Low Level Design (LLD) Spam Ham Classifier

Geetha S

Akanksha Venkatesh V B

Ashok Kumar S

M C Shravan



# Document Version Control

| **Date Issued** | **Version** | **Description** | **Author** |
| --- | --- | --- | --- |
| **29th September 2021** | 1.0 | Complete Architecture | Geetha S  Akanksha Venkatesh V B  Ashok Kumar S  M C Shravan |
|  |  |  |  |



**Contents**

Document version control 2

**Abstract** 4

1. Introduction 5

1.1 Why this Low-Level Design Document 5

1.2 Scope 5

1.3 Constraints 5

1.4 Risks 5

1.5 Out of Scope 6

2. Technical specifications 6

2.1 Dataset 6

2.2 Predicting Spam or Ham 9

2.3 Logging 9

2.4 Deployment 9

3 Technology stack 9

4 Proposed Solution 10

5 Model training/validation workflow 10

6 User I/O workflow 11

7 Key Performance Indicators 11



# Abstract

The high number of unsolicited emails commonly called spam has necessitated the development of an increase in reliable and robust antispam filters. Using the recent machine learning approach we have been successful in detecting and filtering spam emails. We are classifying Spam or ham from the dataset which is a set of SMS tagged messages that have been collected for SMS Spam research.

# Introduction

## Why this Low-Level Design Document?

The purpose of this document is to present a detailed description of the Spam Ham Classifier. It will explain the purpose and features of the system, the interfaces of the system, what the system will do, the constraints under which it must operate and how the system will react to external stimuli. This document is intended for developers of the system and will be proposed to the higher management for its approval.

The main objective of the project is to classify spam and ham messages.

Our datasets consist of information, such as:

1. Label: spam, ham.
2. Message: spam, ham.

## Scope

This software system will be a Web application. This system will be designed to classify spam or ham messages using Machine Learning Techniques.

## Constraints

Variation in accuracy and limited to the given dataset.

## Risks

Document specific risks that have been identified or that should be considered.

## Out of Scope

Delineate specific activities, capabilities, and items that are out of scope for the project.

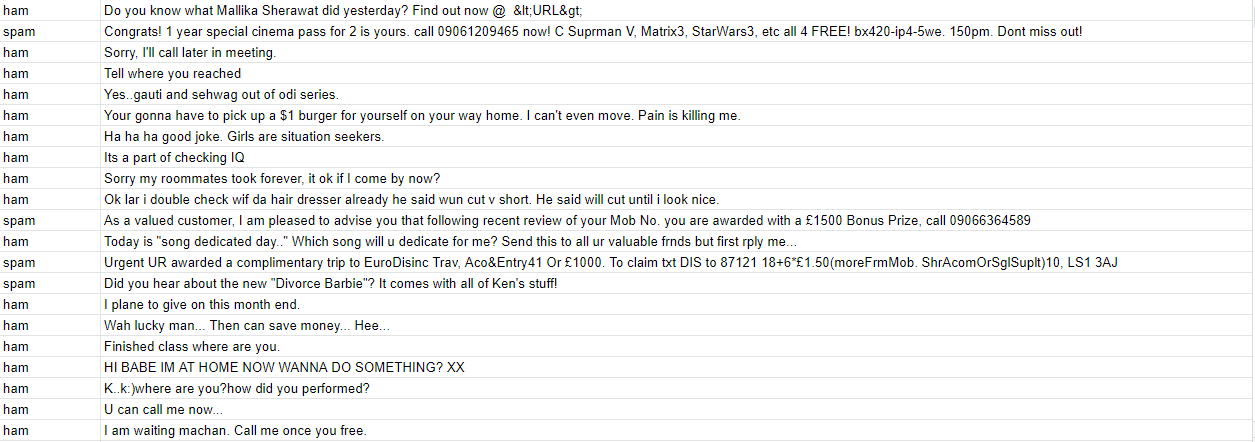
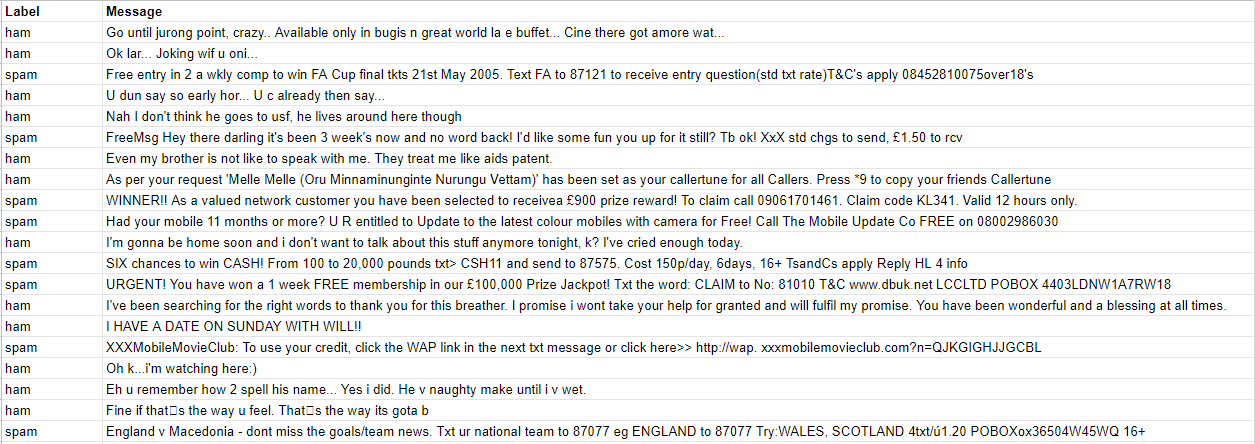
# Technical specifications

## Dataset

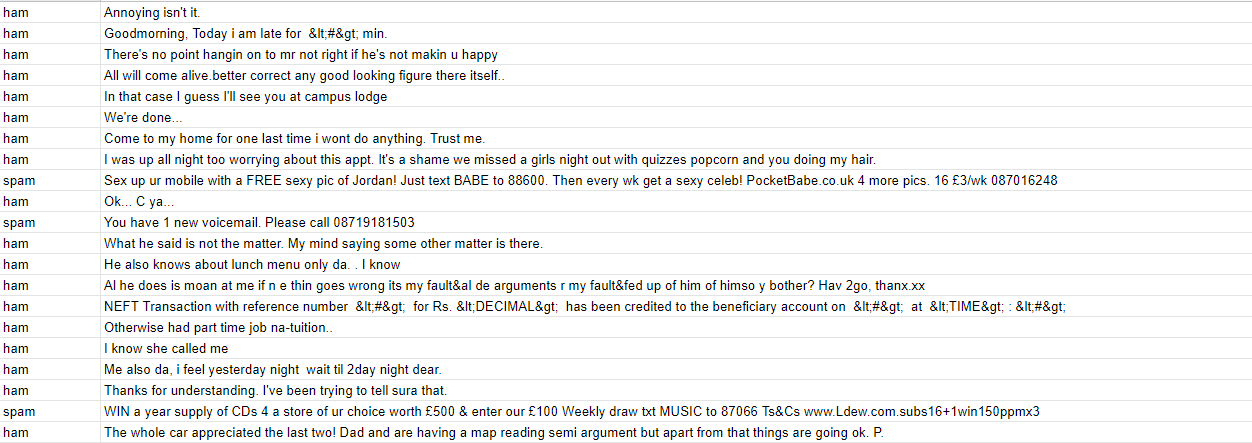
* + 1. **Dataset overview**

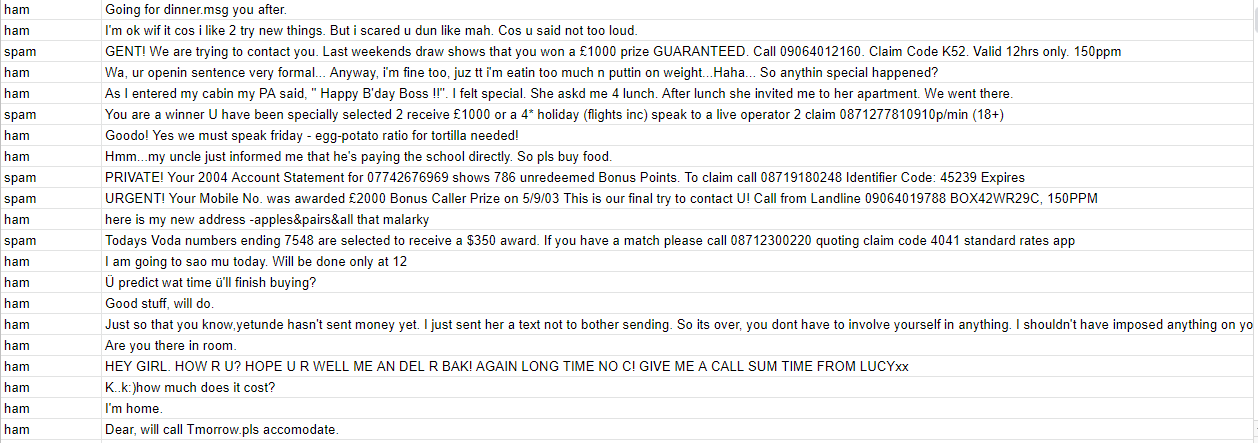
Consists of 2 different tables.

We have a train (5284) and test (1322) datasets, We need to classify Spam or ham from the dataset which is a set of SMS tagged messages.











* + 1. **Input schema**

| **Feature name** | **Datatype** | **Null/Required** |
| --- | --- | --- |
| Message | Object | Required |
| Label | Object | Required |



## Predicting Spam or Ham

* The system displays the output based on entered input.
* The User selects the message.
* The system helps to provide the set of inputs required from the user.
* It gives the required information.
* The system should be able to classify spam and ham.

## Logging

We should be able to log every activity done by the user.

* The System identifies at what step logging required
* The System should be able to log each and every system flow.
* Developers can choose logging methods. You can choose database logging/ File logging as well.
* The system should not be hung even after using so many loggings. Logging just because we can easily debug issues so logging is mandatory to do.

## Deployment

1. Heroku



# Technology stack

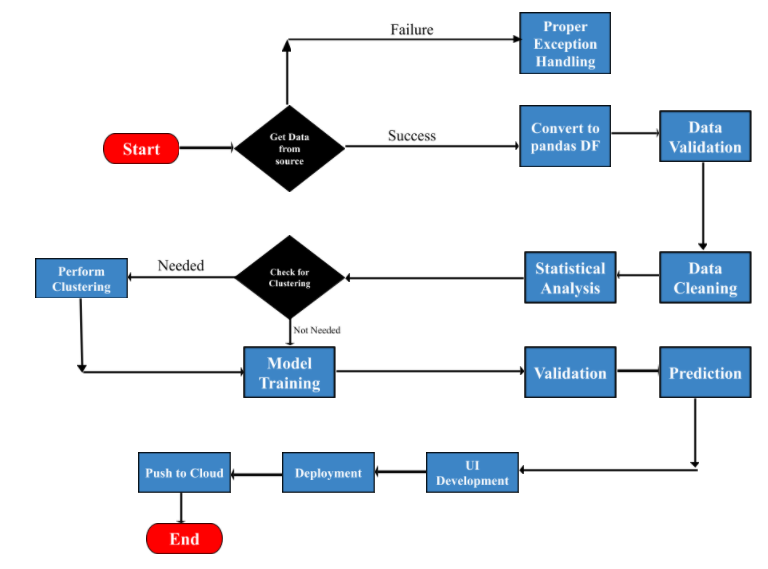
| **Front End** | HTML/CSS |
| --- | --- |
| **Backend** | Python Flask |
| **Deployment** | Heroku |
| **Database** | MySQL |



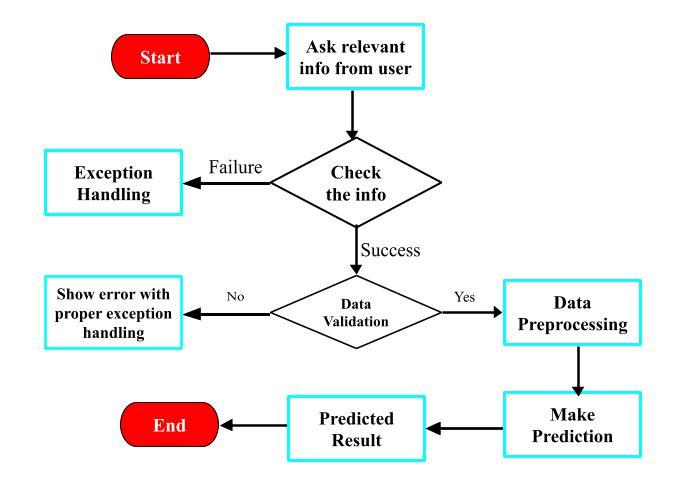
1. **Proposed Solution**

The solution proposed here is a Spam Ham Classifier which will help to classify Spam or ham from the given dataset. Here we have used Logistic regression, Extra trees classifier, Adaboost classifier, XGBoost classifier and Gradient boosting classifier.

# Model training/validation workflow



1. **User I/O workflow**



# Key performance indicators (KPI)

* The output will be provided within seconds
* Provides good accuracy.
* Easy to access.
* The classification of Spam and Ham can be done efficiently.

# Latency

# The output for the given input will be provided within seconds after clicking on the submit button.