Disease diagnosis using classification and NLP

Rebecca Ceppas de Castro, Fulya Tastan, Philip Barron, Mohammad Rafiqul Islam, Nina Adhikari, Viraj Meruliya
Github: https://github.com/nina-adhikari/disease_prediction

Project Overview: Automatic Symptom Detection (ASD) and Automatic Diagnosis (AD) have seen several advances in recent years. Patients and medical professionals would benefit from tools that can aid in diagnosing diseases based on antecedents and presenting symptoms. The lack of quality healthcare in many parts of the world makes solving this problem a matter of utmost urgency. The aim of this project is to build a tool that can diagnose a disease based on a list of symptoms and contribute to our understanding of automatic diagnosis.

Dataset: https://figshare.com/articles/dataset/DDXPlus_Dataset/20043374

- Very large synthetic dataset with over 1 million samples spanning 49 unique pathologies.
- Contains information about patient symptoms, their antecedents and socio-demographic data, a true diagnosis, and a differential diagnosis of their underlying conditions.
- We chose to focus on the following ten diseases:
  - **Infectious Diseases**: HIV (initial infection), Whooping cough, Chagas disease, Tuberculosis, Influenza
  - **Autoimmune and Inflammatory Conditions**: SLE (Systemic Lupus Erythematosus), Sarcoidosis
  - **Allergic Reactions and Related Conditions**: Anaphylaxis, Allergic sinusitis, Localized edema

Stakeholders: Medical professionals, epidemiology experts, health organizations (e.g. WHO and CDC), data scientists and ML researchers, and end users (patients).

Key Performance Indicators (KPIs)

- *Precision*: % of predicted positives that are true positives,
- *Recall*: % of true positives that are predicted as positive, and
- F_1 *score*: harmonic mean of precision and recall; we want our model to correctly identify patients with a certain disease and also be confident in our prediction at the same time.

Modeling

- We built several multiclass classification models that identify the disease based on symptoms and antecedents. For final model selection, we adopted the following strategy for training and evaluation:
  - Features: We used _ number of features including (initial evidence, pain levels, travel details,...)
  - Training: The original had its own train, validation, and test split. To randomize the splits for our models, we combined the training and validation datasets, and then performed an 80% (train) - 20% (validation) split. The different models were trained and validated respectively on these datasets.
  - Test: The best model with its optimal hyperparameters was evaluated on the unseen test data
- We experimented with an alternative approach using natural language processing (NLP), where the data was used to generate 1-3-sentence-long paragraphs of text, and a new dataset was prepared with this text and the disease label. A DistilBERT transformer was fine-tuned on this dataset, and the fine-tuned model was evaluated on the test set.

Results and Outcomes

- Random Forest was found to be the best model, with the following scores:
  F_1: **59.58%** Precision: **75.83%** Recall: **59.04%** Accuracy: **60.40%**
  A similar performance was also achieved by XGBoost. We chose Random Forest since it is simpler and more interpretable, which would be useful to stakeholders. The feature importance function of Random Forest shows that 'INITIAL_EVIDENCE' is the significant input in classifying the disease.
- The fine-tuned text classification transformer achieved an accuracy of **58.68%** on the test set.
- We made a web app that can be used to interact with the models, available at disease-pred.streamlit.app.

Future Directions

- Using more datasets to improve the model performance.
- Building classification models for other diseases in the current dataset.
- Enhancing our app with a chatbot that patients can interact with directly.