uc3m

MACHINE LEARNING FOR BUSINESS DECISION MAKING

ASSIGNMENT #2:
FEATURE SELECTION FOR ATTRITION/BURNOUT PREDICTION

2024-25

SOURCE: MICROSOFT COPILOT USING DALL-E 3

INDEX OF CONTENTS

INTRODUCTION

The aim of this assignment is to try and evaluate several feature selection methods, to see if results of the previous assignment can be improved and also if the features selected automatically give some understanding about the problem . This is a continuation of the previous assignment. We will use the same dataset and the same model evaluation strategy (holdout / train-test) for evaluating each of the feature selection methods.

GENERAL CONSIDERATIONS

1. Results **must be reproducible**. Therefore, set the seed at the appropriate places. But instead of using seed 42, use your **Student ID number**.

STEPS TO FOLLOW

1. Use **the same pipeline / method that worked best** for you in the previous assignment.
2. Add **feature selection to the pipeline** and **create two different pipelines**:
    a. Feature selection with **SelectKBest and criterion f_classif.**
    b. Feature selection with **SelectKBest and criterion mutual_info_classif.**
3. Use **grid search (HPO)** to **tune the number of features** to be selected (k) by each of the pipelines.
    - **Note**: the remaining hyper-parameters will not be tuned, but left to their default values.
4. **Evaluate the models obtained** with the two pipelines on the testing dataset **What is the best feature selection method? How many features are actually selected? Which features are actually selected? Are results improved compared to the previous assignment?** Please**, explain the reasons**.
5. Using the best feature selection technique, **fix the number of features to be selected to the optimal number found in section 4**. That means that you have to define again the pipeline, but now fixing k to the value that worked best in section 4. **Then, do hyper-parameter tuning on the hyper-parameters of the method**.
6. **Are results further improved?** (on the testing dataset) Please, **explain the reasons**.
7. **Make predictions for the competition data, save them, and save the model**.

WHAT TO HAND IN

- A **jupyter notebook** with the code in the proposed order of steps. Please **use some of the cells to comment** about what you are doing and your results. In particular, emphasise your conclusions after each step with short arguments based on your results. **Better and clear reports will get better grades**.
  If it is more convenient, you can also hand in a file with Python code instead and a separate report.
  If you decide to use any AI chatbot, **briefly explain in those commented cells what purpose you used it for and how you used it (for instance, you can quote the prompt and the output used**).
  Please **write the names of the components of your group at the beginning of the notebook**.

- A file containing your **final trained model** in an appropriate ML format.

- A pickle or .csv file containing **your final model's predictions (values of your model's predictions** in the competition set).

DUE DATE

Initially the assignment due date is set for December 13th by 23:59.