EhB – Postgraduate in AI ‘21-’22

Data Science Assignment

Please solve the text data assignment, and one of the others, either time series or tabular data!

Submit your **Jupyter notebook, with explanations, justifications and critical assessment of your work included in the notebook as comments**, on **Canvas, by January 5th, 2022.**

# **Assignment 1: Text data set (compulsory)**

Predict whether an sms message is spam or not

*https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection*

Perform the following steps in a Jupyter notebook and provide your comments, justifications on the outcome for each of the steps towards end objective of creating a good fit model to predict the class of each message.

1. Preliminary descriptive statistics, for example:
   1. Determine the frequency of each word in the collection (logplot?)
   2. Determine the strongest correlations among words
   3. Word frequencies per class etc.
2. Feature engineering
   1. Stop words, stemming, feature selection, etc.
   2. Transformation to bag of words, n-grams, etc.
   3. Normalization of frequencies (with reference corpus, or wrt the corpus itself)
3. Model fitting
   1. Fit a decision tree model to predict the target variable
4. Evaluate
   1. Do a proper, critical evaluation within the business context (at least display the confusion matrix)

# **Assignment 2: Predict forest fires (tabular)**

Predict forest fire

https://archive.ics.uci.edu/ml/datasets/Algerian+Forest+Fires+Dataset++

Perform the following steps in a Jupyter notebook and provide your comments, justifications on the outcome for each of the steps towards end objective of creating a good fit model to predict a forest fire based on available attributes.

* 1. Understand the data: descriptive statistics
     1. Data types, missing data, feature correlations, interpretation etc. with respective visualizations
  2. Pre-processing
     1. normalisation, encoding, transformations, imputations
  3. Feature engineering
     1. Dimensionality reduction, normalisation, etc.
  4. Model fitting
     1. Fit a linear regression model and display the RMSE.

# **Assignment 3: Activity monitoring (time series)**

Predict accelerometer data in the future

[*https://archive.ics.uci.edu/ml/datasets/Activity+Recognition+from+Single+Chest-Mounted+Accelerometer*](https://archive.ics.uci.edu/ml/datasets/Activity+Recognition+from+Single+Chest-Mounted+Accelerometer)#

1. a. Understand the data: descriptive statistics

b. Analyse the time series for each axis x,y,z in terms of trend, seasonality and residuals.

c. Analyse the autocorrelation of the time series.

d. Fit a model to forecast/predict the accelerometer data for 1s, 5s and 1 minute.

e. Evaluate how well the predictions are for the same user doing 1/ the same activity, 2/ different activities, and for the same activities for different users.

1. Like the previous tasks, please provide explanations to your findings, justification for your decisions and a proper critical assessment of your results in a Jupyter notebook along with the code.