Accurate and Personalized Academic Advising
(Meetings Log)

2nd April Meeting :

Some points discussed during the meeting:

- We are going to predict grades instead of probability, since grades is more consistent with the data we have.
- For the prediction, features from the previous paper “**Techniques for Data-Driven Curriculum Analysis**” could be helpful.
- For the data:
    - We start by performing **Data anonymization**
    - We will only consider data from Computer Science department (so the data will be more homogenous and interpretable given previous background knowledge)
- The Basic use case scenario would be:
    - **Input:** a set of courses to be taken by a specific student.
    - **Output:** for each course, a prediction for the grade.
    - The prediction would be based on student’s history and the general history for previous students.
- The expected deliverables for the next meeting on **11th April**:
    - Data anonymization
    - Creating Meeting Logs shared document
    - Environment setup
    - Training set discussion and plan

Meeting April 5th:

- Ask difference between “Fiscal” and “Nacional”
- Extract also the first word/unigram, the first bigram and the last word/unigram of the high school name as features

    - Cities with multi-word names: SANTA ELENA, EL EMPALME, MARCELINO MARIDUEÑA
- Each course becomes a feature accompanied by a boolean variable that says whether the student has taken the lecture or not (to help the model know when to ignore a column)
- We could have multiple training sets:
    - One for lectures taken in the first semester (no academic history)
    - One for lectures taken in from the second to the third semester (short academic history)
    - One for lectures taken from the third semester
    - You have to use the admission year to know when a course was taken
- Paper doing something similar: https://peer.asee.org/regression-models-for-predicting-student-academic-performance-in-an-engineering-dynamics-course.pdf
- Questions:In “SEMESTER” column, we have the following values (1S, 2S, 3S, 1T, 2T, 3T, 4T, 5T, 6T, 7T). What T’s stand for ?
    - By third semester, we mean first semester from second year ?

Meeting April 11 th:

- Answers to questions:
    - S and T are the same thing
    -
    - 3S = break period
- Toy example to predict the grade of Algorithm Analysis (Análisis de Algoritmos - FIEC05066) based on the students’ personal information (P factor, school, etc…) + the grades of Programming Fundamentals (Fundamentos de Programación), Data Structures (Estructura de Datos) and Discrete Mathematics (Matemáticas Discretas)
- We’ll start with basic linear regression https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html#sklearn.linear_model.LinearRegression
- Boolean variable now encodes the multiplicity of the lecture (# times the student took it).
- Lectures not taken by the students will be imputed with average grades
- We could use reduced models (look for good literature)
- ANALISIS DE ALGORITMO; FUNDAMENTOS DE PROGRAMACIÃ“N; ESTRUCTURAS DE DATOS, MATEMÃTICAS DISCRETAS (IEC)

Meeting April 15th

- Add the additional features to the first toy example
- Data Structures and Discrete Mathematics are the best predictors for the performance of Algorithms Analysis
- Second toy example: predict the grade of Statistics (Estadística, ICM00166) based on the grades for Multivariable Calculus (Cálculo de varias variables, ICM01966), Linear Algebra (Álgebra Lineal, ICM00604), Diferential Calculus (ICM01941), Integral Calculus (ICM01958)

Meeting April 17th

- Other multitoken cities: PEDRO CARBO, BAHIA DE CARAQUEZ, SANTO DOMINGO DE LOS COLORADOS
- Add always the P factor
- Take into account the number of times the student took the target lecture (minus 1)
- Try without unigrams/bigrams and see the performance

- Multivariate calculus correlates well with the other lectures
- Linear Algebra is the strongest predictor for performance in Statistics
- Adding the course count, and courses averages as features and target reduced the RMSE by 23%, and the MAE by 16%

Meeting April 18th

- Verify that Linear Algebra was taken before Statistics
- We'll look at the older curriculum to get more data
- Two models: repeaters and non-repeaters
- Additional features: difficult estimator per lecture (consider the professor who taught the lecture), combined difficulty of the lectures of the semester
- For latter: combined difficulty per component
- Predict the performance in the Statistics course from the performance of ICM00646 (Cálculo II), ICM00216 (Cálculo I), ICM00604 (Álgebra Lineal)

Results:

3rd Experiment Results

Meeting April 25th

- Unify data for repeaters and non-repeaters using imputation with the average grade of non-repeaters (since we do not have the latest grade for them).
- Boolean variable stating whether a student has repeated or not the target lecture

- Next stages: include the difficulty estimators to simulate the planification of a semester.

Some results using HIPAR

Best model using LassoLARS as regression method
{count_ICM00646>=2.0} ===> 6.217391 : amax_ICM00166 = 6.2174 | ERROR = 0.62 ---- SUPP : 2.74%
{PrevLatestGrade_ICM00646<6.9 & BETA<4.4253} ===> 6.653778 : amax_ICM00166 = 6.6538 | ERROR = 0.655 ---- SUPP : 46.98%
{PrevLatestGrade_ICM00216>=8.675} ===> 8.097826 : amax_ICM00166 = 8.0978 | ERROR = 0.9849 ---- SUPP : 2.74%
{PrevLatestGrade_ICM00166>=6.7424 & count_ICM00646<1.0} ===> 6.855628 : amax_ICM00166 = 6.8556 | ERROR = 0.7566 ---- SUPP : 54.63%
{PrevLatestGrade_ICM00166<6.7424} ===> 6.685333 : amax_ICM00166 = 6.6853 | ERROR = 0.5789 ---- SUPP : 35.54%
{amax_ICM00604<6.825 & PrevLatestGrade_ICM00604>=6.05} ===> 6.739083 : amax_ICM00166 = 6.7391 | ERROR = 0.5564 ---- SUPP : 27.13%
{(7.45<=amax_ICM00216<8.65) & count_ICM00216<2.0} ===> 7.110764 : amax_ICM00166 = 7.1108 | ERROR = 0.6666 ---- SUPP : 17.01%
{amax_ICM00216<7.45 & PrevLatestGrade_ICM00646>=6.95} ===> 6.890323 : amax_ICM00166 = 6.8903 | ERROR = 0.6131 ---- SUPP : 22.02%
{Default} ===> 6.727955 : amax_ICM00166 = 6.728 | ERROR = 0.8326 ---- SUPP : 100.0%

Performance in cross-validation
RMSE: 0.7960765605252886
MAE: 0.5648708118928444
MeAE: 0.4614665254711303
Performance in test set
RMSE: 0.7992615585535413
MAE: 0.5548563227979878
MeAE: 0.4603782503779059

Meeting May 15th

- Unified model to predict the mark of more than one lecture
- PCA analysis
- Fine-grained alphas/betas
    - Per factors as suggested in Gonzalo's paper
    - Per professor/lecture
- Task: do the same analysis as done for Statistics, for Programming Languages (FIEC01552), with Data Structures (Estructuras de Datos, FIEC03012)
  , Object Oriented Programming (FIEC04622), and Programming Fundamentals (FIEC04341) as prerequisites.

Meeting May 17th

We will predict the performance in "Análisis de Redes Eléctricas" (FIEC01735) using the grades of Physics and Math lectures: ICM00646 (Cálculo II), ICM00216 (Cálculo I), ICM00604 (Álgebra Lineal), ICM00653 (Cálculo III), ICF00463 (Física I), ICF00489 (Física II)

Meeting June 3rd

Few points to explore:

- Sum of Skewness as new feature (**Done**)
- Aggregate alphas as product (**Done**)
- Calculate difficulties estimators (alpha, beta, skew) per course-professor (**Done**)
- Unified model using Basic factors mentioned in Gonzalo's paper

Meeting July 19th

Unified model for the lectures of a given semester. We create an example 2nd semester with the following lectures:
Data Structures (FIEC03012), Linear Algebra (ICM00604), Calculus II (ICM00646), (ICF00489)
The common set of prerequisites consists of:
ICM00216 (Cálculo I) -> Calculus I

ICF00463 (Física I) -> Physics I
Programming Fundamentals (FIEC04341)
Introduction to Informatics (FIEC04358)
Discrete Mathematics (ICM00901)

Target score label
Imputation using average score of the lecture in the past
Difficulty estimators per professor (max alpha, beta, min skewness)

Meeting August 9th

- Sketches Discussion
- Stereo-typical semester mining
- Map the output of Stereo-typical semester to strings
- Automate the process so the prediction algorithm takes transactions(representing stereo-typical semester) as input and build a model for each semester

Meeting August 27th

RMSE unified model

```
Global:  0.7767596722960277
Partial for FIEC02097  =  0.8185435357748588
Partial for FIEC03053  =  0.7066057381354042
Partial for FIEC01545  =  0.7882106380974744
```

RMSE individual models

```
*********************************************************************
***************************
                                            FIEC01545:
*********************************************************************
***************************
Linear Regression:
-----------------------------------------------------
    === Cross-validation ===
    === Summary ===

    Correlation coefficient                0.5223
    Mean absolute error                    0.5891
    Root mean squared error                0.7561
    Relative absolute error                87.2913 %
    Root relative squared error            85.5697 %
    Total Number of Instances              439
-----------------------------------------------------
Random Forest:
-----------------------------------------------------
    === Cross-validation ===
    === Summary ===

    Correlation coefficient                0.5039
    Mean absolute error                    0.5622
    Root mean squared error                0.7642
    Relative absolute error                83.3096 %
    Root relative squared error            86.4873 %
*********************************************************************
***************************
                                            FIEC02097:
*********************************************************************
***************************
Linear Regression:
------------------------------------------------------
    === Cross-validation ===
    === Summary ===

    Correlation coefficient                0.5264
    Mean absolute error                    0.6582
    Root mean squared error                0.8092
    Relative absolute error                84.3762 %
    Root relative squared error            85.284  %
-------------------------------------------------------
Random Forest
-------------------------------------------------------
    === Cross-validation ===
    === Summary ===

    Correlation coefficient                0.5837
    Mean absolute error                    0.6157
    Root mean squared error                0.7692
    Relative absolute error                78.9245 %
    Root relative squared error            81.0686 %
*********************************************************************
*****************************
                                            FIEC03053
```

```
**********************************************************************
******************************
Linear Regression
---------------------------------------------------------
    === Cross-validation ===
    === Summary ===

    Correlation coefficient                0.4157
    Mean absolute error                    0.5518
    Root mean squared error                0.7088
    Relative absolute error                92.7069 %
    Root relative squared error            92.2327 %
---------------------------------------------------------
Random Forest
---------------------------------------------------------
    === Cross-validation ===
    === Summary ===

    Correlation coefficient                0.4124
    Mean absolute error                    0.5317
    Root mean squared error                0.6996
```