CS 210 Project Report for the TheGrandAssemblyOfDataScientists Group

Kanat Özgen

June 7, 2023

1 Introduction

In the current political climate of Türkiye, it is very important to follow the political trends in a chronological context. Changes in the trends of political paries in the lawmaking process is very important, considering the fact that these very laws are used by the populations that elect them.

In this case, our team decided to gather the data of deputies that served between the 22^{th} and 27^{th} installments of the Turkish Grand National Assembly to develop prediction models in order to determine whether the present data of the deputies would be deterministic of their political affiliations, political leanings, etc.

For this, we used various supervised learning models, such as random forest, neural networks and regression. More of this can be found in the last section of this report.

For futher examination of our dataset and findings, you might want to visit the GitHub repository of our project. We have included several interactive maps and graphs for your evaluation.

2 Data Collection

2.1 Scraping of the TBMM website

In the state-affiliated website of TBMM, Turkish National Grand Assembly, there exists page that loads the hyperlink references of deputies in the format of an interactive map. Due to the dynamic nature of the web design of the page, we referred to search engine page instead.

(a) Search Engine Page

(b) Interactive Map

Figure 1: Figure (a) has the capability of showing all the installments whereas Figure (b) only shows the 27^{th} installment.

Also, due to the fact that pages were dynamically loaded using JavaScript, we had to manually extend the search results to get to the individual links of deputies. After having the links extended, we scraped the hyperlink references for the deputies. This list for separate installments can be found in the "data" section of the repository, alongside with the script `save_mv_links.py` that we used to scrape the separate links of the deputies.

After having the links at our hands, we had the problem of getting into separate links that are inside the links that we scraped. For this, we used a selenium webdriver. This was an arduous task, as the process of scraping using automation takes great amount of time. For the sake of optimization, we used Headless Chrome, and disabled CSS for the pages. Due to a problem in the scheduling process of threads, we scraped different installments with our computers.

Even though we had optimized the driver, it took approximately 2 days of scraping. The scraping script `save_mv_links.py` gets the deputy data as JSON objects, and appends them to a text file. The interoperability of JSON objects is very handy in this case.

For preparation for data analysis and models, we use a separate script called `to_csv_file.py` to get the JSON objects into a csv file for further use with `pandas` library.

Our JSON objects have these attributes:

1. `KANUN_TEKLIF_ILK`
2. `KANUN_TEKLIF`
3. `SORU_ONERGELERI_YAZILI`
4. `SORU_ONERGELERI_SOZLU`
5. `GENEL_GORUSME_ILK`
6. `GENEL_GORUSME`
7. `MEC_SORUSTURMA_ILK`
8. `MEC_SORUSTURMA`
9. `MEC_ARASTIRMA_ILK`
10. `MEC_ARASTIRMA`
11. `GENSORU_ILK`
12. `GENSORU`

These attributes carries out the quantity of the said deputy's activeness in that term, for example, the `KANUN_TEKLIF` attribute holds the amount of bills presented to the assembly. `KANUN_TEKLIF_ILK` holds the count where the MP is the first one to sign the bill.

As a bonus, we scraped age data of deputies in order to get an idea of the trends of political parties and the changes of the mean ages of parties throughout the terms. For this, we automated creating search queries on Google and scraped Wikipedia. All of the relevant scripts can be found in the repository. Note that this was done as a bonus and this additional data was not used as a feature in ML models that we prepared.

2.2 Data Completion

Due to the fact that there was maintenance in some pages, some pages of the deputies were not extracted by our script. Due to these erroneous pages, we could have partial data completion. See Table 2 for the proportions.

Installment	MP's Extracted	Percentage (%)
22^{th}	552/550	100.0
23^{th}	550/550	100.0
24^{th}	546/550	99.27
25^{th}	547/550	99.45
26^{th}	548/550	99.63
27^{th}	596/600	99.33

Table 1: Table shows the amount of MP data that were extracted from the website.

Also, in opposition to our proposal, due to the complex nature of the website and the fact that there was an enormous amount of textual non-processed data, we abandoned our concept of adapting an NER model to see the overall topics of interest of MP's. With this in mind, our process of data collection was confined to collecting data pertaining to quantity, not the quality of the efforts made.

It is very important to mention that `GENSORU` and `GENSORU_ILK` attributes became redundant, since they became what is known as "Mülga" in Turkish Constitution. It is not possible for a deputy after 27^{th} installment to give an interpellation. Therefore, this data became obsolete in the later processes.

3 Data Cleaning and Standardization

3.1 Standardization of the Data

A Min-Max scaler for the machine learning applications were put to use in order to use the neural network. We used the number of deputies inside a city as a normalization metric. In lesser crowded cities, there are lesser amounts of deputies sent to the assembly, whereas in larger cities; the contrary is true. And we can say that this can be a way to normalize data. For example, there are a lot of deputies in Istanbul but Mahmut TANAL is a very active deputy, which in turn does not affect the overall average of Istanbul since there are a lot of Istanbul-designated deputies in Istanbul.

3.2 Determining Outliers

There are certain outlying deputies that change the overall data drastically. When we investigate the overall activity averages for the cities, we get certain outliers.

City	Total Activity Average
Ardahan	1105.91
Niğde	856.22
Kırklareli	385.44
Iğdır	342.00
Artvin	284.50

Table 2: Table shows the top 5 average total activity in cities.

These outliers are Niğde and Ardahan, where their overall activity average is three-fold of the average. When we investigate these cities where potential outliers are present, we get the result like this:

Deputy	Total Activity
Ensar ÖĞÜT	13107
Öztürk YILMAZ	90
Orhan ATALAY	44
Kenan ALTUN	24
Saffet KAYA	5

Figure 2: Total Activity counts in Ardahan

Deputy	Total Activity
Ömer Fethi GÜRER	13860
Doğan ŞAFAK	522
Orhan ERASLAN	384
Mümin İNAN	303
Selim GÜLTEKİN	121

Figure 3: Total Activity counts in Niğde

Deputy	Total Activity
Alim IŞIK	7297
Ali Fazıl KASAP	1064
Ahmet ERBAŞ	166
Ceyda ÇETİN ERENLER	87
İdris BAL	62

Figure 4: Total Activity counts in Kütahya

Therefore, the rows of Ömer Fethi Gürer, Ensar Öğüt and Alim IŞIK are dropped from the dataset. Their absence will make the dataset much more concise. Even though there could be other outliers, they are getting normalized by the sheer quantity of the deputies within that city. The other cities are normal after these drops, so theres no more action required.

3.3 The 25th Installment

25^{th} installment of the Turkish Grand National Assembly was a rather short one, that is why there is a drastic change in the statistics of the deputies. That is why, the data that pertains to this very specific installment can be discarded.

4 Hypotheses and Research Questions

Our main hypothesis is that the probability of a supervised classifier model to determine the political affiliation and the political stance is statistically significant.

Is a Random Forest Classifier sufficient to determine whether there is a statistically significant decision model to predict the affiliation of a deputy? Is there a neural network model that can predict the political affiliation and the stance of a deputy with a statistically significant precision? Is there a Naive Bayes model that can also do these? Of certain classifications, which ones yield the most statistically significant results? How are the cities' average activeness averages distributed across Türkiye? Are there any correlations between the attributes of the deputies? Is age a significant data?

5 Findings and Results

There are a lot of things to be inferred from this dataset. Figure below shows the correlation matrix of the attributes of the deputies. First of all, the overall changes in the parties' presence in the assembly and their overall total activity can be investigated.

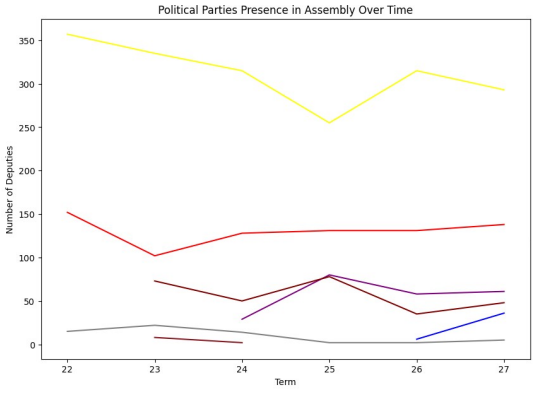

Figure 5: Parties' presence in the assembly.

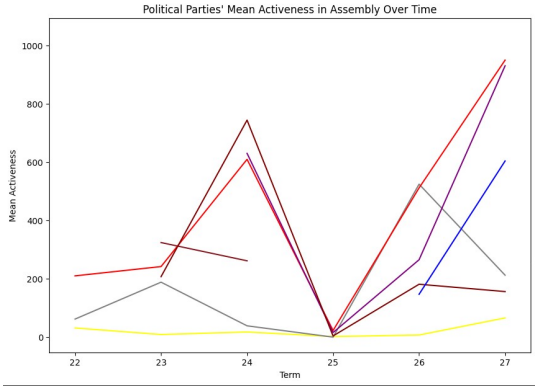

Figure 6: Parties' activity in the assembly.

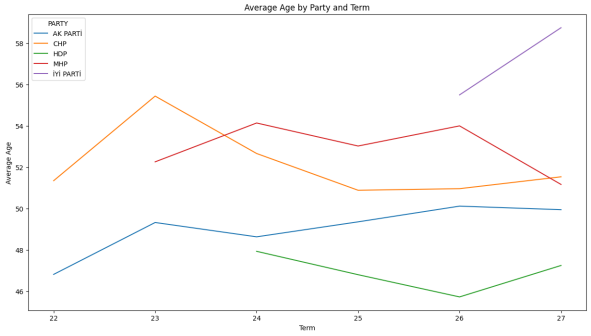

Figure 7: Age distributions and changes.

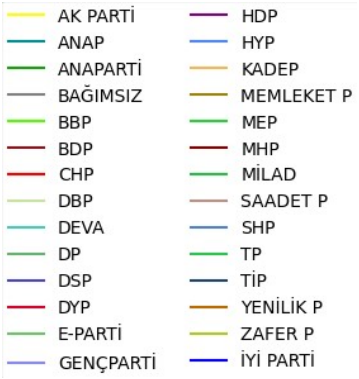

Figure 8: Legend for the party colors.

Even though MHP deputies were the most active ones in the parliement, after having done an alliance with the ruling party AKP, there has been a fall in their overall activeness. This can be explained with the AKP's phenomena since they are allied with the ruling party and ruling parties are not very inclined to be active.

We see a gradual increase in the mean activeness of the opposing parties' deputies where İYİ parti, precursor of which is MHP, looks as if it has taken over the activeness of MHP over time, and considering the fact that majority of İYİ parti deputies are ex-MHP, it is very logical that generally

active deputies left MHP for the sake of İYİP after the decision of MHP to be allied with the governing party.

There is an overall increasing trend for the opposition parties' activeness. Also, there's an increasing trend of nationalist parties for the passage from 26th to 27th. Also, it can be seen that AKP lost some seats in 25th while HDP increased its seats by a great margin. These can be explained by some paradigm shifts, since HDP and MHP, İYİP are polar opposites.

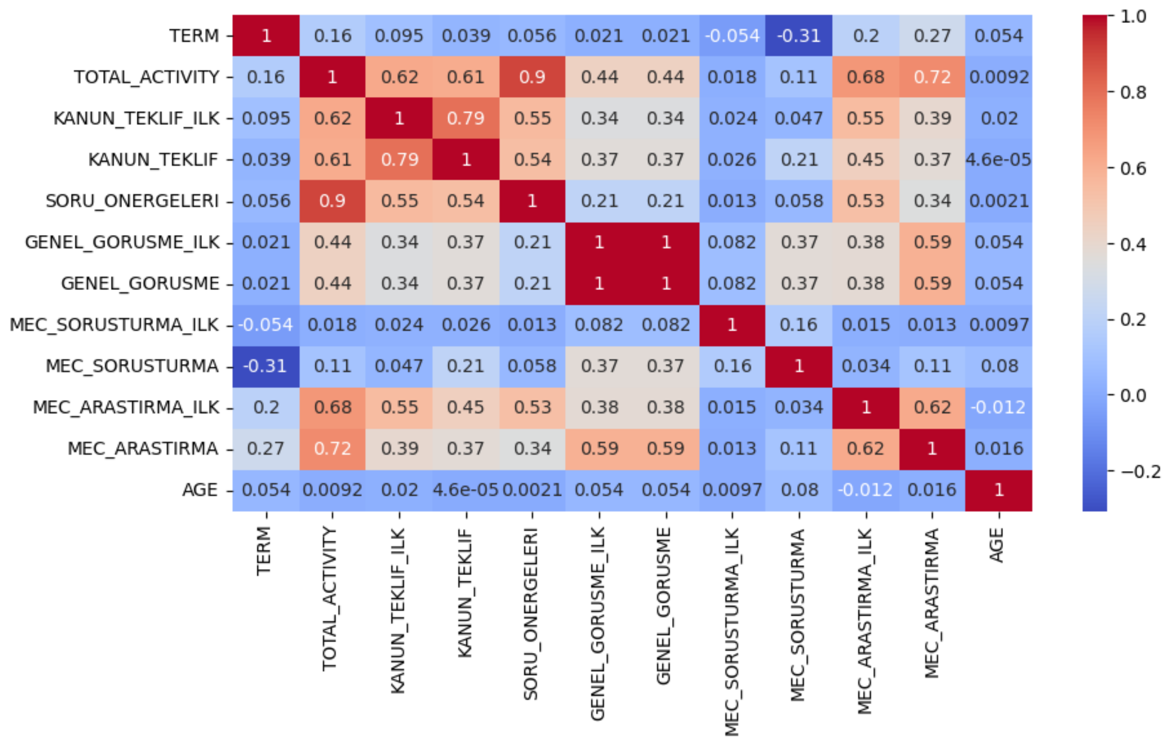

Figure 9: Figure shows the correlation matrix of the attributes.

From this figure, it can be inferred that there exists a positive correlation between Parliementary Questions and Total Activity. This shows that whenever the deputy asks parliementary questions frequently, it means that his/her overall activity is also higher. This can mean that the overall activity is related to it, but also it can mean that the parliementary questions make up the majority of the activity of the deputies.

Also considering the activeness levels of cities, we have this heatmap over the political map of Türkiye, as Figure 10:

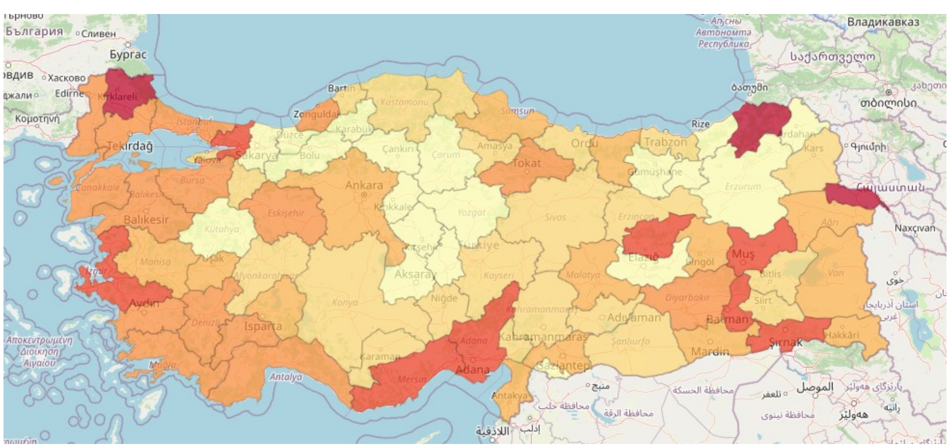

Figure 10: Figure shows the heatmap of mean activity of deputies city-wise.

This heatmap shows that in inner cities, activity gets lower. While sociopolitical factors may explain this phenomenon, it is known that inner cities are generally right-leaning and right-leaning Ak Parti has been ruling since the 22th installment. This can be the reason as to why deputies of such cities not feeling inclined to ask parliementary questions, or be active. It can be expected that opposing parties are better off doing such tasks, as this is done for politically undermining the ruling party.

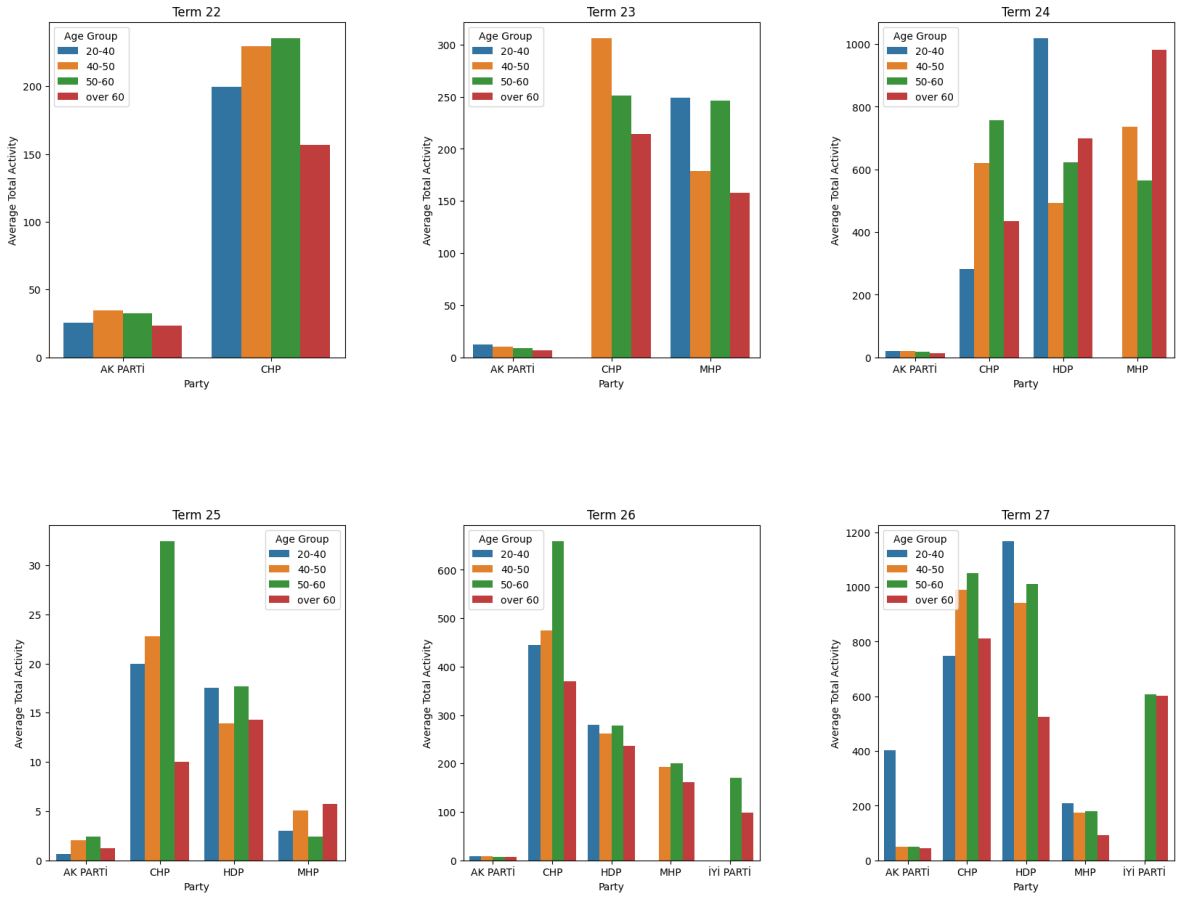

Figure 11: Figure shows the correlation matrix of the attributes.

There are a lot of things to be inferred from this statistic. Especially the cases with the İYİ Parti, HDP and AK Parti. HDP deputies under 40 are generally more active compared to others, whereas in İYİ Parti, we see a non-existence of deputies under 40. Another thing worth mentioning is the sudden increase in activity by AK Parti deputies under 40. This shows that there is

6 Machine Learning Models and Accuracy Results

Due to the complex situation created by the 25th installment of the assembly, two separate classification result sets are prepared for all of the hypotheses. Also, we do not include the `AGE` attribute in our training set because the age values are very close to each other in many sets, and that they do not have any contributions to the valuation process. Also, there's little to none correlation between other metrics. Please note that accuracy values may vary runtime to runtime.

6.1 Political Party Prediction Based On Parliementary Activities

Model	Accuracy
f1-Score of RFC	0.868066
Logistic Regression	0.674662
Naive Bayes	0.559220

Figure 12: With 25th included

Model	Accuracy
f1-Score of RFC	0.870736
Logistic Regression	0.745062
Naive Bayes	0.596050

Figure 13: Without 25th included

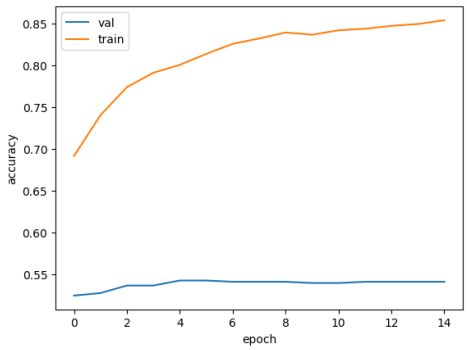

Figure 14: Neural Network With 25th

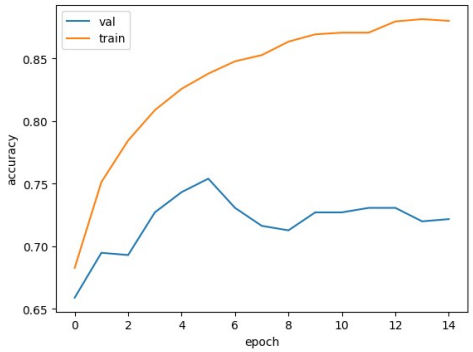

Figure 15: Neural Network Without 25th

It is quite reasonable that the models are having a hard time classifying different parties, since there exists 27 unique parties that had seats since the 22th installment. This is why there exists other labels that may outperform the classification of this particular label. This is why we created two additional labels: `AFFILIATION` and `INCUMBENT`. These are great features since their class are significantly greater than that of parties. And therefore, a more accurate classification will be made with these labels. With these in the features, there was a significant amount of data leakage that ended up increasing the accuracy and f1-scores to absolute statistical significance. In this situation, we performed a feature significance analysis and eliminated these from features.

Also, it seems that the neural network model is suffering from overfitting dearly, in the case of including 25th term of the assembly. This might indicate a lot of things, but firstly, it should be duly noted that the random forest model outperforms the sequential neural network model because neural networks are more susceptible to overfitting compared to random forests, due to the "majority voting" mechanism.

6.2 Political Affiliation Prediction Based On Parliementary Activities

Model	Accuracy
f1-Score of RFC	0.869565
Logistic Regression	0.782608
Naive Bayes	0.655172

Figure 16: With 25^{th} included

Model	Accuracy
f1-Score of RFC	0.879713
Logistic Regression	0.791741
Naive Bayes	0.669658

Figure 17: Without 25^{th} included

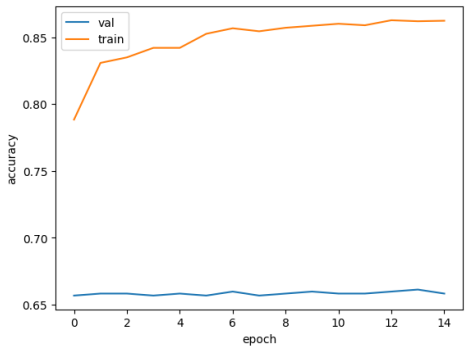

Figure 18: Neural Network With 25^{th}

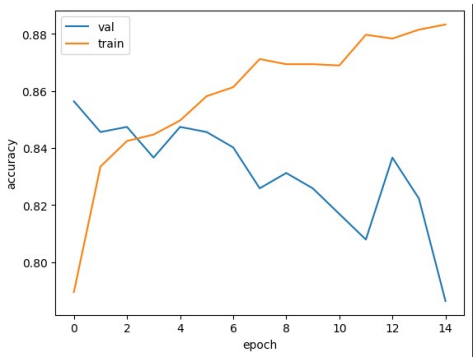

Figure 19: Neural Network Without 25^{th}

This data is very interesting from the perspective of political obligations of MP's. The fact that there's a high level of significance shows that political stances in the spectrum is directly related to how active they are. From looking to any further documentation or outside information, this can infer that there is a correlation in between being the government and having a certain political affiliation.

When we look at the recent history of Türkiye, we can see that there's a reign of AK Parti and sometimes with MHP coalition. Therefore, there exists a near 1.0 correlation between being the government and being a rightist party. However, this correlation is never 1.0 since there exist parties such as İYİ Parti, Saadet Partisi which are predominantly rightist but in opposition.

Also, it seems that the neural network model is suffering from overfitting dearly, in the case of including 25^{th} term of the assembly. This might indicate a lot of things, but firstly, it should be duly noted that the random forest model outperforms the sequential neural network model because neural networks are more susceptible to overfitting compared to random forests, due to the "majority voting" mechanism.

6.3 Political Power Prediction Based On Parliementary Activities

Model	Accuracy
f1-Score of RFC	0.869565
Logistic Regression	0.782608
Naive Bayes	0.655172

Figure 20: With 25^{th} included

Model	Accuracy
f1-Score of RFC	0.935368
Logistic Regression	0.816876
Naive Bayes	0.640933

Figure 21: Without 25^{th} included

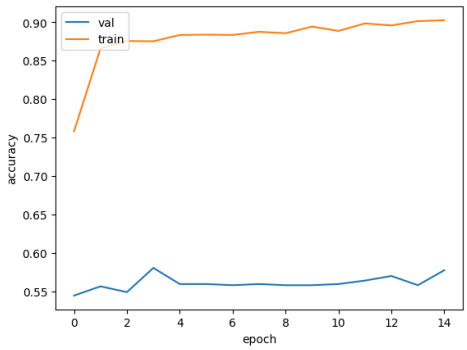

Figure 22: Neural Network With 25^{th}

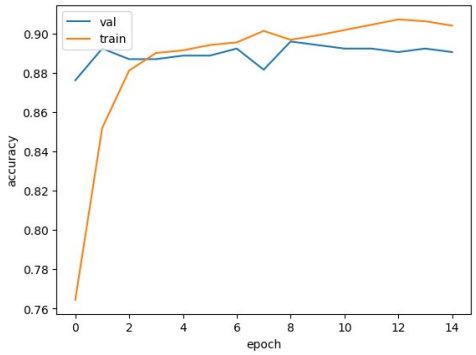

Figure 23: Neural Network Without 25^{th}

It is quite evident that models are more efficient in determining whether the deputy belongs to the ruling party(es). When we look at the models that we trained, it would seem that the model that has 25^{th} installment with it is quite overfit, while the training result of the neural network with the dataset with 25^{th} involved is quite well fit. Also, the f1-score of the random forest classifier is quite high, indicating a near-perfect classifier of deputies.

While the political reasons are evident, this result proves that the overall phenomenon is near statistically significant for the last 5 terms of the assembly.

Also, it seems that the neural network model is suffering from overfitting dearly, in the case of including 25^{th} term of the assembly. This might indicate a lot of things, but firstly, it should be duly noted that the random forest model outperforms the sequential neural network model because neural networks are more susceptible to overfitting compared to random forests, due to the "majority voting" mechanism.