Georgia State University

CSC4780/6780 and DSCI4780 – Fundamentals of Data Science

Fall 2023

Final Project Report

[Disease/Illness Prediction]

Jay Sompalli
Siya Katoch
Rithwik Kamalesh
Jeet Arora

Table of Contents

1 Business Understanding

1.1 Business Problem

The current business challenge revolves around the high volume of appointments faced by physicians, leading to longer patient wait times and potentially reducing the quality of patient care. The inefficiency of the existing system is evident, as physicians often lack sufficient preparation time for each visit, resulting in an unpleasant patient experience. This situation not only impacts the overall satisfaction of patients but also places a significant burden on physicians, hindering their ability to deliver personalized and effective healthcare.

Our objective is to create a solution that enables physicians to be better prepared for appointments. By implementing a machine learning-driven approach that provides physicians with information about patients' potential diagnoses when they arrive, we aim to streamline the appointment process while improving the overall patient experience by reducing wait times, increasing efficiency, and allowing physicians to use their time more effectively.

1.2 Dataset

This report investigates a Kaggle-sourced dataset comprising 4920 rows and 134 columns, with a focus on health-related information. The dataset features 132 binary symptom columns and one categorical prognosis column, encompassing 41 unique diseases. An intriguing aspect of the dataset is a column with entirely missing values, posing a challenge for analysis. Strategies for handling this missing data will be explored to ensure the effectiveness of analysis. The binary symptom variables provide insights into the presence or absence of 132 distinct symptoms, offering a detailed understanding of health conditions. The categorical prognosis column introduces a predictive modeling element, aiming to unveil relationships between symptoms and disease likelihood. Objectives include descriptive analysis, data cleaning to address missing values, feature engineering to enhance predictive modeling, and disease prediction using machine learning algorithms. The study aims to derive meaningful insights, potentially contributing to early diagnosis and targeted interventions. The unique dataset structure and missing values challenge emphasize the need for a meticulous approach to extracting valuable conclusions.

1.3 Proposed Analytics Solution

Through the application of data science techniques, our mission is to use the gathered patient data to construct a predictive model. This model, using machine learning algorithms, is designed to identify potential illnesses based on the presented symptoms with high accuracy. Our main goal is to create a strong prediction system that can accurately determine the likelihood of an illness.

The technology we aim to create, we envision, will be used through a chatbot system. Rather than aiming for a definite diagnosis, our approach is to provide physicians with a baseline assessment before the patient's visit. The insights generated by our predictive model can equip healthcare professionals with crucial information, allowing them to be better prepared for the upcoming patient visit. By offering a baseline understanding of the patient's potential health issues, the model enhances the efficiency of the diagnostic process and enables a more informed and personalized approach to patient care. This innovative integration of technology aims to streamline healthcare practices, fostering a proactive and collaborative environment for improved patient outcomes.

2 Data Exploration and Preprocessing

2.1 Data Quality Report

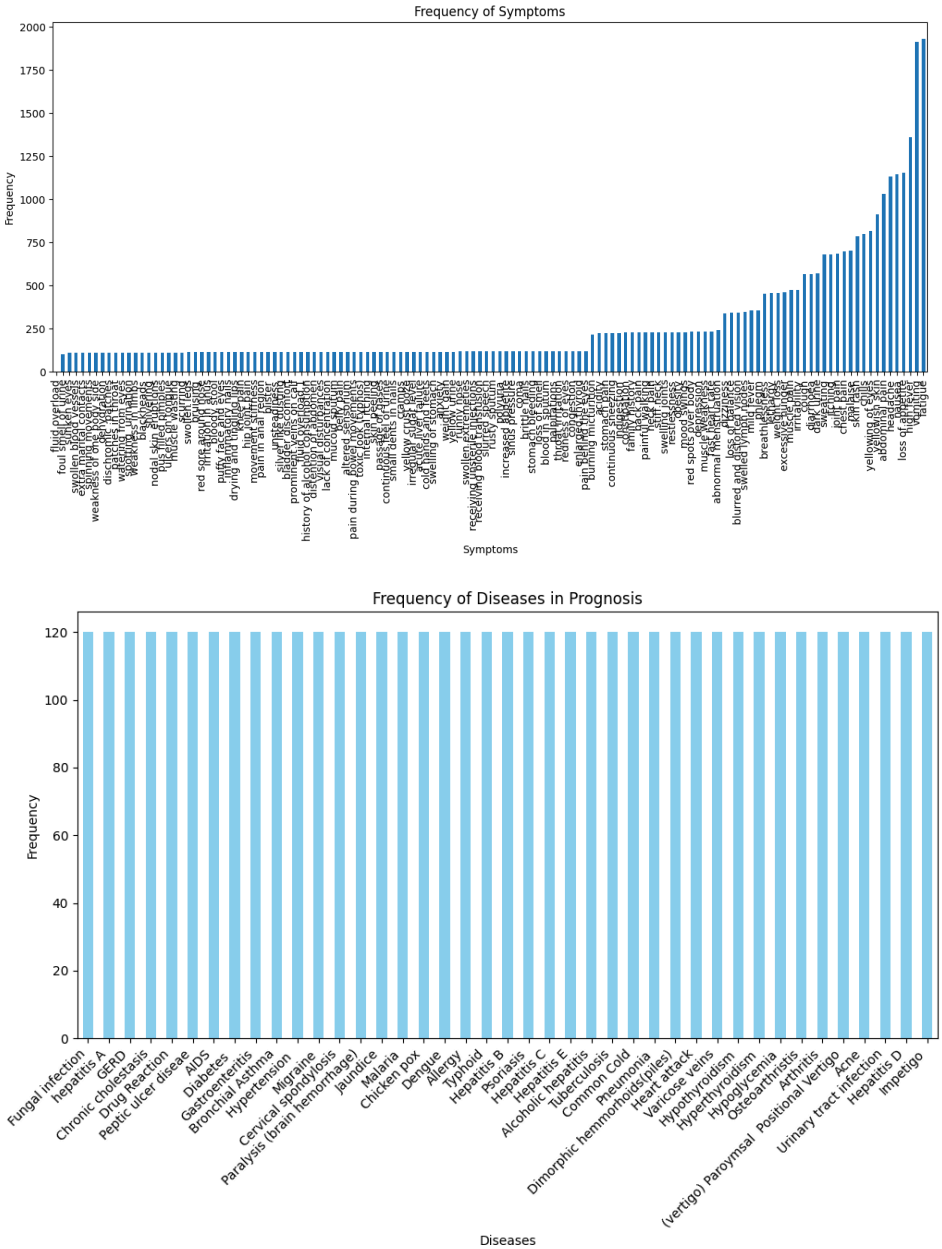
The dataset used for developing the disease/illness prediction model can be considered reasonably good quality. It contains 4920 rows and 134 columns, with each row representing a patient case and each column representing a binary symptom or the categorical prognosis. Some data cleaning was performed, including substituting underscores with spaces in symptom names and encoding categorical features. No major data quality issues are reported, though one column with all missing values was dropped. The models built on this data, including Random Forest and Logistic Regression, were able to achieve exceptional performance with accuracy scores of around 0.99. This indicates that the dataset, though not perfectly clean, contains useful signals and predictive power to classify illnesses based on symptoms. As more data is collected over time, continued monitoring of data quality and augmentation of the dataset could further improve model performance. Overall, the current dataset provides a valuable foundation for developing accurate disease prediction models.

Data Quality Report:

- 132 Symptoms
- No Empty Values
- 41 Illnesses
- No Empty Values

```
array(['Fungal infection', 'Allergy', 'GERD', 'Chronic cholestasis',
       'Drug Reaction', 'Peptic ulcer diseae', 'AIDS', 'Diabetes ',
       'Gastroenteritis', 'Bronchial Asthma', 'Hypertension ', 'Migraine',
       'Cervical spondylosis', 'Paralysis (brain hemorrhage)', 'Jaundice',
       'Malaria', 'Chicken pox', 'Dengue', 'Typhoid', 'hepatitis A',
       'Hepatitis B', 'Hepatitis C', 'Hepatitis D', 'Hepatitis E',
       'Alcoholic hepatitis', 'Tuberculosis', 'Common Cold', 'Pneumonia',
       'Dimorphic hemmorhoids(piles)', 'Heart attack', 'Varicose veins',
       'Hypothyroidism', 'Hyperthyroidism', 'Hypoglycemia',
       'Osteoarthristis', 'Arthritis',
       '(vertigo) Paroymsal  Positional Vertigo', 'Acne',
       'Urinary tract infection', 'Psoriasis', 'Impetigo'], dtype=object)
```

Figure 1. Visualizations of Categorical Features



2.2 Missing Values and Outliers

In addressing missing values, our data preprocessing involved the removal of unnamed columns that contained solely blank entries, resulting in a dataset with 133 rows. This reduction ensures that our analysis focuses on only important information, enhancing the overall quality of the dataset and the predictions given by the model.

Regarding outliers, our dataset only comprises binary and categorical variables. Unlike numerical datasets where outliers can significantly skew results, the nature of our data minimizes concerns related to outliers. Binary and categorical variables represent distinct states and resist the presence of outliers. Based on this outlier detection and removal techniques were unnecessary. In the context of our specific analysis, it is important to employ other cleaning techniques which are covered in the “2.4 Transformations” section of the report.

2.3 Normalization

Due to our dataset's binary and categorical variables, traditional normalization methods were unnecessary. Binary variables were encoded as 0s and 1s, representing "false" and "true" states. The absence of continuous numerical data in these features eliminated the need for normalization techniques. Our data preprocessing prioritized the preservation of categorical information through encoding. This tailored approach ensures the relevance and accuracy of subsequent analytical phases.

2.4 Transformations

In the transformation phase of our data preprocessing, we prioritized readability in data visualizations. We substituted underscores with spaces in symptom names, improving overall clarity. Categorical features transform through encoding, converting them into numerical forms to optimize model functionality. This not only helped the readability of the dataset but also allowed machine learning algorithms to effectively interpret categorical information. Additionally, we introduced a new column dedicated to encoding prognosis data. These transformation steps collectively enhance the dataset's accessibility and tailor it for advanced predictive modeling, allowing the dataset to be more usable for our purpose.

2.5 Feature Selection

In our feature selection strategy, we chose to incorporate all available symptoms for illness classification. This decision stems from the recognition that seemingly less significant symptoms can hold value when diagnosing an illness, especially in healthcare, where seemingly unimportant symptoms can be important. Unlike other approaches that give priority to traits that occur frequently,

ours does not ignore symptoms that, while less prevalent on their own, may together provide important information for precise classification. This aims to prevent potential biases towards more prevalent symptoms. By using all the symptoms, we aim to create a classification that incorporates all the possible symptoms.

3. Model Selection and Evaluation

3.1 Evaluation Metrics

In evaluating our models, we chose a diverse set of metrics. The confusion matrix provides a detailed breakdown of predictions, which is crucial in healthcare for reducing false positives and false negatives. The classification report, including precision, recall, F1 score, support, accuracy, macro average, and weighted average, offers a greater understanding of the model's performance. Precision shows the avoidance of false positives, recall shows the capture of all instances of a class, and the F1 score shows the overall performance of our model. Support reveals class distribution and accuracy, while macro average and weighted average provide balanced metrics. These metrics are important in evaluating our chosen models.

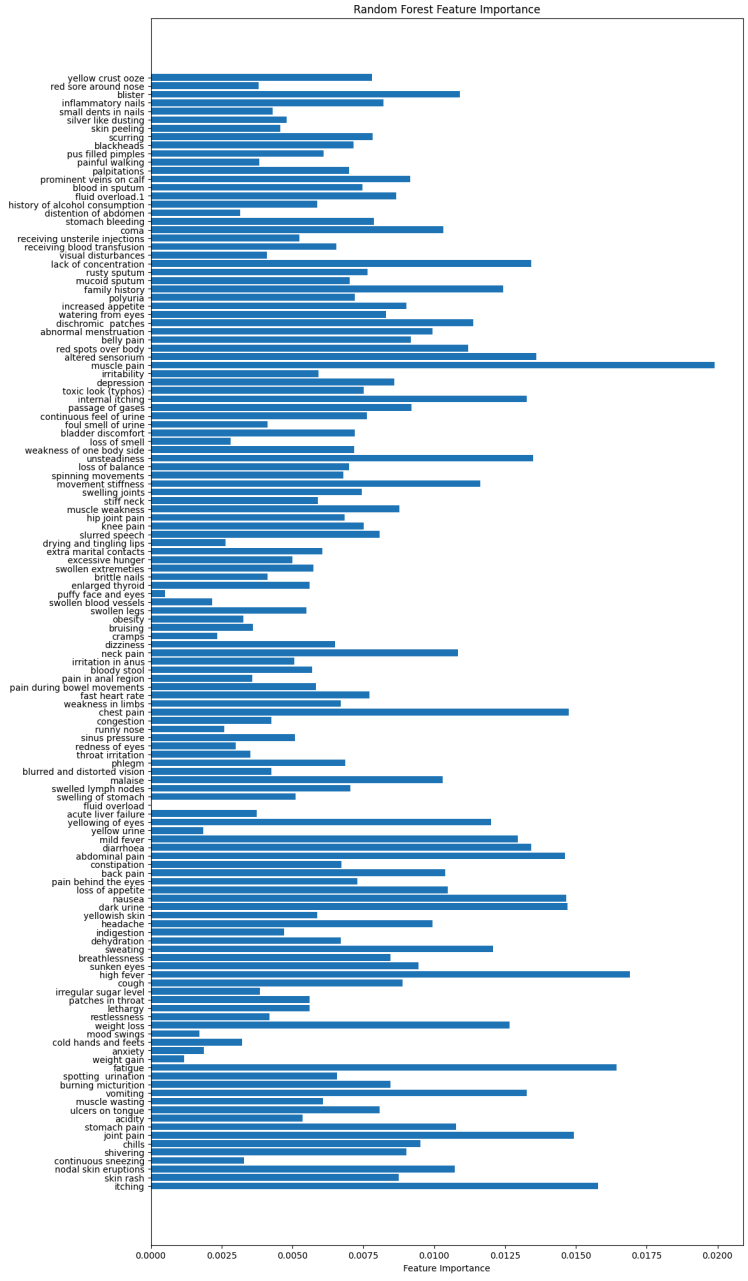
3.2 Models

Two models we chose to use were random forest modeling and logistic regression. The reason why we chose random forest modeling is because this model exhibits superior accuracy in classification tasks when compared to many other models. This is thanks to their ensemble approach, which aggregates predictions from multiple decision trees. For our next model, the importance of logistic regression is ideal for classification due to its ability to model probabilities, simplicity, and interpretability. It works well with both small and large datasets, minimizes overfitting, and provides clear insights into the impact of features on the classification outcome.

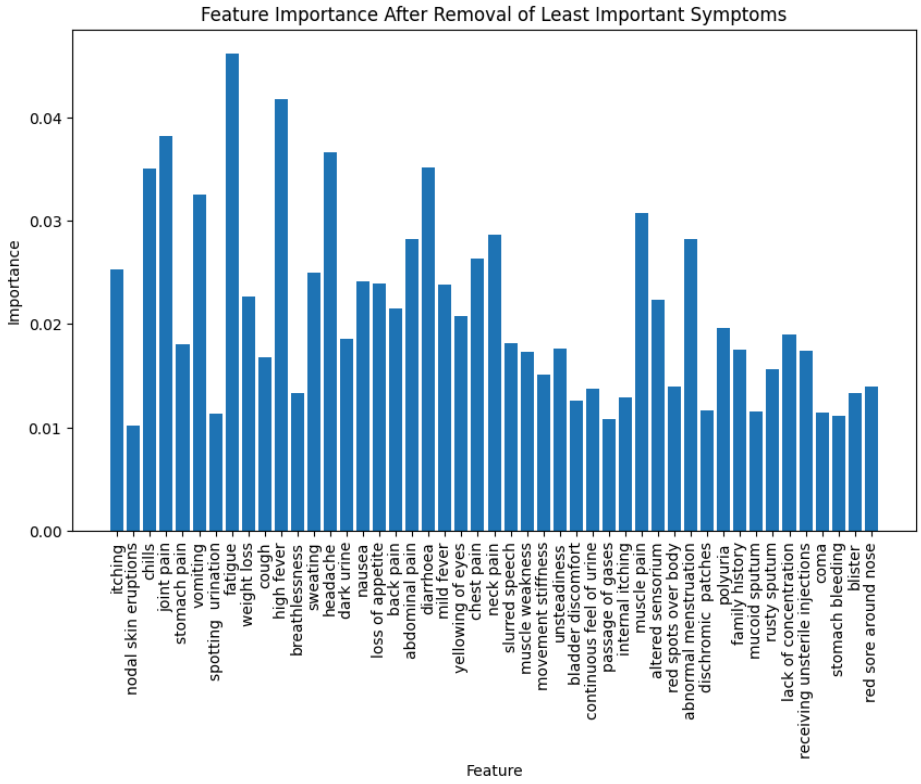
3.3 Evaluation

Random forest combines the output of multiple decision trees into a single decision. In this scenario, random forest translates into medical practice for prioritizing patient visits based on symptoms of higher importance. Logistic regression was also used in this project. It was mainly used to minimize overfitting. The initial accuracy for both models was 1.0 We then reevaluated with the random forest model by using a different testing set, which yielded an accuracy of 97.6%. Because the dataset was so large (132 symptoms), we carefully went through it and eliminated 85 of the symptoms that didn't seem to be very relevant. After the reduction, a dataset with 47 crucial features was obtained. After refining the dataset and retraining the model, the accuracy reached approximately 95%, which is more in line with our expectations and a more credible depiction of the model's predictive abilities. This iterative feature curation process demonstrates our will to improve the model's dependability.

Random forest Feature Importance Visualisation before Re-Tuning:



Random Forest Feature Importance Visualization after Re-Tuning:



3.3.2 Hyper-parameter Optimization

Hyperparameter optimization is crucial for fine-tuning machine learning models and fitting them to the specific characteristics of the data. In this case, the optimized RandomForestClassifier, with hyperparameters {'max_depth': None,'min_samples_leaf': 1,'min_samples_split': 2, 'n_estimators': 50}, achieved an accuracy of 95.24%. The absence of a maximum depth allows flexible tree growth, while minimum leaf samples and split criteria enhance the effectiveness of the model. With 50 estimators, it allows a balance between efficiency and accuracy. These hyperparameters ensure a good-performing model.

3.3.3 Evaluation

In the dataset, we chose a random forest, and logistic regression performed well. This was consistent with our project goal. In particular, the Random Forest model's capacity to yield perceptive data regarding feature importance is in line with our goal of enhancing decision-making during patient interactions. This emphasizes how important it is to use Random Forest's capabilities, which provide essential insights that let medical professionals treat patients with the best care possible.

3.3.4 Evaluation and Additional Information

We encountered some challenges while picking data sets. We went through multiple data sets, but none of them matched our criteria. We found one data set that matched all of our criteria. It consisted of 4920 rows and 18 columns. We tested the models on this dataset, which revealed SVM with 0.98 accuracy, Decision Tree with 0.99 accuracy, and Gaussian Naive Bayes with 0.87 accuracy. Despite its high accuracy, the dataset's structural limitations prompted further exploration, but this was not used because it could not be visualized.

4 Results and Conclusion

To sum up this research, we developed a predictive model for disease and illness prediction to improve doctors' readiness for patient encounters. Using a dataset of 41 diseases and illnesses and 132 symptoms, we focused on the Random Forest model, which produced excellent results and good insights into the significance of the features.

We had some challenges in picking data sets that met our criteria. After many attempts, we found one that fit all our criteria and, most importantly, had really good features. The top five most important features identified—muscle pain, high fever, fatigue, itching, and joint pain—offer actionable insights. Preprocessing steps optimized the dataset.

The models we selected performed exceptionally well, which is in line with our project. Our chosen models demonstrated exceptional performance, aligning with our mission. The integration of the Random Forest model into a chatbot system holds the potential to transform decision-making, patient satisfaction, and overall healthcare outcomes. This initiative signifies a positive stride in leveraging data science for more informed and effective patient care.