# Problem:

At work we receive passenger data for trains collected by ÖBB. This data contains mainly how many passengers have used a train with a certain train number on certain days. This data is collected manually. Therefore, measurements are only collected on some randomly selected days. To make further analyses and evaluations, we would like to predict the passenger numbers on missing days to have a complete dataset of passenger numbers on each day for each train number.

Example:

Data has been collected as follows (in reality, there are a lot more days at our disposal):

|  |  |  |
| --- | --- | --- |
| Date | Train Number | PassengerCount |
| 01.04.2022 | 123 | 250 |
| 03.04.2022 | 123 | 190 |
| 07.04.2022 | 123 | 300 |
| 09.04.2022 | 123 | 200 |
| 11.04.2022 | 123 | 290 |
| 13.04.2022 | 123 | 310 |
| 01.05.2022 | 123 | 150 |

Here we would like to predict the PassengerCount for the missing days (02.04., 04,04, 05.04. 06.04., 08.04., 10.04., 12.04., 14.04. – 30.04.). To do so, we need to “enrich” the date-data, because by itself, a date does not explain the differences in passenger numbers. Therefore, we added information about the day of the week and whether the date corresponds to a holiday or not to account for commuters and possible holiday traffic.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Date | DayOfWeek | IsHoliday | Train Number | PassengerCount |
| 01.04.2022 | 5 | False | 123 | 250 |
| 03.04.2022 | 7 | False | 123 | 190 |
| 07.04.2022 | 4 | False | 123 | 300 |
| 09.04.2022 | 6 | False | 123 | 200 |
| 11.04.2022 | 1 | False | 123 | 290 |
| 13.04.2022 | 3 | False | 123 | 310 |
| 01.05.2022 | 7 | True | 123 | 150 |

Since the date as a whole does not hold much meaning, since one date will only occur once, it would also be necessary to split it up and remove the year part (which provides no additional distinction/explanation regarding passenger numbers – even if a certain year-integer has higher or lower numbers of passengers due to for example corona, this will not manifest in a rule by itself – another variable e.g. IsPandemic would be necessary).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Month | DayOfWeek | IsHoliday | Train Number | PassengerCount |
| 1 | 4 | 5 | False | 123 | 250 |
| 3 | 4 | 7 | False | 123 | 190 |
| 7 | 4 | 4 | False | 123 | 300 |
| 9 | 4 | 6 | False | 123 | 200 |
| 11 | 4 | 1 | False | 123 | 290 |
| 13 | 4 | 3 | False | 123 | 310 |
| 1 | 5 | 7 | True | 123 | 150 |

With this data, it would now be possible to try to predict the passenger count for each day using a method that is capable of regression analysis. Naïve Bayes, therefore, is not an option since it cannot be used for regression.

Viable options include:

* Decision Tree
* Perceptron
* K-NN

For its transparency (white box-algorithm) and ease of understanding, I would start training a decision tree and using the Cross Validation method to find out how well the model fits and how high of a reliability I could expect considering the data. As a quality measure, I would use the accuracy since I want to know how reliable the numbers are that will be predicted.

In a second step, I would consider training a K-NN model. I would test the duration it takes for the K-NN model to be trained and depending on the outcome, I would consider doing Cross Validation again.

# Further Explanation of the real-world problem:

In the real world, the manual data collection is not done on random days, but on fixed days (each Monday, Wednesday and Friday). These days differ by region (VOR, Tirol, Steiermark,…). Therefore, in reality we do not have a sufficiently complete sample of days, but only certain days (which do not change) at our disposal. This means, e.g., for some regions, there is no data on weekends or on holidays. This is a considerable problem, because as it stands now, a reliable prediction can not be expected. A trained model will not be able to consider holidays or weekends at all in some circumstances and others there will not be a distinction possible between Saturdays and Sundays, because there has never been data collected on these days. Therefore, data collection needs to be improved by either making it random and providing at least a chance for all weekdays and possible holidays be considered, or deliberately collect data on these missing days at least at some point in time in one year. This will take time however and collecting data on holidays is highly unlikely due to the fact that data collectors probably will not work on holidays.