# Project timeline overview:

**Project start date:** October 19, 2023

**Project topic finalization with supervisor:** October 25, 2023

**Project End Date:** July/August, 2024

**Total weeks:** 31 weeks (from October 19 to July 19,2024)

|  |  |  |  |
| --- | --- | --- | --- |
| **Work Package** | **Description** | **Duration**  **(Weeks)** | **Status**  **(Timeline)** |
| **1.  Data Collection** | Gather school exit exam data from the  past 9 years ensuring completeness and accuracy. This involves liaising with educational authorities to obtain data.  **Tasks:**   * Organize data files for years (2015 to 2023) * Perform initial assessment to understand the data | 3 | DONE  **(November 10, 2023)**  **(Nov 29, 23)** |
| **2.Data Exploration, preprocessing**  **AND**  **Initial Dashboarding** | Perform Exploratory Data Analysis (EDA) to understand the dataset’s characteristics.  **Tasks:**  **2.1: Understanding dataset characteristics**  ***(Done: Year 2015 to 2020)***   * Structure of dataset, datatype of each column, summary statistics, perform initial analysis to identify missing values & potential area of interest, set the initial base of dashboard)   ***(Done – Dec 15)***  **2.2: Perform in-depth Analysis:**  **-** Identify key features, patterns and correlations.  **2.2.1: Univariate Analysis:**  ***(Done: Year 2015 & 16)***  - Explore individual attributes, display statistics and check for missing values. Like,  - Identify missing values specially in subject marks columns  - Analyze extent of missing values and impact  - Identify key metrics for Dashboard visualization  **2.2.2: Bivariate & Multivariate Analysis:**  - Explore relationship and patterns among different variables. Like,  - Explore how grades are related to each subject  - Analyze distribution of students’ performance in each subject  - Explore distribution of students across different schools  - Analyze role of gender in grades  - Identify ratio of student gender in each school  - Analyze count of students from urban & rural  - Explore relationship between student gender and their grades  **2.2.3: Group-by or Merging:**  Check the relationship between different variables  **Gender wise Distribution:**   * Gender Distribution across different districts * Gender distribution in urban/rural areas * Gender distribution across each tehsil * Gender distribution in Govt./Non-Govt. schools * Male/Female perform in each subject * Identify where one gender performs the other * Identify subjects where both(M/F) performing good * Ratio of (Male/Female) across different districts/tehsil   **Subject-wise Analysis:**   * Pass/fail ratio in each subject * Identify subjects with higher and lower pass rates * Compare subject wise performance in rural/urban areas * Identify subjects where urban/rural areas performing better * Subject-wise performance in Government/Non government schools * Distribution of Science/Humanity/(Deaf/dumb) groups across district/tehsil/urban/rural areas * Top performing subjects in (Urban/rural) areas * District/tehsil wise (Top performing or low performing subjects) * Identify subjects where one type of school performs better than other * Explore subject wise performance variations across different district and ratio of each gender   **Institute performance (Tehsil/District wise):**   * Analyze performance of each institute * Identify top performing/low performing institutes * Identify tehsil/district of top performing or low performing institute * Identify continuously low performing areas (districts/ tehsil) * Any difference in performance between urban and rural areas * Any difference in performance between government and non-government schools   **Overall Insight:**   * Distribution of institutes across districts * Variation in pass/fail among institutes * Pass/fail ratio changed over years * Identify districts with consistent high and low performance * Identify ratio of Female/Male   And much more, will append accordingly……  **2.3: Initial Dashboard Population:**  - Define features and functionalities for interactive dashboard  - Identify and outline key visualization  **2.4: Data Cleaning and Structuring:**  **(Done: Year-2015)**  - Renaming the columns  - Drop irrelevant values  - Clean the data by identifying missing values, outliers and inconsistencies  - Implement mean/median method on numerical missing column data and mode on categorical data  - Perform further analysis steps based on insights gained during EDA **(Continued……)** | 4 | In-Progress  Expected End:  **(Jan, 2024)** |
| **3. Story telling & Dashboard Designing** | Prepare comprehensive report summarizing key findings from EDA.  **Tasks:**  Design user interface for the dashboard, outlining key visualizations such as performance trends over time (gender wise, year wise, subject wise, etc.) and demographic insights. Define features and functionalities of interactive dashboard. Integrate visualization into dashboard interface. | 4 | **Feb, 2024** |
| **4. Feature Engineering and selection** | Create new feature and transform the existing ones to enhance the performance of machine learning models or improve insights from data.  **Tasks:**  Feature engineering involves (Age calculation, subject aggregation, subjects grouping etc) | 2 | (Not started)  **March, 2024** |
| **5. Data Visualization**  **&**  **Descriptive Analytics** | Conduct descriptive analytics to highlight overall performance metrics including,  **Tasks:**  (subject analysis, yearly trend, gender comparison, correlation insights, geographical patterns) over the past decade.  Explore correlation b/w different attributes and highlight insights and patterns in the dashboard. | 2 | **March, 2024** |
| **6. Predictive Modeling**  **Research** | Explore suitable predictive modeling techniques for forecasting future exam performance. This may involve regression analysis, time series forecasting or various other machine learning algorithms. | 4 | **April, 2024** |
| **7. Predictive Modeling**  **Implementation** | Implement selected predictive model to evaluate its performance and fine tune the accuracy. | 2 | **May, 2024** |
| **8. Integration with Dashboard & Translating EDA** | Integrate predictive model results into the dashboard, allowing user to explore both historical data and prediction. Design components within the dashboard to show predictive model results including (graphs, charts) that effectively communicate insights. | 2 | **May, 2024** |
| **9. User testing,** **feedback**  **&**  **Explore other frame works if required** | Conduct user testing sessions to gather feedback on dashboard’s usability and effectiveness of visualization. Make necessary adjustment based on user feedback. | 4 | **June, 2024** |
| **11. Documentation and**  **Reporting** | Prepare comprehensive documentation, outline the project methodology, data sources, analytics techniques and findings. Create final project report and poster for submission. | 4 | **July, 2024** |
| **Incorporation** | Final adjustment |  | **August** |
| **Total Time required** |  | **31 weeks** |  |

# WHAT IS EDA:

Exploratory data analysis is an approach of analyzing data set to summarize their main characteristics, often using statistical graphics and other data visualization methods.

# Importance of EDA:

* Why this feature is in the data?
* How this feature is correlated to other variable?
* How can these variable impact on output variable.
* Is this feature important doing further analysis on data.
* How well features are related to output.
* How data is distributed over the plane.
* It helps to understanding the relationship of feature.
* Finding the patterns within the data.

# Aim Of EDA:

* Asking the right question related to purpose on data analysis.
* Obtaining in depth knowledge about problem domain.
* Setting clear objective that are aligned with desired outcome

# How to do EDA:

* Step1: Univariate Analysis
* Step2: Bivariate and Multivariate analysis
* Step 3: Data Analysis report preparation based on insights collected from data