

Quick intro about the TFM

Master in Data Science, 11th ed. (MAD)

Goals of the TFM

- Students will show their capacity to work as data scientists
- Starting from raw data, all the way up to solving a research question
- TFM will include the following phases:
 - Data acquisition
 - Data cleansing and preparation
 - Analysis
 - Frontend / visualization
- The question can be refined after several iterations working with the data

Deliverables

- Repository with all the code and ready to be executed
 - **Please provide instructions** to install any package or dependency. Easiest to use a dependency/package manager like conda.
- Document explaining the TFM
 - Two objectives to be optimized with the document:
 - As brief as possible
 - But at the same time containing enough details so the work could be replicated
 - Any data scientist with the document and the repository should be able to replicate the TFM
- Frontend or visualization
 - What the customer/user would interact with.
- Presentation: an in person presentation covering the TFM. Explain data, internal structure, design decisions.

Suggested structure of the document

- Introduction: what, why, why is it relevant, any previous related work/state of the art
- Raw data description
- Methodology: machine learning techniques used, statistical methodologies
- Summary of main results
 - Detailed results will be available by running the code in the repo
- Conclusions
 - Not a summary of the work. The problem was relevant, now with your work, what can you say about how the problem is solved?
- User manual for the frontend

Technical restrictions

- None
- You can use any technology
 - As long as the evaluators can get access to that technology too
- The evaluators need to have access to the data, too.
 - NDAs have been signed in the past for sensitive data and projects.

Phases during the master

- Write message in Basecamp specifying topic, link to repo and group
 - **Message topic should start with [TFM]**
 - **1 or 2 people per group**
 - **Deadline: February 10th, 2019**
 - PLEASE DON'T WRITE ME AN EMAIL, post in Basecamp
- Short presentation of the planned project
 - Not graded, helps you stay on track
 - **February 16th, 2019**
- Deliver repo and main document
 - Include the document in the repo
 - **Deadline: 16th May 2019.**
- Presentation: an in person presentation covering the TFM. Explain data, internal structure, design decisions.
 - **May 24th/25th, 2019**

Recommendations with the repo:

- Public repo for greater visibility
 - Better future opportunities
- **Please don't upload private data to the repo**
- Clean repo, with proper README.md
- Preferably in English
 - But not mandatory

Evaluation criteria

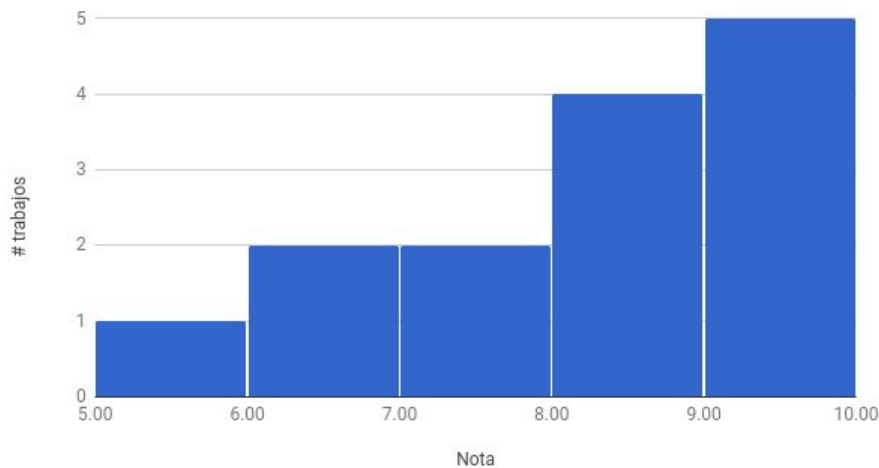
- Clarity in the document, easiness to replicate work
- Complexity
 - In the data and the analysis methodologies
 - More complex is better
- Clarity and correctness of source code
- Relevance and fit-to-purpose of the chosen analytical methods
- Relevance and fit-to-purpose of the chosen technologies
- UX and usability of the frontend
 - We will evaluate as the “consumer” of the report, with zero knowledge of data science
- 0-10 points
 - 0.1 points less per hour of delay

Links

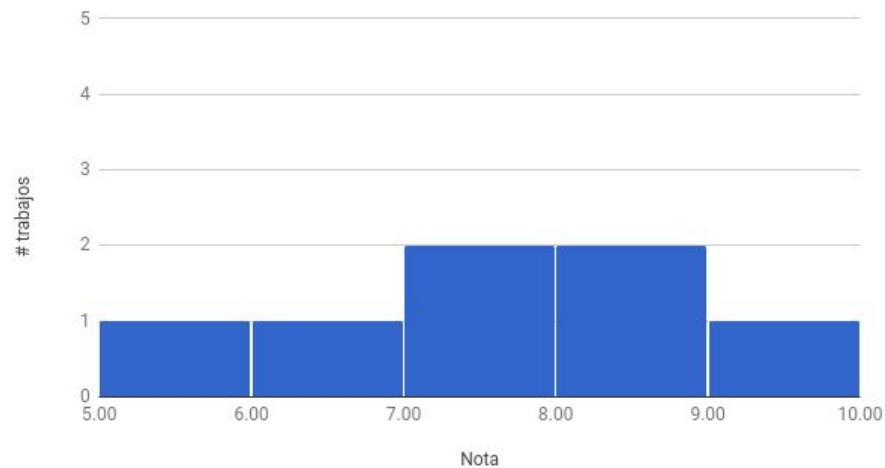
- [Rules for the TFM](#)
- Some links in the KSchool web
 - <http://kschool.com/blog/data-science/martacamarads/>
 - <http://kschool.com/blog/big-data/jmvaldeolmillos/>
 - <http://kschool.com/blog/formacion/gonzalo-sanchez-tfm-data-science/>
 - <http://kschool.com/blog/data-science/jose-manuel-vera-data-scientist/>
 - <http://kschool.com/blog/data-science/tfm-data-science-describiendo-tendencias-busquedas-google-utilizando-tweets-relacionados/>
 - <http://kschool.com/blog/data-science/tfm-data-science-manuel-maestre-estimacion-precios-del-alquiler/>
 - <http://kschool.com/blog/data-science/tfm-data-science-banca/>

Results of previous editions

1st edition

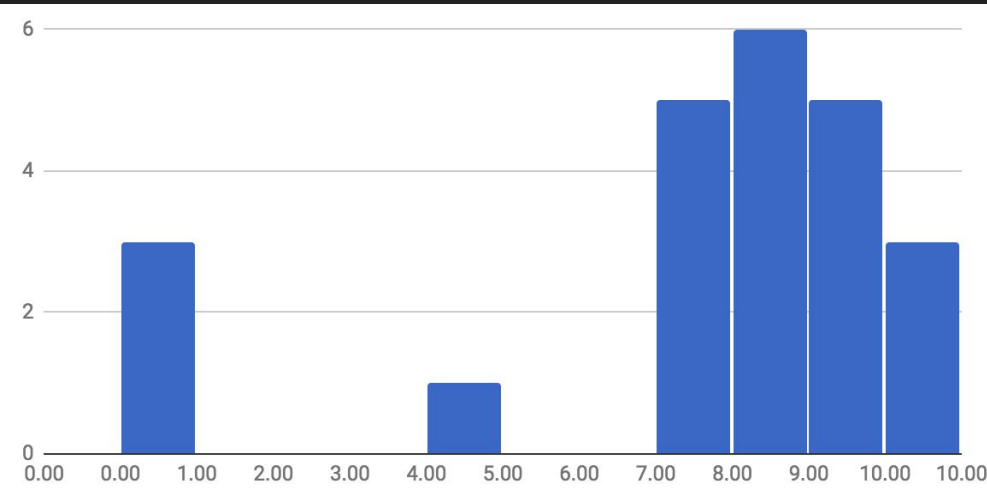


2nd edition

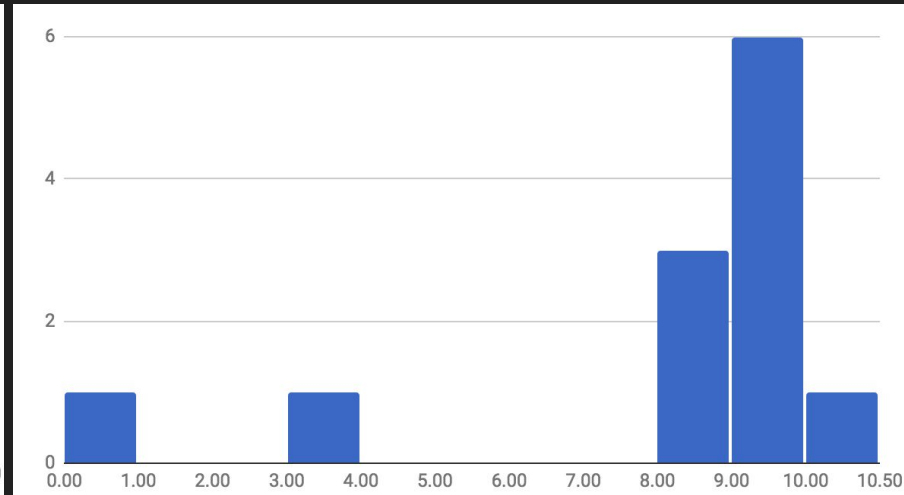


Results of previous editions

3rd edition

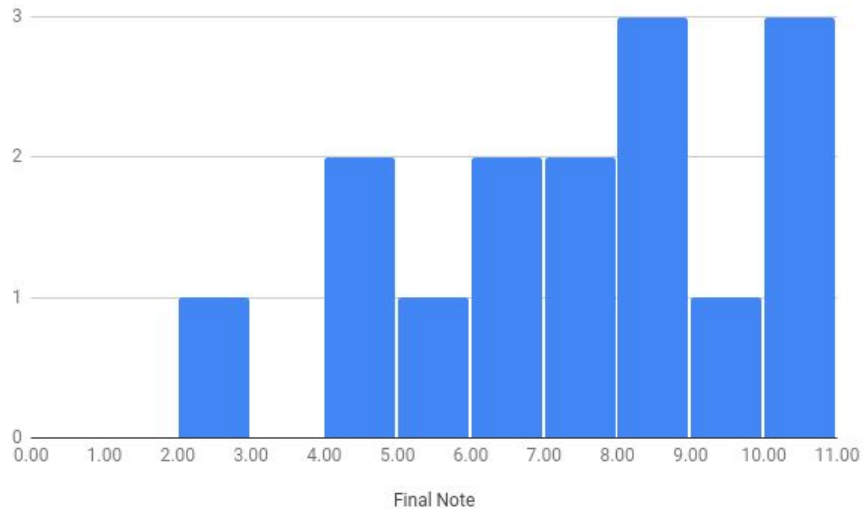


4th edition

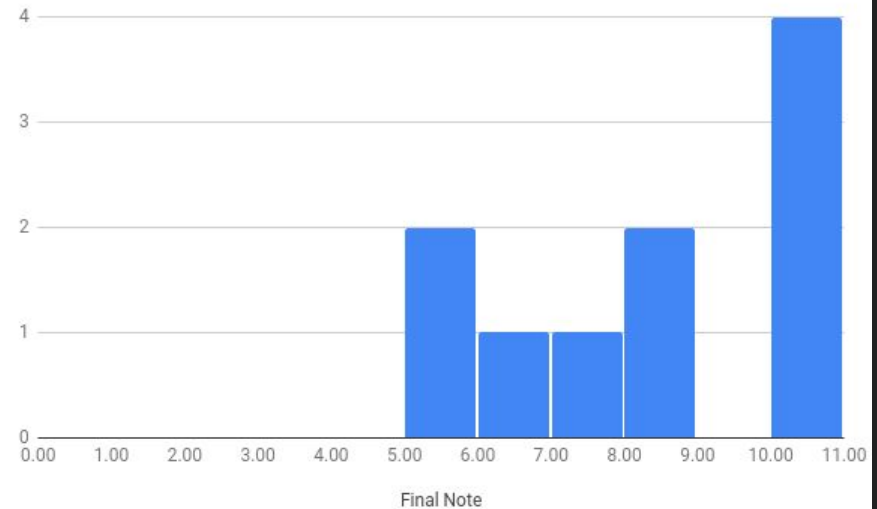


Results of previous editions

5th edition



6th edition



Questions?

- I don't find data for my idea
 - We can share data and propose a topic
 - Airlines' customers segmentation
 - Analysis of factors affecting flight delays
 - Analysis of the Ethereum blockchain
- Can I repeat the same topic of a previous TFM?
 - Yes
- More questions :)