**Data Collection and Preprocessing Phase**

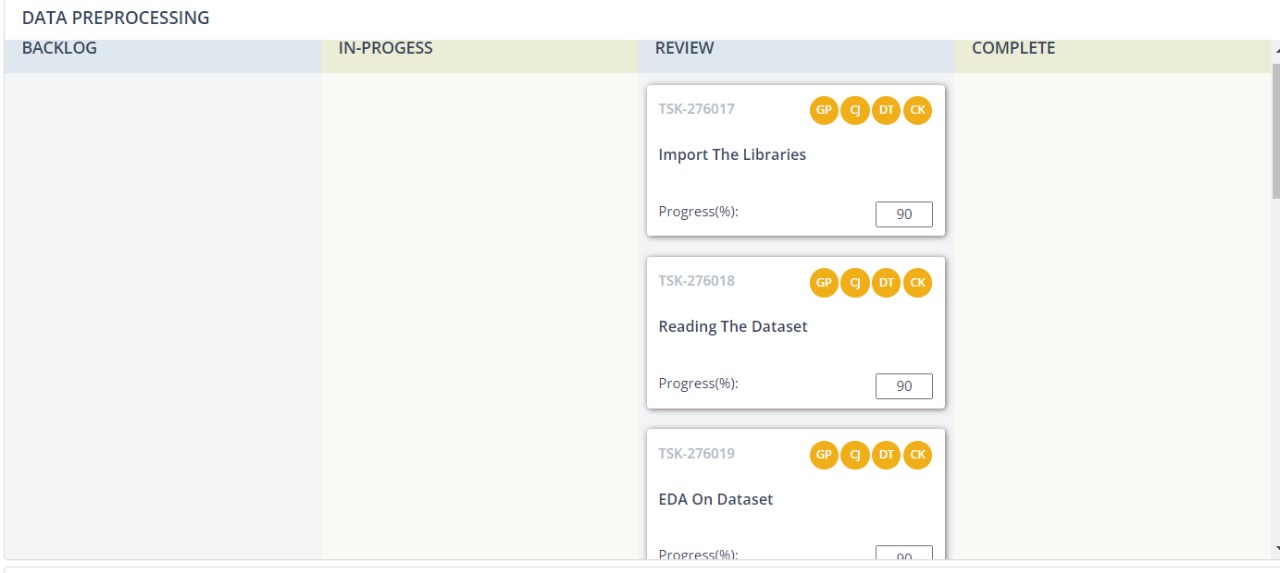
|  |  |
| --- | --- |
| Date | 15 August 2024 |
| Team ID | LTVIP2024TMID24955 |
| Project Title | SMS Spam Detection - AIML |
| Maximum Marks | 6 Marks |

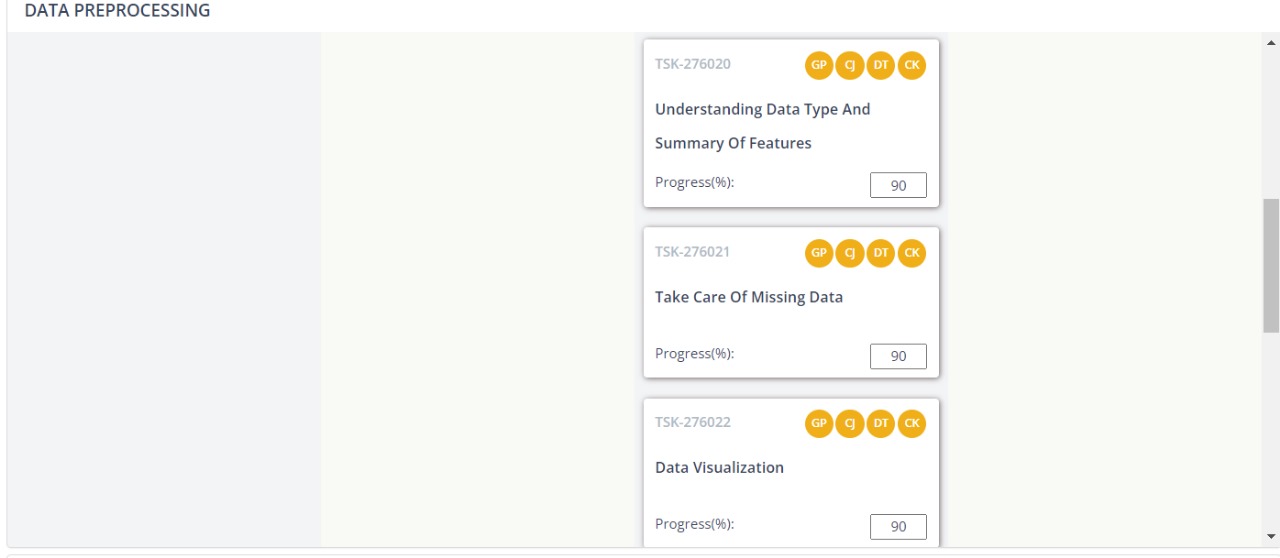
**Data Exploration and Preprocessing Template**

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

|  |  |
| --- | --- |
| **Section** | **Description** |
| 1. Data Collection | Gathering SMS message data from reliable sources such as SMS datasets (e.g., public datasets, user-contributed data, or synthetic data). Includes both spam and non-spam messages for balanced classification.. |
| 2. Data Inspection | Examining the structure of the dataset, including attributes such as message content, labels (spam/not spam), and metadata e.g., time of message, sender info. |
| 3.Exploratory Data Analysis (EDA) | Visualizing and analyzing the distribution of messages, word frequencies, and the relationship between message features (e.g., message length, most common words) and their labels (spam/not spam). |
| 4. Data Cleaning | Removing or correcting noise such as duplicate messages, irrelevant content (e.g., system messages), and non-text characters (special symbols or emojis). Handling missing or incomplete data by filling in, removing, or imputing values. |
| 5. Data Balancing | Addressing any class imbalance between spam and non-spam messages by applying techniques like oversampling |

**Data Preprocessing Code Screenshots**





|  |  |
| --- | --- |
| 6. Text Preprocessing | Processing the raw text of SMS messages by converting to lowercase, removing stop words , punctuation, and stemming/lemmatizing. Transforming the text into a format suitable for machine learning models (e.g., tokenization). |
| 7. Label Encoding | Converting the labels (spam, not spam) into a numerical format (e.g., 0 for not spam, 1 for spam) for use in machine learning models. |
| 8. Data Splitting | Dividing the dataset into training, validation, and test sets to evaluate model performance. |
| 9. Model Building | Developing a machine learning model to classify SMS messages as spam or not spam. This includes selecting appropriate algorithms like Multinomial Naïve base |
| 10.Model Evaluation | Assessing the performance of the trained model using various evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. This involves testing the model on the validation and test datasets |