1. **Process Modeling**

It is used to visually represent the major processes in a business or system and the data that flows through these processes allowing the organizations to optimize their workflows. It includes flowcharts, data flow diagrams and so on.

A diagram of a software system

Description automatically generated

**Figure 8: Context Diagram**

The user uploads a raw data file in csv format into the system. The user can then select what he wants to do with the data using the UI of the system. The system will perform the desired operation and generate a profile or preprocessed data file or a machine learning model.

A diagram of a company's flowchart

Description automatically generated

**Figure 9: Level 0 DFD of the project**

The user uploads the raw data file through the user interface. He can then select what operation needs to be performed on the data. If the data profile option is selected, the application will generate a data profile for the uploaded data. The preprocessing option will simply clean the dirty data using various methods. The application offers both automated cleaning and manual preprocessing. Finally, the model implementation option will train a model of the user’s choice, tune it and generate a classification report to be displayed. The download function will allow the user to download the model file or the clean data file. The developed system does not use any database for storage purposes.

A diagram of data processing

Description automatically generated

**Figure 10: Level 1 DFD of the process 4**

The automated preprocessing of the raw data file (process 4) is expanded to provide a more detailed description. Initially, the user selects the model which the dataset will fit on. Now, the actual preprocessing begins with the data type detection then any date time data are handled. The missing values are detected and taken care of using various methods. Then, the outliers are handled. The imbalance in the dataset is then checked to prevent biases. It is followed by feature scaling and finally, feature selection is performed. After each of these processes, the dataset is evaluated using supervised machine learning models KNN, SVM and decision tree to generate a score. The score is the average accuracy of fitting the dataset on these models.

A diagram of a diagram

Description automatically generated

**Figure 11: Level 2 DFD of process 4.3**

The missing value processing (process 4.3) is further expanded showing various underlying processes. First, the deletion process is used to simply delete all the missing data. Another process of imputation is applied to the dataset. The effect of both these processes on the dataset are evaluated using the model chosen by the user. The process resulting in higher accuracy is selected and the resultant dataset is forwarded to the next process.

A close-up of a stamp

Description automatically generated

**Figure 12: Level 1 DFD of process 6**

The model implementation (process 6) is also explained to understand the underlying processes. The user is allowed to choose the model of his choice. Upon selection, the application hyperparameter optimization on that model using the uploaded data. Here, the model is trained, tuned, and validated. The classification report of the tuned model is generated and the model itself is saved.

1. **System Flowchart**

A diagram of a flowchart

Description automatically generated

**Figure 13: System Flowchart**

This system divides the machine learning pipeline into three parts and automates each of them. The user can generate a data profile or perform data processing in automated or manual way. Similarly, the user can also develop a model automatically. The system allows the users to download the data file as well as the model file.

1. **Algorithm details**
   1. **Automated data preprocessing algorithm.**
2. Start
3. Import required libraries.
4. Accept the uploaded data file.
5. Accept the model selected by the user on which the data is intended to be used.
6. Detect the data types present in the dataset.
7. Handle the datetime data if it is present.
8. Handle missing data either through deletion or imputation. Evaluate the effect of both approaches to select the one that leads to higher accuracy.
9. The outliers are detected and removed using the IQR method.
10. Check if the dataset is imbalanced. If it is, then handle the problem using the oversampling method SMOTE.
11. Perform feature scaling using normalization. Evaluate the effect of normalization with a condition where no feature scaling is used, to select the one that leads to better accuracy.
12. Perform feature selection using correlation-based feature selection approach. Compare the effect of this approach with one where no feature selection is used to identify the optimum approach.
13. The dataset is finally cleaned using the optimal pipeline for that specific model.
14. The dataset is saved for display and download.
    1. **Manual data preprocessing algorithm.**
15. Start
16. Import required libraries.
17. Accept the uploaded data file.
18. Accept the user input from the interface.
19. All the stages in the automated machine preprocessing are performed here but instead of evaluating after each step, the methods selected by the user are performed.
20. The dataset is cleaned and saved for future use.
    1. **Model building algorithm.**
21. Start
22. Import required libraries.
23. Accept the input data file.
24. Accept the model selected by the user.
25. Load the datafile and split the data into training (70%) and testing (30%) sets.
26. Use the randomized search to train and tune the model.
27. Select the tuned model and test it using the test dataset.
28. Generate classification report for the model.
29. Save the model along with the optimum hyperparameters.
    1. **System algorithm**
30. Launch the user interface.
31. Upload the data file.
32. Click on the operation that the users intend to perform on the dataset.
33. If the data profile button is clicked, the application will generate a data profile for the dataset.
34. If the automated preprocess button is clicked, a new UI will open where the users will be able to select a model and the application will preprocess the data to fit on the selected model.
35. If the manual preprocess button is clicked, a new UI will open with a different input field that would allow you preprocess the data manually.
36. If the model implementation button is clicked, a new UI will open where the user can select the model they want to build. Upon selecting, the application will train, tune, and test a model and generate a classification report.