Risk analysis of medical prescriptions(data driven pipeline in python)



Gayathree Basarahalli

EC Utbildning

Python fördjupad

202509

**Abstract:**

# This project focuses on the extraction, analysis, and validation of data from images of medical prescriptions. Traditional manual processing of prescriptions is time-consuming and inefficient, especially when dealing with large volumes. The system leverages Optical Character Recognition (OCR), Regular Expressions (Regex), and Machine Learning (ML) models to automate the reading, categorization, and validation of prescription data.

# Using Python and Streamlit, the application allows users to upload prescriptions, automatically detect key fields, and generate structured reports efficiently. The solution significantly reduces manual workload, improves accuracy, and ensures compliance with healthcare documentation standards.

**Introduction:**

# Medical prescriptions are critical documents that provide detailed instructions on patient medications. Manual processing of these prescriptions is often labour-intensive, prone to human error, and difficult to maintain at scale.

# The objectives of this project are:

1. To implement OCR technology and Regex for accurate extraction of fields of each prescription.
2. To use Machine learning model to validate risk level (high, medium, low).
3. To generate a structured report for health care use.
4. Develop user friendly application for automated image processing.

This project is significant as it improves operational efficiency, ensures higher accuracy, and enables faster processing in medical institutions. The system is designed with scalability in mind, making it adaptable to various document formats and languages.

**Project Process:**

**Data Collection:**

* Gathered a dataset of medical prescriptions from different people in image format.
* The dataset consisted entirely of printed prescriptions.

**Data Processing:**

* Converted images into text using OCR.
* Applied image enhancement techniques to improve OCR performance.
* Removed unnecessary columns from the data set using pandas.
* For a more robust solution, filled missing data with random appropriate values.

**Information Extraction:**

* Utilized Tesseract OCR to extract text from prescription images.
* Applied Regex patterns to identify medication names, dosages, frequency, and patient details.
* Saved all the data in Excel and Json our further processing.

**Machine learning model:**

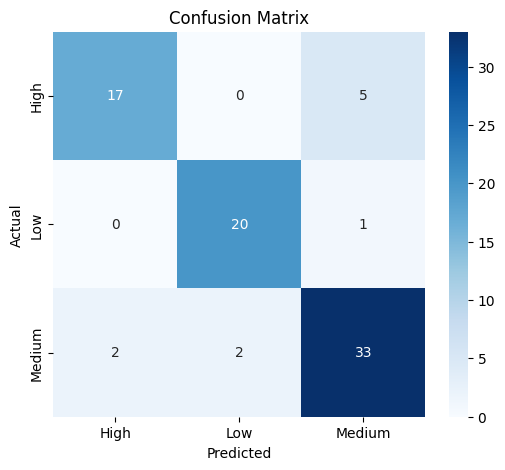
* Trained data to analyse risk based on the patient information.
* Evaluated models using accuracy, precision, recall, and F1-score.

**Streamlit:**

* Developed a Streamlit based interface allowing users to upload prescription files.
* Displayed extraction information in a excel format.
* Provided risk analysis based on patient data.

**Results:**

**Confusion Matrix**:



**Streamlit:**

****

A screenshot of a computer

AI-generated content may be incorrect.

A red rectangular object with text

AI-generated content may be incorrect.