Are you a smoker or non-smoker?

DA 5030 – INTRODUCTION TO MACHINE LEARNING

NORTHEASTERN UNIVERSITY

FALL 2017

Supervised By

MARTIN SCHEDLBAUER

INTRODUCTION TO MACHINE LEARNING

SIGNATURE PROJECT

Contents

1 BUSINESS UNDERSTANDING 1
1.1 Definition of the Problem 1
1.2 Business Success Criteria 2
1.3 Constraints 2
2 DATA UNDERSTANDING 3
2.1 Data Source 3
2.2 Exploring Data 3
2.3 Data Cleaning Assessment 8
2.3.1 Dealing with Outliers 8
2.3.2 Dealing with Missing Data 8
2.4 Assessment of Data Quality 8
2.5 Case Selection 9
2.6 Exploring Data 9
2.7 Data Preparation 10
3 MODELLING 12
3.1 K-Nearest Neighbors 12
3.2 Decision Tree 12
3.3 Support Vector Machine 13
3.4 Clustering 13
3.5 Linear Regression 13
4 EVALUATION 14
5 CONCLUSION 16
6 REFERENCES 17

1 BUSINESS UNDERSTANDING

1.1 Definition of the Problem

Nowadays humankind is under a massive threat. We are in danger of becoming more and more distanced from our beloved ones and having less and less relations. That fact is a serious threat for families because the generations cannot follow each other anymore and the bonds get weaker and weaker. Unfortunately, the cause is not in this paper but reading this paper will enable parents to feel slightly safer and take action if needed.

This increased distance might be unpredictable in 21st century. We are losing the ability to talk and argue because of technology since our best friends are replaced by cell phones. Technology made even one end of the world much closer as it was seen 30 years ago. Children leave their country in much earlier ages to go to universities or even high schools. Especially for younger people that might be the first time they truly discover what is going on outside of their bubble created by their parents. On one hand, it offers a unique opportunity to improve but on the other hand it might be dangerous. Since the protective shield of the parents has disappeared, children can get unhealthy habits easily. The easiest example can be smoking. The effort of healthy parents for all those years might be ruined by a close college friend who is smoking.

Since parents might want to find out whether their children are smoking or not, the machine learning algorithms might assist them achieving this goal. Asking the children straight forward whether they are smoking or not, is basically asking them to lie. This is not realistic to expect the true answer.

In this paper four machine learning algorithms are used to create models to predict the smoking habits of teenagers. If successful, the model will be able to predict and classify the smoking habits of young people to four groups;

- Never smoked
- Tried smoking
- Former smoker
- Current smoker

This paper will allow them to act accordingly and prevent the smoking habits of their children to grow, if there are any. That's why it can help tons of people to stay healthy and functional for a longer period. If the smoking habits are tackled in younger age, person can be prevented from being a long-time smoker.

As it can be seen from the Figure 1, tackling the smoking habits in early ages helps a lot. Considering that the start age is more commonly under 20, parents can help their children and show the way to health. When we consider that our data is collected from participants from age 15 to 30, the graphs below show some similarity.

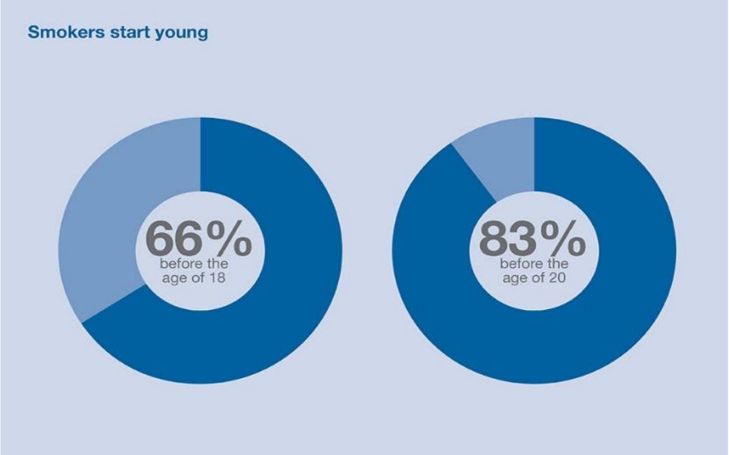


Figure 1: Starting Young to Smoke in England

1.2 Business Success Criteria

Like each model there will be a percent of failure also in this model. Even for the best models, there is a chance of failure. There is no exact simulation of reality. With the skills and knowledge that we possess we can only create a fake reality.

Since the outcome of this model plays an important role on the decisions of parents and their attitudes towards their children, the accuracy must be high. Universally the threshold of being significant is 95%. For this model, we will also adopt this value.

1.3 Constraints

The problems may arise with the issues that lie out of our influence circle. The data is gathered by a Kaggle user. The way data is collected can vary our solutions from the reality. The reasons of the possible problems might be the following points:

- Location: We have very few ideas how the data is collected. It is collected in Slovakia. Depending on the location, there might be only one segment answering the questions and participating. For instance, if the creator of dataset asked the students in Comenius University in Bratislava, the data might reflect the information of just a small area. That makes the result only applicable to that certain location.
- Method: The method of collection also matters. Using a social media source eliminates a certain group of young people. If usage of that particular social media source is collinear with the smoking cigarette habits, that would end up with a low accuracy of the model. We know that data is gathered both electronically and manually. So that created a more diverse group but we cannot know to what extend!

2 DATA UNDERSTANDING

2.1 Data Source

The data is gathered in 2013 by the students who are enrolled for Statistics course offered by Faculty of Social and Economic Science of Comenius University of Bratislava.

2.2 Exploring Data

The data involves two csv files:

- *responses.csv* consists of almost 1000 rows and 150 columns. 139 of these are integer and 11 of them are categorical. To make the analyzing process faster, the names of the columns are shortened.
- The actual question or full name of the columns can be found in *columns.csv*.

Data has missing values and to have a correct prediction in the end we must use one of the imputation techniques. Those values might also vary from group to group. The groups in the data are:

- Music Preferences (19)
- Movie Preferences (12)
- Hobbies and Interests (32)
- Phobias (10)
- Health Habits (3)
- Personality Traits, Views on Life, Opinions (57)
- Spending Habits (7)
- Demographics (10)

The questions are usually answered by integers from 1 to 5. In *kaggle.com*, all the columns are explained like following:

> ***MUSIC PREFERENCES***
> *I enjoy listening to music.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)*
> *I prefer.: Slow paced music 1-2-3-4-5 Fast paced music (integer)*
> *Dance, Disco, Funk: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)*
> *Folk music: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)*
> *Country: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)*
> *Classical: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)*
> *Musicals: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)*
> *Pop: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)*
> *Rock: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)*
> *Metal, Hard rock: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)*
> *Punk: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)*
> *Hip hop, Rap: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)*
> *Reggae, Ska: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)*

Are you a smoker or non-smoker?

Swing, Jazz: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Rock n Roll: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Alternative music: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Latin: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Techno, Trance: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Opera: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)

MOVIE PREFERENCES

I really enjoy watching movies.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Horror movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Thriller movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Comedies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Romantic movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Sci-fi movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
War movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Tales: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Cartoons: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Documentaries: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Western movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)
Action movies: Don't enjoy at all 1-2-3-4-5 Enjoy very much (integer)

HOBBIES & INTERESTS

History: Not interested 1-2-3-4-5 Very interested (integer)
Psychology: Not interested 1-2-3-4-5 Very interested (integer)
Politics: Not interested 1-2-3-4-5 Very interested (integer)
Mathematics: Not interested 1-2-3-4-5 Very interested (integer)
Physics: Not interested 1-2-3-4-5 Very interested (integer)
Internet: Not interested 1-2-3-4-5 Very interested (integer)
PC Software, Hardware: Not interested 1-2-3-4-5 Very interested (integer)
Economy, Management: Not interested 1-2-3-4-5 Very interested (integer)
Biology: Not interested 1-2-3-4-5 Very interested (integer)
Chemistry: Not interested 1-2-3-4-5 Very interested (integer)
Poetry reading: Not interested 1-2-3-4-5 Very interested (integer)
Geography: Not interested 1-2-3-4-5 Very interested (integer)
Foreign languages: Not interested 1-2-3-4-5 Very interested (integer)
Medicine: Not interested 1-2-3-4-5 Very interested (integer)
Law: Not interested 1-2-3-4-5 Very interested (integer)
Cars: Not interested 1-2-3-4-5 Very interested (integer)
Art: Not interested 1-2-3-4-5 Very interested (integer)
Religion: Not interested 1-2-3-4-5 Very interested (integer)
Outdoor activities: Not interested 1-2-3-4-5 Very interested (integer)
Dancing: Not interested 1-2-3-4-5 Very interested (integer)
Playing musical instruments: Not interested 1-2-3-4-5 Very interested (integer)
Poetry writing: Not interested 1-2-3-4-5 Very interested (integer)
Sport and leisure activities: Not interested 1-2-3-4-5 Very interested (integer)
Sport at competitive level: Not interested 1-2-3-4-5 Very interested (integer)

Are you a smoker or non-smoker?

Gardening: Not interested 1-2-3-4-5 Very interested (integer)
Celebrity lifestyle: Not interested 1-2-3-4-5 Very interested (integer)
Shopping: Not interested 1-2-3-4-5 Very interested (integer)
Science and technology: Not interested 1-2-3-4-5 Very interested (integer)
Theatre: Not interested 1-2-3-4-5 Very interested (integer)
Socializing: Not interested 1-2-3-4-5 Very interested (integer)
Adrenaline sports: Not interested 1-2-3-4-5 Very interested (integer)
Pets: Not interested 1-2-3-4-5 Very interested (integer)

PHOBIAS

Flying: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Thunder, lightning: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Darkness: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Heights: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Spiders: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Snakes: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Rats, mice: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Ageing: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Dangerous dogs: Not afraid at all 1-2-3-4-5 Very afraid of (integer)
Public speaking: Not afraid at all 1-2-3-4-5 Very afraid of (integer)

HEALTH HABITS

Smoking habits: Never smoked - Tried smoking - Former smoker - Current smoker (categorical)
Drinking: Never - Social drinker - Drink a lot (categorical)
I live a very healthy lifestyle.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)

PERSONALITY TRAITS, VIEWS ON LIFE & OPINIONS

I take notice of what goes on around me.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I try to do tasks as soon as possible and not leave them until last minute.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I always make a list so I don't forget anything.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I often study or work even in my spare time.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I look at things from all different angles before I go ahead.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I believe that bad people will suffer one day and good people will be rewarded.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I am reliable at work and always complete all tasks given to me.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I always keep my promises.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I can fall for someone very quickly and then completely lose interest.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I would rather have lots of friends than lots of money.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I always try to be the funniest one.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I can be two faced sometimes.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)

Are you a smoker or non-smoker?

I damaged things in the past when angry.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I take my time to make decisions.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I always try to vote in elections.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I often think about and regret the decisions I make.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I can tell if people listen to me or not when I talk to them.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I am a hypochondriac.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I am empathetic person.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I eat because I have to. I don't enjoy food and eat as fast as I can.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I try to give as much as I can to other people at Christmas.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I don't like seeing animals suffering.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I look after things I have borrowed from others.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I feel lonely in life.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I used to cheat at school.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I worry about my health.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I wish I could change the past because of the things I have done.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I believe in God.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I always have good dreams.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I always give to charity.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I have lots of friends.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
Timekeeping.: I am often early. - I am always on time. - I am often running late. (categorical)
Do you lie to others? : Never. - Only to avoid hurting someone. - Sometimes. - Everytime it suits me. (categorical)
I am very patient.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I can quickly adapt to a new environment.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
My moods change quickly.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I am well-mannered and I look after my appearance.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I enjoy meeting new people.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I always let other people know about my achievements.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I think carefully before answering any important letters.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I enjoy children’s' company.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I am not afraid to give my opinion if I feel strongly about something.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I can get angry very easily.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I always make sure I connect with the right people.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)

I have to be well prepared before public speaking.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I will find a fault in myself if people don't like me.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I cry when I feel down or things don't go the right way.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I am 100% happy with my life.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I am always full of life and energy.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I prefer big dangerous dogs to smaller, calmer dogs.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I believe all my personality traits are positive.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
If I find something the doesn't belong to me I will hand it in.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I find it very difficult to get up in the morning.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I have many different hobbies and interests.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I always listen to my parents' advice.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I enjoy taking part in surveys.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
How much time do you spend online? No time at all - Less than an hour a day - Few hours a day - Most of the day (categorical)

SPENDING HABITS

I save all the money I can.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I enjoy going to large shopping centers.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I prefer branded clothing to non-branded.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I spend a lot of money on partying and socializing.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I spend a lot of money on my appearance.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I spend a lot of money on gadgets.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)
I will happily pay more money for good, quality or healthy food.: Strongly disagree 1-2-3-4-5 Strongly agree (integer)

DEMOGRAPHICS

Age: (integer)
Height: (integer)
Weight: (integer)
How many siblings do you have? (integer)
Gender: Female - Male (categorical)
I am: Left handed - Right handed (categorical)
Highest education achieved: Currently a Primary school pupil - Primary school - Secondary school - College/Bachelor degree (categorical)
I am the only child: No - Yes (categorical)
I spent most of my childhood in a: City - village (categorical)
I lived most of my childhood in a: house/bungalow - block of flats (categorical)

2.3 Data Cleaning Assessment

2.3.1 Dealing with Outliers

To clean data there are two steps to be performed. First one is to clean out the outliers and second is the missing values. Some participants may cheat by answering the questions. Since the survey consists of 150 columns, it might take some time to complete. It is a possibility that there can be values entered not consciously. There might be problems also with the entry of data to the excel sheet. That might create missing values.

To tackle the outliers is a challenging issue in this dataset because we have a survey that is answered with the values from 1 to 5. When it is considered, there might be a person who loves music or hates music. This is perfectly normal. So, is 1 or 5 an outlier if the interest in music is asked? In this paper, it is not decided whether a participant's entry should be accepted as an outlier just by checking one feature.

Mainly the numeric columns consist of integer values. The rest is the columns that have a wider range than 1 to 5. These are age, weight, height and number of siblings. First, I checked whether there are age values not in the range of the survey, which is from 15 to 30. They are all in that range. To eliminate the outliers in height, weight and number of siblings, I used z-score method. First the mean and standard deviation of each column are calculated. Then using the z-score formula, the values that are three standard deviations away are detected and labelled as outliers.

The last outlier detection method was an unorthodox technique. Assuming if the person wants to end the survey as fast as possible (s)he selects the same option all the time -more likely 1 or 5. Having this in mind, I calculated the numeric values in the rows and created a summation column. With the help of the z-score test I detected the outliers and deleted them too. In total 26 rows, 2.5% of the whole data, are excluded. (In Appendix)

2.3.2 Dealing with Missing Data

To deal with missing data, I deleted all the missing values in the categorical columns. First it was considered whether to leave some of them but then decided that I need to involve all of the columns. That assessment leaded me to delete almost 9% of the data. When we consider the 9% as a value it might be big but out data is also large. We have more than 950 rows of data which is enough for us to build two subsets to train and then test the model.

For the numerical columns, I replaced the missing cells with the mean of each column. Because in that case, deleting the missing cells and rows would mean a huge data loss which might our analysis. When it is considered, it makes sense because our data only varies from 1 to 5. Replacing them with the mean - arguably close to 2.5- makes sense. (In Appendix)

2.4 Assessment of Data Quality

The data that I am working on is a clean data to a certain extend. It doesn't involve too many rows with missing values although some portion is entered manually. The techniques that I used to get rid of the missing values allowed me to keep more than 90% of the initial dataset.

2.5 Case Selection

Case selection is the part where the main frame of the paper is determined. Mainly I will cover the detailed version of this part in the 3. Modelling chapter but to give a pre-information, four techniques will be used in order to drive the prediction whether the college student smoke cigarettes or not.

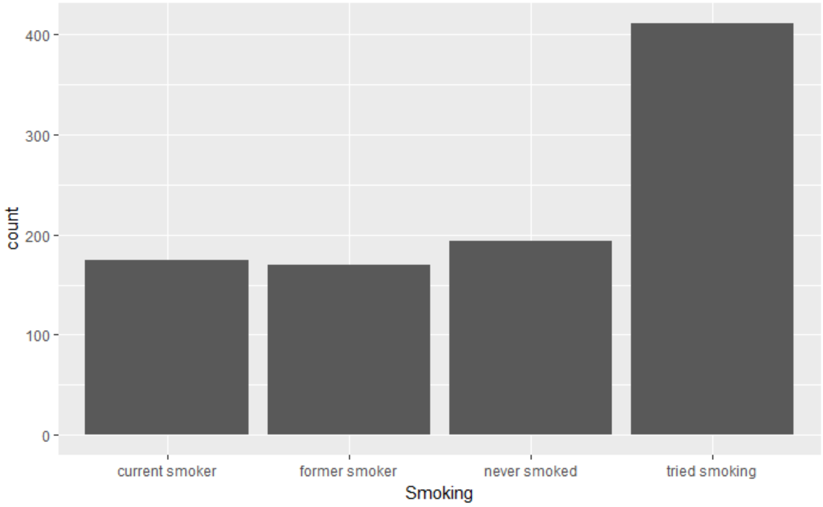
k-Nearest Neighbor will be used initially to drive the class of smoking habit. It can cluster the person into four levels of smoking. The main argument kNN suggests that each entry creates a location in the space. This location will be compared to the k nearest neighbors. From the coordinate of that entry in the space, the distance to each of k nearest neighbors is calculated. The group that has the minimum distance, will be assigned as the group of that entry.

Decision Tree is the second method that I will be using. That will allow me to classify the points with branching. From most to least correlated feature with smoking habits, tree will be shaped. With the help of thresholds that are assigned by background calculations, the user can interpret which category the person falls.

As another algorithm, Support Vector Machine will be used. It is easy to apply in the classification problems. It allows us to use the kernels as we like. When it is compared to a linear regression method, it can easily fit the points to a function because it can be nonlinear. That means instead of having a concrete slope and intercept, the SVM provides a curved function to pass from the closest points in the space.

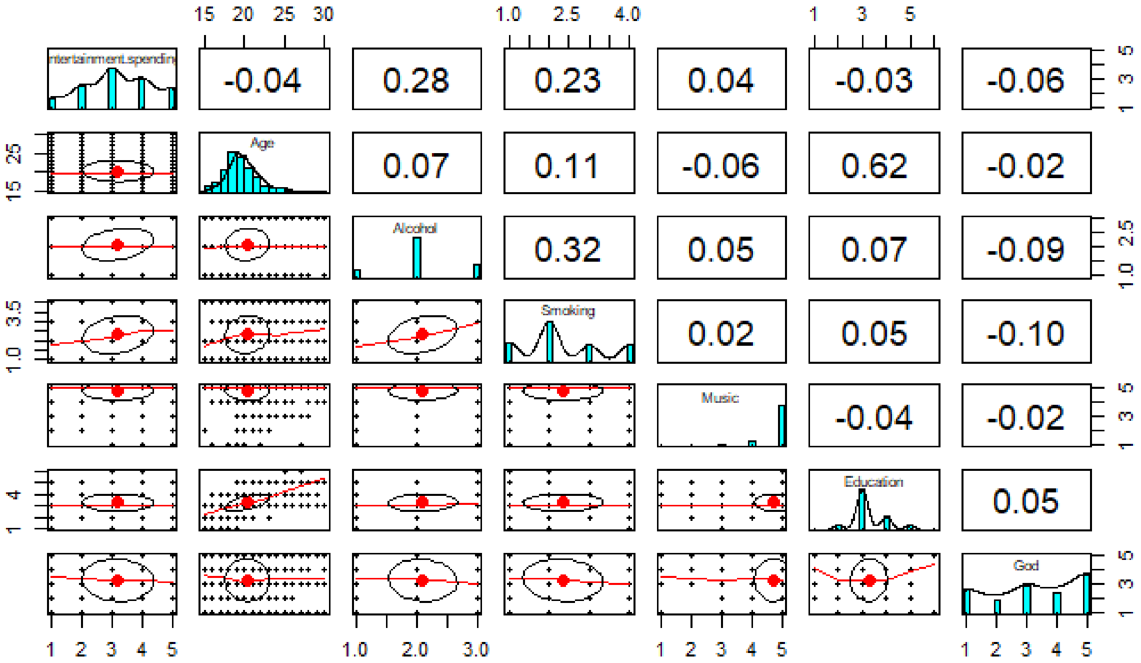
The last method that will be applied will be clustering. That serves our purpose also well, since we aim to basically classify the groups. This is an unsupervised machine learning technique, in which a computing system puts the items in the same groups. Since the algorithm is very flexible, it can be applied in complex datasets and it is expected to have a high accuracy with this technique when we consider our dataset.

2.6 Exploring Data



We will go through these steps in the modelling chapter but to have a general idea, I want to explore the data. I can start checking the data from the proportion of the smoker people. From the graph below, we can see the average proportion of people smoking. Luckily huge group of people are in either tried smoking or never smoked groups. That was basically the reason why I decided to group these two together, whereas former smokers and current smokers build the other one.

Another interesting information that I found out was the correlations between the variables. In order to have a better understanding, I run the linear regression model and realized the important predictors for smoking. One table proved that if a person has high entertainment spending, that person has a tendency to smoke, which we can see from the below graph.



2.7 Data Preparation

Since our data is mainly consists of values varying 1 to 5, it is one level simpler to perform the preparation stage. There are two steps for our data set to become ready for modelling; normalization and creating dummies for the categorical features.

First, we will start by dummies otherwise we don't have numeric values and we cannot normalize. There are two types of categorical values. One is the group that has answers like yes or no; male or female. The associated values can be found in the below table:

Column Name	1	0
Gender	Female	Male
Left/Right Handed	Left	Right
Only Child	Yes	No
Village/Town	Village	Town
House Type	House/Bungalow	Block of Flats

Table 1: Dummy Values of Categorical

The second group is the items that we can compare with themselves. For instance, there is the education column. It consists of values:

- College/bachelor degree
- Currently a primary school pupil
- Doctorate degree
- Master's degree
- Primary school
- Secondary school

We can scale these values saying that highest education level is doctorate degree which can be symbolized by the max value. On the other hand, the current primary school students are the ones with the lowest level of education so they need to get the lowest education score.

After having a dataset full of integers, we need to scale them so that we can insert them to the same model. Otherwise they will have a different effect on the result. At the end of the process each column has a value from 0 to 1. For this purpose, z-score is used again. With this method, I got values that vary between -1 and 1. So my range became [-1,1].

3 MODELLING

3.1 K-Nearest Neighbors

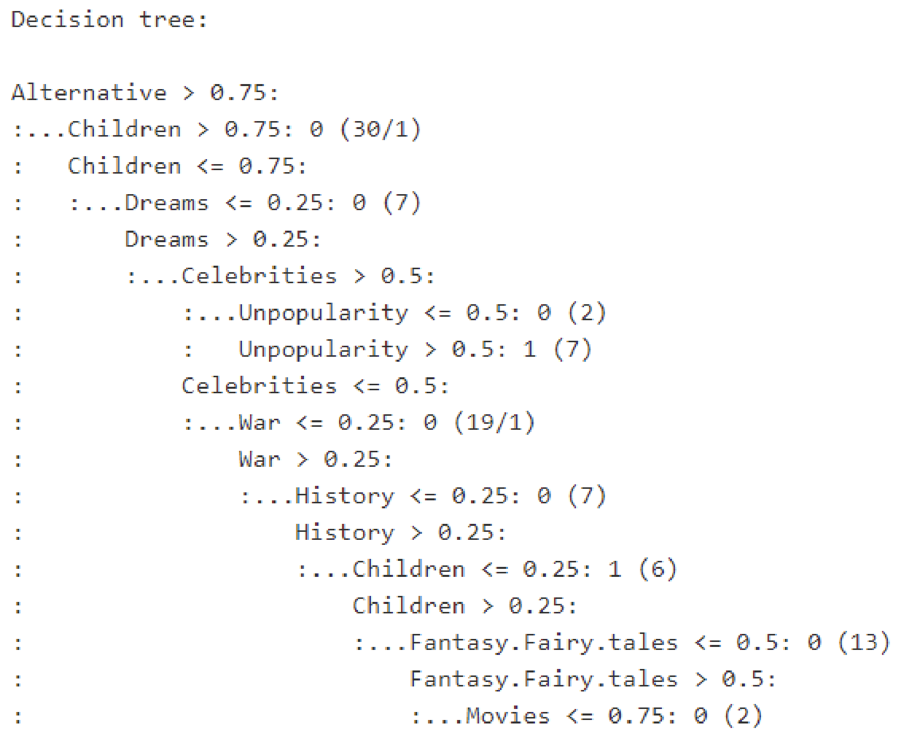
In the training stage of kNN, I came up with 5-fold dataset. So that I can build 5 models and test their accuracies. To create the folds, I basically created index and took 80% of data as training and 20% as testing. That approach is applied in all the models in all the algorithms. To generate this partition, I benefited from the seq() function. By increasing 5 points in one step, I ended up with 5 clusters. When these values are subtracted from the data set, we reached our testing dataset.

With the help of kNN() function, I stated my training dataset, testing dataset and class, which is my target value (smoking habits). At last, k parameter is decided. The square root of my observations is almost 31. So, I picked 31 as k value. In other words, my model became 31-nearest neighbors.

To improve the mode, I modified the k values. That enabled the model to verify the number of closest neighbors to check. For instance, in the case of 31 nearest neighbors, model gets a point in space; detects the closest 31 points and predicts the class of the new entry based on the class of that 31 closest neighbors.

3.2 Decision Tree

Decision tree that we created is a little bit messy but it is close to what we have expected since there are too many features. First, I needed to convert my target values to factors for function to accept them. Then using C5.0() function, I got my decision trees. That is a huge tree with lots of branches. For instance, a part of the tree is shown below:



```
Decision tree:

Alternative > 0.75:
:...Children > 0.75: 0 (30/1)
:   Children <= 0.75:
:   :...Dreams <= 0.25: 0 (7)
:       Dreams > 0.25:
:       :...Celebrities > 0.5:
:           :...Unpopularity <= 0.5: 0 (2)
:           :   Unpopularity > 0.5: 1 (7)
:           Celebrities <= 0.5:
:           :...War <= 0.25: 0 (19/1)
:               War > 0.25:
:               :...History <= 0.25: 0 (7)
:                   History > 0.25:
:                   :...Children <= 0.25: 1 (6)
:                       Children > 0.25:
:                       :...Fantasy.Fairy.tales <= 0.5: 0 (13)
:                           Fantasy.Fairy.tales > 0.5:
:                           :...Movies <= 0.75: 0 (2)
```

What here describes is if the person has an interest and like towards the alternative music and if the person enjoys children; that person belongs to group 0. In other words, that participant is highly a non-smoker. Initial part of the tree is easy to interpret but it gets much harder when the branches get larger and larger. After a point following it gets impossible and not effective. Just like we did in kNN, again the predict() function is used and that helped us to create a prediction on target variable.

To improve the decision tree, trial parameter is added. Number that initiated with trial represents the number of tree that will be taken into consideration. What we managed to have is combining the results of 10 different trees.

3.3 Support Vector Machine

For ksvm() function to run we needed to convert our training and testing datasets into matrixes. We initiated what we address as target value and determined the predictors. Basically, I included all of them. After running the ksvm functions, I was able to have a prediction column. Next step was comparing those columns. To see how many values match I developed a counter mechanism, which counts the true matches between predictions and actual target values.

In the second step, I tried to improve the mode by adding kernel parameters. “rbfdot” provided us a better prediction. The reason is that by Support Vector Machine’s feature, we were able to switch from linear regression to curvature functions that go near the points in the space and fit them better.

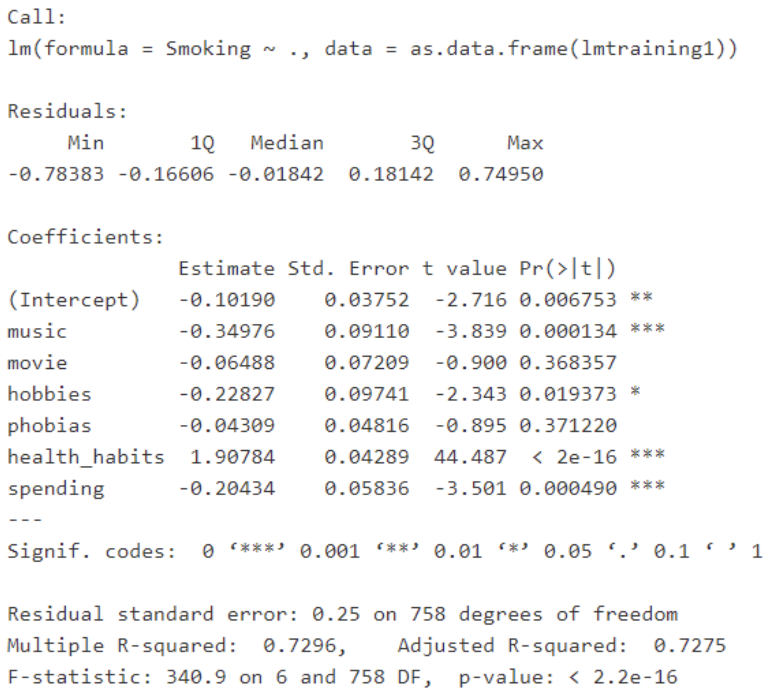
3.4 Clustering

In order to work with clustering algorithm, I revised the testing and training datasets once again since it requires to have the target value in the training set. Then I came up with two clusters, determined their centers and checked how many falls into which cluster. For each and every point, I calculated a distance from each point to two clusters’ centers. According to the comparison that I made, an item falls into a certain cluster if its total distance is smaller to that one.

3.5 Linear Regression

In the last approach, I implemented the linear regression model. First, I revised my datasets to build the function. Then, used the lm() function and created an equation. With the help of that equation, I tried to predict again the smoking habits. This time the linear regression provided me a continuous outcome, which I needed to convert to integers. For a basic classification, I created the if statement that helped me to assign the values larger than 0.5 to 1 and smaller than 0.5 to 0.

To improve the model, I used backwards elimination technique. For this process, I excluded the predictors with a p value higher than 0.05, which suggests that these are not significant estimators. That helped me to revise my accuracies outcomes.

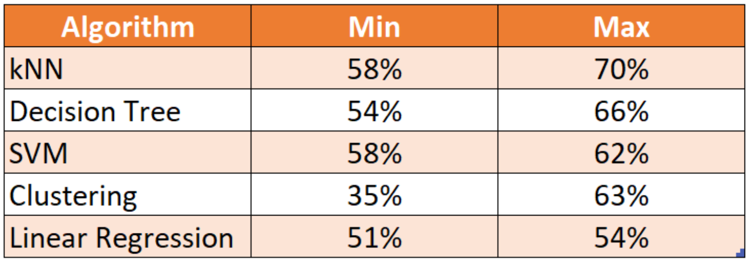


```
Call:
lm(formula = Smoking ~ ., data = as.data.frame(lmtraining1))

Residuals:
     Min       1Q   Median       3Q      Max 
-0.78383 -0.16606 -0.01842  0.18142  0.74950 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -0.10190    0.03752  -2.716 0.006753 ** 
music         -0.34976    0.09110  -3.839 0.000134 ***
movie         -0.06488    0.07209  -0.900 0.368357    
hobbies       -0.22827    0.09741  -2.343 0.019373 *  
phobias       -0.04309    0.04816  -0.895 0.371220    
health_habits  1.90784    0.04289  44.487  < 2e-16 ***
spending      -0.20434    0.05836  -3.501 0.000490 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.25 on 758 degrees of freedom
Multiple R-squared:  0.7296,	Adjusted R-squared:  0.7275 
F-statistic: 340.9 on 6 and 758 DF,  p-value: < 2.2e-16
```

For instance, as we can see from the above table, movie or phobias are unnecessary estimators. That's why they need to be excluded from the model.

4 EVALUATION

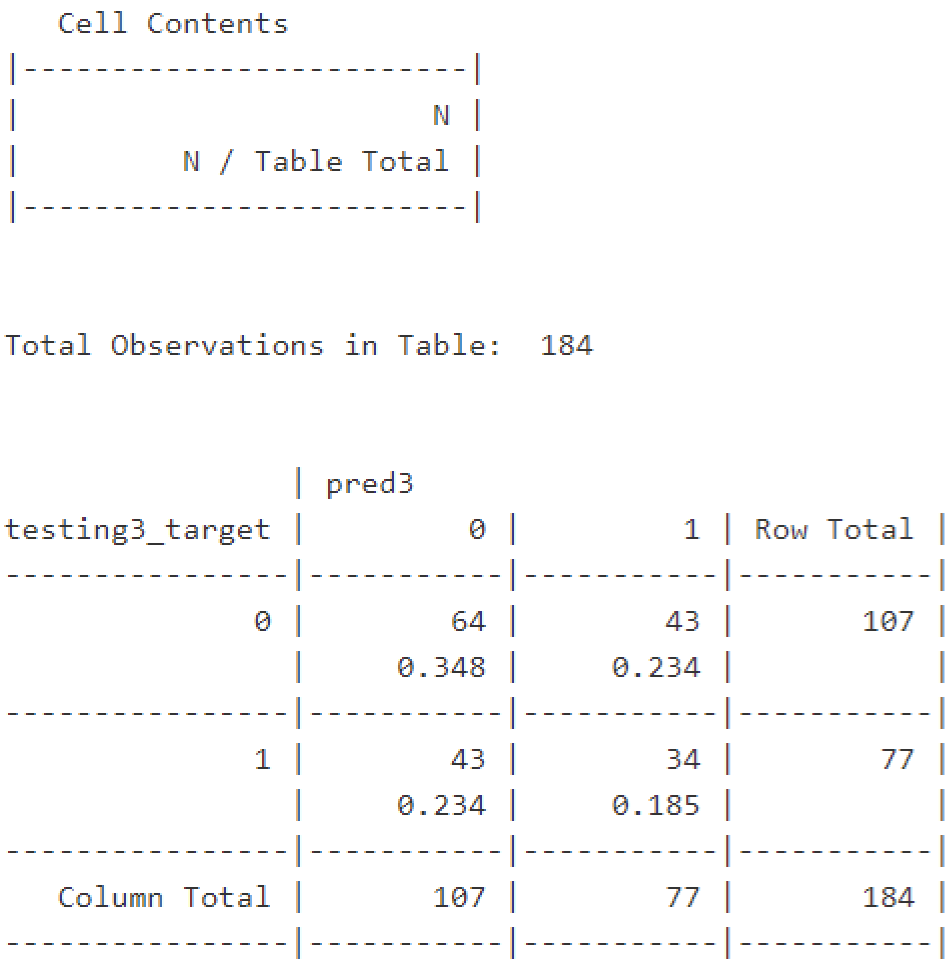
For Evaluating the performances of the models, cross tables or accuracy functions are used. They serve for the same purpose. That means although we used cross table or accuracy functions they have the same scale of information. From the below table we can compare the outcomes of these functions.



Algorithm	Min	Max
kNN	58%	70%
Decision Tree	54%	66%
SVM	58%	62%
Clustering	35%	63%
Linear Regression	51%	54%

Determining the accuracy from the count function, I created a counter variable. That number is increased by one each time the prediction and actual value matches.

Cross table is easier to visualize for sure. There are two labels; actual and predicted. The values that lie on the diagonal are the ones that predicted correctly. The rest are the failures. From the below cross table we can drive the accuracy just by looking. The accuracy here is the ratio of correctly replaced over all of observations, which makes 98/184.

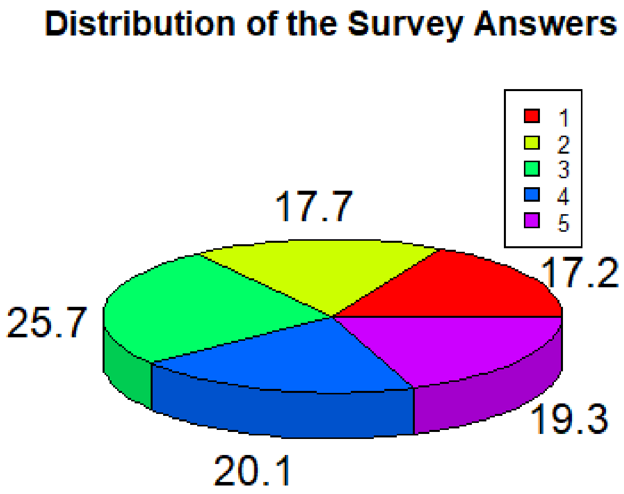


```
   Cell Contents
|-------------------------|
|                       N |
|         N / Table Total |
|-------------------------|

 
Total Observations in Table:  184 

 
                | pred3 
testing3_target |         0 |         1 | Row Total | 
----------------|-----------|-----------|-----------|
              0 |        64 |        43 |       107 | 
                |     0.348 |     0.234 |           | 
----------------|-----------|-----------|-----------|
              1 |        43 |        34 |        77 | 
                |     0.234 |     0.185 |           | 
----------------|-----------|-----------|-----------|
   Column Total |       107 |        77 |       184 | 
----------------|-----------|-----------|-----------|
```

Different from what we have expected the results don’t show any indication of high accuracy. Still if we need to pick one of these approaches that should be kNN. Although it is a simple and even “lazy” algorithm, it is widely accepted.

5 CONCLUSION

70% is still an important measure. Without using any models, we have 50% chance of guessing it. With the help of models, we increase our chance by 20% but still there is a huge room of improvement. Still, we are capable of analyzing as much as data shows us. That made me think about the dataset and I improved an argument saying that if the accuracies are low, that shows that our data is not biased. In other words there might more answer of "3" to scale questions than anticipated. That's why I ran an analysis where I counted all the numbers. The outcome is below:



It is true that there are much more than anticipated "3". They block us from making a proper analysis since they don't show any preference. Maybe next time in a survey analysis, it might be considered to drop all 3's. Since they don't help us. Almost 26% of data being data shows that participants are more hesitant.

As a conclusion, it can be noted that this model of kNN, can be implemented in real life but there needs to be more improvements because the social aspects that the outcome of this model lead is very essential. Parents might shape their attitude towards their children considering this model.

6 REFERENCES

BlackBoard Link to my evaluation:
https://docs.google.com/a/husky.neu.edu/spreadsheets/d/1B98Zd9QeGO0ZdKSFirFyAbeIeyJ6DYTE-7gtvAfkbzk/edit?usp=sharing

https://www.gov.uk/government/publications/health-matters-smoking-and-quitting-in-england/smoking-and-quitting-in-england

https://www.kaggle.com/breinken/young-people/data

http://books.tarsoit.com/Machine%20Learning%20with%20R%20-%20Second%20Edition.pdf