Who’s a parent?

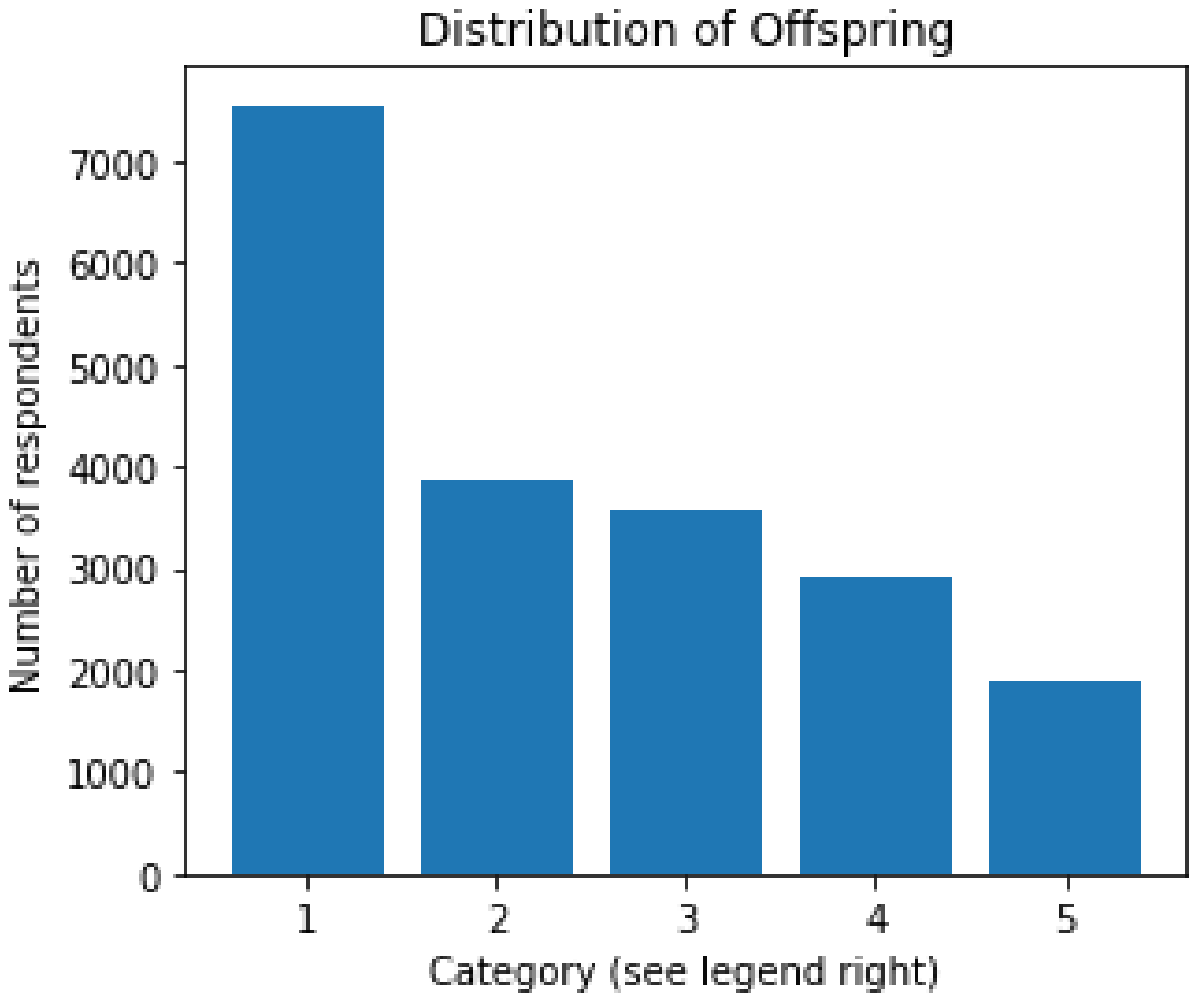
27-5-2022

Stijn Tuijtel

stijntuijtel@gmail.com

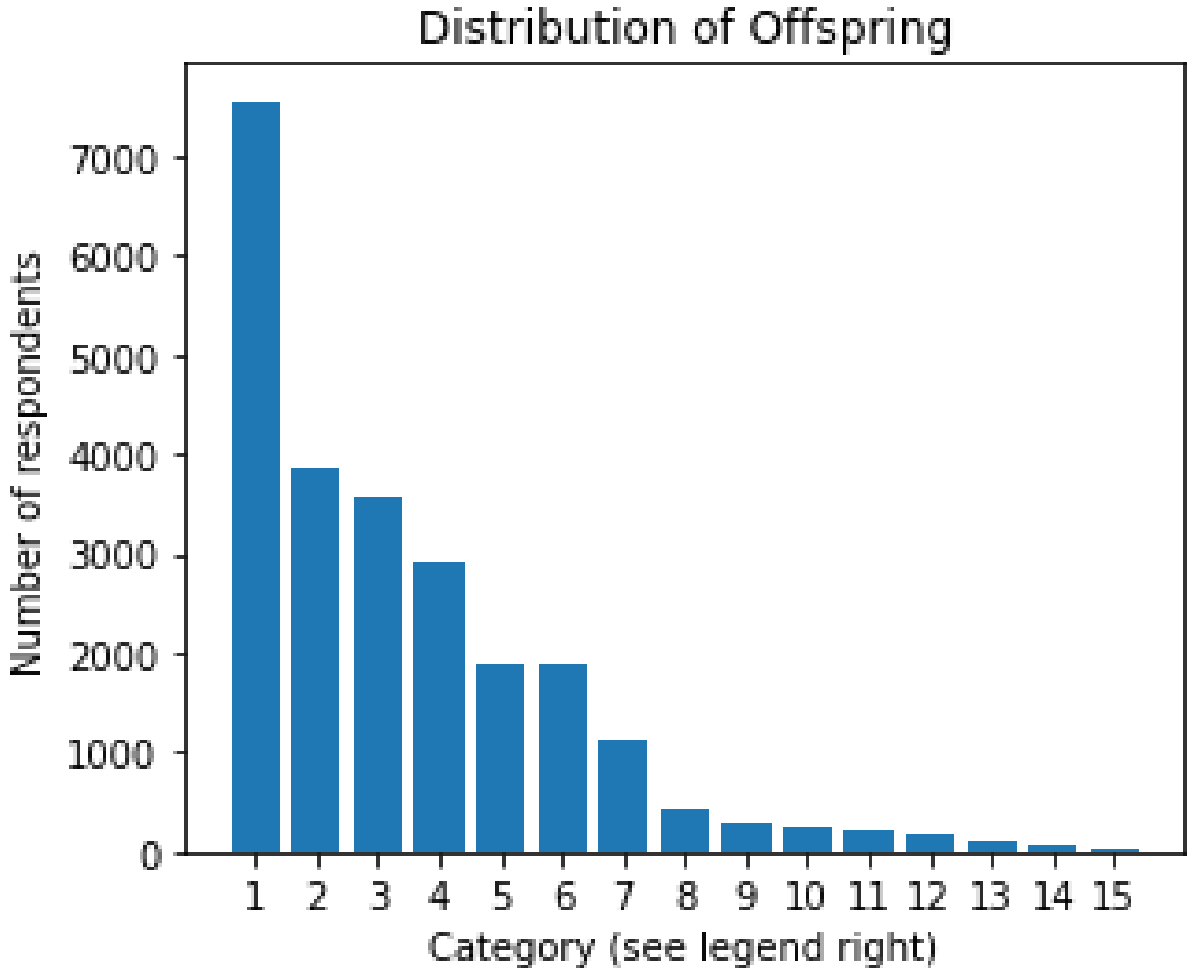
Table of contents

Data exploration – Offspring top 5 values raw



1 - doesn’t have kids
2 - doesn’t have kids, but might want them
3 - doesn’t have kids, but wants them
4 - doesn’t want kids
5 - has kids

Data exploration – Offspring cleaned



1 - does not have kids
2 - does not have kids, but might want them
3 - does not have kids, but wants them
4 - does not want kids
5 - has kids
6 - has a kid
7 - does not have kids, and does not want any
8 - has kids, but does not want more
9 - has a kid, but does not want more
10 - has a kid, and might want more
11 - wants kids
12 - might want kids
13 - has kids, and might want more
14 - has a kid, and wants more
15 - has kids, and wants more

Main question: Who is a parent?

If you become a parent, your whole life turns upside down. Your thoughts, feelings and actions change in a way that was impossible to imagine up front. At least, that is my experience.

I wonder if this experience also applies to other people. Hence I want to investigate, based on their answers only, whether it is possible to predict if someone is a parent or not.

Steps of analysis – Modelling approaches

Via classification techniques K Nearest Neighbor (KNN) and Support Vector Machine (SVM) I want to see to what extent multiple choice answers predict if someone is a parent or not.

In the essays I will count the number of times a respondent names one of the words in Appendix 1. Via logistic regression and K nearest neighbor regression I want to see if the use of these words is related to being a parent or not.

Steps of analysis – feature selection

For the investigation I want to use as many independent variables as possible. The only condition is that the features need to be(come) numerical. This is not the case for diet, orientation, job and pets. As such, the investigation is based on the following features:

Multiple choice: age, drinks, drugs, height, income, last_online, smokes

The following essays I selected to count the parental words:

Essays: 0, 1, 5, 6

Data augmentation – last_online_sec

The column last_online has dtype ‘Object’. To make the values useable for classification, I was interested in their number of seconds.

It was quite a challenge to arrive where I wanted. Eventually, I created a timestamp of the earliest last_online and added the difference with the record value. If I look at it now, an apply function would probably be easier.

```
df['last_online_dt']    = pd.to_datetime(df['last_online'], format='%Y-%m-%d-%H-%M’)
df['datediff']          = df['last_online_dt'] - df['last_online_dt'].min()
last_online_dt_ts       = df['last_online_dt'].min().timestamp()
df['last_online_sec']   = df['datediff'].dt.total_seconds().astype(int)
df['last_online_sec']   = df['last_online_sec'] + last_online_dt_ts
```

Data augmentation – essay_cnt (i)

For regression, I wanted a dataframe column indicating how many times one of the words of Appendix 1 are present in essays 0, 1, 5 and 6. To do so, I replaced NaN-values with ‘’ and concatenated all essays. Afterwards, all records were sent to method ‘Count_parental_words’ (see next slide).

```
essay_cols      = ['essay0', 'essay1', 'essay5', 'essay6']
df[essay_cols]  = df[essay_cols].replace(np.nan, '', regex = True)
df['all_essays’]  = df['essay0'] + df['essay1'] + df['essay5'] + df['essay6’]
df['all_essays’]  = df['all_essays'].astype(str)
df['essay_cnt’]  = df.apply(count_parental_words, axis = 1)
```

Data augmentation – essay_cnt (ii)

```
def count_parental_words(row):

    amount_of_words = 0
    parental_words = ['kid', 'kids', 'child', 'children', \
                      'daughter', 'daughters', 'son', 'sons' \
                      'mom', 'mother', 'dad', 'father', 'baby' \
                      'proud', 'parent']

    for word in row['all_essays'].split():
        if word in parental_words:
            amount_of_words += 1

    return amount_of_words
```

Classification approaches

Data cleaning

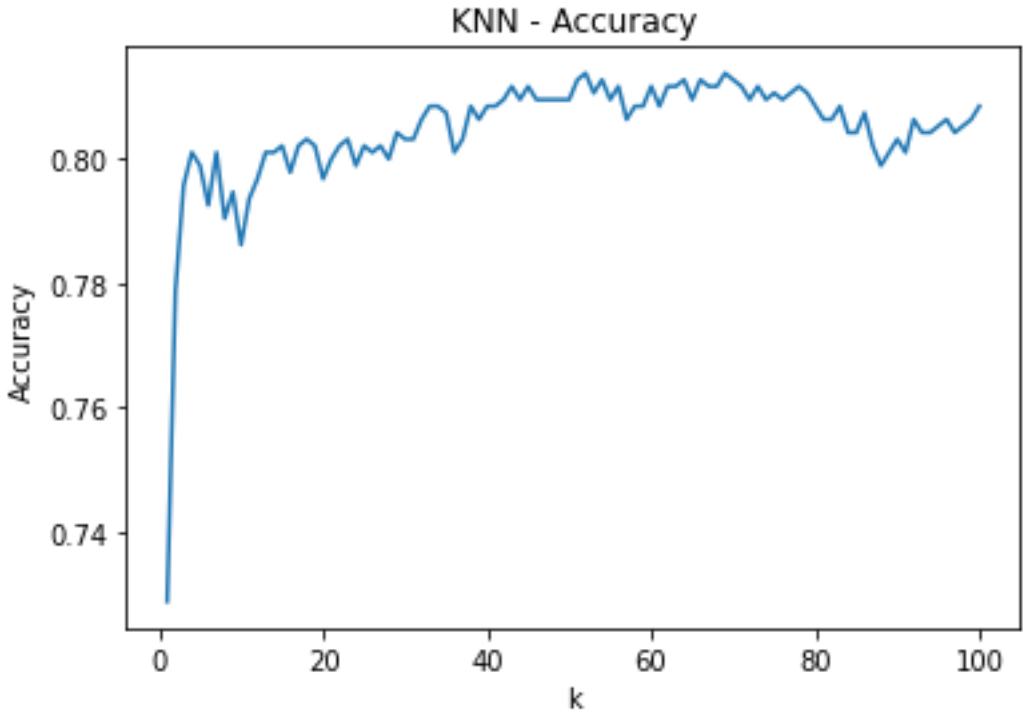
After removing rows with empty values in feature and label column(s), 4.696 records remained. Of these persons 1.113 had kids.

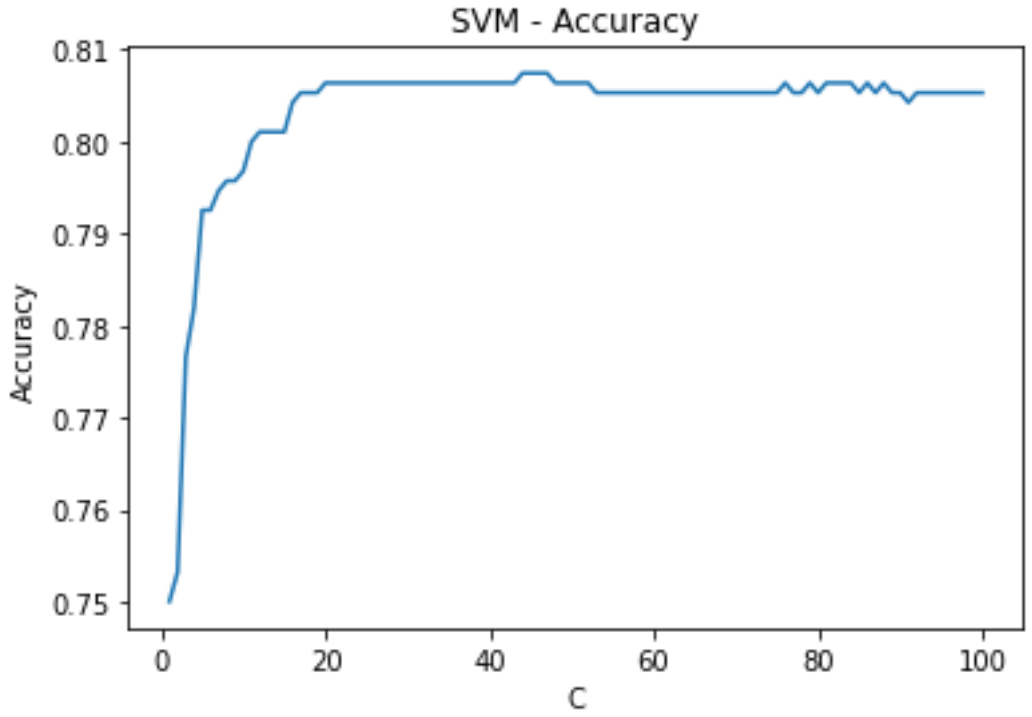
Model description and runtime

Both KNN and SVM were easy to implement via Sklearn. Furthermore, the steps required to build a model are almost identical for the two approaches.

Having said that, there is quite a difference in runtime. Building 101 different models takes 0,7 seconds with KNN and 14,5 seconds with SVM.

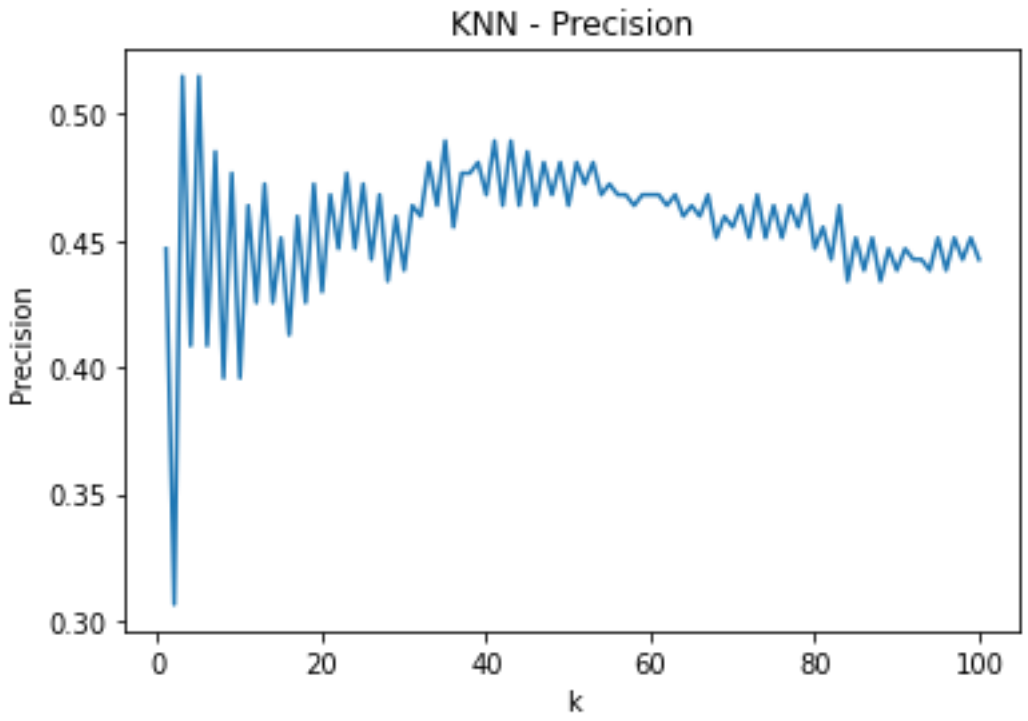
Classification approaches - Accuracy

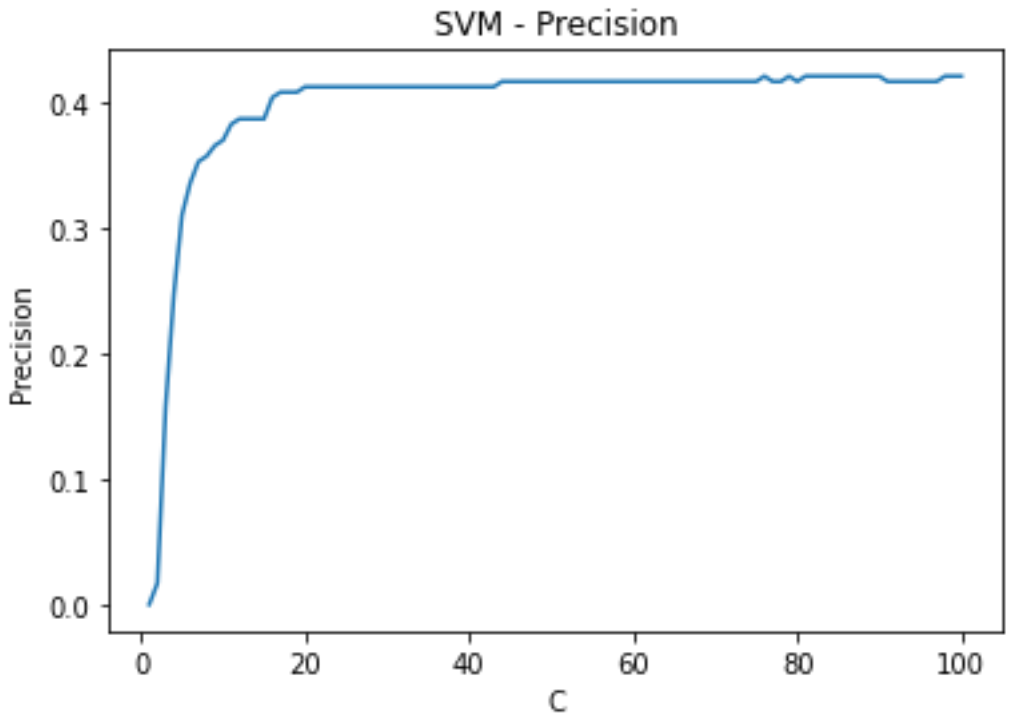




The average accuracy if k or C is higher than 20 is roughly the same between the two approaches.

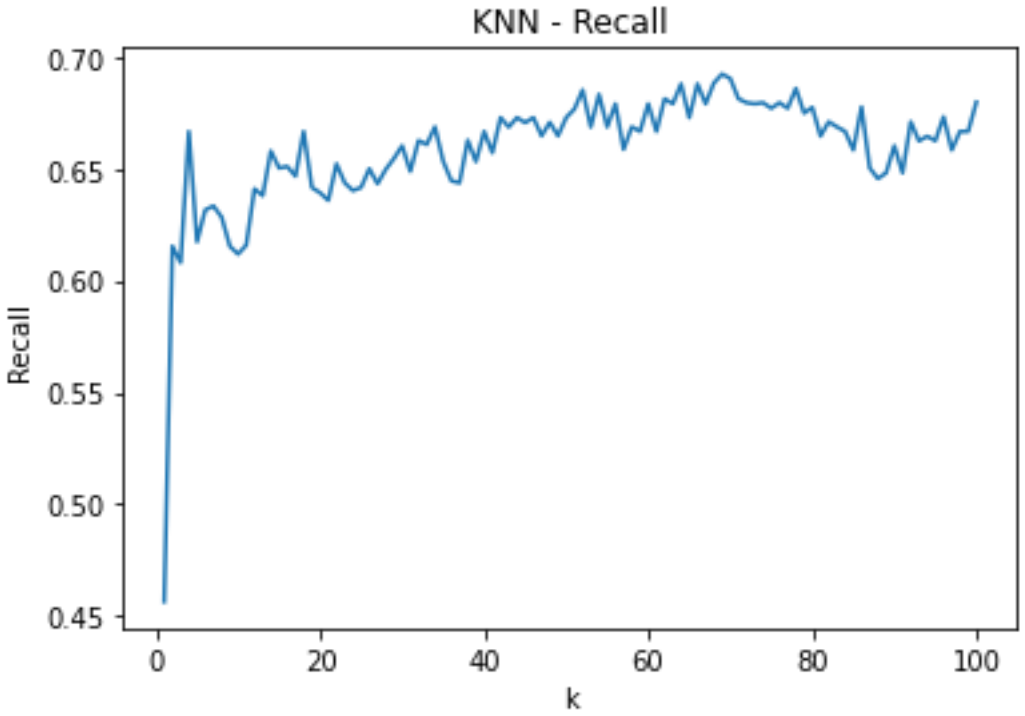
Classification approaches - Precision

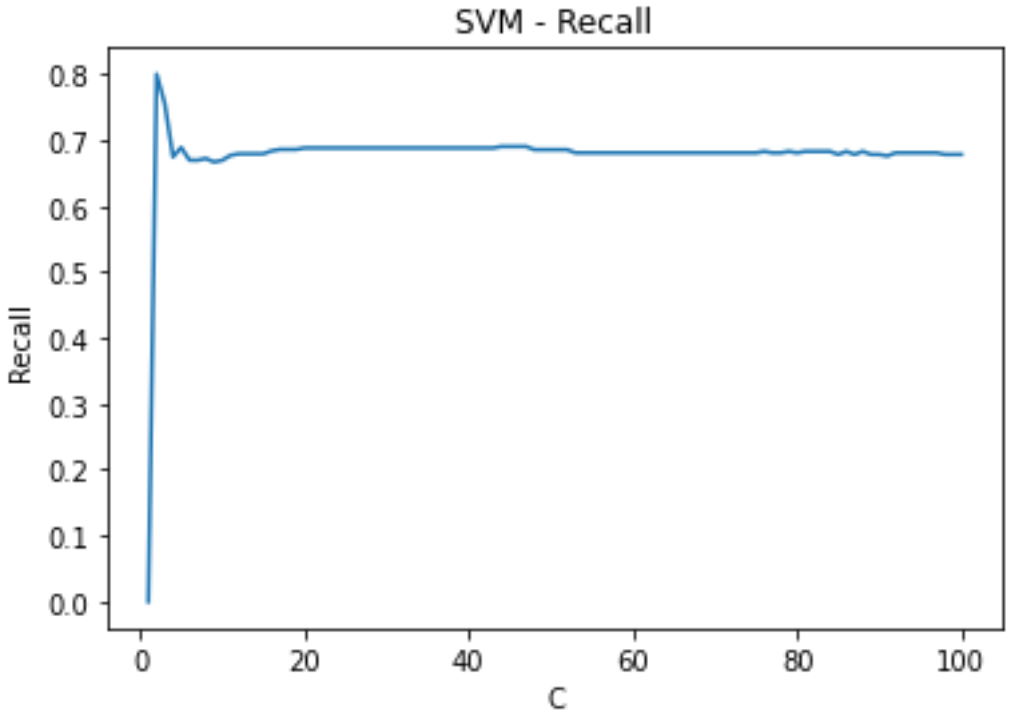




KNN's precision is, in general, slightly higher than that of SVM (0,45 vs 0,40).

Classification approaches - Recall





KNN’s recall is, in general, slightly lower than that of SVM (0,65 vs 0,70).

Classification approaches - conclusion

Both KNN and SVM have an accuracy of at least 0.8. Hence, each technique can predict rather well if someone is a parent, based on the feature columns. This is really great!

The biggest difference between the two approaches is runtime. For testing purposes, certainly when dealing with larger datasets, I would prefer KNN.

Regression approaches

Data cleaning

After removing rows with empty values in feature and label column(s), 24.385 records remained. Of these persons 4.919 had kids. In total, 3.475 essays contained at least one parental word (and 20.910 did not).

Model description and runtime

Both KNN regression and logistic regression were easy to implement via Sklearn. The runtime of each technique was very fast (less than 0,1 seconds).

Regression approaches

Accuracy, recall and precision

	KNN regression	Logistic regression
Accuracy	0,81	0,81
Recall	0,75	0,76
Precision	0,17	0,17

Only recall varies slightly between the two approaches.

Regression approaches - conclusion

Both KNN regression and logistic regression are quite good predictors with an accuracy of 0,81. This means that the set of parental words actually can be used to predict if someone is a parent or not.

The two regression approaches are similar in terms of not only accuracy, recall and precision, but also runtime and implementation. I would not favour one of the techniques above the other, when I would do additional research.

Conclusion

The main question of this analysis was whether it is possible to predict if someone is a parent, based on both quantitative and qualitative data.

It turns out that the classification approaches of KNN and SVM are often able to make such a prediction (having an accuracy of around 0,8), based on multiple choice data. Coincidentally, the same accuracy shows up when creating a KNN regression or a logistic regression model (based on essay data).

Hence, both classification and regression can be used to answer the main question.

Next steps (i)

During this analysis the focus has been on the calculation of a model's accuracy, recall and precision. However, I am also interested in the F1 score and the confusion matrix of each model. Maybe this shows new insights in the difference between models.

Futhermore, I am curious about the correlation between the independent variable(s) and being a parent or not. Is the correlation positive or negative? And, in case of classification, which features have a greater correlation than others?

Next steps (ii)

If I needed to predict whether someone is a parent or not, I would combine the results of both techniques.

For example, if the KNN model predicts respondent X is a parent and the KNN regression model predicts the same, I would be rather sure that this respondent is indeed a parent. Equally, I would be rather sure a respondent is not a parent if the classification and the regression model both predict this.

However, if the two models have a different prediction, I would be hesitant in my prediction.

Discussion (i)

It is nice to see that the accuracy is high for the created models. However, I would be more confident in the models if they would have been trained with (and tested on) better data. To be concrete:

- Offspring, the dependent variable, is NULL in 36K of all 60K cases.
- The variable ‘Income’ has value -1 in 47K cases. It would be nice if all respondents entered their actual income.
- I did not want create an order in the categories of variable ‘Education’ (as I do not think you can compare working and graduating on a college). However, if the feedback options were different for this question I could have integrated this variable as well.

Discussion (ii)

With better -and preferably more- data, it would be possible to examine new questions. For example:

- Do the results of the analysis change depending on the sex? Maybe the accuracy of the regression models is higher for women than for men (or vice versa).
- Are there other parental words that better predict if someone has children?
- Is it possible to predict how many children someone has?
- Can we prove not only correlation, but also causation between the chosen feature(s) and being a parent or not?

Appendix 1: Parental words

During the regression analyses the following words are counted:

kid, kids, child, children,
daughter, daughters, son, sons,
mom, mother, dad, father,
baby, proud, parent