Troll or Not?

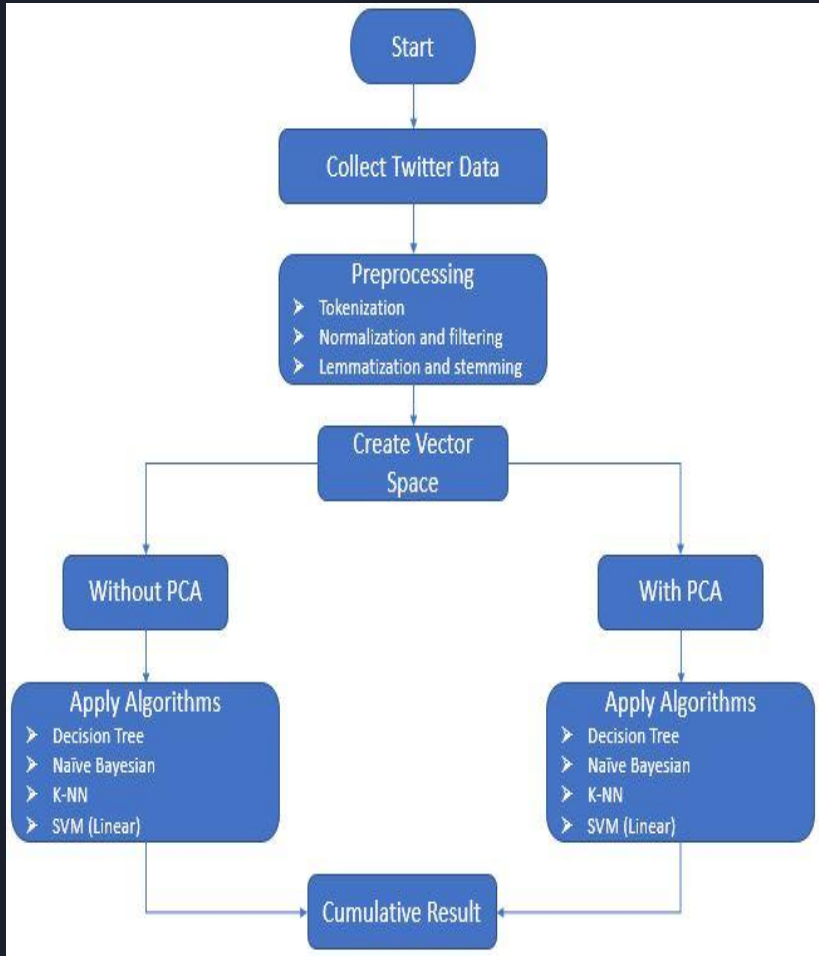
Tweet Classifier Study

MOTIVATION

- Twitter analytics has gained popularity in recent days.
- Political issues arose from fake Twitter accounts.
- Fake news is a prevalent issue nowadays in all forms of social media.
- Text based dataset
  - High feature number
  - Ideal classifier
  - Accuracy

OVERALL PROJECT FLOW

- We propose a system that analyzes the tweets extracted from twitter to classify a tweet as a troll tweet or not.
- After collecting the data, we perform several preprocessing steps in order to clean the data and convert it as a viable data set that can be fed to the machine learning algorithms. The various steps in preprocessing are,
  - Tokenization
  - Normalization and Filtering
  - Lemmatization and Stemming
- The cleaned dataset is then fed to the algorithms and compare the result of each algorithm with PCA and without PCA.
- The results of every algorithm is then accumulated and the accuracy of each algorithm is tested to see which algorithm yields the best accuracy.



Source: *Kaggle.com*

DATASET

Real democrats and republicans dataset

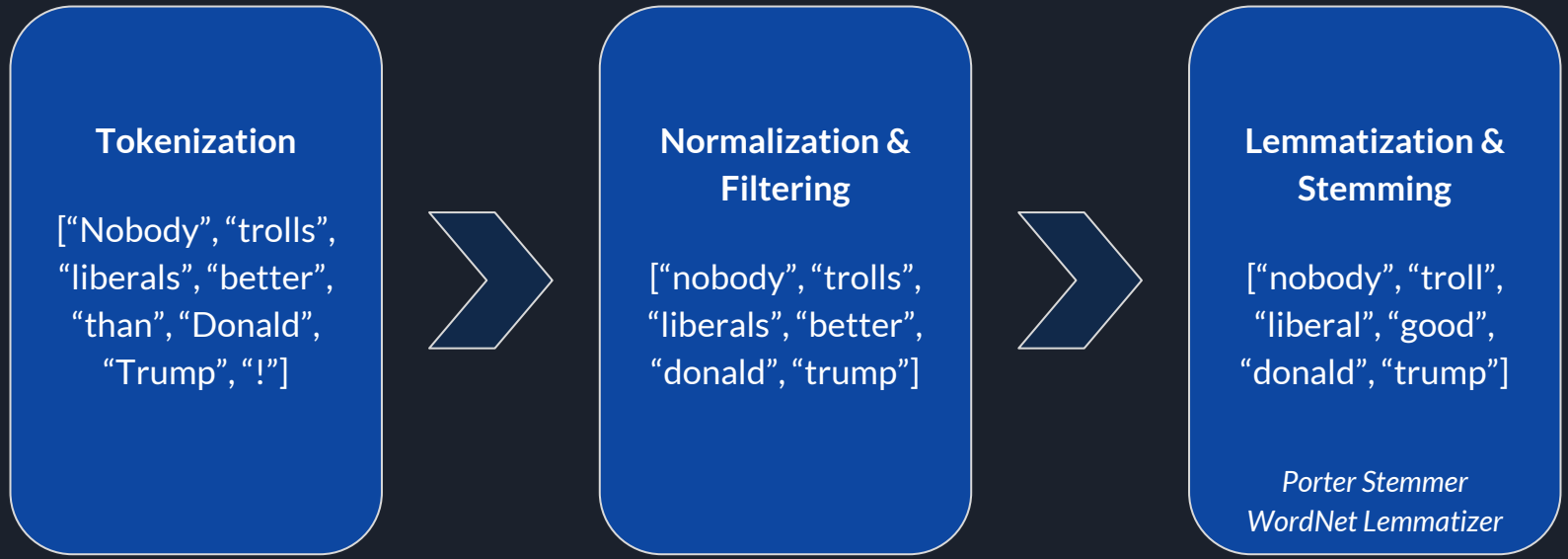
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Democrat	Today, Senate Dems vote to #SaveTheInternet. Proud to support similar #NetNeutrality legislation here in the Houseâ€¦ https://t.co/n3tggDLU1L														
2	Democrat	RT @WinterHavenSun: Winter Haven resident / Alta Vista teacher is one of several recognized by @RepDarrenSoto for National Teacher Appreciaâ€¦														
3	Democrat	RT @NALCABPolicy: Meeting with @RepDarrenSoto . Thanks for taking the time to meet with @LatinoLeader ED Marucci Guzman. #NALCABPolicy2018.â€¦														
4	Democrat	RT @Vegalteno: Hurricane season starts on June 1st; Puerto Ricoâ€™s readiness...well ðŸ¤¦ðŸ¼â€â™‚ï¸ðŸ˜¡@Pwr4PuertoRico @RepDarrenSoto @EspaillatNY														
5	Democrat	RT @EmgageActionFL: Thank you to all who came out to our Orlando gala! It was a successful night that would not have been possible withoutâ€¦														
6	Democrat	Hurricane Maria left approx $90 billion in damages, yet only $1 billion was allocated for rebuilding grid. No surprâ€¦ https://t.co/2kU8BcKwUh														
7	Democrat	RT @Tharryry: I am delighted that @RepDarrenSoto will be voting for the CRA to overrule the FCC and save our #NetNeutrality rules. Find outâ€¦														
8	Democrat	RT @HispanicCaucus: Trump's anti-immigrant policies are hurting small businesses across the country that canâ€™t find Americans willing to doâ€¦														
9	Democrat	RT @RepStephMurphy: Great joining @WeAreUnidosUS and @RepDarrenSoto for a roundtable in #Orlando on federal issues affecting central Floridâ€¦														

Fake democrats and republicans dataset

	A	B	C	D
1	"We have a sitting Democrat US Senator on trial for corruption and you've barely heard a peep from the mainstream medi	RightTroll		
2	Marshawn Lynch arrives to game in anti-Trump shirt. Judging by his sagging pants the shirt should say Lynch vs. belt https:	RightTroll		
3	Daughter of fallen Navy Sailor delivers powerful monologue on anthem protests, burns her NFL packers gear. #BoycottNFL	RightTroll		
4	JUST IN: President Trump dedicates Presidents Cup golf tournament trophy to the people of Florida, Texas and Puerto Rico	RightTroll		
5	19,000 RESPECTING our National Anthem! #StandForOurAnthemðŸ‡ºðŸ‡¸ https://t.co/czutyGaMQV	RightTroll		
6	Dan Bongino: "Nobody trolls liberals better than Donald Trump." Exactly! https://t.co/AigV93aC8J	RightTroll		
7	'@SenatorMenendez @CarmenYulinCruz Doesn't matter that CNN doesn't report on your crimes. This won't change the fac	RightTroll		
8	As much as I hate promoting CNN article, here they are admitting EVERYTHING Trump said about PR relief two days ago. h	RightTroll		

PRE-PROCESSING

Original: “Nobody trolls liberals better than Donald Trump!”



Python library: *NLTK*

VECTOR SPACE

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	senat	dem	vote	#savethein	proud	support	similar	#netneutr	legisl	winterhave	winter	haven	resid	alta	vista	teacher	one	sever	recogn
2	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	2	1	1	1
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Why PCA?

PCA is a dimensionality reduction technique used for denoising the dataset and include those features which retain the maximal information.

Motivation behind using PCA

- Too many features in the dataset (8000).
- Most of the features are really sparse.
- To measure the effect of PCA on the algorithms we have used.

Choosing the number of components.

- We used 1600 numeric components / features because it retained almost close to 65-70 % of the information.

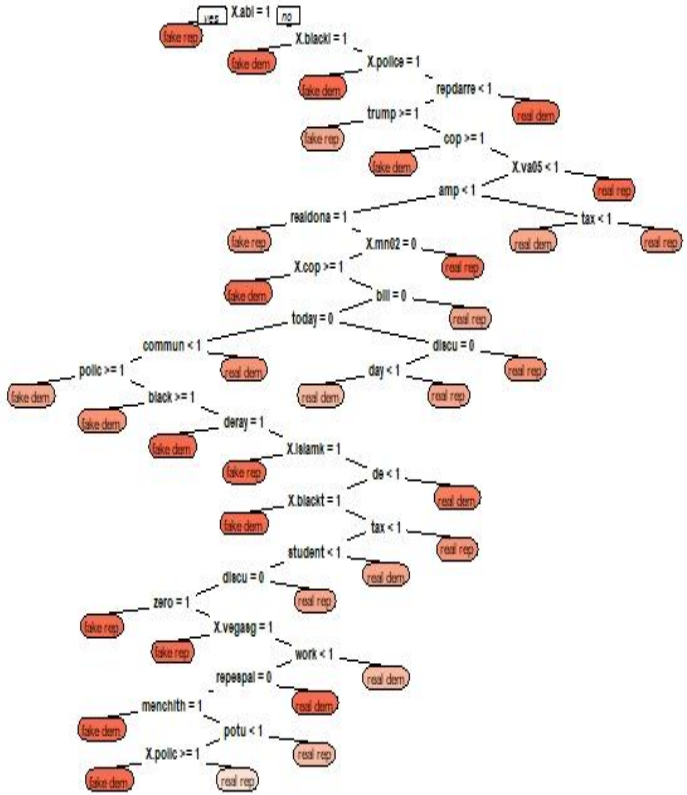
DECISION TREE

- Decision tree learning uses a decision tree to go from observations about an item to conclusions about the item's target value. It is one of the predictive modelling approaches used in statistics, data mining and machine learning.
- The goal is to create a model that predicts the value of a target variable based on several input variables. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf.
- A tree can be "learned" by splitting the source into subsets based on an attribute value test. This process is repeated on each derived subset in a recursive manner called recursive partitioning.
- The entire dataset is split into two, making 70% of the data as training dataset and the remaining 30% as testing dataset.
- In the case of our proposed system, we have a dataset that consists of a result column that has 4 classification variables based on which the the algorithm classifies the entire data and constructs the decision tree.

RESULT OF DECISION TREE

- We feed the data set to the algorithm once with PCA and another time without PCA.
- Here we can see that the decision tree algorithm has learnt the data set and it has plotted a decision tree.
- Based on the confusion matrix obtained after the prediction, we get an accuracy of 66.4% without PCA and an accuracy of 55.64% with PCA.

Confusion Matrix	Fake Democrat	Fake Republican	Real Democrat	Real Republican
Fake Democrat	221	5	22	30
Fake Republican	52	188	30	28
Real Democrat	34	49	129	115
Real Republican	58	32	208	135



NAIVE BAYESIAN

- This classifier works on conditional probability. Conditional probability is the probability that something will happen, given that something else has already occured. Using this we could calculate the probability of an event using prior knowledge.
- We use 70-30 split for training and testing respectively. We also check whether application of PCA has any effect on the classifier. For this experiment we use the gaussian Naive Bayes for classification.

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

RESULT OF NAIVE BAYES CLASSIFIER

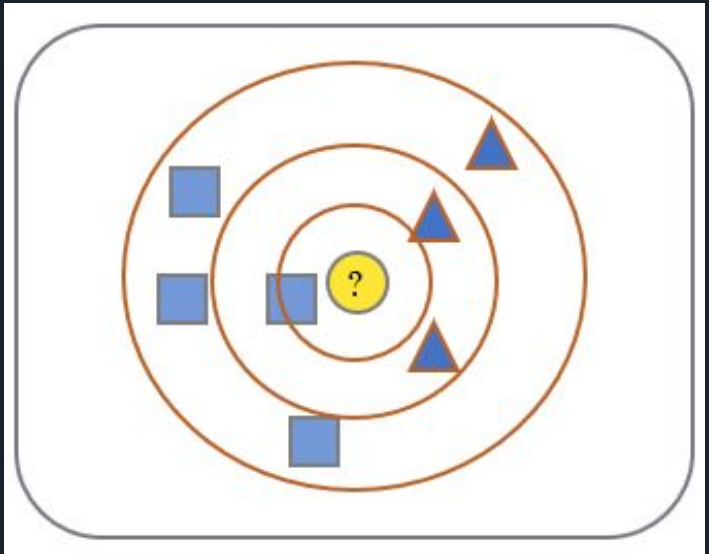
Without PCA: The accuracy of the classifier is 75%.

With PCA: When PCA is applied the accuracy is almost 44% which is very low.

Why? The main disadvantage with Naive Bayes is that it assumes all the features are independent.

Confusion Matrix	**Fake Democrat**	**Fake Republican**	**Real Democrat**	**Real Republican**
Fake Democrat	167	5	1	3
Fake Republican	1	147	11	3
Real Democrat	0	7	92	29
Real Republican	144	153	208	277

K-NN



- Feature Space (Training Data)
- Distance Metric (Cosine, Euclidean, …)
- The value of k (the number of nearest neighbors retrieve from which to get majority class)
- To classify an unknown record:
  - Compute distance to other training records
  - Identify k nearest neighbors
  - Use class labels of nearest neighbors to determine the class label of unknown record

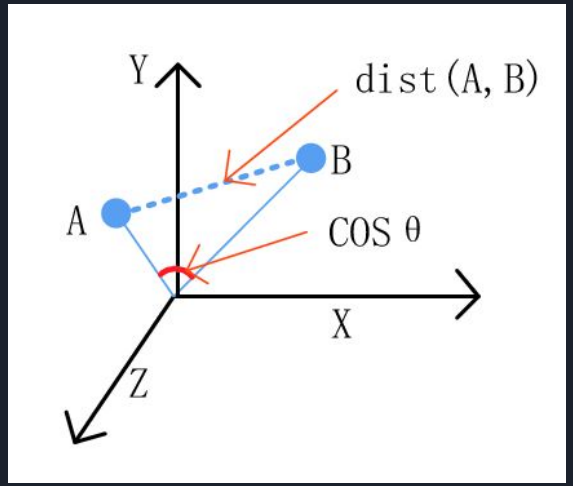
Choosing the value of k:

- If k is too small, sensitive to noise points
- If k is too large, neighborhood may include points from other classes
- Choose an odd value for k, to eliminate ties

KNN-COSINE SIMILARITY

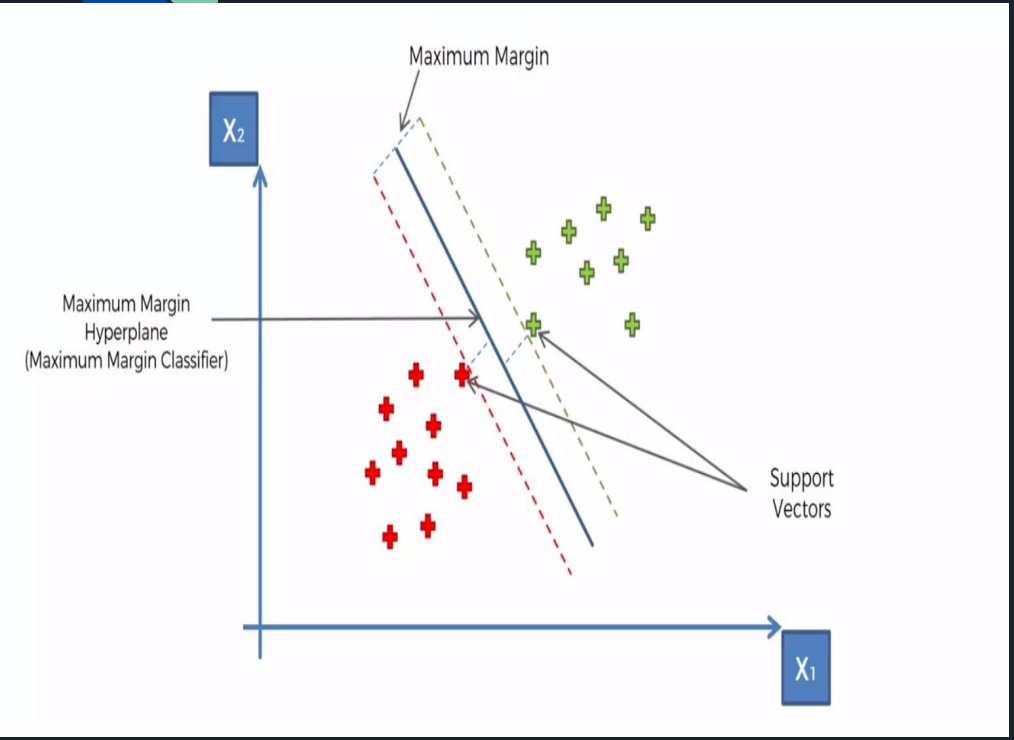
KNN (Euclidean)	With PCA	Without PCA	k=10	k=15	k=20	k=39	k=100
39.0%	69.3%	68.2%	67.4%	68.2%	68.8%	69.3%	67.1%

Cosine-similarity is prefered when working with text data



	Fake Demo	Fake Rep	Real Demo	Real Rep
Fake Demo	329	21	19	7
Fake Rep	83	253	19	25
Real Demo	69	17	230	78
Real Rep	35	22	65	228

SVM



- Support Vector Machine is a supervised machine learning algorithm which can be used for classification.
- SVM is used to classify the data points based on the hyperplane.
- A hyperplane is decided such that the distance between the vectors is maximum.
- Kernels used for the project :
  - Linear
  - Sigmoid
  - RBF

RESULT OF SVM

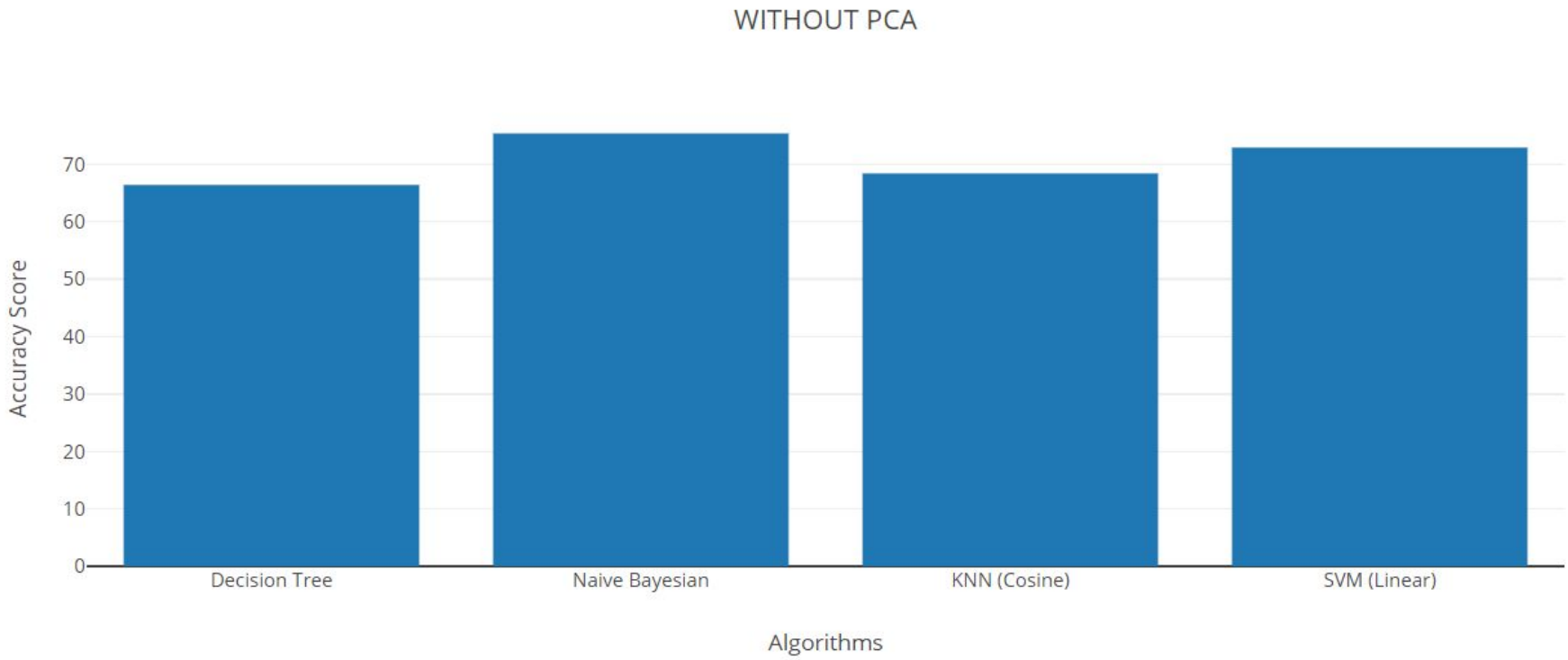
ACCURACY SCORES WITH DIFFERENT KERNELS.

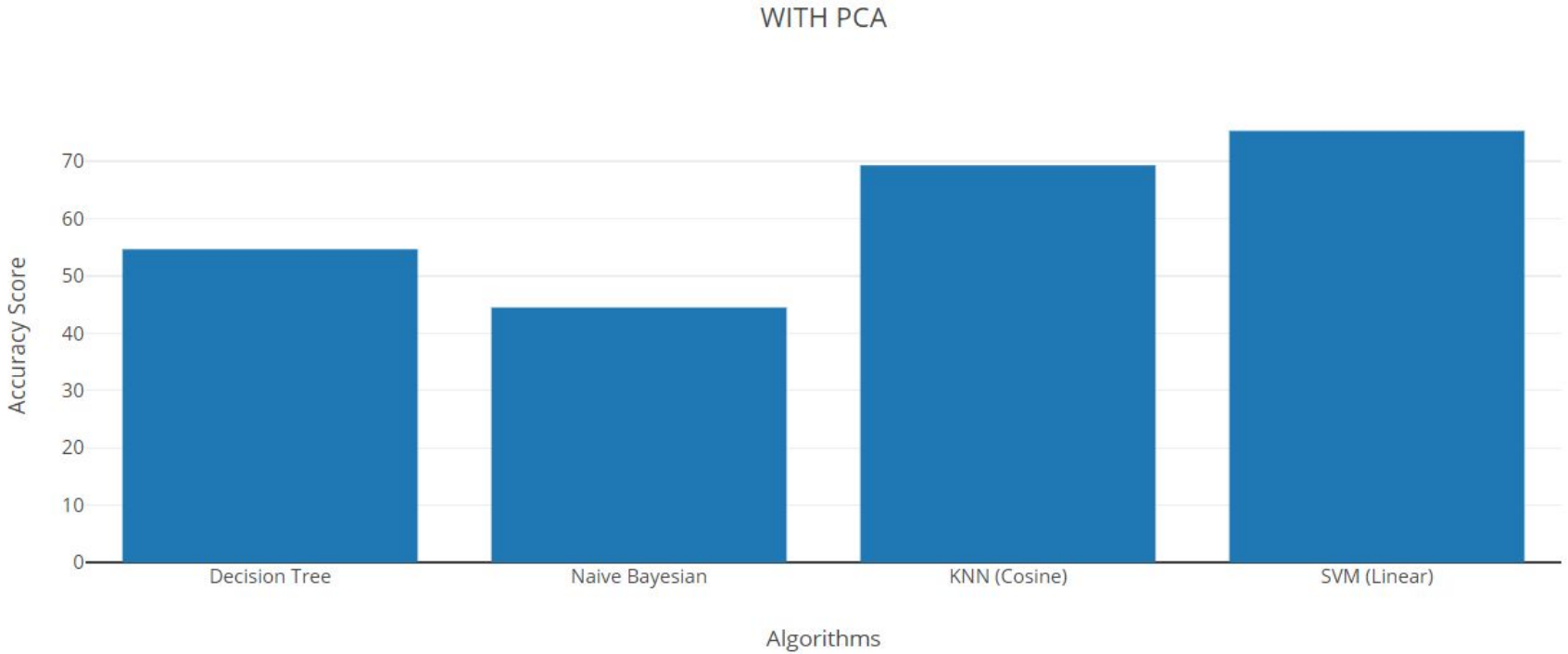
- The accuracy scores of the algorithm have been measured for 3 different kernels.
- The effect of PCA on the accuracy of the classification have been recorded.
- Linear SVM performed the best among the variants of SVM.
- Accuracy gain was possible after the PCA dimensionality reduction.

Kernel	**Without PCA**	**With PCA**
LINEAR	72.9%	75.44%
RBF	23.8%	36.4%
SIGMOID	23.67%	24.16%

SUMMARY OF CLASSIFICATION MODEL AND RESULTS

Accuracy Of Different Algorithms (%)		
Algorithm	**With PCA**	**Without PCA**
Decision Tree	54.64	66.4
Naive Bayesian	44.48	75.36
KNN (Cosine)	69.3	68.4
SVM (Linear)	75.44	72.9


WITHOUT PCA
Accuracy Score
70
60
50
40
30
20
10
0
Decision Tree
Naive Bayesian
KNN (Cosine)
SVM (Linear)
Algorithms


WITH PCA
Accuracy Score
70
60
50
40
30
20
10
0
Decision Tree
Naive Bayesian
KNN (Cosine)
SVM (Linear)
Algorithms

THANK YOU!