Predicting the characteristics of patients in treatment facilities using Machine Learning algorithms

H.S. Reefman

19/11/2021

Introduction

The research topic of this project is the treatment episodes of patients. The dataset used in this project is from the Treatment Episode Data Set (TEDS), an American data system of annual discharges from substance use treatment facilities. The used dataset is a TEDS-D dataset, which means it includes discharges from substance use treatment facilities. TEDS-D contains records on admissions of people of 12-years and older, and includes information of admission, substance use characteristics and discharges. The dataset includes information from patients who where at the facility in 2019. For this project some personal and substance use characteristics information are collected from the dataset. With this dataset a model can be made. In this project is researched whether the length of treatment can be predicted. The model is made by the following question: "Can the length of treatment be predicted based on the substance use, frequency of use and the age at first use by using machine learning?" To answer this question information is collected from the dataset. The personal information about the patients that is used is age. Other information is about substance characteristics, for example; substance use type, substance use at admission and discharge and route of administration. The information about treatment episodes are also covered, this includes length of stay in treatment. In this research multiple Machine Learning algorithms are explored, with using the best performing algorithm to make the model. The data is used to train and test the algorithms.

Materials and Methods

Exploratory data analysis and data preparation is done with the R programming language using Rstudio. Plots were made making use of the 'ggplot2' packages. For the layout of the plots the packages 'grid' and 'gridExtra' were used. Other packages that were used include 'tidyverse', 'tidyr', '(d)plyr', with as function to format tables. The model was built with Weka (version 3.8.5), a free data mining software written in Java. To built the model the clean dataset was inserted into the Weka Explorer and the performance of all standard machine learning algorithms were investigated. The machine learning algorithms investigated are:

- ZeroR
- OneR
- Naïve Bayes
- Simple Logistic
- Nearest Neighbor (IBK)
- J48

The classifications were carried out using 10-fold cross validations and the relevant quality metrics were recorded. These include: speed, accuracy, true positives (TP), false positives (TP), true negatives (TN) and the false negatives (FN). This is done in the Weka Experimenter.

Statistical tests were applied making use of the Weka Experimenter again. And some Meta learners were investigated. The findings were recorded and discussed.

In the end a ROC curve visualization and a learning curve were created.

Because the dataset included so many observations (around the 1.7 milion) a sample of the dataset was taken. To make sure this sample was a representative sample of the dataset, first the full dataset was investigated with the machine learning ZeroR. Also the sample was investigated with ZeroR (see Results).

The logbook and data of this research can be found in a repository on Github.com: https://github.com/Susanreefman/Themaopdracht09

At last a command-line application (wrapper) was made in the Java programming language. Code for the wrapper can be found in a repository on Github.com: https://github.com/Susanreefman/JavaWrapperThema09

Results

EDA

The distribution of the patients is first explored. In figure 1 depicts the histogram the age of patients in groups. The yougest group being 12 to 14 years and the oldest group over 65 years. The age seems normally distributed across the patients.

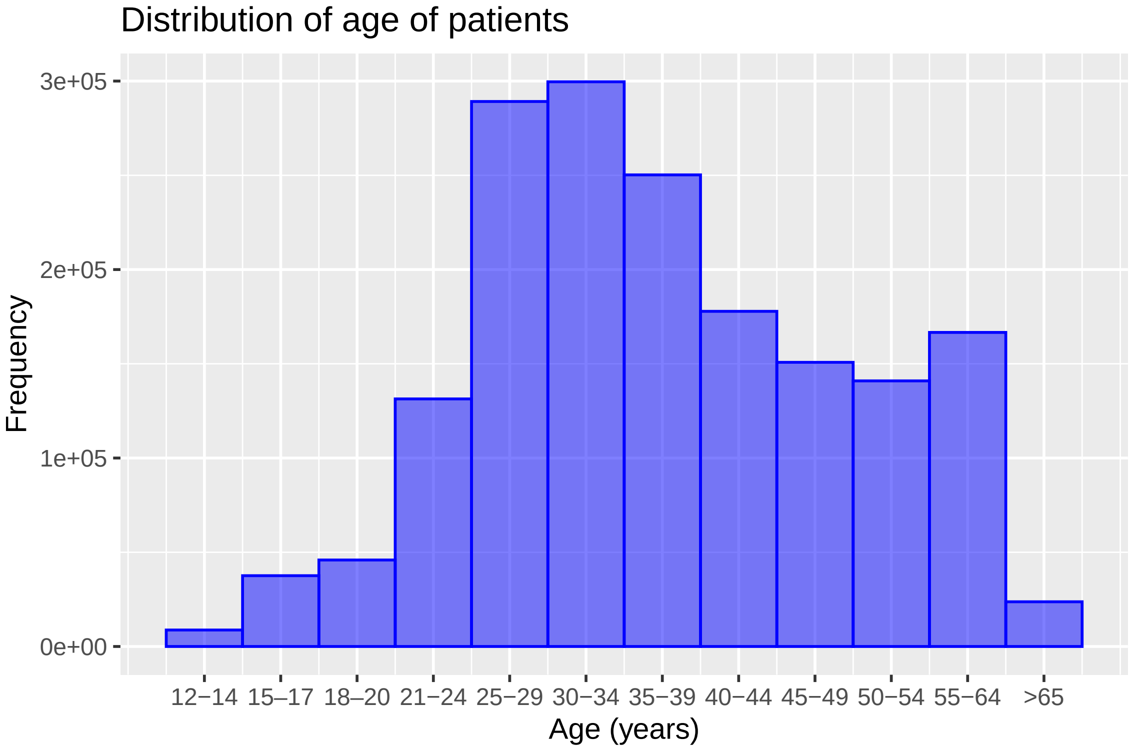

Figure 1: Figure 1: Histogram age of patients in groups

In figure 2 multiple stacked barplots are shown. The top barplot depicts the age of first use of the patients. The age is divided in multiple groups, in which the first groups is 11 years of younger and the last group is 30 years of older. Every bar is divided into 3 classes. The pink class is the age of first using the primary substance of the patients. The blue class is the age of first using the secondary substance, and the green class the tertiary substance. The results seem normally distributed with a large amount of patients who started between 15 and 17 years old with using their primary, secondary or tertiary substance.

The second barplot shows the frequency of use of patients. Unlike the groups the frequency of use in which the data is divided, the 'NA' group is extremely large. Out of proportion is are the missing values of the frequency of the tertiary substance used. This could be because patients do not have a tertiary substance that is used. The frequency for the secondary and tertiary substance is lower, most patients just use one substance. The most patients use their primary substance daily.

The third barplot depicts the usual route of administration of patients. Most patients use the oral route of administration for their primary substance use. Most patients that have a secondary substance use, use the smoking route of administration of that substance. Again large amount of patients have missing information for the route of the secondary and tertiary substance, these are the patients that said they did not have a secondary or tertiary substance they use.

Figure 3 shows the number of patients and their substance use. Not surprisingly is the blue (tertiary substance) very high for the None of substance of use. Not all patients have multiple substances they are using. However, there are a few patients that filled in that the have none as primary substance of use (pink). Alcohol, Heroin and Methamphetamine are the most primary substances of patients. Cocaine and Marijuana or Hashish are mostly the secondary substance used.

The boxplot in figure 4 shows the distribution of patients length of stay in treatment in days versus the substance use type which they get treated for. The mean of length of stay when they indicated "none" as substance use type is 9 days. For alcohol and other drugs the treatment is much longer. For alcohol as substance use type the average length of stay is 22 days. For the patients with a substance use type of only drugs or alcohol and other drugs the average length of stay is 28 days. The length of stay has a large difference in all substance use type groups.

Cleaning dataset

Because the dataset is very large and most of the data is not included in this project. Only the characteristics of the use of patients is necessary, which include: Type of use (alcohol, drugs or both), the substance of use, the frequency of use, the route of inhalation and the age of first use. The length of stay is also necessary in this project. Because the dataset is this large, the patients which have information that is not available is removed from the dataset. This includes around 200.000 patients, the dataset still includes more than 1.5 million patients. Removing the missing values makes it also easier for the machine learning algorithm to classify instances. Thereafter a sample of the dataset is taken. Both the dataset and the sample are written to csv files, to be uploaded to Weka.

Weka

The first action that was taken in Weka is validating the sample, both datasets were investigated with the ZeroR machine learning algorithm. In the Table 1, can be seen that the percentage of instances (in)correctly classified are not significant different. Taking the sample as the dataset to investigate further, the performance of standard machine learning algorithms was investigated. The classifier and attribute investigated were noted as well as the percentages (in)correctly classified instances (Table 2).

Table 1: Table 1: Validation sample of dataset with ZeroR Machine Learning algorithm

Attributes	Dataset correct	Dataset incorrect	Sample correct	Sample incorrect
type drug	53.9%	46.1%	53.8%	46.2%
sub	33.7%	67.3%	33.6%	67.4%
freq	45.5%	54.5%	44.5%	56.5%
frstuse	24.4%	75.6%	24.5%	75.5%
los	13.6%	86.4%	13.8%	86.2%

Table 2: Table 2: Results standard Machine Learning algorithms

Classifier	Attribute	Correct	Incorrect
ZeroR	los	13.6%	86.4%
ZeroR	sub	33.7%	67.3%
ZeroR	freq	45.5%	54.5%
ZeroR	frstuse	24.4%	75.6%
ZeroR	type_drug	53.9%	46.1%
OneR	los	16.9%	83.1%
OneR	sub	54.1%	45.9%
OneR	freq	53.7%	46.3%
OneR	frstuse	32.2%	67.8%
OneR	type_drug	71.4%	28.6%
J48	los	44.7%	55.3%
J48	sub	62.5%	37.5%
J48	freq	53.9%	46.1%
J48	frstuse	32.8%	67.2%
J48	type_drug	72.6%	27.4%
NaÃ¯ve Bayes	los	14.3%	85.7%
NaÃ¯ve Bayes	sub	58.6%	41.4%
NaÃ¯ve Bayes	freq	51.8%	48.2%
NaÃ¯ve Bayes	frstuse	26.9%	73.1%

Classifier	Attribute	Correct	Incorrect
NaÃ¯ve Bayes	type_drug	70.7%	29.3%
SimpleLogistics	los	14.6%	85.4%
SimpleLogistics	sub	58.9%	41.1%
SimpleLogistics	freq	52.3%	47.7%
SimpleLogistics	frstuse	27.6%	72.4%
SimpleLogistics	type_drug	72.5%	27.5%
NearestNeighbor	los	98%	2%
NearestNeighbor	sub	99%	1%
NearestNeighbor	freq	100%	0%
NearestNeighbor	frstuse	100%	0%
NearestNeighbor	type_drug	100%	0%

In table 2 are the standard Machine Learning algorithms and their results. The attribute `type_drug` is performing best of the attributes, while `los` is performing lowest. The Nearest-Neighbor algorithm performs almost for all attributes 100%, but this algorithm takes hours to compute a model. This was also the case for the J48 tree algorithm.

Machine learning

After using the training set, 10-fold cross-validation was used as test option. In table 3 are the results of the investigation. The time to execute the Machine Learning algorithms diverge from eachother. The ZeroR, OneR, Naive Bayes and Simple Logistics algorithms had a time span from 1 second to 10 minutes to train and to test the model. The Ikb and SMO algorithms took hours to compute results. Because Naive Bayes and SimpleLogistics have the highest percentage correclty classified instances and are rather fast in testing and training the model, these are taken into further research.

Table 3: Table 3: Results standard Machine Learning algorithms with 10-fold cross-validation as test option

	Correctly classified	Incorrectly classified
ZeroR	13.6%	86.4%
OneR	10.1%	89.9%
Naive Bayes	85.7%	14.3%
Ikb	17.1%	82.9%
SimpleLogistics	73.0%	27.0%
SMO	11.5%	88.5%

Optimization

Next, optimization of the Naive Bayes and Simple Logistics algorithms was tried to achieve. Because the 10-fold cross-validation did not improve the algorithms and took more time to process, percentage split was used as test option. Both algorithms were run with 4 different percentage splits: 40%, 66%, 80% and 90%. The Naive Bayes algorithm classified between 70.66% (40% split percentage) and 70.74% (66% split percentage). This is significant lower than the original (85.7%) model with using a training set. Also for the Simple Logistics algorithm was no optimization reached. With the different percentage splits, the algorithm correctly classified 72% of the instances each time. This is no decline but certainly no optimization.

At last, some meta learners, stacking, bagging and boosting, were examined on the Naive Bayes algorithm. For the model a training set was used as test model, the results can be seen in table 4. Only Bagging (43%) shows a reasonably amount of correctly classified instances. Thereafter the test option was again switched to

10-fold cross-validation, but none of the meta learners improved the model (table 5). This could be because the meta learners take the results of other algorithms and use those to train the model. In this case, previous runs of algorithms scored low on the correctly classified instances. Taking these algorithms as base learner, the model can not learn to correctly classify instances.

Table 4: Table 4: Meta learners with using training set as test option

	Correctly classified	Incorrectly classified
Bagging	43.3%	56.7%
Stacking	13.6%	86.4%
Boosting	13.6%	86.4%

Table 5: Table 5: Meta learners with using 10-fold cross-validation as test option

	Correctly classified	Incorrectly classified
Bagging	9.6%	90.3%
Stacking	13.6%	86.4%
Boosting	13.6%	86.4%

Curves

Visualizing the results in a ROC curve, figure 5, by running the Naive Bayes algorithm with 10-fold cross-validation, shows a area under the curve of 0.6664. Which indicates that the performance of the model at distinguishing positive and negative classes is little more than half of the time. After taking the data and error rate out of the Weka application a learning curve is made. As you can see in figure 6 the algorithm learns most when 90% of the sample of the data is used. The algorithm starts to learn slow when less than 20% is used.

Conclusion and Discussion

After performing multiple Machine Learning algorithms with different test options, the Naive Bayes and Simple Logistic algorithms were taken into further examination. Optimization of both these algorithms was not achieved, resulting in Naive Bayes with 10-fold cross-validation as test option being the best algorithm to make a model on this dataset. With that model 85% of the instances could be correctly classified. Assuming that the length of stay of patients in treatment can be predicted based on the substance use, frequency of use and the age at first use. However, there might be other attributes that could exceed the model. In this research only the substance, substance use type, frequency of use and age at first use was taken in account, although the route of administration of the substance and for example employment status could have indications for the length of stay. Another problem could be the amount of data in the dataset. Not all the algorithms could be thoroughly examined because of the time to compute the model. In following research the dataset could be split in more samples, differ in attributes and amount of instances. The algorithms could be more intesive tested on those samples.

Project proposal for minor

This research could be taken into a new project on larger scale. The goal would be to examine this large dataset thorougly and make one model of the entire dataset that can predict some attributes. This technology should be able to work as an Desktop app, which could be used by treatment facilities during the intake of patients.