Using Machine Learning techniques to predict the demand for youth care based on neighbourhood-characteristics

submitted in partial fulfillment for the degree of master of science

Jop Hoenderdos
11066881

master information studies
data science
faculty of science
university of amsterdam

Your Date of defence in the format 2020-12-01

	First Supervisor	Second Supervisor	External Supervisor
Title, Name	Ms D. Danielle Sent	Dr Maarten Marx	Dr. Max C. Keuken
Affiliation	Uva, AMC	UvA, FNWI, IvI	Muncipality of Amsterdam
Email	d.sent@uva.nl	maartenmarx@uva.nl	m.keuken@amsterdam.nl

Identifying predictive sociodemographic and neighborhood features for youth care demand - machine learning approach

Jop Hoenderdos
jophoenderdos@hotmail.com
University of Amsterdam

ABSTRACT

In 2015, the organization of the Dutch youth care system was changed radically by transferring all responsibility to the municipalities. This new way of working was introduced to make youth care more efficient, coherent, and cost-effective. However, the demand for youth care continued to increase, and as a result the waiting lists are getting longer and longer. Within the municipality, there is a need clear need to understand what contributes to the demand as to ensure optimal use of the limited resources available in the youth care system. Therefore, using three machine learning algorithms (support vector machine, Decision Trees, Gradient Boost) we will predict the demand for specialist youth care based on the demographic and neighborhood characteristics. Based on the used models, the XGBoost was the model with the highest F1-score. In this model the most import feature in predicting youth care demand was the feature "amount approved" and reflects the cost of a given treatment.

KEYWORDS

Youth Care, machine learning, Neighborhood characteristics, Decision Tree Classifier

INTRODUCTION

The Netherlands has the second-best health care system of Europe. [2] This quality of healthcare is reflected in the well-being of children. UNICEF investigated this well-being of children in rich nations. In this report, it became clear that Dutch children are the healthiest and happiest children in the world [1]. Yet, a minority of children require additional youth care.

Youth care in the Netherlands covers all forms of care available to parents and children to help parents with their challenges and children with their development. Therefore, clients of youth care are those who have problems with their development. Depending on the severity, the clients will either of treated with basic or specialized youth care. In 2019, there were 4,4 million Dutch citizens between the age of 0 and 22 years old. Of this group, approximately 10% received a form of youth care. [1] However, given the existence of long waiting lists, the demand for youth care is higher than the resources; thus, not all children and adolescents receive the care they need. [21]

Before 2015, the financing and responsibilities for youth care were fragmented over different laws and governmental levels. In the past, several evaluations were conducted as there were clear signals that the youth care system was not performing optimal. [22] A common finding was that the previous system resulted in an increased usage and costs of specialized youth care.

To address the challenges that were identified a new youth care act was created. This act entailed that most youth care tasks were transferred to the local municipalities and that the families and social networks would play a larger role in the care process. The goal of this new act was that it would result in more coherent, cost-effective, and transparent services for children and their families. [22]

Youth Care in Amsterdam

Contrary to what was intended with Child and Youth Act the youth care costs have continued to increase since the implementation. [11] Between 2015 and 2018, in particular, the cost of specialized youth care has increased substantially. In these three years, the cost of specialized youth care increased by almost 40 % (see Figure 1). Not only are the costs increasing, but also the amount of people who are using the specialized youth care. In 2015, 10.886 young people received specialized youth care, where this number increased by 14% to 12.412 people in 2018. This recent increase in cost and use can be partly explained by an attempt to provide more care for more children by removing the budget for youth care in 2018. As a result, the budget was overrun substantially and the budget ceilings were reinstated in 2019. Even with these restrictions, an overrun of the budget cannot be ruled out for 2020. [11]

The municipality of Amsterdam implemented a system to categorize the care needed for a person, which includes agreements over the duration and the costs of that care, called SPICs (in Dutch: Segment Profiel Intensiteit-Combinatie).

[1] https://longreads.cbs.nl/jeugdhulp2019/jongeren-met-jeugdhulp/

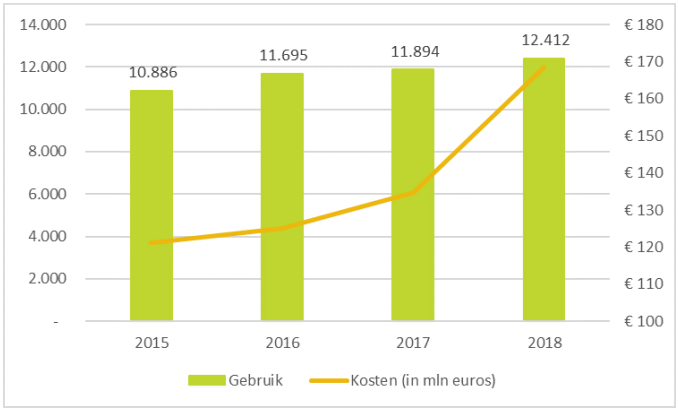


Figure 1: Development of specialist youth assistance from 2015 till 2018

On the left y-axis the amount of usage is shown. In the figure these are the green colums. On the right y-axis the money (in mln euro) can be found, which is made visible in the figure with the orange line. The x-axis shows the years in which the municipality is responsible for organizing the youth care. Adapted from [11].

With this system, it is easier to estimate and control the duration, cost, intensity of the youth care. [2] In total, there are three segment which indicate the extent of the comprehensive nature of (specialist) youth care:

- **Segment A:** contains basic youth care. Preventive, light outpatient youth care: parenting, and family support. This is freely accessible.
- **Segment B:** contains single specialist youth assistance, where it is reasonably defined what kind of support is offered and what result is intended. Not freely accessible.
- **Segment C:** contains the complex (and more expensive) specialist help. This concerns multiple, comprehensive, specialist youth assistance (for one child). Not freely accessible.

One of the challenges for the municipality is to use their limited resources as efficiently as possible. One of the goals is to increase the use of preventive interventions (segment A and B), thereby reducing the need for specialized youth care (segment C). As such freeing resources that could be used specifically for children who are most at risk. [16]. The demand for youth care will, therefore, be lower, and this will translate to a reduction in costs and waiting lists. It is however clear from the current waiting lists that this goal has not yet been achieved. Waiting lists can have severe consequences for children who require immediate care. Due to the long wait and the lack of understanding, some people get even more sicker or became suicidal. [3] In 2019 alone, 127 and 136 clients had to wait longer than 10 weeks before the care in segment B and C could commence [9].

Understanding the potential factors for the demand of specialized youth care will help the municipality of Amsterdam to allocate their resources more efficiently. A recent theoretical model has shown that neighborhood characteristics can improve prediction models such as the risk equalization model for using health care. [20]

In line with these suggestions, we set out to investigate the value of demographic and neighborhood characteristics in predicting youth care need through the use of machine learning (ML). Three ML algorithms will be used to create a predictive model for youth care needs. Specifically, we will predict the type of youth care use of Amsterdam clients in the years 2018 and 2019 on the basis of their demographic and neighborhood characteristics.

Concretely in the current thesis, we will address the following questions:

To what extent can prediction models based on Support vector machine, Decision Tree Classifier, or Gradient Boosting Machine contribute to predicting the demand for the specialist youth care in Amsterdam?

To answer this question, we defined the following subquestions:

- Which of the tested models has the highest score perform metric (f1 score) in predicting the youth care use?
- Which (neighborhood) characteristics are predictive for the use of youth care?

RELATED WORK

The thesis uses existing ML methods to identify predictive demographic and neighbourhood characteristics for youth care need. As such there is already related work on what the potential demographic and neighbourhood risk factors might be.

Demographic risk factors youth care

One of the largest tasks of specialized youth care is providing support and treatment for children's mental health problems. A study by Willie et al, investigated which risk and protective factors are relevant for developing mental health problems. Mental health problems and their assumed features were examined in a representative sub-sample of 2,863 families with children and adolescents aged 7–17. The authors conclude that having conflicts in the family, mental disorder of parent, conflicts in partnership, single parent, low

[2] https://www.zorgomregioamsterdam.nl/jeugdhulp/spic/

[3] https://www.volkskrant.nl/nieuws-achtergrond/psychiaters-slaan-alarm-over-hulp-aan-suicidale-kinderen bbe32a8e/

SES (socio-economic status), step-parent, unwanted pregnancy, low social support in the first year, chronic disease parent, unemployment, parental strain and parental psychiatric symptoms are the most important risk factors for the development of mental health problems in children. [30]

Not only is it important to identify the factors that contribute to developing mental health problems. De Haan and others investigate the risk factors to dropout of psychotherapy for child and adolescent people. [5] Several factors include ethnic minority status, a lower SES, and severity of the mental health problems. Data from almost 400 children and approximately 350 adolescents were used in this study. The authors identified a number of specific demographic groups that had a higher risk of dropout and concluded that therapy compliance was influenced by a number of demographic factors. Considering the previous results, we will incorporate a large number of demographic factors that might be predictive of youth care demand.

Neighbourhood characteristics and machine learning

Recent studies have concluded that a number of neighborhood characteristics can influence the mental health of the population [8, 12, 26] For example, safety concerns, noise, air pollution, and urbanicity [13, 24, 25, 31] had an effect on mental health. Which can lead to a more depressive mood. [12]

A recent study in the Netherlands, try to find which physical and social neighborhood characteristics influenced depression. [15] This study incorporated data from two sources: 1) a survey comprised of various question on sociodemographic, mental health, and perception of the residential neighbourhood factors and 2) registry data which was made available through Statistics Netherlands. Using a ML approach, the authors assessed how the different factors correlated with depression severity while controlling for individual differences in sociodemographic factors. While the results will need to be validated in a within subject longitudinal design, the results suggests that modification of physical and social neighbourhood characteristics could represent an effective intervention to promote mental health. As the potential predictive neighborhood characteristics are quite diverse we will incorporate a wide range of factors using standardized registry data. [12]

METHODOLOGY

This section contains three parts: a data description, Data Cleaning and pre-processing, and model fitting and evaluation methods.

DATA

The data used in this thesis originates from two different sources. The youth care data is made available from the data-team of the social cluster within the municipality of Amsterdam. The youth care data is on an individual level (i.e., one row is one child) and due to the privacy-sensitive nature of the data confidential. As a result of this, any value based on less than ten observations cannot be made public. The neighborhood characteristics dataset is open-source and available through the data website of the municipality of Amsterdam [4] In the next two upcoming sections, we will describe used datasets more extensively.

Youth care data

The youth care dataset contains data about clients who received a form of youth care as organized by the municipality of Amsterdam and has the following columns: *Sex, Date of Birth, Zipcode, Year of treatment, hashID, Amount approved, Product category, Supply type and Services*

To prevent the direct identification of these clients, the personal number has been hashed to an hashID. The column Amount approved is the cost of the care that the municipality has approved. The three columns Product category (8 categories), supply type (7 categories), and services (176 categories) provide hierarchical information about the kind of care the client has received. To illustrate the number of given categories, the different categories and their amount of product categories are in table 2 and table 3. Some categories in both tables are not shown, due to privacy reasons of the data. Depending on how well our model fits, we should see if we can use the services level data. This data set contains 34.557 total rows of data. In table1, the demographic descriptives are given where M is for Male, F for female, and O unknown. Some services will only be finished during the last months of a year. Therefore we are only interested in data which belongs to a full year. 2018 and 2019 are the only two years fully available in the data set. It is therefore decided to structure this table in these two years. Because a person can receive care several times in the same year, the number of rows differs from the number of client.

[4] https://data.amsterdam.nl/datasets/G5JpqNbhweXZSw/basisbestandgebieden-amsterdam-bbga/

Table 1: Data description based on year in Youth Care Data

Variable	Year	
	2018	2019
Number of Rows	10874	14014
Unique Client ID	8996	11246
Average age	12.9	12
Sex: M/F/O	5162/3834/0	6502/4743/1

Table 2: Description and amount of product categories

Variable Name	Year	
	2018	2019
Maatwerkarrangementen jeugd	8962	11886
Specialistische ggz	805	926
Jeugdhulp crisis	802	886
Landelijk ingekochte zorg	253	289
(2015) Zonder verblijf:	37	-
Jeugdhulp verblijf Overig residentieel	13	-
Jeugdhulp verblijf: (excl. behandeling)	-	27

Only samples above n = 10 are visible

Table 3: Description and amount of supply type

Variable Name	Year	
	2018	2019
Jeugd (2018)-Segment B	5169	7211
Jeugd (2018)-Segment C	4418	5588
Jeugd - Specialistische GGZ	849	920
Jeugd (2018)-Landelijk ingekochte zorg	253	289
Jeugd (2018)-Conversie 2018 afwijkende prijzen	177	-

Only samples above n = 10 are visible

Neighborhood data

The "Basisbestand Gebieden Amsterdam" (BBGA) contains key statistics of the municipality at several city division levels citywide, city district-level (8 values), area-level (22 values), neighborhood-level (98 values), and vicinity-level (477 values) from 2001-2020. For each of these levels, the dataset contains around 800 variables with the following themes:

- Urban development and living
- Traffic and public space
- Economy and culture
- Well-being, care and sport
- Education, youth and diversity
- Work, income, and participation
- Sustainability and water
- Services and information
- Social strength

Classification Algorithms

In the next sections, we will give a short description of the chosen algorithms and an argumentation why these algorithms were selected. This thesis will use three different supervised ML algorithms: Support vector machine, Decision Tree Classifier, and Gradient Boosting Machine.

Support vector machine. Support vector machine (SVM) is a supervised learning model that is used to analyze data for classification and regression purposes. The main concept behind SVM is to fit a hyper-plane between the labelled data point, separating the different data points into different groups. Consider a simplistic example as illustrated in Figure 2. Each data point on either side of the hyperplane will be classified into a different group (circle or star). So when a new point enters the model, this hyper-plane is used to decide to which group the new data point belongs.

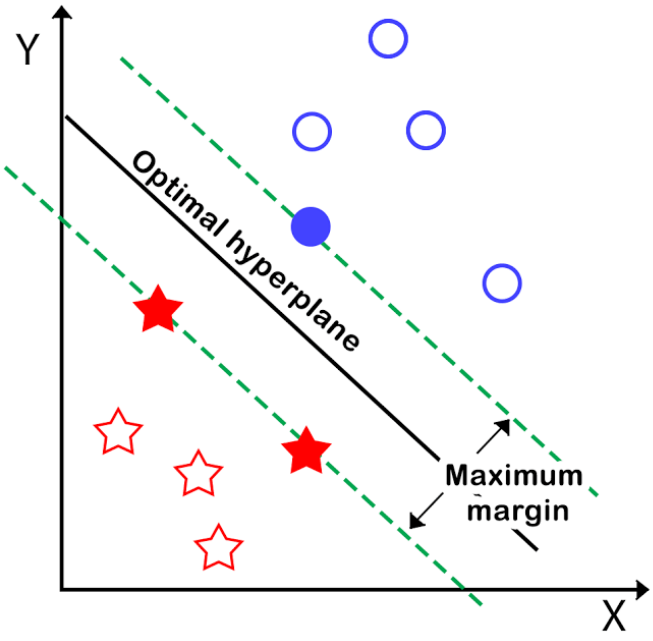


Figure 2: Support Vector Machine Example
Data points are shown in this figure from two different groups (circles and stars). By fitting an optimal hyper plane the membership is determined. Adapted from [9]

To fit the best hyper-plane line, SVM tries to take the points closest to the line from both classes. These points are called support vectors and are filled in figure 2. After the support vectors are determined, the distance is calculated

between the line and the support vectors. The distance between these support vectors and the hyper-plane is called the margin and the goal is to maximize this parameter. The benefit of SVM models is that it is able to generate robust predictions with limited training samples, making is the case in the youth care data set used in this thesis [23]

Decision Tree Classifier. Decision trees (DT) are supervised machine learning techniques which are frequently used for regression and classification problems. The idea behind the DT algorithm is simple, but therefore very powerful. The aim of classification trees is to split the data into smaller, more homogeneous groups. For each attribute in the dataset, the DT algorithm forms a node, where the most important attribute is placed at the root node. For evaluation, we start at the root node and work our way down the tree by following the corresponding node that meets our condition or "decision". This process continues until a leaf node is reached, which contains the prediction or the outcome of the DT. [18]

Gradient Boosting Machine. Gradient Boosting is a popular ensemble DT algorithm which is less prone to overfitting than a single DT. The idea of gradient boosting is that boosting can be interpreted as an optimization algorithm on a suitable cost function. [3] Boosting is a technique where models are built sequentially, aiming to minimize the errors from the previous models while increasing the influence of high-performing models. In this thesis, we will use XGBoost as it one of the fastest implementations of gradient boosted trees.[4]

Random Undersample

When an imbalanced dataset is used, which is the case an has been described, there are too few data points of the class with the fewest data points to learn the decision boundary effectively. [14] Therefore, balanced data will improve model performance. One solution to imbalanced data is to either oversample the smallest class or undersample the largest classes, which will result in a balanced dataset. Building a Support Vector Machine will increase with $O(N^3)$ time and $O(N^2)$ space complexity where N is training set size. [6] Given the large number of classes present in the youth care data set and the number of samples in the largest class, oversampling was computationally not feasible given the available resources. This applies to SVM but also to all other models. Instead we focused on undersampling using the functions implemented in the python package imblearn. The disadvantage with undersampling is that all classes have the same amount of samples as the smallest class, resulting in a considerably lower number of overall samples to train the model on. To quantify this trade-off we trained each of the three models with the full and undersampled dataset.

Data Cleaning and pre-processing

By removing incorrect, incorrectly formatted, or incomplete data from both datasets we will prepare the datasets for the different planned analysis. This comes with a number of steps, which are different for both datasets.

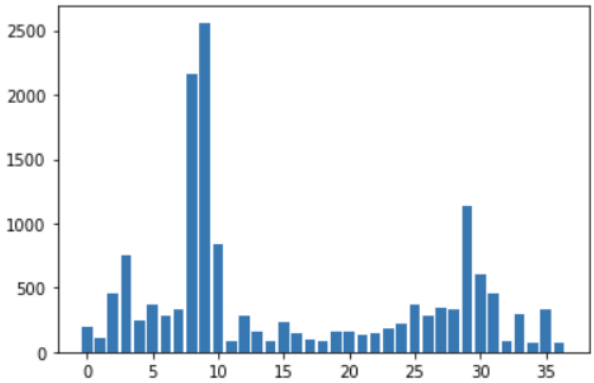
In the **Youth care data**, we performed the following steps to get a dataset that could be joined with the neighborhood data. Only healthcare usage in 2018 and 2019 was included in the analysis. Every client that was included had to have a valid zip code. A valid zip code was necessary to be able to merge the youth care data with their neighborhood characteristics. With this, we removed 343 rows from the dataset. The column Date of Birth is too specific for our needs. Therefore we calculated the age of the client and saved this in a newly created column.

The most detailed information regarding the care a client has received is at the level of the individual SPICs. Every SPIC has a particular code and is structured in the following manner: The specialist youth care is divided into specialist youth assistance (segment B), highly specialized youth assistance (segment C). This is the first letter in the SPIC code. In addition to the segments, a distinction is made according to a number of profiles. A total of eleven profiles are defined on the type of care and the desired outcome. These eleven profiles are indicated with a number. Finally, a distinction is made on the intensity of the care: perspective (P), intensive (I), durable light (DL), durable medium (DM), and durable heavy (DZ) [10] All other non-SPIC services were removed. The used regex can be found on GitHub.

A categorical variable is a variable that takes a fixed number of possible values. This is the sex variable in the youth care dataset. Machine learning models require numerical input and output variables, therefor categorical values must be one- hot encoded. [18] This is where the integer encoded variable is removed, and one new binary variable is added for each unique integer value in the variable. In the "sex" variable example, there are two categories (Female and Male), and therefore two binary variables are needed. A "1" value is placed in the binary variable for the sex and "0" value for the other sex.

In the **Neighborhood Data**, only data from 2018 and 2019 were included. The dataset was cleaned by filtering on the zip code (thus removing all other city division levels). Any feature in the neighborhood dataset which had NaN values for more than half of the zip code were removed. As we employed a data driven approach to identify predictive neighborhood characteristics no futher feature selection was done. The final neighborhood dataset included 187 features with 164 rows of data. The two datasets were merged using the zip code and year.

Figure 3: Frequency of services



Looking at the x-axis you can find the frequency of each individual service. Looking at the y-axis you can find the frequency of each individual service Numbers on the y-axis can be ignored. Due to readability not every singe service is being plot.

For categories that have so few samples it can create challenges in getting accurate predictions. Literature shows no information on the minimum number of samples required, given the number of categories to classify. Instead we tried to determine if we could use the 95% confidence interval. Based on the mean and the standard deviation it would result in excluding a large proportion of the data. Therefore we made an decision to take 95% of the values with the largest sample size. This resulted in removing 34 categories with 795 samples in total. A benefit of only including the 95% largest categories was that the resulting selection with the privacy requirements of the municipality of Amsterdam. The final selection of services is shown in figure 3.

For algorithms that measure the distance between data points, which is the case in our study for the support vector machine, it is necessary to scale the data. This is needed since variables with higher values, will influence the outcome of a prediction more, while they are not necessarily more important as a predictor. [18] We will scale the data with the in-built function of Sklearn.

The final description of the data can be found in table 4. A histogram of all the available features is made, and can be found on the GitHub repository. All by all, we have an data set which have 37 unique SPIC, with 14.826 rows of data. Which matches with the result of table 4.

A histogram of all the available features is made and can be found on the GitHub repository.[5]

[5] [6]

Table 4: Data description of final dataset

Variable Name	Year	
	2018	2019
Product Type		
Maatwerkarrangementen jeugd	6723	8103
Supply Type		
Jeugd (2018)-Segment B	3447	5690
Jeugd (2018)-Segment C	3276	2413
Clients		
Number of rows	6723	8103
Unique Client ID	6149	7333
Average Age	13.8	12.8
Sex: M/F	3956/2767	4746/3357

Model

Model evaluation. In this study, we investigate which of the three chosen algorithms is the most suitable for predicting which features play an important role in using specialized youth care. It is therefore necessary to determine which metrics are used to compare the different algorithms with one another. In binary classification, the predictions can be labelled one of four ways, as shown in table 5. [17]

Table 5: Confusion Matrix for Binary Classification

	Actual Positive Class	**Actual Positive Class**
Predicted Positive Class	True Positive (tp)	False negative (fn)
Predicted Negative Class	False positive (fp)	True negative (tn)

Based on this table, we can come up with four different metrics to evaluate the algorithm: [17]

$$\frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$\frac{TP}{TP + FP} \tag{2}$$

$$\frac{TP}{TP + FN} \tag{3}$$

$$\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The most well know metric is equation 1, and it is known as accuracy. The metric accuracy measures the ratio of correct predictions over the total number of instances evaluated. Unfortunately, it makes no distinction between classes; correct answers for each category are treated equally, which is

fine for balanced data. Our study, however, uses an imbalanced dataset, and accuracy is therefore not suited for our purposes.

Other frequently used metrics are precision and recall and are shown in equation 2 and 3. The precision metric indicates how precise your model is as indicated by the ratio of false positives and true positives. Precision is an excellent metric to use when the costs of false-positive are high. Putting it more in the context of our study, precision can be seen as how efficiently resources will be used. A higher precision results in fewer resources that are wasted on households/children that do not require youth care. The metric recall is useful when there is a high cost of a false negative. Finally the F1 score, which can be found in equation 4 is a combination of precision and recall and is, therefore, suitable for our needs.

These metrics only show us the performance of the model but is not showing us the quantify the uncertainty of the outcome. Mean Square Error (MSE) is popular metric to evaluate machine learning. In short: MSE measures the difference between the predicted solutions and desired solutions. Like accuracy, the main limitation of MSE is that this metric does not provide the trade-off information between class data and will therefore not be used in our study. [17] The area under the ROC Curve is one other popular ranking type. This metric is designed for binary classification but can be used for multiclass classification. [19] The authors however state that this generalization is useful for problems with a low number of classes. Considering our dataset with 37 different classes, we would not see this as a low number. Another reason why we did not use this metric is that the computational cost of AUC is high. [17]

In summary, in the model evaluation, the following metrics will be used: precision, recall, and F1. These metrics are essentially defined for classification tasks. [7] But the sklearn library is providing an in-built function to calculate these scores with multiclass data. [29] By using multiclass data, an extra parameter is required. Of the five possible options, only two are applicable: micro and macro. A short explanation of the macro and the micro parameter: macro-average will compute the F1 metric independently for each class and then take the average whereas Micro-average will weigh the different contributions of F1 per class and then computes the average metric. This because we are dealing with an imbalanced data set.

Choosing for this "micro" will lead to the same score for all of the selected metrics.

$$P = \frac{\sum_c TP_c}{\sum_c TP_c + \sum_c FP_c}$$
$$R = \frac{\sum_c TP_c}{\sum_c TP_c + \sum_c FN_c} \tag{5}$$

Where c is the class label. Since in a multi-class setting you count all false instances it turns out that:

$$\sum_c FP_c = \sum_c FN_c \tag{6}$$

In other words, every single false prediction will be a false positive for a class, and every single negative will be a false negative for a class. Therefore, we will only give the F1 scores of each algorithm.

Models making. Support Vector Machine and the Decision Tree Classifier model were build using the Scikit-learn. This is an open-source Python library for various machine learning models. For the gradient boosting, we will use an XGBoost algorithm. This algorithm is not included in the scikit-learn library and therefore imported separately from the xgboost library.

First, the datasets are loaded into pandas. Second, a baseline model is created. This baseline model is using the default hyperparameters from each algorithm. The data was trained using the train-test-split method of 75%/25% for the train- and test data. The dependent variable is the SPICs used by a given client. For reproducibility, we set the random state variable. By controlling the random state variable, we will get the same results when running the code multiple times. The performance metrics, in our case F1, are evaluated using cross-validation (CV). With the use of cross-validation, we reduce the bias of the model. With a CV of 5, the training data is divided into five different folds, resulting in different train and test data for every run. [18] The reported cross-validated results are the mean of each of the performance metrics .

An imbalanced dataset can influence the predictions of ML algorithms. In figure 3 it is visible how imbalance the data set is. On the y-axis, every different service is shown. On the X-axis, the frequency is shown.

As stated in the introduction imbalanced datasets are challenging for ML models. To quantify the effects we also created identical models as above but then using random undersampled data. This resulted in an dataset with equal number of samples of the minorty class of youth care. The result can be seen in Figure 4 As with the imbalanced dataset cross-validation was used. For each of the three classes, two models were therefore created (resp. with imbalanced and undersampled data). The model with the highest F1-score was then selected to further optimize it by tuning the hyperparameters using GridSearchCV. This is a method in the Scikit-learn library that randomly searches combinations of hyperparameters within a given grid. The best scoring combination is then provided. The hyperparameters are chosen, and the values in the grids are based on conventions from literature.

[18] The hyperparameters (with their used parameters set) are shown table 6.

Table 6: Tuned hyperparameters

SVM	Decision Tree	XGBoost
Kernel: [Linear, RBF, Poly, Sigmoid]	criterion: ['gini', 'entropy']	max_depth: range(4,26,4)
C: [0.1,1,10,100]	max_depth: range(4,26.4)	scale_pos_weight: [1,25,50,75,100]
Gamma: [0.001,0.01,0.1,1,10]	min_samples_split: range(1,10,2)	colsample_bytree': arrange(0.5,1.0,0.3)
	min_samples_leaf: range(1,5)	

Confusion Matrix. A way to visualize the performance of a classification model is to use a confusion matrix. A confusion matrix visualized a table that enables the user to summarize and visualize a classification model's performance. The number of correct and incorrect predictions are summarized with count values for each class. Which means that all the elements on the diagonal represent the number of data points that are predicted correctly, while off-diagonal elements are data points that have not been predicted correctly. As a result, the higher the diagonal values are, the better the algorithm is performing.

GitHub Code

All code used in this thesis to clean and pre-process the data as well as fit and evaluate the models are made available in the GitHub repository. [6] While the notebooks include all the output generated by the syntax it is not possible to include all used data. As stated in the data description, the BBGA data set is freely available, but due to privacy reasons, the youth care dataset is not.

RESULTS AND EVALUATION

Model Performance

The three different algorithms were first trained and tested with the full dataset and the performance metric F1 are given in the first column of table 7. Behind every F1-given score, the standard deviation is given. The model which performed best was based on the XGBoost algorithm. In this table, with the best model highlighted in bold, it is also clear that the Support Vector Machine performs poorly compared to the other tested algorithms. Based on the low scores, we conducted an additional analysis for the SVM model by varying the different kernel types. Based on the results in table 8 the SVM model with a linear kernel improves the F1 score substantially compared to the default RBF, but still underperforms compared to the Decision Tree classifier and XGBoost algorithms.

Table 7: The F1 Scores for the different algorithms

	Baseline Model	Random Undersample	GridSearchCV
Support Vector Machine	23% (0.004)	6% (0.02)	-
Decision Tree Classifier	51% (0.01)	**31%** (0.04)	**57%** (0.01)
XGBoost	**60%** (0.006)	25% (0.005)	58% (0.002)

For every tested model the average F1 scores of the CV. The standard deviation is provided between brackets. Every model was trained on the entire dataset (Baseline model), on the undersampled but balanced dataset (Random Undersample) and finally, where computationally possible, the parameters of the Baseline model were optimized by a grid search approach (GridSearchCV). The model that scored the highest given the used dataset is marked in bold.

Table 8: SVM scores over different kernel. Marked = highest score of column

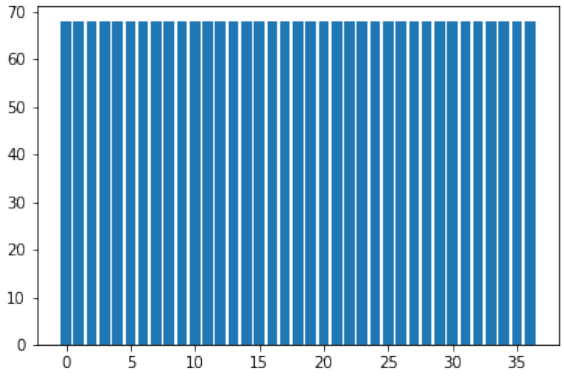
	F1-score
Linear	**32%**
RBF	24%
Poly	22%
Sigmoid	16%

As stated in the introduction and method section, imbalanced data might be detrimental to the overall performance of ML models. To quantify this we used a random undersample technique and fit the three models on the reduced dataset. To visualize the result of the random undersampling, and therefore see which data we used, we made a figure which can be seen in Figure 4. You can see that all the services, which are on the X-axis, have the same amount of samples (Y-axis). A big difference compared with figure we made before in Figure 3. The F1 scores are shown in the second column of table 7 and it is clear that all models performed substantially worse when using an undersampled but balanced dataset compared to a full but imbalanced dataset. Therefore we decided not to further optimize the models based on the undersampled dataset and continue to optimize the full datatset models.

A final method to optimize the parameters of the three models is by performing a grid search. This was computationally possible for two of the three models (Decision Tree classifier and XGBoost). A grid search was not possible because of computational limitations. For the Support Vector

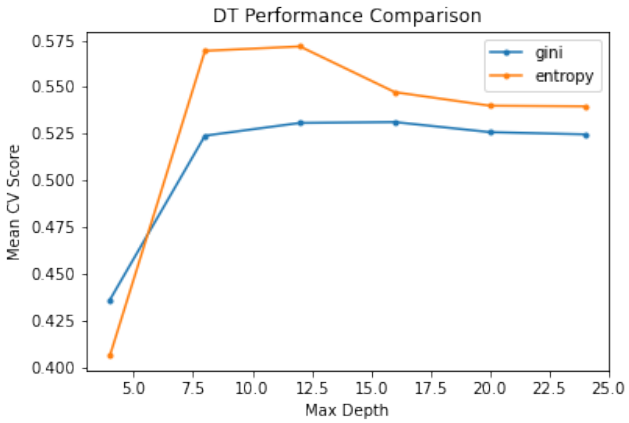
[6] https://github.com/jtothehoenderdos/MasterThesis

Figure 4: Frequency of after Random Undersample



Looking at the x-axis you can find the frequency of each individual service. Looking at the y-axis you can find the frequency of each individual service. Numbers on the y-axis can be ignored. Due to readability not every singe service is being plot.

Figure 5: Performance Comparison Decision tree



On the horizontal axis, the different values of the parameter "Max Depth" can be found. Looking at the vertical axis, the mean CV scores can be found. These values are plotted for the two different possible criterion of Decision Tree.

Machine model a single fit took approximately 7 minutes to perform on the used PC (Intel I5, 8GB RAM). Given the number of potential grid parameter combinations and the use of cross validation the required computational time was not feasible within the current project. Of the two models for which grid search was performed, the F1-score only improved for the Decision Tree Classifier model.

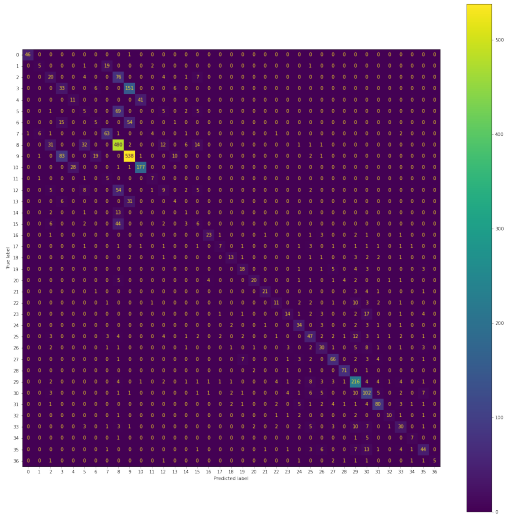
One of the parameters that were fine-tuned with grid search for the Decision Tree Classifier model are the criterion which is used to split the different nodes and the maximum depth of the decision tree. The effect of the different parameter settings on the mean CV score (which corresponds to the mean F1-scores of the multiple cross validations) are shown in Figure 5.

The Confusion Matrix of the winning XGboost model can be found in figure 6. A bigger version can be found on the given GitHub pages.

Based on the optimization of the different models, the model that performed best is the XGBoost classifier using the full dataset with the following parameter settings: criterion: 'colsample_bytree': 0.8, 'max_depth': 4, 'scale_pos_weight': 1.

Based on these metrics, the XGBoost model shows the best performance in the context of this study, thus its feature importance will be examined.

Figure 6: Confusion Matrix

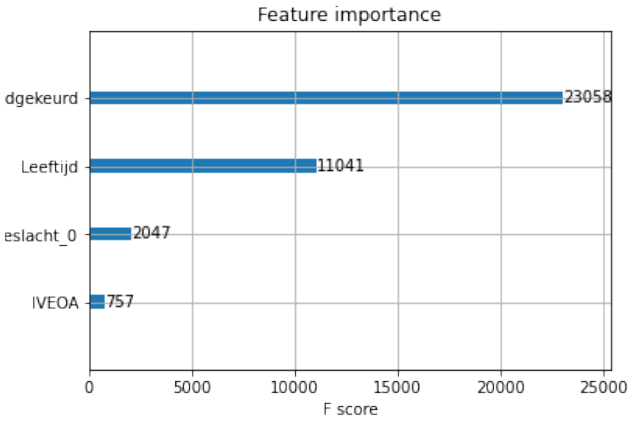


The left axis is showing us the True labels, where on the horizontal axis the predicted values are show. Each individual box in this matrix is showing us the amount of predicted sample vs the actual sample. The lighter the color, the higher the number of samples is in that box.

Features Importance

Based on the parameter optimization we further investigated what the most important features were for the XGBoost algorithm. The predictive values of the individual features that had the largest feature importance of the XGBoost algorithm can be found in Figure 7.

Figure 7: Feature Importance



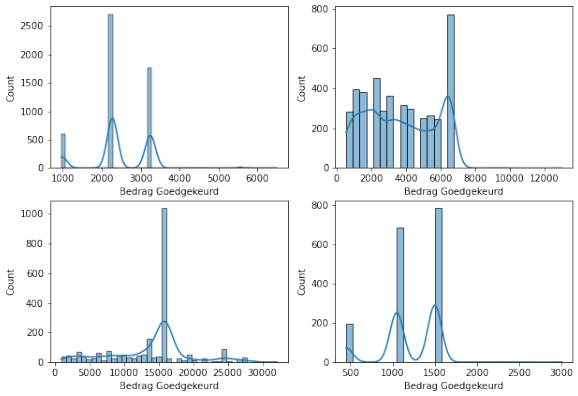
The different features are shown shown on the left, togheter with their importance which can be found on the other axis. "Bedrag goedgekeurd" is thus the most important feature.

All other features had too low importance in be visualized meaningfully in this figure.

Starting from the bottom, the "IVEOA" feature need some explanation. "IVEOA" is the number of notifications of the team "Vroeg Erop Af". This is a team tries to detect early detection for young people with starting money problems. If someone has money problems, this team receives a notification. [7] If problems are still minor, advice or a light intervention may be enough. Although it has a very low feature importance score, it has some predicting properties implying that the usage of a number of SPICs might be influenced by starting money problems. The second-best feature to predict youth care is age ('Leeftijd') and indicates that a number of SPICs are more frequently used by specific age groups. For example, care given in profile 10, is only for a young person up till the age of 6 years. [10] Finally, 'Bedrag Goedgekeurd is the most important feature of this model and stands for the amount of money the municipality has approved to pay the healthcare provider.

Since the amount of money approved for a certain SPICs is based on agreements between the municipality and the healthcare provider there might be a one-to-one relationship between the SPIC label and the feature 'Bedrag Goedgekeurd", defeating the whole reason why one would include this feature to start with. For the four most frequently used SPICs we have visualized the distribution of the amount of money approved in figure 8. If there was a clear one-to-one mapping there would be no variation in the amount of money approved. Based on figure 8 this is clearly not the case. In other words, when a client receives a specific SPICs, you cannot directly infer what the approved amount of money will be. To further investigate the importance of this feature we re-ran the winning model but now without the feature 'Bedrag Goedgekeurd'. As expected the F1-score drops dramatically to 16%. The result of the feature importance of the model without the feature 'Bedrag Goedkeurd' can be found in figure 9.

Figure 8: Bedrag goedgekeurd disttribution



4 different histograms are plotted against the "bedrag goedgekeurd", which represents the distribution of this feature.

CONCLUSION AND DISCUSSION

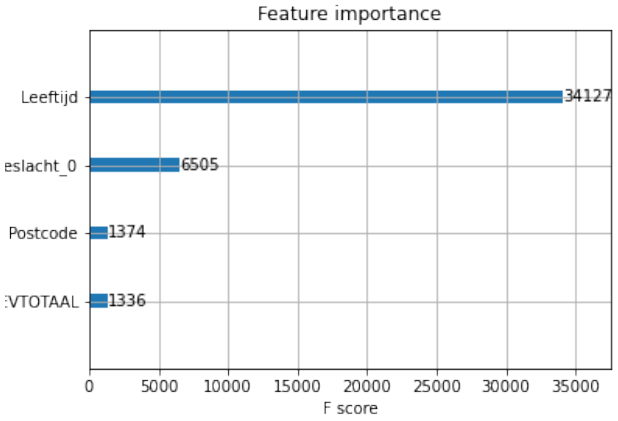
In this thesis, we tried to answer the following research question:

To what extent can Support vector machine, Decision Tree Classifier, or Gradient Boosting Machine contribute to predicting the demand for the specialist youth care in Amsterdam?

- Which of the tested models has the highest f1 score in predicting the youth care use?
- Which (neighborhood) characteristics are predictive for the use of youth care?

[7] https://www.amsterdam.nl/sociaaldomein/voor-intermediairs-werk-participatie-en/schuldhulpverlening-amsterdam/vroegsignalering/h28b60cfa-dd7d-4246-ad51-6bb6e3117f3b

Figure 9: Feature Importance removed bedrag



The different features are shown shown on the left, together with their importance which can be found on the other axis. "Leeftijd" is thus the most important feature.

For answering these questions: Three different ML algorithms were built for predicting the demand for specialized youth care. To improve the prediction model, different approaches were tested: cross-validation, GridSearchCV and Random undersample. We assessed the models on one performance metric, the F1 score. For all the machine learning algorithms, we created a baseline model. Were the XGboost performed at best. After the baseline models were created, random-undersample and a gridScearchCV were performed in order to get the best model for predicting the specialized youth care. XGBoost had the highest F1 score in with the use of the normal full dataset.

Having a look at this result, we can see which characteristics are predictive for the use of youth care. *Bedrag Goedgekeurd, Age and IVEOA* were features which involves the model at most for predicting the youth care. Where Bedrag Goedgekeurd has the highest feature of importance, and therefore could be seen as a value which can predict the youth care. Bedrag Goedgekeurd is a feature which has been set after a type of care has been given, but there is no direct link between the price and the given youth care. Due to agreements between the municipality and suppliers of the given care, there is a range of prices which are set for a particular given care. By removing this feature, the F1 score dropped dramatically.

Based on the results of this thesis, the value of the (neighborhood) characteristics for predicting the youth care in Amsterdam are limited.

However, better results can be obtained by including data on young healthy people to prevent bias. The data contains only data about youth persons in Amsterdam which had received youth care, and not of young healthy people. Which is not a true representation of the real world and can create bias in the models. Most of the models had reasonable f1-scores, deemed the most important metrics. However, this was mainly because most models were biased towards a positive prediction for youth care need and showed poor performance on other metrics. Due to CPU limitation we could not train the model on a data set which was over-sampled by the minority class. Over-sampled data, normally performs better then the used under sample technique. Translating this to practicality, it decently predicts youth care need, but does not enable a more cost-effective use of resources. Despite these limitations, this research does serve as an exploratory insight in what the possibilities of predictive modeling for youth care need are.

Future research. Future work on data-driven analysis of youth care data could focus on collecting and including more features that could explain and predict the demand and the costs of youth care. This is not an easy task, since many of these features would privacy-sensitive data. Which is, therefore, hard to get. Through literature review, the paper by [28] showed that there could be a relationship between neighborhood characteristics and the use of youth care. As said before, the main difference is the amount of features in this study. How the different neighborhood characteristics are exactly related to the use of youth care is sometimes unclear.

Feature works could also be found in the way of using the given data set. In this thesis, we used the random undersample technique to get a balanced data set. When a computer is used with more CPU power, it is possible to make an model that is based on a dataset which is balanced by oversampling. An intermediate step is the use of a combination of these two techniques. This could contribute to better model performance, but not having to much disadvantages of both options. Given the information we had on forehand, we choose for this three algorithms.Future research might also consider the use of other ML algorithms than the ones included in the current thesis. A method that comes to mind is Logistic Regression (LR), which would be an interesting alternative compared to SVM for imbalanced datasets. However, given the number of features, training samples and number of classes, it is unlikely that LR will outperform the SVM with the youth care data.[27] We used a data driven approach to identify predictive neighborhood characteristics, so a future scientist could make a feature selection beforehand. By removing the "bedrag goedgekeurd" column, the model changes dramatically. The selection of features beforehand could change the data set performance. Looking at which feature is important for predicting the demand, we looked at just one factor. Coming out of the related work

part, next researchers could try for fining combinations of factors in this study.

In conclusion, we have shown that it is possible to make a ML model which can make a prediction based on the given data. The amount approved is the feature which has the highest future importance. By adding even more features, it is possible to have a better insight in what factors influence the demand for care the most. Having an better insight in the cause of using the care you can avoid long waiting times and resulting suffering for young people from Amsterdam can be effectively combated.

REFERENCES

[1] Peter Adamson et al. Child well-being in rich countries: A comparative overview. Technical report, 2013.

[2] A Bjornberg et al. 2017 euro health consumer index. *PharmacoEconomics & Outcomes News*, 796:31–10, 2018.

[3] Leo Breiman. Arcing the edge. Technical report, Technical Report 486, Statistics Department, University of California at ..., 1997.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[5] Anna M de Haan, Albert E Boon, Robert RJM Vermeiren, Machteld Hoeve, and Joop TVM de Jong. Ethnic background, socioeconomic status, and problem severity as dropout risk factors in psychotherapy with youth. In *Child & youth care forum*, volume 44, pages 1–16. Springer, 2015.

[6] Scikit-Learn Developers. 1.4. support vector machines. In *Scikit-Learn 0.22. 1 Documentation. 2019*. 2019.

[7] Scikit-Learn Developers. 3.3. metrics and scoring: quantifying the quality of predictions. In *Scikit-Learn 0.22. 1 Documentation. 2019*. 2019.

[8] Annahita M Ehsan and Mary J De Silva. Social capital and common mental disorder: a systematic review. *J Epidemiol Community Health*, 69(10):1021–1028, 2015.

[9] DL Eliot. Support vector machines (svm) for ai self-driving cars. *Retrieved from AITrends: https://aitrends. com/ai-insider/support-vector-machines-svm-ai-self-driving-cars*, 2018.

[10] Gemeente Amsterdam. Bestuursrapportage jeugdstelstel 1e helft 2019. https://www.amsterdam.nl/sociaaldomein/beleid-jeugdhulp/artikelen/bestuursrapportage-jeugdstelstel/, 2019. [Online; accessed 10-September-2020].

[11] Gemeente amsterdam. Jeugdhulp in amsterdam. https://publicaties.rekenkamer.amsterdam.nl/jeugdhulp-in-amsterdam/, 2020. [Online; accessed 10-September-2020].

[12] Yi Gong, Stephen Palmer, John Gallacher, Terry Marsden, and David Fone. A systematic review of the relationship between objective measurements of the urban environment and psychological distress. *Environment international*, 96:48–57, 2016.

[13] Xuelin Gu, Qisijing Liu, Furong Deng, Xueqin Wang, Hualiang Lin, Xinbiao Guo, and Shaowei Wu. Association between particulate matter air pollution and risk of depression and suicide: systematic review and meta-analysis. *The British Journal of Psychiatry*, 215(2):456–467, 2019.

[14] Haibo He and Yunqian Ma. *Imbalanced learning: foundations, algorithms, and applications*. John Wiley & Sons, 2013.

[15] Marco Helbich, Julian Hagenauer, and Hannah Roberts. Relative importance of perceived physical and social neighborhood characteristics for depression: a machine learning approach. *Social psychiatry and psychiatric epidemiology*, pages 1–12, 2019.

[16] Clemens MH Hosman, Karin TM van Doesum, and Floor van Santvoort. Prevention of emotional problems and psychiatric risks in children of parents with a mental illness in the netherlands: I. the scientific basis to a comprehensive approach. *Australian e-Journal for the Advancement of Mental health*, 8(3):250–263, 2009.

[17] Mohammad Hossin and MN Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

[18] Max Kuhn, Kjell Johnson, et al. *Applied predictive modeling*, volume 26. Springer, 2013.

[19] Thomas Landgrebe and R Duin. A simplified extension of the area under the roc to the multiclass domain. In *Seventeenth annual symposium of the pattern recognition association of South Africa*, pages 241–245, 2006.

[20] Sigrid M Mohnen, Sven Schneider, and Mariël Droomers. Neighborhood characteristics as determinants of healthcare utilization–a theoretical model. *Health economics review*, 9(1):7, 2019.

[21] Netherlands Youth Institute. Wacht maar. https://vng.nl/sites/default/files/publicaties/2017/201705_wacht_maar_nji_onderzoek.pdf, 2017. [Online; accessed 23-november-2020].

[22] Netherlands Youth Institute. Reform of the dutch system for child and youth care. http://www.youthpolicy.nl/en/Download-NJi/Publicatie-NJi/Evaluation-of-the-Youth-Act-4-years-later.pdf, 2019. [Online; accessed 10-September-2020].

[23] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006.

[24] Ester Orban, Kelsey McDonald, Robynne Sutcliffe, Barbara Hoffmann, Kateryna B Fuks, Nico Dragano, Anja Viehmann, Raimund Erbel, Karl-Heinz Jöckel, Noreen Pundt, et al. Residential road traffic noise and high depressive symptoms after five years of follow-up: results from the heinz nixdorf recall study. *Environmental health perspectives*, 124(5):578–585, 2016.

[25] Jonathan Purtle, Katherine L Nelson, Yong Yang, Brent Langellier, Ivana Stankov, and Ana V Diez Roux. Urban–rural differences in older adult depression: a systematic review and meta-analysis of comparative studies. *American journal of preventive medicine*, 56(4):603–613, 2019.

[26] Robin Richardson, Tracy Westley, Geneviève Gariépy, Nichole Austin, and Arijit Nandi. Neighborhood socioeconomic conditions and depression: a systematic review and meta-analysis. *Social psychiatry and psychiatric epidemiology*, 50(11):1641–1656, 2015.

[27] Diego Alejandro Salazar, Jorge Iván Vélez, and Juan Carlos Salazar. Comparison between svm and logistic regression: Which one is better to discriminate? *Revista Colombiana de Estadística*, 35(SPE2):223–237, 2012.

[28] Roelof Schellingerhout, Ingrid Ooms, Evelien Eggink, and Jeroen Boelhouwer. Jeugdhulp in de wijk. 2020.

[29] Sklearn. f1_score. https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html, 2020. [Online; accessed 11-november-2020].

[30] Nora Wille, Susanne Bettge, Ulrike Ravens-Sieberer, BELLA Study Group, et al. Risk and protective factors for children's and adolescents' mental health: results of the bella study. *European child & adolescent psychiatry*, 17(1):133–147, 2008.

[31] Maureen Wilson-Genderson and Rachel Pruchno. Effects of neighborhood violence and perceptions of neighborhood safety on depressive symptoms of older adults. *Social science & medicine*, 85:43–49, 2013.