



Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

*Distributed
Computing*



A Machine Learning Analysis of the Swiss Political Spectrum and Candidate Recommendation Process

Group Project

Sebastian Bensland, Karim Saleh and Pietro Ronchetti

`sbenslan@student.ethz.ch`

`salehk@student.ethz.ch`

`pietroro@student.ethz.ch`

Distributed Computing Group
Computer Engineering and Networks Laboratory
ETH Zürich

Supervisors:

Darya Melnyk, Devillez Henri

Prof. Dr. Roger Wattenhofer

July 23, 2020

Acknowledgements

First and foremost we would like to thank Smartvote for providing us with this unique opportunity. We are aware that the dataset they decided to share with us is a significant asset for them, we thank them for trusting us with such precious information, since nothing in this text would have been possible without their support. We would also like to thank our two great supervisors Darya Melnyk and Henri Devillez, they were always there to help us with their experience and to challenge us to look beyond simple solutions.

Abstract

This year, the Distributed Computing Group (DisCo) offered the opportunity to work with an exclusive dataset provided by Smartvote. Smartvote is an online voting advice platform that matches voters to candidates and parties that share their political views on policies. They do that by comparing the answers that candidates and voters give to the same questionnaire on a wide range of policy issues. The more similar the answers, the better the match between voter and candidate.

Smartvote's dataset contains the answers of thousands of politicians and hundreds of thousands of voters to the questionnaire. It provides a unique look into the Swiss political picture.

The goal of this project was twofold: to gain insight into the Swiss political landscape and to refine Smartvote's process. To address the former goal we performed a market fit analysis. It shows how well the Swiss electorate is represented by the various Swiss parties. As for the latter we came up with techniques to reduce the questionnaire size. Wanting to ensure that the impact on the final results caused by the reduction would be minimised, we employed a validation system for our results.

Introduction

Every four years, the national council is elected in a nationwide election by the Swiss people. The national council consists of 200 politicians. Because of the diverse political landscape in Switzerland, twelve different parties are represented in the national council (Nationalrat), as well as one independent politician. In addition to that, there are also cantonal elections. In 2019 for the national council elections, a total of 4652 politicians stood for election. How do voters find their best representative in such a diverse and complicated landscape without having to invest too much time?

Smartvote¹ is trying to solve this problem. Smartvote is an online platform that matches voters to candidates and parties that share their political views on policies. Founded by Politools in 2003 they have been providing their services in over 200 elections throughout Switzerland and around the world.

Smartvote works by letting political candidates answer a questionnaire on a wide range of current policy issues. A voter then answers the same questions on the Smartvote website. Judging by how much the answers from politicians and voter overlap, a ranked list of candidates and parties that best match the voter's political profile is provided.

Smartvote is trying to improve its services all the time. Smartvote's deluxe questionnaire includes 75 questions. It takes a lot of time to go through each question and answer it thoroughly. The user might be put off by the amount of time it will take and might lose interest during the questionnaire and start skipping questions. One huge step to improve their user experience would be to reduce the number of questions used in the questionnaire without losing relevant information about the voter and candidate. In this thesis, we tried different methods to reduce the number of questions and analyzed the information loss. We additionally build custom classifiers to give Smartvote a different alternative and perhaps improve their euclidean distance classifier, which they use to give

¹www.smartvote.ch

the best matching electoral lists for the user.

With a constantly changing political landscape, politicians could use a market fit tool to better align themselves with their voters for future elections. It could also give them feedback on which policies their voters find more important than others. Smartvote could publish this market fit after every big election, giving the voter an opportunity to reflect on their political orientation and get a better overview of the political landscape. After analyzing the data set in Section 3 we performed a market fit analysis between politicians and the voters in Section 4. This represents a market fit evaluation for the 2019 national council election.

Current State of Smartvote

How does the questionnaire actually look like? First, the user has to choose between a deluxe and a rapid version. The deluxe one is made up of 75 different questions. The rapid version contains only 30 (handpicked by Smartvote) of the 75 questions from the deluxe questionnaire. The questionnaire for recording political positions is the central element of Smartvote. It is individually adapted for each election and must meet several quality criteria. Political neutrality and the widest possible coverage of the questionnaire is essential. At the same time, the questionnaire must take into account current political discussions as well as take up topics that will be relevant in the coming years. Finally, the questions have to be formulated clearly and understandably.

1. Welfare state & family (5/6) ▾

1. Do you support an increase in the retirement age (e.g. to 67)?

Yes Rather yes Rather no No No answer

Weight answer: - = +

Figure 2.1: Example of a question in the deluxe questionnaire of Smartvote

Depending on the type of question the user can choose between different answers and not answering at all. Additionally, the user can weight the answer to mark its perceived importance. Before receiving its final politician recommendations, the user has to choose a district. This is because, for example, a person living in Zürich is only allowed to vote for a Representative from Zürich itself.

Smartvote also gives the voter the option to receive a ranking of the top parties matching the user's questionnaire best, as depicted in the Figure 2.3 below.

Your voting advice

Election: National Council ele... | District: Zurich | Candidates | Lists

Additional search criteria

Matching | smartmap | My smartspider | Share | Voting advice (PDF)

	1. Thomas Hug 1991 glp 29.26	<div style="width: 71.7%;"></div>	71.7% >
	2. Ernst Müllhaupt 1950 glp 23.19	<div style="width: 71.5%;"></div>	71.5% >
	3. Ueli Lott 1964 glp 23.25	<div style="width: 70.5%;"></div>	70.5% >
	4. Jonathan Felix Landolt 1998 BDP 07.26	<div style="width: 70.2%;"></div>	70.2% >
	5. Urs Bernasconi 1947 glp 23.31	<div style="width: 70.0%;"></div>	70.0% >

Figure 2.2: The top 5 politicians for the example user

Your voting advice

Election: National Council ele... | District: Zurich | Candidates | Lists

Matching | My smartspider | Share | Voting advice (PDF)

	1. Grünliberale, Junge Grünliberale Participation: 34 / 35	<div style="width: 64.6%;"></div>	64.6% >
	2. Grünliberale, UnternehmerInnen Participation: 31 / 35	<div style="width: 64.2%;"></div>	64.2% >
	3. Grünliberale, senior GLP Participation: 23 / 35	<div style="width: 63.9%;"></div>	63.9% >
	4. Grünliberale Participation: 35 / 35	<div style="width: 63.7%;"></div>	63.7% >
	5. Bürgerlich-Demokratische Partei Participation: 28 / 35	<div style="width: 61.2%;"></div>	61.2% >

Figure 2.3: The top 5 electoral lists for the example user

2.1 Calculation of the recommendations

To remain transparent with their users, Smartvote provides a detailed documentation¹ of the method they use to compute the recommendations on their website. The election recommendation is based on the measurement of the "political" distance or proximity between the candidates and the voters. The distance, $Dist(v, c)$ between candidate (c) and voter (v), is calculated using all (n) questions answered by the voter. For each question (i) the difference between the voter's answer (v_i) and the candidate's one (c_i) is weighted by (w_i) (the weight given by the voter), before computing the Euclidean distance.

$$Dist(v, c) = \sqrt{\sum_{i=1}^n (w_i * (v_i - c_i))^2}$$

To act as a normalization constant, the maximum possible distance between candidate (c) and voter (v) is calculated using all questions answered by the voter. The maximum possible distance is therefore simply the square root of the sum of the weight for every question answered by the voter (v_i) times 100 (If the reader questions where this apparently arbitrary value of 100 comes from we encourage them to hold off until Section 3.1 better describes the question formats).

$$MaxDist = \sqrt{\sum_{i=1}^n (100 * w_i)^2}$$

The last step is to convert the distance into a measure of proximity and display it as a percentage between 0 and 100. This is done by normalizing the calculated distance by the maximum distance, which is then subtracted from 1 and multiplied by 100.

$$Matching(v, c) = 100 * \left(1 - \frac{Dist(v, c)}{MaxDist}\right)$$

The political proximity to the position of a voter is then calculated this way for all candidates. The values can lie between 0 (no overlapping positions) and 100 (completely congruent positions).

¹www.smartvote.ch/method-description

Voters can also have election recommendations made for lists of candidates. The same method is used in principle. For all candidates on a list, the *Matching* value between candidate and voter is computed. The list's final score is then the mean of the list's politicians' *Matching* values.

Dataset Analysis

In this first Chapter, we want to familiarize the reader with what the dataset looks like and what it contains. It is important for the reader as much as it was for us to have a sense of what kind of data we were dealing with before diving into the more technical aspects of our work. In the following sections, we will first give a brief explanation of the dataset's contents, followed by a few important considerations and insights that the reader should be aware of.

3.1 Dataset description

The data package we received includes the collected questionnaires for the 2019 national council election in Switzerland. The package contained two main documents on which we focused our efforts: The candidate and voter datasets containing the questionnaire answers compiled by the politicians and voters respectively. The questionnaire for the voters and the candidates is composed of the same 75 questions (ordered differently). The two datasets also contained the following additional information:

- The candidate dataset includes information about the candidate, such as age, party affiliation, gender, occupation and other details. Additionally, it gives every candidate the option to add a comment on their answers, an option that voters do not have. The dataset contains the answers of more than 4600 politicians from 69 parties.
- The voter dataset, to preserve anonymity, does not include personal details. Voters, as opposed to the candidates, have the option to weight their answer to indicate their perceived importance of the issue.

Smartvote offers two questionnaire options for voters: an integral one (which they labelled "deluxe"), composed of 75 questions and a reduced version ("rapid")

for the voters in a hurry, with only 30 questions. Therefore, the voter dataset is internally split up into two different classes of voters. Overall the voter dataset contains the questionnaire answers of more than 427000 voters. Roughly 63% of these are deluxe and the rest are rapid.

Out of the datasets we extracted both the voter's and politician's answers to Smartvote's questionnaire: the answers are numerical and all on the same scale. Each question can be either answered with 0, 25, 50, 75 or 100, where 0 means no, not applicable or totally disagree and 100 representing yes, applicable or totally agree.

Additional information about the voters and politicians as the zip code, comments, age and gender was disregarded for two reasons:

- The candidate and voter data sets contain different information about the questionnaire taker. Including this information might introduce a different bias into the two data sets.
- Only using the answers to the 75 Smartvote questions gives us a standardized numerical data set on which we could easily implement various analytical algorithms. Taking into account comments or otherwise nonnumerical data would have required an additional error prone effort that was not required for our purposes.

3.2 Dealing with a sparse dataset

Voters as well as candidates are not obligated to answer each question. They have the option to skip some. Surprisingly, there are approximately 25% of candidates who refuse to fill out the questionnaire. Those candidates are still included in the candidate data set, but their answers are left completely blank. Even though Smartvote gives the option to fill out a shorter questionnaire, many voters still decided to skip a large number of questions.

It is interesting to mention that, when looking only at the voters' deluxe version, one can see that there are questions that are skipped more frequently than others. Especially questions regarding the state budgeting are left unanswered. An explanation for this could be, that to assess a budget (e.g. the military one) one needs far more background information on the topic. Another possibility is that for a group of people with a specific political orientation certain groups of questions are not interesting and are therefore skipped. Or even further, someone with a certain socioeconomic background or profession might not be concerned with certain lines of questioning.

All this leads to big gaps in both the voter and politician's data.

We decided to deal with the problem by inserting the constant number value of (50) in place of all "Not A Number" (NaN) values. (50) is one of the possible choices of answers and always the neutral answer at that. This is important as to not affect the data set in an unforeseen way. Filling in the gaps with a constant value simplifies further work without distorting valuable information. This is one of the common practices in data analysis to deal with "Not A Number" entries in data sets.

3.3 Dimensionality Reduction

Once the missing values have been dealt with we are faced with the problem of the datapoints' dimensions. As previously discussed, every voter and politician is represented by a point in a 75-dimensional space.

Visualising the data is impossible in such a space. We therefore decided to apply a dimensionality reduction technique. We considered both Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE). The t-SNE technique works by modelling each high-dimensional point into a two- or three-dimensional one in such a way that similar points are likely to be near each other and dissimilar ones are likely to be far away [1].

Unfortunately t-SNE can easily produce fallacious clusterings. One would need to manually verify the findings to validate the results [2], something we couldn't do since our voter set was unlabelled. Instead, we decided to settle on PCA [3].

PCA works by rotating the 75-dimensional frame of reference in such a way that only a few of the new orthogonal basis vectors contain most of the information about the data points. We can therefore only focus on the "Principal Components" and discard the remaining dimensions without fear of losing a lot of information. Because of this relatively easy concept, results can be easily interpreted and performed on both voter and politician sets.

After performing PCA on the politicians' dataset we project the politicians onto our new coordinate system. We then plot the datapoints along the first two principal components (PCs). The two-dimensional representation of the politicians in Figure 3.1 is the one that loses the least amount of information.

An intuitive explanation on how PCA can represent a data point in so few dimensions without losing too much information: One could predict what kind of answers a candidate would give to certain questions by knowing their political

alignment (For example a right-wing politician could be less open to immigration than a left-wing one, without explicitly knowing their answer).

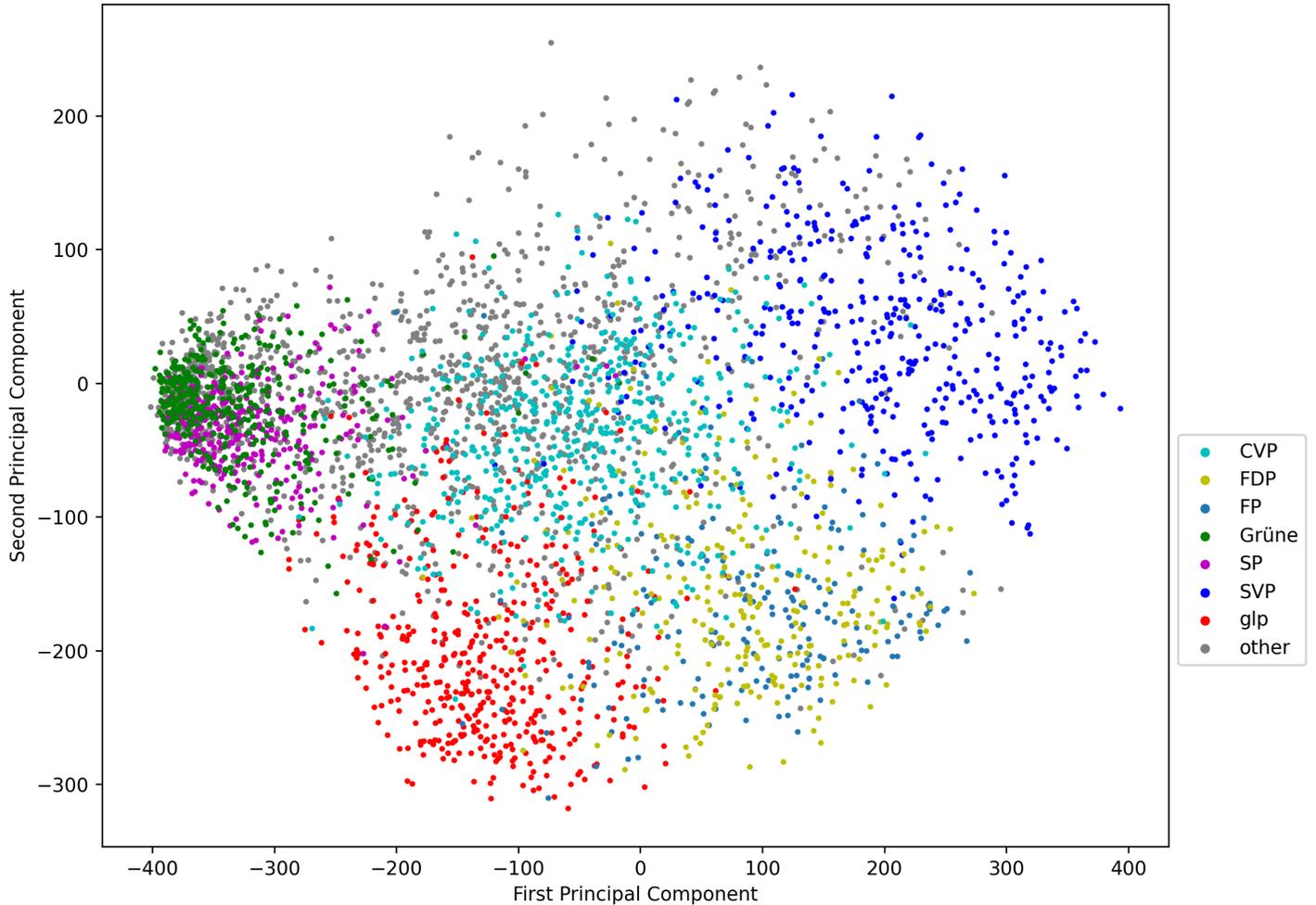


Figure 3.1: Politicians projected onto the first two Principal Components, coloured by party

Figure 3.1 shows politicians' views in a two-dimensional space. The position on the two axes represents their political alignment. Deconstructing the meaning of the two axes by using the individual components of the two PCs would be a very tedious task. Luckily the politician data is labelled, therefore we can infer the meaning by looking at where the ideologies of the different parties are most represented. For example we see how Grüne and SP tend to cluster up on the left side of the plot, whereas FP, SVP and FDP take up the right. From this we can assume that the first axis gives an indication to the left/right inclination of a party. Based on this we can see that the left-oriented parties are more compact

and that there is a lot of overlap in their ideologies (the difference between Grüne and SP is almost negligible). The parties on the right on the other hand tend to be more sparse and cover more ground.

The first PC explains 37% of the total variance in the answers, while the second PC already explains only 9%, the third PC 5% and the fourth 3%. This means that a more meaningful interpretation of the other directions is less likely.

Performing PCA on the voters on the other hand yielded results that were harder to interpret. This is mainly due to the fact that voter data is unlabelled. Figure 3.2 represents the projection of the voter dataset onto the newly computed PC basis for voters. We can notice how in this case the voters are more concentrated on the left side of the plot. Unfortunately, because the PCs are not the same as the ones we computed for the politicians, we cannot make the assumption that the dense spot corresponds to left-wing voters.

One insight that Figure 3.2 provides is that there is a very clear separation between deluxe and rapid questionnaire takers. Given the higher percentage of deluxe takers and the missing data in the rapid part we decide to only use the deluxe part going forward. No clear clusters seem to emerge from the PCA, which suggests that the voters tend to be more homogeneous and less polarized than the politicians. Given the lack of meaningful results on the voters' political we decided to apply some clustering techniques to better understand the voter set.

3.4 The "elbow" method

One important aspect of analysing the data set is understanding how the data is clustered. Before finding any kind of clusters we needed to find out what the ideal number of clusters would be. We chose to solve this problem by using K-Means clustering to partition the voters into their respective groups, since the algorithm aims to choose centroids which minimize the inertia or within-cluster sum-of-squares (see below). It was for us the ideal choice, since it would group voters with similar answers together. The K -means algorithm aims to choose centroids which minimize the inertia

$$\sum_{i=0}^n \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

To identify the ideal number of clusters we have implemented the "elbow" method [4]. The "elbow" method helps us to identify the optimal number of clusters into which the data may be clustered by fitting the K-means clustering model with a range of values for K . For each K , we calculate the sum of squared

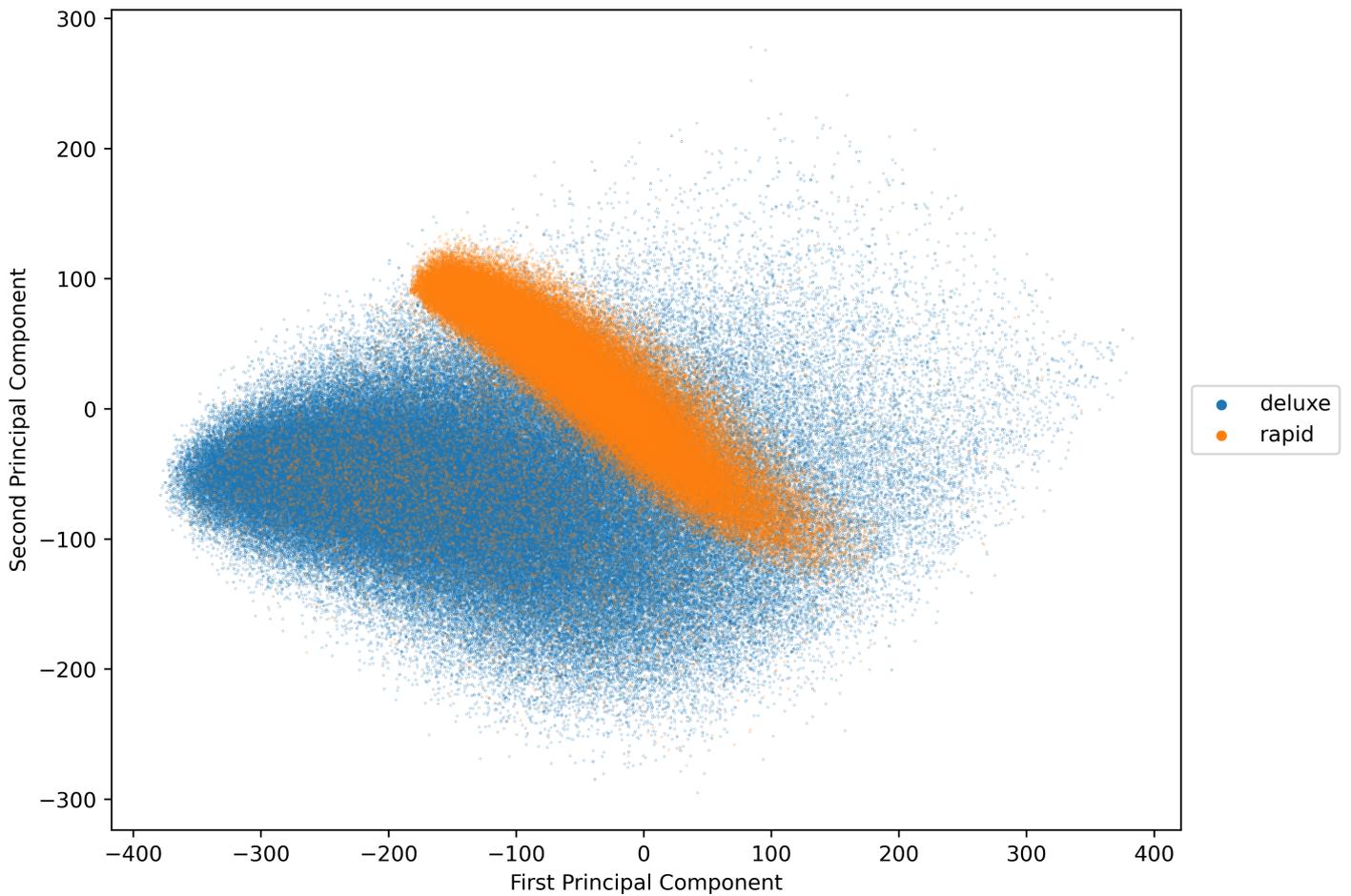


Figure 3.2: Voters projected onto the two principal components, coloured by the questionnaire type

distances of samples to their assigned cluster center also referred to as the inertia. Figure 3.3 displays this method for the voters data.

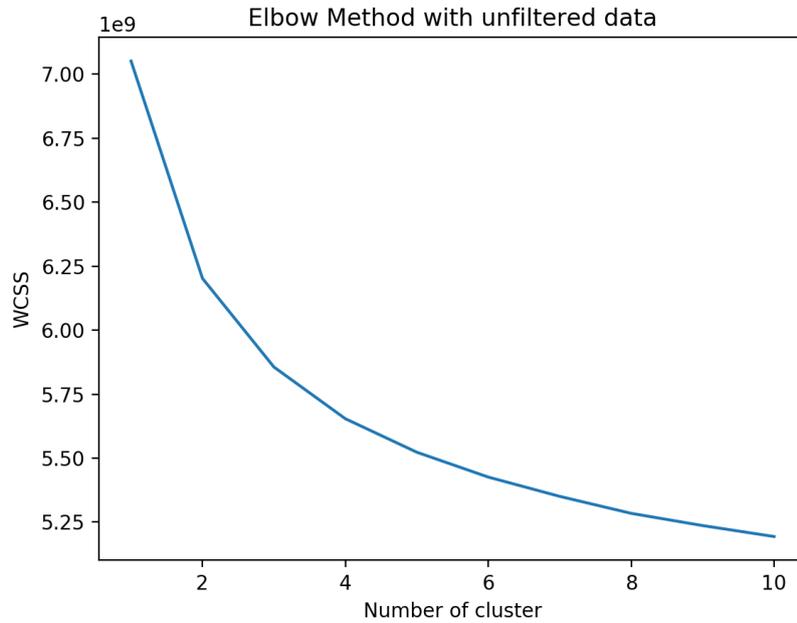


Figure 3.3: Elbow method for K in range [1,10]

To determine the optimal number of clusters, we have to select the value of K at the “elbow” i.e the point after which the inertia start decreasing in a linear fashion. Thus observing Figure 3.3 we conclude that for the given data set the optimal number of clusters is approximately around 4. This means that based on the voters answers to Smartvote’s questionnaire, the voters can be partitioned into 4 groups. More information about the nature of these clusters will be given in Chapter 4.

Market Fit Analysis

This chapter focuses on the market fit analysis, also known as product-market fit. It is the degree to which a product satisfies a strong market demand. If we treat representation as a product supplied by politicians and demanded by the voters we can perform the same analysis on the "market of political views". In essence, we want to see which ideologies are better represented (politicians and voters whose political orientation matches well) and which are less well represented (less overlap). Please keep in mind that in reality things are more complicated. A politician cannot just change their opinion just to cater to more voters. Nonetheless this analysis is interesting, as it shows the gaps in ideology coverage.

4.1 Clustering the voters

In section 3.4 we determined that the optimal number of clusters in which the data should be clustered is 4. Therefore, to partition the voters into their respective groups we ran the K-Means Clustering algorithm with $K=4$. The K-Means algorithm returns 4 centroids, one for each cluster, with answers also ranging from 0 to 100. Analyzing the answers of each centroid, enables us to quantify where each of the 4 voter clusters stands towards the questions asked by Smartvote. This analysis gives us further understanding of how the 4 different voter groups are positioned on the political spectrum, since Smartvote's questions tackle current policy issues.

Figure 4.1 depicts the answers of each of the 4 cluster centers (red, blue, green, orange). For each question (the x-ticks represent the question ID) we plotted the mean answer of each cluster in the corresponding colour. Looking at Figure 4.1 one could easily notice that the blue and red clusters are usually at the opposite end of the spectrum, which suggests an opposing political ideology. These cluster centers give an overview and insight into each voter group, which can be used to

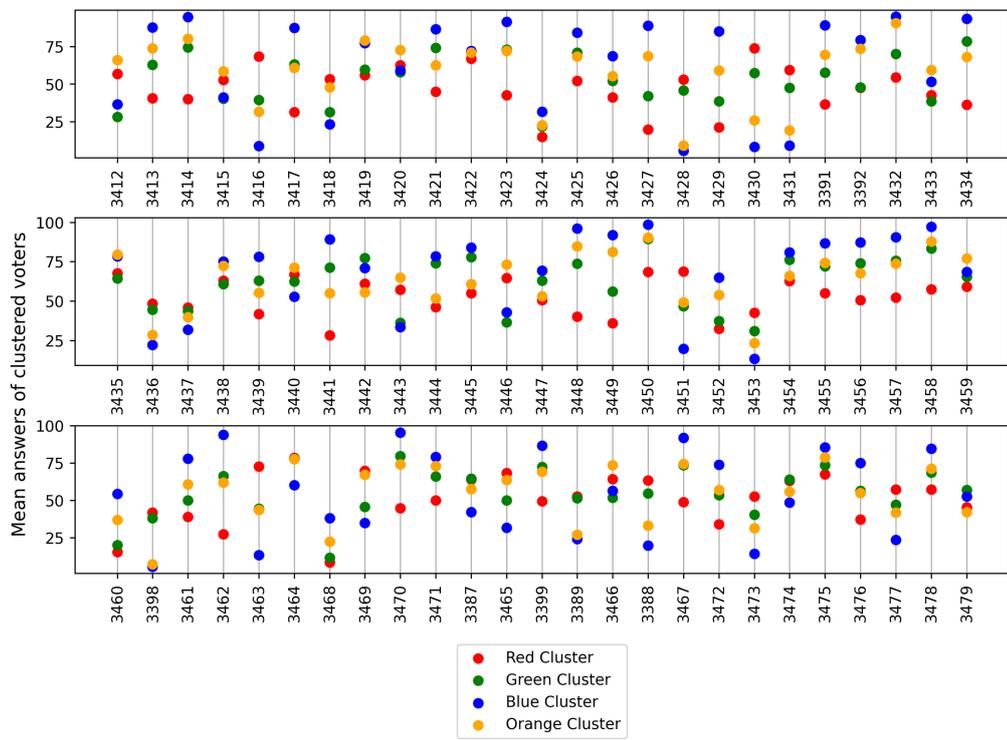


Figure 4.1: Visualising the clusters' political position by plotting the mean answers to questions with ID x

compare them with the political parties in Switzerland and determine a market fit.

4.2 Analysing the parties' coverage of voter base

We then want to see how well these four clusters are represented by Swiss parties. The four clusters represent four unlabelled voter groups (red, blue, green, orange) that share the same political views. To find an overlap between those groups and the political parties that should represent them we compute the euclidean squared distance between the party centers and the voter cluster centers. This gives us an insight into which cluster shares the same ideology of which parties.

Computing the party centers for the main parties yields the results in Figure 4.2. (Because of the little differences in their party centers shown here we decided to bundle Grüne and SP together)

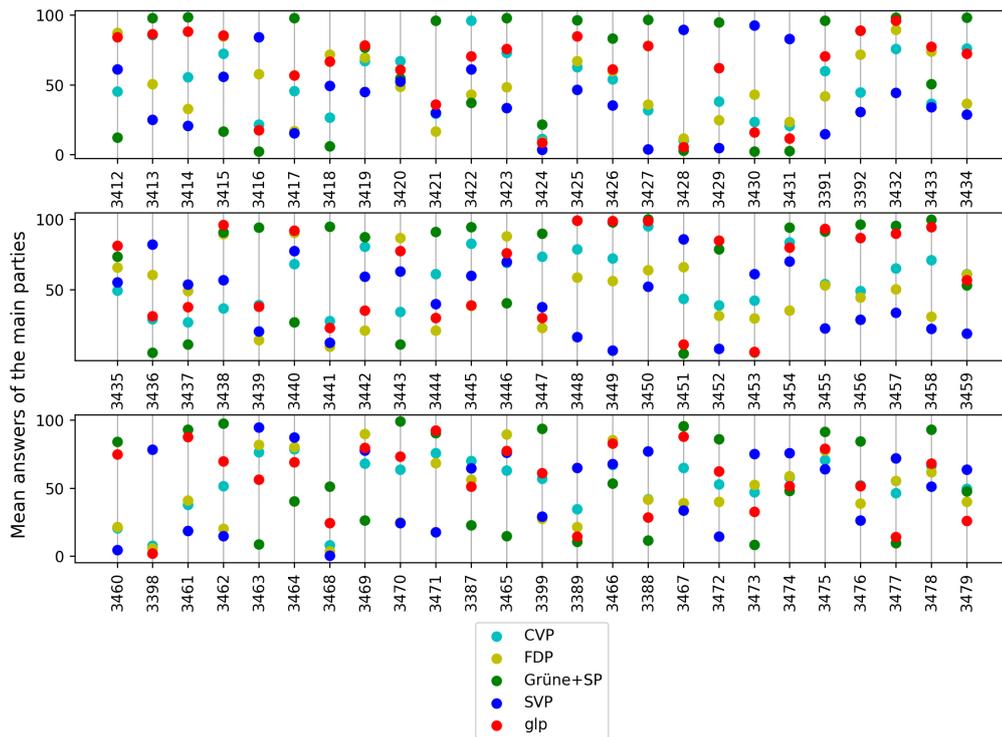


Figure 4.2: Visualisation of the mean value of the main parties' answers for every question

The party centers represent the mean value of the answers given by all members of the main Swiss parties. As in the previous section (4.1), we can draw

some interesting conclusions based on the party centers relative to each other. We notice how, with a few exceptions, Grüne+SP and glp tend to find themselves on the same part of the spectrum. The CVP tends to be the more moderate party: its mean answer is rarely the one at the edges. We can also see how the SVP's answers are in contrast with Grüne+SP's ones most of the time. All these findings seem to accurately depict those parties' political positions.

Now that we have an idea of how both the voter cluster centers and party centers relate to one another we need to forge a connection between the two. To do that we compute, for each cluster center in Figure 4.1, the euclidean distance between it and the party centers in Figure 4.2. That gives us a good idea of how well the ideologies of parties and clusters overlap. A good overlap occurs when the cluster center is near a party center and ideally far away from the others. To get more insights from the plots we inverted the distance, so that the highest bar in Figure 4.3 intuitively signals the best fit.

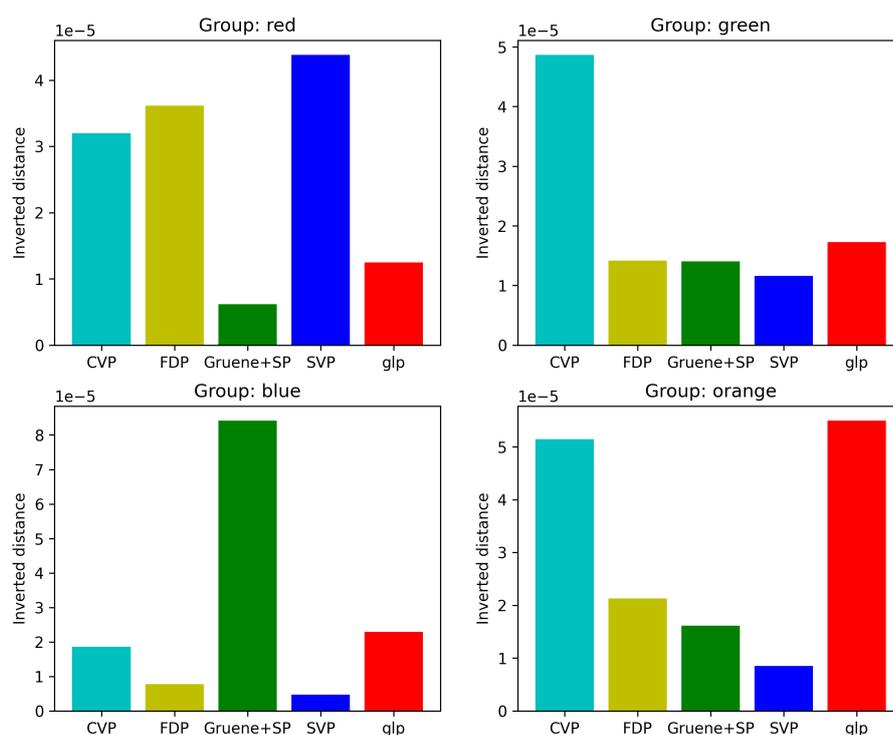


Figure 4.3: Ideology overlap between parties and clusters as shown by the inverse of the distance between the clusters' and the main parties' centers.

As we can see in Figure 4.3 group blue has a very convincing overlap with the Grüne and SP parties. This means that the distance between the blue cluster's and the parties' center is very low. Our interpretation is that the blue cluster represents the left/liberal voter group and that those two parties have a pretty

tight grasp of their voters. The same can be said about how the overlap of group green and CVP allows us to label the green group as the right/liberal voter cluster.

On the other hand the red and orange groups do not quite have the same overlap with a single party. This means that:

- The other parties are not quite representing the two other major clusters and they end up having to split the voters that make those up.
- When looking for a representative the voters out of the red and orange clusters are not going to be satisfied with the available options.

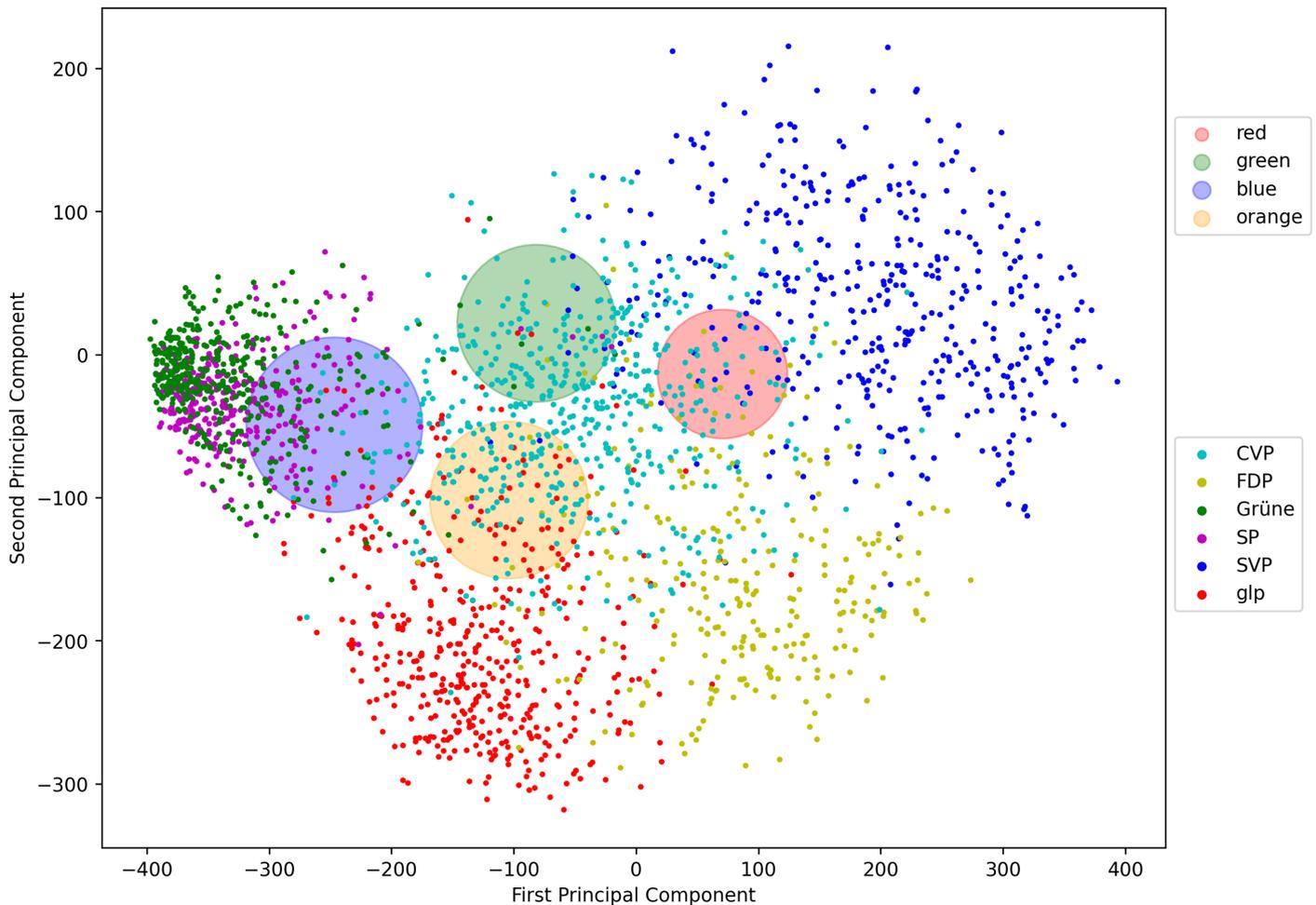


Figure 4.4: Overlaying the PC politicians representation with the voters' cluster centers

This point is further illustrated in Figure 4.4, where we plotted the cluster centers from Figure 3.1. The circle sizes are proportional to the amount of voters contained in the cluster. We can see how the voters tend to be biased towards the left side of the spectrum.

Question Reduction

To further enhance the user experience it is vital to optimize the quality and reduce the amount of unnecessary questions in the set. Additionally it could improve the quality of collected responses by the user.

To give Smartvote valuable information about how to reduce the number of questions they ask in their questionnaire we followed a two step approach:

- We first aimed to identify the questions which are most important. We extracted a few question rankings based on different analytical methods. The questions are ranked in terms of how useful for classification they are. This first step is outlined in Section 5.1.
- To validate the practical usefulness of our rankings we had to get a bit creative. Our goal was to demonstrate that even when taking into account only the most important questions (as given by our rankings) it is still possible to accurately identify the questionnaire-taker's political stance. We therefore had to use some kind of classifier and test its accuracy with the reduced questionnaires. Since we didn't have access to Smartvote's algorithm and the voters' data set was unlabelled, we decided to perform various supervised learning techniques on the politician data set. In this analysis, the answers were considered to be the features and the political party of the candidate the label. For each technique the full set of features was considered as the baseline performance against which we could test the jumps in accuracy when adding questions. This approach enabled us to observe at which point the accuracy surpassed a specific threshold and thus estimate what the minimal number of questions is. This is discussed in Section 5.2.

5.1 Ranking question importance

We explore different methods to extract question rankings that we will then test with different classifiers for their effectiveness.

5.1.1 Principal Component Analysis - Explained Variance

One way of ranking the importance of the individual questions is the following: Ideally we would like to avoid asking questions that are answered in the same way by most people, since this does not really tell us anything about their individual political orientation. Therefore we should try and focus on questions that clearly polarize the voters.

This is where we introduce the concept of variance within the answers. An answer with high variance is interesting, since it allows us to split the voter base in two or more groups on that dimension. That was the basis of our reasoning when applying this method.

Fortunately we have already been using the same reasoning when applying Principal Component Analysis in Section 3.3: PCA is based on the co-variance between the answers. How likely is someone who gave answer A to question x to give answer B to question y ? The PCs are ranked using their Explained Variance. In other words: it is how much of the total variance is explained by each of the PCs with respect to the whole variance of the dataset. Therefore, the PCs that explain more variance are the ones that contain more information about a voter's political orientation.

We therefore first scale the PCs by their explained variance. We then add up all scaled directional components (The components of the Principal Components if that makes more sense) to look at the Explained Variance contribution of the single answers. This gives us an idea of how much variance each question carries with it. For politicians, when performing this method taking into account all PCs, we get the importances outlined in Figure 5.1.

We can see how for example the last few questions seem to be less important overall, mainly because a question that is answered in the same way by everyone is less useful for classification. These are the answers to the budget questions, which seem to be the least polarizing among politicians.

Scaling all principal components by their explained variance and summing their scaled directional components seems like an overly complicated process. After all, the resulting "explained variance per question" is going to be proportional to the variance of the answers, so why not just compute that?

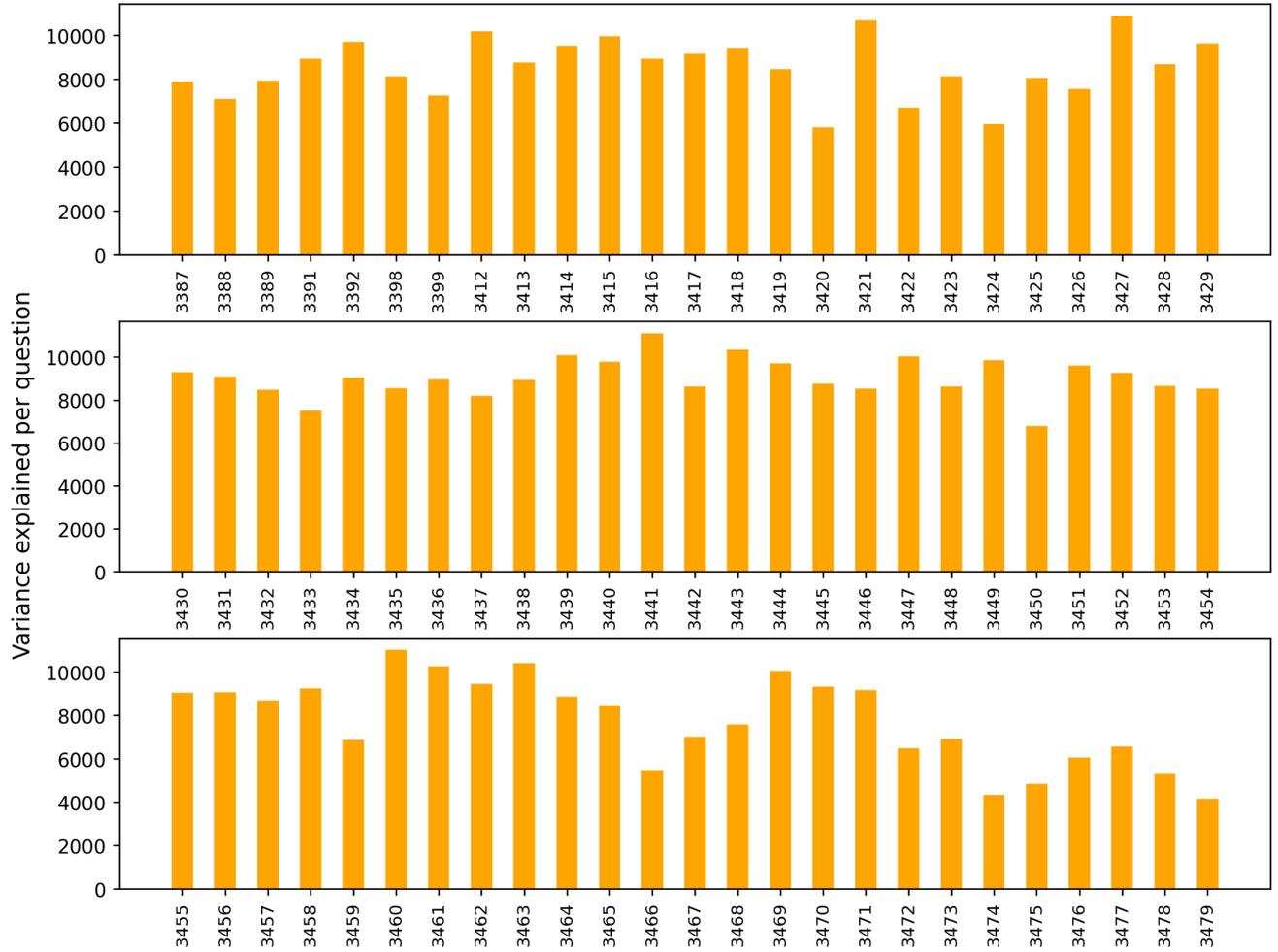


Figure 5.1: Amount of explained variance per question for politicians

While true, our more long winded approach enables us to consider the contributions of only a few PCs, essentially including only the most important principal components are included in our search. This gives an insight on which questions built the axis on which we can differentiate most voters in the best way.

In Figure 5.2 the orange bars tell us that using only the first 15 components the most important questions from before are still the most important, but the relative importance is greater between answers.

To the orange bars we compare the blue ones. These are the result of applying the same process (only take into account the first 15 PCs) on the voters. Their order differs slightly, which means that the most polarizing answers for voters are not necessarily the most polarizing answers for politicians. Figure 5.2 also shows that the explained variance seems to always be a bit higher for politicians than for voters. This is due to the fact that the first 15 PCs of the politician dataset contain more relative variance than their voter counterparts. Which in turn supports our theory that the voters tend to be more homogeneous and less polarized than the politicians.

From these results we extracted two lists in which we rank the question importance based on the criteria outlined in this section. In Section 5.2 the two lists are named "politicians" and "voters" respectively when testing them for their effectiveness. In the next Section 5.1.2 we explore another way of ranking questions using decision trees.

5.1.2 Decision Trees

Decision trees are one of the most powerful machine learning algorithms. Compared to different machine learning algorithms, they can be easily visualized so that a human can understand what's going on. Imagine a flowchart where each level is a question with a simple yes or no answer. In the end, you get a solution to the initial problem. The idea here is to use the decision tree to get a ranking of the questions, which we then can use for our Classifiers in Section 5.2.

The decision tree analyses a data set to construct a set of rules, or questions, which are used to predict a class. In our case, we are working with the politician questionnaire from Smartvote. Each politician has to answer 75 questions, which gives us the 75 features to differentiate the politicians. The problem for the decision tree algorithm is stated as follows, classify each politician into one of the six biggest parties. The biggest parties are chosen by the number of politicians they have in the dataset. Politicians with a different party are removed from the dataset. This leaves us with around 2500 Politicians. The dataset then can be split up randomly into a train and test set. The test set is used to build the

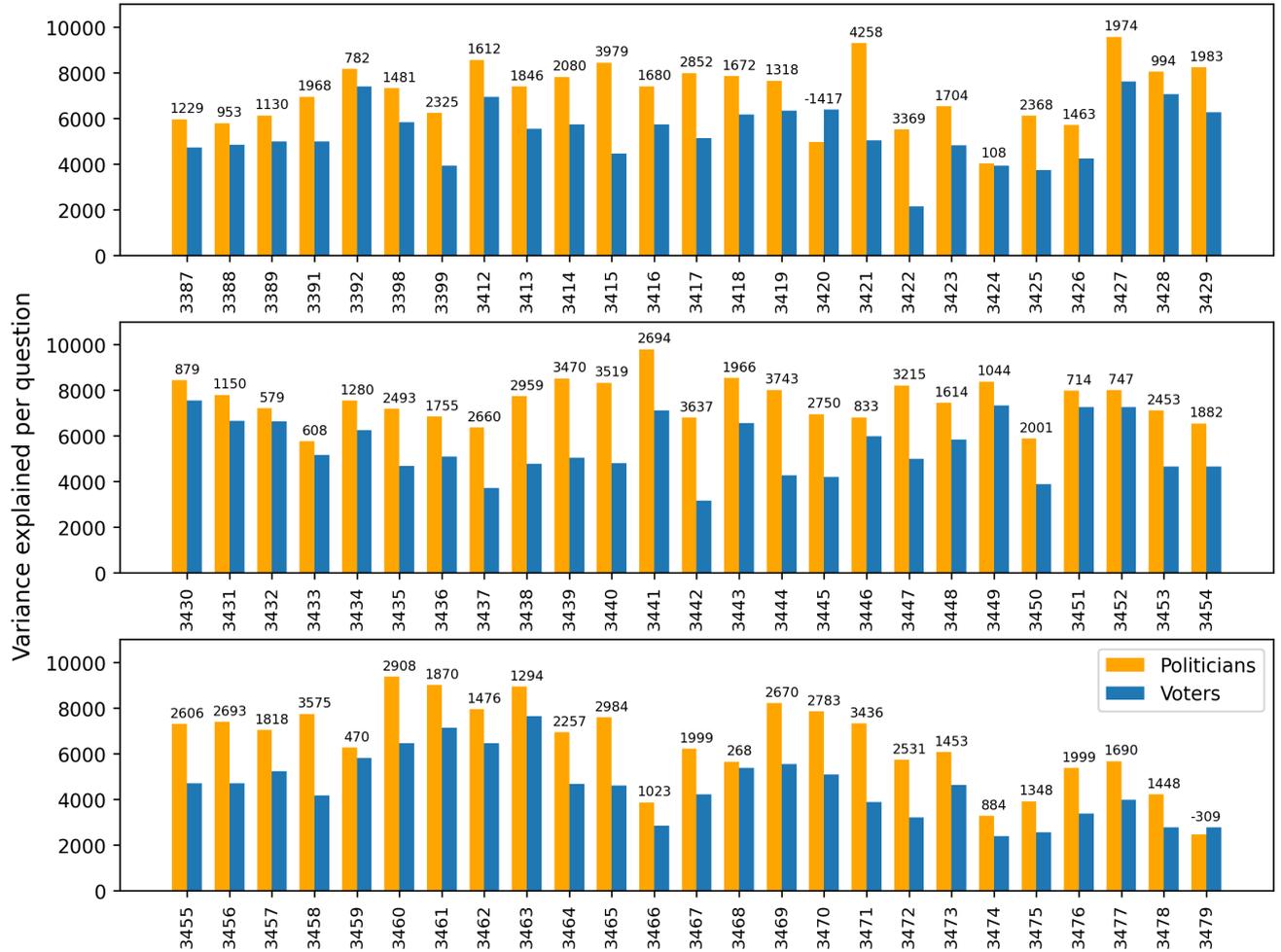


Figure 5.2: Explained variance per question, differences between voters and politicians, normalized to account for greater number of voter datapoints

decision tree and the test set can then be applied to give an accuracy score for the final classification model. In Section 5.2 it is explained why the classification problem was chosen in this manner.

The Algorithm starts with all samples in the top node. It then searches for one feature to split the node into two children. This feature is chosen by the amount of impurity it reduces. The Impurity in our case is the probability of allocating the wrong party to a random politician inside the node.

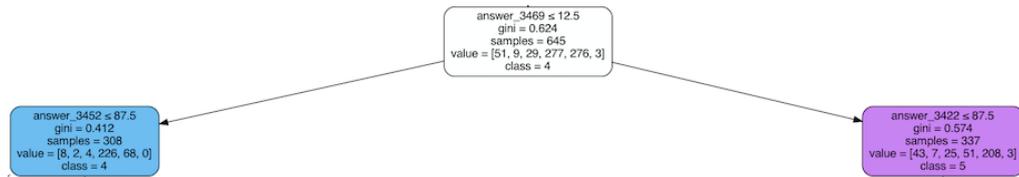


Figure 5.3: Example of one step in the decision tree algorithm

In figure 5.3 one step of the decision tree algorithm is depicted. There are 645 samples in the parent node. The value shows, how many politicians of each party are situated in this node. This means that there are 51 from CVP, 9 from FDP, 29 from glp, 277 from Grüne, 276 from SP, and 3 Politicians from SVP inside this node. You can see that the Gini impurity index shows a value of 0.624. This means that there is a 0.624 probability of choosing the wrong party for a random politician of this node. Now it selects a question that reduces the Gini index the most. In this feature, answer-3469 was selected. If a politician has a value of smaller or equal to 12.5 for question-3469, the politician goes inside the left child. If the sample has a greater value it goes into the right child. Now the process starts again for the child nodes. Where previously two classes were dominant inside of one node, the algorithm managed to separate the two classes 4 and 5 from each other. This process will go on until no further information can be gained or a preset rule is met, e.g. maximum depth is reached. In the end, the algorithm returns a nicely documented binary tree depicted in figure 5.4. The entire decision tree can be viewed in detail in the Appendix A.

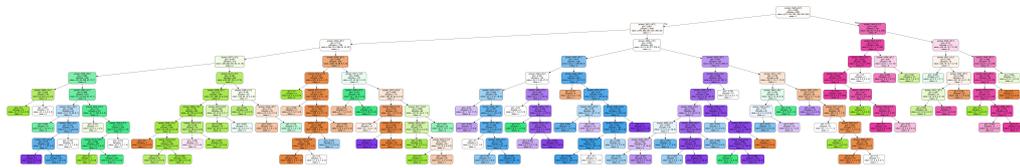


Figure 5.4: Final decision tree for the classification problem of the 6 biggest parties, full size figure in Appendix A

After creating the decision tree, we then can calculate a ranking of the questions as follows. First, we have to calculate the importance of each node. Every single node accounts for one specific question used in this node to further create

the next two children.

$$n_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)}$$

The importance of $node_j$ calculates itself from the total number of samples inside the node times C_j the Gini impurity value of the $node_j$ minus the amount of sample in the left child $w_{left(j)}$ times the impurity of the left child $C_{left(j)}$ and minus the same for the right child. This is done for every node in the binary decision tree.

Because each question can be used not once but multiple times, we have to account for that in our calculations.

$$q_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} n_j}{\sum_{k \in all\ nodes} n_k}$$

The importance for a single question q_i is calculated as the fraction of the sum of every node importance, where this question is used $\sum_{j:node\ j\ splits\ on\ feature\ i} n_j$, divided by the sum of every node importance in the whole decision tree $\sum_{k \in all\ nodes} n_k$

This leaves us with a value for each question used to build the binary tree. The question with the highest value is the most important question to classify a party for a politician.

5.1.3 Classifier Specific Rankings

The final ranking we will test is a bit different from the ones we have seen before. It is based on the performance of one of the classifiers instead of mathematical methods.

When using the Explicit Classifier (Sec. 5.2.2) we discovered that we could rank the question importance according to the performance of the classifier itself.

It was computed in the following way: After having calculated a baseline accuracy we would drop each of the questions from the dataset individually, then rerun the classification algorithm on the reduced dataset. With a question not taken into consideration we would register a drop (sometimes even a jump) in accuracy. We then used those variations to compute an optimal question importance ranking, where the most important questions were the ones that had the least negative impact on the accuracy.

Out of curiosity we also validated this ranking using the classifiers. The tag we chose for it is "custom", since it consists of the optimal solution for only a very specific classification algorithm.

5.2 Classifiers

In this section we analyze how different classifiers perform when the number of questions used as features is increased. The goal of such an analysis is to show that the number of questions Smartvote uses can in fact be reduced and to validate the ranking of questions we obtained in the section above.

5.2.1 Support Vector Machines

In the first subsection of Section 5.2 we take a closer look in to how the different question rankings affect the performance of Support vector machines (SVMs) classifiers.

SVMs are a set of supervised learning methods used for classification, regression and outliers detection [5]. One of the key advantages of SVMs is that they are very effective in high dimensional space, which makes them very effective with Smartvote's data set. To analyze how the number of questions affects the performance, we developed a model which assigns every politician a party based on his/her answers to Smartvote's questionnaire.

A challenge which we encountered in doing so is the fact that the data set is unbalanced; there are 69 parties of different sizes from which 50 percent do not have more than 5 members. Since we do not care about the true classification of every politician to his party but rather the change of performance in terms of the number of features, we worked around this problem by analyzing 3 custom cases. For each case, we added questions in 5 different orders all based on question importance from most to least important. These 5 orders were determined by the rankings of; the Decision Tree algorithm Section 5.1.2 ("decision tree" label), PCA analysis of the voter data ("voters" label) Section 5.1.1, PCA analysis of the candidate data Section 5.1.1 ("politicians" label), our custom algorithm in Section 5.2.2 ("custom" label) and a random ordering ("random" ordering):

- Using only the candidates of the 6 largest parties in Switzerland (CVP, FDP, glp, Grüne, SP, SVP). Figure 5.5 depicts how the accuracy for the 6 class classifier changes, when the number of questions is increased.

Setting the acceptable threshold at -10 percent from the baseline accuracy

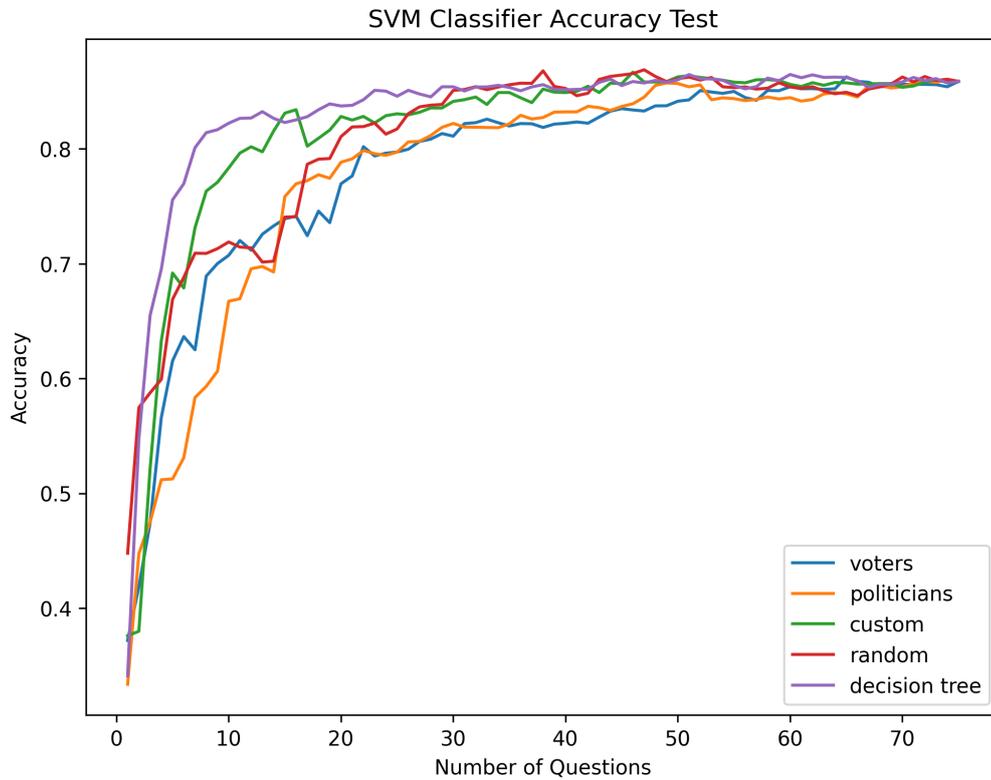


Figure 5.5: Accuracy in terms of the number of questions used for the 6 class classifier

we observe in Figure 5.5 that for the 6 class classifier the minimum number of questions acceptable is around 15 questions. 15 questions would be sufficient enough only if they're chosen based on the ordering obtained by the Decision Tree algorithm. From Figure 5.5 it becomes clear that the questions used influence the accuracy and that the Decision Tree algorithm performs best in identifying the most important questions. This is a clear improvement, to understand if further improvement is possible we took a look at the confusion matrix of this model. A confusion matrix also known as error matrix, is often used to describe the performance of a classifier. The number of correct and incorrect predictions are summarized with count values and broken down by each class, providing us with information about which classes our model miss classifies the most. Figure 5.6 visualizes the confusion matrix for the 6 party classifier.

In Figure 5.6 it becomes clear that the highest miss classification rate is between CVP - FDP and Grüne - SP.

- Due to the fact that Grüne and SP are very similar in their political ori-

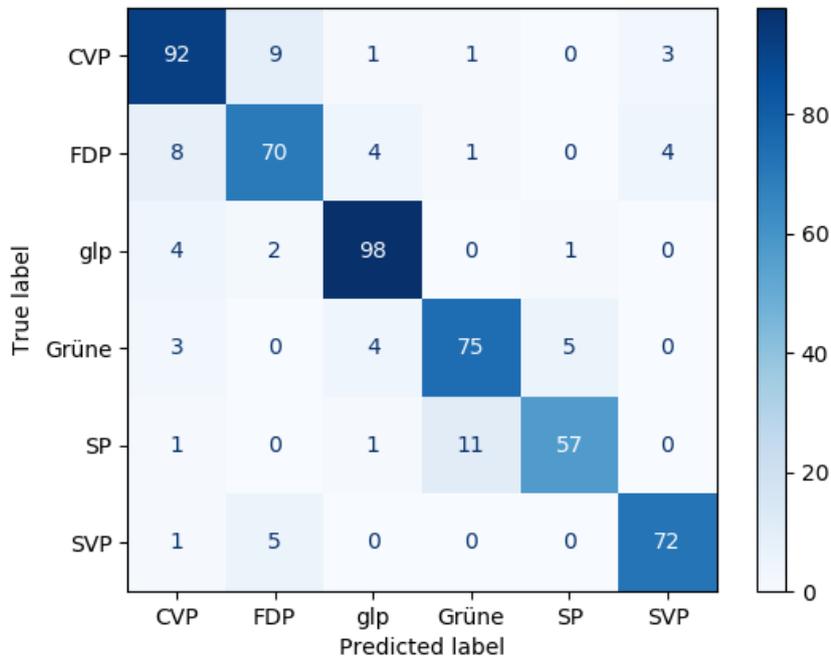


Figure 5.6: Confusion matrix of the 6 party classifier

entation and Figure 5.6 displayed their miss classification, we chose the second case to be a 5 class classifier with CVP, FDP, glp, SVP and SP and Grüne combined. Figure 5.7 depicts how the accuracy for the 5 class classifier changes, when the number of questions is increased.

Figure 5.7 shows that combining SP and Grüne improves the performance drastically and gives a new minimum for the number of question at 10 questions. This new minimum already captures the accuracy of the full data set.

- To further analyze how the classifier preforms on SP and Grüne we used the SP and Grüne classifier as the third case. Figure 5.8 depicts how the accuracy for the 2 class classifier changes, when the number of questions is increased.

Figure 5.8 shows that the SP and Grüne classifier preforms very badly and that the accuracy doesn't converge. This finding implies that using Smartvote's current set of questions it is hard to distinguish between the Grüne and SP and a small reduction of questions would make this feat even harder.

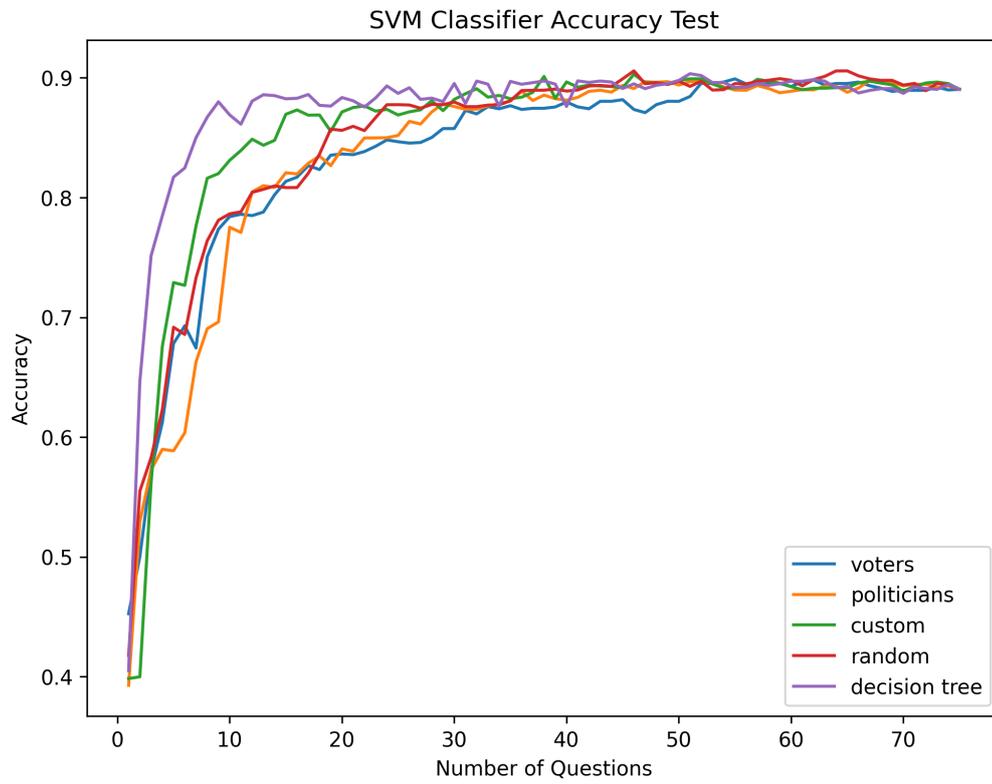


Figure 5.7: Accuracy in terms of the number of questions used for the 5 class classifier

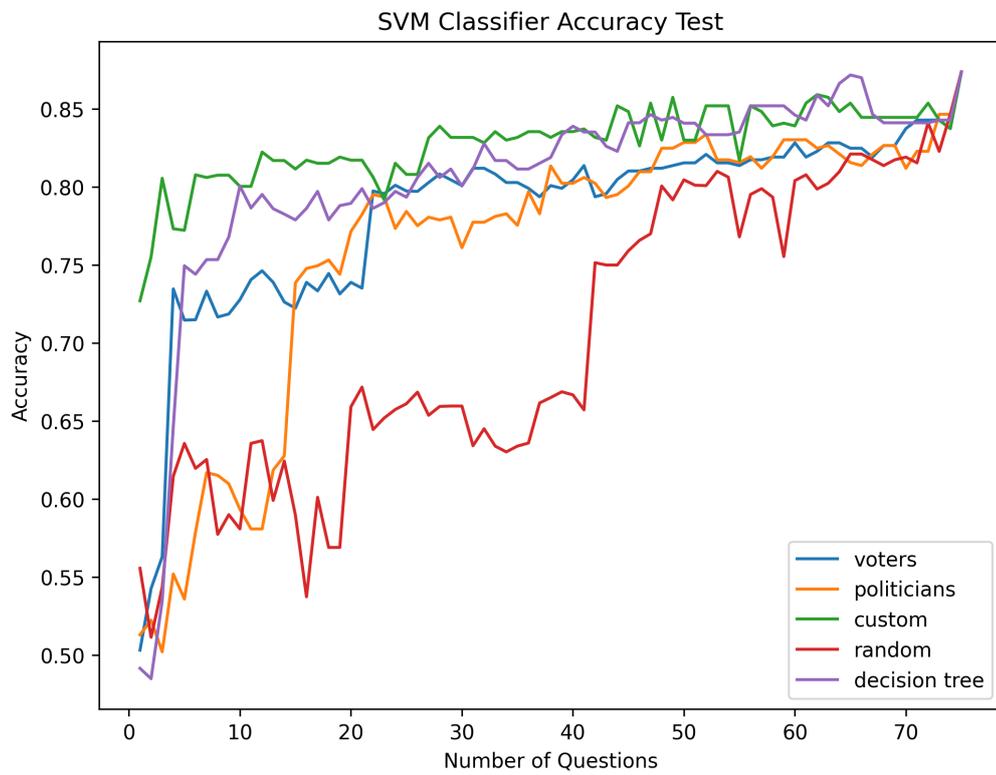


Figure 5.8: Accuracy in terms of the number of questions used for the 2 class classifier

Summarizing the performance of the different SVM classifiers, we conclude that the number of questions can be reduced by more than 60 percent. This is mainly due to the fact that the most important questions hold the largest weight in the decision making. The ranking given by the decision tree algorithm consistently outperforms the other rankings, except when differentiating between Grüne and SP. This is probably due to the decision tree deciding to split the parties into a right/left scheme before differentiating two parties that are very similar. We have also noticed that to be able to differentiate between Grüne and SP and thus their politicians, which is essential for Smartvote to give accurate recommendations, new questions need to be identified. These questions should tackle the main policy differences between SP and Grüne. In the following subsections we enquire if our Explicit Classifier (Section 5.2.2) and the K-Nearest-Neighbour classifier (Section 5.2.3) yield a similar conclusion.

5.2.2 Explicit Classifier

Support Vector Machines, while very efficient, mostly deliver "Black Box" results, where the decision boundaries are optimised but often difficult to interpret.

Since we wanted a bit more transparency on how our classification method operated we also decided to design a classifier from scratch.

The idea behind it is that entire parties can be represented by a 75-dimensional point that represents the average answer given by the candidates of that party to every question. To make the representation clearer we can also include the standard deviation for every question. That way, a party can be represented by a 75 dimensional line instead of just a point. In Figure 5.9 we give the reader a graphical example of the explicit classification process. We illustrate it with the aid of two fictitious parties and their answers to six hypothetical questions.

The blue area's (representing *Party_a*) center line is the party's mean answer and its bounds are given by the standard deviation of all answers by the members of the party. When looking at an individual that needs to be classified (in the example represented by the red curve) the classifier computes a cost for each party and chooses the party with the lowest cost. The cost is increased heavily every time the red curve exits the predefined party area, otherwise it only incurs a small penalty proportional to the distance from the mean answer. In the example we can see that the individual has a high cost in the orange party, since his views don't align with both *answer_d* and *answer_c*, whereas in the blue party it only slightly deviates from the *answer_b* mean.

The idea behind this is that as individuals we tend to choose the political party or candidate that disagrees with us less instead of the candidate that agrees with

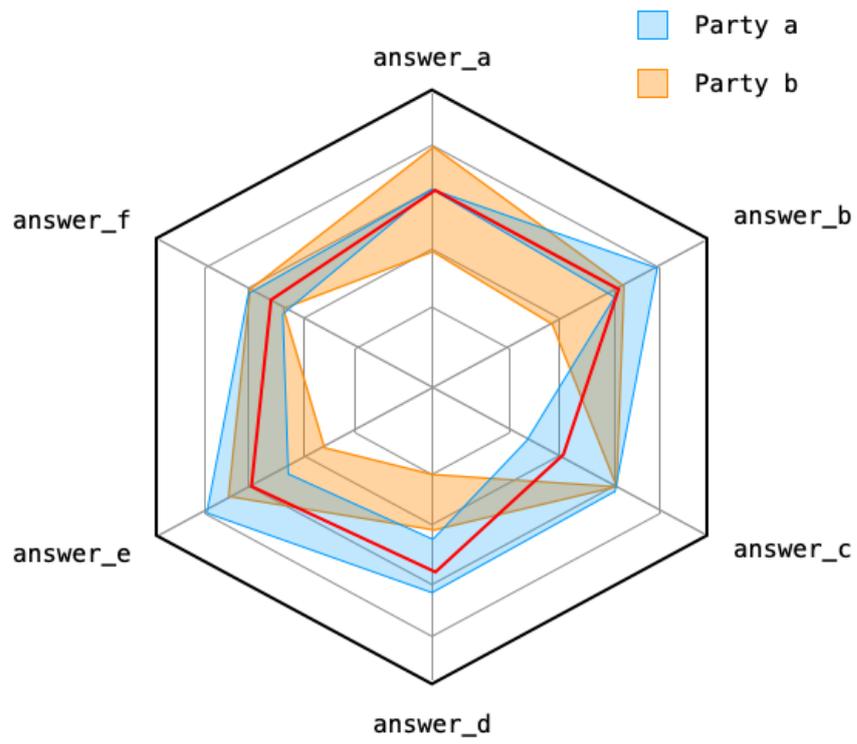


Figure 5.9: Illustrating the explicit classifier using some hypothetical values for two example parties and their collective answers to six example questions

us more. For example party a clearly puts a lot of weight on *answer_a*, since all members respond unanimously. This classifier takes this implicitly into account.

This led to an accuracy of (88%) when classifying among the six biggest parties in Switzerland (CVP, FDP, glp, Grüne, SP, SVP). After following the same steps gathered when building better SVMs (dropping politicians with more than 45 unanswered questions and considering Grüne and SP to be the same party) the accuracy increased to roughly 96%. This jump is due to the large number of misclassifications of Grüne/SP (see Fig.5.10). This could be explained by their areas virtually overlapping each other.

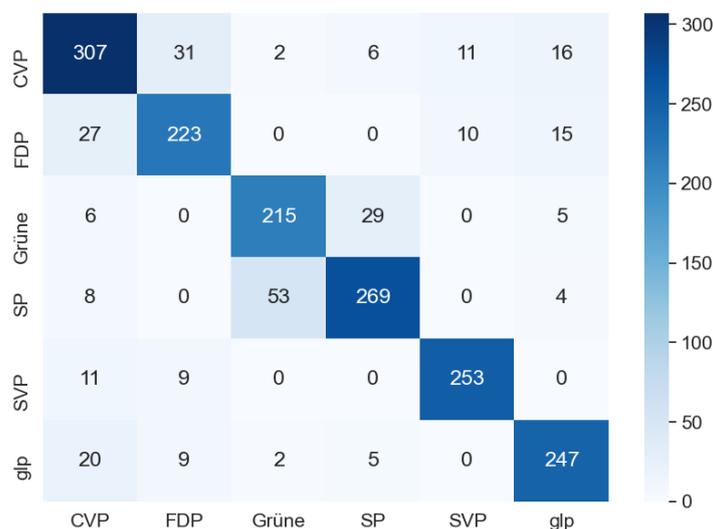


Figure 5.10: Confusion matrix for the Explicit Classifier

Once the classifier had been validated we proceeded to investigate its performance when adding questions in the order of their importance described by the different lists. For completion's sake we are still treating Grüne and SP as separate parties in Figure 5.11

As in the results for SVM classifiers (Sec.5.2.1), we conclude that the number of questions can be reduced by more than 60 percent. As before, the most confusion is due to the impossibility to efficiently distinguish between Grüne and SP politicians. Not surprisingly the ranking tailored to be applied to the custom classifier outperforms every other ranking in the long run. However the best initial performance is given by the decision tree ranking.

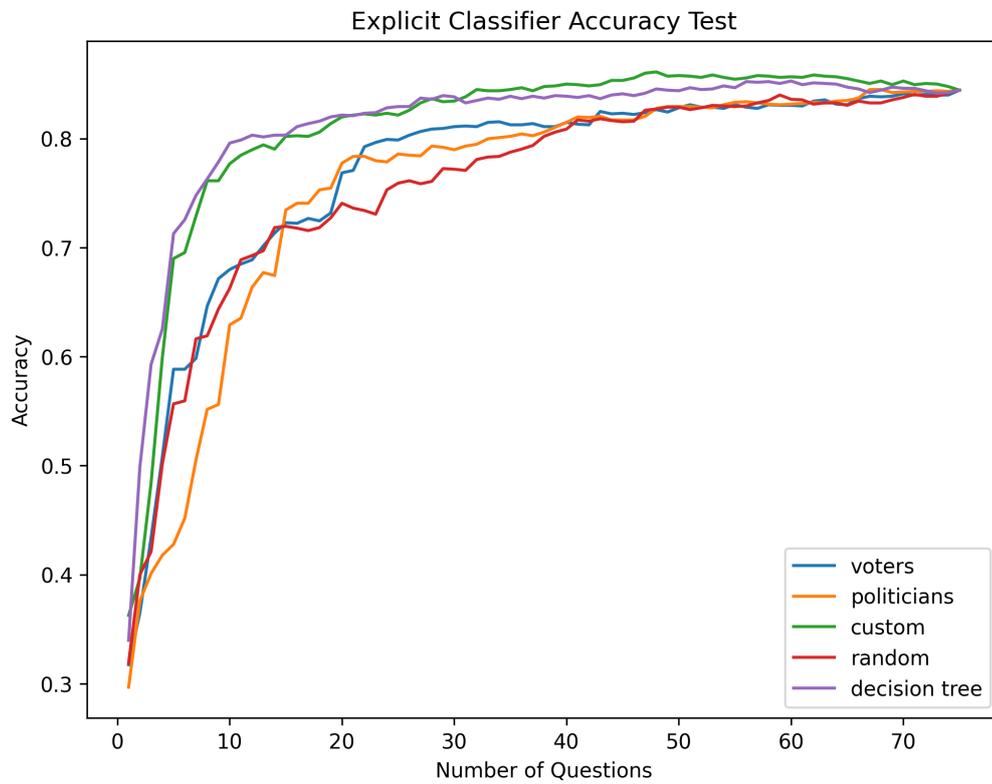


Figure 5.11: Jumps in accuracy of the explicit classifier when adding questions

5.2.3 K-Nearest Neighbours

The third and final classifier we used was a simple K-Nearest Neighbours (KNN) classifier. KNN works by classifying a data point by interrogating its neighbours on their class. The point is assigned to the class most common among its K-Nearest Neighbors [6].

We chose it because of the ease with which it could be configured using the scikit-learn library and because of the fact that it requires no training or explicitly designed parts.

A quick cross-validation resulted in an optimal value of k (number of nearest neighbours to interrogate) of 10. We chose a 25% test size.

Adding the questions bit by bit as in the previous subsections we got to the results depicted in Figure 5.12.

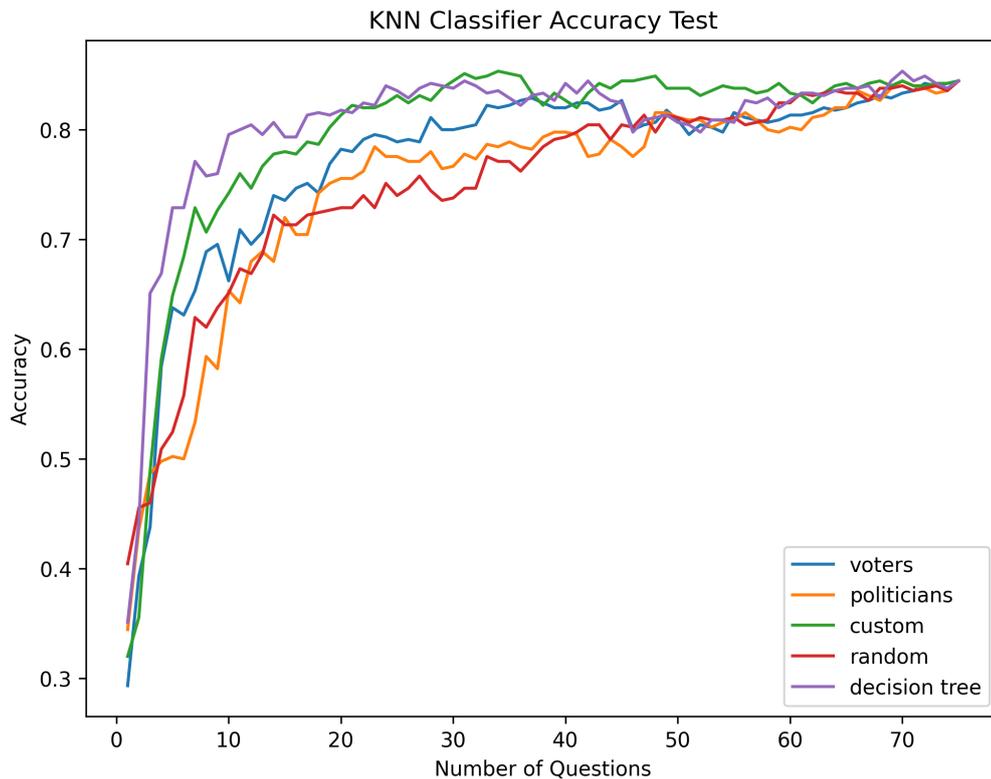


Figure 5.12: Jumps in accuracy of the KNN classifier when adding questions

As with the two previous classifiers the accuracy saturates after roughly 40% of questions. Again, the best performance is given by the decision tree ranking.

Conclusions and Future Work

We were given a dataset by the Swiss electoral advisory company Smartvote. Smartvote offers a service to voters with which, upon answering a questionnaire, voters can make more informed decisions. The dataset provided to us contained the the questionnaire answers of hundreds of thousands of voters and thousands of politicians.

Going into the project we wanted to tackle two main questions: How well is the Swiss electorate represented by the various Swiss parties? How can Smartvote improve their process to deliver more accurate suggestions?

In a couple first steps, outlined in Chapter 3, we analysed the dataset provided to us. We discovered that in many cases, voters as well as politicians skipped some questions, resulting in a sparse dataset. We devised a way to impute the missing data in such a way that unanswered questions equaled neutral answers. Given the high-dimensional nature of the data points we applied a dimensionality reduction technique (PCA) to be able to represent both the voter and politician dataset in two dimensions (Sec.3.3).

In the case of politicians this allowed us to attach a semantical meaning to the first two principal components computed from the politicians' dataset. We interpreted the first PC to be related to the left/right, the second PC to the liberal/conservative alignment of a party. We could also see how the two Swiss parties Grüne and SP have a lot of overlap in their candidates.

Regarding the voters it was made clear that the two types of voter questionnaires, deluxe (Containing all 75 questions) and rapid (Containing less questions) create a clear split in the dataset. Given the relative size of the two sets and the clear separation we decided to only proceed with the deluxe dataset, as it presents a clearer picture than its counterpart.

After having a better idea of what the datasets looked like we decided to proceed with the first question. In Chapter 4 we analysed how well the Swiss

electorate is represented by the various Swiss parties. We conducted a market fit analysis. We first clustered the voters into 4 clusters, an optimal number found by applying the elbow method on the set. We found the mean answers of those clusters for every question and compared them to the mean answers of the largest Swiss parties.

We discovered that two voter clusters resonated very well with some parties. There is a lot of overlap in ideology between the leftist cluster and the Grüne and SP parties and the centre-right/liberal cluster and the CVP party. On the other hand the voters of the other two clusters are not going to be satisfied with the available options, since the other parties have to suboptimally share the voter clusters.

How can Smartvote improve their process to deliver more accurate suggestions? We came up with a way to test different rankings of question importance. We deemed a question to be important when it helped in the task of classifying an individual to a party. The different lists we decided to test came from different ideas and analytical methods outlined in Section 5.1.

We then used different classifiers to test the accuracy of those rankings (Sec.5.2). Our results were confirmed by each of the different classifying methods:

- The number of questions can be reduced by more than 60 percent
- The ranking given by the decision tree algorithm consistently outperforms the other rankings, except when differentiating between Grüne and SP
- To be able to differentiate between Grüne and SP and thus their politicians, which is essential for Smartvote to give accurate recommendations, new questions need to be identified by Smartvote

In the future it might be interesting to add a time variable to our research, creating different voter sets at different points in time, say each year. This would allow for our methods, especially the market fit analysis to be applied on a time series. This would provide meaningful insight into the temporal changes the political landscape goes through.

Appendix

A.1 Questionnaire:

Table A.1: All Questions from the Questionnaire:

ID	Type	Question
3387	Slider-7	What is your position the following statement: "Someone who is not guilty, has nothing to fear from state security measures."."
3388	Slider-7	What is your position the following statement: "Punishing criminals is more important than reintegrating them into society."
3389	Slider-7	What is your position the following statement: "It is best for a child, when one parent stays home full-time for childcare."
3391	Standard-4	Should the federal government provide more support for the integration of foreigners?
3392	Standard-4	Should cannabis use be legalized?
3398	Standard-4	Should Switzerland terminate the Schengen Agreement with the EU, in order to reintroduce more security checks directly on the border?
3399	Slider-7	What is your position the following statement: "Wealthy individuals should contribute more to the funding of the state."
3412	Standard-4	Do you support an increase in the retirement age (e.g. to 67)?
3413	Standard-4	Should the federal government provide more financial support for the creation of childcare facilities outside the family?

- 3414 Standard-4 An initiative calls for the introduction of paid paternity leave for four weeks. Do you support this proposal?
- 3415 Standard-4 Should the conversion rate of the occupational pension fund be reduced in order to adjust for increases in life expectancy?
- 3416 Standard-4 Do you support cantonal efforts to reduce social welfare benefits?
- 3417 Standard-4 Should the federal government provide more support for the construction of non-profit housing?
- 3418 Standard-4 Should insured persons contribute more to healthcare costs (e.g. by increasing the minimal deductible)?
- 3419 Standard-4 Would you support the introduction of an opt-out solution of for organ donation?
- 3420 Standard-4 Should compulsory vaccination of children be introduced based on the Swiss vaccination plan?
- 3421 Standard-4 An initiative calls for health insurance subsidies to be designed so that no one needs to spend more than ten percent of their disposable income on health insurance premiums. Do you support this proposal?
- 3422 Standard-4 An initiative wants to give the federal government more powers to introduce measures to reduce healthcare costs (Introduction of a cost barrier). Do you support this proposal?
- 3423 Standard-4 Should the government increase its efforts to support equal education opportunities (e.g. through vouchers for private tutoring for students from low-income families)?
- 3424 Standard-4 Are you in favour of schools granting/allowing exemptions from individual subjects or events for religious reasons (e.g. PE/swimming, sex education, etc.)?
- 3425 Standard-4 Should the federal government expand its financial support for continued education and retraining?
- 3426 Standard-4 According to the Swiss integrated schooling concept, children with learning difficulties or disabilities should be taught in regular classes. Do you approve of this concept?

- 3427 Standard-4 Should foreigners who have lived in Switzerland for at least ten years be given the right to vote and be elected at the municipal level?
- 3428 Standard-4 Is limiting immigration more important to you than maintaining the bilateral treaties with the EU?
- 3429 Standard-4 Should sans-papiers be able to obtain a regularized residence status more easily?
- 3430 Standard-4 Are you in favor of further tightening the asylum law?
- 3431 Standard-4 Should the requirements for naturalization be increased?
- 3432 Standard-4 Should same-sex couples have the same rights as heterosexual couples in all areas?
- 3433 Standard-4 Should the rules for reproductive medicine be further relaxed?
- 3434 Standard-4 Are you in favour of stricter monitoring of pay equity for women and men?
- 3435 Standard-4 Would you be in favour of a doctor being allowed to administer direct active euthanasia in Switzerland?
- 3436 Standard-4 In your opinion, is lowering taxes at the federal level a priority for the next four years?
- 3437 Standard-4 Do you support a further reduction in contributions paid by financially strong cantons to financially weak cantons within the framework of financial equalisation (NFA)?
- 3438 Standard-4 Should married couples be taxed separately (individual taxation)?
- 3439 Standard-4 Are you in favour of restricting competition between the cantons with regard to corporate tax rates?
- 3440 Standard-4 Should private households be free to choose their electricity supplier (complete liberalisation of the electricity market)?
- 3441 Standard-4 Are you in favour of introducing a general minimum wage of CHF 4'000 for all employees for full-time employment?
- 3442 Standard-4 Should investment controls be introduced in order to better protect Swiss companies from takeovers by foreign investors?
- 3443 Standard-4 Are you in favour of a complete liberalisation of business hours for shops?

- 3444 Standard-4 Should the protection against dismissal for older employees be extended?
- 3445 Standard-4 Should the federal government provide more support for public services (e.g. public transport, post offices) in rural regions?
- 3446 Standard-4 Should the expansion of the mobile network according to the 5G standard continue?
- 3447 Standard-4 Should online brokerage services (e.g. "Airbnb" accommodations, "Uber" taxi services) be regulated more strongly?
- 3448 Standard-4 An initiative calls for Switzerland to stop using fossil fuels by 2050. Do you support this proposal?
- 3449 Standard-4 Currently, a CO2 charge is levied on fossil combustibles (e.g. heating oil, natural gas). Should this charge be extended to motor fuels (e.g. petrol, diesel)?
- 3450 Standard-4 Should the federal government provide more support for renewable energies?
- 3451 Standard-4 Should high traffic motorways be expanded to six lanes?
- 3452 Standard-4 Are you in favour of introducing "Road Pricing" for motorised individual transport on busy roads?
- 3453 Standard-4 Do you support the relaxation of the current measures to protect large predators (lynx, wolves, bears)?
- 3454 Standard-4 Should the current moratorium on genetically modified plants and animals in Swiss agriculture be extended beyond 2021?
- 3455 Standard-4 Should direct payments only be granted to farmers that provide an extended ecological performance record (e.g. no synthetic pesticides and limited use of antibiotics)?
- 3456 Standard-4 Are you in favour of extending landscape protection (e.g. stricter rules for building outside existing building zones)?
- 3457 Standard-4 Are you in favour of stricter animal welfare regulations for livestock (e.g. permanent access to outdoor areas)?
- 3458 Standard-4 Should campaign finance for political parties and referendums be openly declared?

- 3459 Standard-4 Should the introduction of electronic voting in elections and referendums (e-voting) be further pursued?
- 3460 Standard-4 Are you in favour of lowering the voting age to 16?
- 3461 Standard-4 Should the Federal Council's proposal to tighten the conditions for admission to the civil service be abandoned?
- 3462 Standard-4 Should the export of war materials from Switzerland be banned?
- 3463 Standard-4 Are you in favour of Switzerland acquiring new fighter jets for the armed forces?
- 3464 Standard-4 Do you support an expansion of the legal possibilities for using DNA analysis in investigations?
- 3465 Slider-7 What is your position the following statement: "In the long term, everyone benefits from a free market economy in the long term."
- 3466 Slider-7 What is your position the following statement: "The ongoing digitalization offers significantly more opportunities than risks."
- 3467 Slider-7 What is your position the following statement: "Stronger environmental protection is necessary, even if its application limits economic growth."
- 3468 Standard-4 Should Switzerland start membership negotiations with the EU?
- 3469 Standard-4 Should Switzerland strive for a free trade agreement with the USA?
- 3470 Standard-4 An initiative calls for liability rules for Swiss companies with regard to compliance with human rights and environmental standards abroad to be tightened. Do you support this proposal?
- 3471 Standard-4 Are you in favour of Switzerland's candidacy for a seat on the UN Security Council?
- 3472 Budget-5 Should the federal government spend more or less in the area of "Development assistance"?
- 3473 Budget-5 Should the federal government spend more or less in the area of "National defence"?
- 3474 Budget-5 Should the federal government spend more or less in the area of "Public security"?
- 3475 Budget-5 Should the federal government spend more or less in the area of "Education and research"?
- 3476 Budget-5 Should the federal government spend more or less in the area of "Social services"?

- | | | |
|------|----------|--|
| 3477 | Budget-5 | Should the federal government spend more or less in the area of "Road traffic (motorised individual transport)"? |
| 3478 | Budget-5 | Should the federal government spend more or less in the area of "Public transport"? |
| 3479 | Budget-5 | Should the federal government spend more or less in the area of "Agriculture"? |

A.2 Decisiontree

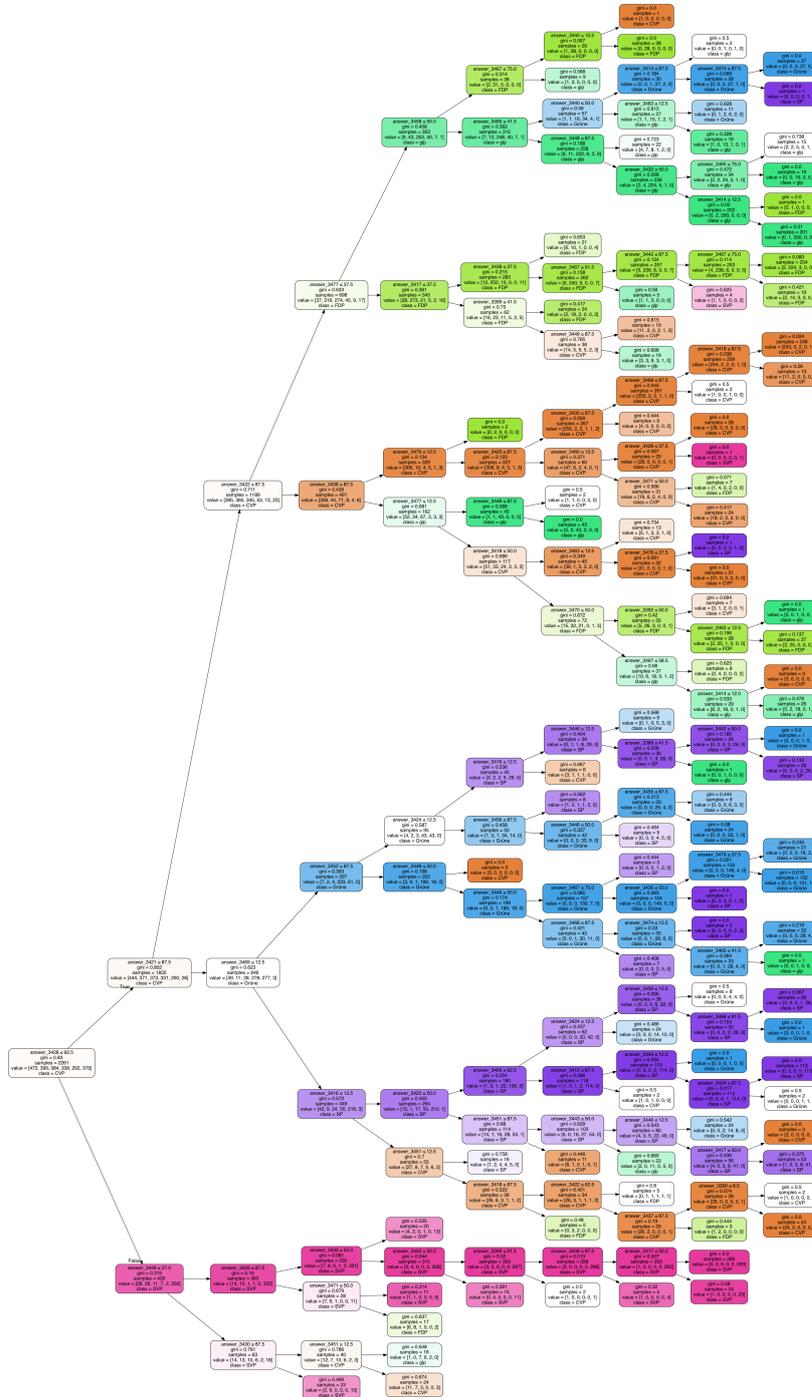


Figure A.1: The whole decision tree for the 6 biggest parties

Bibliography

- [1] L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: <http://www.jmlr.org/papers/v9/vandermaaten08a.html>
- [2] M. Wattenberg, F. Viégas, and I. Johnson, “How to use t-sne effectively,” *Distill*, 2016. [Online]. Available: <http://distill.pub/2016/misread-tsne>
- [3] K. Pearson, “LIII. On lines and planes of closest fit to systems of points in space,” Nov. 1901. [Online]. Available: <https://doi.org/10.1080/14786440109462720>
- [4] T. M. Kodinariya and P. R. Makwana, “Review on determining number of cluster in k-means clustering,” *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.
- [5] S. R. Gunn *et al.*, “Support vector machines for classification and regression,” *ISIS technical report*, vol. 14, no. 1, pp. 5–16, 1998.
- [6] N. S. Altman, “An introduction to kernel and nearest-neighbor nonparametric regression,” *The American Statistician*, vol. 46, no. 3, pp. 175–185, 1992. [Online]. Available: <https://www.tandfonline.com/doi/abs/10.1080/00031305.1992.10475879>