02/02/2024

PROJECT PROPOSAL

Students:

Name: Lucas Jakin

St. ID: 89211034

Faculty: UP FAMNIT, RIN3-EN

Name: Saša Nanut

St. ID: 89211085

Faculty: UP FAMNIT, RIN2-EN

Selected dataset (from Kaggle): Rain in Australia - Predict next-day rain in Australia
https://www.kaggle.com/datasets/jsphyg/weather-dataset-rattle-package

Project Description (Selection):

As a group, we decided to take on the **first project type**, which can be found on the project registration document. The project focuses on utilizing DM and ML algorithms to address a specific problem chosen from Kaggle. We chose three different datasets that are aligned with our interests and expertise. Each one of these represents a classification problem, meaning that we will try to solve the problem and predict the required information. Moreover, the goal of the project is to address the classification problem by utilizing more than one classification algorithm, in order to do a systematic experimentation with different algorithms to identify in what they differ and which one is the most effective one for the given dataset. Initially, we will perform some data cleaning and preprocessing to handle the missing values and other inconsistencies, if needed depending on the chosen dataset. We will consider different classification algorithms for the experimentation, including Artificial Neural Networks, KNN, Random Forests, CatBoost, Logistic Regression etc. We will implement these algorithms and make the comparison between two of them at least (depending on how many algorithms we'll be able to successfully implement). The algorithms will be evaluated and compared using various evaluation metrics such as accuracy, precision, time of computation and so on. The implementation will be primarily carried out using Python.

Expected Outcome:

The goal is to develop a comprehensive understanding of the classification problems and to apply data mining and machine learning techniques to find solutions. Moreover, we aim to identify the most effective approach between different algorithms.

Division of work:

The project will be undertaken in pairs, meaning that the work will be divided in two. Each of us is going to implement one classification algorithm (if possible also two each). The two algorithms with higher priority of implementation are the Neural Network algorithm and the Catboost algorithm. At the end, we will collaborate on the experimental setup, evaluation metrics and comparative analysis. Both of us will contribute to the interpretation of results.