National College of Ireland

Gender bias analysis in Machine Learning prediction for the secondary school exit exam in Colombia

MSc Research Project
Data Analytics

Karen Samantha Pinzon Velandia
Student ID: x22144137

School of Computing
National College of Ireland

Supervisor: Vladimir Milosavljevic

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Karen Samantha Pinzon Velandia
Student ID:	x22144137
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Vladimir Milosavljevic
Submission Due Date:	16/09/2024
Project Title:	Gender bias analysis in Machine Learning prediction for the secondary school exit exam in Colombia
Word Count:	6464
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	14th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Gender bias analysis in Machine Learning prediction for the secondary school exit exam in Colombia

Karen Samantha Pinzon Velandia
x22144137

Abstract

This research looks at the potential gender bias by analysing the performance prediction using machine learning models applied to the high school exit exam in Colombia focused on the Boyaca region. Using the exam data for the year 2022, the study compares the accuracy of Random Forest Classifier and Logistic Regression Tree for different gender-based experiments and concludes that there is no significant difference in the accuracy between training the models with female and male data, or training with only one of the genders. The significance was determined by a one-way ANOVA with a 0.05 tolerance. The average accuracy for both models was 66% showing no difference between the models applied. However, the models show a better accuracy when predicting the true negative values.

1 Introduction

The Colombian Institute for the Evaluation of Education (ICFES for their Spanish acronym) with over 50 years of experience, conducts the creation and administration of the education level exams in Colombia. The exam with the highest affluence is the exit exam from high school which is a requirement in all schools to be able to graduate.

Also, one of the recent issues mentioned by UNESCO (United Nations Educational, Scientific and Cultural Organization) is the ethical dilemmas that Artificial Intelligence represents in terms of gender bias [1]. Being machine learning the starting point of Artificial Intelligence, and with the importance of education in the progress of society, analysing educational information such as the results of the high school exit exam as it is the Saber 11 exam, adding with a cornerstone in the gender, allows to uncover patterns in the dataset while studying methods to prevent the inclusion of bias in education analysis.
With that context in mind, this research pretends to answer the question: Which machine learning models would provide less gender bias when used to predict and analyse the Colombian educational dataset for Saber 11 results for the Boyaca region?

To achieve this, a feature selector and a machine learning model suggester are leveraged, along with the Random Forest Classifier and Logistic Regression Tree models which are tested against different data combinations based on gender. The study takes all the demographic information compiled for each student which captures aspects such as age,

[1] reference:`https://www.unesco.org/en/artificial-intelligence/recommendation-ethics/cases`

family level of education, home commodities and use them to predicts the final score of the exam to determine if the student will perform above or below the median.

The structure of the next sections in the document is as follows: Section 2. Literature review: Provides some context about Colombia along with the selected department, the importance of the Saber 11 exam and its structure, the educational exploration with machine learning models and the gender bias in new technologies. Section 3. Methodology: Covers the process carried out with the data explicitly from the beginning to the obtained outcomes. It also gives an overview on the data characteristics. Section 4. Design and Implementation: Describes what was implemented and the evaluation metrics. The last 2 sections will show the results and discuss them in base of the research question.

2 Literature Review

The economy of a country is tied to its level of education measured by its capacity to grow through the knowledge and implementation of new technologies, which ultimately leads to an economical growth Hanushek and Woessmann (2021). Although the intention of this research is not to impact the economy of Colombia, it pretends with the inclusion of new technologies to critically analyse some of the aspects within the education perspective that align with the current society goals to ensure a sustainable and development-oriented world Nations (2015).
With the use of machine learning, this study proposes to consider three of the UN Sustainability Goals: the gender equality, quality education and reduced inequalities, all of this by analysing the gender bias in the prediction of the performance in the secondary school exit exam in one of the middle-sized departments in Colombia which are normally not a great focus for this type of research. To fully understand the importance of this study, it is necessary to introduce the information used as a basis for this analysis. With that in mind, the next section provides more context on the dataset

2.1 Saber 11 exam

In many countries the education level acquired during the high school years is assessed through an exit exam, which also measures the student's readiness to face a higher education. It also can be used as a reference to a vocational look into the future of the assessed depending on the scores obtained and the areas that they stand out in. In some countries, getting a good score is even necessary to graduate from high school or can be used as a starting point to get a scholarship in a higher education institution.
In terms of the context of the country where this study is being presented, the equivalent to the evaluated exam is the Leaving Certificate, while in the analysed context, the exam is known as the Saber 11 test. This name is given since the last grade in high school in Colombia for most schools is 11th grade. Although other assessments are run throughout the high school years, the Saber 11 is the one that allows the institutions to know how prepared the students are for a higher level of education which is the main goal of the schools.

During the six grades of high school in Colombia several subjects are covered, one of which is a secondary language, to include the five main components of the exam: mathematics, natural sciences, social and citizen sciences, critical reading and English. The

maximum score for each of these sections is 100, as shown in the Table 1 [2]
Some additional questions that are not included in the table above, are the socioeco-

Table 1: Saber 11 - five components distribution

Component	Questions	Max score
Critical reading	41	100
Mathematics	50	100
social and citizen sciences	50	100
Natural science	58	100
English	55	100
Global	254	500

nomic characteristics of the person presenting the test. There's a total of 24 questions that allow the government to get the demographics of the people presenting the test and draw conclusions from that data too.
The ICFES indicates that the test has been applied for over 50 years in Colombia, however, it wasn't until 2016 that they started generating statistical information for its analysis which led to them being invited to join the ECTeI (Education, Science, Technology and innovation, for their Spanish Acronym) in 2020 with the purpose of providing timely information to strengthen public policy in the technology sector ECTel: Education and innovation (2023) based on the information gathered and allowing the analysis of the same.
ICFES also discusses in one of their main pages that in order to align with the public data schema suggested by the Colombian government, the institution decided to present and collect the results and the additional information in a different way from 2021 to make sure it's more readable and adaptable in terms of the analysis of the same. In the light of complying with the access to public data imposed, the institution shares their data in the Colombian open data site. However, this data is posted after the internal teams have fully analysed it and anonymised it as necessary. This publication changes meant that the current data available for analysis is the data for the year 2022. Barragán and Marcelo (2023) In their paper the authors utilized the data available from 2012 to 2022 for the capital of the country to analyse the impact of the public policies in terms of education in the test results itself. Nonetheless, the results were that despite the implementation new government plans with education policies, the results were still average even in Bogota that being the capital, a clear difference was expected.
The data used in this research given its easy access and reliability as it comes from the institution directly is the data for Boyaca, Colombia from 2022.

One of the gaps this study closes is that after the change in reporting in 2021, no additional studies have been done to analyse the results of this exam with a gender perspective in mind. Abadía and Bernal (2017) presented a study in which a professor in economy applied a statistical analysis to the results from back to 2014. Those results were selected as they showed a higher trend in the global score compared to the past 10 years, concluding that men performed better than women in the high school exit exam. Their results covered different regions and concluded that there is a difference between regions

[2]Created based on: `https://www.icfes.gov.co/documents/39286/14390199/02+Febrero_Infografa+Generalidades+Saber+11.+2023-1.pd`

as it wasn't in every region that men outperformed women. Abadía et al. (2020), a more recent paper analysed the how the result of the high school exit exam impacted the women in higher education in STEM programs. They found that there was a gap in math and critical reading scores that amplifies after college, where once more it is men who obtain the higher scores, and it also introduces another aspect which is the fact that these results are more notorious on students from public institutions.

With this understanding of the exam, it is time to explore the population that will be analysed in this study.

2.2 Boyaca, a Colombian department

Colombia is a South American country, with 51,6 million inhabitants. It is one of the most bio-diverse countries in the world leading with its flora and fauna. Colombia consists of 32 departments, 1123 municipalities and 5 districts. This means that when compared to a European country, Colombia can be considered as a big country. Compared to Ireland, as shown in Figure 1, Colombia is 16 times bigger than Ireland. [3]

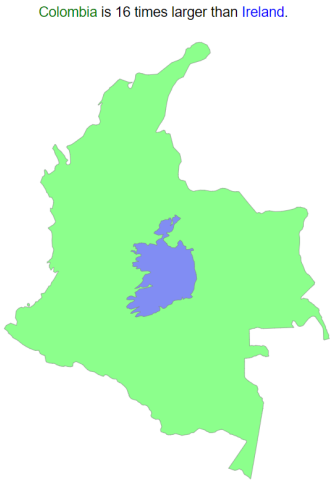


Figure 1: Colombian size vs Ireland size

In 2022, the Global Green Growth Institute (GGGI) [4] reported that Colombia has the lowest average salary in the world, this was determined based on the PIB for that year which didn't increase significantly for 2023 either. The institute also Colombia as one of the countries with the highest unemployment rates. These aspects indicate that compared to other countries in the GGGI list, Colombia population are considered to have a low level of wealth.
It comes as no surprise when studies to analyse the Saber 11 results focus on aspects such as the impact that access to internet has in the scores obtained by the participants Barrios Aguirre et al. (2021) where a pooled two-stage least squares model was applied, or the relation between owning a gaming console and the gender gap in STEM Bustamante-Barreto et al. (2024). The mentioned studies concluded that having access to internet, a computer and even to video games provided a positive impact in the results of the test.

[3] Taken from: `https://www.comparea.org/COL+IRL`
[4] reference:`https://datosmacro.expansion.com/paises/colombia`

The second study also showed how the gender gap is less between the participants with a gaming console.

Although the mentioned studies focused on the whole country, Colombia is divided in different regions for which it is not uncommon for researchers to select and analyse one region at a time, as is the case in this current study. To understand what is behind this approach, it is necessary to know more about the distribution of the country. Figure 2 shows the 6 regions of Colombia.

Figure 2: Colombian regions

On some regional studies Cárcamo Vergara and Mola Ávila (2019) focused on finding the difference in the performance for men and women for two of the sections of the exam, separating them per region, authors concluded that men perform generally better in every region of the country, except for the Caribbean region were women performed better. Also in the Caribbean region and comparing the results obtained from 2017 to 2019, another research found that the performance of the region was under the median of the national scores Solano et al. (2022).
With the idea to maintain a regional approach, this researcher decided to focus on a much less notorious department in the Andean region: Boyaca. Looking at Figure 2 closely, the Andean region shows two of the most known cities of the country, its capital Bogota, and Medellin, a city that very often is seen in movies and tv shows as it holds a big part of the Colombian past. However, having such big cities in the region, does not create space for other departments to stand out and be a main focus on a study. However, it doesn't mean that this part of the country is ignored.

According to the Colombia travel site [5] the Andean region has different varieties of landscapes, and in the case of Boyaca, they are known for their interest industrial interest and their land work, as most of the Boyaca region is agricultural. This in turn means

[5]refer to: `https://colombia.travel/en/blog/travel-magazine/country-regions`

that most of the areas are a mix between the urban and rural.
To show how prominent the agriculture is in the area, there are investigations where deep learning has been applied to potato crop irrigation with the main goal of improving water usage in the studied crop. The authors obtained 0,067 and lower for MSE and 0,258 or less in terms of RMSE with the CNN-LSTM modelJiménez-López et al. (2021). Other founded studies from Boyaca universities worked on creating a IoT system for monitoring urban crops with the use of deep learning Tovar-Soto et al. (2022).
Other studies acknowledge the reduction in farming workforce in the region, which causes food production insufficiency, while suggesting technological solutions for newer generations to maintain the industry by learning from these approaches and utilize them as a starting point to improve production and avoid any more reduction Nuncira et al. (2023).

Bringing surging technologies to areas that require an improvement is the core of research and development. The next section refers to how machine learning can be used to improve education.

2.3 Machine Learning in Education

Predictive analysis is widely used in educational datasets. The implementation of machine learning to reveal trends in the data and the possibility to predict performance has increased the use of this technology in education. Having a view to the past and the future at the same time, permits stakeholders such as universities to be informed and to help refine their educational level.
Pallathadka, et al. in their research about classification of students based on their performance prediction found that after applying models such SVM, Nave Bayes, ID3 and C4.5 and comparing them, concluded that SVM provided the best accuracy, around 90% indicating that when classifying the students in two categories between fail or pass, the model would more accurately predict the number of passing students, allowing the teachers to focus on the failing students while helping the passing students to stand out and focus on better opportunities Pallathadka et al. (2023).

Other studies find it useful to use machine learning to analyse the performance of students in big countries as is the case of a student performance analysis in India, where the support vector machine (SVM) model gave them an accuracy of 84.3% Pande (2023). The authors also used a dataset that included similar aspects to the ones considered in this study, such as mother education, job and age of the student.
In terms of dealing with large datasets which is normally the case with educational information, Ghanbari, et al. using an engineering dataset and supervised machine learning algorithms such as a variation of the logistic regression named multinomial logistic regression as their target was not a binary classification problem. Within their exploration, the authors included a graph comparing the results for both men and women where for fail and pass men received a better score than women with a gap of about 2.5 points average Poudyal et al. (2020). Nonetheless, their accuracy for the decision tree and KNN was of 99% which may be an indication of over fitting.
All the previous studies coincide that applying machine learning to educational data help in making decisions based on the prediction of the performance of students. Khan et al. (2022), agree with this opinion after applying different machine learning models to a secondary school dataset. Their best performance model was random forest with an

accuracy of 93.7%.

Although the accuracy of the studies mentioned was high, and they analyse general differences between men and women, there's no accuracy comparison or in-depth analysis of the gender gap or if there is any sort of gender bias. Knowing how gender bias may affect the latest technologies is crucial to Understanding the niche of this research.

2.4 New technologies and gender bias

Human rights and equality organizations such as UNESCO and UNWOMEN present their concerns about the potential inequality that technologies such as Artificial Intelligence cause with its wide use [6], and how this behaviour is a reflection of the current society [7], given that AI is created by humans and based on human data as well [8]. These affirmations are a call to action from researchers and media to not only study the current impact of the bias, but to find solutions to the problem.

Smith and Rustagi address their article to developers and world leaders on an analysis of what may be causing the gender bias and how to fix it from both perspectives. From social indicators such as or the lesser number of women that are in the data industry, and the access to internet for women. For the developers, the suggestion is to pay a closer attention to the feature selection and even to separate medical data in terms of women and men and this may create an inadequate representation in the results. The main aspect that they mention is how developers need to recognize that the algorithms are not neutral which in turn requires of consciously prevent them from creating any bias. Smith and Rustagi (2021).

Most studies focus their efforts on artificial intelligence given its sudden rise. Ferrara (2024) explains how the use of AI can now be seen in healthcare, credit scoring, employment and even in generated data and addresses the aspects that introduce bias in AI dividing them into user, data and algorithmic bias and including ways to reduce them. In terms of data, the quality of the data as is the accuracy and groups representation are some the authors mention, while a mix between the algorithm bias and user bias is explained saying that they can happen when the criteria and assumptions inserted in the machine learning algorithm can be biased by the creator.

With the importance allocated to the data used in quality and in training processes, one of the research projects that was used as a main reference was "AI Gender Bias, Disparities, and Fairness: Does Training Data Matter?" in which the authors studied the possible introduction of a gender bias to an automatic written scoring for students. The study didn't show any significant differences between the male, female and mixed data trained models Latif et al. (2023).

Having the context exposed in the previous sections of this document, the main limitation found in previous studies is that none of the educational research took an approach to

[6] https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressi

[7] https://www.unwomen.org/en/news-stories/explainer/2024/05/artificial-intelligence-and-gender-equality

[8] https://www.internationalwomensday.com/Missions/14458/Gender-and-AI-Addressing-bias-in-artificia

identify the bias in the training based on gender. As the aim is to identify if the machine learning models can introduce any gender bias, and assuming that the data pre-processing is done correctly, this research pretends to overcome the mentioned constraint by exploring the results of different machine learning algorithm that are trained with mixed data, and then compare it in training the model with male and female data separately. The models to be compared are the Lazy predict (LP), Random Forest Classifier (RFC) and The Logistic Regression Tree (LRT). This approach will simultaneously helps to answer the research question focused on the accuracy of the models as the more accurate model is the one with less bias.
The methodology section will show the steps carried out in the data to and the aspects that were considered during the development of the project.

3 Methodology

This section outlines the steps carried out with the data and the reasoning behind the decisions made during the project progression.
The base methodology for this project is as described in Figure 3. In the next subsections, the details of each stage will be shared, and the final system designed is included in the implementation section 4.

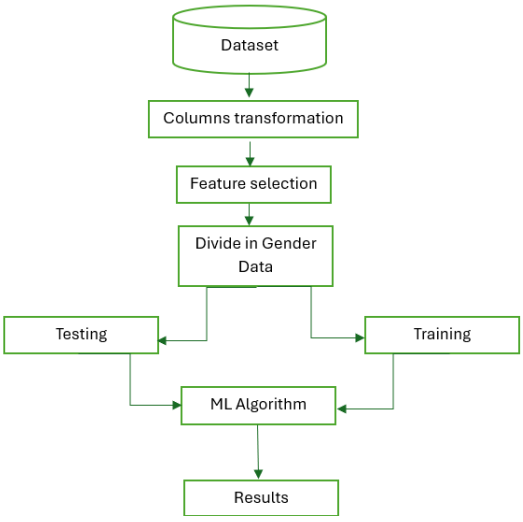


Figure 3: Methodology

3.1 Dataset

The data was collected from the Colombian open data site from which a CSV file is obtained. The link to this dataset can be found in the configuration manual attached to this document. The full dataset contains 7.1 M rows and 51 columns.
The data was filtered based on the department column to obtain only the data from Boyaca. This returned a working dataset with 25K rows and 42 columns.
The data dictionary for the final dataset can be found in the configuration manual attached to this document.

3.2 Data Exploration

One of the most important parts of a project is to know the data that the researcher is working with. With that in mind, once the data is loaded, graphs can be created from them.
The dataset has 3 main sets of features: student related, family or home related, and school related. Based on that, random features within those categories are selected and plotted for reference as how the data looks like. The Figure 4 shows the age and gender as part of the student category. From the age group, the ages 16 and 17 have the greatest distribution with 51.2% and 38.9% respectively.

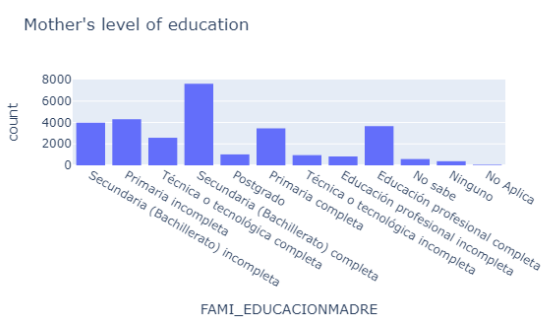
(a) Student's Age (b) Gender distribution

Figure 4: Student demographics

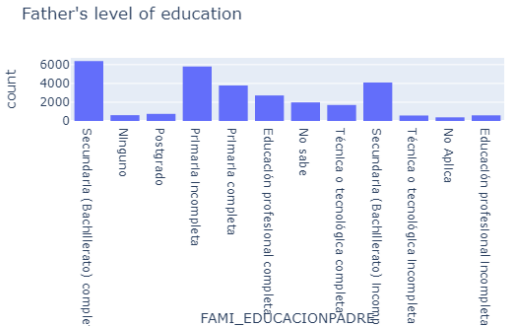
In terms of the family and home related, some of the main factors to look at are in Figure 5. The image shows that for the level of education of the father, the most common answers are basic education complete, and high school complete and incomplete. For the mother's level of education the higher count belong to High school complete and incomplete, and incomplete basic education. On top of that, the families that own a car are half of the ones that don't. These aspects disclose the level of education and wealth in Boyaca's families.
The next set of features in Figure 6 show some home aspects such as the number of rooms and people in the house, and if there is a washing machine at the student's home. As it can be seen, the highest number of families are conformed by 4 people, and the more frequent number of rooms is 3. Also, more than double of the families own a washing machine. Another important aspect is the home stratification, where the most homes are stratus 2 or 1 which are the lowest on a scale that in Colombia goes from 1 to 7.
The last set of features this researcher wanted to include in this document, are the school aspects to consider.Figure 7 shows that in Boyaca, the greatest number of schools are Urban but there is a significant amount of rural schools as well. The type of school graph shows the type of degree that the students receive at the end of their high school years. The bilingual schools in the area are also low compared to the ones with only a Spanish emphasis.

3.3 Data transformation

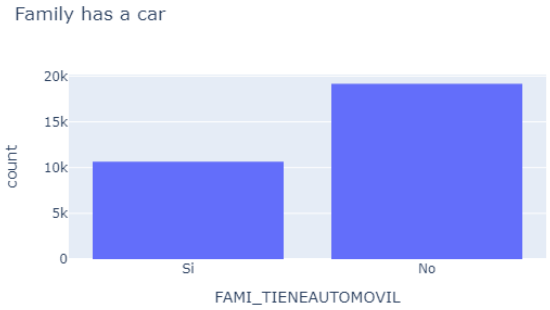
For this study, the factors captured in the socio-economic questionnaire are the ones used to predict the outcome of the classification into med-high or med-low. However, since the



(a) Mother level of education



(b) Father level of education



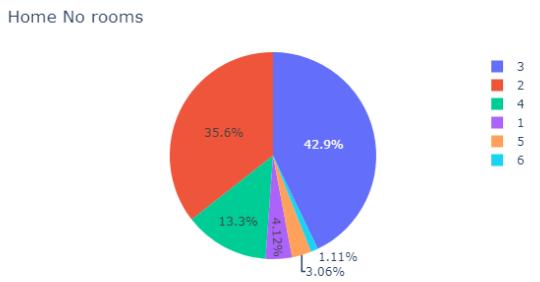
(c) Family owns a car

Figure 5: Family features

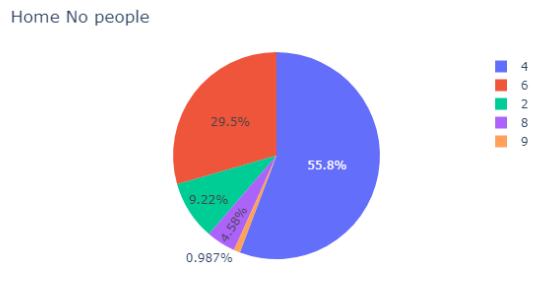
features captured through the exam and during the socioeconomic questionnaire cause for the dataset to have a total of 51 columns before any manipulation, a feature selection is necessary. The transformation process was divided in:

- Identifiable information: The first step was to drop all the columns that contained identifiable information.
- Missing values: all the observations with missing values were removed from the dataset as there was enough data for the analysis and prediction.
- Outliers: All of them were kept as they represent a normal variation in the population.
- Categorical values: for this, the columns were assigned a string data type and then converted to dummies. The details of this can be found in the configuration manual.
- Other transformations: These transformations are discussed below.
- Duplicated information: after the columns transformations, the duplicated columns were dropped.

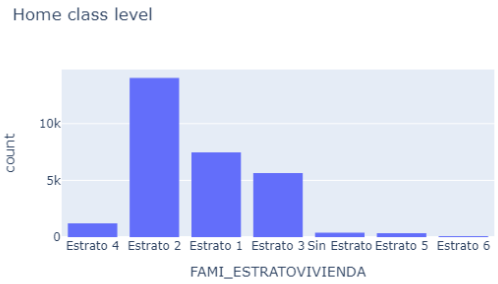
For the process to be seamless, some of the columns require a transformation. One of them was the date of birth which was used to create an age column. However, there were ages of less than 14 years which is very unlikely for this exam, for which the most repeated values were taken into account to create an average and the values under 14 were approximated to the average value which was 18. All other ages were kept as part



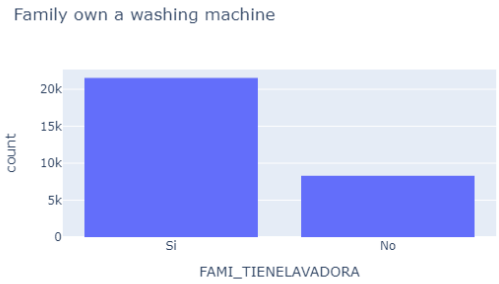
(a) No of rooms



(b) No people



(c) Home stratification



(d) Washing machine

Figure 6: Home features

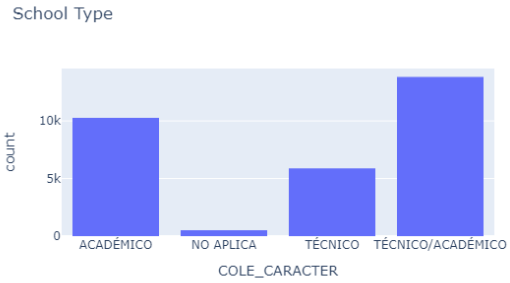
of population representation which lead to the column ranging from 16 to 56 years old. Another column that requires a transformation is the people at home. For this column the options weren't numerical but more of a range. When transforming, the lower value of the range was selected, and the selection was changed by a numerical value. So, if a student had selected that their home was conformed by 3 to 4 people, it would reflect the number 3.
The stratification column was also transformed from string to a number, so instead of "stratus 1", it would just be the numerical value "1". This same transformation is applied to the number of rooms which were in a string writing in the original dataset.

One of the decisions made with the dataset was to create a target column that was the focus of the prediction done in this analysis. The column was created by taking the global score and classified it in medium-high or medium-low since each of the sections in the exam have classification levels. The performance levels are shown below in Table 2, 1 being the lowest and 4 the highest band [9]:
Then, ICFES in its informative material explains that the global score is calculated with a weighted average formula as shown below, where GS is the global score:

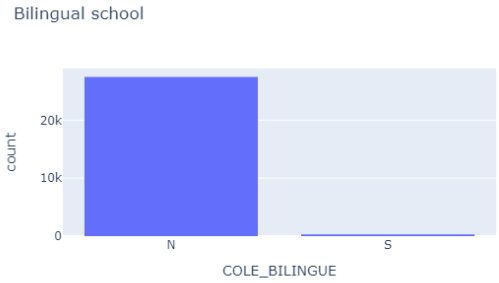
$$GS = \left(\frac{CR \times 3 + MAT \times 3 + SC \times 3 + NS \times 3 + En \times 1}{13} \right) \times 5 \qquad (1)$$

Being 3 and 1 the weights for each section, while 13 is the sum of the weights to have an average, which then is multiplied by the number of components. Then, to calculate what the score for high and low was going to be, it was decided to take the lowest score in the

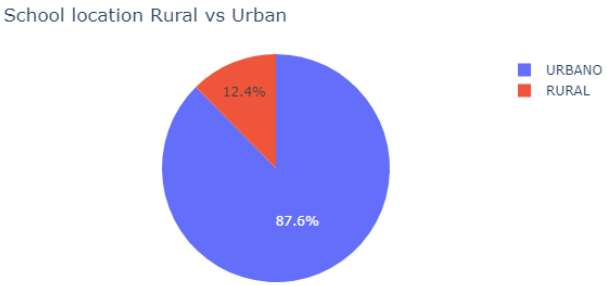
[9]Created based on: `https://www.icfes.gov.co/acerca-del-examen-saber-11f`



(a) school type



(b) Bilingual schools



(c) Urban vs Rural schools

Figure 7: Family features

Table 2: Saber 11 - Point distribution

Component	Level 1	Level 2	Level 3	Level 4
Critical reading	0-35	36-50	51-65	66-100
Mathematics	0-35	36-50	51-70	71-100
Social and citizen sciences	0-40	41-55	56-70	71-100
Natural science	0-40	41-55	56-70	71-100
English	0-36	37-57	58-70	71-100

level 3 band and calculate everything with it, which meant for example that for critical reading the band was 51-65, being the lowest score 51 points and thus that was selected for the formula. The end formula was:

$$MH = \left(\frac{51 \times 3 + 51 \times 3 + 56 \times 3 + 56 \times 3 + 58 \times 1}{13} \right) \times 5 = 269.2 \tag{2}$$

Meaning that all values from 269 and above are classified as a med-high score, while 268 and below were classified as low score. Taking this threshold into account, the global score column was transformed into a binary High-low score column.
After the transformation is completed, the duplicated columns are deleted. This includes the score obtained in each component of the exam as this study is focused only on the global score. The next step is to look for the most relevant features for the prediction.

3.4 Feature reduction

The study by Li, et al. reveals that applying the same models to a small dataset before and after doing feature reduction show opposite results where before the reduction SVM performed better than Linear Regression, while after the reduction, it was linear regression the one with better accuracy, although both cut short against random forest which got a MAE of 3, and a RMSE of less than 4 Li et al. (2021).
Rodas-Silva et. al obtained an accuracy of 75.24% in their use of Boruta + Random Forest when predicting the student performance of students with low income. As the idea of the mentioned authors was to analyse the contributing aspects to the academic success of low-income students in online universities in Ecuador, having a feature selector provided that factors such as the grade in the leveling course and the age were in the main focus for the student's success Rodas-Silva and Parraga-Alava (2023). As shown in Figure 6 the number of students in the dataset that belong to lower classes is high for this department.

Nonetheless, given the number of features that the dataset contains and knowing that after transforming the amount is higher due to categorical variable transformation, it was decided to use Boruta as the feature selection algorithm for speed in feature analysis.
With Boruta, the Random Forest Classifier was used, and two tests were run:

- Global score
- High - Low score

This author expected to confirm if having the dataset with the original global score column would produce results that would truly differ from the newly created High-low target column.

3.5 Gender bias Analysis

In a gender bias analysis for an automated scoring system focused on written responses by Zhai et. al, the authors concluded that when the models were trained with one gender data, the scoring showed a gender gap which was called a mean score gap by the researchers Latif et al. (2023).

To ensure no bias is introduced unintentionally in this study, and that accurate results could be compared, it was decided to create 3 sub-datasets based on the gender column in the main Boyaca resulting dataset after the transformation of the columns. The 3 sub-datasets were:

- Both genders: 25,000+ rows
- Female: 13,800+ rows
- Male: 11,400+ rows

After those copies and partitions are created, depending on each machine learning model, the data is split into test and training. 80% for training, 20% for training. For each algorithm, the tests run was:

- Mixed data: both genders included for both training an testing the model.

- Male data: using just the male partition for both training an testing the model.
- Female data: using just the female partition for both training an testing the model.
- Male - female data: using the male data to train the model and the female data to test the model.
- Female - male data: using the female data to train the model and the male data to test the model.
- Male - mixed: using the male data to train the model and both the genders data to test the model.
- Female - mixed: using the female data to train the model and both the genders data to test the model.

The purpose of the model is to predict the target variable which as described before would be if a student gets a high or a low score. In other words, the model would predict a 1 for a high score, and 0 for a low score.
Each model provides an accuracy that is compared between the same data type, and then an average accuracy is calculated which is compared with other data types.

4 Design and Implementation

The Figure 8 shows the system design of this project. It indicates that the project was implemented in a local environment and the stages of the project itself which were covered in section 3 as well.

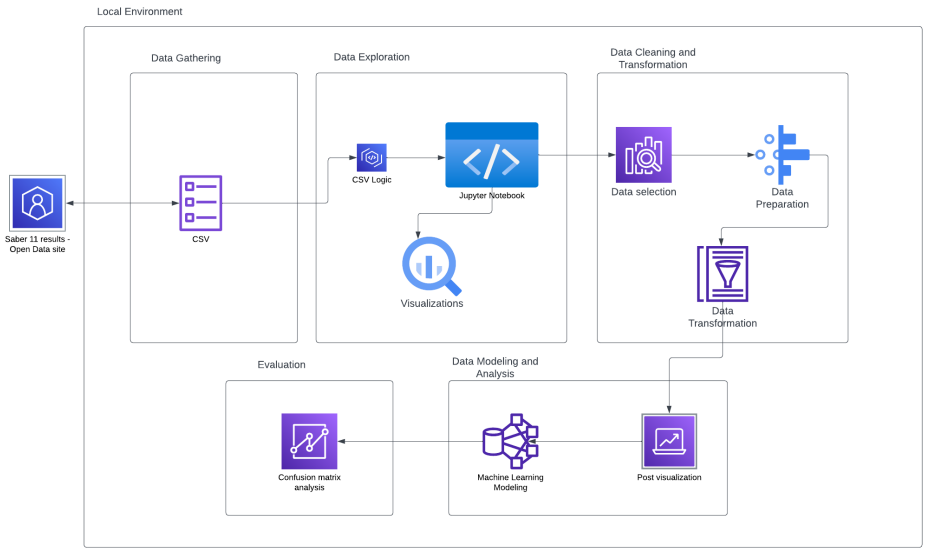


Figure 8: System Design - Methodology

After the data is gathered and loaded into Jupyter, visualized and transformed, the feature selection takes part. As this dataset initially contained 51 columns, with the

transformations applied to the columns, the total number of features were 78. To decide what the best features were, two tests were run, one with each final score column, the original and the transformed one. The results were similar, so the features for the transformed column were taken and the final data dictionary was created for the columns selected.
Boruta doesn't eliminate the features, but suggests the most relevant combination of them, opening up the option for this researcher to make the final decision about the features to keep as part of the analysis. The total selected features were 42 which still indicates a big dataset but with more relevant data for the decision.
Before moving on, a confirmation that the data is not correlated is run. This is done through a heat map between the features. The heat map in Figure 9 shows no strong

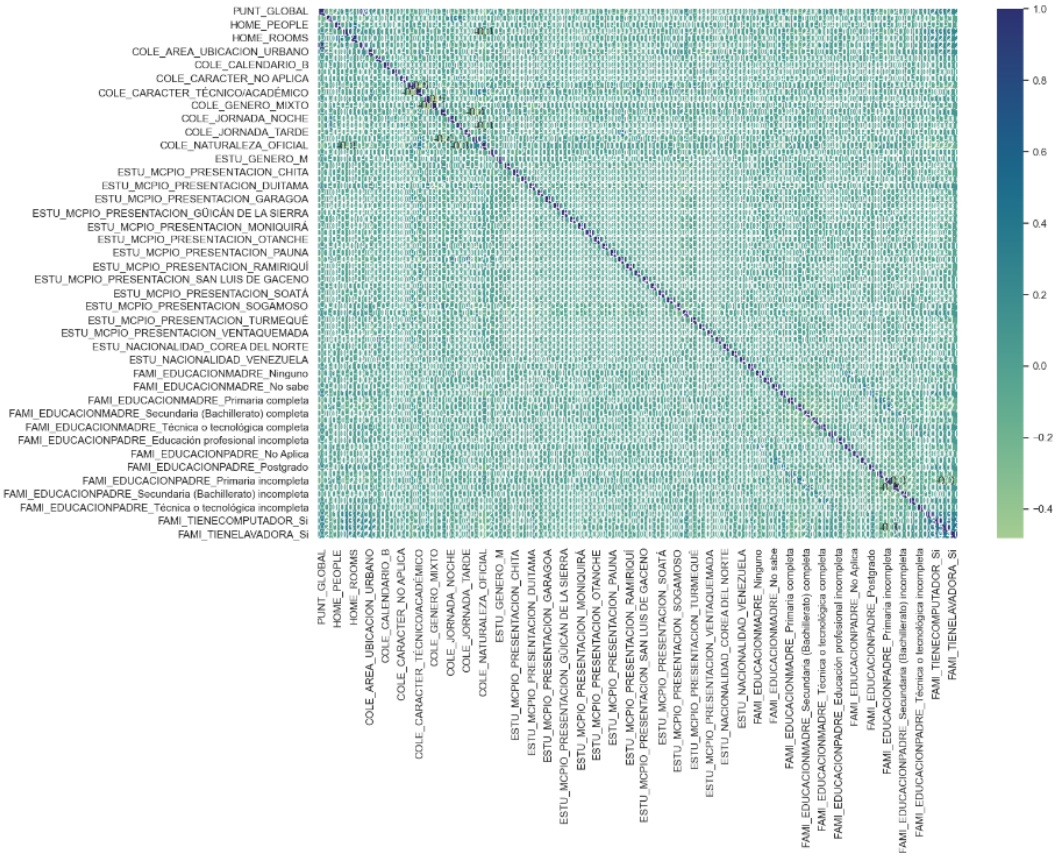


Figure 9: System Design - Methodology

between the features.

The next step is then testing each gender partition through Boruta and see how many features are relevant depending on the type of data. Table 3 shows the results obtained, where ranking with score of 1 indicates the number of most relevant features in the dataset. As seen in the table, with the data that contains both genders, the features are still relevant, while for only male data the relevant features go down to 29 and for only female data 34 features are found to be the most important for the prediction. As the intention in this experiment is only to compare results, no additional features were removed to ensure that the same conditions are kept during the trials and avoid any unintentional bias on the data.

Table 3: Boruta number of features per ranking

Boruta			
Ranking	Mixed data	Male	Female
Score 1	41	29	34
Score 2-5	1	6	5
Score 5-10	0	4	2
Score >10	0	2	0

For the Machine Learning models applied, a model selector was applied to the mixed dataset and with that, the same models were applied to the other partitions and test combinations. Nonetheless, for each combination Boruta was run to check which were the most accurate models for that data. The model selector used was Lazy Predict, which for the mixed dataset showed that the most accurate models were the Random Forest Classifier and the Decision Tree Classifier with an accuracy of 93% and 91% respectively. With that in mind, the two models applied were the RF Classifier, and the Logistic Regression Tree as it is one decision tree. In total 7 tests were performed for Lazy predict, for RF classifier and for LRT.
Although each method produces a measure of accuracy as it compares the predicted values with the actual values, this research uses the confusion matrix to have a visual representation of the results. As shown in Figure 10, this performance measure divides

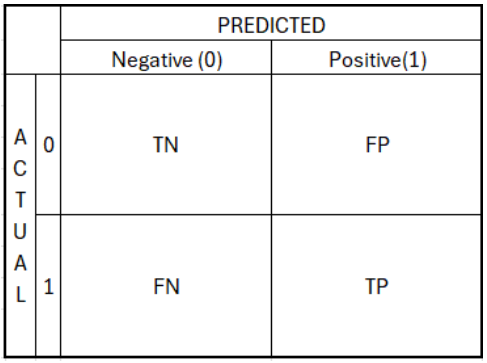


Figure 10: Confusion Matrix

into 4 quadrants.

- True Positive (TP): Number of accurate positive "1" predictions.
- False Positive (FP): Number of predicted positive "1" that were actually a negative "0".
- False Negative (FN): Number of predicted "0" that were actually a positive "1".
- True Negative (TN): Number of accurate "0" predictions.

For this research positive "1" results refer to the student classified as a high score, while a negative "0" result refers to the student classified as a low score.
Last, to confirm if there is a significant difference between the results obtained for each

model on each dataset, an ANOVA test is performed. This test analyses if there are differences between the means of the groups.

$$H_0 = \mu_1 = ...\mu_k \tag{3}$$

$$H_1 \neq (\mu_1 = ... = \mu_k) \tag{4}$$

In the equation above H0 is the null hypothesis which means that all the means are equal, while H1 is the alternative hypothesis that means there is at least one mean that is different than the rest.
In the case of this research, performing a one-way ANOVA would indicate if there is one data type that outperforms the other data type groups or in other words if there is a significant difference between the performances obtained based on the type of data used for training and testing purposes.

5 Results

Taking into account that the purpose of the methodology applied was to identify if the training data had a significant impact when tested against the same type of data, this section shows the results of the main 3 tests: both genders used for test and training which is called mixed data, male data used for training and mixed data used for testing, and female data used for training and mixed data used for testing.

5.1 Mixed Data

Figure 11 shows that for the 5000 predicted values, the Logistic Regression Tree has 1797 TN, and 1690 TP, while also having a lower rate in FN values compared to Random Forest which obtained 1815 TN, 1553 TP and 936 FN, which concluded with LRT having the highest accuracy for this data with 69%. Interestingly, according to Lazy predict, the accuracy of Random Forest was the highest for this data, but it's actual accuracy for RF was considerably lower than expected.

5.2 Male training - mixed testing

In Figure 12 the colours show a clear difference between the Random Forest and the Logistic Regression Tree predictions. For the Random Forest the scale of colours vary more in the 4 quadrants, while for LRT, the colours are a representation of high values for the TP and TN, and low for the FN, and FP quadrants. Nonetheless, in terms of the accuracy, both models obtained the same score of 66%.

5.3 Female training - mixed testing

In Figure 13 both TN quadrants obtained the highest values for both models. For this experiment, LTR was the best performing model with an accuracy of 68% obtaining 1944 TN, 1500 TP, 610 FP and 989 FN.

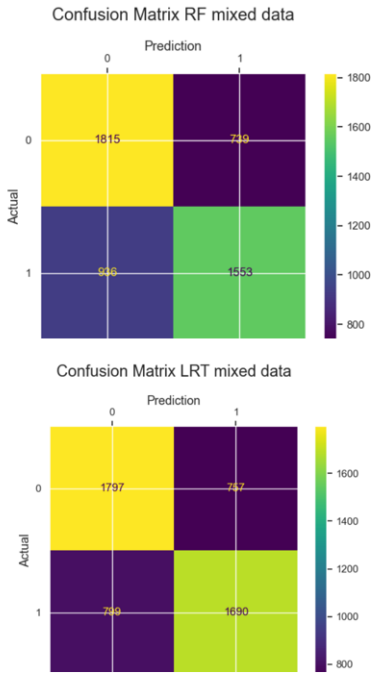


Figure 11: Mixed data-confusion matrix

Model		Accuracy
		Mixed data
Lazy predict	RF	93%
	DTC	91%
RF Classifier		66%
LRT		69%
Average		80%

Table 4: Accuracy mixed data

5.4 Discussion

Based on the percentages shown in table Table 7, The models that Lazy predict applied obtained better results than the unique models itself on every experiment except for when the training of the model was done with one gender and tested on the other gender. Also, for the Lazy predict accuracy, there can be seen values over 80 and 90% while for the models themselves, the values oscillate between the 60 and 70%. This indicates that Lazy predict can be taken as a reference to decide the models, but it's accuracy should not be assumed as completely correct.

Looking at the averages, and comparing them with their same category, when using male data to train the model and other data to test it, the accuracy was lower compared to the female data. This can be explained by the amount of data, as the male data was the smallest sample, although it continues to be a big dataset in general. It is well known that the more data a model has available in the training stage, the better accuracy its predictions will produce.

Based on the confusion matrices, there is also a trend in the True Negative values, where the predictive power of the models was higher for this quadrant, as the color yellow which indicates high value, is more present in this quadrant than in others for each gender test.

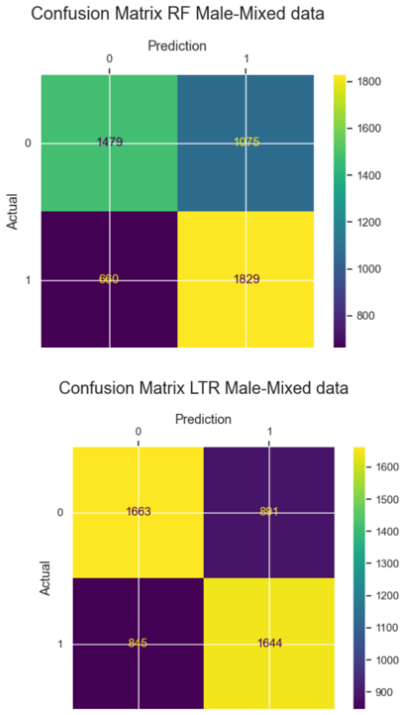


Figure 12: Male - mixed results

Model		Accuracy
		Male-Mixed
Lazy predict	RF	79%
	DTC	75%
RF Classifier		66%
LRT		66%
Average		71%

Table 5: Accuracy male_mixed data

Table 7: Accuracy based on type of data

Model		Type of data						
		Mixed data	Male	Female	Male tr - Female t	Female tr - Male t	Male tr - Mixed t	Female tr - Mixed t
Lazy predict	RF	93%	91%	93%	65%	64%	79%	82%
	DTC	91%	90%	91%	57%	59%	75%	80%
RF Classifier		66%	67%	69%	65%	64%	66%	65%
LRT		69%	65%	70%	62%	63%	66%	68%
Average		80%	78%	81%	62%	63%	71%	74%

To answer to the methodology approach for this research, and analysing if there is a significant difference between the accuracy obtained based on the type of data used for the experiments, Table 8 illustrates that the tolerance error also named as α was defined as 0.05. However, for this research, the p-value obtained was 0.65 which is higher than α indicating that there is not a significant difference between the accuracy based on the data used.

Table 8: ANOVA test

ANOVA: For accuracy data types. α=0.05	
F-statistic	p-value
2.38	0.065

Confirming the first paragraph in this discussion section, Table 9 and Table 10 show how including lazy predict to identify a difference in the models, show very different results when lazy is included between the analyzed groups.

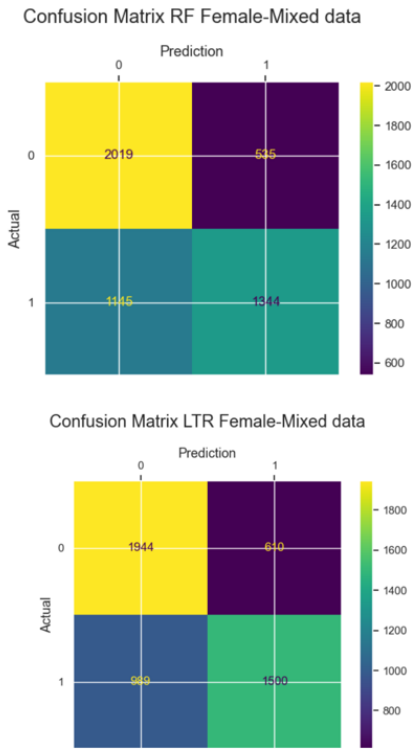


Figure 13: Female - mixed results

Model		Accuracy
		Female - Mixed
Lazy predict	RF	82%
	DTC	80%
RF Classifier		65%
LRT		68%
Average		74%

Table 6: Accuracy female_mixed data

Table 9: ANOVA comparing with Lazy predict

ANOVA with lazy. α=0.05	
F-statistic	p-value
6.2	0.008

Table 10: ANOVA for Machine Learning model

ANOVA without lazy. α=0.05	
F-statistic	p-value
0.01	0.9

Focusing on the accuracy based on the model types were the performance for Random Forest and Logistic Regression Tree are compared including all the 7 tests performed and shown in the main accuracy table, there is no significant difference between the model performance to conclude that either of them introduce a bias during the analysis and prediction of the data.

6 Conclusion and Future Work

As the aim of this research was to identify which machine learning models provide the least gender bias when used to predict and analyse the Colombian educational dataset for Boyaca Saber 11 results, in terms of both the data type used to train and test the model and the machine learning models used to predict the outcome which were RF and LRT, there is no conclusive proof that either approach is introducing a gender bias for

the results obtained in this region.

Additionally, there should be noted that with a bigger amount of data, the models tend to perform better as it was seen when comparing the accuracy of the models between male trained data and female trained data.
With that, as future work it is proposed to use the data of 2022 and 2023 together, once it's publicly available, bto train the model and the 2024 data to test the models, once the information becomes publicly available. Another approach would be to apply the same methodology and models to other regions of the country and compare with the obtained results in this research.

Lastly, as in the experiments of this research each feature is given an equal importance, exploring weighted machine learning models would yield different results.

References

Abadía, L. K. and Bernal, G. (2017). A widening gap? a gender-based analysis of performance on the colombian high school exit examination, *Revista Economía del Rosario* **20**: 5–31.
URL: *https://revistas.urosario.edu.co/index.php/economia/article/view/6144*

Abadía, L. K., Bernal, G. and Gomez Soler, S. C. (2020). Women in stem: does college boost their performance, *Higher Education* **79**: 849–866.
URL: *https://doi.org/10.1007/s10734-019-00441-0*

Barragán, S. and Marcelo, E. (2023). Results of standardized government tests: an educational quality indicator, *Frontiers in Education* **8**.

Barrios Aguirre, F., Forero, D. A., Castellanos Saavedra, M. P. and Mora Malagón, S. Y. (2021). The impact of computer and internet at home on academic results of the saber 11 national exam in colombia, *SAGE Publications* **11**.

Bustamante-Barreto, A., Corredor, J. and Hernandez-Posada, J. D. (2024). The association between owning a videogame console and the gender gap in stem: an instrumental variable approach, *Journal of Computers in Education* **11**: 51–74.

Cárcamo Vergara, C. and Mola Ávila, J. A. (2019). Diferencias por sexo en el desempeño académico en colombia: Un análisis regional, *Economía & Región* **6**(1): 133–169.
URL: *https://revistas.utb.edu.co/economiayregion/article/view/137*

ECTel: Education, Science, T. and innovation (2023). First statistical report.
URL: *https://www.dane.gov.co/index.php/estadisticas-por-tema/educacion/poblacion-escolarizada/educacion-formal?highlight=WyJlY3RlaSJd#eduacion-ciencia-tecnologia-e-innovacion-ectei*

Ferrara, E. (2024). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies, *Sci* **6**(1).
URL: *https://www.mdpi.com/2413-4155/6/1/3*

Hanushek, E. A. and Woessmann, L. (2021). Education and economic growth.
URL: *https://oxfordre.com/economics/view/10.1093/acrefore/9780190625979.001.0001/acrefore-9780190625979-e-651*

Jiménez-López, F.-R., Ruge-Ruge, I.-A. and Jiménez-López, A.-F. (2021). Deep learning techniques applied to predict the irrigation prescription for potato crops in boyacá, *2021 IEEE CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON)*, pp. 1–8.

Khan, M. I., Khan, Z. A., Imran, A., Khan, A. H. and Ahmed, S. (2022). Student performance prediction in secondary school education using machine learning, *2022 8th International Conference on Information Technology Trends (ITT)*, pp. 94–101.

Latif, E., Zhai, X. and Liu, L. (2023). Ai gender bias, disparities, and fairness: Does training data matter?, *arXiv preprint arXiv:2312.10833* .

Li, H., Li, W., Zhang, Z., Yuan, H. and Wan, Y. (2021). Machine learning analysis and inference of student performance and visualization of data results based on a small dataset of student information, *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, pp. 117–122.

Nations, U. (2015). Sustainable development goalsh.
URL: *https://www.undp.org/sustainable-development-goals*

Nuncira, T. A., Rodriguez-Hdez, N., Echeverry, G., Paramo, L., Ramirez, J., Agosto Cintron, L., Andrea Lopez, M., Nuncira, S. A., Javier Fonseca, Y. and Lambertinez, M. E. (2023). Implementation of the steam method, through emerging technologies such as the metaverse, to motivate high school students to investigate and use technology to solve problems in their environment, *2023 IEEE Global Humanitarian Technology Conference (GHTC)*, pp. 393–396.

Pallathadka, H., Wenda, A., Ramirez-Asís, E., Asís-López, M., Flores-Albornoz, J. and Phasinam, K. (2023). Classification and prediction of student performance data using various machine learning algorithms, *Materials Today: Proceedings* **80**: 3782–3785. SI:5 NANO 2021.
URL: *https://www.sciencedirect.com/science/article/pii/S221478532105241X*

Pande, S. M. (2023). Machine learning models for student performance prediction, *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, pp. 27–32.

Poudyal, S., Nagahi, M., Nagahisarchoghaei, M. and Ghanbari, G. (2020). Machine learning techniques for determining students' academic performance: A sustainable development case for engineering education, *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pp. 920–924.

Rodas-Silva, J. and Parraga-Alava, J. (2023). Predicting academic performance of low-income students in public ecuadorian online universities: An educational data mining approach, *Predicting Academic Performance of Low-Income Students in Public Ecuadorian Online Universities: An Educational Data Mining Approach*, pp. 52–63.

Smith, G. and Rustagi, I. (2021). When good algorithms go sexist: Why and how to advance ai gender equity, *Stanford Social Innovation Review* **80**.
URL: *https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity*

Solano, J. A., Lancheros Cuesta, D. J., Umaña Ibáñez, S. F. and Coronado-Hernández, J. R. (2022). Predictive models assessment based on crisp-dm methodology for students performance in colombia - saber 11 test, *Procedia Computer Science* **198**: 512–517. 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare.
URL: *https://www.sciencedirect.com/science/article/pii/S1877050921025175*

Tovar-Soto, J. P., González, M. O. and Sánchez, J. A. S. (2022). Digital agriculture for urban crops: design of an iot platform for monitoring variables, *2022 IEEE International Conference on Automation/XXV Congress of the Chilean Association of Automatic Control (ICA-ACCA)*, pp. 1–6.