UNIVERSIDAD DE GRANADA

Master Erasmums Mundus COSI

# Data Science

## Practical Assignment (2023-2024)

# 1. Objective

The objective of this assignment is to enforce the practical application of the knowledge and abilities gathered by students along the course on “Data Science” in the program of the COSI Master. The student should prove that he or she is able to comply with the whole process of data analysis and predictive modeling.

# 2. Problem: Student Performance Prediction

The objective is to develop a predictive model to predict a student’s success or risk of dropping out based on a number of information known at the time of student enrollment — academic trajectory, demographics and social-economic factors. These data are collected from students of a polytechnical University.

The dataset is composed of 3,500 instances each with 36 attributes along with the target value. It is available as `dataset.csv` at the Prado Platform.

The feature columns are:

1. `Marital status`: The marital status of the student.
2. `Application mode`: The method of application used by the student.
3. `Application order`: The order in which the student applied.
4. `Course`: The course taken by the student.
5. `Daytime/evening attendance`: Whether the student attends classes during the day or in the evening.
6. `Previous qualification`: Grade of previous qualification.
7. `Nationality`: Nationality of the student.
8. `Mother’s qualification`: The qualification of the student’s mother.
9. `Father’s qualification`: The qualification of the student’s father.
10. `Mother’s occupation`: The occupation of the student’s mother.

11. `Father's occupation`: The occupation of the student's father.
12. `Admission grade`: Student admission grade.
13. `Displaced`: Whether the student is a displaced person.
14. `Educational special needs`: Whether the student has any special educational needs.
15. `Debtor`: Whether the student is a debtor.
16. `Tuition fees up to date`: Whether the student's tuition fees are up to date.
17. `Gender`: Gender of the student.
18. `Scholarship holder`: Whether the student holds a scholarship.
19. `Age at enrollment`: The age of the student at the time of enrollment.
20. `International`: Whether the student is an international student.
21. `Curricular units 1st sem (credited)`: The number of curricular units enrolled by the student in the first semester.
22. `Curricular units 1st sem (evaluations)`: The number of curricular units evaluated by the student in the first semester.
23. `Curricular units 1st sem (approved)`: The number of curricular units approved by the student in the first semester.
24. `Curricular units 1st sem (grade)`: The number of curricular units grade by the student in the first semester.
25. `Curricular units 1st sem (without evaluations)`: The number of curricular units excluding grades, in the first semester.
26. `Curricular units 2nd sem (credited)`: The number of curricular units enrolled by the student in the second semester.
27. `Curricular units 2nd sem (evaluations)`: The number of curricular units evaluated by the student in the second semester.

28. `Curricular units 2nd sem (approved)`: The number of curricular units approved by the student in the second semester.
29. `Curricular units 2nd sem (grade)`: The number of curricular units grade by the student in the second semester.
30. `Curricular units 2nd sem (without evaluations)`: The number of curricular units excluding grades, in the second semester.
31. `Unemployment rate`: Unemployment rate.
32. `Inflation rate`: Inflation rate.
33. `GDP`: Gross Domestic Product.

The goal is to predict the `Target` value for each student. Proper performance metrics must be defined.

# 3. Tasks

The student has to analyze the data and build effective predictive models. All the techniques studied along the course should be considered, namely:

1. KNN (`mknn`)
2. Linear Discriminant Analysis (`mlda`)
3. Logistic Regression (`mlr`)
4. Classification trees (`mtree`)
5. Random Forests (`mrf`)
6. Artificial Neural Networks (Multi-layered perceptrons) (`mann`)
7. Support Vector Machines (`msvm`)
8. Clustering techniques.

The student should assess wether the technique is suitable for the task or not. Then he or she must define how to assess the models, stating the measures to be used. Both, accuracy and complexity of the models should be taking into account.

Then he or she must apply each technique by:

- Doing preprocessing as necessary.
- Exploring as many configurations as needed.
- Building an effective model.

The the overall results should be analyzed and propose a final model. Justify the choice.

As extra work the student can include extra sections considering additional techniques or preprocessing. For example,

- Boosted trees
- Additional ANNs
- Unsupervised methods
- Using other data analysis tools, e.g. python.
- . . .

# 4. Evaluation

For the evaluation of the developed work, the student has to deliver a report describing the work done. The report should have the following structure:

- Title page, including course title and student name
- Problem description
- Section for technique 1

  . . .
- Section for technique $m$
- Result analysis
- Final model selected
- Result analysis
- Conclusions

The report should be in pdf format. In addition, an R script including all the commands to obtain all the final models should be produced. The name of the models are indicated previously, e.g. `mknn`, . . . These objects should be available at the end of the script.

Both the report (in pdf) and the R script should be uploaded, within a zip file, to the “Practical Assignment” task on the Prado platform no later than **May 15th, 2024**.