**Instituto Tecnológico**

**y de Estudios Superiores de Occidente**

Department of Electronics, Systems and Informatics

**Master in Computer Systems**



**Final Project**

**Assignment: Management and Analysis of Massive Information**

Presented by:

Carlos Alberto Cordero Robles (MS197686)  
75% Credits Approved

Jehovany Alejandro Rodríguez Vázquez (MS678536)  
25% Credits Approved

Professor: Dr. Alberto De Obeso Orendain

Tlaquepaque, Jalisco. May 2020

# Content

[I. Content 2](#_Toc39775993)

[II. Introduction 3](#_Toc39775994)

[III. Data Integration 5](#_Toc39775995)

[IV. Clustering 6](#_Toc39775996)

[V. Dimensionality Reduction 9](#_Toc39775997)

[VI. Neural Network 10](#_Toc39775998)

[VII. Genetic Algorithm 12](#_Toc39775999)

[VIII. Conclusions 15](#_Toc39776000)

[IX. References 16](#_Toc39776001)

# Introduction

The objective of this document is to explain the implementation made during the assignment “Management and Administration of Massive Information” as final project.

Before starting the description and the objective of this project is important to know the concept of data mining and big data in order to do that, we’ll use Data Science.

Data Science is a field with inter-disciplines that mostly use scientific methods, processes and system to somehow extract knowledge and insights from many structural and unstructured data. There’s a well know Venn diagram that represents how the disciplines interact:

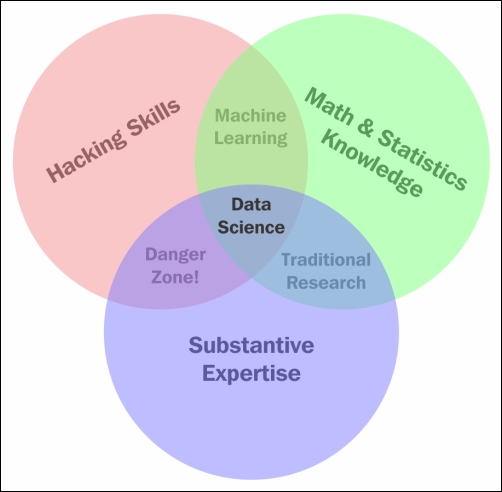


Figure 1. The Venn diagram of data science[1]

Understanding data science begins with three basic areas[1]:

**Math/statistics:** This is the use of equations and formulas to perform analysis.

**Computer programming:** This is the ability to use code to create outcomes on the computer.

**Domain knowledge:** This refers to understanding the problem domain (medicine, finance, social science, and so on).

This project consist in a deep analysis of a provided dataset with diverse student information in order to apply all the Data Science tools and techniques learned and practiced during the course. The main objective is to analyze the impact and possible desertion causes in order to provide with a prediction and best probable solution to the still rescuable cases so the student can finish school and obtaining a grade.

Now, let’s talk about the problem that we’re trying to analyze, for that let’s talk about scholar desertion, specifically, in Mexican universities.

Scholar desertion is a problem that concerns all superior education institutes in Mexico, according to ANUIES (National Association of Universities and Higher Education Institutions) in a research made in 2000 from 100 students coursing bachelor, 60 graduate and only 20 of those 60 actually obtain a degree. According to this study the first year during a student career is crucial to their permanence[2].

This project uses data science tools and techniques that facilitate the analysis of the information in order to determine the relationship between the different desertion cases and how this data can be considered for the superior education institutes to reduce desertion in Mexico.

# Data Integration

We’ve included as a single Data Frame using R Studio the following Data Set:

AsistenciasTotales.R, perfilAlumnos.R, ResultadosExamenes.R, ResultadoTrabajos.R, UsoBiblioteca.R, UsoPlataforma.R, ApartadoDeLibros.R, Becas.R, HistorialPagos.R, EvaluacionProfesorMateria.R, CambioCarrera.R.

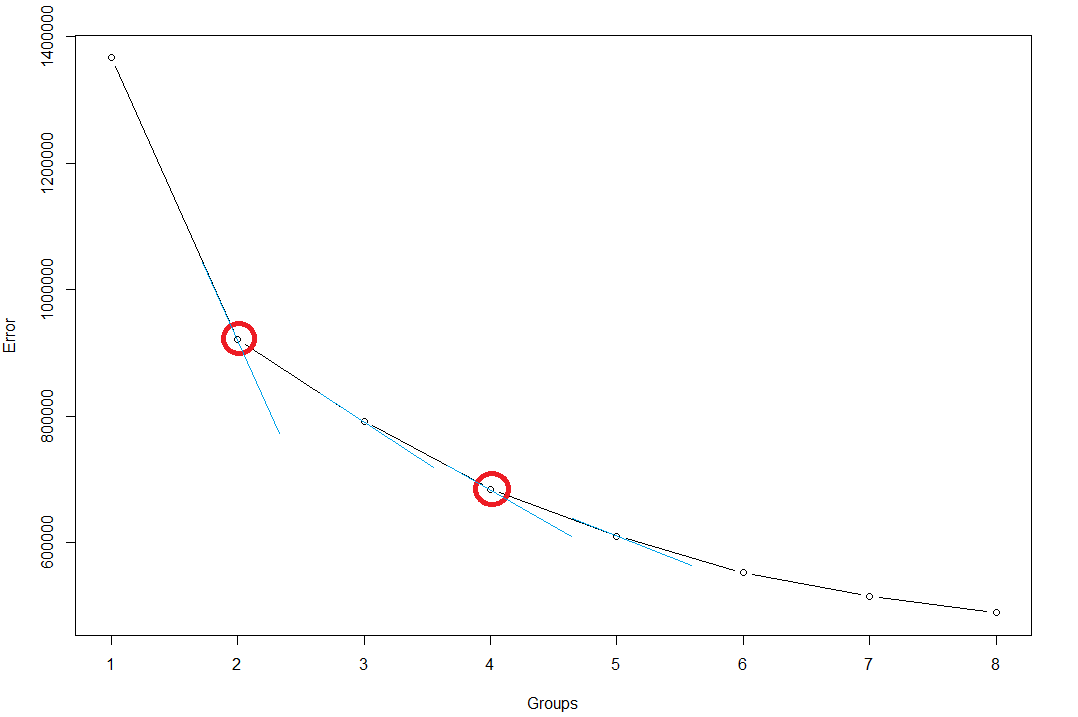
Getting a dataset with the following elements to analyze:

|  |  |
| --- | --- |
| Colname | Description/Values |
| genero | 1 genderA 2 genderB |
| admision.letras | % obtained |
| admision.numeros | % obtained |
| promedio.preparatoria | % obtained |
| edad.ingreso | Student age |
| nota.conducta | conduct grade over 20 |
| scholarship | 1 Yes 0 No |
| prev.change | Amount of changes |
| economic.level.4 | Low |
| economic.level.3 | Medium-Low |
| economic.level.2 | Medium-High |
| economic.level.1 | High |
| payment.Semester1 | 2 paid on time 1 late 0 not paid or paid too late |
| payment.Semester2 | 2 paid on time 1 late 0 not paid or paid too late |
| payment.Semester3 | 2 paid on time 1 late 0 not paid or paid too late |
| assistance1 | 0 miss 1 late 2 Present |
| assistance2 | 0 miss 1 late 2 Present |
| assistance3 | 0 miss 1 late 2 Present |
| exams1 | exam grade over 20 |
| exams2 | exam grade over 20 |
| exams3 | exam grade over 20 |
| works1 | work grade over 20 |
| works2 | work grade over 20 |
| works3 | work grade over 20 |
| library.entrances1 | times library enters per course |
| library.entrances2 | times library enters per course |
| library.entrances3 | times library enters per course |
| platform.times1 | times platform is used per course |
| platform.times2 | times platform is used per course |
| platform.times3 | times platform is used per course |
| books.reserved1 | times a book is reserved per course |
| books.reserved2 | times a book is reserved per course |
| books.reserved3 | times a book is reserved per course |
| desertion | 1 Yes 0 No |

Based on the paper from M. F. Á. Icaza, S. G. Silva, and A. T. España where s is stated that “the first year during a student career is crucial to their permanence”[2] and in order to obtain more information, we decided to use 3 semesters only for this research.

# Clustering

Having the data frame as described in section III of this document, we’ve found couple points of inflection as represented in the following graph:

  
Figure 2. This image represents the point of inflection in Group 2 and 4

As represented in Figure 2 and using k-means to find the best possible centers, we evaluated our data using 4 groups with the following findings:

|  |  |
| --- | --- |
| Group | Description |
| 1 | Best profile, more previews changes than others, group with the highest economic level, the rest of the economic levels are balanced, balanced in semester payments, balanced in assistances all the semesters, balanced in exams and works, balanced in library assistance and platform usage, balanced in books reserved. |
| 2 | The profile is balanced, balanced at all economic levels, the one with more payment complications every semester, the worst in assistance every semester, the worst in exams every semester, the worst in works every semester, the worst in library assistance and platform usage, worst in books reserved. |
| 3 | The worst profile, many previews changes, the one with more scholarships, most of them in middle economic group, balanced in exams and works, balanced in library and platform usage, balanced in book reservations. |
| 4 | Balanced profile, the one with less scholarships and previews changes, the group with lowest economic level, the rest of the economic levels are balanced, balanced in payments, the one with better assistance, the best in exams and works, the one with more library entrances, platform usage and books reserved. |

Since there are many dimensions for this model, is hard to represent graphically the centroids in a way that helps us to select the best group or at least differentiate them as you can see the Figure 3:

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generatedA close up of a map

Description automatically generated

Figure 3. This image represents the centroids for each pair of features selected

Here we are doing a bi-dimensional representation of some of the characteristics of the students. The groups are represented with the following colors:

* Group 1: Black
* Group 2: Red
* Group 3: Green
* Group 4: Blue

The centroid of every cluster is represented with a triangle shape and using the same color of its group.

However, if we review the data in the table, we can easily notice that group 2 is the worst in almost every feature.

A close up of text on a white background

Description automatically generated

Table 1. Table that represents K-means model centers for each value.

Based on above analysis we have decided to take group 2 as risk group and set the decisions in the model accordingly.

# Dimensionality Reduction

To facilitate the analysis of the information and in order to maintain only the information than can provide tangible results to this analysis, we’ve decided to do the following:

1. Remove “gender” from the model as the economic level is correlated.
2. Represent the following fields with one-hot codification\*:
   1. “economic level”
3. We took the average to represent the following fields:
   1. “payment per semester”
   2. “assistance per semester”
   3. “exams per semester”
   4. “works per semester”
   5. “library.entrance per semester”
   6. “platform.times per semester”
   7. “books.reserved per semester”
4. Due to a correlation higher than 80%, the following fields were removed:

|  |  |
| --- | --- |
| **Fields removed:** | **Correlation with:** |
| “works1”, “library.entrance.1”, “platform.times.1”, “books.reserved.1” | “exams1” |
| “works2”, “library.entrance.2”, “platform.times.2”, “books.reserved.2” | “exams2” |
| “works3”, “library.entrance.3”, “platform.times.3”, “books.reserved.3” | “exams3” |
| “admission.letras”, “admision.numeros”, “edad.ingreso”, “nota.conducta” | “promedio.preparatoria” |

Leaving the elements to analyze as below:

|  |  |
| --- | --- |
| **Colname** | **Description/Values** |
| promedio.preparatoria | % obtained |
| scholarship | 1 Yes 0 No |
| prev.change | Amount of changes |
| economic.level (4,3,2,1) | Low, Medium-Low, Medium-High, High |
| Payment (1,2,3) | Average per semester |
| Exams (1,2,3) | Average per semester |
| Assistance (1,2,3) | Average per semester |
| desertion | 1 Yes 0 No |

\* About one-hot encoding, using this technique a separate bit of state is used for each state. It is called one-hot because only one bit is “hot” or TRUE at any time. For example, a one-hot encoded FSM with three states would have state encodings of 001, 010, and 100[3].

\* One important note is that the correlation reduction mentioned in the point 4 was applied just for the neural network. For data integration and clustering we didn’t apply the data reduction based on correlation.

# Neural Network

Now let’s train a Neural Network using neuralnet with the following parameters:

1. In the Input Layer we took into account:
   1. promedio.preparatoria
   2. scholarship
   3. prev.change
   4. economic.level (4,3,2,1)
   5. Payment (1,2,3)
   6. Exams (1,2,3)
   7. Assistance (1,2,3)

A close up of a device

Description automatically generated

Figure 3. Neural Network for our model

1. Splitting the data\*[4]:
   1. Training dataset: it refers to the data used to fit the model and train it, for this section we used 700 elements from the dataset.
   2. Validation dataset: it refers to the data to evaluate a given model, normally used to refine or update parameters in order to obtain better results. For this section we used 200 elements from the dataset.
   3. Test dataset: finally this section it refers to just a sample of the dataset to be used as validation for the final fitted model, for this final section we decided to use 100 elements from the dataset.

\*All elements were selected randomly.

1. Hidden Layers:
   1. Based on the parameters received (input layer), we decided to use two hidden layers, one with 32 neurons and another one with 8 (we tried also with 10 neurons in a hidden layer, however the training was extremely slow: 2x times more than with 32 and 8).
2. Threshold for this model is .001 since there’re a lot of input values and we’d like our model to be as accurate as possible.
3. The maximum number of steps is set as
4. Linear output as True to eliminate the activation function for the output layer.
5. And finally this model is represented by a full connected network.

After performing the training, we stimulated the model with the validation dataset to check the efficiency of the model. We saw that 171 out of the 200 samples were evaluated well, that means an efficiency over the validation dataset of 85.5%. Since the validation set is out of the training set, we expect a similar efficiency over the Test dataset that we are going to use for our investigation.

# Genetic Algorithm

In this last step of our analysis, we are implementing an evolutionary algorithm in order to optimize the results and provide the best proposal given the population received.

An evolutionary algorithm (Population-based, Fitness-oriented and Variation-Driven) is dynamic and can evolve over time, in this case, we’re using a Genetic algorithm (a random-based evolutionary algorithm) that is based in Darwin’s theory of evolution, it uses a slow and gradual process to slightly and randomly change the population getting the best solution.

In order to find the best solution we defined a fitness function fed with the following parameters:

|  |  |  |
| --- | --- | --- |
| Item | Cost | Benefit |
| 1 - Dar Beca Estudiantil | 500L | 100L |
| 2 - Vales de Transporte | 100L | 50L |
| 3 - Mentaría | 200L | 70L |
| 4 - Consultoría Psicológica | 400L | 75L |
| 5 - Boleto a evento Integración | 50L | 40L |
| 6 - Asesor Individual | 250L | 80L |
| 7 - Cursos Remediales | 2500L | 80L |
| 8 - Visita a empresa | 50L | 40L |
| 9 - Examen Extemporáneo | 100L | 70L |
| 10 - Platica Motivacional | 50L | 30L |
| 11 - Viaje recreativo | 10L | 25L |

And following these rules to discard the population that won’t work in our solution.

1. If a student is from economic level 1 (the highest one) and the chromosome suggest that either assigns another scholarship or assign transportation vouchers, then there would be a penalty for this chromosome (80 for scholarship and 40 for transportation vouchers) as you can see, we’re not penalizing with the full amount of the benefit so this student can still be eligible for both items, but at a lower probability.
2. If a student is above 13.9 in the third semester exams average value (meaning his grade is high) and the chromosome suggest that either assigns remedial courses or assigns extemporaneous examination, then there would be a penalty for this chromosome (70 for scholarship and 70 for transportation vouchers) as you can see, we’re not penalizing with the full amount of the benefit in remedial courses as this student might be disappointed being a straight A student so he/she can still be eligible for this item, but at a lower probability. We have selected the value 13.9 because the groups 4 and 1 have the grade over such value in the third exams average and we are considering that groups 1 and 4 are out of the scope.
3. If the student already has a scholarship and the chromosome suggest to assign another scholarship, then the penalty for this population is the worst one.

This fitness function follows the parameters listed below:

1. Population size: 100
2. Generations: 1000
3. Mutation Chance: 0.01
4. Elitism: 20

We selected 1000 generations because we saw that at least 600 generations were required to make the genetic algorithm converge to stable benefits. We can see this behavior in the Figure 4.

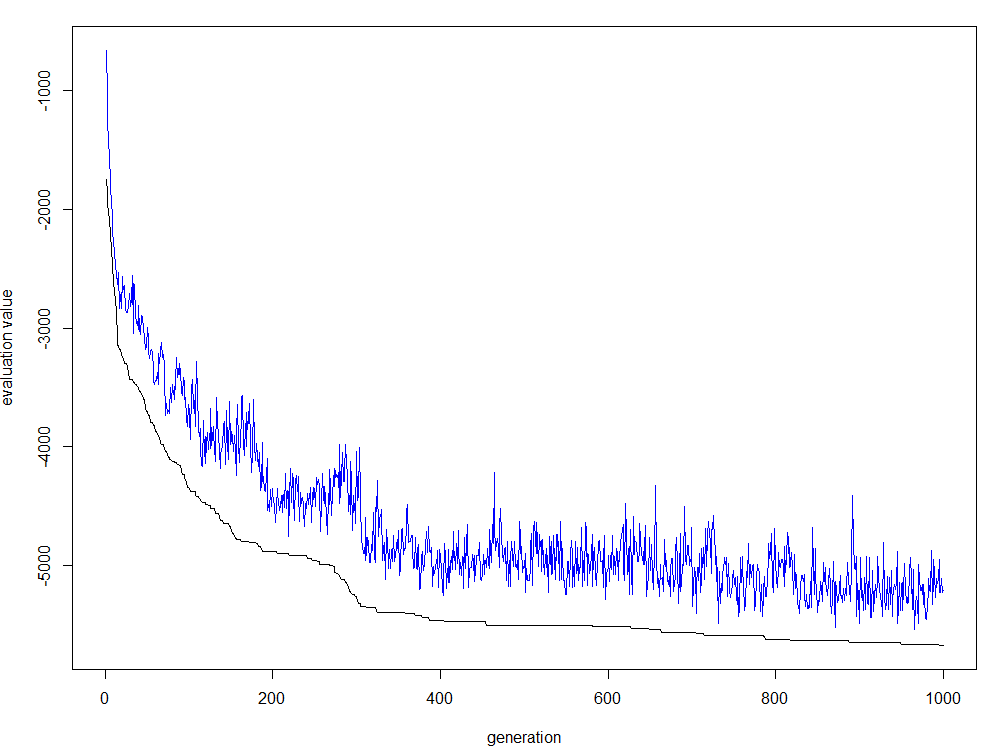


Figure 4. Genetic Algorithm benefit vs generations graph

Final results:

From the 100 students that were validated using our model, 22 are in desertion risk, our model is recommending the following actions in order to help them to continue studying or providing the needed tools to reduce the possibility of quit:

|  |  |  |  |
| --- | --- | --- | --- |
| Student | Recommendation | Student | Recommendation |
| 103 | 2 - Vales de Transporte  5 - Boleto a evento Integración  8 - Visita a empresa  10 - Platica Motivacional  11 - Viaje recreativo | 561 | 2 - Vales de Transporte  5 - Boleto a evento Integración  6 - Asesor Individual  8 - Visita a empresa  9 - Examen Extemporáneo  11 - Viaje recreativo |
| 974 | 2 - Vales de Transporte  5 - Boleto a evento Integración  8 - Visita a empresa  9 - Examen Extemporáneo  10 - Platica Motivacional  11 - Viaje recreativo | 123 | 2 - Vales de Transporte  3 - Mentoría  5 - Boleto a evento Integración  8 - Visita a empresa  9 - Examen Extemporaneo  10 - Platica Motivacional  11 - Viaje recreativo |
| 382 | 5 - Boleto a evento Integración  8 - Visita a empresa  9 - Examen Extemporáneo  11 - Viaje recreativo | 608 | 3 - Mentoría  5 - Boleto a evento Integración  8 - Visita a empresa  9 - Examen Extemporáneo  10 - Platica Motivacional  11 - Viaje recreativo |
| 952 | 5 - Boleto a evento Integración  8 - Visita a empresa  9 - Examen Extemporáneo  11 - Viaje recreativo | 854 | 5 - Boleto a evento Integración  8 - Visita a empresa  11 - Viaje recreativo |
| 379 | 2 - Vales de Transporte  5 - Boleto a evento Integración  6 - Asesor Individual  8 - Visita a empresa  9 - Examen Extemporáneo  10 - Plática Motivacional  11 - Viaje recreativo | 569 | 3 - Mentoría  5 - Boleto a evento Integración  6 - Asesor Individual  8 - Visita a empresa  9 - Examen Extemporáneo  11 - Viaje recreativo |
| 343 | 2 - Vales de Transporte  3 - Mentoría  5 - Boleto a evento Integración  8 - Visita a empresa  9 - Examen Extemporáneo  11 - Viaje recreativo | 665 | 5 - Boleto a evento Integración  8 - Visita a empresa  9 - Examen Extemporáneo  11 - Viaje recreativo |
| 740 | 5 - Boleto a evento Integración  6 - Asesor Individual  8 - Visita a empresa  9 - Examen Extemporáneo  11 - Viaje recreativo | 434 | 2 - Vales de Transporte  3 - Mentoría  5 - Boleto a evento Integración  9 - Examen Extemporáneo  11 - Viaje recreativo |
| 840 | 2 - Vales de Transporte  4 - Consultoría Psicológica  5 - Boleto a evento Integración  6 - Asesor Individual  8 - Visita a empresa  9 - Examen Extemporáneo  10 - Plática Motivacional  11 - Viaje recreativo | 757 | 2 - Vales de Transporte  5 - Boleto a evento Integracion  8 - Visita a empresa  9 - Examen Extemporaneo  11 - Viaje recreativo |
| 258 | 2 - Vales de Transporte  5 - Boleto a evento Integración  8 - Visita a empresa  9 - Examen Extemporáneo  10 - Plática Motivacional  11 - Viaje recreativo | 508 | 3 - Mentoría  5 - Boleto a evento Integración  6 - Asesor Individual  8 - Visita a empresa  9 - Examen Extemporáneo  10 - Platica Motivacional  11 - Viaje recreativo |
| 696 | 2 - Vales de Transporte  3 - Mentoría  5 - Boleto a evento Integración  8 - Visita a empresa  10 - Platica Motivacional  11 - Viaje recreativo | 169 | 2 - Vales de Transporte  3 - Mentoría  5 - Boleto a evento Integración  8 - Visita a empresa  9 - Examen Extemporáneo  11 - Viaje recreativo |
| 946 | 5 - Boleto a evento Integración  8 - Visita a empresa  10 - Plática Motivacional  11 - Viaje recreativo | 986 | 1 - Dar Beca Estudiantil  5 - Boleto a evento Integración  8 - Visita a empresa  9 - Examen Extemporáneo  10 - Platica Motivacional  11 - Viaje recreativo |

With this result, the total cost to apply all recommendations for each student is $9,920.00 USD that is within budget ($10,000.00 USD) and there is a chance to save the students from desertion.

# Conclusions

Jehovany’s:

To be a data scientist is not an easy work, even when you get a result that might fit your expectations, is only one solution. The data science practice as we described previously is the usage of the best judgment in terms of tools to be used, analysis, exploration, test-failure process and other combinations of tools and or techniques that might help you get better results every time. The process should not end with the first result as this might lead you to a different analysis but also faces you into a new challenge, always getting more and more data to be analyzed and complicating things a little more.

In this project we handled only a few information from 1000 students, however in a more robust project there are external factors to be considered in order to develop a better model and that can be cause of desertion and not only “internal” factors like grades, assistances, we might face in the future environmental factors that can completely change this model and provide different solutions but also in the line of the ethics, privacy and other blockers to continue and that might be taken in consideration for data scientist before proceeding with the analysis.

Carlos’:

* R language is a specialized language to process information in a rapidly way.
* R language has many libraries that contains APIs to manage the Data analysis.
* The first step to start the data analysis is obtain clean and gather the data if possible in a database or dataset.
* Then the information has to be processed to have meaning full compressed data, in our cases we took the average value of many cases and one hot representation for economic level.
* Making a clustering using K-means help the data analyzer to separate the data in groups of analysis.
* Make a representation of the K-means centroids when having more than two dimensions is complicated, but we can select two features an make a graph with them.
* One popular activation function of the neural networks is the sigmoid function that is why it is practical to use normalized values when training.
* Genetic algorithms are very useful to select a good set of options according to its importance and cost.
* The fitness function of the genetic algorithm

# References

[1] S. B. Online, “The data science Venn diagram - Principles of Data Science.” https://learning.oreilly.com/library/view/principles-of-data/9781785887918/ch01s02.html (accessed May 07, 2020).

[2] M. F. Á. Icaza, S. G. Silva, and A. T. España, “DESERCIÓN EN EL PRIMER AÑO DE LICENCIATURA: RESULTADOS PRELIMINARES,” p. 12.

[3] “One-Hot Encoding - an overview | ScienceDirect Topics.” https://www.sciencedirect.com/topics/computer-science/one-hot-encoding (accessed May 07, 2020).

[4] T. Shah, “About Train, Validation and Test Sets in Machine Learning,” *Medium*, Dec. 10, 2017. https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7 (accessed May 07, 2020).