# DS PROJECT DOCUMENTATION

**Project Name :** Medical Sample Collection Process Streamline.

**Project Batch ID :** DS\_Project\_Team\_69

**Participant Name :** Prahasith Sai Dachapally

**Problem Statement:**

Determining whether the samples of patient collected by the agent for test can deliver to the lab within time or not.

**Survey:**

Collecting the relevant open-source data available on Internet for solving the problem.

**Tasks:**

* Data Pre-processing - You will need to prepare the data set so you can analyze it (e.g. Convert string variables to numerical variables).

• Look at what factors influence with both

a) A continuous measure, and

b) A discrete measure.

• Predict the output using a data-driven / machine learning approach (e.g. Decision trees)

• Present and interpret your findings.

**Assessment Criteria:**

* If you/your team use more than one language (e.g., BI, Python and SQL) or multiple files, please create a zip file containing all the scripts and submit the zip file instead.

• Assessment will be based on the structural approach taken by the team for problem solution including the dashboard (reporting), documentation, and codes developed (optimized).

• Recommended reporting structure to follow: Introduction, Methods, Results, Discussion

**Sol:**

**a) Business Objective:** To determine whether the agent can deliver the samples of the patient to the lab within time or not.

**Business constraints:** Choosing the top most relevant data.

**b) Data dictionary: (Secondary Data Source)**

| **S No** | **Name of feature** | **Data Type** | | **Relevance** |
| --- | --- | --- | --- | --- |
| 1 | Patient ID | Discrete | Nominal | Irrelevant |
| 2 | Patient Age | Numeric | Ratio | Relevant |
| 3 | Patient Gender | Discrete | Binary | Relevant |
| 4 | Test Name | Discrete | Nominal | Relevant |
| 5 | Sample | Discrete | Nominal | Relevant |
| 6 | Way of Storage of Sample | Discrete | Nominal | Relevant |
| 7 | Test Booking Date | Datetime64 | | Irrelevant |
| 8 | Test Booking Time HH MM | Numeric | Ratio | Relevant |
| 9 | Sample Collection Date | Datetime64 | | Irrelevant |
| 10 | Scheduled Sample Collection Time HH MM | Numeric | Ratio | Relevant |
| 11 | Cut-off Schedule | Discrete | Ordinal | Irrelevant |
| 12 | Cut-off time HH MM | Numeric | Ratio | Irrelevant |
| 13 | Agent ID | Discrete | Ordinal | Relevant |
| 14 | Traffic Conditions | Discrete | Ordinal | Relevant |
| 15 | Agent Location KM | Numeric | Ratio | Irrelevant |
| 16 | Time Taken To Reach Patient MM | Numeric | Ratio | Relevant |
| 17 | Time For Sample Collection MM | Numeric | Ratio | Relevant |
| 18 | Lab Location KM | Numeric | Ratio | Irrelevant |
| 19 | Time Taken To Reach Lab MM | Numeric | Ratio | Relevant |
| 20 | Mode of Transport | Discrete | Ordinal | Irrelevant |
| 21 | Reached on Time | Discrete | Binary | Relevant |

**c) Data Preprocessing:**

* Drop the attributes which are irrelevant from the dataset.
* Customize the column names
* Perform Text Mining to get relevant information for corresponding columns
* Map the labels as per the their scores in the given columns of the dataset
* Perform label encoding for the categorical features where ever it is applicable
* Perform all the required operations for the given columns as mentioned in the Sheet2 of the given dataset
* Check for NaN values, if any present perform imputation.
* Check for zero variance & near zero variance features and drop it.

**d) Exploratory Data Analysis (EDA):**

* Perform pair plot for the finalized dataset & identify for co linearity between predictors.
* Used Bar plot for plotting accuracy scores & f1-scores of different models.
* Build correlation coefficient matrix for dataset for checking of co linearity problem.

**e) Feature Engineering:**

* Using "K-Best & Chi2" Algorithm - Gives most significant features with respect to target variable
* Identify the features which are most relevant for model building based on their scores & drop the remaining features from the dataset
* Check the correlation coefficient,(|r|) between input features, if any present (|r| > 0.85) drop those features which has least impact on target variable

**e) Model Building:**

* Classification Model used is:

1. Decision Tree Classifier

* Import the required libraries for model building
* Split the entire dataset into train & test (with test size = 0.3)
* Use the Grid Search CV technique to choose the best hyper parameters in Decision Tree model.
* Run the model using test data & compare the prediction values with actual values of test data
* Check for accuracy scores & f1-scores for goodness of the different models.
* Store the file into pickle format viz., while using it in creating flask API, HTML, CSS.
* Deploy the model & evaluate it’s performance.
* Attachment of python code for model building is available with documentation

**f) Results:**

* **Evaluation Metrics:**

| **S No** | **Model** | **Accuracy** | | **f1-score** | |
| --- | --- | --- | --- | --- | --- |
| **Train** | **Test** | **Train** | **Test** |
| 1 | Decision Tree Classifier with (Grid Search CV) | 1.0 | 0.9673 | 1.0 | 0.97 |

* **Confusion Matrix:**

| **Train Dataset** | **Predicted Values** | | **Test Dataset** | **Predicted Values** | |
| --- | --- | --- | --- | --- | --- |
| **Actual Values** | **No** | **Yes** | **Actual Values** | **No** | **Yes** |
| **No** | 139 | 0 | **No** | 139 | 0 |
| **Yes** | 0 | 574 | **Yes** | 0 | 574 |

* **Highly Significant features:**

1. After performing Correlation coefficient between input features
2. Using "K-Best & Chi2" Algorithm

| **S No** | **Features** | **Score** |
| --- | --- | --- |
| 1 | Time\_Taken\_To\_Reach\_Lab\_MM | 50047.80 |
| 2 | Time\_Taken\_To\_Reach\_Patient\_MM | 2887.70 |
| 3 | Time\_For\_Sample\_Collection\_MM | 381.19 |
| 4 | Scheduled\_Sample\_Collection\_Time\_HH\_MM | 10.99 |
| 5 | Traffic\_Conditions | 7.46 |
| 6 | Agent\_ID | 5.04 |
| 7 | Test\_Name | 3.25 |
| 8 | Patient\_Age | 1.65 |
| 9 | Sample | 0.99 |
| 10 | Way\_Of\_Storage\_Of\_Sample | 0.33 |
| 11 | Test\_Booking\_Time\_HH\_MM | 0.13 |
| 12 | Patient\_Gender | 0.01 |

**Summary:** The above features should be given most importance in order to reach samples of patient within time.