DATA MINING

Predicting wine quality

Reflection

Name:	Inger Dekker (1651461)
	Tara Maarse (2136582)
Class:	MDD-01
Data:	19-10-2023
Class:	Data Mining

Table of contents

1. Dataset

We chose the dataset to predict wine quality. The following link will take you to this dataset: Wine Quality Dataset (Wine Quality Dataset (kaggle.com)). With this dataset you can use the components like fixed acidity, residual sugar, chlorides etc. to be able to predict wine quality.

2. Models

To start predicting wine quality, we chose the KNN model and the Naive Bayes model. These models can both predict in a separate way. After we run both of these on the dataset, we will also compare them with each other to give a conclusion as to which is ultimately the best one to use for predicting quality of wine. We started with the KNN model and then we did the Naive Bayes.

2.1 K-Nearest Neighbors (KNN)

KNN works by finding the 'k' nearest data points (neighbors) in the training dataset to a new input data point and then making predictions based on the majority class (for classification) or the average value (for regression) among those neighbors.

Key steps in the KNN process:

1. Choose a value for 'k', which determines how many nearest neighbors will be considered.
2. Calculate the distance between the input data point and all data points in the training dataset (commonly using Euclidean distance).
3. Select the 'k' data points with the shortest distances.
4. For classification, assign the class that appears most frequently among the 'k' neighbors as the predicted class for the input data point.

This model is sensitive to the choice of ‘k’ and the scaling, so we also tried different numbers for this.

2.2 Naive Bayes

Naive Bayes Classification is a probabilistic machine learning algorithm used for classifying data points into predefined categories. Here is how it works for predictions:

1. **Training Phase:**
   - Collect a labelled dataset, where each data point is associated with a category (class).
   - Calculate the prior probabilities of each class, which represent the likelihood of each class occurring in the dataset.
   - Create a Naive Bayes Classifier and Train It.

2. **Prediction Phase:**
   - Given a new, unlabelled data point, calculate the likelihood of the data point's features under each class using the conditional probabilities.
   - Let the model predict some new data points.
   - Evaluate the model’s performance.

Naive Bayes Classification is "naive" because it assumes that all features are independent of each other, which simplifies the calculations.

2.3 Conclusion of the models

K-Nearest Neighbors model (KNN)
The KNN model achieved an accuracy of approximately 65%, indicating that it correctly predicted wine quality for about 65% of the test samples. The F1-scores for different wine quality classes varied, with the highest F1-score of 0.69 for class 5 and the lowest F1-score of 0.00 for classes 4 and 8. This suggests that the model performed well for some classes but struggled to make correct predictions for others due to a lack of training examples. The average F1 score was 0.61 which indicates the average performance across all classes.

Naive Bayes model
The NB model achieved an accuracy of approximately 61%, indicating that it correctly predicted wine quality for about 61% of the test samples. The F1-scores for different wine quality classes also varied, with the highest F1-score of 0.70 for class 5 and the lowest F1-score of 0.00 for class 4. Similar to the KNN model, the NB has problems making predictions in certain classes.

Overall conclusion
The KNN model has a slightly higher accuracy (65%) compared to the NB model (61%).

- Both models have problems making predictions for classes 4 and 8 due too little to no training data
- Both models achieved a moderate overall accuracy but no outstanding performance.

We can conclude that the dataset is imbalanced. Classes 4, 7 and 8 have less data available meaning that these classes have fewer examples to train the model. This makes it difficult for the model to learn meaningful patterns and make correct predictions. Models like KNN and NB rely on learning patterns from training data so these models may not be the right models to use for this specific data set given the fact that the data in this set is not distributed evenly.

3. Reflection Inger

The different steps that I took with this assignment.

Practise with coding

I started by mostly watching videos and doing exercises with programming and coding in python. For this I also asked some fellow students and teacher for different videos and exercises for beginners. From myself, I do not have any experience with programming and with the program Python. Because of this, I also started watching a video and joining and writing every time to learn the basics of Python. Through this I already understood more about the different variables and what some different errors said. In addition, at a later time I had also received from a fellow student a good site where you could do exercises every time with short explanation videos. I must say that I had quite some trouble understanding what I was doing instead of just writing down the codes. This got better and better and I understood every time more.

Steps with dataset

When me and my fellow student had both looked at some information ourselves about the different models, we wanted to use, we got started. We first started by choosing a dataset. The first dataset was also not good to use. This was also a time to learn more about the models and what exactly we needed. Later we came out to use the wine quality dataset.

First, I used the steps in the teaching material what we had to do. In addition, I also used ChatGPT to ask what steps we needed to do. These combined with each other we started to perform. Since in the beginning the steps did match well with what I had learned myself through videos I understood them quite well. Until you get to the point where you get outcomes and my fellow student, and I did not quite understand what it meant. Every time that happened, we looked at it together and asked ChatGPT for an explanation. So that each time we did understand what the outcome meant.

I also found the lessons Mathematics and statistics to be very helpful in learning more about coding. I had the idea that with these lessons it was explained more from the beginning, so I could combine everything together into more knowledge.

Conclusion

I must say that sometimes I am quite surprised at how far I have come at this point and how much I already understand in a short time. I found it very difficult in the beginning and spent a lot of time after school trying to understand as much as possible. Because of this I can currently say that I often understand the bigger picture and sometimes can write down some steps of coding already in one go. ChatGPT helps a lot to build on the good steps and when I get an error that I do not understand I can ask what I should do to solve it.
I think I can use this quite well in maybe my graduate internship and my later work. I also find it quite interesting to continue learning about this and to develop this further in order to become even better. It is also precisely because of the other lessons that you see what you can eventually achieve with coding. Also, every time I get something out of it or make a visualization, it is a small victory that I get to enjoy.

The conclusion is that I have come quite far, but I want to practice a lot more and get better at it. I found it an interesting task to learn more about Python, but it took a lot of time to understand the basics and to continue.

4. Reflection Tara

What I have learned:

At the beginning of Data mining, the world of coding and programming seemed like the most difficult thing for me that made me really overwhelmed. Having zero prior experience in coding and being entirely unfamiliar with Python did not help. I wanted to immediately understand all of it, but I later realised this was simply not realistic and I had to take it step by step.

My initial approach was to dive into Python by watching videos and tutorials and completing simple exercises to grasp the basics. My teammate and I decided to start by focusing on understanding the purpose and inner workings of regression models to be familiar with the bigger context. After this we have chosen a data set for our models. At every step, we made a conscious effort to comprehend what was happening and why. Whenever we encountered errors or unfamiliar terms, we used ChatGPT for added context and clarity. In the face of errors and issues within our code, we adopted a problem-solving mindset. Instead of relying solely on ChatGPT we really wanted to understand why something went wrong and what we where doing. This taught me a lesson of the importance of analytical thinking in coding. I have developed a more comprehensive and global perspective on coding and programming. This has boosted my confidence in this field. I have learned that coding as a skill requires patience and a lot of practice and is not learned overnight. I also like to acknowledge that ChatGPT is a remarkable tool that acted as a tutor-like companion throughout this journey. It was there to provide assistance and clarification whenever I needed it.

I also learned how both models (KNN and Naive Base) work and how you can train these models using Python and a dataset. It was interesting for me to sometimes encounter errors in the code because this resulted in me having to immerse myself in the code, which ultimately made me understand coding a lot better.