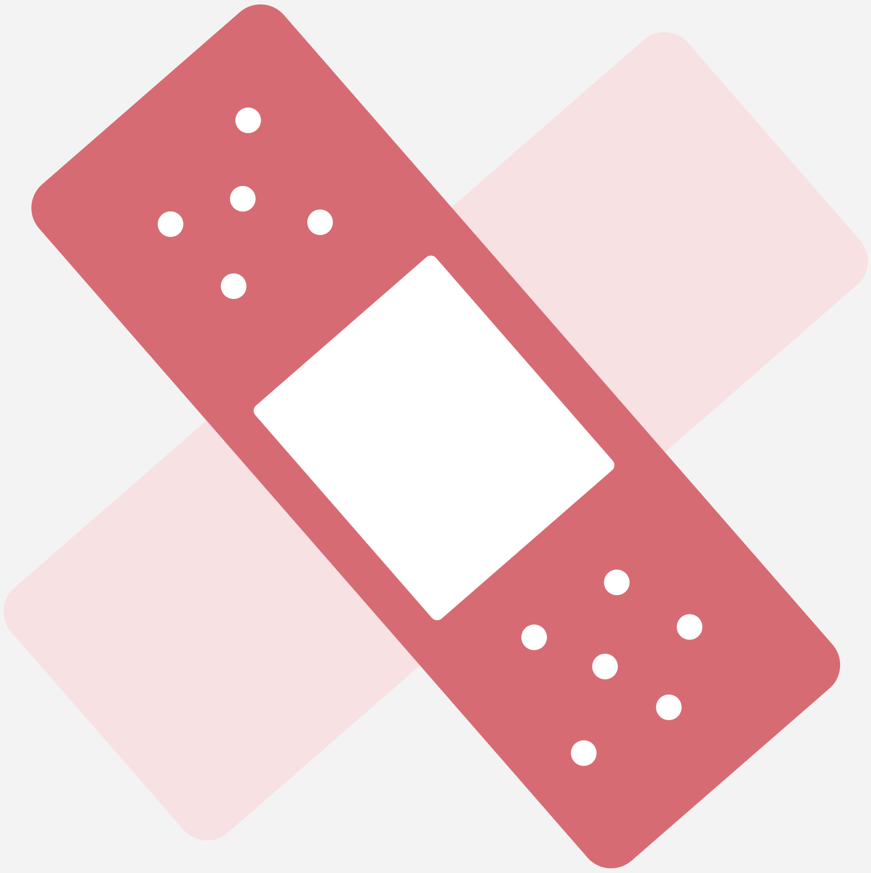


Disease Prediction

BCH 339N - DR. SARINAY CENIK

BY FILINA AND KALEI

Outline

1. Introduction
2. Dataset Overview and Data Cleaning
3. Machine Learning Model
4. Importance/Conclusions

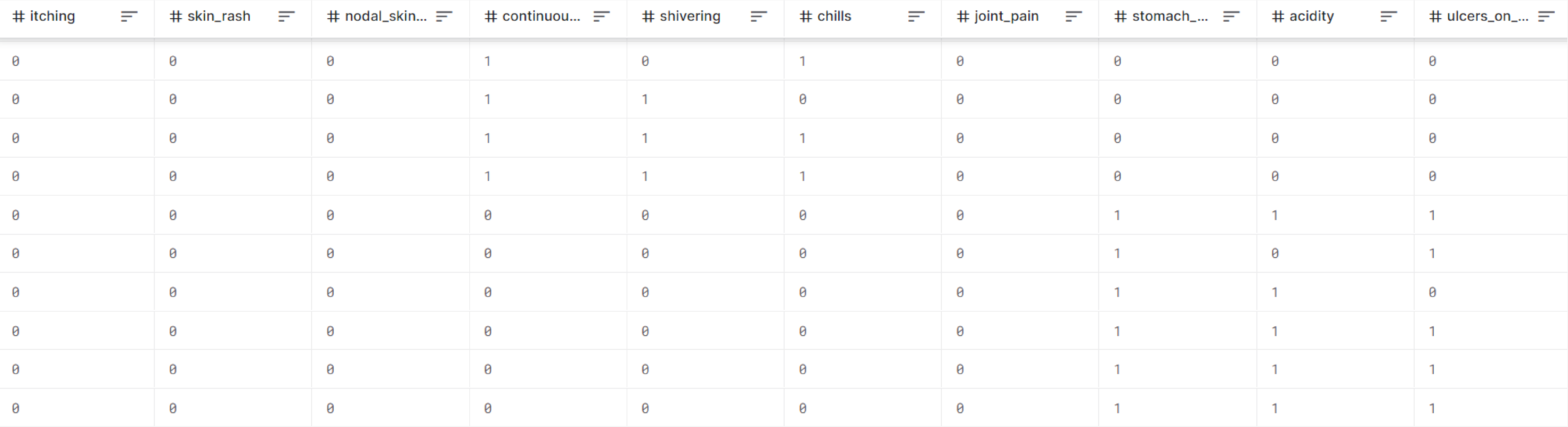
Introduction

- Machine learning model to predict a patient's prognosis using various symptoms.
- Common problem - misdiagnosis within healthcare.
- Build a model that could help doctors make better diagnoses.

Questions we are trying to ask?

- Can diseases be predicted with a ML model?
- Which model would be the best for this purpose?
- How can we use this to benefit healthcare?

Dataset

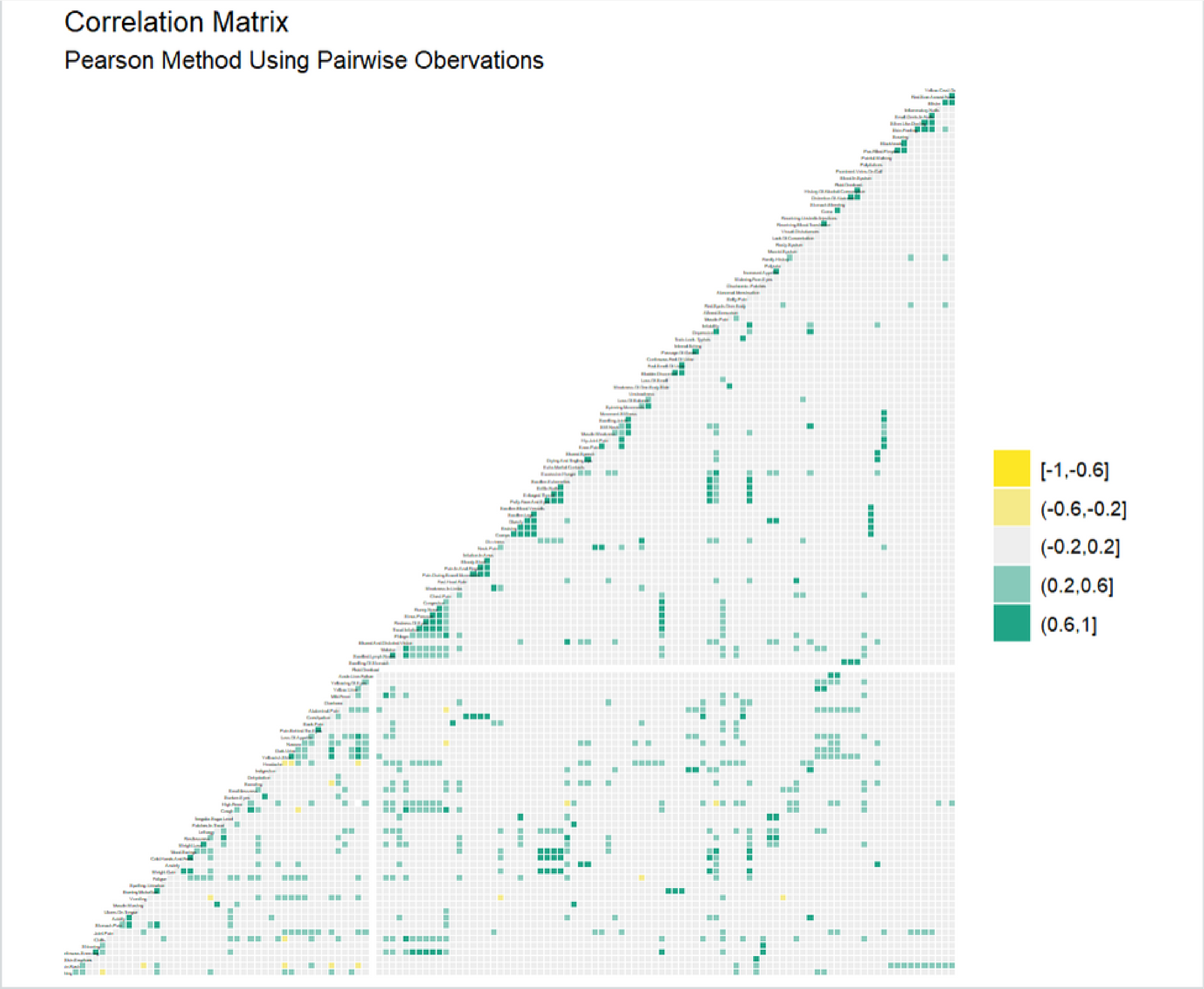


# itching	# skin_rash	# nodal_skin...	# continuou...	# shivering	# chills	# joint_pain	# stomach_...	# acidity	# ulcers_on_...
0	0	0	1	0	1	0	0	0	0
0	0	0	1	1	0	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	1	1	1	0	0	0	0
0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	1	1	0
0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	1	1	1
0	0	0	0	0	0	0	1	1	1

- 42 possible diseases based on 132 symptoms.
- Binary variables with categorical responses.
- Testing set: 42 observations
- Training set: 4920 observations

Exploratory Analysis

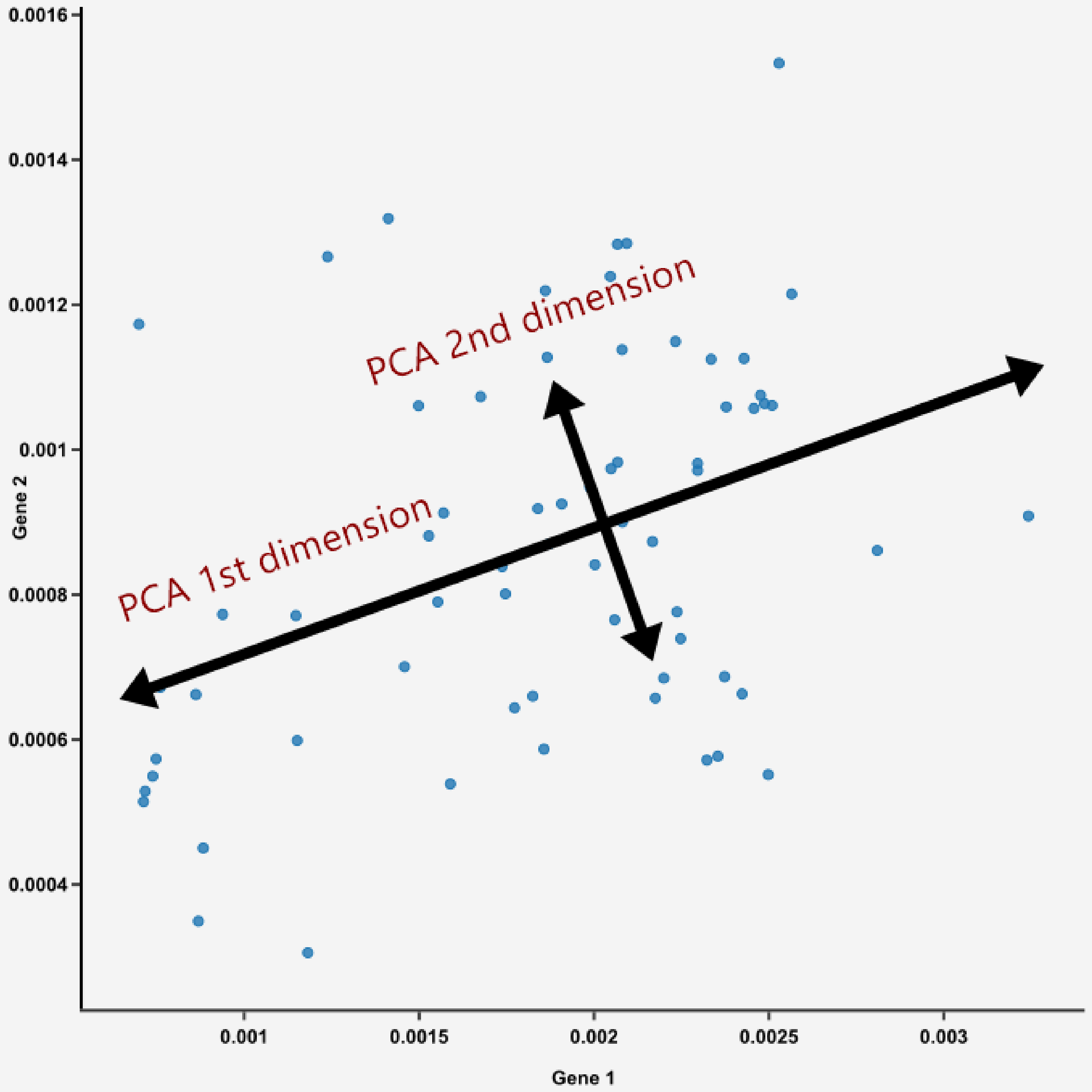
- Variables don't appear to be highly correlated



Correlation Matrix

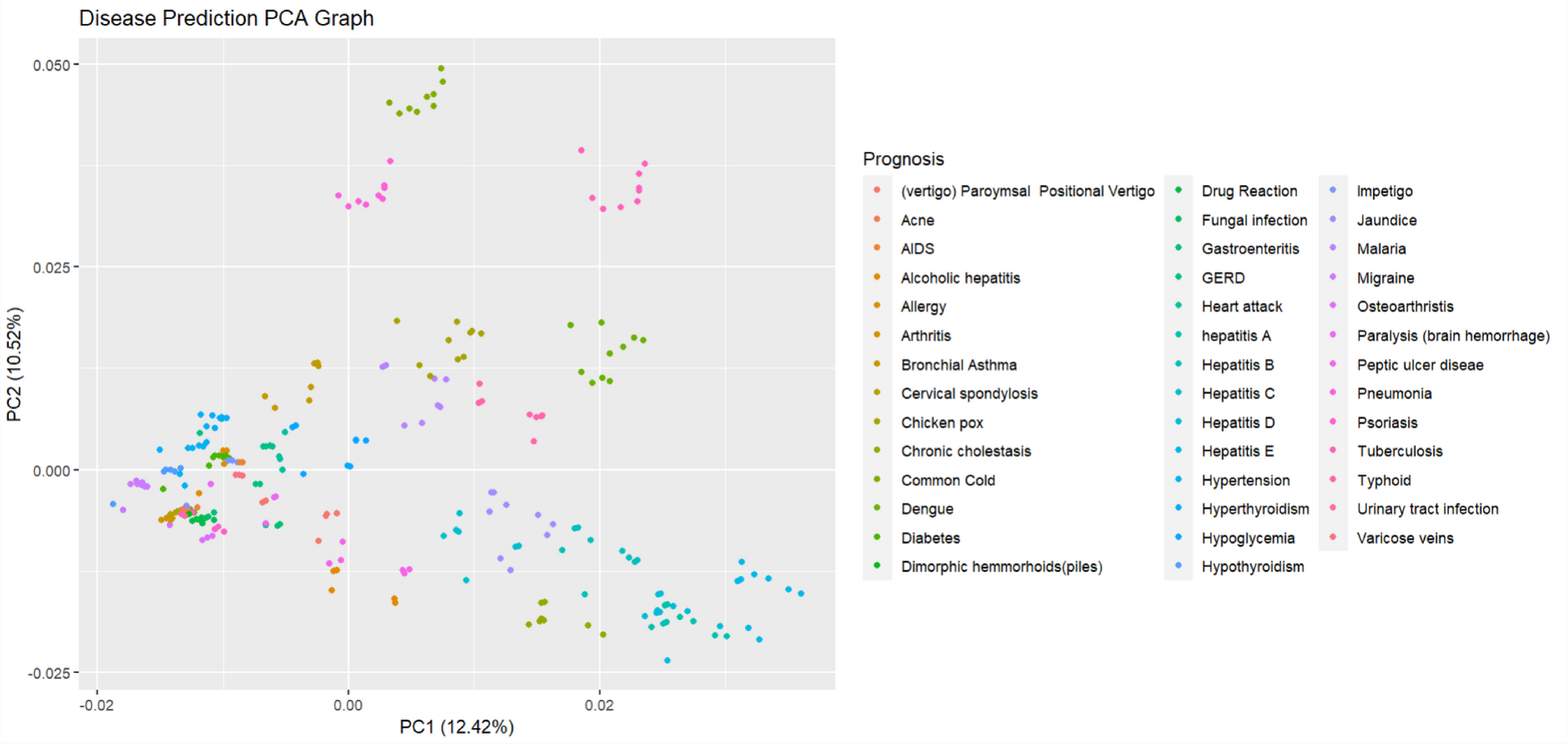
Pearson Method Using Pairwise Obervations

PCA Analysis



- Why PCA? Easy visualization of the variation present in a dataset with many variables.

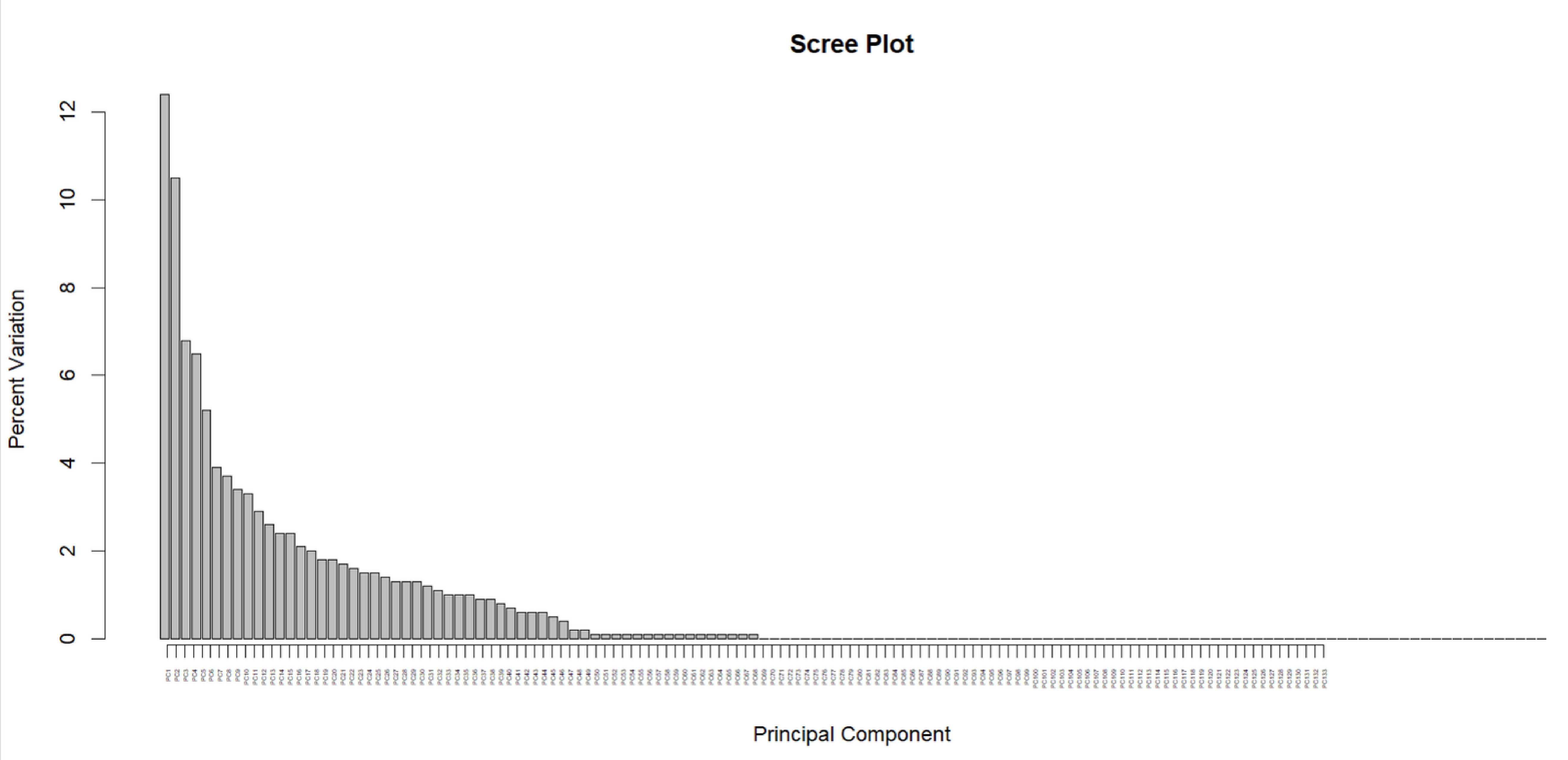
PCA Analysis



Disease Prediction PCA Graph

- Reduced the dimensions of data and analyzed clusters of data to determine their ability to be categorized.
- Between diseases, there are clear separations in symptoms.

PCA Analysis



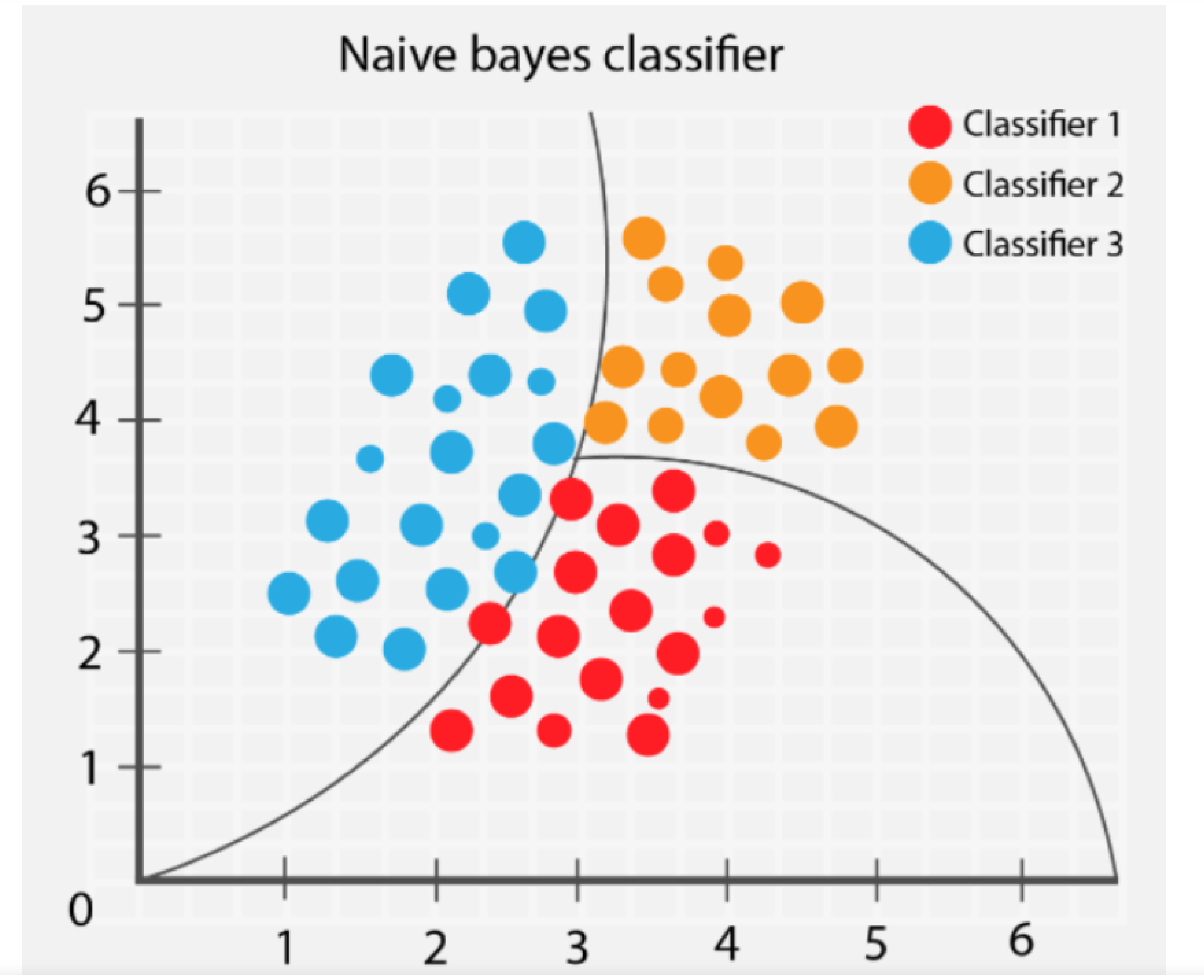
- First two PC components are used in PCA graph, and while it does not constitute a majority, we can still get a good representation of clustering

Possible Model? Random Forest

- Constructs a multitude of decision trees and chooses one.
- Efficient and Robust.
- Large Dataset.
- Able to do regression and classification.
- Able to produce predictions that are easy to understand.

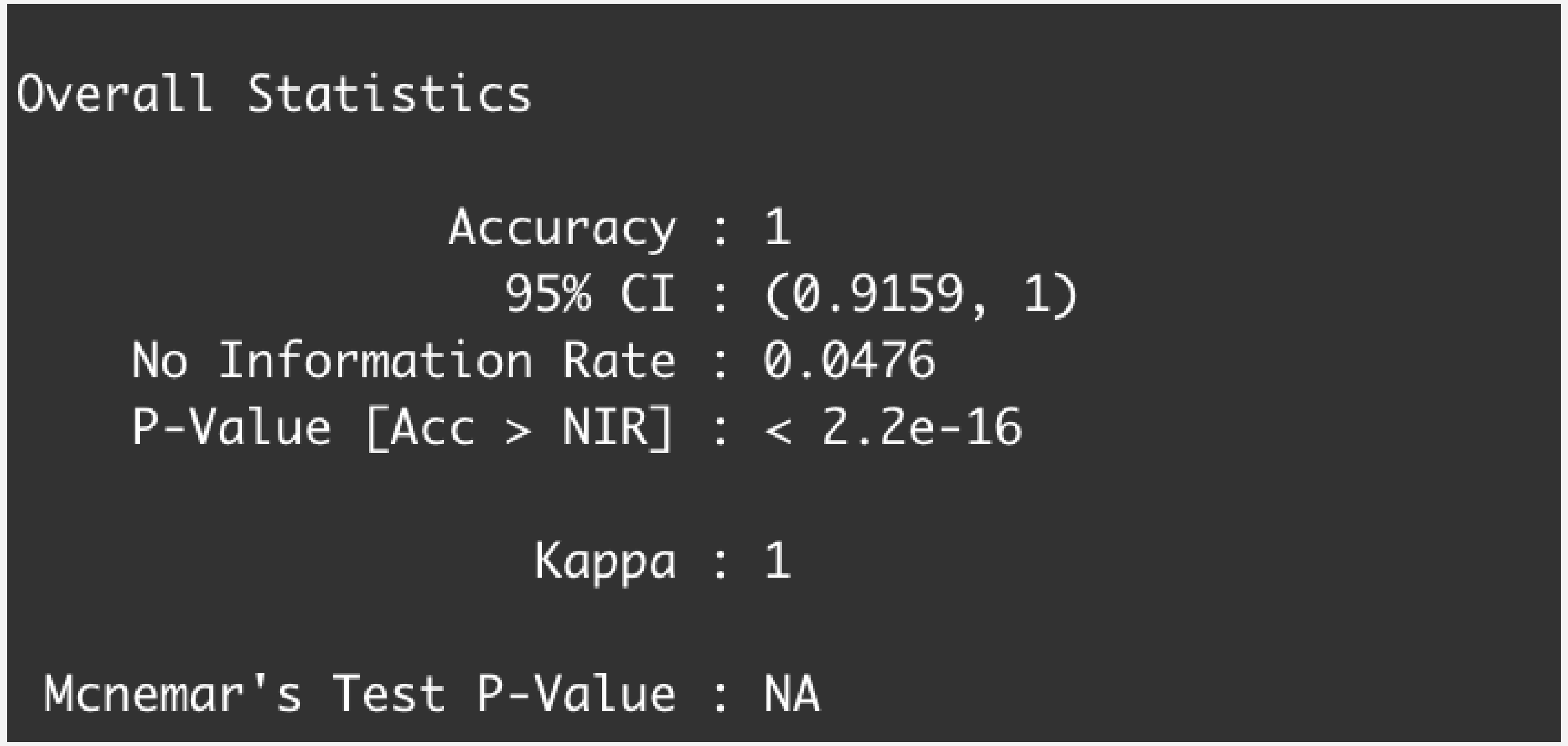
Naive Bayes

- Classifier that uses Bayes Theorem.
- Is able to handle categorical data.
- Large Dataset.
- Able to do the classification we are looking for.
- Able to produce predictions that are easy to understand.



Our Model

- Our Naive Bayes model generated an accuracy score of 100%.
- Runs quickly compared to Random Forest model.
- Can predict diseases from all 132 variables.

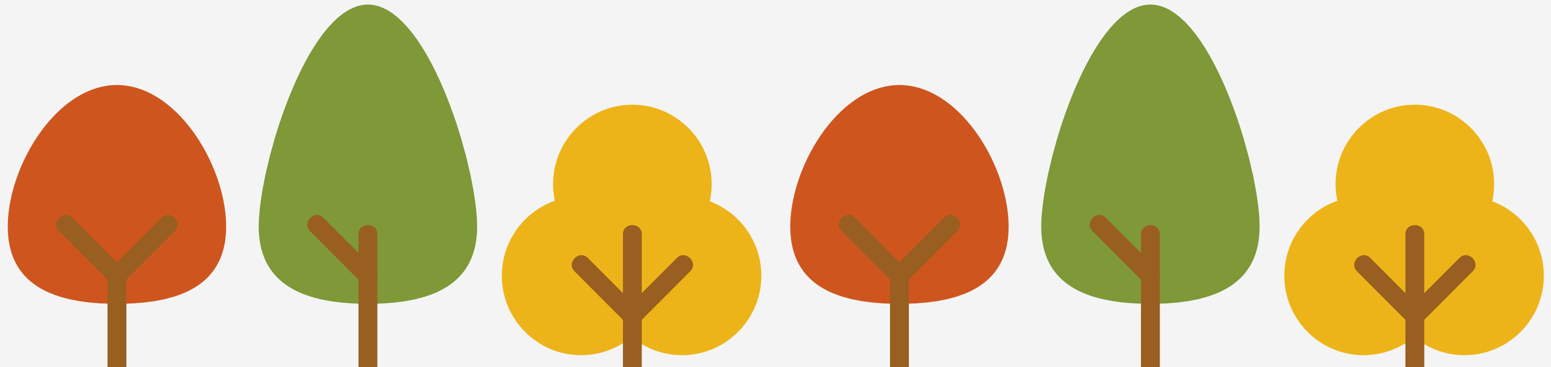


```
Overall Statistics

               Accuracy : 1
                 95% CI : (0.9159, 1)
    No Information Rate : 0.0476
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 1

 Mcnemar's Test P-Value : NA
```



Application

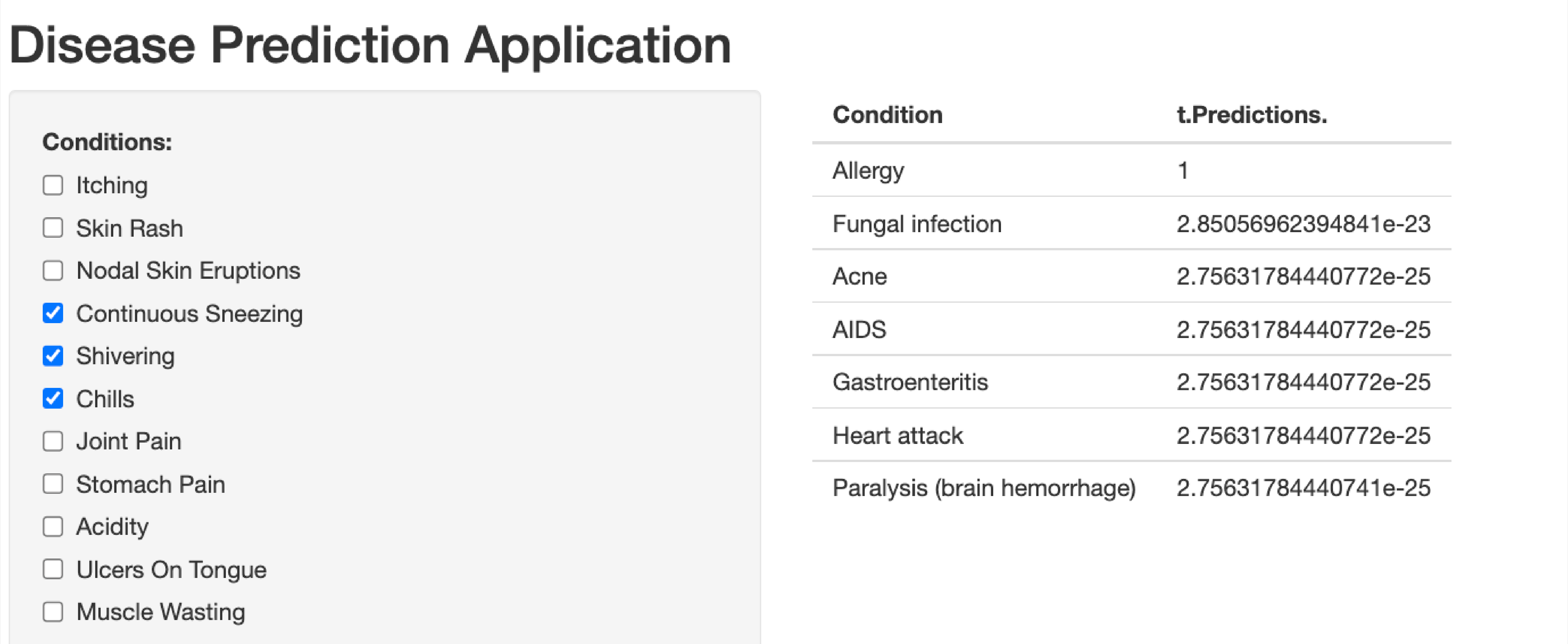
with RShiny !

- Created an web app to represent our disease prediction model!

Allergies

Symptoms Include :

- Sneezing
- Chills
- Shivering



Disease Prediction Application

Conditions:

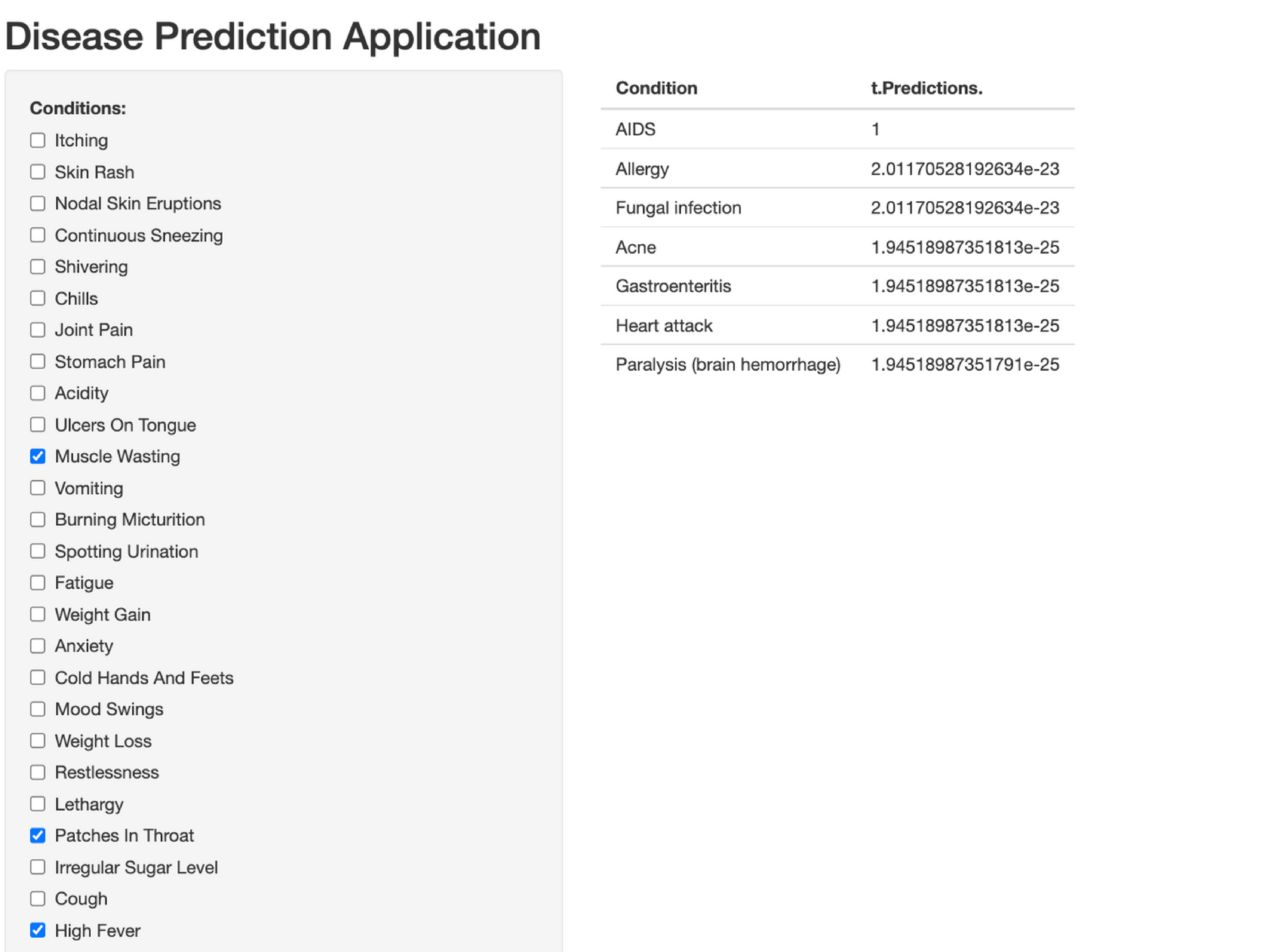
□ Itching
□ Skin Rash
□ Nodal Skin Eruptions
☑ Continuous Sneezing
☑ Shivering
☑ Chills
□ Joint Pain
□ Stomach Pain
□ Acidity
□ Ulcers On Tongue
□ Muscle Wasting

Condition	t.Predictions.
Allergy	1
Fungal infection	2.85056962394841e-23
Acne	2.75631784440772e-25
AIDS	2.75631784440772e-25
Gastroenteritis	2.75631784440772e-25
Heart attack	2.75631784440772e-25
Paralysis (brain hemorrhage)	2.75631784440741e-25

AIDS

Symptoms Include :

- Extramarital Contacts
- Muscle Wasting
- Patches in Throat



Disease Prediction Application

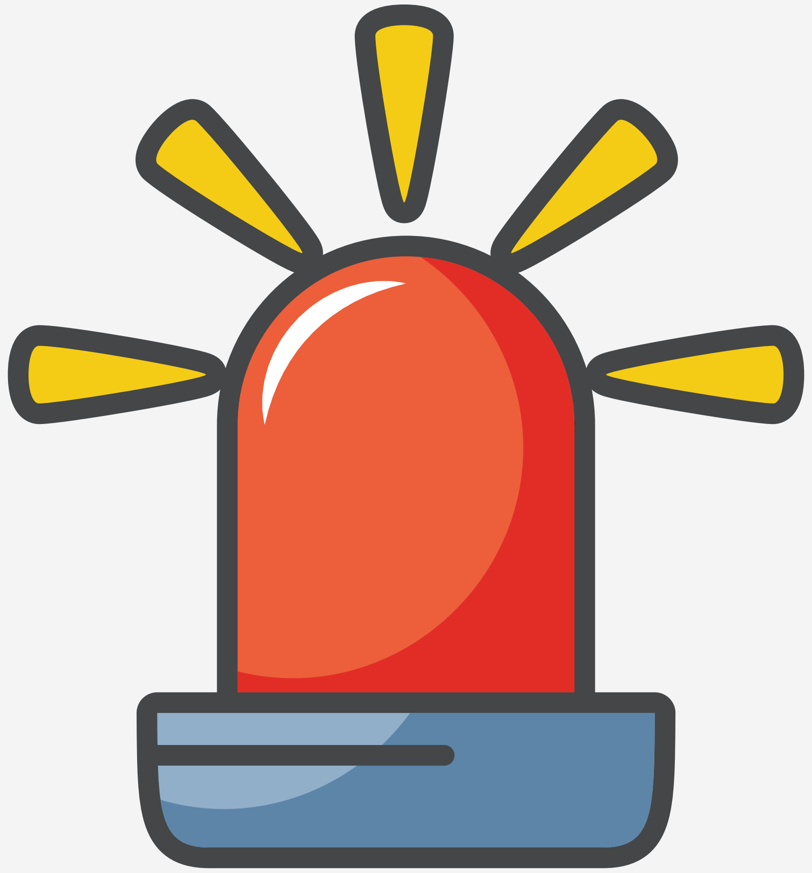
Conditions:

☐ Itching
☐ Skin Rash
☐ Nodal Skin Eruptions
☐ Continuous Sneezing
☐ Shivering
☐ Chills
☐ Joint Pain
☐ Stomach Pain
☐ Acidity
☐ Ulcers On Tongue
☑ Muscle Wasting
☐ Vomiting
☐ Burning Micturition
☐ Spotting Urination
☐ Fatigue
☐ Weight Gain
☐ Anxiety
☐ Cold Hands And Feets
☐ Mood Swings
☐ Weight Loss
☐ Restlessness
☐ Lethargy
☑ Patches In Throat
☐ Irregular Sugar Level
☐ Cough
☑ High Fever

Condition	t.Predictions.
AIDS	1
Allergy	2.01170528192634e-23
Fungal infection	2.01170528192634e-23
Acne	1.94518987351813e-25
Gastroenteritis	1.94518987351813e-25
Heart attack	1.94518987351813e-25
Paralysis (brain hemorrhage)	1.94518987351791e-25

Limitations

- Only works with data seen before
- Doesn't account for multiple diseases.
- It can only work with a certain number of symptoms (depends on prognosis).



Improvements

- Categorize symptoms.
- Include visuals like a pie chart.
- Add more diseases
- Be able to account for more than 1 disease at a time.

Conclusion/Importance

- Utilizing technology can help improve patient outcomes.
- We should invest more in technology in the medical field.
- There are machine learning models to help doctors solve problems.
- Naive Bayes and Random Forests are a good way to predict diseases.

Thank you!

Resources

KAUSHIL268. (2020, May 15). Disease prediction using machine learning. Kaggle. Retrieved from https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning

Mastroianni, B. (2020, February 22). Why getting medically misdiagnosed is more common than you may think. Healthline. Retrieved from https://www.healthline.com/health-news/many-people-experience-getting-misdiagnosed

Singh, H., Meyer, A. N., & Thomas, E. J. (2014). The frequency of diagnostic errors in outpatient care: Estimations from three large observational studies involving us adult populations. BMJ Quality & Safety, 23(9), 727–731. https://doi.org/10.1136/bmjqs-2013-002627