**“AI-DRIVEN INSIGHTS:**

**UNDERSTANDING EDUCATIONAL PATTERNS THROUGH VISUALIZATION & ANALYTICS”**

### Final Year Project Initial Report

## C:\Users\rmate\OneDrive\Desktop\misc-docs\iub logo.jpg

**Supervisor**

**Dr. Muhammad Ateeq**

**Submitted by**

**Hamna Qaseem**

**{F20BDATS1E01012}**

**[GROUP-ID: MA-1]**

**Department of Data Science**

The Islamia University of Bahawalpur.

**[January 24th, 2024]**

# **DECLARATION**

I hereby declare that the report of the project entitled “**AI-Driven Insight: Understanding educational patterns through visualization and analytics**” which is being submitted to the Department of Data Science, Islamia university of Bahawalpur, in the fulfillment of the requirements that are required for my final year project. This report has been prepared on the basis of my own work, and where I used helping sources these have been acknowledged.

# **CERTIFICATE OF APPROVAL**

The undersigned certify that they have read, and recommended to the Department of Data Science for acceptance, a final year project report entitled **’AI-Driven Insight: Understanding educational patterns through visualization and analytics’** submitted by **Hamna Qaseem** in fulfillment of the requirements of final year project.

**Supervisor: Dr. Muhammad Ateeq**

Department of Data Science

Faculty of Computing, The Islamia Univeristy of Bahawalpur

**Head of Department: Dr. Muhammad Akmal Khan**

Department of Data Science

Faculty of Computing, The Islamia Univeristy of Bahawalpur

**DATE OF APPROVAL:** November, 2023

# **ACKNOWLEDGEMENT**

This project is undertaken as part of my academic curriculum for the Bachelor’s degree in Data science. I express my sincere appreciation to the **Department of Data Science (faculty of computing)**, for granting me the opportunity to work on this project.

I would like to extend my sincere gratitude to my supervisor, **Dr. Muhammad Ateeq**, for their invaluable guidance throughout this project. Their expertise and support have played a crucial role in shaping my understanding of the subject matter. I am grateful for their continuous encouragement and insightful feedback, which have contributed significantly to my learning experience.

Development of interactive dashboardhas given me the opportunity to delve into the field of exploratory data analysis and gain hands-on experience with visualization and development techniques.

I am confident that the skill and knowledge required through this project will prove bene ficial in my future academic and professional endeavors.

Table of Contents

[1. Introduction 2](#_Toc154570867)

[1.1 Background 2](#_Toc154570868)

[1.2 Motivation 3](#_Toc154570869)

[1.3 Scope and Limitations 3](#_Toc154570870)

[1.4 Significance of the Problem 3](#_Toc154570871)

[1.5 Research Questions or Hypotheses 3](#_Toc154570872)

[1.6 Target Audience 3](#_Toc154570873)

[1.7 Organization of the Document 3](#_Toc154570874)

[1.8 Novelty and Uniqueness 3](#_Toc154570875)

[1.9 Technical and Non-Technical Goals 3](#_Toc154570876)

[1.10 State of Technology 3](#_Toc154570877)

[2. Literature Review 3](#_Toc154570878)

[2.1 Traditional Methods 3](#_Toc154570879)

[2.2 Advanced Techniques and Models 4](#_Toc154570880)

[2.3 Comparative Studies 4](#_Toc154570881)

[2.4 Challenges and Limitations 4](#_Toc154570882)

[3. Methodology 4](#_Toc154570883)

[3.1 Research Design 4](#_Toc154570884)

[3.2 Data Collection 4](#_Toc154570885)

[3.3 Data Preprocessing 4](#_Toc154570886)

[3.4 Feature Engineering 4](#_Toc154570887)

[4. Data Exploration and Analysis 4](#_Toc154570888)

[4.1 Dataset Overview 4](#_Toc154570889)

[4.2 Exploratory Data Analysis (EDA) 5](#_Toc154570890)

[4.3 Statistical Summaries 5](#_Toc154570891)

[4.4 Class Distribution 5](#_Toc154570892)

[4.5 Feature Correlations 5](#_Toc154570893)

[5. Model Development 5](#_Toc154570894)

[5.1 Model Selection Criteria 5](#_Toc154570895)

[5.2 Model Architecture 5](#_Toc154570896)

[6. Results 5](#_Toc154570897)

[6.1 Performance Metrics 5](#_Toc154570898)

[6.2 Comparative Analysis 5](#_Toc154570899)

[7. Deployment 6](#_Toc154570900)

[7.1 Deployment Strategy 6](#_Toc154570901)

[7.2 Technology Stack 6](#_Toc154570902)

[7.3 Integration Steps 6](#_Toc154570903)

[7.4 Cross-Platform Compatibility 6](#_Toc154570904)

[8. Conclusion 6](#_Toc154570905)

[9. References 6](#_Toc154570906)

[10. Acknowledgments 6](#_Toc154570907)

[11. Appendix 6](#_Toc154570908)

[11.1 Code Snippets 7](#_Toc154570909)

[11.2 Data Privacy Policy 7](#_Toc154570910)

[11.3 Model Architecture Diagrams 7](#_Toc154570911)

[11.4 User Interface Screenshots 7](#_Toc154570912)

[11.5 User Guide or Documentation 7](#_Toc154570913)

[11.6 Version Control Information 7](#_Toc154570914)

[11.7 Contributions of FYP Group members 7](#_Toc154570915)

# **Introduction**

A nation’s socio economic progress revolves around its educational system. It plays critical role in building human capabilities and accelerating economic growth. Education leads to positive outcomes like poverty reduction, improved health and good governance in socio economic policy implementation[1]. Secondary education is a fundamental stage in determining a national education system’s effectiveness. Developed nations including those in Europe, are concentrating all their attentions on and research on finding better solutions to the growing issues that young people in secondary school encounter. However, the field of educational data mining is emerging precisely because valuable pedagogical is gained from deep analysis of the data [2].

## **1.1 Background**

When it comes to educational research, knowing how students performance is essential. We investigate the dynamics of academic performance, focusing on different research question like differences between urban and non-urban institutions, how does the gender distribution vary across different institutes?, average age of students taking the exit exam?, any significant differences in performance between male and female students?, subjects are most commonly chosen by students? etc. With the goal of exposing significant variables influencing student achievement, this study will provide insightful information for educational practices and policy. With this contribution, we hope to follow the evolution of education by assessing student achievement based on the results of the Secondary School Certificate Exam. We may observe how education changes in the modern era in response to societal disturbances and the emergence of new technology. It's like traveling through time, as we continue this journey, we've determined that educational departments need a tool to assist them discover and predict where the education is heading by tracing the impact of time on shape of education.

## **1.2. Motivation**

The motivation behind this research lies in the potential to enhance educational outcomes. By analyzing the nuances of student performance in different settings, we aspire to identify areas for improvement and implement targeted interventions, ultimately fostering a more equitable and effective education system.

## **1.3. Scope and Limitations**

Our research uses the outcomes of school exit exams to conduct a thorough analysis of student performance, concentrating on important variables such as gender, topics, and institutes. We can identify complex trends in academic outcomes because to the dataset's comprehensiveness. The constantly changing nature of education and possible data constraints, such as missing values, are the sources of limits. The dynamic nature of education may eventually affect the generalizability of our findings, even though thorough pre-processing tackles some difficulties. Despite these factors, the goal of our research is to offer insightful information on how well students do within the specified parameters.

## **1.4. Significance of Problems**

The role of examination in students’ life has great impact in their academic life and in future for professional life. Without conducting examination and tracking the performance of every individual at particular level cannot be determined and the goal of developing quality in education cannot be achieved. That is why there is need to explore and investigate the link between Secondary School Exam outcomes and performance [3], by understanding how student performance evolves over time and provides insights into the effectiveness of educational interventions and highlight areas where improvement required. Our study holds significant value as it empowers informed decision-making in education and fostering improvements in overall educational outcomes.

## **1.5 Research Questions or Hypotheses**

Our project aims to address specific research inquiries such as, few of discussed below:

1. What is the gender distribution of students across the years?
2. Are there any noticeable trends or changes in gender representation over time
3. How is the age distribution of students across different years?
4. Are there any patterns in the age distribution based on regions or institutions?
5. How does the academic performance vary across different regions (tehsils)?
6. Are there regions consistently performing better or worse than others?
7. What are the most popular subjects among students over the years?
8. How do students from different institutions (Inst\_name) perform in comparison to each other
9. Are there any significant differences in performance between male and female students?
10. Are there any observable differences in performance between urban and rural schools within each tehsil?

## **1.6 Target Audience**

For educators, policymakers, and academics looking for data-driven insights regarding educational disparities, this research is designed specifically for them. The purpose of the findings is to suggest actions that support student achievement.

**1.7 Organization of the Document**

Our project's entire scope is covered in this document's organizational structure. It includes parts which guide readers through the development and consequences of our system step-by-step via the introduction, methodology, dashboard integration, and conclusion.To take readers on a thorough investigation, this report has been thoughtfully prepared. Foundational background data is presented first, followed by in-depth analyses and recommendations that can be put into practice.

## **1.8 Novelty and Uniqueness**

We differentiate our research by exploring multiple dimensions of educational dynamics, such as the differences in gender distribution between institutions, the number of male and female institutions in each district and tehsil, the average grades for each subject, and the average age of exit exam takers. We investigate how student performance varies in both urban and non-urban settings among tehsils, institutes, and governance systems (government or non-government). We also look closely at pass and fail rates on a district and tehsil level as well as pass and fail rates by gender. Our method goes beyond traditional evaluations by examining the popularity of subjects and student choices, offering researchers, educators, and policymakers a comprehensive viewpoint. Our approach provides a broader and more comprehensive perspective than related publications, which focus just on disengagement behaviors or examination issues. The distinctive combination of variables in our study seeks to enable focused interventions to support student performance by offering educators, policymakers, and researchers meaningful insights.

## **1.9 Technical and Non-Technical Goals**

Our technical goals are twofold: firstly, we aim to enhance the dataset through effective data pre-processing techniques and feature engineering, ensuring optimal quality for subsequent analyses. Simultaneously, we aspire to develop a robust machine learning models capable of extracting valuable insights from the educational data. As a non-technical objective, we are committed to translating these advanced capabilities into actionable recommendations for educational practitioners, facilitating informed decision-making. Additionally, our focus is on developing an interactive Streamlit dashboard, providing an intuitive interface for users to explore and comprehend the educational patterns revealed by our analyses.

## **1.10 State of Technology**

In the domain of educational data analysis, modern approaches prominently feature in-depth analysis, machine learning and statistical modeling. With the use of these technologies, we are able to examine complex patterns in the dataset and get a more comprehensive understanding of the dynamics of student performance. As we progress towards our end goal of creating a Streamlit dashboard, these technological advancements will play a pivotal role in crafting an engaging and interactive platform. The dashboard will serve as a dynamic tool, allowing educational stakeholders to visualize, interpret, and draw meaningful conclusions from the rich insights extracted from the dataset.

# **2. Literature Review**

## **2.1 Traditional Methods**

In the past, traditional statistical methods have been widely used in studies on educational data analysis, such as regression analysis to gain detailed understanding of students performance. The use of analytics in education is by no means new, as MacNeill (2012) emphasizes; data gathering, utilization, and sharing are well-established practices in the field.[2] Despite producing ever-increasing amounts of data, very few institutions have up until recently been able to make use of the abundance of information that they regularly get from their main operations, which are teaching and examining. Descriptive statistics and visualizations are used by researchers to find patterns and trends in the dataset, which offers important insights into the dynamics of student performance. This method has been very helpful in highlighting the main variables affecting academic performance.

## **2.2 Advanced Techniques and Models**

Recent advancements in educational data analysis have witnessed a significant shift towards leveraging machine learning models, various tools and by using fuzzy logic.[3] These advanced models provide a more sophisticated analysis of student performance, allowing for the identification of strong relationships and trends. Our project is mainly focusing on in-depth exploratory data analysis to find patterns and trends in the dataset, which offers important insights into the dynamics of student performance. This method has been very helpful in highlighting the main variables affecting academic performance.

## **2.3 Comparative Studies**

The review of multiple research papers emphasizes how important it is to perform thorough exploratory data analysis (EDA) prior to diving into machine learning models[3]. Additionally, the data show a consistent increase in student enrollment in specific STEM fields, which is consistent with these fields being required at some educational levels. But as the results show, there are variations in student performance, which emphasize the need for a more complex understanding of academic outcomes.[4] This is consistent with our project's methodology, which puts an emphasis on in-depth exploratory data analysis before using machine learning models in order to identify patterns and trends in educational outcomes. The adoption of machine learning techniques, such as Random Forest, KNN, and Decision Tree Classifier, is seen as a subsequent step before EDA, [5] so our project will first focusing on exploring the rich information from comprehensive dataset and populate it on dashboard before diving into any machine learning models.

## **2.4 Challenges and Limitations**

Even though exploratory data analysis offers insightful information, issues with data quality concerns, privacy, merging and the changing nature of educational institutions continue to be obstacles. To improve the EDA procedure, guarantee the accuracy of insights, and direct evidence-based decision-making in the field of education, it is imperative to recognize these difficulties.

# **3. Methodology**

## **3.1 Research Design**

We have designed our observational study methodology to identify complex trends in student performance, topic preferences, and institute disparities across various dimensions. Our objective is to investigate the hidden variations in educational results with respect to urbanization, governance kinds (government/non-government), and geographic locations (tehsil/district). This multifaceted strategy guarantees a thorough comprehension of the educational environment.

## **3.2 Data Collection**

The dataset utilized for this project comprises a comprehensive collection of student performance data in secondary school exit exams from the years 2015 to 2023. The collection comprises data from many institutes, including both government and non-government organizations, in order provide an extensive representation.

## **3.3 Data Preprocessing**

We used a multi-step data preparation approach for ensuring the reliability of our study. This included handling missing data by employing mean/median imputation for numerical columns and mode imputation for categorical columns. Outliers were identified and addressed to prevent skewing of results.The dataset included a thorough cleaning process that involved eliminating redundant values and maintaining consistency in column naming conventions.

## **3.4 Feature Engineering**

Feature engineering played a pivotal role in enhancing the dataset's quality and relevance to our research questions. Techniques such as age calculation, subject aggregation, and grouping would be applied. To create a more complete foundation for further analysis, this involves creating features that capture the key elements of student performance and subject preferences particular to government/non-government, urban/non-urban, and tehsil/district contexts. The transformation of features aiming to align the dataset with the study's objectives and ensuring optimal model performance.

# **4. Data Exploration and Analysis**

## **4.1 Dataset Overview**

The dataset utilized for this project comprises a comprehensive collection of student performance data in secondary school exit exams from the years 2015 to 2023. The collection comprises data from many institutes, including both government and non-government organizations, in order provide an extensive representation.

## **4.2 Exploratory Data Analysis (EDA)**

## In our Exploratory Data Analysis (EDA), we look deeply into the dataset using various visualizations to uncover meaningful patterns and insights. Explore EDA as our research activity, where we investigate the perspective that the data is presenting to us.

## **Histograms:**

## Think of histograms as bar graphs showing the distribution of marks in each subject. They help us see how many students scored within different score ranges.

## For example, in English, we can observe if more students scored high or low.

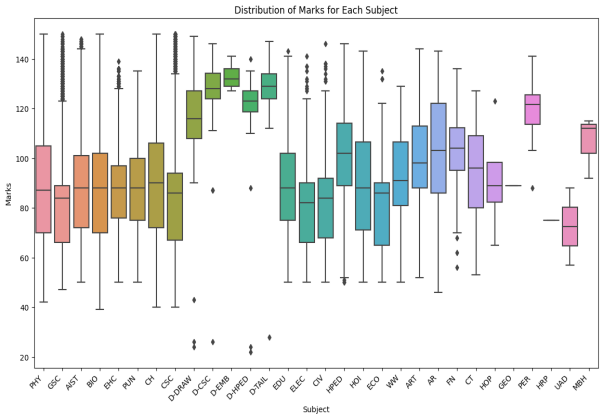
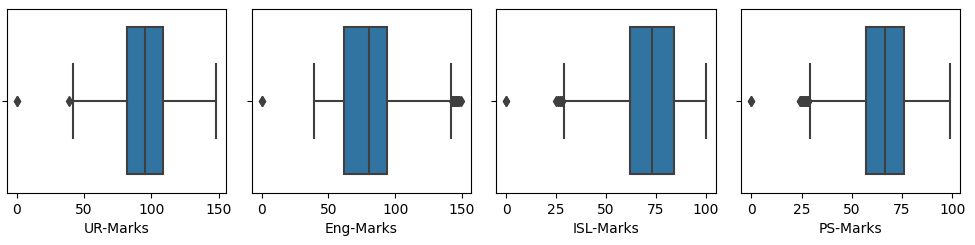
## download (8)

## Fig. 4.2

**Box Plots:**

Box plots give us a summary of the distribution of marks, highlighting key statistics. They help identify outliers – those exceptional cases that stand out.

For instance, a box plot can reveal if there are students with unusually high or low marks in a subject.

**Pie chart:**

The pie chart provides a quick snapshot of the overall success landscape. It helps us understand the distribution of students who conquered the exam and those who faced hurdles. A large slice of Passing indicates a substantial number of students succeeded. A smaller slice of Fail suggests some students encountered difficulties

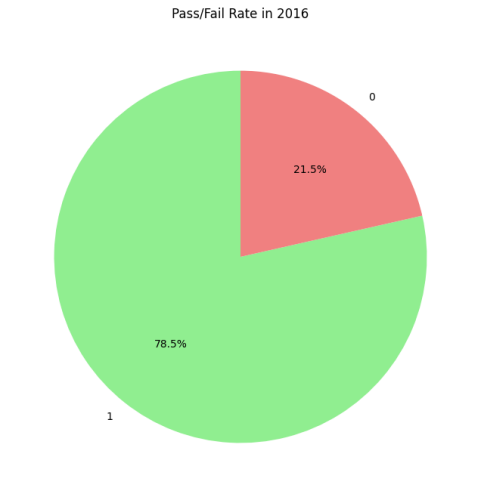


Fig. 4.2

## **4.3 Statistical Summaries**

We used statistical summaries to extract important indicators of central tendency and dispersion in our investigation to identify the most important insights from the educational dataset. Using the "describe()" function, our research offered a thorough overview of the statistical landscape of the dataset. Important statistics for every variable were included, including the count, mean, standard deviation, minimum, quartiles, and maximum values.

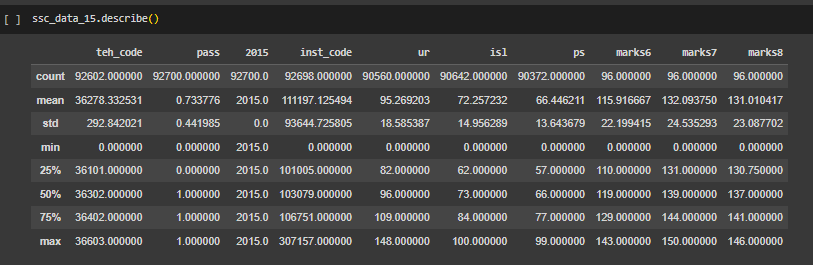


Fig. 4.3

These summaries provide a statistical view through which we could examine subject-wise distributions, variability evaluations, and gender-based performance comparisons. These fundamental statistical findings act as an indicator for us as we move forward with our exploratory journey, pointing us in the direction of deeper features of the educational data environment.

## **4.4 Class Distribution**

The analysis of class distribution focused on understanding success and challenges, identifying strengths and areas for improvement. It highlighted subjects with higher pass rates and those with lower rates, enabling the strategies of interventions tailored to specific needs. During analysis, we identified the subjects with higher pass rates and lower pass rates. We investigated subjects, institutes, and demographics to identify areas of strength and potential improvement.

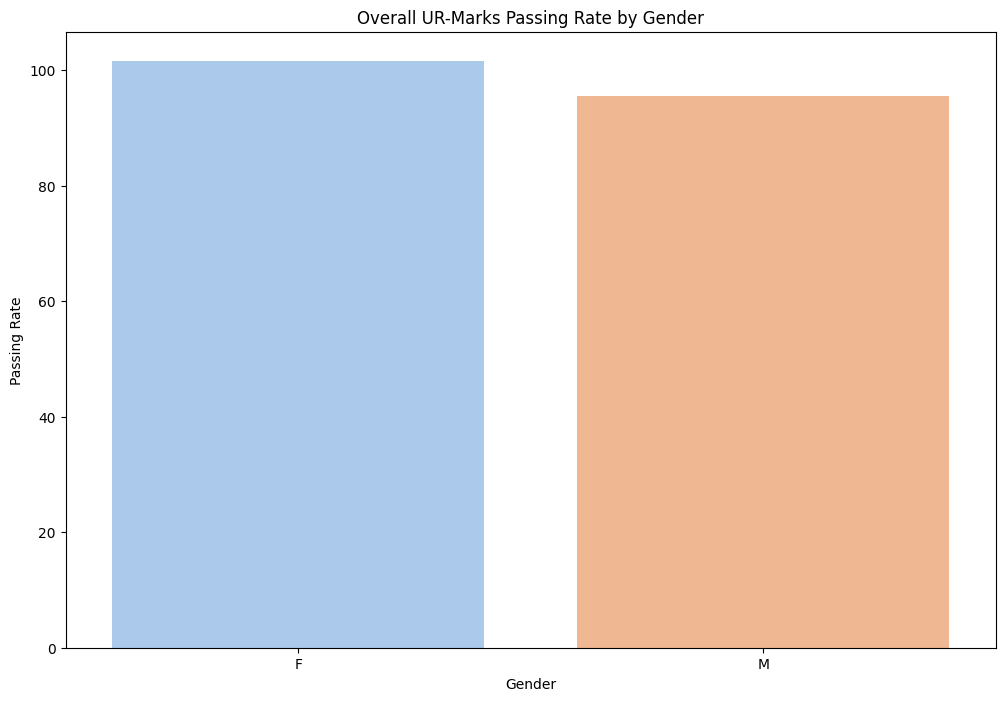


Fig. 4.4

## **4.5 Feature Correlations**

We investigated and planning to examine various correlations between gender, location, governance type, age, subject choice, institute performance, and performance trends over multiple years. It focuses on identifying patterns in student performance, examining the influence of gender on specific subjects, the relationship between geographical and governance types, examining the correlation between age and performance, the correlation between subject choice and overall performance, and examining the correlation between institute performance and student distribution.

# 

# **(These steps of Model development will be filled upon implementation)**

# 5. Model Development

## 5.1 Model Selection Criteria

Clearly define the criteria used for selecting specific models or algorithms. Discuss considerations such as model complexity, interpretability, and suitability for the task.

## 5.2 Model Architecture

Provide a detailed description of the architecture of the selected model(s). Include information on the layers, nodes, and connections in your models.

# 6. Results

## 6.1 Performance Metrics

Present detailed quantitative performance metrics for your models. Include metrics such as accuracy, precision, recall, F1 score, and area under the ROC curve.

# **7. Deployment**

## **7.1 Deployment Strategy**

For deploying our educational performance analysis models, we adopt a web application strategy, leveraging the Streamlit framework. Our data-driven insights may be easily transformed into an interactive and intuitive dashboard with Streamlit's simple method. The deployment will be cloud-based, making use of easily accessible services like Streamlit Sharing or Heroku.

## **7.2 Technology Stack**

Our chosen technology stack for deployment revolves around the following key components:

Streamlit Framework: As the core of our web application, Streamlit facilitates rapid development and deployment of data applications.

Heroku or Streamlit Sharing: Cloud-based platforms that host our application seamlessly. Heroku offers scalability and easy integration, while Streamlit Sharing provides a dedicated platform for Streamlit apps.

## **7.3 Integration Steps**

Integrating our insights into the Streamlit dashboard involves the following steps:

Prepare Streamlit App: Develop the Streamlit app script, embedding visualizations, and interactive components.

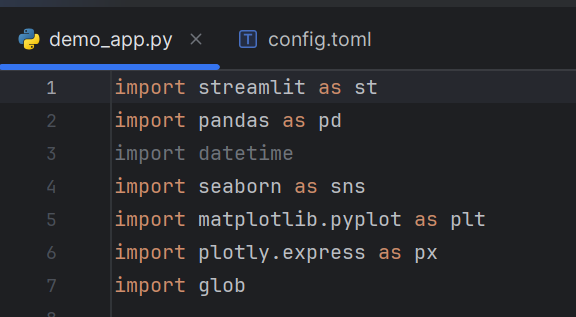


Fig. 7.3

**Main Page of Dashboard**

Below is the demo main page of dashboard that have been created yet, we will adjust the navigation, styling, components accordingly.

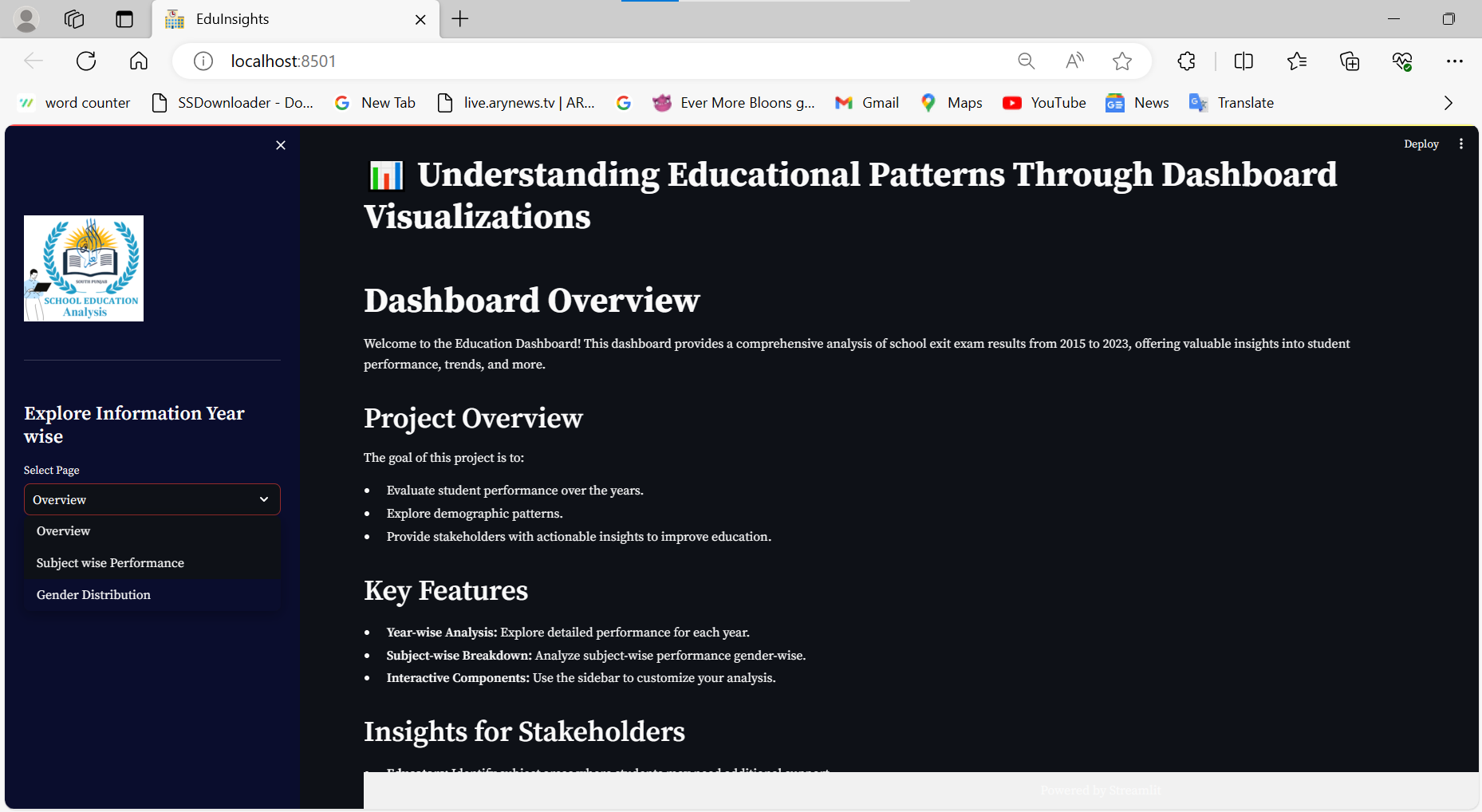


Fig. 7.3

# 

# **8. Conclusion**

Our project is currently in progress. The initial findings are helpful and we are excited to translate these and further insights into user-friendly dashboard. Our objective is to give educators and policymakers useful information to improve the educational system as we proceed with deployment and more analysis.

# **9. References**

|  |  |
| --- | --- |
| [1] | \*. M. A. \*. M. K. DR. SAFDAR REHMAN GHAZI, "CAUSES OF HIGH, AVERAGE AND LOW ACADEMIC," *Journal of Social Sciences,* p. 62, 2009. |
| [2] | 2. A. M. André Krüger1, "A Data Model to Ease Analysis and Mining of Educational data," *3rd International conference on educational data mining Pittsburgh, PA, USA,* Vols. ISBN: 978-0-615-37529-8, no. JUNE 11-13-2010, p. 345, 2010. |
| [3] | A. L. J. T. Anna SVIRINA1, "ANALYSIS OF STUDENTS PERFORMANCE IN RELATION TO THE RESULTS OF STATE UNIFIED EXAM: THE CASE OF RUSSIAN UNIVERSITY," *Business, Management and Economics Engineering,* vol. Volume 19 , no. 1: 170–179, p. 170, 2021 . |
| [4] | S. K. Meenal Joshi1, "Prediction and Analysis of Student Performance in Secondary Education Based on Data mining and Machine learning techniques," *International journal of scientific Research in computer science,* vol. 6, no. 5, pp. 294-301, 2020. |
| [5] | M. F. L.-V. D. F. &. V. C. Fidel Cacheda, "EARLY IDENTIFICATION OF ACADEMIC FAILURE ON HIGHER EDUCATION: PREDICT STUDENT PERFORMANCE USING AI," *Center for Information and Communications Technologies Research (CITIC),* vol. 1, 2023. |