FEDERAL OF UNIVERSITY GOIAS
INSTITUTE OF MATHEMATICS AND STATISTICS

SYLLABUS

1. General Information:

Course: Introduction to data science using R
Lecture Code: IME0344
Credit Hours: 64h
Year/Semester: 2020/1
Course Designation: Undergraduate Students
Hours: 24N45
Professor: Renato Rodrigues Silva

2. Summary

Introduction to R; Data manipulation; Exploratory data analysis; Statistical Modelling in R; Effective visualization for communication.

3. Course Outline

1. Introduction to R: Basic Data Types: Vector, Matrix, Data Frame and lists. Programming e.g. functions, loops, control structures.

2. Data manipulation: Data manipulation with tidyverse, Tidy data, relational data and data import.

3. Exploratory data analysis: Visualization with ggplot2, Summary Statistics with R.

4. Statistical Modelling in R: Multiple Regression, Logistic Regression, Classification and Regression Trees, Bagging, Random Forests, Boosting and Neural Networks.

5. Effective visualization for communication: R Markdown and Interactive Graphics with shiny.

4. Schedule

Lecture 1 - An Introduction To Data Science (synchronous activities)	March 2	2 credit hours
Lecture 2 - An Introduction To R (synchronous activities)	March 4	2 credit hours

Lecture 3 - An Introduction to Data Science part II (synchronous activities)	March 9	2 credit hours
Lecture 4 - Tibbles and Data Visualization (synchronous activities)	March 11	2 credit hours
Lecture 5 - Visualization of the data (synchronous activities)	March 16	2 credit hours
No Classes (Covid19 Outbreak)	March 17 - August 30	-
Lecture 6 - General Information (synchronous activities)	August 31	2 credit hours
Lecture 7 - An introduction to R (synchronous activities)	September 2	2 credit hours
Lecture 8 - Video Activities - Simple Linear Regression (synchronous activities)	September 9	2 credit hours
Lecture 9 - Simple Linear Regression (synchronous activities)	September 14	2 credit hours
Lecture 10 - Video Activities - Multiple Linear Regression (synchronous activities)	September 16	2 credit hours
Lecture 11 - Multiple Linear Regression (synchronous activities)	September 21	2 credit hours
Lecture 12 - Lecture Exercises I (synchronous activities)	September 23	2 credit hours
Lecture 13 - Assignment I (asynchronous activities)	September 28	2 credit hours
Lecture 14 - Assignment I (asynchronous activities)	September 30	2 credit hours
Lecture 15 - Video Activities - Logistic Regression (synchronous activities)	October 5	2 credit hours
Lecture 16 - Logistic Regression (synchronous activities)	October 7	2 credit hours
Lecture 17 - Lecture Exercises II (synchronous activities)	October 14	2 credit hours
Lecture 18 - Assignment II (asynchronous activities)	October 19	4 credit hours
Lecture 19 - Regularization (synchronous activities)	October 21	2 credit hours
Lecture 20 - Cross Validation and Accuracy (synchronous activities)	October 26	2 credit hours
Lecture 21 - Lecture Exercises III (synchronous activities)	November 4	2 credit hours
Lecture 22 - Assignment III (asynchronous activities)	November 9	2 credit hours

Lecture 23 - Video Activities - Classification and Regression Trees (synchronous activities)	November 11	2 credit hours
Lecture 24 - Classification and Regression Trees part I (synchronous activities)	November 16	2 credit hours
Lecture 25 - Classification and Regression Trees part II (synchronous activities)	November 23	2 credit hours
Lecture 26 - An introduction to bootstrap (synchronous activities)	November 25	2 credit hours
Lecture 27 - Video Activities - Bagging and Random Forest (synchronous activities)	November 30	2 credit hours
Lecture 28 - Bagging and Random Forest (synchronous activities)	December 2	2 credit hours
Lecture 29 - Assignment IV (asynchronous activities)	December 7	2 credit hours
Lecture 30 - Video Activities - An introduction to neural network (synchronous activities)	December 9	2 credit hours
Lecture 31 - An introduction to neural network (synchronous activities)	December 14	2 credit hours
Lecture 32 - Assignment V (synchronous activities)	December 16	2 credit hours

This syllabus is a guide and every attempt is made to provide an accurate overview of the course. However, circumstances and events may make it necessary for the instructor to modify the syllabus during the semester and may depend, in part, on the progress, needs, and experiences of the students. Changes to the syllabus will be made with advance notice.

5. General objectives of teaching

Develop a theoretical-practical understanding of data science and R skills so that students to be able to solve practical problems.

6. Learning Objective

At the end of the course, students should be able to:

- Use R to import, clean, process, model and visualize data.
- Apply statistical and computational methods to make inference, predictions and or classification based on data.
- Effectively communicate the outcome of data analysis using R tools.

7. Methodology

- The professor will use the SIGAA platform for communication and Google Classroom for classes online (synchronous activities). Class notes will be available on the GitHub and SIGAA repository.
- The classes online will be in Portuguese. Additional resources might be used such as other bibliographies, educational websites and etc.
- Asynchronous activities will be class videos and assignments.
- The videos recommended were made in English.The professor will provide more details about that during the course.

WARNING !!!

- Classes will not be recorded. However, if the student does so, the reproduction, distribution and / or publication of excerpts or the entire content of the recorded classes is prohibited.
- Only the professor and students regularly enrolled in this discipline will have access to the virtual teaching environment. It depends on the authorization of the professor the access of others to the virtual environment, who may not be directly involved with the activities developed in it.
- The students must use the institutional email during the class.

8. Grading

- Assignment one : Multiple and or Simple Regression (A1) is due to October 5.
- Assignment two : Logistic Regression (A2) is due to October 19.
- Assignment three : Regularization (A3) is due to November 9.
- Assignment four : Classification and Regression Trees and Random Forest (A4) is due to December 7.
- Assignment five : Neural Network (A5) is due to December 16.
- Final Grade (MF): $MF = 0.2A1 + 0.2A2 + 0.2A3 + 0.2A4 + 0.2A5$
- Grading criteria follows academic regulations of Federal University of Goias. The grades will be posted via SIGAA.

It is **mandatory to do the assignments in small workgroups (minimal 2 students, and maximal 5)**. Otherwise, they will be not accepted.
Submitting assignments: It is could be in the video, audio, and report format. maximum upload file size: 2 MB (for pdf file report). Videos must be submitted by google drive platform.

9. Bibliography

[1] WICKHAM H., GROLEMUND, G. *R for Data Science.* O'Really, 2016.

[2] BERK, A. R. *Statistical Learning from a Regression Perspective* . Springer, 2008.

[3] WICKMAN, H. ggplot2: *Elegant Graphics for Data Analysis (Use R!).* Springer, 2009.

[1] HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of. Statistical Learning: Data Mining, Inference, and Prediction.* Second Edition. February 2009.

10. Complementary bibliography

[1] DALGAARD, P. *Introductory Statistics with R.* Springer, 2nd Edition 2008.

[2] GENTLEMAN.*R Programming for Bioinformatics.* CRC Press 2008.

[3] CHAMBERS, J.M. *Programming with Data: A Guide to the S Language.* New York: Springer, 2008.

11 Course Textbook

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of. Statistical Learning: Data Mining, Inference, and Prediction.* Second Edition. February 2009.

12. Office hours

There will be not an individual appointment in my office, the student will able to be schedule an individual appointment via Google Meeting to discuss their draft assignment a few days prior to the deadline or any concerns they have about it. Monday 6:00- 6:50 pm.