Homework 10

Team: B5

Team members:

- Karl-Christofer Veske
- Kristjan Siim
- Karl Tomasson

First step - Repository

The repository is here, and Carel Kuusk has been invited to it.

- https://github.com/karlveske/DS_Mental_Health

Second step - Business understanding

Business goals

Background:

As the level of recorded counts of people with depression has increased drastically over the years, we wanted to analyze how certain lifestyle conditions might affect people's chances of having depression and whether it can be confidently predicted. .

Goal

We want to analyze what lifestyle factors might affect depression, find patterns, and build a model that confidently predicts whether the person might have depression if we know about the same type of information that's used in the training data set.

Therefore, it can help to find risk groups and help individuals make decisions in order to decrease their odds of getting depression.

Business success criteria

For us, the business success comes in two parts.

1. We find interesting practical insights and patterns for every working/studying professional in order to decrease their odds of getting depression.

2. When testing our ML models in the Kaggle competition, where this data is from, it performs with ~90% confidence.

Assesing our situation

Resources

We have 3 brilliant aspiring data scientists in our team who are eager to make some sense of this data.

We also have two datasets, training and test datasets. Both of these have columns like profession, job/study satisfaction, sleep duration, etc.

They are almost identical, but the test data set doesn't have the data on depression(0 or 1 values), so we can use it to test our ML models in the Kaggle competition.

Risks, requirements, constraints.

The project must be completed by the 9th of December.

No out-of-the-ordinary risks or requirements might hold us back from achieving this goal by this deadline. There might be the usuals: lost internet connection, getting sick, or a dog eating the cable. All of these challenges we plan to resolve using common sense.

Costs and benefits

As the data is free, electricity, time, and coffee are the only costs.

The benefit of this project comes from the mouth of the late Charlie Munger: "All I want to know is where I'm going to die so I'll never go there". Meaning it's a good piece of information to know how to live a happier life.

Terminology.

In the communication of the findings, we plan to explain all of the terminology. Right now, don't know what's relevant in the business context, as technicalities are not worthy of discussing.

Defining the data-mining goals

Data-mining goals

The end deliverables for this project:

- A Jupyter notebook with all the work
    - visualization of interesting insights and patterns
    - A machine learning model that predicts confidently whether the person has depression or not if we give it inputs on same variables as the model was trained on.

- A poster introducing all of the work we did and what we discovered

Data-mining success criteria

- For us the project is successful, if we build a model that predicts depression with around ~90% confidence

Third step - Data understanding

Data gathering

The data is from a Kaggle competition. It has two datasets: train and data. It's in the .csv format.

When selecting the data, it was relevant for us that it had a lot of rows, a lot of different variables, and a clear 0 or 1 verdict on whether the person has depression.

Data description

The training dataset contains 20 columns and 140700 rows. The test data is 93800 rows and 19 columns.

Both of these have columns like id, Name, Gender, Age, City, Working Professional or Student, Profession, Academic Pressure, Work Pressure, CGPA, Study Satisfaction, Job Satisfaction, Sleep Duration, Dietary Habits, Degree, Have you ever had suicidal thoughts?, Work/Study Hours, Financial Stress, Family History of Mental Illness, Depression

The difference is that the test data set doesn't have the depression(0 or 1 values) column, so we can use it to test our ML models.

Data exploration and quality

Data seems to have relatively good quality. From the training data set, we dropped 224 corrupted rows (out of 140700), which doesn't affect the dataset quality.

For better use in training the ML model, we used one-hot encoding to turn columns with multiple values into separate columns containing 1s or 0. And we combined some similar information.

- For example, there were various degrees in the dataset, we divided into 4 different categories,
- Degree_12, Degree_Masters, Degree_PHD. And if all of these are 0, then the individual has Bachelors.

We combined certain metrics that had been recorded in separate columns into one.
For example

- work pressure column, study pressure column → one column: work & study pressure
- work hours column, study hours column → one column work & study hours

And we categorized

Fourth step - Project plan.

Since this data is from a Kaggle competition, we don't have to spend much time preparing the data.

Here's our plan with estimated time commitments.

1. Removing the bad data - Kristjan & Karl(5 hours)
    a. There is some random information and broken rows.

2. Preparing the data - Karl (5 hours)
    a. We are doing one-hot encoding to build better ML models.

3. Putting together a plan of work – we all (1 hour)
    a. Dividing who will do what and how we share the results with each other.

4. Exploratory analysis, pattern mining and visualization - Karl-Christofer(15 hours)
    a. We'll try to find interesting insights from the data we have.

5. Developing machine learning models - Kristjan & Karl (15 hours)

6. Designing the poster - Karl-Christofer (6 hours)

Tools we plan to use:

- Excel - for initial data evaluation and some cleanup
- Jupyter Notebook with Python for all the development and analysis
- Canva for creating a poster.

Methods we plan to use

- For ML, we plan to test out different models
    - KNN
    - Linear regression
    - Decision tree
    - Random Forest
- And try out under and overfitting.

As of writing this, we are still open to exploring throughout the work what more we could do with this data and the ML models we are developing.