Technical Solution

This section addresses solution for the problems discussed in the previous section.

System overview

Data polish is a web-based application primarily for students who want to perform data science on dataset. Our application aids by cleaning datasets prior to any machine learning or deep learning techniques are performed on their dataset. This application has very simple user interface which is easy to operate.

Below is a high-level of the system flow of our application:

A diagram of a flowchart

Description automatically generated

Next, we designed a high-level use case diagram as shown below:

A diagram of a data flow

Description automatically generated

The site map diagram attached below shows a hierarchical view of the pages on our web application, starting with the home page at the top.

A screenshot of a computer

Description automatically generated

Data import

To start with the ‘Data Polish’ journey, end user first needs to upload their dataset on the application. In this page, they can upload their file using drag and drop feature. Also, using the ‘+ Choose’ button they can browse their device to upload the file. At present, we are accepting .csv file format only but we have plans of including more file formats in upcoming releases. After choosing the file, user can click ‘Upload’ button to upload it or if they have chosen a wrong file they can click on ‘Cancel’. It will be saved in the cloud and user will get an unique ID (job ID) to track the progress.

A computer screen shot of a cloud

Description automatically generated

Data Profile

In this page, end user will be able to get the current data quality index and full profile of the dataset they have uploaded. This will show some analysis of their data quality which will help to understand what the issues with it are. By clicking on the ‘Missing Values’, ‘Outlier - Scatter’, ‘Outlier – Table’ and ‘Unique Values’, users can check the particular analysis presented in those diagrams which pops up when they click it. We have a plan of replacing this page to an interactive dashboard in next month.

A blue background with a square with black text

Description automatically generated with medium confidence

A screen shot of a graph

Description automatically generated

Data cleaning

In this page, end user will be able to click the button to initiate data cleaning process in the backend. For now, there is only this button on the page, and we are only removing duplicates from the dataset as the cleaning process. In upcoming releases, we will give the user the flexibility to choose which columns and which options of data cleaning they want the application to perform or keep as it as. This feature is referred to as ‘Column Specific Data Cleaning’.

A blue and white screen

Description automatically generated

Export Data

In this page, user can enter the Job ID they received from after uploading the file and check status of it whether that file has been cleaning or not by clicking on the ‘Check Status’ button. If data cleaning process of that file is completed, they will be able to download the cleaned data file by exporting as CSV file.

A computer screen shot of a computer screen

Description automatically generated