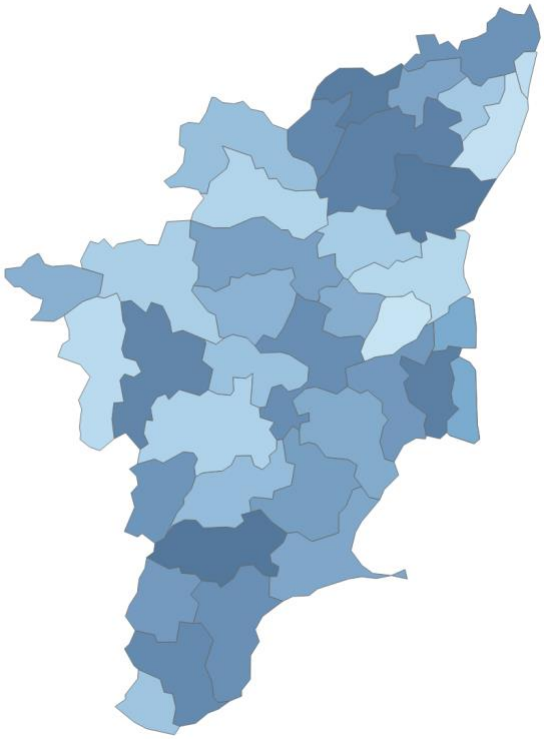
POLITICAL ANALYSIS AND PREDICTIONS



TAMIL NADU LEGISLATIVE ASSEMBLY ELECTION, 2021

2048006 – ARCHER ROZARIO J
2048015 – MANOJ KUMAR
2048057 – SOUNDARYA G

MACHINE LEARNING PROJECT REPORT

TABLE OF CONTENT

Political Analysis and Predictions concerning Tamil Nadu Legislative Assembly election, 2021

I) INTRODUCTION

In the present scenario, the world faces a pandemic situation due to an outbreak of coronavirus, which has affected almost all countries' economies, Especially in India. The economy has gone out of control, imposing a lockdown last year, 2020. which tends to unemployment 6.7% to 26% by the end of March. This year 2021, is also the year where the Legislative assembly general election occurs in Tamil Nadu, Pondicherry, Kerala, and West Bengal. So this year's election has a wide variety of influential factors that affect the election results.

A wide variety of influential factors affect the results of the election. Data analytics and predictive analytics, which involve data collection, extracting meaningful insights from various sources, forecasting the future based on past patterns, have a scope and need in every field, including Politics. Political Analysis nowadays plays a crucial role in every election which happens in and around the world. One such example is how Cambridge Analytica provided a clear blueprint for Trump to win in the 2017 US Presidential Election. The number of Twitter users in India has gone up to 17.5 million in 2021. Most people rely on Twitter to get the fastest and facts and information, with the amount of data generated every day in social media, online blog posts. An online survey assists the analytical companies in using a concept called "*Opinion Mining*" to plan the strategic technique to guide the respective party to micro-target the voters to win the election.

1. Problem Statement

I-PAC Indian Political Action Committee is the platform of choice for students and young professionals to participate in and make a meaningful contribution to political affairs and governance of the country without necessarily being part of a political party. Started as Citizens for Accountable Governance (CAG) in 2013, I-PAC has brought some of the best minds from diverse academic and professional backgrounds together and provided them

with a unique opportunity to become a part of the election process and influence policymaking in India. I-PAC works with visionary leaders with a proven track record. In the process, the group helps them set a citizen-centric plan and partners with them to conceptualize and implement the most effective methods of taking it to the public and gathering mass support.

2. Works of I-PAC

The table represents the I-PAC's association with the respective political party, position, and their year of victory. The first breakthrough of I-PAC was in 2014 were they associated with BJP for the Lok Sabha election. In the year 2014, BJP was the largest NDA party in the parliament.

TABLE 1, I-PAC Services.

CANDIDATE	YEAR	PARTY	POSITION
Narendra Modi	2014	BJP	Prime Minister
Nitish Kumar	2015	Janata Dal	CM of Bihar
Amarinder Singh	2017	INC	CM of Punjab
Arvind Kejriwal	2020	Aam Aadmi	CM of Delhi

In 2021 DMK (Dravida Munnetra Kazhagam) has associated with I-PAC for the Legislative assembly general election in Tamil Nadu. So we have taken a scenario as Archer Rozario J, Manoj Kumar, and Soundarya G are a part of I-PAC's Data analysis team; we aim to analyze the final ground survey as in which we have to get the voter's opinion in Tamil Nadu as well as the opinions from Twitter in the last week of the rally to analyze people's mindset about the election as well the candidates who stand representing their party.

II) ABOUT THE DATA

The dataset for this report comes from different data sources. Our primary data acquisition model comes under Social networks, where Tweets were pulled from Twitter using web scrapping methodology. The appropriate permission was reserved from Twitter. The Tweets dataset includes information about the user name, Location or active Location of the user, and the posted tweets. Twitter dataset metadata are shown in table 2.

TABLE 2, Details of twitter dataset.

S.NO	CRITIERIA	DETAILS
1	**Name**	Twitter Dataset
2	Type	Text data
3	No. of rows	~2500
4	No. of columns	3
5	Missing values	No
6	Target Type	Data Mining
7	Applicable Technique	NLP
8	Data repository	Social Media (Public domain)

This data source model is considered public data, where anyone can access the trending tweets using the internet. The geographical survey data acquisition is considered our secondary data source. The survey conducted 'Survey concerning Tamil Nadu Legislative Assembly election, 2021' by approaching local people and circulated the link through WhatsApp, Instagram, and other social media. Survey data metadata are shown in table 3.

TABLE 3, Details of survey dataset.

S.NO	CRITIERIA	DETAILS
1	Name	Survey Dataset
2	Type	Multivariate Dataset
3	No. of rows	~407
4	No. of columns	24
5	Missing values	No
6	Target Type	Ordinal/Nominal/Categorical
7	Applicable Technique	Classification
8	Data repository	Survey dataset (Owned)

This data source model is considered private data. Since chosen topic requires historical data for statistical and other geographical analysis, we're taking past and current election

details from the governmental repository. Which is accessible by every citizen with the help of the internet. Historical dataset metadata were shown in Table 4.

TABLE 4, Details of Historical dataset.

S.NO	CRITIERIA	DETAILS
1	Name	Election Dataset
2	Type	Multivariate Dataset
3	No. of rows	1638
4	No. of columns	9
5	Missing values	No
6	Target Type	Categorical
7	Applicable Technique	Statistical Analysis
8	Data repository	Governmental repository

Those datasets include information about the election, Public opinions, Assembly constituency details, Parties & Alliances details, Party details, Candidate profile, Past result, Public twitter tweets, and another necessary questionnaire with enough pollster's opinions. Those features descriptions for all the dataset were detailly discussed in following table 4,5,6.

TABLE 5, Features descriptions of twitter data source.

S.NO	FEATURE NAME	DESCRIPTION
1	User	Registered User name of the twitter user.
2	Location	Registered Location, or active location of the twitter user, while posting the tweets.
3	Tweets	Posted tweets in form of Text, Emojis or an Image.

TABLE 6, Features descriptions of survey data source.

S.NO	FEATURE NAME	DESCRIPTION
1	Which of the following best describes your age?	This fields help to identify the age grouping of pollsters. In generally, '18 – 24', '25 – 40', '41 – 50', and 'Above 50' are the provided options.
2	What gender do you most identify with?	To indicate the gender of the pollsters. 'Male', 'Female' and 'Other' are fixed options.

3	Are you a registered to vote?	To identify whether pollsters are registered to vote or not? It's an Yes or No Mandatory field.
4	Are you registered to vote at the current address you reside at?	To identify whether pollsters are registered to vote at the current address?
5	Do you feel that you fully understand the election process?	This fields help to identify how pollsters understand the election process. 'Little', 'Lot' and 'Moderate' are some of the options provided for pollsters.
6	In the last 5 years did you vote in a local election? This includes voting for Councillor & mayors	This fields help to identify whether pollsters has voted in any local election in the last 5 years. It's an 'Yes' or 'No' mandatory field.
7	Did you voted in the 2021 elections?	This fields help to identify whether pollsters voted in 2021 election and It's an 'Yes' or 'No' mandatory field.
8	If not voted, why ?	If Question 7, is answered No. This field capture the reasons why pollster did not voted in 2021 election. FAQ was provided as options.
9	On what basis do you assess a political candidate?	To identify how assessable are the political candidate over the election period. In general- poster ad, Digital ad, Newspaper, TV ad are some common options provided for pollsters.
10	Which of the following best describes your decision to vote in the 2021 election?	To identify how assessable are the political candidate during/over the election period.
11	On what basis you select your political candidate.	This fields help to identify how pollster selecting their political candidate. Example here some of the provided options. - Die-hard fan of the Party/ Candidate - On past experience/Leadership - On their promises - Influenced by others
12	Contributed or collected money	To identify whether pollsters has Contributed or collected money? and It's an 'Yes' or 'No' mandatory field
13	Attended election meetings/rallies	To identify whether pollsters has Attended election meetings/rallies?
14	Participated in door to door canvassing	To identify whether pollsters has Participated in door to door canvassing and It's an 'Yes' or 'No' mandatory field
15	Distributed election leaflets or put up posters	To identify whether pollsters has Distributed election leaflets or put up posters and It's an 'Yes' or 'No' mandatory field
16	Do you think the existing government is going in the right direction to benefit Tamil Nadu's people?	To identify whether the existing government is going in the right direction to benefit people and It's an 'Yes' or 'No' mandatory field

17	How would you rate Edappadi Palanisamy's performance as the Chief Minister of Tamil Nadu	To analysis the performance rating for our Chief Minister of Tamil Nadu. In general, Very good, Good, Neutral, Bad, and Very Bad are the provided option to choose.
18	What is your assessment of the performance of the AIADMK government in Tamil Nadu in the last five years? Would you say that you have been satisfied or dissatisfied with it?	To analysis the performance rating for ruling party of Tamil Nadu. In general, 'Fully satisfied', 'somewhat satisfied', 'somewhat dissatisfied', 'Fully dissatisfied' and 'I'm not sure' are the provided option to choose.
19	During the last two- three years have you or any of your family members benefited from any Government scheme ?	To identify whether the pollster or any of the family members received any governmental scheme during the last two- three years. It's an 'Yes' or 'No' mandatory field.
20	On the day of voting will you vote for the same party which you voted now or your decision may change?	This filed help to analysis the decision of the pollsters on their day of voting.
21	Which party do you support in your ward?	This fields help to identify the party which is having majority in their constituency.
22	What is your opinion about the candidate?	Text descriptive field to process the pollsters opinion about their candidate.
23	Which party will rule Tamil Nadu for the next five years?	To identify Public concerns to determine which party rule Tamil Nadu for the next five years. Options provided with 5 majority part in TN.
24	Why do you think your candidate will win?	To identify Public concerns to determine why pollster wanted his/her candidate to win. Reason were collected in text format and analysed for NLP

TABLE 7, Features descriptions of Historical data source.

S.NO	FEATURE NAME	DESCRIPTION
1	S.no	Unique identity number for each record, ranging from 1 to 1638.
2	Election year	Historical year, only for last two successful election and current 2021 election. (2011, 2016, 2021)
3	District	Name of the 38 District inside Tamil Nadu
4	AC No.	Unique assembly constituency number for each assembly constituency. Ranging from 1 to 234.
5	Assembly Constituency	Given name of the Assembly Constituency under each district. Each of 234 AC named uniquely.
6	Parties & Alliances	Details of Parties & Alliances group

7	Party	Details of Party
8	Candidate	Party-wise name of the candidate for each Assembly Constituency.
9	Result	Assembly Constituency-wise election result for past record and 'pending' status for ongoing election, 2021.

III) IMPLEMENTATION

A) EXPLORATORY DATA ANALYSIS

Exploratory data analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. Here the survey data has 24 attributes out of that 3 are ordinal and rest of 21 are nominal variables. These 24 attributes are the number of questions answered by each surveyee. 25 percentage of the nominal data are Yes/ No question type. Below shown four pie charts are of nominal data.

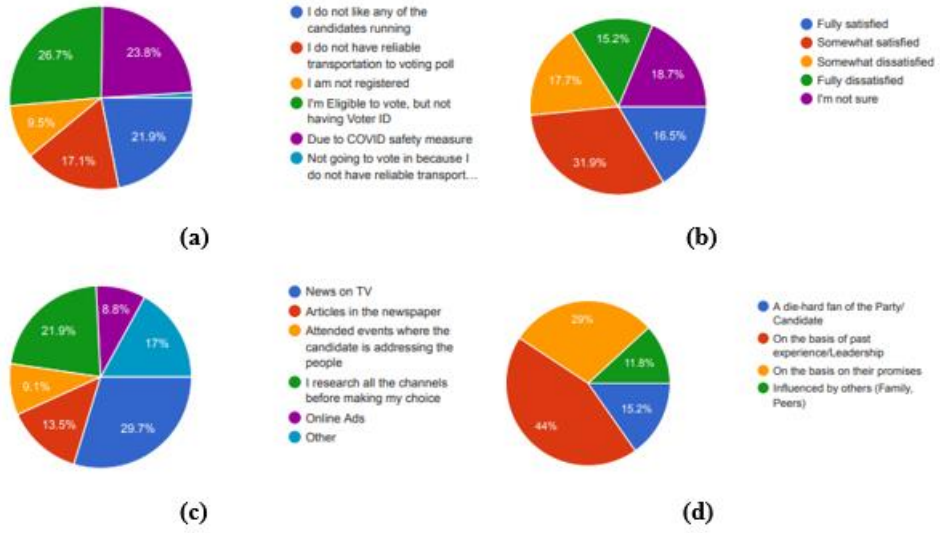


FIGURE 1, Distribution Pie Chart

Figure 1(a), represents the pie chart of "If not voted, why?", where the maximum percentage (26.7%) of surveyee are eligible to vote but doesn't have Voter ID while the minimum percentage (9.5%) of surveyee are not registered to vote.

Figure 1(b), represents the pie chart of "What is your assessment of the performance of AIADMK government in Tamil Nadu in last five years? Would you say that you have

been satisfied or dissatisfied with it?", at most 31.9% of surveyee are somewhat satisfied while 15.2% of surveyee are fully dissatisfied.

Figure 1(c), represents the pie chart of "On what basis do you assess a political candidate?", where most of the surveyee have voted on basis of News on TV and rest on basis of research on all channels, articles and online Ads.

Figure 1(d), represents the pie chart of "On what basis you select your political candidate?", where most of the surveyee have voted on basis of past experience/ leadership and rest on basis of their promises, influenced by others and a die-hard fan of the party/ candidate.

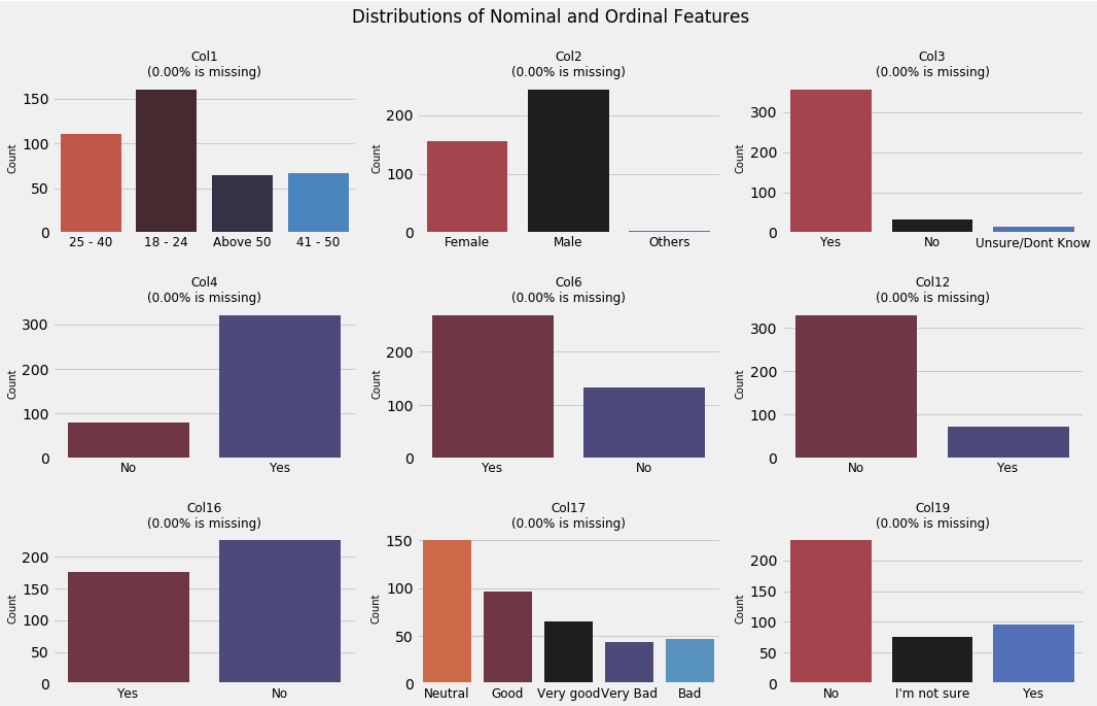


FIGURE 2, Distribution of Nominal/ Ordinal Features

In *Figure 2*, Col1 and Col17 are ordinal variables, whereas Col2, Col3, Col4, Col6, Col12, Col16 and Col19 are nominal variables. From Col1, we can understand that maximum number of surveyees are of 18-24 age and have equal number of surveyees in above 50 and 41-50 age category. From Col2, we can notice that males have responded the survey more than female and others count.

From Col3 its clear that most of them are aware of their registration while less than 50% aren't aware of it. We can observe a similarity between Col4, Col6, Col12 and

Col16, that is all represents yes/no type of questions. From Col4 it says most of them have registered to from current address. From Col6 its clear that most of the surveyees have already taken part in previous elections. From Col12, we can understand that there is small percentage of surveyees who have involved in collecting money for their votes.

From Col17, it's clear that most of the surveyee have neutral opinion on current CM's performance. It looks like most of them are satisfied by his work as CM, since majority of surveyees have chosen good, very good and neutral.

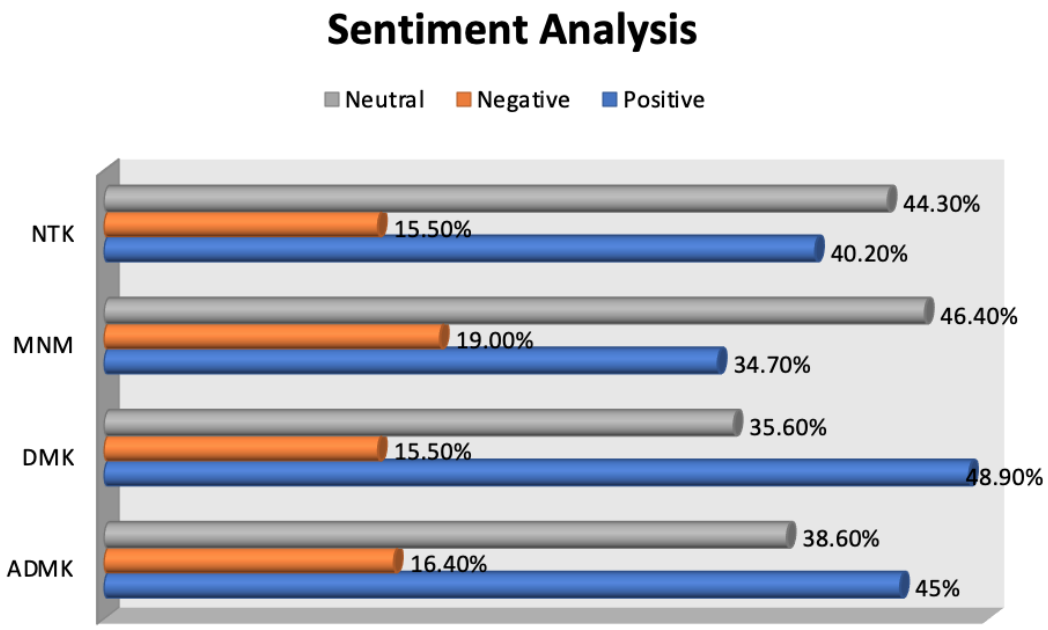


FIGURE 3, Sentiment Analysis

Based on the polarity of the extracted tweets we are labelling the tweets as positive, negative and neutral. From the above chart we can see that DMK has the most positive tweets (48.9%) when compared to the other major partied. MNM has the least positive percentage (34.7%) and also it has the highest negative and well as neutral tweet percentage (19.00% & 46.40) respectively.

There is a close fight between ADMK and DMK with 3.9% difference in positive tweets percentage giving a upper hand for DMK for the majority. Though MNM and NTK doesn't have a majority in the polling there will definitely be huge portion of voters voting for them which leads to increase in distribution of votes.

Initially we performed three different word clouds based on positive tweets, negative tweets and neutral tweets. After that we manually picked the most reflected words

considering in all the three clouds and represented it as a overall word cloud for that particular party

FIGURE 4, NTK Word Cloud

In the above word cloud we can see words such as *"bjp", "fake"," b-team"* which represents that they is conspiracy that NTK which is lead by Seeman is an alternate team of BJP to increase split ratio of the voters. Along with that we can see *"youth", "change", "no alliance"* there are key features of NTK party they never go with another party for alliance and his speech always attracts youth. "*Dictatorship*" in all his speech's we can see a shade of one people one leader and concept of no alliance with any other party leads to a think that NTK party may possess a dictatorship rule

FIGURE 5, ADMK Word Cloud

In the ADMK word cloud we can see "*north", "alliance"," money"* since it is the ruling party in Tamil Nadu with alliance with BJP people think of that ADMK is directly influenced by the Central Government.

They have enough money so that each voters in specific wards were given Rs.2000 for a vote. *"Constituency"* whoever the candidate may be there are few places called with high "*mukkulathor"* cast ratio so in those places ADMK will always be the winners. "Better Candidate" in 2021 election there are many wards ADMK alliance have placed unfamiliar candidates to fight against popular leaders of other major political parties.

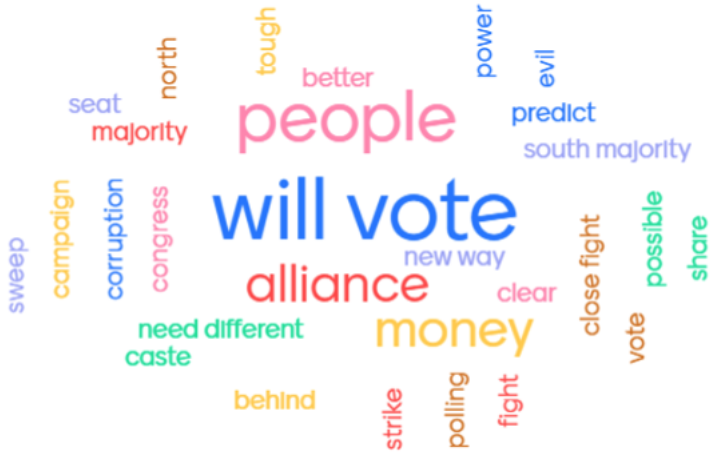


FIGURE 6, DMK Word Cloud

In the above DMK word cloud we can see *"alliance", "congress", "new way", "need different" ,"south majority"* words represent the promises he made during the election campaign and rally. Alliance with congress gives DMK an upper hand over the southern coastal area of Tamil Nadu (*Kanniyakumari). "Corruption", "evil*" denote that during their last ruling period DMK leader M. Karunanidhi and his family members were accused over the 2G scam but the verdict was favourable to DMK.

FIGURE 7, MNM Word Cloud

In the MNM word cloud we can see very few words when compared to the other three major parties *"Controller", "better", "speaker", "casual" ,"helping"* denotes the characteristics of MNM party leader Kamal Haasan though he doesn't have much experience in politics he has his control over his party members and candidates he is better thinker and better speaker when compared to the other party leaders.

"Casual", "helping" in Coimbatore South(his ward) there were few places where people didn't have proper sewage system and water facility he approached those places casually without any media coverage and promised them that he will consider to renovating those place as the first priority when he wins the election

B) DATA CLEANING

Data cleaning is a critically important step in any machine learning project. Data cleaning refers to all kinds of tasks and activities to detect and repair errors in the data. Comparing Tweets and survey datasets, we found tweets from the messiest dataset since we restricted our survey questioners with required fields. In contrast, Twitter web scrap pulls all the tweets without any necessity.

Further, we identified and removed the columns that contain a single value in survey data and duplicated tweets from the Twitter dataset as the measure of the data clearing process.

1) TWITTER DATA CLEANUP

Twitter Cleaning The tweets column in the twitter dataset, all the tweets are cleaned by removing *'@,#,RT, http:(hyperlink)', 'emojis', 'stop words', 'punctuations', 'digits'* and *'non-English'* words using all possible UNICODE value and regular expression.

2) MISSING DATA & REMOVING OUTLIERS

Removing outliers- Understanding the core domain of our project, we identified the uncertain situation and handled it as an outliers.

[Tableau Dashboard 1 - Link]

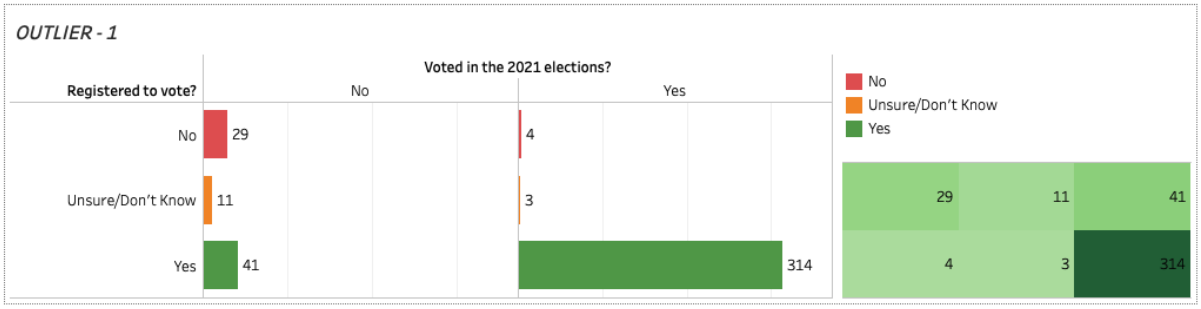


FIGURE 8, Outlier 1

Narrowing the questioners to 'Registered to vote?' and 'Voted in the 2021 elections?' we spotted out an outlier. We have undergone a situation where a citizen has not registered to vote but has voted in the 2021 election, which is not possible. Hence we recognize four records as outliers and handled them.

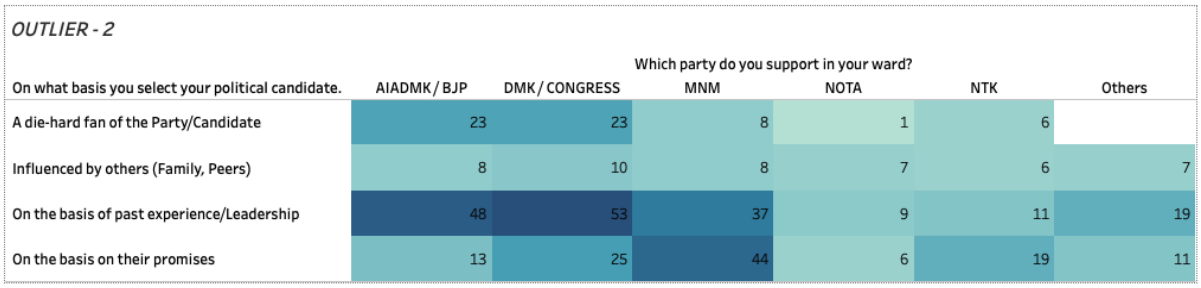


FIGURE 9, Outlier 2

Here, we can see that more people have voted based on their political candidate's past experience and leadership skills. Their support level has cast them a high number of votes for DMK of 53 votes, and AIADMK of 48 votes have been recorded based on past experience and leadership. On the line, MNM has 44 votes based on their promises. The data also shows that AIADMK and DMK received more support and votes from their die-hard fans.

The Outlier that can be identified here is The die-hard fans who have voted for NOTA. NOTA is not a party, and it indicates that the voter has not chosen to vote for any of the parties. Therefore, we cannot proceed with this DATA as it is not logical that a die-hard fan has voted for NOTA.

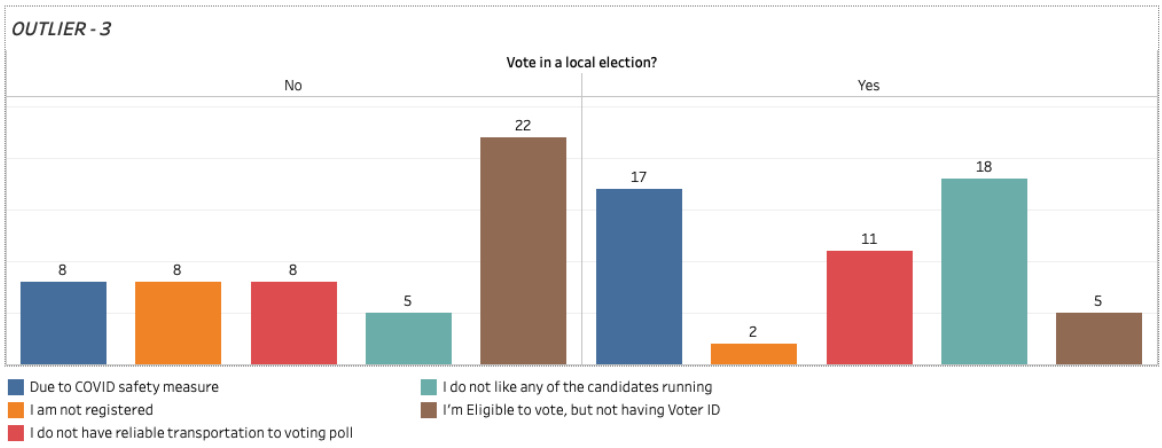


FIGURE 10, Outlier 3

In this Data, we could see those who have voted and not voted in the local election. We can see that people have not voted due to factors like no reliable transportation, COVID safety measures, and others by analysing the data. On the other side, we can see the data of people who have voted. The outlier spotted here is '2' people have said that they have voted, but they are not registered to vote, and '5'have voted, but they do not have Voter Id to vote. Without registration and Voter ID, people cannot cast their votes. Therefore, the Data is not logical and unclear. So, we reject this Data before proceeding to the next step.

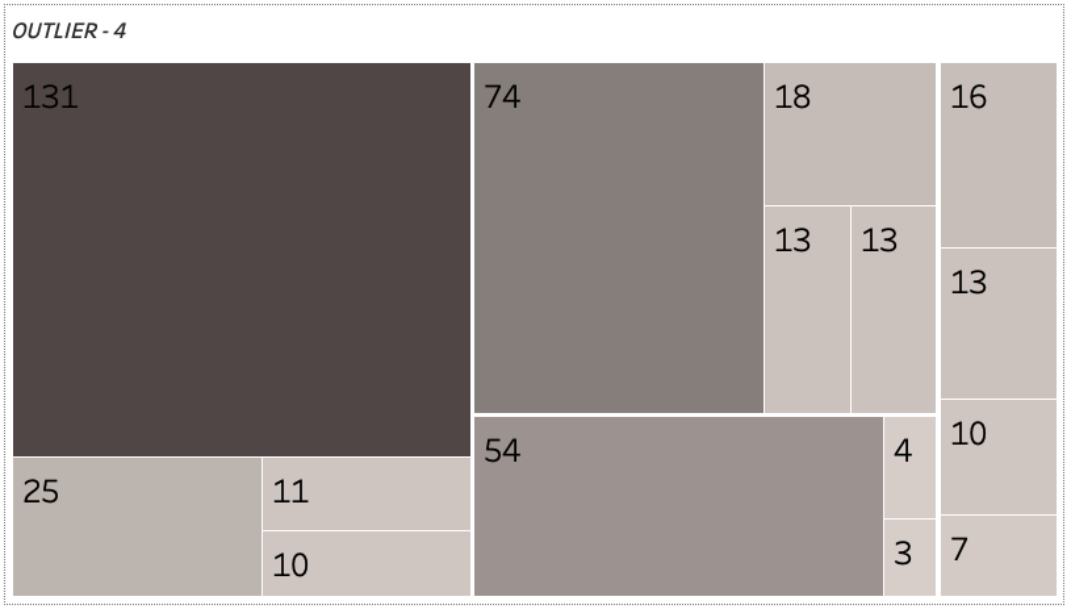


FIGURE 11, Outlier 4

Here we analyze that whether the Government is going in the right direction. People answering this should also have to give ratings for Edappadi Palanisamy. '111' People

who have said that it’s not going in the right direction have given a neutral rating, and ‘44’ people have rated it to be bad, and ‘44’ have said that it is very bad. The Outlier spotted here is ‘3’ People who have said that the Government is going in the right direction have also given bad and very bad ratings to Edappadi. This is because that they like the Government but personally do not like Edappadi. There is a bias in the dataset. So, we remove this data before proceeding to the next step.

FIGURE 11, Outlier 5

This data shows how people have cast votes based on specific criteria. A high number of '131' people have voted based on past experience of the political candidate, and '74' people have voted based on the candidate's promises. '54'die-hard fans voted to the specific party they like and also were sure about whom they were voting for. The Outlier that can be identified here is '3' die-hard fan of the particular party said that they would vote for the party they like, but they also said they don't care about who wins. Being a die-hard fan for a particular party but doesn't care who wins is not logical. So, we reject this data before proceeding to the next step.

C) DATA PREPROCESSING

Data Pre-processing is a technique for converting raw/unclean data set into a clean data set. In other words, the data gathered from various sources are of raw format which is not feasible for the analysis. The need to pre-process the data are to make database more accurate, boost consistency, and smooth the data.

1) WORKAROUND WITH COLUMNS

In the survey dataset, we can observe the column names are too lengthy which will be trouble for the programmer. Hence, the column names “Which of the following best describes your age?”,” What gender do you most identify with?”, “Are you a registered to vote?” are changed into Col1, Col2, Col3. Similarly other columns are also renamed. This renaming of columns makes programmer more convenient to code.

TABLE 8,Workaround With Columns

ORIGINAL COLUMN NAME	RENAMED NAME
Which of the following best describes your age?	Col1
What gender do you most identify with?	Col2
Are you a registered to vote?	Col3
Are you registered to vote at the current address you reside at?	Col4
Do you feel that you fully understand the election process?	Col5
In the last 5 years did you vote in a local election? This includes voting for Councillor & mayors	Col6
Did you vote in the 2021 elections?	Col7
If not voted, why?	Col8
On what basis do you assess a political candidate?	Col9
Which of the following best describes your decision to vote in the 2021 election?	Col10
On what basis you select your political candidate.	Col11
Contributed or collected money	Col12
Attended election meetings/rallies	Col13
Participated in door-to-door canvassing	Col14
Distributed election leaflets or put-up posters	Col15
Do you think the existing government is going in the right direction to benefit Tamil Nadu's people?	Col16
How would you rate Edappadi Palanisamy’s performance as the Chief Minister of Tamil Nadu	Col17
What is your assessment of the performance of the AIADMK government in Tamil Nadu in the last five years? Would you say that you have been satisfied or dissatisfied with it?	Col18
During the last two- three years have you or any of your family members benefited from any Government scheme?	Col19
On the day of voting will you vote for the same party which you voted now or your decision may change?	Col20

Which party do you support in your ward?	Col21
What is your opinion about the candidate?	Col22
Which party will rule Tamil Nadu for the next five years?	Col23
Why do you think your candidate will win?	Col24

2) WORKAROUND WITH CATEGORICAL RECORDS

The survey dataset is combination of both nominal and ordinal data, where Col1, Col17 and Col18 are ordinal data and rest all nominal data. Hence Label encoding has to be done to carry with further analysis. Label Encoding is used to encode nominal and categorical values. It is simply converting each value in a column to a number. For example, the Col2 describes the gender of surveyee [Female – 0, Male- 1 and Others-2]. Similarly, rest of the columns are encoded except Col22 and Col24.

TABLE 9,Workaround With Columns

COLUMNS	OPTIONS	ENCODED
Col1	18-24	0
	25-40	1
	41-50	2
	Above 50	3
Col2	Female	0
	Male	1
	Others	2
Col3, Col4	No	0
	Yes	1
	Unsure/Don’t Know	2
Col5	None at all	0
	A little	1
	A moderate amount	2
	A lot	3
Col6, Col7	No	0
	Yes	1
Col8	nan	0
	Eligible to vote, but not having Voter ID	1

	I do not have reliable transportation to voting poll	2
	I am not registered	3
	I do not like any of the candidates running	4
	Due to COVID safety measure	5
Col9	Articles in the newspaper	0
	News on TV	1
	I research all the channels before making my choice	2
	Attended events where the candidate is addressing the people	3
	Online Ads	4
	Other	5
Col10	I am going to vote, and I know which candidate I am voting for.	0
	I am going to vote, and I don't care who wins.	1
	I research all the channels before making my choice	2
	I am going to vote, but I am not sure who I will vote for yet.	3
	Other	4
Col11	A die-hard fan of the Party/Candidate	0
	On the basis on their promises	1
	On the basis of past experience/Leadership	2
	Influenced by others (Family, Peers)	3
Col12,Col13, Col14,Col15, Col16	No	0
	Yes	1
Col17	Very Bad	1
	Bad	2
	Neutral	3
	Good	4
	Very Good	5
Col18	Fully dissatisfied	1

	Somewhat dissatisfied	2
	I’m not sure	3
	Somewhat satisfied	4
	Fully satisfied	5
Col19	No	0
	I’m not sure	1
	Yes	2
Col20	Not prefer to say	0
	Vote for the same party	1
	May change	2
Col21	NOTA	5
	MNM	3
	AIADMK / BJP	1
	DMK / CONGRESS	2
	NTK	4
	Others	0
Col23	Edappadi Palanisamy	1
	M. A. Stalin	2
	Kamal Haasan	3
	Seeman	4
	T. T. V. Dhinakaran	5
	Others	0

D) MODEL BUILDING

Building machine learning models that have the ability to generalize well on future data requires thoughtful consideration of the data at hand and of assumptions about various available training algorithms. Machine learning consists of algorithms that can automate analytical model building. Model Building Consists of 6 steps, collect pre-processed data, feature selection, choose the right model, train data with chosen model, parameter tuning and evaluate the model.

Here the pre-processed data has gone under feature selection to get most essential features based on sequential forward selection. Later, the data is split into train and test data sets. In order to find the best model to our dataset, we have taken 8

classification models: Gaussian Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, XGBoost, Support vector classifier, ANN and KNN.

These 8 models are compared with hyper parameter tuning, without hyper parameter and with PCA implementation. The measure used to compare models were accuracy in percentage and execution time/ time complexity. The below table shows the detailed comparison between the eight models.

IV)DEPLOYMENT

To visualize our report we're using Tableau Public, It is a free platform to explore, create, and publicly share data visualizations online. Visualizations that have been published to Tableau Public ("vizzes") can be embedded into web pages and blogs, they can be shared via social media or email, and they can be made available for download and exploration by other users.

*TABLE 10, **ASSEMBLY CONSTITURNCY COMPARISON MAP***

S.NO	LAYOUT TYPE	COUNT	FEATURE DESCRIPTION
1	Card	4	Highlighted key values in terms of Tamil Nadu election, 2021.
2	Geographical Tamil Nadu map	3	Comparative Tamil Nadu map Assembly constituency by district. 2011, 2016, and 2021 election were plotted.
3	Interactive legends	1	Party & Alliance legends.
4	Highlight filters	1	Highlight search filter for picking selected district-wise comparison.
5	Table	1	Raw data which table contains all the raw records.

Based on our text mining and statistical analysis we're developed an comparison dashboard using Tableau Public and was deployed in Tableau web server.

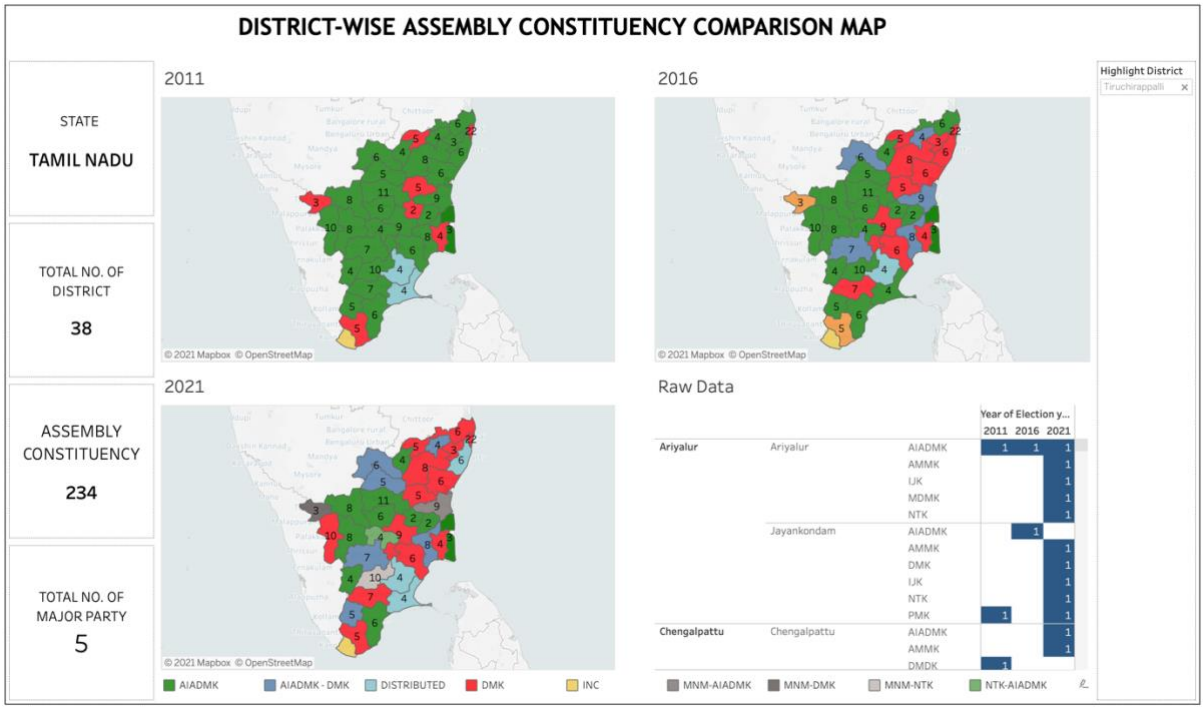


FIGURE 12, Assembly Constituency Dashboard

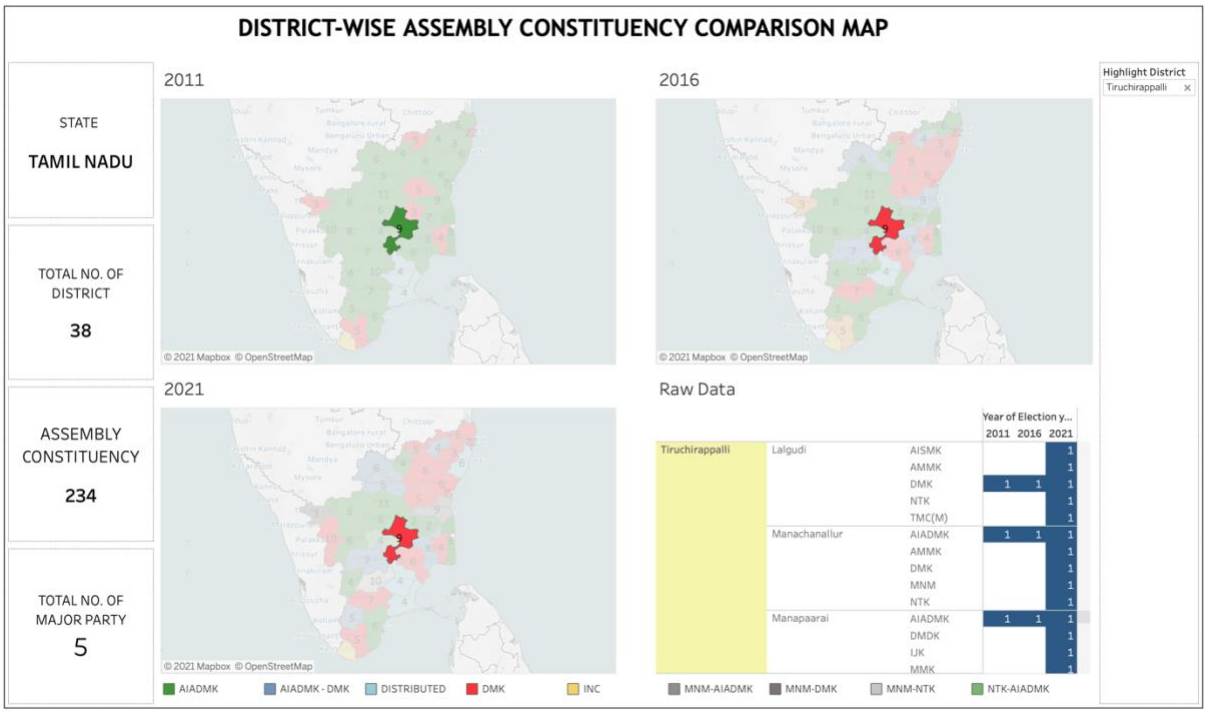


FIGURE 13, Highlight Search Filter

Above Figure , Highlight search filter is used to toggle between different district-wise comparison on Party and alliance over the period of 2011, 2016, and 2021. 38 individual district was provided as the option in the search field, where user can select any one district at a time for comparison.

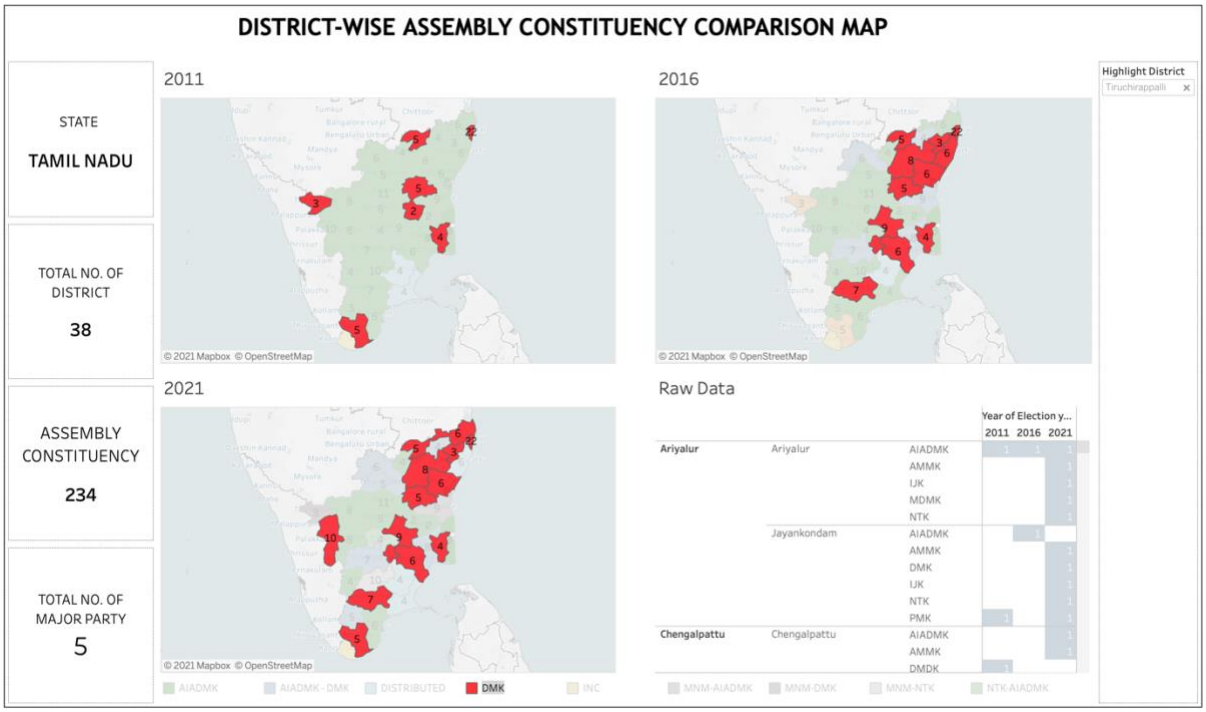


FIGURE 14, Interactive Party & Alliance legends.

Above Figure , Interactive Party & Alliance legends is used to toggle between different Party -wise comparison on the map over the period of 2011, 2016, and 2021. where user can select any one Party & Alliance at a time of comparison.

V) RESULT AND DISCUSSION

PCA Implementation reduced the accuracy of the models drastically except for the Gaussian Naïve Bayes model, which shows that PCA implementation is not recommended for survey datasets.

XGBoost Classifier is the best performing model both with or without hyperparameter tuning.

There is no vast change in accuracy value before and after hyperparameter tuning. Without feature selection the accuracy value tends to fall down expect for Gaussian Naïve Bayes model.

The following two high-accuracy models with good time complexity, with or without hyperparameter tuning, are Random Forest and SVC.

As per the survey and sentimental analysis results, it is evident that DMK and its alliance will win the election with the majority.

TABLE 11, Model Comparisons

MODELS	Without Hyper Tuning and PCA		With Hyper Tuning		With PCA Implementation	
	Accuracy (%)	Execution time (secs)	Accuracy (%)	Execution time (secs)	Accuracy (%)	Execution time (secs)
Gaussian Naïve Bayes	30	0.006	30	0.007	44	0.012
Logistic Regression	55	0.027	56	2.97	46	0.027
Decision Tree	56	0.002	50	0.446	40	0.005
Random Forest	64	0.234	51	9.824	43	0.337
XGB Classifier	67	0.402	61	10.013	34	0.467
Support Vector Classifier	59	0.014	55	1.9	42	0.01
Artificial Neural Network	56	0.766	56	20.824	40	0.78
KNN	50	0.002	60	47.102	35	0.002

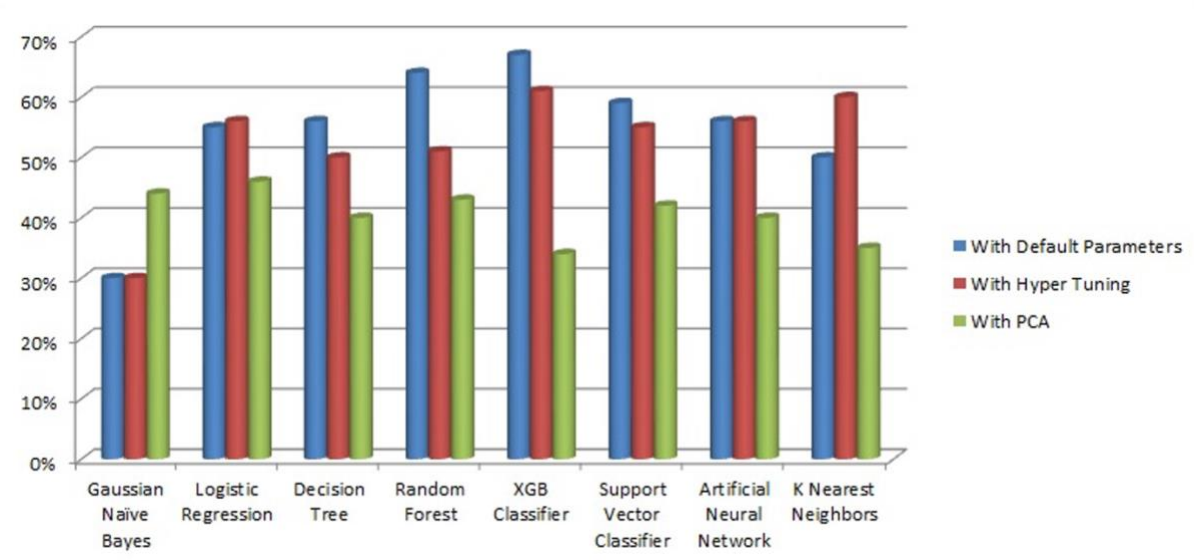


FIGURE 15, Model Accuracy Comparison

VI) LIMITATIONS

The sample size of the dataset was not collected based on population size and Didn't check for the statistical compatibility between survey questionnaire questions. So, we consider there is some lack of data acquisition in Survey dataset.

The collected survey data doesn't represent entire constituencies due to limited time constrains, ongoing pandemic, and privacy concerns of public people for answering some of the survey questions.

We lacked with continuous numeric variables in captured dataset to explore complete Exploratory data analysis before proceeding with model building.

We have faced several issues while translating native(Tamil) and Tanglish tweets into meaningful English sentences using Google API. Because of this translating limitation, we stick on our tweets collection to English only.

Comparison and representation of geographical Party information were limited to District-wise, since Tableau Public Map service didn't provide an option to represent the party's by constituency-wise. Which made us unable to project more accurate prediction and comparison report.

This project backup with three different data sources namely Tweets (Text data), Survey dataset, and Historical dataset. It made us complicated while integrating, building generic model and interpreting the conclusions.

The predication was based on multi-classification(more than 2 majority Parties) which impacted the model accuracy values when compared with binary classification dataset.

VII) CONCLUSION

In 2021 election is the end and beginning of a new era for Tamil Nadu Politics. Two great leaders of the two powerful parties in Tamil Nadu are no more. They are arising of new political parties, which led to a diverse split of the vote ratio. Young, Experienced, and Well Educated candidates are volunteering themselves into politics, especially in parties like MNM (MAKKAL NEEDHI MAIAM) and NTK(NAAM TAMILAR).

Though there are many new parties still ADMK and DMK, hold the majority in Tamil Nadu. The loss of late Chief Minister SELVI. J. JAYALALITHA has put the ADMK government and party in a vulnerable position.

From the Predicted graph of the 2021 Tamil Nadu election, we can see that in the northern part of Tamil Nadu and the districts near the Western Ghats, the majority is now under DMK and MNM, NTK respectively, which ADMK initially dominated.

Except for the place where ADMK can win based on castes such as "*mukkulathor*" and "*vanniyar*" almost in all other districts, most ADMK has now turned into the distributed ratio between ADMK and DMK.

As per the survey and sentimental analysis results, it is evident that DMK and its alliance will win the election with the majority.

VIII) FUTURE WORK

Expanding the result and comparisons of major parties projected instead 38 district by 234 constituencies, to get gain more appropriate result predictions.

The complete political analysis and prediction project acts as the historical data support for future expansion and other political related projects predictions in Tamil Nadu.

Expanding data acquisition with different types of data sources representing entire state and 234 constituencies.

BIBLIOGRAPHY & REFERENCES

[1] https://www.elections.tn.gov.in/

[2] https://en.wikipedia.org/wiki/2021_Tamil_Nadu_Legislative_Assembly_election

[3] https://www.elections.tn.gov.in/TNLA2021.aspx

[4] https://scikit-learn.org/stable/supervised_learning.html

[5] https://www.kdnuggets.com/2020/04/hyperparameter-tuning-python.html

[6] https://www.tableau.com/community

[7] https://public.tableau.com/en-us/s/