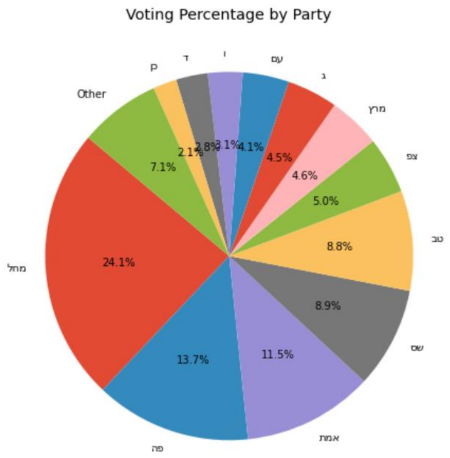
Contents

Data used in project

- expb_.csv – Number of votes per party per ballot box.
- Bycode.xls – Metadata of the cities and towns in Israel

A pie chart of the voting percentage per party on a national scale:



Data preprocessing

While the voting data frame does not contain missing values, this is not the case for the metadata data frame, named *city_info*, in which many features have a high percentage of missing values as the following table shows:

Dealing with feature: Uninformative, went through feature engineering, removed because highly correlated to other features, removed because a high percentage of missing values

feature	Nan percentage
שם יישוב	0%
סמל	0%
תעתיק	14%
מחוז	0%
נפה	0%
אזור טבעי	9%
מעמד מוניציפאלי	4.8%
שיוך מטרופוליני	62.18%
דת יישוב	15.12%
סך הכל אוכלוסייה 2013	17.6%
יהודים ואחרים	26.27%
יהודים	26.4%
ערבים	90.77%
שנת ייסוד	25.72%
צורת יישוב שוטפת	0%
השתייכות ארגונית	43.57%

קואורדינטות	1.16%
גובה ממוצע	16.48%
ועדת תכנון	10.33%
מרחב משטרה	10.05%
תעתיק פרסומים	15.94%
שנת עיבוד	15.94%
שדה לקישור	16.48%

While some features do not seem to contain valuable information, such as שדה לקישור and תעתיק, others are still important even with their high percentage of nan values.

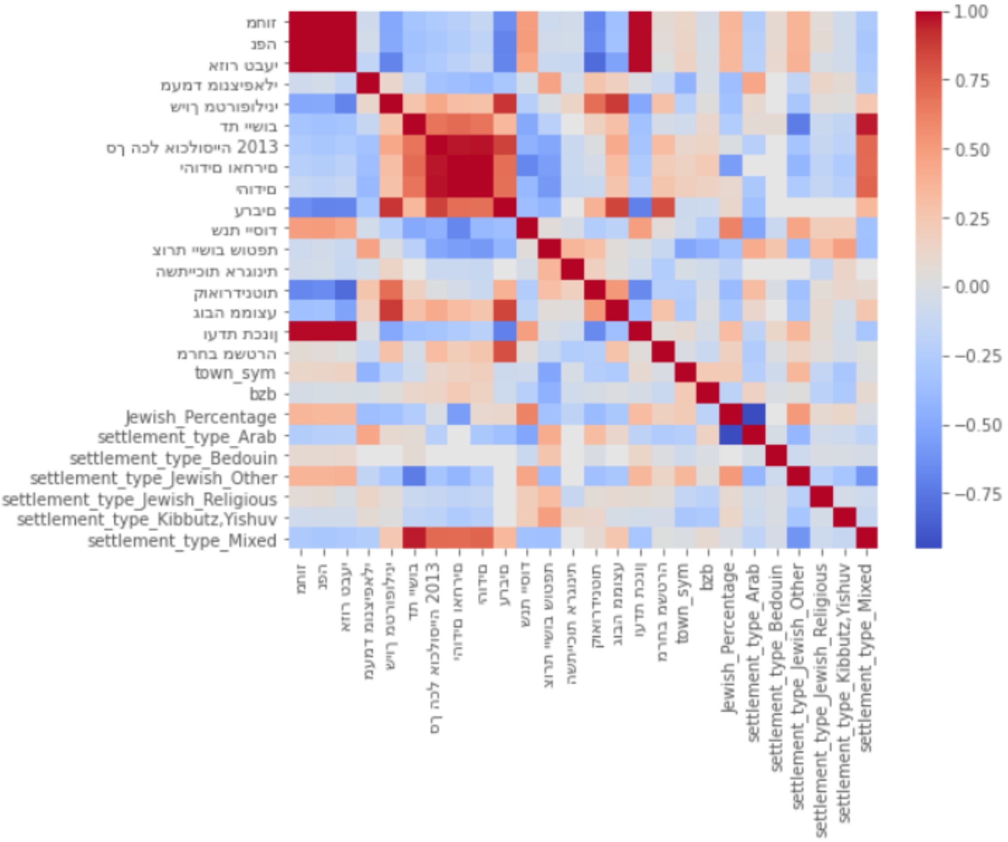
Feature engineering

- The feature צורת התיישבות ארגונית contains information about the type of the settlement (Kibbutz, Yishuv, Moshav, Religious Kibbutz, West Bank/ J&S settlements, etc.) which is valuable regarding voting patterns in smaller settlements. This feature contains "nan" for a high percentage of data points, mainly cities and Arab villages. Therefore, a new feature was created using the data from this feature and דת יישוב, which sorts towns into 4 categories: Jewish, Arab, Bedouin, and Mixed.
  The new feature is named "settlement_type", is based on the information provided by those two features, has no nan values, is categorial, and can get one of the following values:
  - Arab
  - Bedouin
  - Mixed
  - Kibbutz_Yishuv
  - Jewish religious
  - Jewish other
- There are several features containing demographic information, such as ערבים, יהודים, יהודים ואחרים each has the absolute number of residents belonging to each group, and the number appears only if the group exists in the data. Therefore, a new feature was created: Jewish Percentage, based on those features as well as 2013 סך הכל אוכלוסייה.
- The שנת ייסוד feature has missing values for Arab and Bedouin towns. Therefore, all missing values were replaced by the year 1850. This seems like an arbitral choice, but since most of these towns existed before the establishment of the state of Israel it is a choice that made sense.

Correlation between features

To find highly correlated features, the data was normalized and a correlation heatmap matrix was computed:

(I apologize for the inverse Hebrew script, my computer absolutely refused to download the package that would have fixed this, so this is the case for some of the plots)

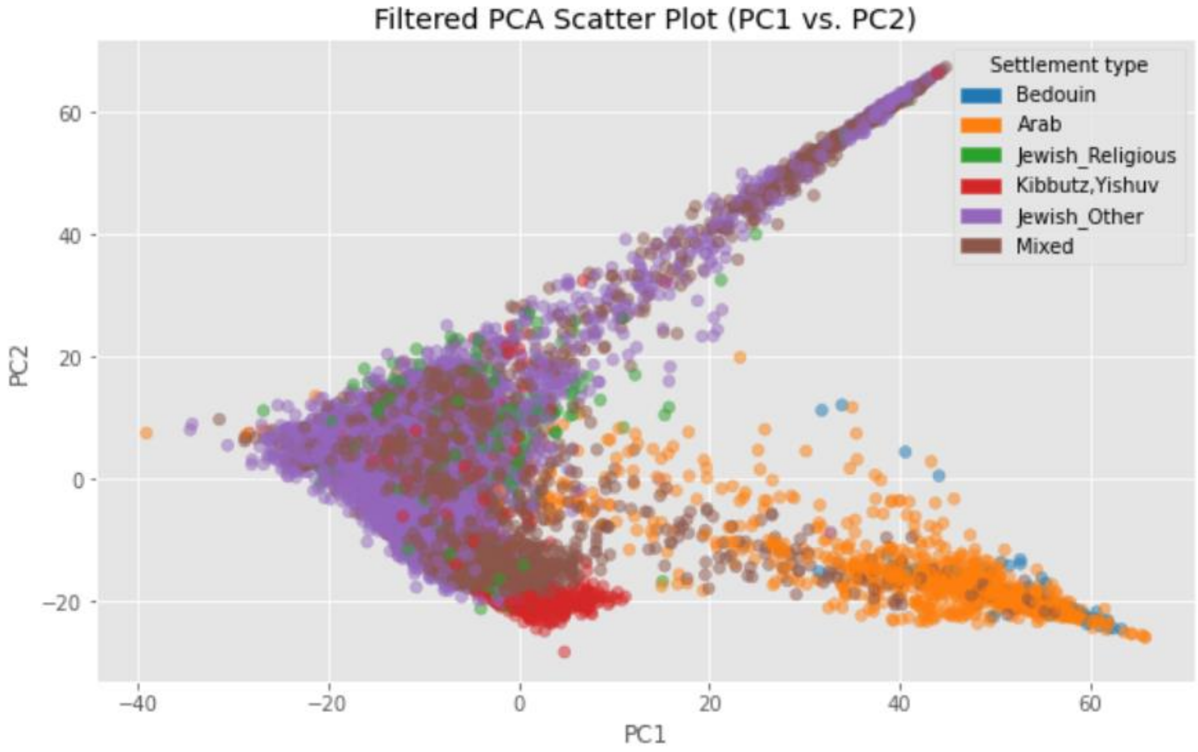


This has led to the conclusion that some sets of features are highly positively or negatively correlated. Of the following sets of correlated features, the one marked in green was kept (mainly because it was found to better improve the model or was more informative):

- settlement_type_mixed and דת יישוב
- אזור טבעי, נפה, מחוז, ועדת תכנון (The information in מחוז is contained in נפה which is contained in אזור טבעי and ועדת תכנון. Of the two, ועדת תכנון was found to contribute more to the model.)
- יהודים, ערבים, יהודים ואחרים, סך אוכלוסיה (Jewish Percentage feature contains the information of the other three features)

Cluster ballot boxes according to vote distribution

- The data was normalized – instead of having votes count, it has vote percentage.
- PCA dimension reduction was applied to reduce the data to two features which are a linear combination of the others.
- The new data was plotted, and these are the results after the data was colored according to the newly created feature of "settlement_type":



We can observe a clear distinction between the **Arab** and **Bedouin** towns (orange and blue respectively) compared to a Jewish majority. There is also a clear cluster of **Kibbutz** and **Yishuv**, colored in red, which I assume in general vote for a more left-leaning party.

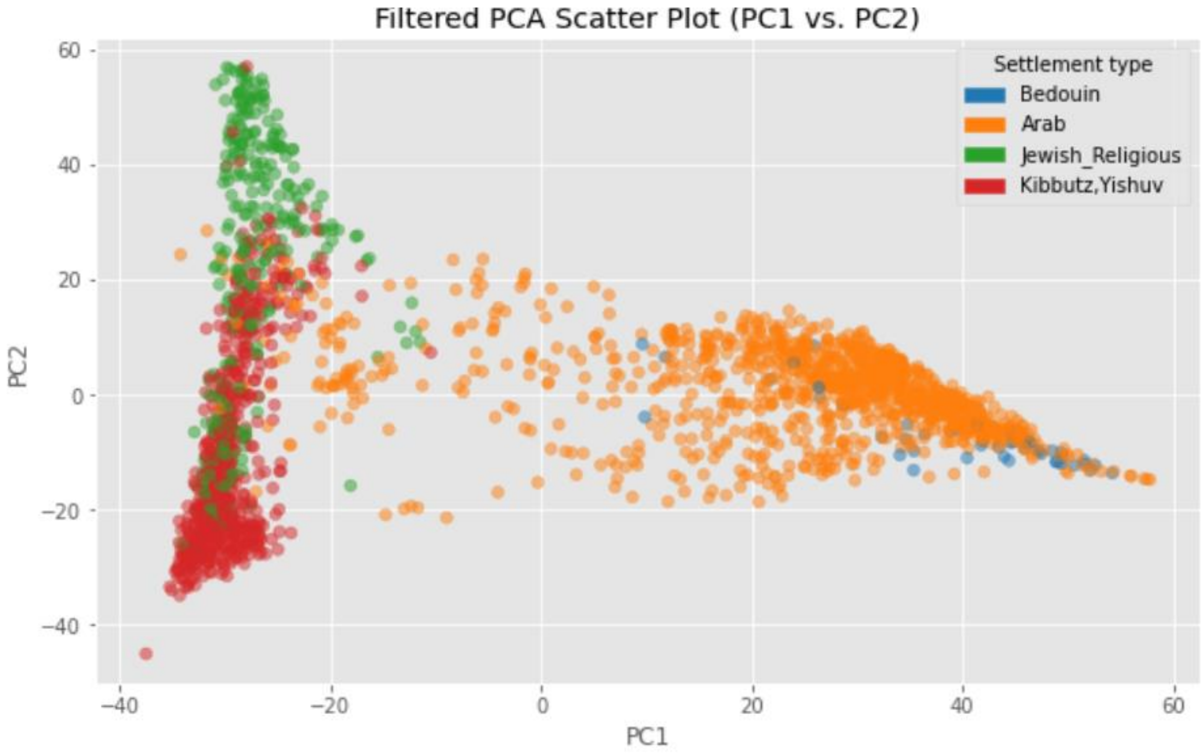
In **Mixed** cities, we can observe all types of voting patterns, some are similar to Arab towns, some are similar to the Kibbutz-Yishuv cluster (probably left-leaning), and some form another clear cluster at the top right part of the plot with Jewish majority towns, corresponding to the following ranges: $20 < PC1 < 50\ and\ 30 < PC2 < 70$

Upon inspecting these ballot boxes, I found out that they consist of **Ultra-Orthodox** towns and cities (see table below), or ballot boxes from Jerusalem, which is a mixed city and has the largest Ultra-Orthodox community.

City	Count of Ballot boxes in a cluster
Jerusalem	172
Bnei Brak	126
Modi'in Ilit	34
Beitar Ilit	30
Beit Shemesh	28
Ashdod	24
Elad	18

Unfortunately, I did not find a fitting feature to distinguish between the Ultra-Orthodox majority towns and the other Jewish majority, which could have been a great contribution to the algorithm I am about to present in the next section. I would, however, have added manually to the settlement_type feature a category of Ultra-Orthodox based on my knowledge and the towns in this cluster.

The **religious Jewish** towns and settlements (mostly West Bank settlements and religious Kibbutzim), colored in green, are contained in a cluster with other types of Jewish-majority towns, and I would assume has a more right-winged leaning ballots. This cluster becomes clearer if we remove Jewish_other and Mixed datapoint:



*Please remember this plot and the data points from Arab towns which are clustered closer to Jewish towns ($-40 < PC1 < -10$), we will look into those in the anomaly detection part.

Predictive model

The question: Using the metadata about each ballot box, can we classify the party that scored the most votes?

Type of problem: classification

Chosen model: Random forest is highly suitable for the task as a robust ensemble model for multiclass classification.

Features and predicted value

Chosen features:

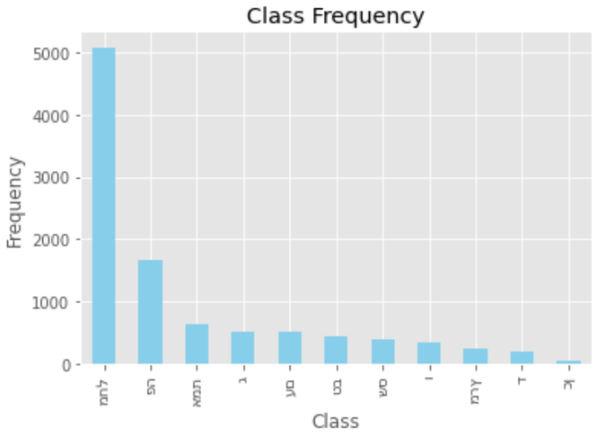
- ועדת תכנון
- בזב
- מעמד מוניפיאלי
- דת יישוב
- סך הכל אוכלוסייה 2013
- שנת ייסוד
- צורת יישוב שוטפת
- מרחב משטרה
- Jewish percentage
- Settlement_type_Jewish_Religious
- Settlement_type_Jewish_other
- Settlement_type_Arab
- Settlement_type_Kibbutz,Yishuv
- Settlement_type_Bedouin

Predicted value: Winning – a class of 11 categories (political parties)

(11 parties remained after I removed data points with parties voted by less than 20 ballot boxes.)

Preparations before model

- One hot encoding – we have one categorial feature, the settlement_type, which was separated into different one-hot encoded vectors.
- Normalization – each feature has a different range, therefore all were normalized to be between 0 and 1
- Data imbalance:
  Our data is very unbalanced, as can be observed in the following plot:



The most frequent category by far is מחל, more than twice as frequent as the following one - פה.

To deal with the imbalance problem, I tried three approaches:

  - The SMOTE method which over samples minority classes by creating synthetic data. This approach is not ideal, as many synthetic data points will have to be created, and as we've seen with the PCA, there isn’t always a clear distinction between the different data points.

- Weighted samples – give each class a weight. This approach reduced significantly the model performance and therefore was abandoned.
- Not dealing with the problem – the model tended to vote for מחל and have lower score.

Hyperparameter tuning

I used a combination of RandomizedSearchCV and GridSearchCV.

RandomizedSearchCV helped to randomly find a set of hyperparameters that resulted in good performance, while Grid Search was used to explore similar sets of hyperparameters to those I detected with the random approach, to find the most suitable.

The CV (cross validation) should help avoiding overfitting of the parameters to the training data.

The following parameters were chosen for Random Forest model:

- N_estimators = 20
- Min_samples_split = 5
- Min_samples_lead = 3
- Max_features = 3
- Max_depth = 30
- Bootstrap = True

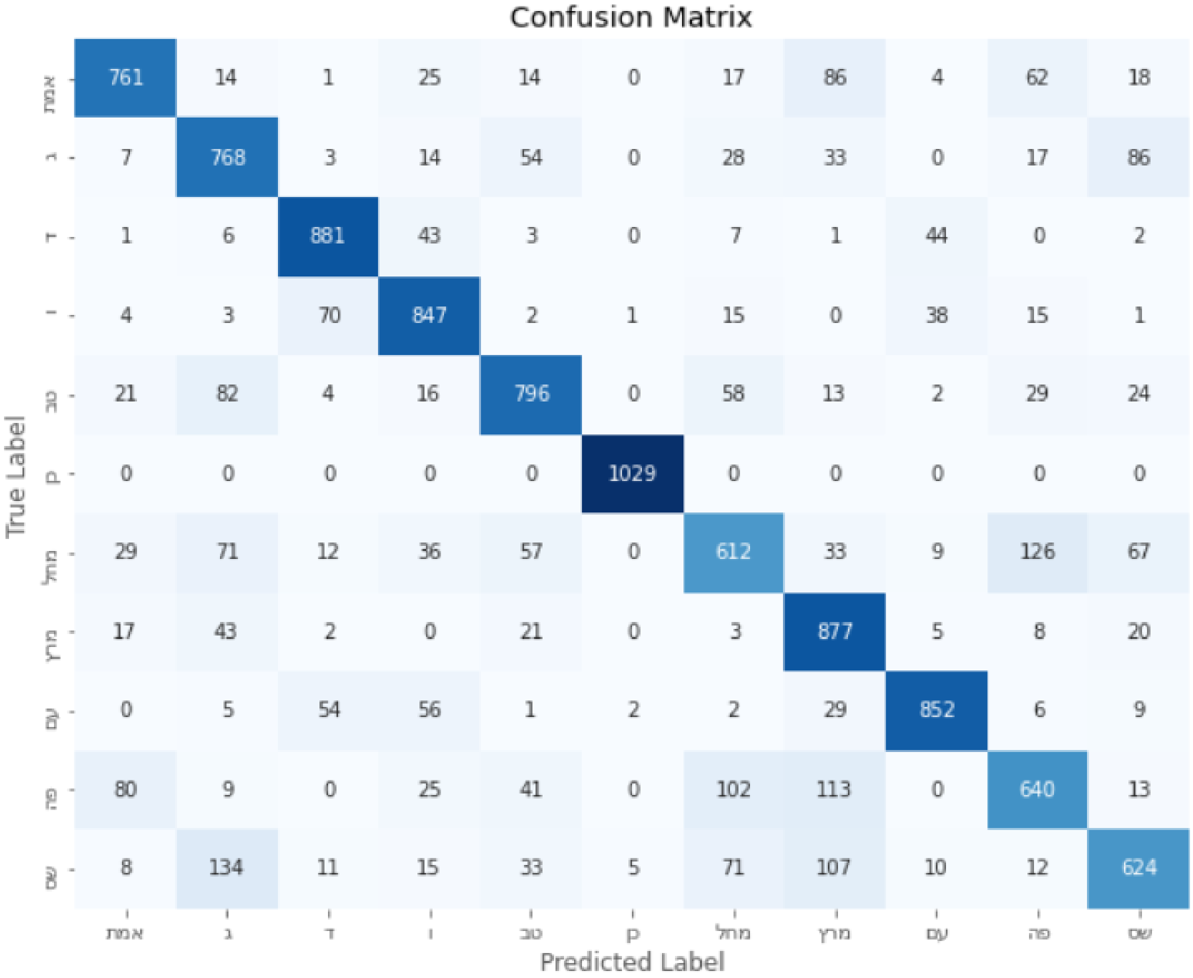
Model results

- Overall: Accuracy – 78%, Precision – 78%, Recall – 78%, F1 score – 78% (same values due to class balance)
  - Training: Accuracy – 86%
  - Without SMOTE: Accuracy – 65%, Precision – 64%, Recall – 65%, F1-score-64%
- Performance varied between the different classes. I recognized two groups of parties that tended to have lower results:
  - **Large parties** such as מחל and פה, probably because their characteristics overlap other parties, and are not "fringe parties".
  - **Ultra-Orthodox parties** ג (יהדות התורה) and שס which as I mentioned earlier, I wasn’t able to create a feature that separates Ultra-Orthodox majority ballot boxes. In comparison, **Arab parties** tended to have good performance, as there were quite enough features to detect Arab-majority towns.
- By observing the confusion matrix (see on the next page), we can see that the mistakes that the model makes tended to "make sense", for example:
  - the Arab party ד (בלד) most common misclassification is to the other two Arab parties.

    - The Likkud party, מחל misclassifies the most to טב (הבית היהודי) שס and פה (יש עתיד).
    - אמת (העבודה) misclassifies the most to פה.
- The five most important features according to the random forest model were:
    - בזב (I have no idea what it is and at this point I'm afraid to ask)
    - Jewish_Percentage
    - מרחב משטרה
    - וועדת תכנון
    - סך כל האוכלוסיה 2013
    - צורת יישוב שוטפת
    - שנת ייסוד

All of this leads to the conclusion that **many characteristics of voters are overlapping, and there is a need for more features to make a better classification and detect nuance**. A simple statistic such as mean household size or mean income per household in each city could be of great improvement for small, homogenous towns. Still, large cities have so much diversity between neighborhoods, (Jerusalem for example), so predicting by the city statistics alone is not enough.

I assume that if I filtered out ballots from large cities (we have a feature for that) I could have shown that the model's performance improves significantly, but I didn't have time for that.



	אמת	ג	ד	ו	טב	ק	מחל	מרץ	עם	פה	שס
אמת	761	14	1	25	14	0	17	86	4	62	18
ג	7	768	3	14	54	0	28	33	0	17	86
ד	1	6	881	43	3	0	7	1	44	0	2
ו	4	3	70	847	2	1	15	0	38	15	1
טב	21	82	4	16	796	0	58	13	2	29	24
ק	0	0	0	0	0	1029	0	0	0	0	0
מחל	29	71	12	36	57	0	612	33	9	126	67
מרץ	17	43	2	0	21	0	3	877	5	8	20
עם	0	5	54	56	1	2	2	29	852	6	9
פה	80	9	0	25	41	0	102	113	0	640	13
שס	8	134	11	15	33	5	71	107	10	12	624

Anomaly detection

Two methods were applied to detect anomalies within the data:

- Isolation forest – Using this method with a parameter of contamination of 0.005 I detected 51 outliers. Some of these were ballot boxes from Arab and Bedouin towns voting for Jewish parties such as שס, ,אמת and מחל.
- Looking into the data points of Arab towns within the Jewish cluster in the PCA plot shows that these villages are in the majority **Druze** (Beit Jan, Yanukh Jeit, Julis, Horfeish, Sajour, etc.) but also **Alawi** (Ghajar) and **Circassian** (Rihania), which is of no surprise, but the data does not distinct between them and Arab villages.