Student: Lana Ali

Lecturer: Witek Ten Hove



Introduction to Data Mining

Contents

[Personal reflection 2](#_Toc149559091)

[Data conclusion 2](#_Toc149559092)

A diagram of data processing

Description automatically generated

# Personal reflection

During the first two weeks of the Data Mining course, I was having some difficulties. Due to some health issues, I had to follow the classes online, which left me feeling like I was already behind and struggling to understand whole concepts. As this was my first encounter with working with data and data programs, I was not really confident.

It wasn't until a conversation with Witek in week 5 that my perspective began to shift. I came to the realization that it really wasn't as complex as I thought it was. I understood that the expectation wasn't for us to be instant experts, but rather to understand on how to initiate the process. It was also then that Witek showed me ‘how’ to use ChatGPT for this course. I had heard from classmates that they had been using this, but I never knew it could also be used in this certain way. This was so helpful - suddenly, everything seemed more attainable. I found myself being more productive in class than I had in the preceding weeks. I realized that I had really been overcomplicating this course in my mind.

Not to mention, Jason and Kevin also were more than helpful in week 5. They went out of their way to help me, ensuring that I not only understood how to navigate the tools but also understand the meaning of the tests. Since they did not have to do all of this, I really appreciate their help a lot.  
Now that I have a better understanding of Data Mining works, I find myself enjoying this course more!

# Data conclusion

Together with my classmate Rutger Bles, we had a look into some data to predict office occupancy based on factors like temperature, humidity, light, CO2, and humidity-ratio. We both used different models to check the accuracy of this predictive model. Where he used the KNN-model and I used the Logistic Regression Model.

I will now further inform you about my experience with the logistic Regression model.   
When put to the test, the model performed very well. It achieved an accuracy score of 0.988 on the test, indicating that it is quite adept at processing new data. So, the features selected are definitely giving us some meaningful insights into occupancy status.

The classification report digs even deeper, breaking down precision, recall, and F1-score for each class. This gives us a thorough understanding of just how well the model can make its predictions.

Of course, there's always room for improvement. We might want to explore different features or even try out different algorithms down the line.

The Logistic Regression model is an excellent tool for forecasting occupancy based on environmental factors. Its accuracy and extensive categorization report demonstrate its real-world use.

In conclusion, our task centered around training both a KNN-model and a Logistic Regression model for occupancy prediction. The results were exceptionally promising, boasting nearly flawless accuracy rates, accompanied by an impressively low count of false positives and false negatives. This underscores the remarkable reliability of the dataset and the remarkable predictive capabilities exhibited by our models.