Final Assignment Introduction to Modelling semester 2 2024-2025

As a final assignment, you will have to write a report in which you show that you are able to analyze data, present and visualize the results, and predict an outcome. The assessment consists of four assignments that you have to complete. For each assignment you can earn a minimum of 1 point and a maximum of 4 points.

The deadline for the report is Friday 28 March at 23:59. The report must be uploaded to hand-in as a word or pdf file. The results of your analyses should be clearly presented in a word or pdf file. The python files used for analyses should also be uploaded. Upload them as separate files and not as a zip file. Also, please do not copy text from the assignments into your report as it triggers the plagiarism alarm.

The data

The file **turnover.csv** contains data about employees of a very large delivery service company. Unfortunately, this company has a very high turnover rate, and the management wants to know what they can do to reduce this. To do so, they gathered data about employees who recently quit their job. Your job is to create a regression model that can predict how many months an employee is likely to stay before quitting.

Below you can find an explanation of what the variables mean:

- **Months_active:** how many months the employee was active before quitting (this is the dependent variable).
- **Distance_from_work:** how many kilometers the employee lived from the place of work.
- **Age:** The age of the employee.
- **Disciplined:** whether the employee has been disciplined for bad behavior.
- **Children:** How many children the employee had.
- **Social_drinker:** whether the employee went out drinking with colleagues.
- **Social_smoker:** whether the employee was seen smoking with colleagues.
- **Pets:** how many pets the employee had.
- **Weight:** weight of the employee in kilograms
- **Height**: height if the employee in centimeters
- **BMI:** body mass index of the employee (weight / height^2)
- **Absent_hours:** how many hours the employee was absent from their work in an average month, due to sickness or otherwise.

Assignment 1 (15%)

The management first wants to know if there are any important correlations between pairs of variables. Specifically, they want you to check the variables Months_active, Distance_from_work, Age, and Children.

First create a correlation heatmap
Second create a scatterplot matrix
Third explain what the correlations mean in a way that the management can understand (they don't know anything about statistics)

Assignment 2 (20%)

The management also wants to see an explanatory regression model where you predict Months_active using the other variables. So, Months_active should be the dependent variable.

First, create dummy variables and create a regression model with all independent variables together.

Second, check if there is multicollinearity between any of the independent variables. If so, explain how you dealt with it.

Third, present the results in a proper APA table. Make sure to include all the necessary elements.

Assignment 3 (20%)

First, now that you have the model, explain what the meaning is of the independent variables in the models. Do so in such a way that the management can understand it (so in non-statistical language).

Second, there is a discussion in the management team about which variable has the strongest relationship with months_active, after you use multiple regression to control for all the other variables. One group believes that Distance from work has the strongest relationship with months_active. After all, if someone lives far from work this might be a reason to quit. A second group believes that it is actually age that has the strongest relationship. After all, young people are more likely to switch jobs to something better. A third group believes that it is social_drinker that has the strongest relationship. After all, employees who go out drinking with colleagues must have a good bond with them and thereby stay longer. Who do you think is correct? Use the correct statistical techniques to draw your conclusion and make sure to include all the same variables as in Assignment 2-3.

Assignment 4 (45%)

Finally, the management wants you to create a machine learning model that actually makes a prediction on how long an employee is expected to stay before quitting their job. They ask again for your data analysis skills to do this.

First, make a prediction using normal linear regression. Use cross validation with 5 folds and mean absolute deviation. What is the average prediction error?

Second, make a prediction using ridge regression. Don't forget to normalize the data first. Use grid search with cross validation with 5 folds to get the optimal alpha values. What is the best alpha value you found?

Third, create a neural network that can predict the selling price using the following hyperparameters:

Hidden layers	5
Number of nodes per hidden layer	512
Activation function	Relu
Number of Epochs	100
Batch Size	16
Loss function	Mean absolute error
optimizer	Adam

Then evaluate the results with cross validation with 5 folds and mean absolute error scoring. What is the average prediction error? How well does this model work compared to normal linear regression?

END OF THE ASSESSMENT