

# Spam Ham Classifier

A series of several thin, white, parallel diagonal lines extending from the bottom left towards the top right of the slide, adding a modern, geometric design element.


## Objective:

Classification of spam or ham from the provided dataset. This classifier will predict if the particular messages is a spam or ham.

## Benefits:

- ❑ Protection against Malware.
- ❑ Reduces the risk of users clicking on something they shouldn't.
- ❑ Keeps hackers at bay.
- ❑ Phishing scams that attempt to get information can be avoided .

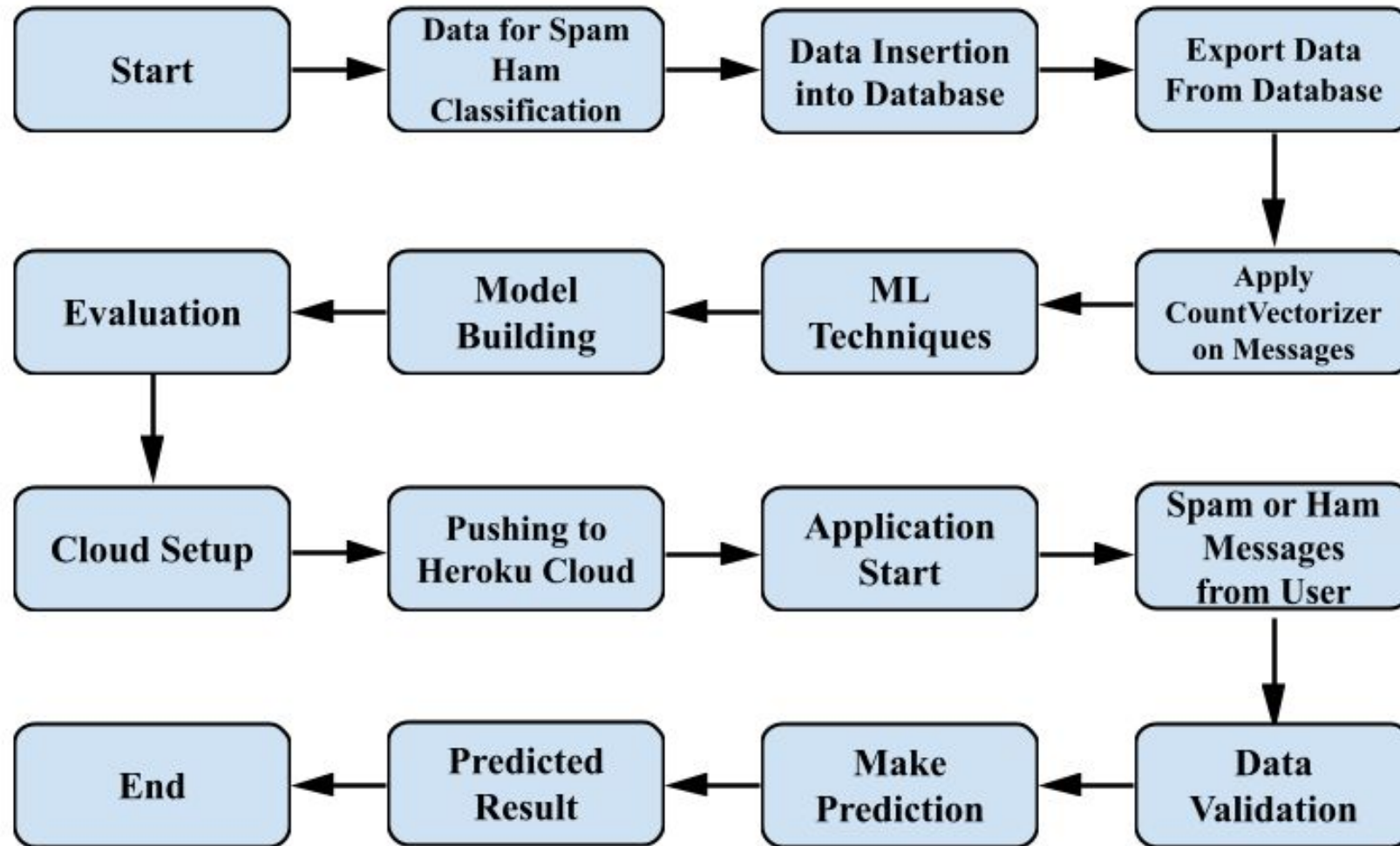
## Data Sharing Agreement :

- ❑ Sample file name (Spam\_Ham.txt)
  - ❑ Number of Columns
  - ❑ Column names
  - ❑ Column data type
- 
- A series of several parallel white diagonal lines of varying lengths and positions, located in the bottom right corner of the slide, creating a modern, abstract graphic element.

# Data Schema

Feature name	Datatype	Null/Required
Message	Object	Required
Label	Object	Required

# Architecture



# Architecture Description

## Data Description

The collection is composed by just one text file, where each line has the correct class followed by the raw message. The identification of the text of spam messages in the claims is a very hard and time-consuming task, and it involves carefully scanning hundreds of web page's. The 425 SMS spam messages are manually extracted from the Grumble text Web site and a subset of 3,375 SMS have been randomly chosen as the ham messages of the NUS SMS Corpus (NSC), which is a dataset of about 10,000 legitimate. These messages were collected from volunteers who were made aware that their contributions were going to be made publicly available.

## Data Insertion into Database

- a) In this project we have used MySQL database.
- b) spam\_ham database has been created
- c) Using that database we have created a table called as data.
- d) For that table we have inserted data.

## Export data from Database

- a) We connect to spam\_ham database.
- b) From that database we export the data

## **Apply CountVectorizer for messages**

Based on the problem statement and requirements we apply CountVectorizer to transform a given text message into a vector on the basis of count of each word that occurs in the text

## **Machine Learning Techniques**

Based on the problem statement and requirements we can use supervised or unsupervised technique which fits the project.

## **Model Building**

Depending on the data type of the target variable we are either going to be building a classification model. The main aspect of machine learning model building is to obtain actionable insights and in order to achieve that it is important to be able to select a subset of important features from the vast number.

## Evaluation

The Evaluation of accuracy can be done using the test data. Confusion Matrix, Accuracy Score and Classification Report can be found using test data and Prediction data.

## Cloud Setup

The cloud deployment platform this platform is setup for deploying the virtual app.

## Pushing to Heroku Cloud

Pushing workloads to the cloud requires significant planning, testing, expertise so once the cloud is setup, the virtual app created will be pushed to the cloud and will finally be deployed into the cloud .

## Application Start

Once the virtual app is deployed in to the cloud we can open the web application using any web browser



## **Spam or ham from user**

Using a web browser we open the web application and provide the necessary information as the input for prediction.

## **Data Validation**

Once the input is provided and we click on the submit button, the system will provide the output based on its requirements.

## **Result Prediction**

Once the data validation is completed the prediction will be done for the type of product in Stores and Big Marts provided in the input.

## Q & A

Q1) What's the source of data?

The data for training is provided by the client in the form a link which takes us to the site containing the datasets.

Q 2) What was the type of data?

The data is of Categorical values.

Q 3) What's the complete flow you followed in this Project?

Refer from slide 6 for better Understanding

Q 4) How logs are managed?

We are using different logs as per the steps that we follow in validation and modeling like Data Validation log ,Info log, Error log , Data Insertion ,Model Training log , prediction log etc.

Q 5) What techniques were you using for data pre-processing?

- ▶ Removing unwanted attributes
- ▶ Visualizing relation of independent variables with each other and output variables
- ▶ Checking and changing Distribution of continuous values
- ▶ Converting categorical data into numeric values.
- ▶ Transforming the data based on the program requirements.

Q 6) How training was done or what models were used?

- ▶ Before diving the data in training and validation set we performed data pre-processing, exploratory data analysis and feature selection.
- ▶ Based on the client given dataset, the training and validation data were divided.
- ▶ The label encoder and one hot encoding was performed over training and validation data
- ▶ Algorithms like Logistic Regression, Extra Tree Classifier, Gradient Boosting Classifier, Random Forest Classifier, XGBoost Classifier and K-neighbours Classifier were used and we saved that model .

Q 7) How Prediction was done?

The test data was shared by the client .We performed pre processing, EDA for given test data, then with this test data, prediction was performed. In the end we get the accumulated data of predictions.

Q 8) What was the platform used for deploying the project ?

We used a cloud servicing platform named Heroku for deploying the project into the cloud.

Q 9) What are the advantages of the platform used for deploying the project?

- ▶ The advantages of Heroku are:
- ▶ It is free of cost.
- ▶ It is easy to use.
- ▶ Developer Centric.
- ▶ Easy to scale.
- ▶ Provides security.
- ▶ Powerful CLI

Q 10) What was the Framework used for doing the backend?

We used Flask Framework for completing the backend.

Q 11) How were the errors removed from the program?

There were no errors left by the end of the program execution because all of the errors were identified and debugged during the execution of the program itself.

Q 12) What is the future scope of the project?

- ▶ Use multiple algorithms.
- ▶ Optimize flask app.py and Stores Sales Prediction.ipynb
- ▶ The front end can be developed even more.

Q 13) Which Database is used ?

MySQL.

