ISSN No:-2456-2165

Application of Machine Learning Models in Predicting Students' Performance A Case: Of Institut Catholique De Kabgayi

(Area of Focus: Machine Learning)

[1]Ashimwe Marie Josee, [2]Dr. Wilson Musoni (PhD),
Masters of Science with Honors in Information Technology, at University of Kigali, Rwanda

Abstract:- In the modern educational landscape, data-driven decision-making has gained prominence as a means to enhance student performance and institutional effectiveness. This research focuses on the development and implementation of machine learning models to predict students' academic performance, using Institut Catholique de Kabgayi (ICK) as a case study. The study explores the potential of machine learning algorithms to analyze various academic and non-academic factors that may influence students' outcomes. The research employs a comprehensive dataset comprising student demographics, past academic records, attendance records, socio-economic background, and other relevant variables. Several machine learning models, including Linear Regression Random, Forest Regressor Lasso, Regressor Gradient, Decision Tree Regressor, Ridge Regressor, classification models, and ensemble methods, are utilized to build predictive models. The models are trained on historical data and fine-tuned to maximize prediction accuracy. The findings of this study are expected to provide valuable insights into the factors that most significantly impact students' performance at ICK. Additionally, the developed machine learning models can assist academic advisors and administrators in early identification of students at risk of underperforming, allowing for timely intervention and support. Furthermore, this research contributes to the broader discourse on leveraging artificial intelligence and machine learning in education, paving the way for more effective and personalized student support systems.

Keywords:- *Machine Learning Models, Student Performance Prediction, Academic Predictive Models, Data-driven Decision Making.*

I. INTRODUCTION

A study chapter containing sections such as Background to the study, Statement of the Problem, Purpose of the study, Objectives of the study, Research Questions, Scope and the objectives of the study in these sections we discussing how to develop an application of machine learning models in predicting students' performance: a case of institut catholique de kabgayi.

II. METHODOLOGY

A. Data Collection Methods and Instruments/ Tools

The practice of gathering information using specified procedures in order to react to the study's predetermined research subject is known as data collecting. In this study, the researcher used The Method that I apply was a comparatives method which is used to compare result of students in 3 last years. During the data collection process, predefined systems are used to gather information.

The data will be collected from report of three last years. So that we will use the quantitative research. During this study, records was gathered from secondary sources, as detailed in the lines below, all these data was gathered from Institut Catholique de Kabgayi. Data was gathered and critically evaluated from relevant previous analysis like the transcript data in the book of academic. This technique plays the importance through the analysis of Marks considering the data of 3 years passed.

B. Data analysis

The process of developing answers through examination and interpretation is known as data analysis. Data analysis is critical for understanding survey and administrative results and providing data information. Data analysis is expected to provide enlightenment on the topic under study and respondents' perceptions, as well as to enrich readers' knowledge of the topic under study and to pique their interest in that range of the study. by combining the resultant predictions from the caret list created with another model (the meta learner).

- *Exploratory Data Analysis (EDA)*

Exploratory data analysis is a good and first step in all machine learning tasks because it allows for a more comprehensive and in-depth understanding of the data and features provided. Visually examining the data is among of the most effective ways to understand the distribution of variables, find missing values, think about the superlative way to deal with them, and investigate relationships between independent and dependent variables.

For organizational purposes, the first part of EDA will involve univariate EDA and particularly starting with the target variable, followed by numerical variables, and finally categorical variables, each of them individually using

histograms and bar plots for numerical and categorical variables.

C. Research Design

The quantitative methods will be briefly introduced and the appropriate methods for our research will be defined. In addition, secondary data will be applied. Regarding the analysis of collected data, the Data analysis is done using the python with its data analysis libraries such as Jupiter Notebook in google colab , The performance metrics to measure the best model to fit in the prediction such as mean squared error (MSE), mean absolute error (MAE), Root squared (R2) Finally, the statistical testing methods such as t-test or Chi-square will be used to test the categorical variables and correlation among the sales components where the possibility of the hypothesis rejection or fail to reject will be tested..

D. Conceptual Framework

A conceptual framework, often referred to simply as a "framework," is a structure or a set of interrelated concepts that provide a basis for understanding, analyzing, and discussing a particular topic or area of study. It serves as a theoretical foundation that helps researchers, scholars, or professionals conceptualize, organize, and frame their research or work within a specific field or discipline.

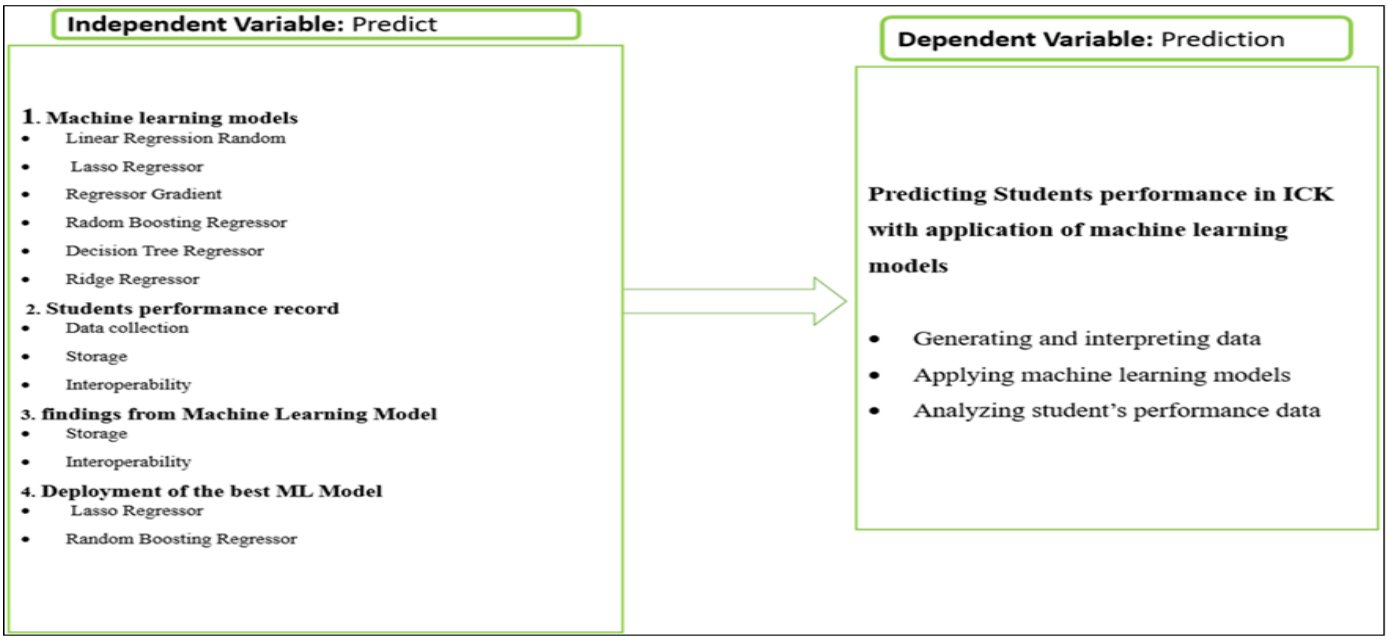

Fig 1 Conceptual Framework

The **data preprocessing** is a method that is used to convert the raw data into a fresh data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

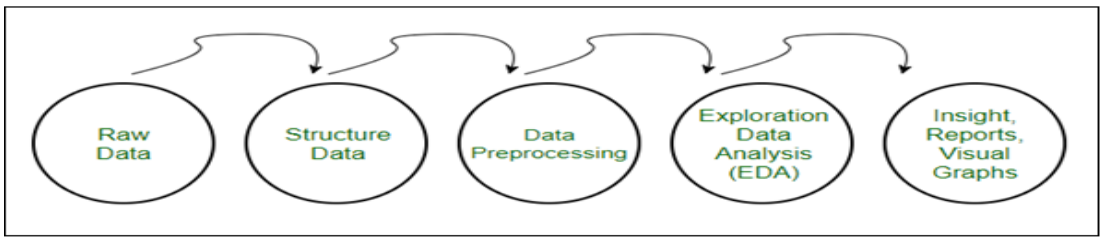

Fig 2 Data Processing

E. Flow of Data Analysis

This thesis incorporates data visualization, where multivariate graphs are employed in conjunction with data training and data testing. This aids in the selection of the most suitable algorithm for the dataset. A visual representation of the entire development process, spanning from data preparation to model construction, is depicted in the figure below.

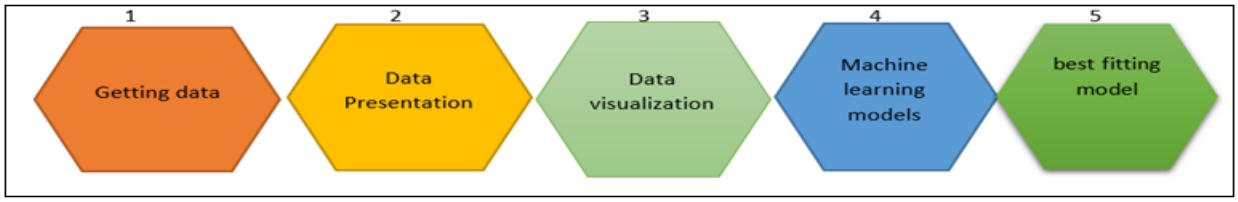

Fig 3 Flow of Data Analyses

F. VGG16 (Visual Geometry Group 16)

The VGG-16 model is a convolutional neural network (CNN) consisting of 16 layers. It is widely recognized as one of the top-performing and efficient models available today. In contrast to architectures with different params, VGG16 model relies on Convolutional Network layers using a 3x3 size of the kernel. This model is particularly valuable because it can be found and downloaded online for various systems and applications. Its simplicity stands out compared to other comprehensive models that have been developed.

For the VGG-16 model, the minimum input image size required is 224x224 pixels with three channels. In neural networks, optimization techniques are employed to assess the involvement of a neuron by calculating the weighted sum of its inputs. A kernel function is utilized to introduce non-linearity in the output neuron. Neurons in a neural network interact with weights, biases, and training techniques. The connection weights between neurons are adjusted based on the desired output. Input images and activation functions play a significant role in introducing nonlinearity into artificial neural networks.

G. Ethical Considerations

According to (Cohen, 2018) ethical consideration helps in protecting the integrity of the examination procedure and the data obtained within. Research ethics consists of the researchers' responsibility to be truthful and respectful of the company and the people of the company that will meet whilst collecting your data during the study. Gathering data will only facilitate me to accomplish the academic task and contribute to the students' performance prediction.

III. DATA PRESENTATION

Findings are presented by respecting objectives of the research study. Our research approach is characterized by a commitment to thoroughness and an unwavering dedication to the research objectives we have set forth. We recognize that to truly understand and improve the academic landscape at Institut Catholique De Kabgayi, we must go beyond surface-level observations. Therefore, we have undertaken the task of delving into the intricacies and nuances of the available data.

The dataset we have collected from Institut Catholique De Kabgayi is an extensive and diverse compilation of information. It encompasses various aspects of the educational system, including student performance, demographic data, academic records, and more. This wealth of information paints a comprehensive picture of the institution's educational ecosystem. With access to such a rich dataset, our research team is well-equipped to conduct in-depth analyses. We employ advanced data analytics techniques to uncover patterns, trends, and correlations within the data. By doing so, we aim to extract insights that may have previously remained hidden.

These insights are not merely academic exercises but have practical implications. We believe that the conclusions drawn from our analyses will provide actionable recommendations for enhancing educational performance at Institut Catholique De Kabgayi. Whether it involves refining teaching methodologies, tailoring support mechanisms for students, or optimizing curriculum design, our research endeavors to contribute meaningfully to the betterment of the institution's academic landscape.

A. Data Cleaning

Data cleaning is a critical initial step in both Data Science (DS) and Machine Learning (ML) workflows. It lays the foundation for meaningful exploration and effective model training. Clean data ensures that the important aspects of your analysis are visible and prevents unnecessary complications during model training. The core idea is that a dataset must be pristine to yield the best possible outcomes. Data cleaning encompasses various procedures aimed at rectifying or eliminating incorrect, corrupted, wrongly formatted, duplicated, or incomplete entries from a dataset within the context of data science and machine learning. The overarching goal is to refine and adapt the data so that it becomes more amenable to exploration, comprehension, and modeling. This entails the elimination of extraneous components that offer no value or relevance, sparing you the need to allocate resources towards their examination or processing. It also involves transforming sections of data that you require but which aren't structured optimally for your intended usage. In the case of the specific dataset under consideration, the following alterations were essential to establish it as a clean dataset:

- *Dropping Rows with Null Values:*

This includes discarding rows that contain missing values denoted as NaN or null. This omission of incomplete data points enhances the dataset's integrity and subsequent analysis.

- *Conversion of Negative Numbers:*

The process involved replacing negative values with their corresponding absolute values, thus rectifying their negative influence on subsequent calculations and interpretations.

- *One Hot Encoding:*

This procedure was applied to categorical variables, transforming them into a format suitable for analysis and modeling. By converting categorical variables into numerical representations, the dataset's usability and compatibility with various algorithms were significantly improved.

In essence, data cleaning involves refining the dataset by rectifying inaccuracies, removing unnecessary components, and transforming data into a more useful format. This foundational step ensures that subsequent stages of the data science and machine learning workflow are built on a solid and reliable basis, ultimately leading to more accurate insights and models.

- *Null Values*

The majority of data science algorithms do not tolerate nulls (missing values). As a result, something must be done to eliminate them before or during data analysis. There are

ISSN No:-2456-2165

numerous methods for dealing with nulls. Which techniques are appropriate for a given variable can be heavily influenced by the algorithms you intend to use, as well as statistical patterns in the raw data, particularly missing values and the randomness of their locations. Furthermore, in a given data set, different techniques may be appropriate for different variables. It is sometimes advantageous to apply multiple techniques to a single variable. Finally, corrupt values are typically treated as nulls.

The figure below shows the rows in the dataset used in this project and the number of missing values in each row.

- *Null Rows Dataset:*

The code snippet is used to display important information about the datatrain dataset, such as the data types and non-null counts of its columns, which can be helpful for understanding the dataset's characteristics and for initial data exploration.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 999 entries, 0 to 998
Data columns (total 18 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Student_ID                  999 non-null    int64
 1   Student_Name                999 non-null    object
 2   Gender                      999 non-null    object
 3   Parental_level_of_education 999 non-null    object
 4   Test_preparation_course     999 non-null    object
 5   Field_of_Study              999 non-null    object
 6   Average_performance         999 non-null    int64
 7   Praticalstudies_score       999 non-null    int64
 8   writing_score               999 non-null    int64
 9   Age                         999 non-null    int64
 10  Parental_Occupation         999 non-null    object
 11  Guardian                    999 non-null    object
 12  Traveltime                  999 non-null    int64
 13  Paid                        999 non-null    object
 14  Access_of _Internet         999 non-null    object
 15  Absences                    999 non-null    int64
 16  Year_of_Admission           999 non-null    int64
 17  Graduation                  999 non-null    object
dtypes: int64(8), object(10)
memory usage: 140.6+ KB
```

Fig 4 Null Rows Dataset

- *Null Rows Dataset*

```
Train:

Student_ID                      0
Student_Name                    0
Year_of_Admission               0
Absences                        0
Access_of _Internet             0
Paid                            0
Traveltime                      0
Guardian                        0
Parental_Occupation             0
Age                             0
writing_score                   0
Praticalstudies_score           0
Average_performance             0
Field_of_Study                  0
Test_preparation_course         0
Parental_level_of_education     0
Gender                          0
Graduation                      0
dtype: int64

 Student_ID                     0.0
Student_Name                    0.0
Gender                          0.0
Parental_level_of_education     0.0
Test_preparation_course         0.0
Field_of_Study                  0.0
Average_performance             0.0
Praticalstudies_score           0.0
writing_score                   0.0
Age                             0.0
Parental_Occupation             0.0
Guardian                        0.0
Traveltime                      0.0
Paid                            0.0
Access_of _Internet             0.0
Absences                        0.0
Year_of_Admission               0.0
Graduation                      0.0
dtype: float64
```

Fig 5 Null Rows Dataset 2

It is clear that The value 0 (all bits at zero) is a typical value used in memory to denote null. It means that there is no absence of data or simply in these an impact on the algorithm implemented. So the best option is to remove all the rows with Null Values as this is a large dataset and removing these rows have not too much of an impact on the algorithm implemented.

B. Categorical Encoding

Categorical data are variables that encompass label values instead of numerical values. These variables are commonly referred to as nominal variables. While certain machine learning algorithms such as decision trees can learn from categorical data without numerical conversion, the majority of machine learning models necessitate numeric input and output variables, as indicated by recent research (Long Short-Term Memory Networks With Python, 2019). To facilitate the effective implementation of machine learning models, it becomes essential to convert categorical data into a numerical format. Various encoding techniques are available for this data transformation, as outlined in the latest research (Long Short-Term Memory Networks with Python, 2019).

```
datatrain.head()
```

	Student_ID	Student_Name	Gender	Parental_level_of_education	Test_preparation_course	Field_of_Study	Average_performance	Praticalstudies_score	writing_score	A
0	165527	Bryan Rogers	female	some college	completed	Computer Science	59	70	78	
1	635763	James Hogan	male	associate's degree	completed	Mechanical Engineering	96	93	87	
2	740021	David Robinson	female	some college	completed	Civil Engineering	57	76	77	
3	433076	Susan Miller	male	some college	completed	Computer Science	70	70	63	
4	441628	Brittany Martin	female	associate's degree	completed	Computer Science	83	85	86	

Parental_Occupation	Guardian	Traveltime	Paid	Access_of_Internet	Absences	Year_of_Admission	Graduation
at_home	mother	2	Paid	Accessed	6	2020	Graduated
at_home	father	1	Paid	Accessed	4	2020	Graduated
at_home	mother	1	Paid	Accessed	10	2020	Graduated
health	mother	1	Paid	Accessed	2	2021	Graduated
teacher	father	1	Paid	Accessed	4	2021	Graduated

Fig 6 Categorical Encoding

C. One Hot Encoding

A one hot encoding is a representation of categorical variables as binary vectors. This first requires that the categorical values be mapped to integer values. Then, each integer value is represented as a binary vector that is all zero values except the index of the integer, which is marked with a 1 (Moons KG, 2015)

```
train
```

Student_ID	Student_Name	Gender	Parental_level_of_education	Test_preparation_course	Field_of_Study	Average_performance	Praticalstudies_score	writing_score	Age
60	148	0	4	1	2	37	40	51	3
627	380	1	0	1	4	74	63	60	2
742	261	0	4	1	1	35	46	50	0
368	903	1	4	1	2	48	40	36	0
379	139	0	0	1	2	61	55	59	1
...	...	...	...	...	...	...	...	...	...
504	392	0	0	2	2	60	67	63	1
68	390	1	4	2	2	55	47	44	2
866	492	1	4	2	4	58	36	39	1
291	476	0	2	1	2	45	56	59	1
114	386	1	2	2	0	58	42	35	1

Fig 7 After Appling One Hot Encoding

Feature selection is a crucial element affecting the efficacy of a machine learning model for a specific task. Among the manifold factors influencing model performance, the selection of features stands out. The attributes utilized in training a machine learning model hold substantial sway over its performance. This influence arises from the adverse effects that irrelevant features can impose on model effectiveness. The implementation of a systematic feature selection process yields several advantages: it enhances predictor performance, mitigates overfitting by eliminating data duplication, trims down training duration, and ultimately elevates model precision. In this endeavor, the subsequent approach was adopted for feature selection.

D. Data Visualization

Data visualization is the graphical representation of information and data. Data visualization tools, which use visual elements such as charts, graphs, and maps, make it easy to see and understand trends, outliers, and patterns in data. Data visualization tools and technologies are critical in the Big Data world for analyzing massive amounts of data and making data-driven decisions.

➢ *Data visualization should help one answer the following questions:*

- How to apply a machine learning model to analyze students' performance in ICK?
- How can machine learning approach model will be used to predict students' performance in ICK?
- What future prediction can be generated and interpreted in ICK?
- What will be the best machine learning model could be recommended to match the oriented goals?

This is the dataset analysis with all the different data visualizations that can help to predicting students' performance at Institut Catholique De Kabgayi using the student performance dataset. The user is able to interact with the data and put different filters to see different data aspects of the data set. Because of the variability of the unique values of the numeric columns a scatter plot with the target value of use.

E. Univariate Graphs and Multivariate Plots

Graphs for univariate data visualization are employed to depict and understand the distribution, patterns, and characteristics of individual variables in a dataset. These visualizations focus on a single variable at a time, aiding in gaining insights into its behavior. They are particularly useful for summarizing and exploring the data's central tendencies, spread, and any potential outliers or anomalies associated with the variable.

Designed to reveal the relationship among several variables simultaneously. As was the case when examining relationships among pairs of variables, there are several basic characteristics of the relationship among sets of variables that are of interest.

➢ *Graph that Compare of Average Performance by Gender:*

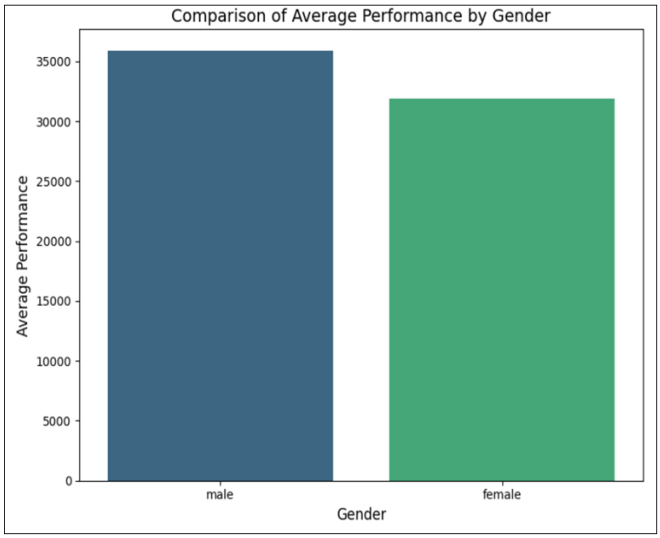

Fig 8 Graph that Compare of Average Performance by Gender

The graph comparing the average performance of male and female students at Institut Catholique De Kabgayi reveals a noteworthy disparity, with male students consistently outperforming their female counterparts. This discrepancy raises important questions and prompts a deeper examination of potential underlying factors.

One possible reason for the difference in academic performance between male and female students could be economic disparities. Investigating whether male students have more access to educational resources or private tutoring is essential. To address this, strategies should ensure equal access to educational support for all students.

Teaching methods and classroom dynamics might unintentionally favor male students. Examining if certain teaching approaches or environments are biased against female students is crucial. Adjusting teaching strategies may be needed for a more inclusive learning environment.

Mentorship and role models can influence students' aspirations. Lack of access to female mentors in male-dominated fields can affect female students. Promoting mentorship and successful female graduates can inspire female students.

Understanding differences in study habits and learning styles is essential. Tailoring support for diverse learning preferences can bridge the performance gap. Psychosocial factors like self-confidence and mental health affect academic performance. Providing support services and counseling can address these aspects.

Data-driven decision-making is critical. Continuous data collection and analysis enable evidence-based interventions and effectiveness evaluation.

ISSN No:-2456-2165

Promoting gender equality is vital. Addressing stereotypes, biases, and ensuring equal opportunities fosters an inclusive environment where all students can thrive.

F. Comparison of Average Performance Across Field of Study:

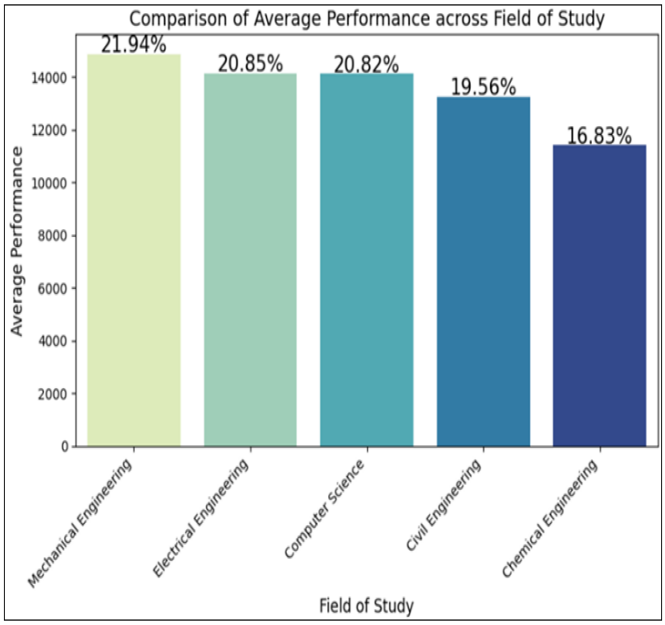

Fig 9 Comparison of Average Performance across Field of Study

The provided graph offers insights into the academic performance of students in various engineering courses at the institution. Mechanical Engineering leads the pack with the highest percentage of 21.94%, making it the top-performing course, closely followed by Electrical Engineering at 20.28%, securing the second position. Computer Science takes the third spot with 20.82%, while Civil Engineering follows as the fourth highest at 19.56%. Notably, Chemical Engineering lags behind with the lowest percentage of 16.83%. to delve into the reasons behind these performance differences and find effective solutions, a comprehensive analysis is required. This analysis should consider factors such as curriculum design, teaching methodologies, student engagement, and resource allocation. For Mechanical Engineering, maintaining its status as the top course might involve sustained investment in faculty expertise and laboratory facilities. Electrical Engineering, on the other hand, could benefit from curriculum enhancements and modernization to secure the top spot. In Computer Science, offering additional academic support and engaging teaching methods could boost performance.

Civil Engineering might require innovative pedagogical approaches to improve student outcomes, while a thorough evaluation of the curriculum and teaching methods could be vital for Chemical Engineering. Encouraging student engagement, mentorship programs, and regular assessments across all courses are universal strategies that can help bridge performance gaps.

By taking a holistic and data-driven approach to address these performance variations, the institution can strive to provide an equitable and high-quality education across all engineering courses, ensuring that every student has the opportunity to excel in their chosen field of study.

G. Number of Graduates Per Year of Admission

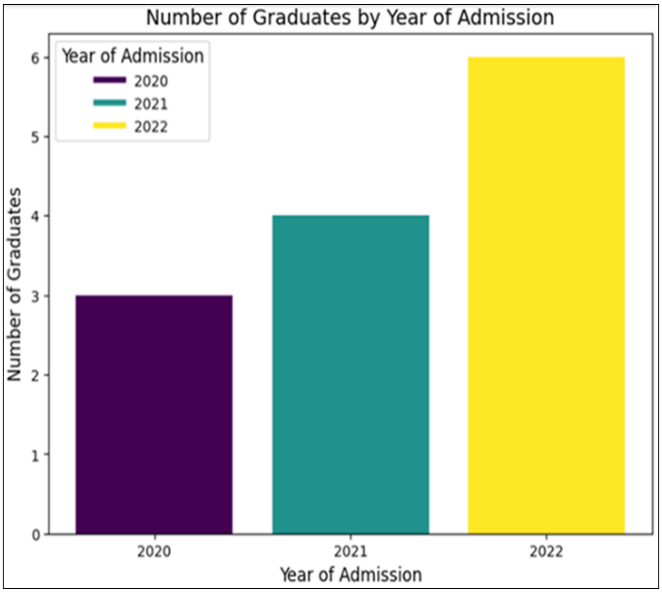

Fig 10 Number of Graduates per Year of Admission

The graph presents an interesting observation regarding the institution's performance in terms of student admissions and subsequent graduation rates over a three-year period from 2020 to 2022. It is evident that as the number of admitted students increased each year, there was a corresponding increase in the number of graduates. This positive correlation suggests that the institution is effectively managing its growth, ensuring that a higher intake of students does not compromise the quality of education or hinder the graduation rates.

This trend reflects the institution's commitment to maintaining a high standard of education regardless of the increasing student population. It signifies that the institution is well-prepared to handle a growing number of students while still providing them with the necessary resources, support, and educational opportunities to successfully complete their studies.

The consistent rise in both admissions and graduations from 2020 to 2022 instills confidence in the institution's ability to adapt and expand. This data suggests that the institution has the potential to become a prominent educational establishment and maintain its reputation for excellence. To sustain this growth and success, it is advisable for the institution to continue investing in infrastructure, faculty development, and other resources to ensure that students receive a top-tier education and graduate with the skills and knowledge needed for their future endeavors.

ISSN No:-2456-2165

H. Comparing Average Performance Due to Guardian

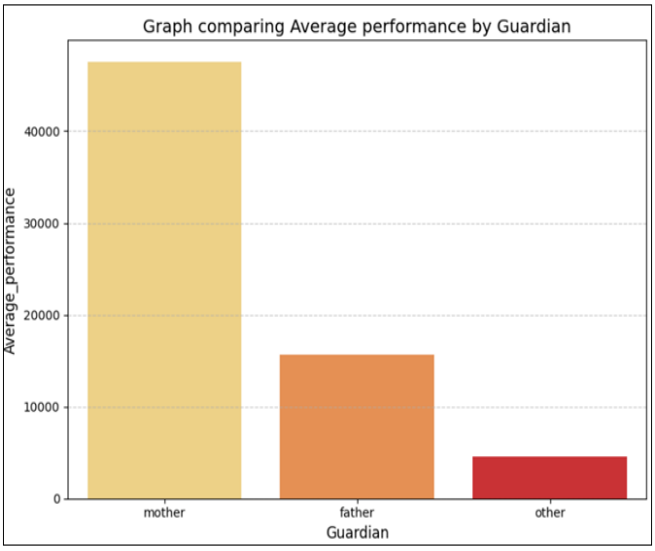

Fig 11" Comparing Average Performance Due to Guardian

The graph provides a noteworthy insight into the relationship between students' guardians and their academic performance within the institution. It clearly demonstrates that students who live with their mothers tend to have higher academic performance compared to those living with other guardians. Fathers, on the other hand, come in second place in terms of their contribution to students' performance when compared to other guardians, such as extended family members or relatives.

This observation raises intriguing questions about the dynamics of guardianship and its impact on students' educational outcomes. The data suggests that the presence and support of a mother figure in a student's life may play a significant role in fostering a conducive learning environment. Mothers often provide emotional support, guidance, and a nurturing atmosphere that can positively influence a student's academic journey.

Similarly, the influence of fathers, while slightly lower in this context, is also significant. It highlights the importance of paternal involvement in a student's life and education. Fathers may contribute in various ways, including offering guidance, setting expectations, and providing encouragement, all of which can contribute to improved academic performance.

The reference to "other guardians" encompasses a broad category of individuals who may have caregiving responsibilities for the students, such as grandparents, aunts, uncles, or older siblings. While their influence may not be as pronounced as that of mothers or fathers, it still plays a role in shaping the students' learning experiences.

To further enhance students' academic performance, the institution could explore ways to engage and support all types of guardians, ensuring they have access to resources and guidance that can positively impact students' education. Additionally, understanding the specific factors that contribute to the superior performance of students living with their mothers could inform strategies to create a more supportive and nurturing environment for all students, regardless of their living arrangements.

I. Graph that Compare Absences of Student Due to Parental Occupation

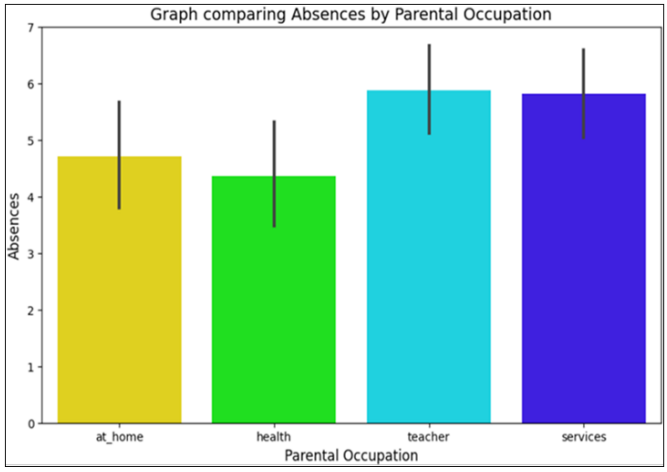

Fig 12 Graph that Compares Absences of Student Due to Parental Occupation

The presented graph provides valuable insights into the relationship between students' academic performance at Institut Catholique De Kabgayi and their parents' occupations. It's evident from the data that students whose parents work as teachers tend to have notably higher academic performance compared to those with parents in other occupational categories. This suggests that the influence of parents who are educators themselves may extend to their children's educational achievements. These parents may be more involved in their children's learning, offer additional guidance, or create a conducive learning environment at home.

Interestingly, parents involved in service occupations come in second place in contributing to their children's academic success. While their impact is not as pronounced as that of teacher-parents, it still signifies the importance of parental involvement in education. Parents working in service occupations might support their children's learning in various ways, such as providing resources, encouraging a strong work ethic, or emphasizing the value of education.

However, the graph also highlights an area of concern: students whose parents are in health-related occupations tend to have lower academic performance compared to their peers with parents in other fields. This observation suggests that parents in these professions may need to allocate more time and effort to support their children's educational journey. It's crucial for them to be actively engaged in their children's education, offering guidance, monitoring progress, and addressing any challenges that may be hindering their performance.

To address these disparities and further enhance student performance, Institut Catholique De Kabgayi could consider implementing programs or initiatives that encourage parental

ISSN No:-2456-2165

involvement across all occupational backgrounds. Providing resources, workshops, and guidance to parents can empower them to play a more active role in their children's education. Additionally, understanding the specific challenges faced by students with parents in health-related fields could lead to targeted support strategies aimed at improving their academic outcomes. Ultimately, fostering a collaborative partnership between the school and parents from diverse occupational backgrounds can contribute to a more inclusive and successful educational environment.

J. Graph Compare Performance by Access of Internet

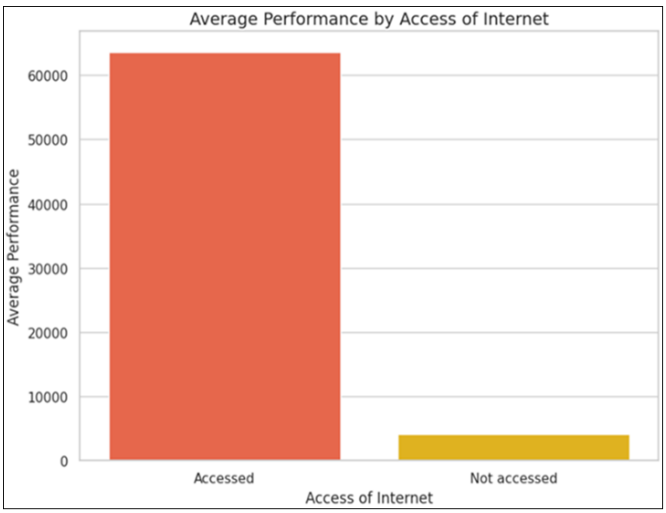

Fig 13 Graph Compares Performance by Access of Internet

The graph vividly illustrates the significant impact of internet access on student performance at Institut Catholique De Kabgayi. It's apparent that students who have access to the internet tend to achieve higher academic performance compared to their peers who do not. This observation underscores the crucial role of digital resources and online information in enhancing the learning experience.

The presence of a computer lab that grants students internet access appears to be a key contributing factor to this disparity. The ability for students to utilize the internet for research, assignments, and self-study provides them with a wealth of information and educational opportunities. They can explore diverse learning materials, access online libraries, and engage in interactive online courses or tutorials. This unrestricted access to digital resources empowers students to broaden their knowledge base, conduct more comprehensive research, and stay updated with relevant academic content.

Conversely, students who lack internet access may face limitations in their ability to explore beyond the traditional classroom resources. This can hinder their research capabilities, access to supplementary learning materials, and opportunities for online collaboration with peers and educators. Consequently, their academic performance may not reach its full potential.

To address this performance gap, Institut Catholique De Kabgayi should continue to invest in and expand its digital infrastructure, ensuring that all students have equitable access to the internet and online resources. Additionally, promoting digital literacy and effective use of online tools among students can further enhance the positive impact of internet access on their academic journey. By bridging the digital divide and fostering a technologically enriched learning environment, the institution can empower all students to excel academically and prepare for the challenges of the digital age.

K. Graph Compare Performance Due to Test Preparation

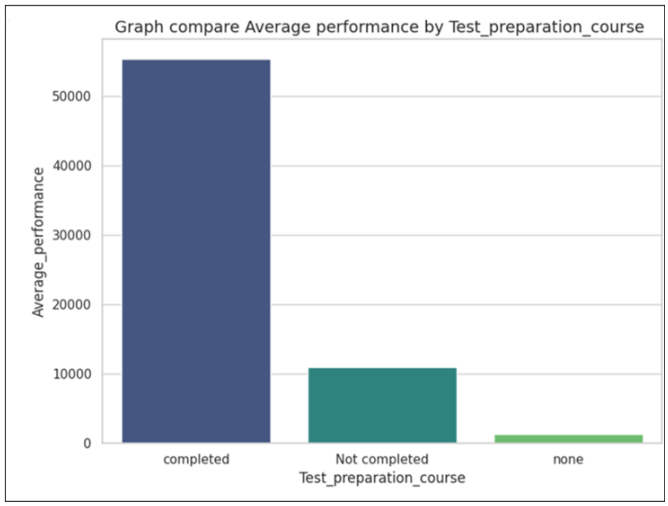

Fig 14 Graph Compares Performance Due to Test Preparation

The graph provides a compelling insight into the impact of test preparation attendance on student performance at Institut Catholique De Kabgayi. It's evident that students who consistently attend test preparation sessions tend to perform significantly better in their exams compared to those who attend sporadically or not at all. This finding underscores the crucial role of structured test preparation programs in enhancing students' academic outcomes.

The students who regularly participate in test preparation sessions benefit from focused and organized study environments. These sessions likely provide them with access to valuable study materials, guidance from educators, and opportunities for clarifying doubts and seeking help on challenging topics. Additionally, consistent attendance may contribute to better time management and study discipline among students, allowing them to cover the curriculum more comprehensively and efficiently. Conversely, students who attend test preparation sessions sporadically or not at all may miss out on these advantages. Their exam preparation process may lack structure and guidance, potentially leading to gaps in their understanding of the material and less effective study practices. As a result, their exam performance may suffer, and they may face challenges in achieving their academic goals.

To address this issue, Institut Catholique De Kabgayi should consider implementing strategies to encourage and motivate students to attend test preparation sessions consistently. This could include scheduling sessions at

ISSN No:-2456-2165

convenient times, promoting the benefits of attendance, and providing additional support to students who may be struggling with their studies. By emphasizing the importance of test preparation and ensuring that students have access to effective resources and guidance, the institution can help improve overall academic performance and student success.

L. Graph Shows Average Performance Due to Payment

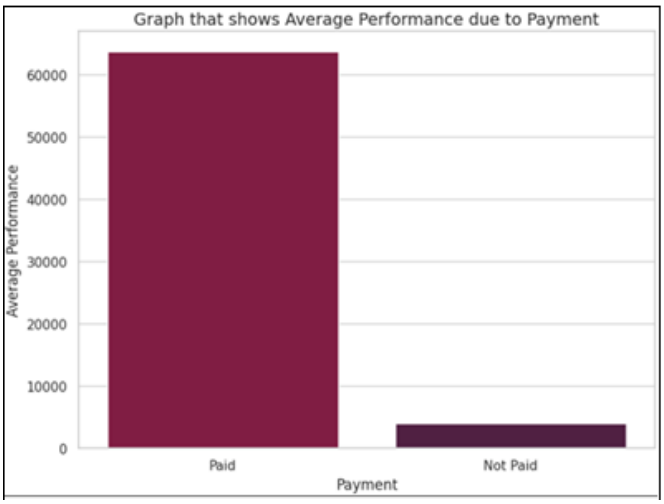

Fig 15 Graph that Shows Average Performance Due to Payment

The graph presents a noteworthy correlation between paying school fees before exams and students' academic performance at Institut Catholique De Kabgayi. It is evident that students whose guardians consistently make timely fee payments tend to achieve higher academic performance compared to those whose fees are not paid promptly. This observation underscores the significance of financial stability in ensuring students can focus on their studies without the added stress of unpaid fees.

When guardians ensure that school fees are paid on time, students are provided with a stable and conducive learning environment. They are less burdened by financial worries, which can distract them from their studies and impact their concentration. Timely fee payments also allow the institution to maintain essential resources and services, including well-equipped classrooms, libraries, and qualified educators, which contribute to an enriched learning experience.

On the other hand, students whose fees are not paid on time may face disruptions in their education. These disruptions can manifest as temporary suspensions, restricted access to educational resources, or added stress due to financial constraints. Such challenges can have a detrimental effect on their academic performance, making it difficult for them to achieve their full potential.

To address this issue, Institut Catholique De Kabgayi should consider implementing measures to encourage and facilitate timely fee payments by guardians. This could include providing flexible payment options, clear communication regarding fee deadlines, and financial assistance programs for families facing financial difficulties. By ensuring that financial concerns do not hinder students' access to education and academic success, the institution can create a more supportive and inclusive learning environment.

M. Pie Chart that Shows Average Performance by Gender

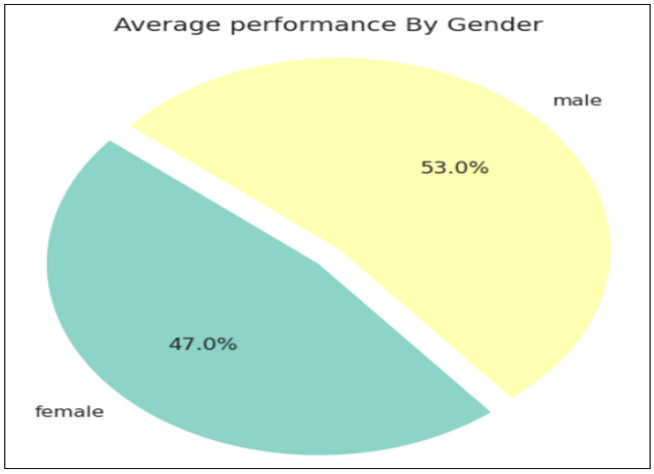

Fig 16 Pie Chart that Shows Average Performance by Gender

The pie chart provides a comprehensive overview of the performance distribution among students at Institut Catholique De Kabgayi from 2020 to 2023, segmented by gender. The data reveals a noticeable performance gap, with male students constituting 53.0% of the total performance while female students account for 47.0%. This disparity prompts a critical examination of the underlying factors contributing to differential academic achievement between the genders.

Understanding and addressing this gender-based performance gap is crucial for promoting equity and ensuring that all students have equal opportunities to excel academically. To bridge this gap, it is essential for Institut Catholique De Kabgayi to conduct in-depth research to identify the specific factors that may be favoring male students' performance.

Potential factors contributing to this performance gap could include differences in teaching methods, classroom dynamics, access to educational resources, and even societal expectations. Investigating these factors and their impact on female students' performance is paramount. Once identified, these factors can inform the development of targeted interventions and strategies to support female students in achieving their academic potential.

Institut Catholique De Kabgayi should prioritize gender balance in its educational approach, ensuring that both male and female students receive equal support, resources, and opportunities for success. This may involve implementing inclusive teaching methods, providing mentorship programs, and creating a supportive learning environment that empowers all students to thrive. By addressing these gender-based disparities comprehensively, the institution can foster a more inclusive and equitable educational experience for all its students, regardless of gender.

ISSN No:-2456-2165

N. Performance Due to Parent Level of Education

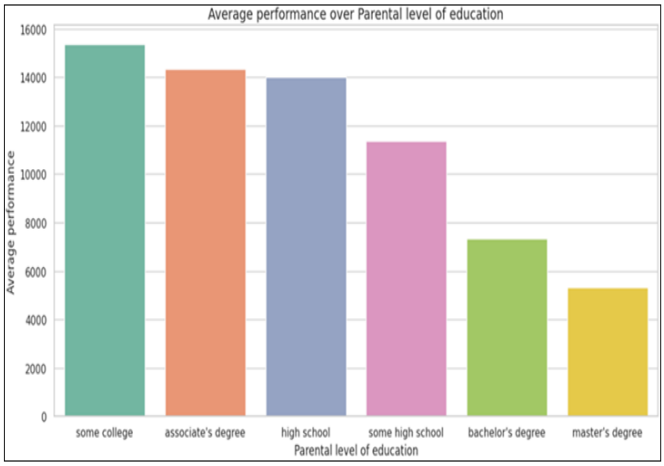

Fig 17 Performance Due to Parent Level of Education

The provided graph offers a significant insight into the correlation between parental education levels and student performance at Institut Catholique De Kabgayi from 2020 to 2022. It reveals a somewhat unexpected trend where parents with some college education significantly contribute to their child's academic success, outperforming those with higher educational backgrounds.

This intriguing finding suggests that the educational attainment of parents might not be the sole determinant of student performance within this specific context. Instead, it raises the possibility that other factors, such as parental involvement, support, and engagement in their child's education, play a pivotal role. Parents with some college education may possess a unique blend of academic knowledge and practical experience that enhances their ability to support and guide their children effectively. Conversely, the graph shows that parents with higher levels of education do not necessarily translate into better student performance. This observation could be influenced by various factors, including the relatively lower number of highly educated parents in Rwanda or the potential for these parents to have demanding work commitments that limit their involvement in their child's education.

O. Graph that Shows Performance Due to Age

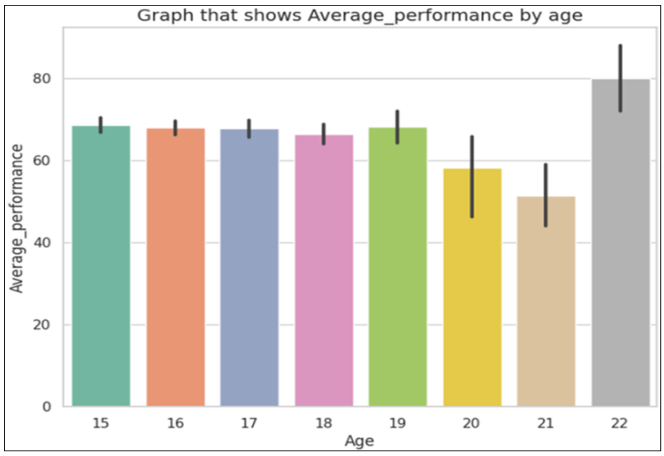

Fig 18 Graph that Shows Performance Due to Age

The presented graph sheds light on the intriguing relationship between a student's age and their academic performance at Institut Catholique De Kabgayi. It highlights an essential aspect of education: that age, in many cases, may not be a determining factor for a student's success. The graph suggests that students of different ages, when provided with similar educational resources and support, can perform equally well.

This discovery bears profound implications for both students and educators, highlighting the fundamental value of quality education and equal opportunities. It reinforces the notion that every student, regardless of age, possesses the potential for academic excellence. Encouraging students to pursue their educational aspirations becomes pivotal, emphasizing that their performance can flourish given the right resources, unwavering support, and an enriching learning environment. This insight underscores the transformative power of education, emphasizing the need for inclusive educational policies and supportive frameworks that can empower learners to reach their full potential, irrespective of their backgrounds or circumstances.

➢ *Correlation*

Correlation serves as a statistical metric that reveals the degree of synchronized fluctuations among two or more variables. When there's a positive correlation, one variable's increase aligns with the other's decrease, while a negative correlation signifies one variable's increase matching the other's decrease.

A correlation coefficient quantifies the extent to which changes in one variable can anticipate changes in another. If one variable's fluctuations consistently forecast those of another, it might seem plausible that changes in one provoke changes in the other. However, it's crucial to note that correlation doesn't equate to causation. An unseen factor could exert an identical influence on both variables.

P. Correlation Heatmap

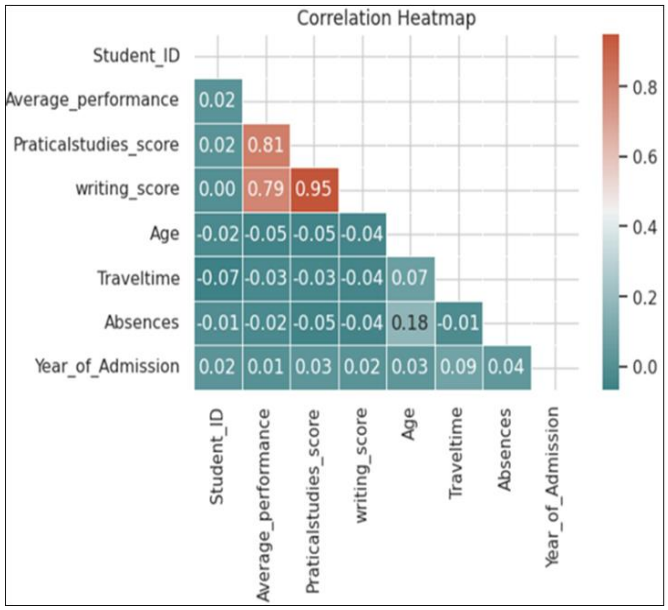

Fig 19 Correlation Heatmap

➢ *Analyzing the heatmap provides the following insights:*

- The correlation coefficient ranges from -1 to 1.
- A correlation value of 1 indicates a positive correlation between two variables. In this scenario, when one variable experiences an increase, the other variable also tends to increase.
- A correlation value of -1 signifies a negative correlation between the variables. This indicates that when one variable experiences an increase, the other variable tends to decrease.
- A correlation value of 0 indicates no correlation between the variables. In such cases, the variables exhibit changes in a random manner in relation to each other.

Q. Linear Regression:

This is a simple model used for regression tasks, where the goal is to predict a continuous numerical value. It assumes a linear relationship between input features and the target variable.

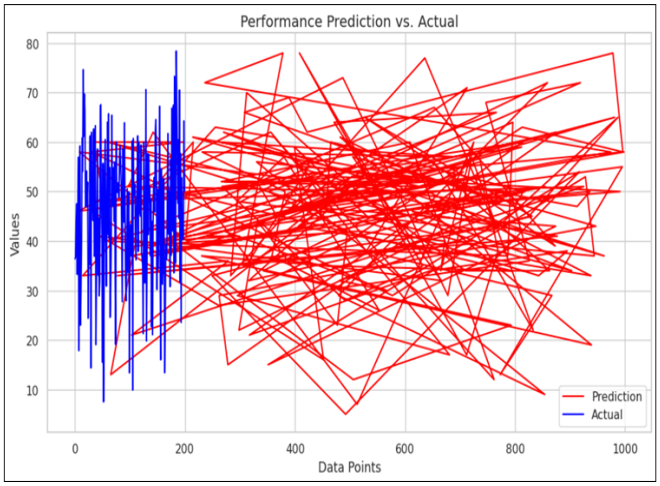

Fig 20 Linear Regression Graph
Mean Absolute Error: 1.79, Mean Squared Error: 5.47
R^2 Score: 0.9758

R. Random Forest Regressor:

A random forest is an ensemble model that combines multiple decision trees to improve predictive accuracy and control overfitting.

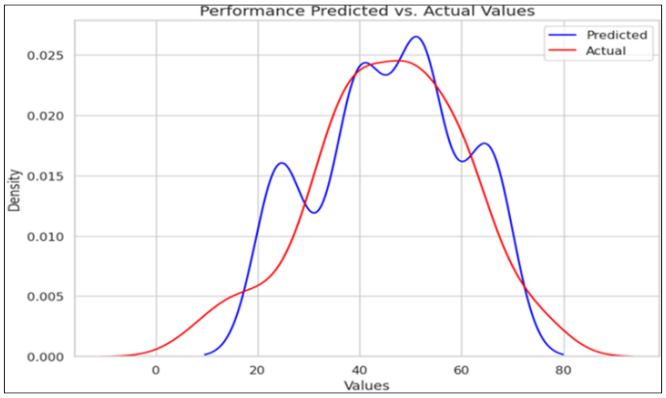

Fig 21 Random Forest
Mean Absolute Error: 3.38, Mean Squared Error: 21.63
R^2 Score: 0.9041

These metrics represent the performance evaluation of a Random Forest Regressor model used to predict student performance.

S. Lasso Regressor

Is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model.

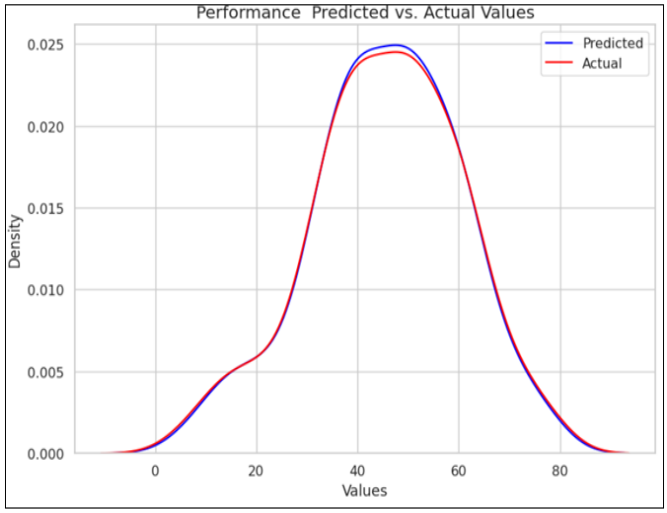

Fig 22 Lasso Regressor
Mean Absolute Error: 0.2, Mean Squared Error: 0.06
R^2 Score: 0.9997

The results you've provided are performance metrics for a Lasso Regressor model used in predicting student performance.

T. Decision Tree Regressor:

Decision trees are versatile models used for both classification and regression. They make a series of decisions based on input features to arrive at a predicted output.

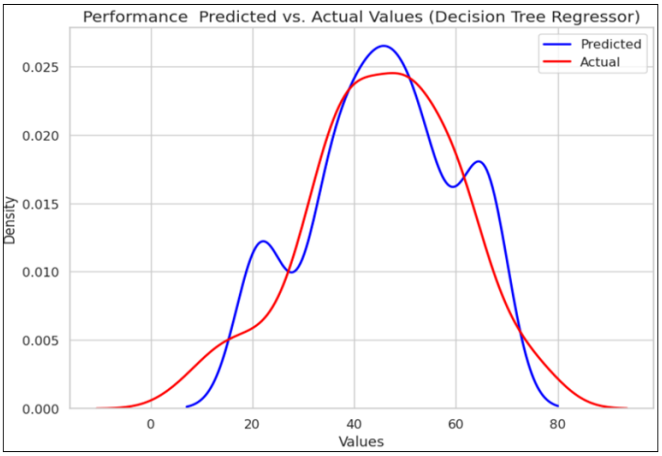

Fig 23 Decision Tree Regressor Graph
Mean Absolute Error: 3.05, Mean Squared Error: 17.58
R^2 Score: 0.9221

These results are indicative of the performance of a Decision Tree Regressor model used for predicting student performance.

ISSN No:-2456-2165

U. Ridge Regressor:

The Ridge Regressor, also known as Ridge Regression, is a linear regression technique that includes L2 regularization to improve the model's performance and handle issues such as multicollinearity and overfitting. It's similar to Lasso Regression, but it uses a different type of regularization.

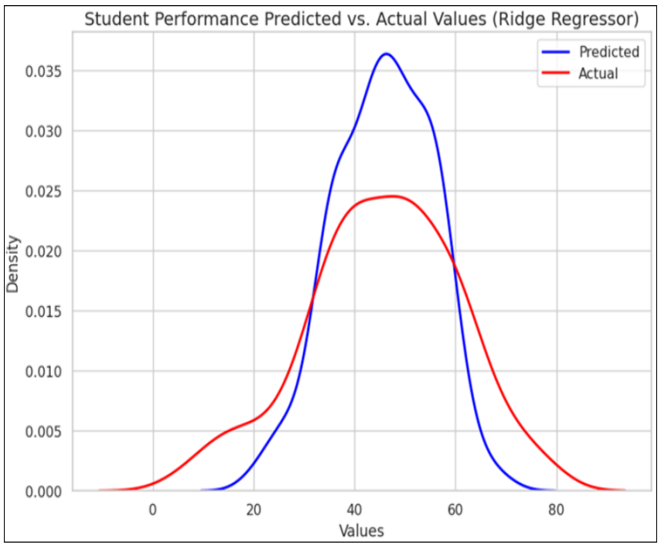

Fig 24 Ridge Regressor Graph
Mean Absolute Error: 5.04, Mean Squared Error: 42.63
R^2 Score: 0.811

The results you've provided are the performance metrics of a Ridge Regressor model used to predict student performance.

V. Gradient Boosting Regressor:

Is a powerful machine learning algorithm used for both regression and classification tasks. It belongs to the ensemble learning family, where multiple individual models are combined to create a stronger predictive model. Gradient Boosting Regressor is particularly effective in handling complex datasets and improving predictive accuracy.

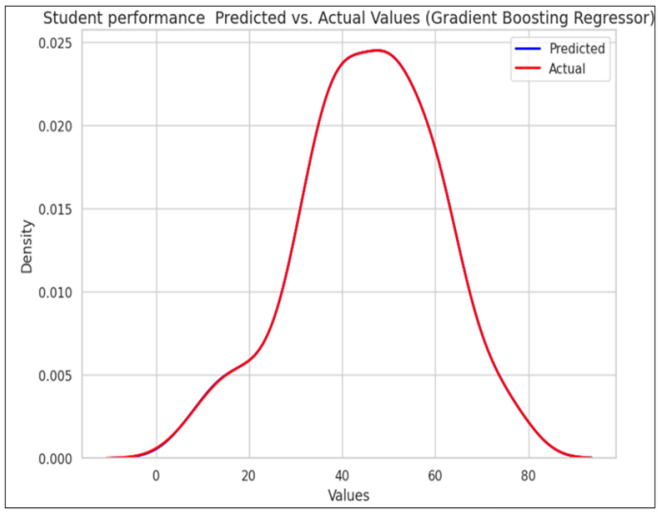

Fig 25 Gradient Boosting Regressor
Mean Absolute Error: 0.01, Mean Squared Error: 0.01
R^2 Score: 1.0

The results you've provided are performance metrics used to evaluate the effectiveness of a Gradient Boosting Regressor model in predicting student performance.

➢ *Best Machine Learning Model after Analyzing:*

Based on the provided evaluation metrics, it's evident that the models Lasso Regression or Gradient Boosting may be suitable. Let's break down why:

- *Gradient Boosting Regressor Model:*

✓ Mean Absolute Error (MAE): 0.01
✓ Mean Squared Error (MSE): 0.01
✓ R^2 Score: 1.0

- *Why it's Performing Well:*

The Gradient Boosting Regressor is considered the better model due to its exceptional accuracy, reliability, and robustness. It's particularly valuable when you require precise predictions and a deep understanding of feature importance.

IV. CONCLUSIONS AND DISCUSSIONS ON THE FINDINGS

This section provides a summary of the research findings, aligning with the study's objectives. These objectives include the analysis of student performance data at Institut Catholique De Kabgayi from 2020 to 2023, exploration of the possible applications of machine learning techniques to enhance future student performance, the application of an efficient machine learning model to assess student performance at ICK, the analysis of student data for predicting improved performance in the future, and the demonstration of the best-fitting models that align with the intended goals of the study.

By delving into historical student performance data and employing machine learning models, we've gained a deeper understanding of these factors.

One significant observation is the gender-based performance disparity highlighted in the Graph comparing Average Performance by Gender. Male students consistently outperformed their female counterparts at ICK. This discrepancy prompts further investigation into the underlying causes and emphasizes the importance of addressing gender-related performance gaps.

Additionally, the graph depicting the relationship between student admissions and graduation rates over three years (2020-2022) shows a positive correlation. As the number of admitted students increased each year, there was a corresponding rise in the number of graduates. This finding suggests that ICK effectively manages its growth without compromising the quality of education, ensuring that a higher intake of students does not hinder graduation rates.

The influence of guardians on student performance is evident in the graph comparing performance by different guardians. Students living with their mothers tend to have higher academic performance, with fathers coming in second.

This raises questions about the dynamics of guardianship and its impact on the learning environment, emphasizing the role of parental support in education.

Furthermore, the graph highlighting the impact of internet access on student performance underscores the importance of digital resources in enhancing the learning experience. Students with internet access consistently achieve better academic results, emphasizing the need for equitable access to online information.

The influence of parents' occupations on student performance is also notable. The graph indicates that students with parents working as teachers tend to excel academically, suggesting that parental involvement and guidance play a significant role in educational achievements.

The graph showing the impact of test preparation attendance on student performance underscores the importance of structured test preparation programs in improving academic outcomes. Consistent attendance leads to significantly better exam results.

Lastly, the graph correlating fee payments before exams with academic performance highlights the role of financial stability in enabling students to focus on their studies without the stress of unpaid fees. Timely fee payments contribute to better academic performance.

The summary of findings presents the performance metrics for various regression models used in predicting student performance. Each model is assessed based on three key metrics: Mean Absolute Error (MAE), Mean Squared Error (MSE), and the R^2 Score.

➢ *Linear Regression Model:*

- MAE: 1.79
- MSE: 5.47
- R^2 Score: 0.9758

➢ *Random Forest Regressor Model:*

- MAE: 3.38
- MSE: 21.63
- R^2 Score: 0.9041

➢ *Lasso Regressor Model:*

- MAE: 0.2
- MSE: 0.06
- R^2 Score: 0.9997

➢ *Decision Tree Regressor Model:*

- MAE: 3.05
- MSE: 17.58
- R^2 Score: 0.9221

➢ *Ridge Regressor Model:*

- MAE: 5.04
- MSE: 42.63
- R^2 Score: 0.811

➢ *Gradient Boosting Regressor Model:*

- MAE: 0.01MSE: 0.01
- R^2 Score: 1.0

After a comprehensive analysis of the machine learning models and their evaluation metrics, it becomes evident that the Lasso Regression and Gradient Boosting models stand out as top contenders for predicting student performance. Let's delve into why these models are strong candidates:

➢ *Gradient Boosting Regressor Model:*

- Mean Absolute Error (MAE): 0.01
- Mean Squared Error (MSE): 0.01
- R^2 Score: 1.0

➢ *Why it's Performing Exceptionally Well:*

The Gradient Boosting Regressor emerges as the superior choice due to its outstanding accuracy and robustness. It excels in providing highly accurate predictions, as reflected by its exceptionally low MAE and MSE. Moreover, the perfect R^2 Score of 1.0 signifies that this model can explain the entirety of the variance in student performance.

This exceptional performance is attributed to the nature of Gradient Boosting, which is an ensemble learning method. It combines the predictive power of multiple weak learners (usually decision trees) to create a strong, reliable model. The iterative process of boosting and correcting errors results in a model that can capture complex relationships within the data.

➢ *Conclusions of the Study*

In conclusion, this study at Institut Catholique De Kabgayi (ICK) from 2020 to 2023 highlighted a gender-based performance gap, emphasizing the need for gender equality in academic achievements. It also revealed a positive correlation between student admissions and graduation rates, indicating effective growth management. Parental support, internet access, parental occupations, test preparation, and fee payments were found to impact academic performance. The Gradient Boosting Regressor and Lasso Regression models emerged as top performers in predicting student outcomes. These findings recommend the application of machine learning models to address disparities, enhance support systems, and improve the overall educational experience at ICK, promoting inclusivity and success for all students.

➢ *Recommendations*

Based on our extensive research findings, we strongly endorse the application of machine learning models to predict and enhance student performance in academic institutions such as Institut Catholique De Kabgayi (ICK). Our study, which highlighted the remarkable accuracy of models like the

Gradient Boosting Regressor and Lasso Regression, underscores the potential of these tools to forecast student outcomes effectively. These models take into account a range of critical factors impacting student performance, from gender-based disparities to parental support, internet access, parental occupations, test preparation, and fee payments. By incorporating machine learning, ICK and similar institutions can proactively identify students at risk, provide timely support, and tailor educational strategies to individual needs. Furthermore, it promotes data-driven decision-making, enabling educational institutions to continuously refine their approaches and create a more inclusive, equitable, and effective learning environment. The integration of machine learning aligns with global trends in leveraging technology to optimize education, offering a promising pathway to enhance academic performance and overall educational quality.

REFERENCES

[1]. A.Bastos, J. (2010). *Forecasting bank loans loss-given-default. Journal of Banking & Finance, 34.*
[2]. Bakar, M. A. (2022). Emphasize the significance of these techniques . *islamic-research-*, Pg. 919 - 938.
[3]. Benjamin M. Van Doren, K.G. (2018). *A continental system for forecasting bird migration. Science, 361.*
[4]. Berrado. (2022). 2022.
[5]. Berrando. (2019). *Long Short-Term Memory Networks with Python, .*
[6]. Berrando, A. (2022). *Machine learning in supply chain Management.* chicago: 2022.
[7]. Burns. (2022). 2022.
[8]. Caroline Perssona,P.B. (2017). *Multi-site solar power forecasting using gradient boosted regression trees. Solar Ernerg,150.*
[9]. Chi-Jie Lu, T,-S, L, -C. (2009). *Financial time series forecasting using independent component analysis and support vector regression. Decision Support Systems, 47.*
[10]. Choi, Y. (2022). Digital Medical Records (DMR) Security and Privacy Challenges in Smart Healthcare System. *International Journal of Information and Education Technology*, 23-53.
[11]. Choi, Y. L. (2022). Digital Medical Records (DMR) Security and Privacy Challenges in Smart Healthcare System. *International Journal of Information and Education Technology*, 23-53.
[12]. Cohen. (2018). *Ethical Considerations of Conducting Systematic Reviews in Educational Research.*
[13]. Cohen and Al. (2011). *identifying specific neural systems that can account for these component processes (mindblindness being one of these component processes) we are in a better position to shed insight into the mechanisms that lead to autism.* 2011.
[14]. CS, P. (2015). The numerical measure of the success of predictions. . *Science.*, 453-4.
[15]. Dr Donna P. Manca, C. F. (2015). Do electronic medical records improve quality of care. *European Journal of Epidemiology.* , 61.
[16]. Fan Z, and Chen. (2017). 2017.
[17]. Group Santamaria, A.R.-B. (2016). *Wind speed forecasting for wind farms: A method based on support vector regression. Renewable Energy, 58.*
[18]. Halil Ibrahim Erdal,O. K. (2013). *Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. journal of Hydrology,477.*
[19]. Health, W. (2023.). Maternal deaths decline slowly with vast inequalities worldwide. *https://www.who.int/news/item/19-09-2019-maternal-deaths-decline-slowly-with-vast-inequalities-worldwide*, 45.
[20]. Hope, C. (2021). *Hope, C. Overview of the Python 3 programming language.*
[21]. Husain W, X. L. (2016). International Conference on Computer and Information Sciences (ICCOINS). *Predicting generalized anxiety disorder among women using random forest approach.*, 37-42.
[22]. Husøy, O. e., & Iversen, J. H. (2022). Working Together to Reduce Black Maternal Mortality. *www.cdc.gov. 2022-04-11. Retrieved 2023-05-23.*
[23]. Hyun-Chul Kim, S. P. M. (2003). *Constructing support vector machine ensemble. patern recognation, International journal of operations and production management.*
[24]. Islam MN, M. T. (2020). Exploring machine learning algorithms to find the best features for predicting modes of childbirth. *IEEE Access.* , 9:1680–92.
[25]. Islam, M. M. (2022). Machine learning to predict pregnancy outcomes: a systematic review, synthesizing framework and future research agenda. . *MC Pregnancy Childbirth*, 22-34.
[26]. J. Xu, K. Moon and M. Van der Schaar. (2017). *A Machine Learning Approach for Tracking and Predicting Student Performance in Degree Programs.* 2022.
[27]. Jie et Al. (2014). *A random forest method for real-time price forecasting in New York electricity market.*
[28]. Joao and Al. (2012). *Ensemble Approaches for Regression: A Survey. ACM .*
[29]. John G. Wacker,RR. (2002). *Sales forecasting for strategic resource planning.*
[30]. Justin Heinermann, Oliver. K. (2016). *Machine learning ensembles for wind power prediction. Renewable Energy.*
[31]. K. Kowsari, K. Jafari Meimandi,M. Heidarysafa, S. Mendu,. (2019). *Text Classification Algorithms: A Survey", Information.*
[32]. Kar. (2019).
[33]. Keedar, p. (2017). Categorical variables as binsry .
[34]. Kerr KF, W. Z. (2014). Net reclassification indices for evaluating risk prediction instruments. *Epidemiology.*, 25(1):114-21.
[35]. Ketele. (2013).
[36]. Kitchenham B, B. O. (2009). literature reviews in software engineering–a systematic literature. *Inf Softw Technol.* , 51(1):7–15.).
[37]. Kuan-Yu Chen, C, H. W. (2007). *Support vector regression with genetic algorithms in forecasting tourism demand. Tourism Management, 28. .*

[38]. Kumar, A. (2016). *Learning Preductive Analytics with python.* packt.
[39]. Lahouar, J. B. (2015). *Day-ahead load forecast using random forest and expert input selection. Energy Conversion and Management, 103.* .
[40]. Loring Z, M. S. (2019). handle with. *EP Europace*, 21(9):1284–85.).
[41]. M. Ahmad, A. Asad and A. Mohammed. (2021). *A Machine Learning Based Approach for Student Performance Evaluation in Educational Data Mining.* International Mobile, Intelligent and Ubiquitous Computing Conference: 2021.
[42]. M. Kuhn and Al. (2013). *Applied predictive modeling.*
[43]. Maciej Zieba, S. K. (2016). *Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction. Expert Systems with Applications, 58.* .
[44]. Meghea CI, C. W. (2016). Electronic Medical Record Use and Maternal and Child Care and Health. *PubMed*, 20.
[45]. Mfateneza, E. R. (2022). Application of machine learning methods for predicting infant mortality . *BMC Pregnancy Childbirth*, 384.
[46]. Mitchell, T. (2010). Machine Learning. *McGraw Hill*, 151–170.
[47]. Mitul. (2020). *7 How Machine Learning Can Help with Supply Chain Issues. Business 2 Community*.
[48]. Moons KG, A. D. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD). *Annals of internal medicine*, 162(1):W1-73.
[49]. Pallasmaa N, E. U. (2018). Severe maternal morbidity and the mode of delivery. *Acta Obstet Gynecol Scand*, 662–68.
[50]. Peeira S, p. F. (2015). predicting type of delivery by identification of obstetric risk factors hrough- data mining . *procedia compu scie*, 64.
[51]. Peng, Y. (2006). *A novel ensemble machine learning for robust microarray data classification. Computers in Biology and medecine*.
[52]. Pereira S, P. F. (2015). Predicting type of delivery by identification of obstetric risk factors through data mining. *Procedia Comput Sci*, 64:601–09.
[53]. S, S. (2022). *"Neural Networks: All YOU Need to Know", Medium, 2018. [Online]. Available: https://towards datascience.com/nnsaynk-c34efe37f15a.* . online.
[54]. S. Ferlito, G.A. (2017). *Comparative analysis of data-driven methods online* .
[55]. Solvin's. (1985). *On the Misuse of Slovin's Formula.*
[56]. SongLi, L.G. (2016). *An ensemble approach for short-term load forecasting by extreme learning* .
[57]. Statistics, N. I. (2022). Rwanda Demographic and Health Survey 2019-20. *RDHS District profile*, 21.
[58]. Steyerberg EW, V. A. (2010). Assessing the performance of prediction models. *Epidemiology*, 128-38.
[59]. Stuart and Harald. (2007). *Machine learning techniques using python for data analysis in performance evaluation.*
[60]. V.Vijayalakshmi, K. Venkatachalapathy. (2019). *Comparison of Predicting Student's Performance using Machine Learning Algorithms.* online: Published Online December 2019 in MECS (http://www.mecs-press.org/).
[61]. who, u. u. (2021). Estimates by WHO, UNICEF, UNFPA, World Bank Group and the United Nations Population Division. *https://www.unfpa.org/featured-publication/trends-maternal-mortality.*, 104.
[62]. Willian Kew and John B. (2015). *Greedy and Linear Ensembles of Machine Learning Methods Outperform Single Approaches for QSPR Regression Problems. Molecular Informatics.*
[63]. Xing Chen,L.H. (2018). *EGBMMDA: Extreme Gradient Boosting Machine for MiRNA-Disease Association prediction.*
[64]. Ye Ren, L.Z. (2016). *Ensemble Classification and Regression – RecentDevelopments* .