Portfolio assignment 101 – Data Science

Data Science wants to find anomalies, patterns and correlations in data to then use these results to either predict future events, find correlations which can then be used to classify future data based on the properties those correlations are based on. It is used to answer research questions using mathematical equations, statistics and models. Often an error is also taken into account, that says something about whether a result can be accepted or not.

In the assignments we often looked at columns or combinations of columns to see if we could find anomalies like outliers that might need to be removed (assignment 6), patterns like a confidence interval in which most values are found (assignment 8) or correlations between 2 or more columns (assignment 9 to 19). For most of these assignments we also checked if the result was likely to be accurate or not likely to be accurate using various error measurements (standard deviation, RMSE).

Data Science stretches over many fields, in organisations it is usually related to transforming a large volume of data into usable information and sometimes creating software and algorithms that help organisations make informed decisions. Data Science is also used in scientific research where the goal is not so much to make informed decisions, rather to increase knowledge and describe phenomena.

As mentioned before, Data Science can be used by organisations to get usable information and make more informed decisions. In assignment 5 we performed an analysis to find the countries with the highest and lowest life expectancies. In BI similar analyses are often performed to see how well the company is performing, or which strategies work and which do not have any positive effect. Just finding the highest and lowest life expectancy may not say too much, but you could also look at whether there is a different relation between countries with similar life expectancies and if this relation gives any new insights that can be used.

Artificial Intelligence is a field that can be used in the context of Data Science to do predictive analyses. One of the fundamental concepts of AI is Machine Learning, the study of computer algorithms that improve automatically through experience. In assignments 15 to 18 we used machine learning algorithms (decision trees) to build a model that could either predict classifications or a numerical value. We also checked the accuracy for the set the algorithm trained on and a separate test set to see if the model found a proper fit.

For assignment 19 we used the KMeans algorithm to find clusters. The accuracy of these clusters was a little harder to determine, so that also involved manually checking if the found clusters made sense with what we already know about the data (like classifications).

Portfolio assignment 102 – Tasks and process of a Data Scientist

The process of a Data Scientist often starts with accumulating data as we did in assignment 4, either to answer a particular question, or use given data to find anomalies, patterns and correlations. The data sets used are often quite large and cannot be easily viewed in full or filtered manually.

After retrieving the data a data scientist first needs to understand the structure and variables in the set and how they might be used. Following this, a data scientist checks if the data can be used as it is or if any adjustments need to be made, like removing outliers. This can be done by checking for outliers (comparing mean and median and concluding whether any data is either unlikely to be true or will have negative effects on any analyses, as done in assignment 6), or looking at the distributions of records (assignment 7). If the set follows a normal distribution a confidence interval can be calculated that tells something about how often a value will fall within which range (assignment 8).

These are all examples of data processing using only one variable. Often though, it is necessary to use two or more variables. To find correlations, two or more variables are compared, which can give insight in whether variables are related. Assignments 9 to 14 use bivariate analyses to compare combinations of numerical and categorical variables and assignments 15 to 19 make use of multivariate analyses to compare two or more variables. Using more than one variable you can make predictions of values based on other known variables. You can create a classification which records can fall into, or predict numerical values using regression. This is often done using machine learning algorithms, in our case decision trees in assignments 15 to 18. A machine learning algorithm was also used to create clusters in the dataset from several variables in assignment 19.

Thus a data scientist needs to find a proper combination of statistics and mathematical equations to transform data into the desired format. They might also need to write / choose algorithms that further process data and give insights for example by making models. The output of those algorithms does not always make sense, so it is also essential that a data scientist checks the results and compares to what is already known (for example comparing clusters to known classifications to see if the algorithm works) and adjust models (retraining) if necessary.

Another thing that is apparent over all assignments is the need to visualize the data and its transformations. The visualization needs to be clear to understand and properly show what information was found. The end user might need to be taken into account to make visualizations that make sense to them.