138
4	26
5	10
6	18
7	1
8	2
9	1

	precision	recall	f1-score	support
0	0.50	1.00	0.67	3770
1	0.00	0.00	0.00	3185
2	0.00	0.00	0.00	352
3	0.00	0.00	0.00	138
4	0.00	0.00	0.00	26
5	0.00	0.00	0.00	10
6	0.00	0.00	0.00	18
7	0.00	0.00	0.00	1
8	0.00	0.00	0.00	2
9	0.00	0.00	0.00	1
accuracy			0.50	7503
macro avg	0.05	0.10	0.07	7503
weighted avg	0.25	0.50	0.34	7503

F1 Score: 0.668854785771312

MODELS CONSIDERED



Decision Tree

https://towardsdatascience.com/decision-tree-in-machine-learning-e380942a4c96

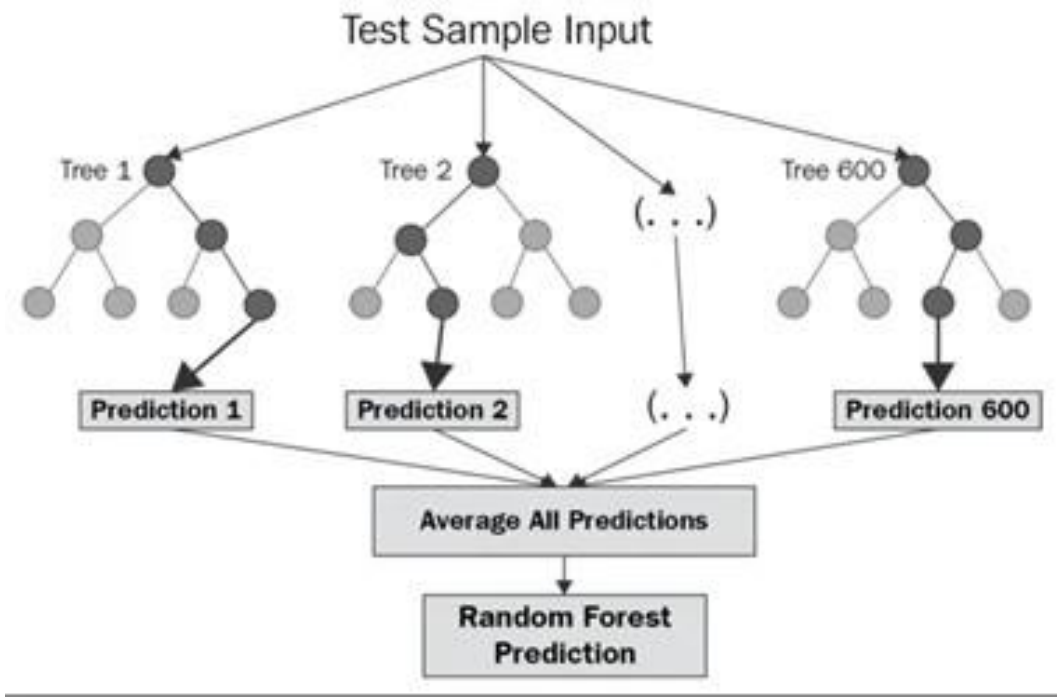
Accuracy Score of DecisionTree : 0.4571504731440757

Predicted Result	0	1	2	3	4	5	6	7	8	9
Actual Result										
0	2015	1482	171	70	18	9	5	0	0	0
1	1426	1366	253	104	17	8	7	2	0	2
2	140	157	34	15	1	2	2	1	0	0
3	49	60	13	13	3	0	0	0	0	0
4	10	12	2	1	1	0	0	0	0	0
5	5	3	0	0	0	1	0	0	1	0
6	3	11	2	2	0	0	0	0	0	0
7	0	1	0	0	0	0	0	0	0	0
8	1	0	0	0	1	0	0	0	0	0
9	0	1	0	0	0	0	0	0	0	0

	precision	recall	f1-score	support
0	0.55	0.53	0.54	3770
1	0.44	0.43	0.44	3185
2	0.07	0.10	0.08	352
3	0.06	0.09	0.08	138
4	0.02	0.04	0.03	26
5	0.05	0.10	0.07	10
6	0.00	0.00	0.00	18
7	0.00	0.00	0.00	1
8	0.00	0.00	0.00	2
9	0.00	0.00	0.00	1
accuracy			0.46	7503
macro avg	0.12	0.13	0.12	7503
weighted avg	0.47	0.46	0.46	7503

F1 Score: 0.4631118258162234

MODELS CONSIDERED



Random Forest

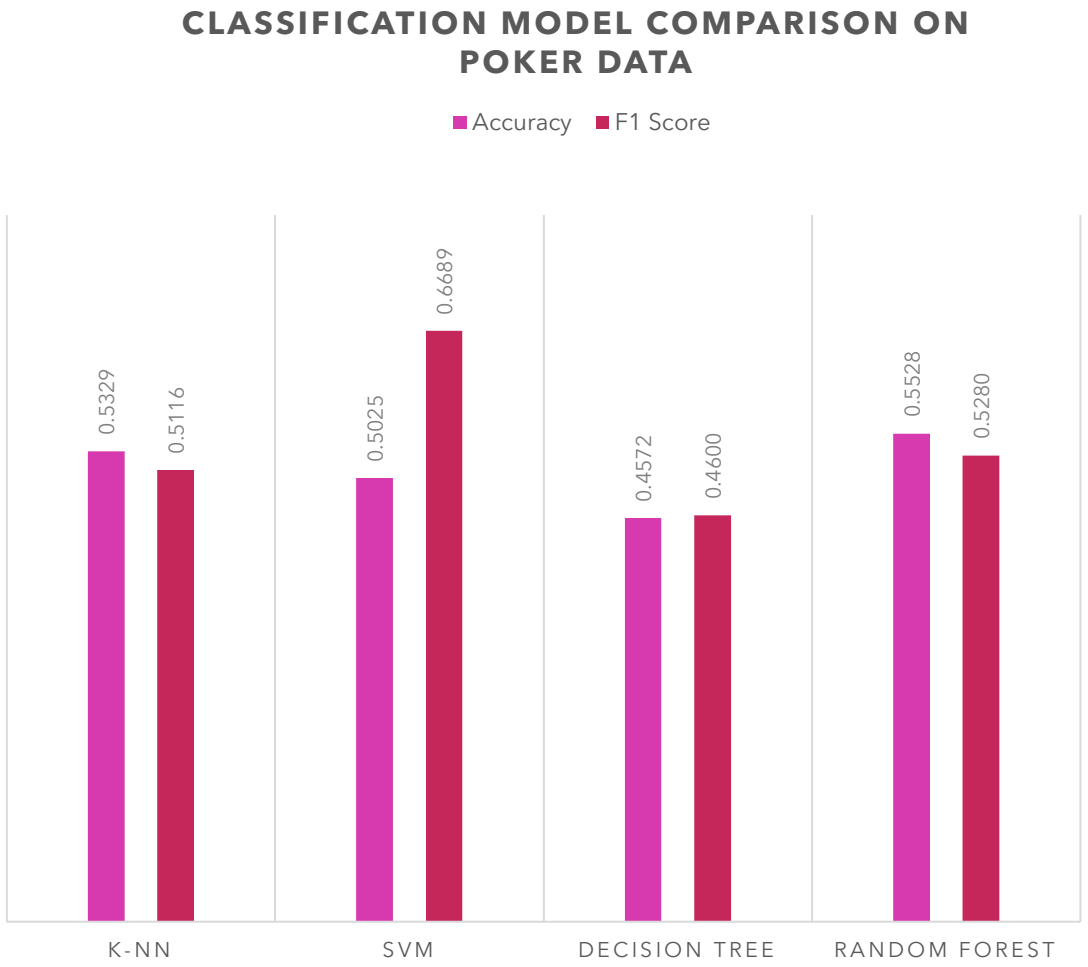
Accuracy Score of RandomForest : 0.5528455284552846

Predicted Result	0	1	2	3	5
Actual Result					
0	2757	1012	1	0	0
1	1786	1390	7	1	1
2	130	220	1	1	0
3	36	102	0	0	0
4	12	14	0	0	0
5	9	1	0	0	0
6	1	17	0	0	0
7	0	1	0	0	0
8	0	2	0	0	0
9	0	1	0	0	0

	precision	recall	f1-score	support
0	0.58	0.73	0.65	3770
1	0.50	0.44	0.47	3185
2	0.11	0.00	0.01	352
3	0.00	0.00	0.00	138
4	0.00	0.00	0.00	26
5	0.00	0.00	0.00	10
6	0.00	0.00	0.00	18
7	0.00	0.00	0.00	1
8	0.00	0.00	0.00	2
9	0.00	0.00	0.00	1
accuracy			0.55	7503
macro avg	0.12	0.12	0.11	7503
weighted avg	0.51	0.55	0.52	7503

F1 Score: 0.52805537847242

WHICH MODEL TO CHOOSE?



THE PROBLEM WITH SVM

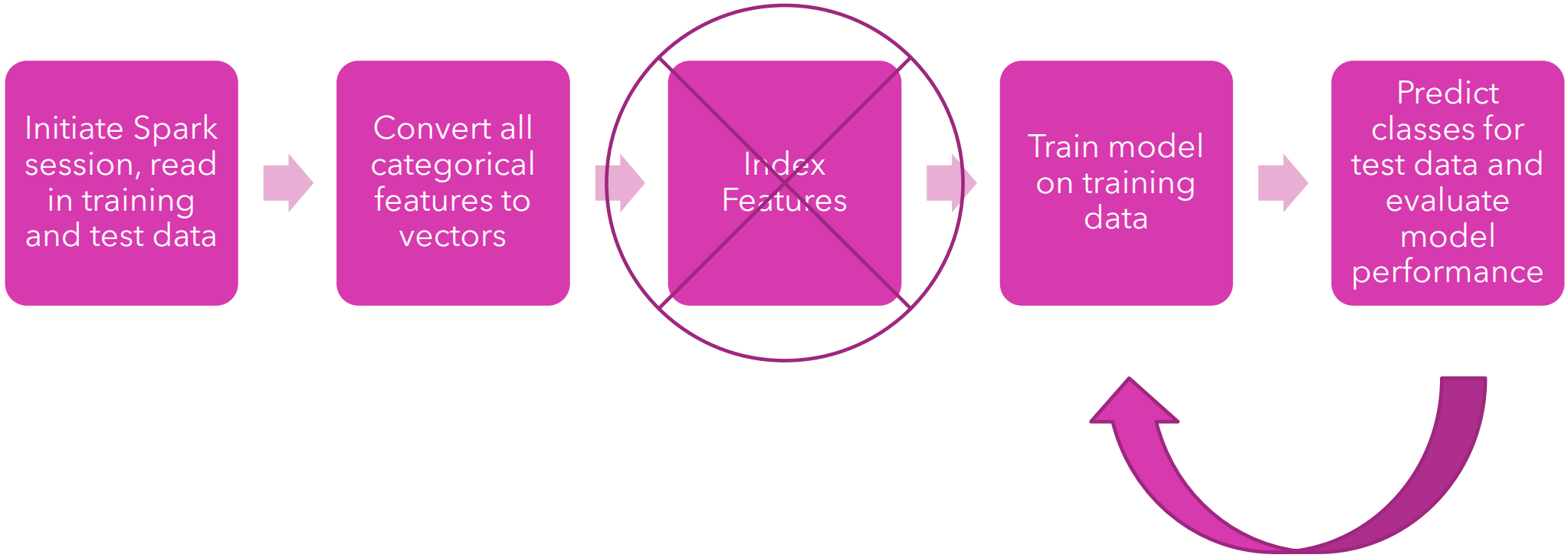
Accuracy Score of SVM : 0.5024656803945089

Predicted Result	0
Actual Result	
0	3770
1	3185
2	352
3	138
4	26
5	10
6	18
7	1
8	2
9	1

	precision	recall	f1-score	support
0	0.50	1.00	0.67	3770
1	0.00	0.00	0.00	3185
2	0.00	0.00	0.00	352
3	0.00	0.00	0.00	138
4	0.00	0.00	0.00	26
5	0.00	0.00	0.00	10
6	0.00	0.00	0.00	18
7	0.00	0.00	0.00	1
8	0.00	0.00	0.00	2
9	0.00	0.00	0.00	1
accuracy			0.50	7503
macro avg	0.05	0.10	0.07	7503
weighted avg	0.25	0.50	0.34	7503

F1 Score: 0.668854785771312

RANDOM FOREST MODEL TRAINING: PROCESS WITH SPARK MLLIB



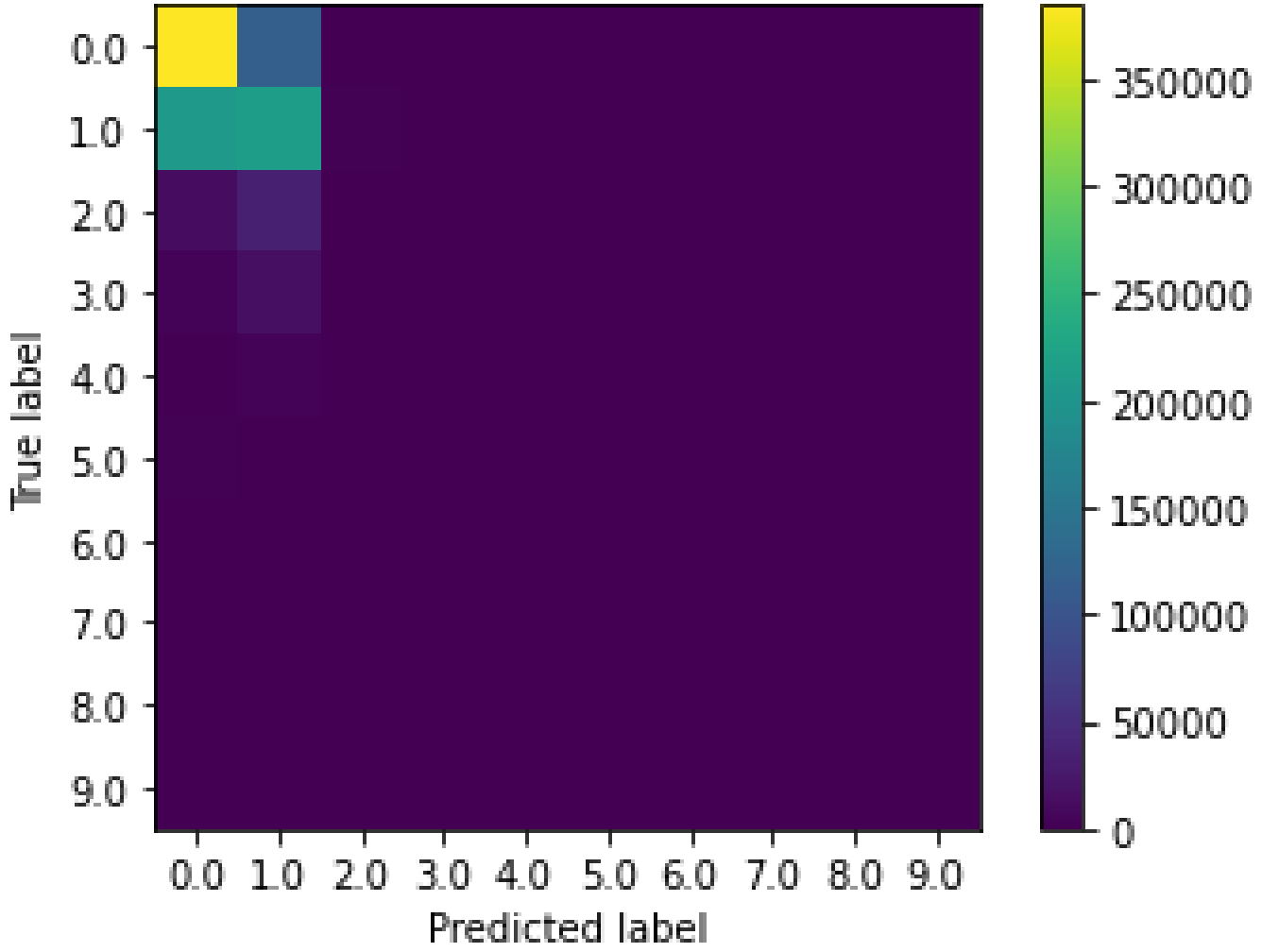
RANDOM FOREST PERFORMANCE METRICS

Random Forest – Best Performance	
Accuracy	0.599915
F1 Score	0.57336
Number of Classes Predicted	9

CONFUSION MATRIX – RAW DATA

Predicted Result / Actual Result	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0
0	384449	116599	129	17	7	7	1	0	0	0
1	206508	213963	1737	251	34	1	2	0	2	0
2	12840	33520	1131	121	4	0	6	0	0	0
3	4365	16007	393	349	3	0	1	3	0	0
4	606	3223	41	3	11	0	0	0	0	1
5	1645	342	0	0	0	9	0	0	0	0
6	97	1188	113	22	1	0	3	0	0	0
7	8	183	24	15	0	0	0	0	0	0
8	3	7	0	0	2	0	0	0	0	0
9	0	3	0	0	0	0	0	0	0	0

CONFUSION MATRIX



CONCLUSIONS

- Tie to sequence in data representation added needless complexity
- Imbalanced distribution of classes (class 0 and 1 = 92% of classes) made classification difficult. No model correctly predicted classes 7 – 9
- Indexing didn't add to performance
- Because of the class imbalance, F1 score more telling than accuracy
- Data may be better used for a different task or with additional transformations in the data
- Don't bet on this model!