Multiclass Classification - Predicting Disease given Symptoms

Allison Peng, Jasper Dong, Wilson Zhou, Eric Sun

Introduction - Problem Statement

- **September 2023 Census Bureau report: 26 million Americans lack health facility access or insurance affordability.**
- **Machine learning in healthcare is increasing in popularity due to improved prediction technology.**
- **Growing availability of healthcare data provides a foundation for accurate prediction models.**

Data sourced from the 'Disease Prediction Using Machine Learning' dataset on Kaggle, offering a symptom-diagnosis data frame.

Our goal

Construct a model that can forecast the disease associated with a patient's symptoms.

Exploratory Data Analysis

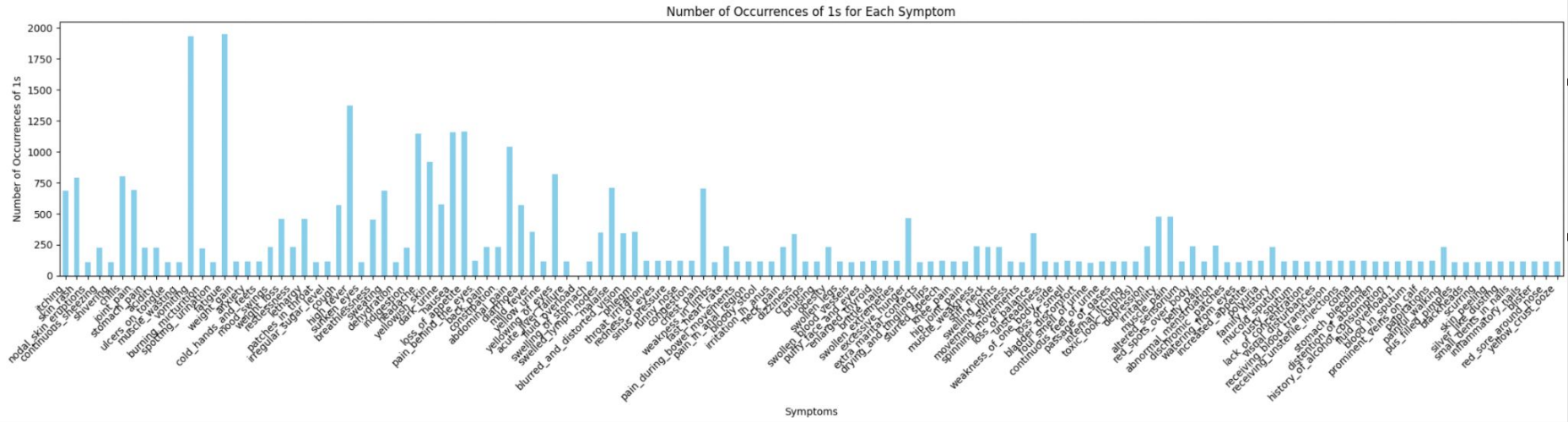
Exploratory Data Analysis

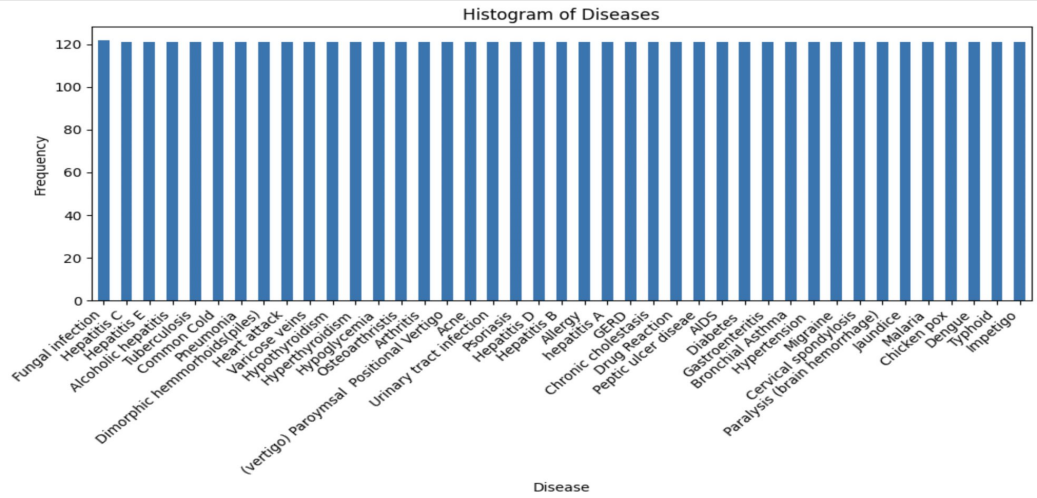
Number of Symptoms/Features	132
Number of Observations	4692
Number of Diseases	42

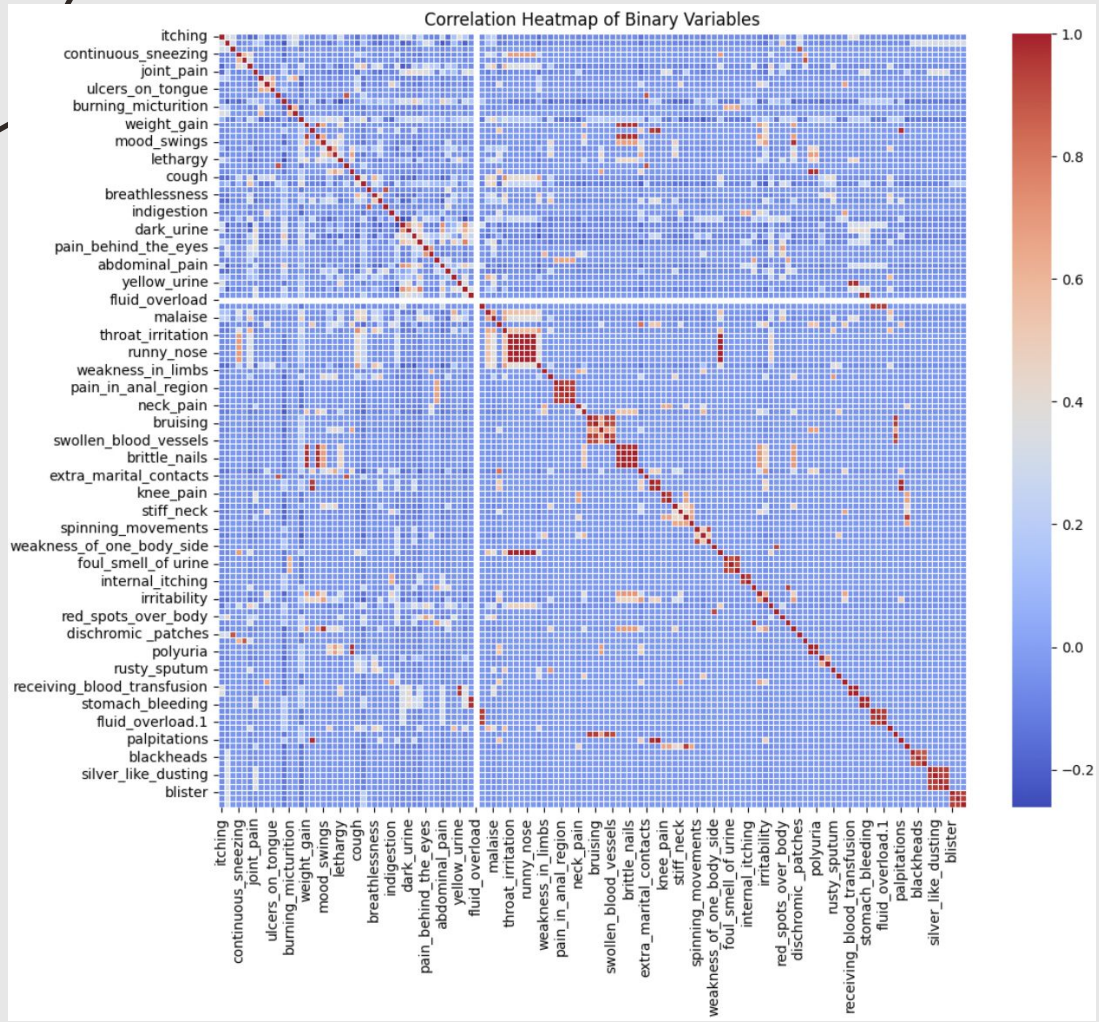
Top 10 symptoms:
Fatigue, vomiting, high_fever, loss_of_appetite, nausea, headache, abdominal_pain, yellowish_skin yellowing_of_eyes, chills

Top 10 diseases:
Fungal infection, Hepatitis C, Hepatitis E, Alcoholic hepatitis, Tuberculosis, Common Cold, Pneumonia, Dimorphic hemmorhoids(piles), Heart attack, Varicose veins

small_dents_in_nails	inflammatory_nails	blister	red_sore_around_nose	yellow_crust_ooze	prognosis
0	0	0	0	0	Fungal infection
0	0	0	0	0	Fungal infection
0	0	0	0	0	Fungal infection
0	0	0	0	0	Fungal infection
0	0	0	0	0	Fungal infection
0	0	0	0	0	Fungal infection
0	0	0	0	0	Fungal infection
0	0	0	0	0	Fungal infection
0	0	0	0	0	Fungal infection
0	0	0	0	0	Fungal infection
0	0	0	0	0	Allergy
0	0	0	0	0	Allergy
0	0	0	0	0	Allergy
0	0	0	0	0	Allergy
0	0	0	0	0	Allergy
0	0	0	0	0	Allergy
0	0	0	0	0	Allergy


Number of Occurrences of 1s for Each Symptom
Number of Occurrences of 1s
Symptoms


Histogram of Diseases
Frequency
Disease
Fungal infection
Hepatitis C
Hepatitis E
Alcoholic Hepatitis
Tuberculosis
Common Cold
Pneumonia
Dimorphic hemmorhoids(piles)
Heart attack
Varicose veins
Hypothyroidism
Hyperthyroidism
Hypoglycemia
Osteoarthritis
Arthritis
(vertigo) Paroymsal Positional Vertigo
Urinary tract infection
Acne
Psoriasis
Hepatitis D
Hepatitis B
Allergy
hepatitis A
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
Impetigo



Correlation heatmap: we calculated the correlation matrix and created a heatmap to visualize the numbers.

With so many predictor variables, it is hard to see the specific correlations. However, we see that a majority of the predictor variables are not correlated with each other.

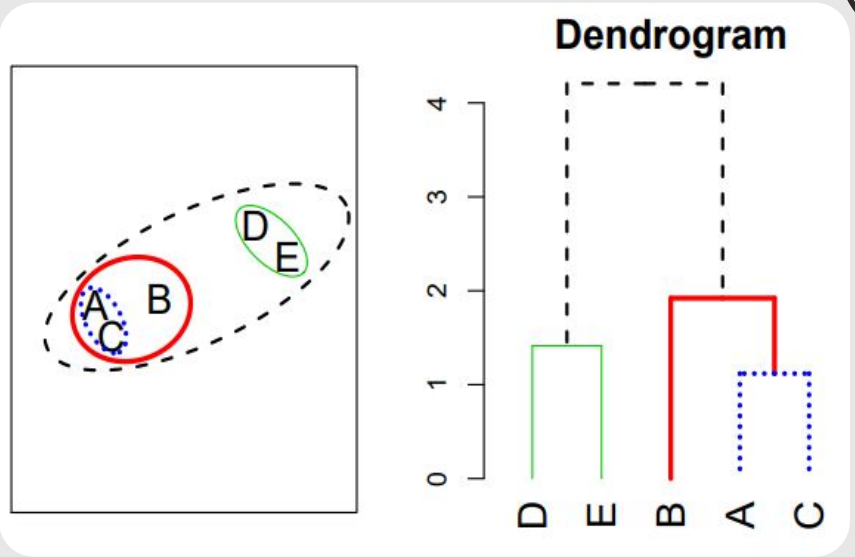
We looked at the top predictor variables that are correlated with each other. Redness_of_eyes, throat_irritation, sinus_pressure, loss_of_smell, congestion, runny_nose, phlegm, blurred_and_distorted_vision, malaise

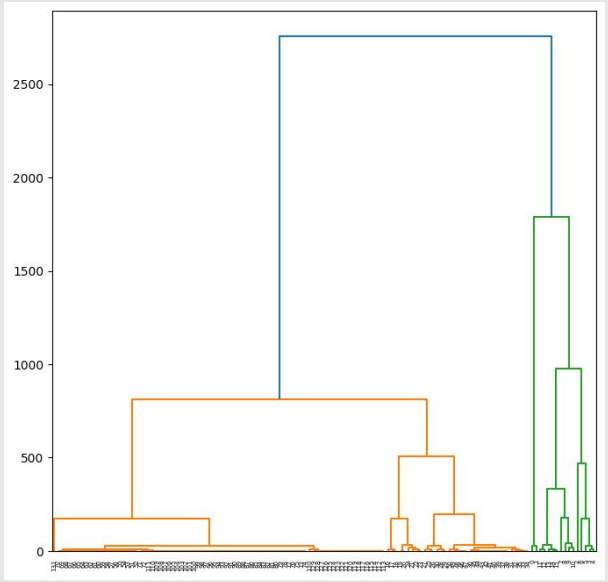
Clustering

Clustering can be used to see if we can group similar symptoms together.

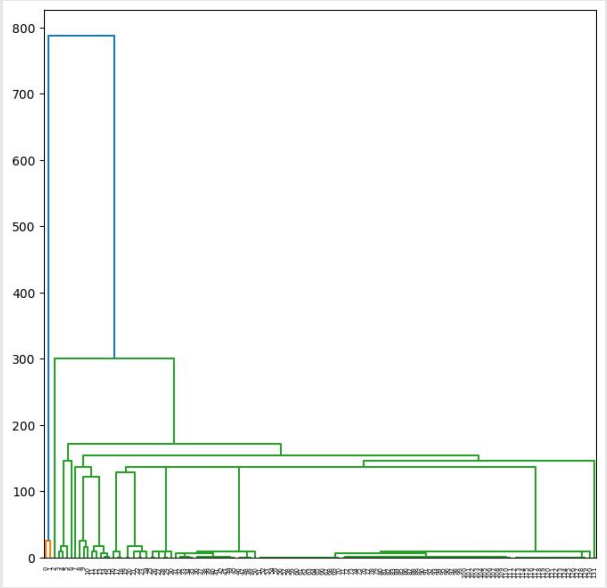
Hierarchical Clustering: Doesn't need a predefined number of clusters, allows us to create dendrograms based on Euclidean distance.

Height of each fusion in dendrogram can indicate similarities/dissimilarities between symptoms.

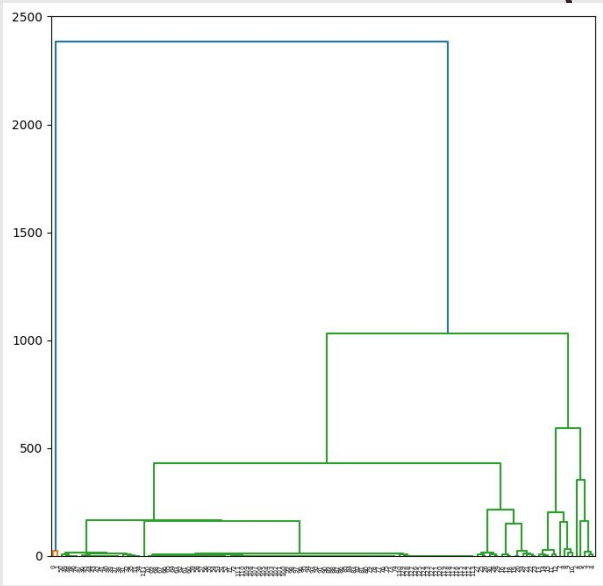




Complete Linkage



Single Linkage



Average Linkage

Methodology

Methodology

Data Preprocessing:
NA/Missing values, Types, Label encoding, Correlated Features

Random Forest Model:
Bagging

Higher prediction accuracy than individual DTL, DTL are prone to overfitting

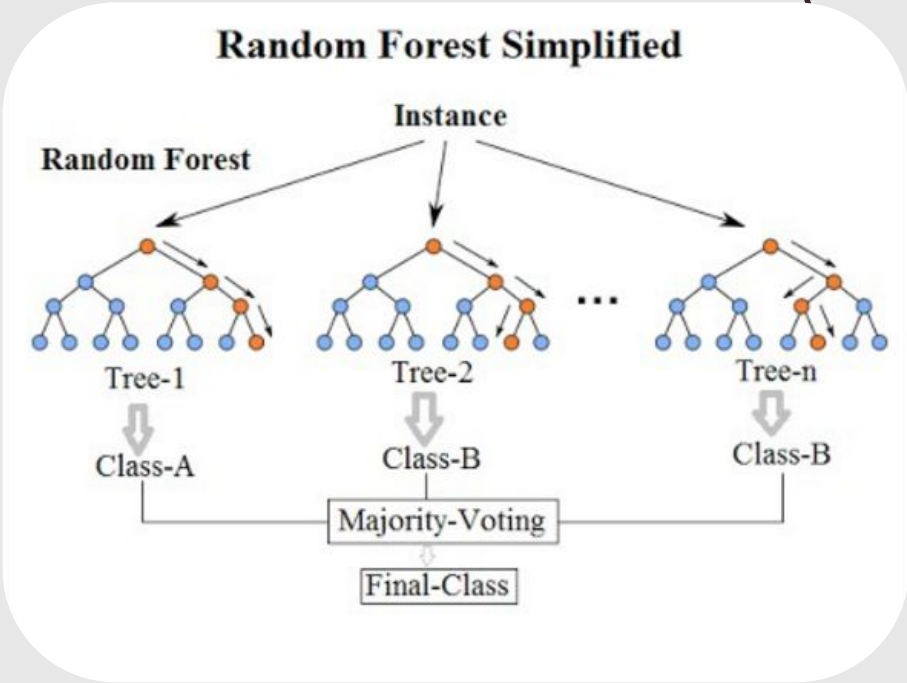
Error Calculation: Out of Bag
Observations not seen in bootstrapped samples.

Validity: ⅓ observations are unseen, unfitted data

Error: Measured through majority vote. Predictions obtained through passing observation -> RF
Majority vote classification of aggregated individual trees

Random Forest

- Why RF?
  - Ensemble of trees using **bootstrapped training samples**
  - Binary data
- Why RF > DT?
  - Decision Tree
    - Interpretability, easier feature split, less complex
    - **OVERFIT**
  - Random Forest
    - Larger, complex
    - Aggregated trees lead to less overfitting, lower variance, decorrelation



Main Results

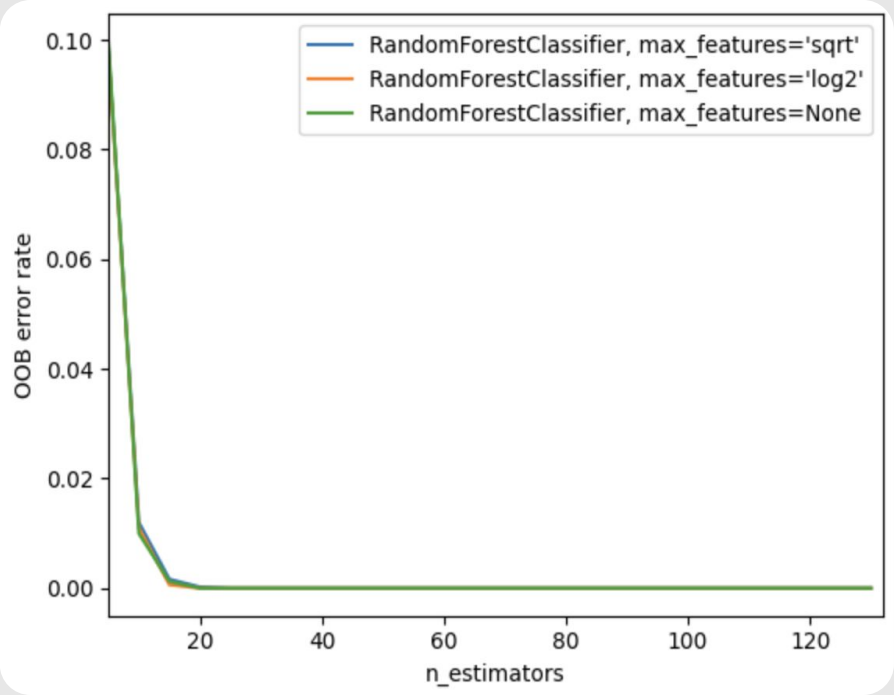
🎯 Accuracy Measurements

	precision	recall	f1-score	support
accuracy	0.976190	0.976190	0.976190	0.97619
macro avg	0.987805	0.987805	0.983740	42.00000
weighted avg	0.988095	0.976190	0.976190	42.00000

SAME RESULTS FOR ALL FEATURE SIZES

[M = P, M = P/2, M = SQRT(P)]

Error w/Out of Bag



How many trees is *good?*

Out of Bag: The out-of-bag (OOB) error represents the mean error of each observation, utilizing predictions from trees that exclude that observation in their bootstrap samples. This feature enables the RandomForestClassifier to be trained and validated simultaneously.

Discussion/ Outlook

Strengths

- High accuracy
- Random Forest performed well given the binary data

Future Work

- Compare other tree based methods
- Eliminate more variables
- Change the train/test split of the dataset

Conclusion

Final Remarks

- Our fitted model turned out to perform really well in classifying the symptoms into diseases

- Clustering also helped in indicating similarities and dissimilarities between symptoms

- However, because our accuracy was so unusually high, this could be an indicator of bias present in our model

- This could be due to a poor split of between the training and testing data, or our model is insufficient in predicting symptoms with diseases in this particular dataset