■ Spam Classifier

A machine learning project that classifies emails and SMS messages as **Spam** or **Ham**. Built using **Python, scikit-learn, and Streamlit**, with my own reusable toolkits for data cleaning and PCA.

■ Features

- Trains on the SMS Spam Collection Dataset (spam.csv).
- Text preprocessing using TF-IDF vectorization with uni- and bi-grams.
- Logistic Regression classifier with balanced weights for imbalanced data.
- Interactive Streamlit app (app.py) for real-time spam detection.
- Shows predictions with confidence scores and probability bars.
- Includes personal toolkits:
- • data_clean.py: custom DataCleaner (DC) class for NaNs, duplicates, text cleaning, and exports.
- • my_Pca.py: custom MY_PCA class for scaling, dimensionality reduction, and explained variance.

■ Project Structure

■■■ app.py # Streamlit app for predictions ■■■ model_.py # Training script with pipeline & evaluation ■■■ spam.csv # Dataset (SMS Spam Collection) ■■■ spam_model.pkl # Saved trained model ■■■ data_clean.py # Personal data cleaning toolkit ■■■ my_Pca.py # Personal PCA toolkit ■■■ README.md # Documentation

■■ Installation

Clone the repository and install dependencies: git clone https://github.com/harshalj786/Spam-classifier.git cd Spam-classifier pip install -r requirements.txt

■■ Usage

1. Train the Model python model_.py This will load the dataset, train the classifier, and save the model. **2. Run the Web App** streamlit run app.py Paste any email/SMS into the text box and click Classify.

■ My Toolkits

data_clean.py (DC class) - View & drop NaNs - Fill missing values with mean, median, mode, or custom rules - Clean text (remove duplicates, links, etc.) - Standardize column names - Export cleaned datasets **my_Pca.py (MY_PCA class)** - Scale features automatically - Apply PCA with customizable components - Return results as a Pandas DataFrame - Retrieve explained variance

■ Future Improvements

- Add deep learning models (LSTMs, Transformers).
- Build multilingual spam detection.
- Add explainability (highlight words that triggered 'spam').
- Deploy as an API with FastAPI or Flask.

■ Credits

- Dataset: SMS Spam Collection Dataset - Libraries: Python, scikit-learn, Streamlit, pandas, numpy
- Designed and developed by **Harshal**