Data Science - Final Project

Phase 1 - Goal **2**

Phase 2 - Data **4**

Phase 3 - Build Model **6**

Phase 4 - Evaluation of the model **7**

Predictive 7

Descriptive 9

Phase 1 - Goal

For our final project, we collected relevant data concerning an innovative dating app that some of our group members are currently developing, named Kukumbo.

Our SMART goal is to **Find and define an efficient target market segmentation** for our product.
We want to analyze our potential customers and understand their main characteristics, to optimize the marketing process accordingly. Specifically, by deeply understanding the different target market segments, we can focus our marketing resources on the most potential customers and run effective advertising campaigns.

In addition to the main goal, we want to train a model that will be able to predict whether a person would want to use the app or not, based on several “basic” questions.

SMART goal parameters

- **Specific**

  Define multiple (3+) segments of potential users out of the general population based on their characteristics.

  - Build a model that will predict if an instance would use the app with an accuracy of above 85%.

- **Measurable**

  Find sets of characteristics that define the different segments of the target market. We will measure our success based on the following:
  Identify sets that contain no more than four characteristics that define segments within the target market. Accordingly, a person who belongs to one of the above segments is likely to be interested in using the app with a probability of at least 0.7.

- **Assignable** – we divided the tasks as follows:

Task	Assigned to
Data Mining	Kai & Daniel
Data Preparing	Daniel & Arielle
Training Models	Kai
Analysis	Daniel & Gal
Summarize	Gal
Report	Arielle

- **Realistic**

  We built our database for the research, based on the results of a large online survey we conducted, thus - the data exists.

  The team is capable of carrying out this research project because we have expertise in:

  1. Python & Excel data management.
  2. Python Decision Tree training.
  3. Analysis of model's performance.
  4. Summing up and delivering results, for both:
     a. An expert who understands the algorithm and the mathematical background.
     b. An entrepreneur who only cares about the bottom line.

- **Time-related**

  We defined milestones based on the division of the tasks mentioned in the table above. In addition, we pinned minor deadlines accordingly, so tasks would be completed in chronological order and the project would be delivered on time. We also prepared the schedule for dealing with unexpected events, so it is guaranteed that we hand in the project on time.

Phase 2 - Data

Collecting the data - We obtained data by surveying in the IDC and outside of it. The survey includes questions about the participants' personal lives and their preferences concerning dating and dating apps (Click here to access the survey form). The instances we selected are as heterogeneous as possible, so they will represent a wide and realistic sample of the Israeli population.
As a result, we gained insight into the user's diversity, his romantic needs and desires, and the way he uses other dating apps. Initially, we chose these questions because we believed they would provide the most valuable information. We had an assumption that the characteristics of a person are highly correlated to our goal, and thus we'll be able to create a clear and relevant segmentation.

Preparing the data (data specifics)-

"K_dirty_data.csv" contains the raw data as downloaded from the Google Form. We wrote the *"General_clean.ipnb"* scripts, which cleans the raw data of irrelevant parts, such as a column of *"time_stamp"* which indicates when the surgery was submitted. The script then extracts new features we thought would contain valuable information. Sometimes we did it towards dimensional reduction, to achieve simpler models but without losing information. We also used a Pearson Correlation Matrix to identify redundant features. Additionally, it includes detailed explanations for what we did there. The Last step was saving the clean and improved data, which is called - "*K_clean_data.csv*". To see the full explanation of the columns' values and meaning go to "*Dictionary.xlsx*"

Our data consists of the following columns:

1. Location - Place of living (whether in Israel or abroad)
2. Age - by groups
3. Gender
4. Interested_in - men, women, etc
5. Current_status - regarding relationship status
6. Usually - same as (5)
7. Confidence - level of self-confidence
8. Rate of different ways a potential partner can be met:
    a. Bars
    b. Social_network
    c. Dating_apps
    d. Through_a_friend

    e. Street9
9. experience - the user's rating of their experience of using dating apps so far
10. frequently_use - How frequently do they use dating apps
11. paid - if someone has paid for a dating app
12. swipes_are - the attitude towards swipes
13. look_for_short_term - Are you seeking a short-term relationship?
14. look_for_long_term - Would you like a long-term relationship?
15. Did you use this app:
    a. Tinder
    b. Bumble
    c. OkCupid
    d. Grinder
    e. Hinge
    f. Atraf
    g. Badoo
16. num_apps - number of apps you've used, we calculated it
17. use_for_Long_term - did you use dating apps in order to find a long term relationship
18. use_for_Short_term - did you use dating apps in order to find a short term relationship
19. use_for_Friends - did you use dating apps in order to find new friends
20. Is this a reason you don't like/use dating apps:
    a. affraid_bad_people
    b. not_seen_there
    c. too_much_not_myType
    d. cannot_find_somebody_I_like

Phase 3 - Build Model

To split the database into multiple segments, we created a Decision Tree model.
The major reasons we considered when choosing the DT model were its ability to-

1. Deeply understand the branchings of the population into different segments according to the features.
2. Effectively rank the different features by importance level.
3. Using the DT visualization, explain the results clearly to potential business partners and investors who had no prior knowledge of AI or even mathematics.

We had two phases of research, the first one was to find the best predictive model - we wanted to be able to predict whether a person would use the app or not, with as high accuracy as possible. We made two technical decisions regarding training the model:

1. Using a large range of max depths to show a graph of the accuracies of each of them on the train and the test sets. This way we could find, for each data set, the “sweet spot”, which is the point just before overfitting.
2. Train-test - we used 30% of the data as a test set for evaluating the model’s behavior in the “real world”, which is what matters, and also for finding the best model, regarding the max length aspect.

In the second phase, our goal was to descriptively analyze the data using DT. We made two technical decisions regarding training the model:

1. Max depth - we trained a DT with a max depth of 1,2,3 because we wanted to define strong characteristics and segments, if we use deeper trees they will be more complex and thus harder to interpret and understand, and even harder to use for business purposes.
2. No train-test - we used the whole data, assuming it contains more information. We’ve done it because our data is relatively small and we wanted our analysis to be as accurate as possible, and also it’s not important, in this phase, what are the predictive power of the models.

The progress and documentation can be found in “*DT.ipynb*”.

Phase 4 - Evaluation of the model

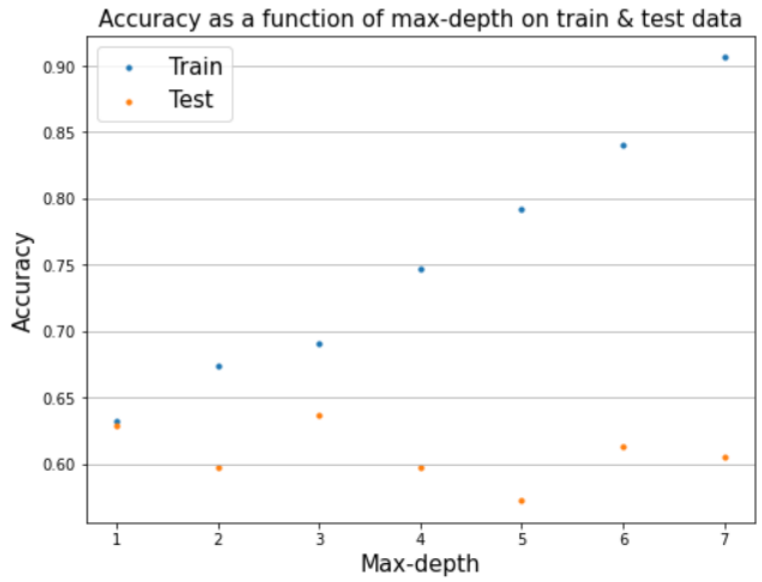
Predictive

We trained many models on many combinations of different datasets and features sets, here we show the most valuable ones. Note that by choosing some features to use while training a model, we got models that are at most not better than when using all the features. It makes sense and follows the theory (maybe not for every case) of Decision Trees, that is because the DT chooses the most valuable features, regarding decreasing the impurity of the data. As described in Phase 3, we chose the best tree using a graph of the accuracies on the train and test sets, that is the most accurate tree, before overfitting.

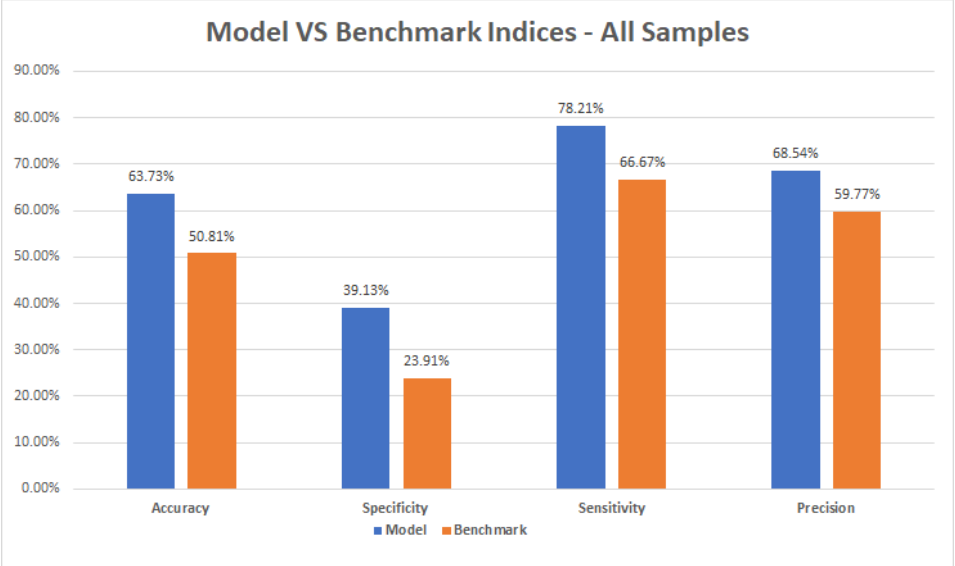
The accuracy is important to us because it represents the general prediction power of the model, but we chose a deeper estimation of the model, which suits our business goals better - precision. We try to predict whether a person is likely to use our app so we use our marketing resources efficiently, therefore it is crucial for us to define an instance positively with high probability of correctness probability. Recall that $Presicion = TP/(TP + FP)$.

Besides looking at the accuracy, precision, etc, we evaluated the model's performance next to those of a benchmark model of the same positively labeled instances ratio.

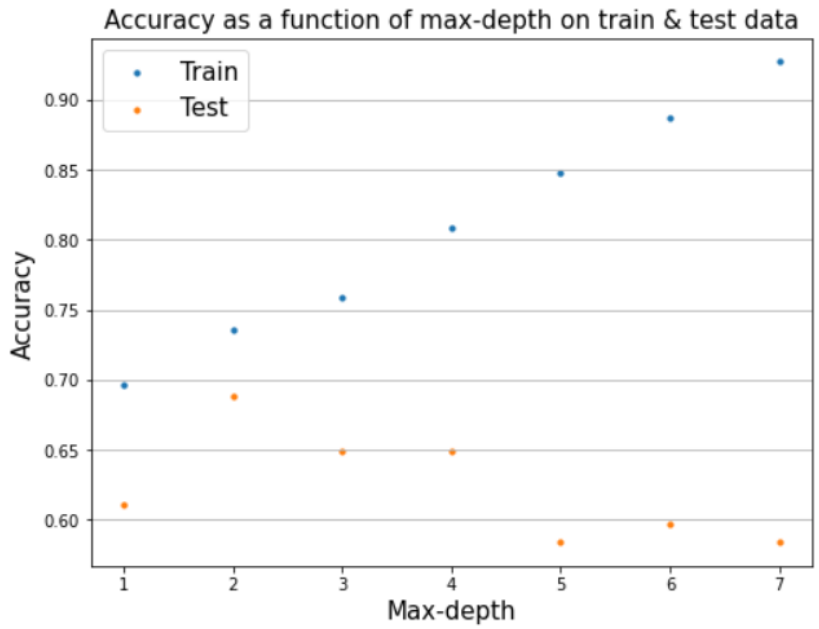
The first run is on the whole dataset, it can be seen that the best tree is achieved with max depth of 3 with an accuracy of ~64% on the test set.



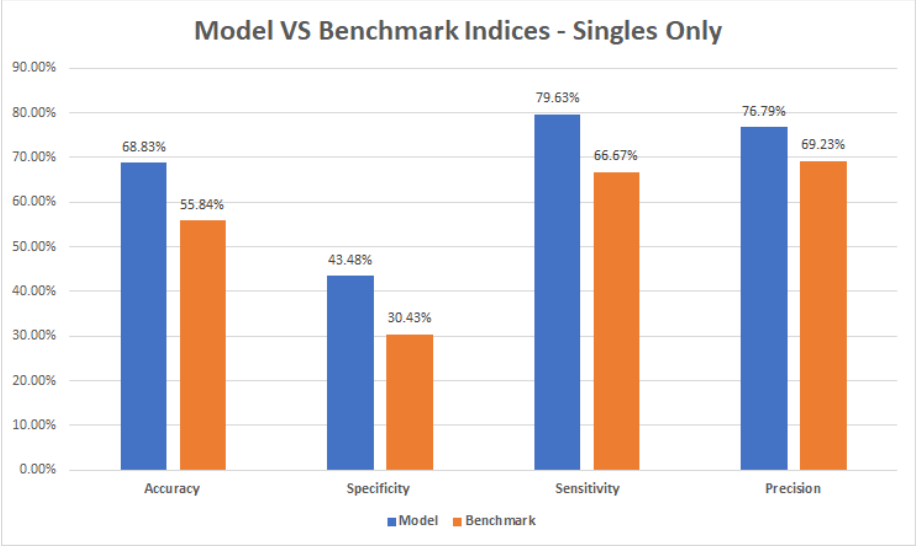
We examine this model (tree) more closely, as shown in the graph below. First of all our model is better at all means than the benchmark, which is good. The Precision is ~68% which is not as high as we wanted it to be, but note that the Sensitivity here is high as well. The combination between them may lead to relatively good performance.



The Second run is done the same way, using only the singles among the data, we've done it because they are much more likely to use our app, so we wanted to analyze them closely. This time the best accuracy is almost 70%, achieved using max depth of 2.



The analysis and explanation of this model is pretty similar to the one above, with the whole data, with two main differences. The Accuracy is a bit higher and the precision is a bit lower. Note that the Sensitivity and, especially, the Specificity are also greater, so we assume that this model may perform better on the general population. This is a meaningful insight for us - if we make further research on this topic, we'll focus on singles, as it seems (may indicate) that the instances "in a relationship" creates noise when training the model, and again, they are not really potential users to begin with.

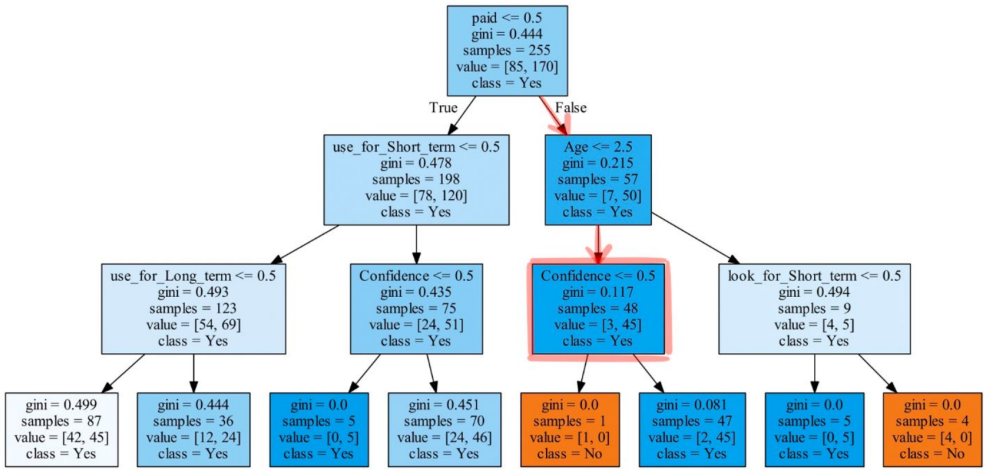


Descriptive

As a means of achieving our goal and defining the different segments of our target market, we used the Decision Tree model. We ran the algorithme many times and created many different classifiers, among which we selected three trees as the most significant ones. The trees we chose are those that clearly mark the segments of users who are likely to be interested in our application - Kukumbo. Our goal was to find trees whose segments are defined by distinct sets of attributes that define different segments.

Therefore, we searched the tree for nodes with low Gini scores (which indicate a high level of certainty), that would reflect segments with a high likelihood of being interested in our application. Furthermore, we chose segments with a sufficient amount of instances.

An additional measure was to choose nodes with higher Gini scores than some of their descendants if the split in values was small, so as to avoid overfitting. For example:



** Our segment (segment #3, see results next page) was defined by the marked node above, even though its child node had a lower Gini score, as the partition of values was minor and might have resulted in an overfitting effect.*

RESULTS

Based on the analysis, the following target segments emerged:

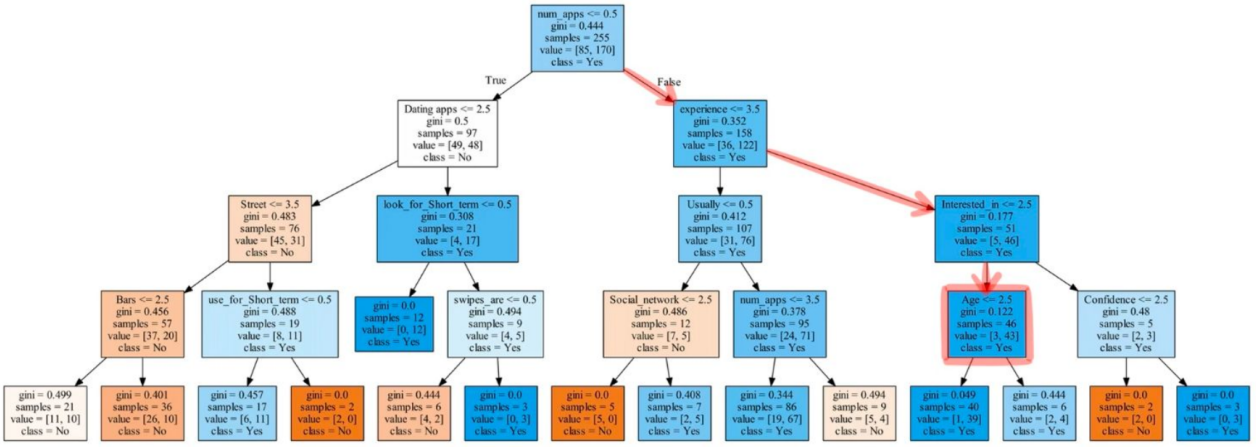
Segment #1 -

- Defined by the following path:
    1. num_apps > 0.5
    2. experience > 3.5
- Values = [3,43]
- Probability = 43 / (3 + 43) = 0.934

Meaning: This segment is made up of people who use one or more dating applications and have had positive experiences with them.

93.4% of the segment is likely to be interested in the application.

Visually:



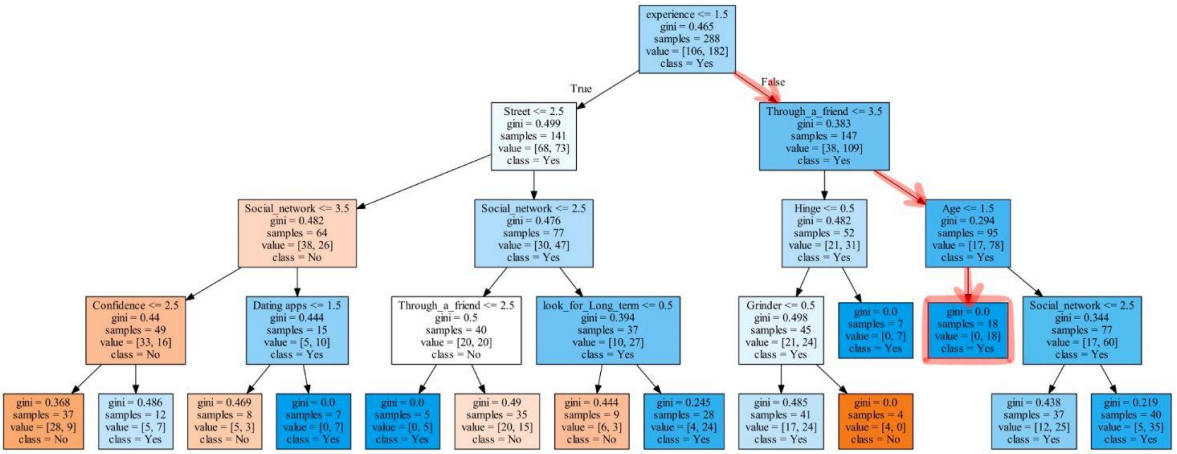
Segment #2 -

- Defined by the following path:
    1. experience > 1.5
    2. Through_a_friend > 3.5
    3. Age <= 1.5
- Values = [0,18]
- Probability = 18 / (0 +18) = 1

Meaning: This market segment is composed of people in the age group 18-22 who would rather find a partner through mutual friends and haven't had a bad experience with online dating applications.

100% of the segment is likely to be interested in the application.

Visually:



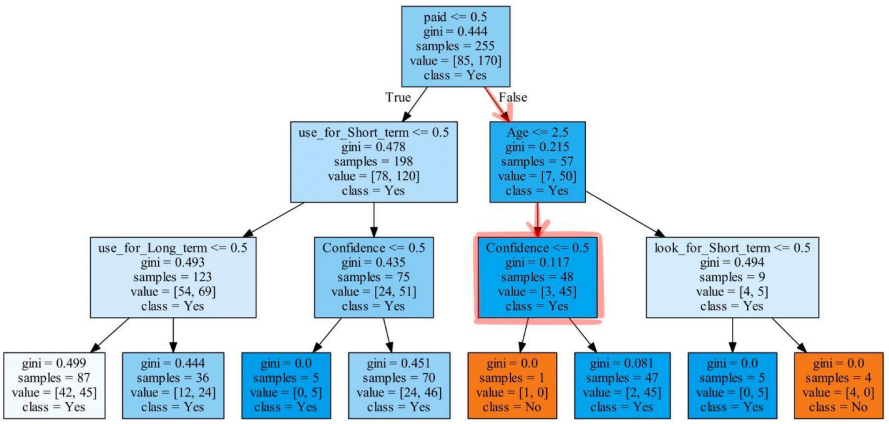
<u>Segment #3</u> -

- <u>Defined by the following path:</u>
    1. paid > 0.5
    2. Age <= 2.5
- Values = [3,45]
- Probability = 45 / (3 + 45) = 0.937

<u>Meaning:</u> This segment is a target market composed of people aged 18-27 who have paid for dating applications in the past.

93.7% of the segment is likely to be interested in the application.

<u>Visually:</u>



<u>Segments to avoid -</u>

Our primary goal was to define segments of the population which are more likely to be interested in our product, but we also took note of less likely segments. For example:

- <u>Defined by the following path:</u>
    1. frequently_use <= 0.5
    2. Dating_apps <= 2.5

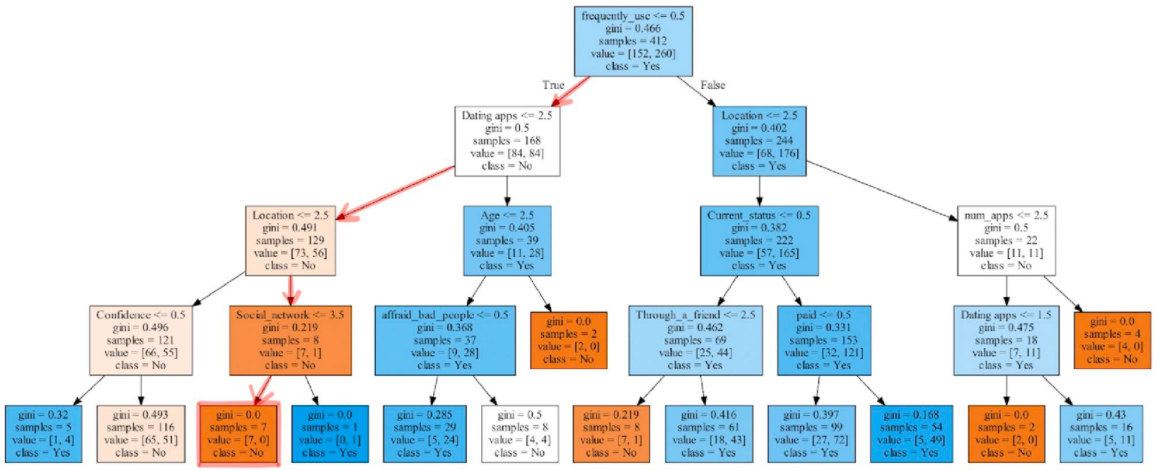
3. Location > 2.5
4. Social_network <= 3.5

- Values = [7,0]
- Probability = 7 / (7 + 0) = 1

Meaning: This group consists of people who don't use dating apps, don't like meeting partners via dating apps or social networks, and live in the south of Israel or abroad.

100% percent of the segment is unlikely to be interested in the application.

Visually:



Note: If we know which segments are unlikely to be interested in our product, it could be both practical and financially beneficial because we could avoid wasting finances on advertising to these segments. We should therefore consider further research to identify these segments with low probability of interest in our application.