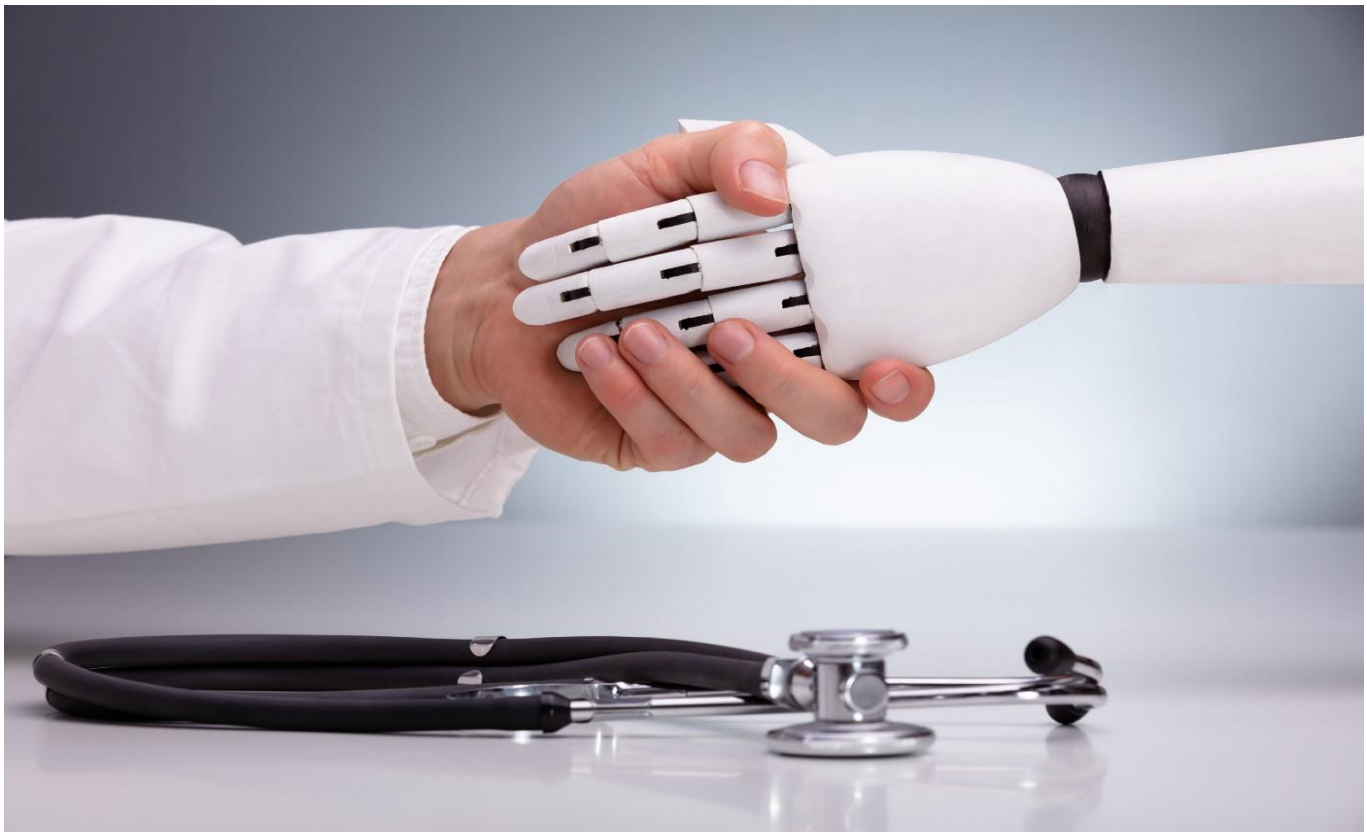
Machine Learning Public Health Project
Disease Prediction
Spring 2022

Pranjal Srivastava (ps4379) & Irving Angeles (ia2246)

March 09, 2022



A. Introduction

Machine learning application in the public health field is a growing field that can increase and improve the patient's health outcomes and even have a positive impact in the health system. For instance, there are diseases that require diagnosis from specialists (ie. pathologist or radiologist) and human error can lead to false positives or false negatives[1]. In scenarios like this machine learning can be used to account for these errors and help improve accuracy of diagnosis. In addition, machine learning can help ease the burden of medical infrastructures due to a shortage of healthcare providers. Studies have shown that repetitive tasks such as reviewing lab results, imaging, medication adherence can be automated through machine learning and thus reduce the burden on health care staff[2]. This allows for the better allocation of resources and maximizes patient care. In our project we will conduct an example of using health data in order to predict a diagnosis to simulate an application of machine learning in public health. We obtained a dataset containing prognosis of various diseases along with various symptoms with a binomial value of symptom present or not for each prognosis.

B. Related work

Our project follows the trajectory of using health data to predict medical diagnosis in unsupervised machine learning. Our data set can be viewed an example of using electronic health records to approach precision medicine by combining various health data to better diagnosis accuracy and therefore treatment[3]. An example of machine learning methods that have been researched are decision tree models, these models have been studied thoroughly to explore the automation of medical diagnosis. One such study used single decision tree, boosted decision tree and decision tree forest to classify breast cancer and each model performance was assessed[4]. Since the application of unsupervised learning can be done in various specializations of medicine our project focuses on diagnosis of a specific disease, fungal infection. There are some drawbacks to unsupervised machine learning for instance clinical expertise is still needed in order to choose variables that are the most related for the diagnosis and at times these type of models are unable to differentiate between diagnosis that are not clinically possible[5]. This is why machine learning is a tool that must be used in conjunction with clinical expertise in order to have predictive models that are interpretable.

---

[1] Ahuja, Abhimanyu S. "The impact of artificial intelligence in medicine on the future role of the physician." *PeerJ* 7 (2019): e7702.

[2] Ahmed, Zeeshan, et al. "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine." *Database* 2020 (2020)

[3] Nayak, Losiana, Indrani Ray, and Rajat K. De. "Precision medicine with electronic medical records: from the patients and for the patients." *Annals of Translational Medicine* 4.Suppl 1 (2016).

[4] Azar, Ahmad Taher, and Shereen M. El-Metwally. "Decision tree classifiers for automated medical diagnosis." *Neural Computing and Applications* 23.7 (2013): 2387-2403.

[5] Kononenko, Igor. "Machine learning for medical diagnosis: history, state of the art and perspective." *Artificial Intelligence in medicine* 23.1 (2001): 89-109.

C. Methods

In this machine learning project we chose prognosis as our response variable and the predictors include: itching, skin rash nodal skin eruptions, shivering and chills. We chose these predictors because these are symptoms associated with fungal infections. Our data set included many other predictors but in order to assess the accuracy of our models we identified a single prognosis.

Classification machine learning models were conducted in our project: decision tree, gradient boosted decision tree and bagging.

D. Data And Experiment

Data Preparation and Cleaning

The most important part of this project is to import and clean the data as needed.The dataset contains the variables as various clinical symptoms and prognosis as a result of combination of symptoms. The data is originally taken from Kaggle data source: (https://www.kaggle.com/datasets/kaushil268/disease-prediction-using-machine-learning)

Examining the raw dataframe

Later upon examining the data we got to know that there is one unknown variable named X with 4290 missing values. But as we will be filtering the data frame with our variables of interest, these missing values will not be problematic. All other variables are well structured with no missing values.

So, as we discussed earlier, for our project and research question, we need only 6 variables out of the raw data to build our model and perform analysis on it. We will be filtering the variables Prognosis, Itching, Skin rash, Nodal skin eruptions, Shivering, Chills. Using these variables we will form a new data frame Disease which will be our final data frame to work upon. Also, the values in the target variable Prognosis are of object datatype, but for proper modeling and prediction, we will have to convert them into factors. There is a limitation with the R studio, that it cannot handle the variables with more than 21 categories for the machine learning prediction, so we have to unfortunately remove some of the unique values(Diseases) from the variable prognosis but this will not affect our models or prediction but will also improve the results by making them clear and explicit.

Now that we have our final data frame Disease, we performed some exploratory data analysis.

We did not see any missing values in our final dataframe. Prognosis has multiple categories describing names as various diseases. The symptom variables are valued either 1 or 0, with 1 meaning symptom is present and 0 meaning it symptom is not present.

We have a total of 19 Diseases in our data frame. These are 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**.**

Splitting the Data

We then split the data into 50% training and 50% testing data. We randomly split our data set and so our training and testing data set have random chances of having the prognosis.

E. Results

After splitting the data into training and testing data, we applied various machine learning algorithms to make various models, and validating them by checking testing and training errors.

Decision Tree

Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously split according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves. The leaves are the decisions or the final outcomes. And the decision nodes are where the data is split. We made the decision tree of our model using all 5 variables of symptoms as predictors and prognosis as response variable.We also did pruning of trees to make the tree better visualized. We also calculated training and testing errors for the same.

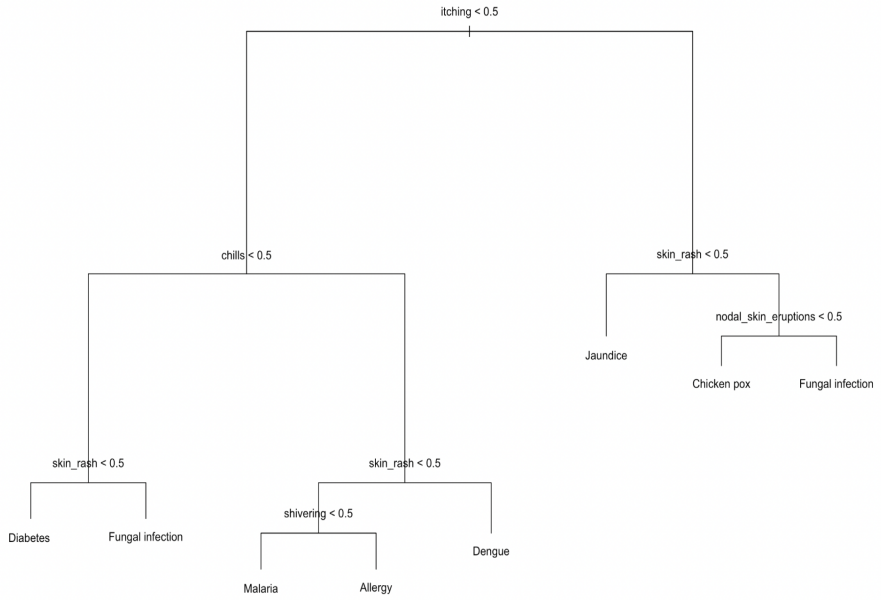


Figure 1. Decision Tree

Training Error	0.6552632
Test Error	0.6815789

Results

Since, the predictors variables are binary, "<0.5 shows 0", " >0.5 shows 1", respectively this means no presence of symptom and presence of symptom. This classification tree depicts clearly that if a person has itching, skin rash and nodal skin eruption, he might have Fungal Infection. Meanwhile, if a person has itching, but no skin rash, he might have Jaundice. In the same way if the person has itching, skin rash, but no nodal eruption, chances are that he has chicken pox.

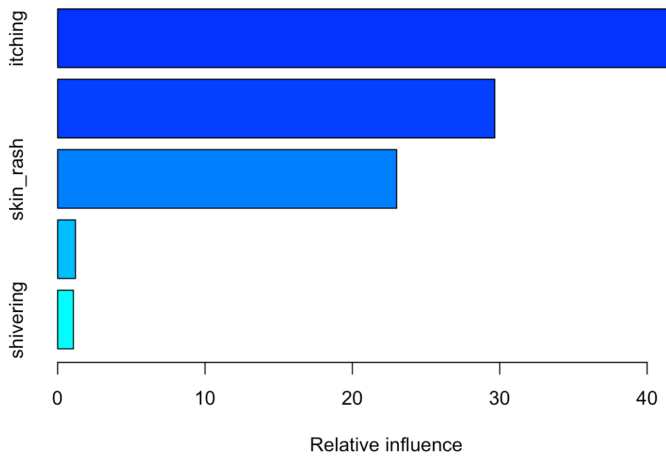
Coming to the non itching disease, if a person has chills and skin rash, he should have Dengue, but he has chills but no skin rash, then he might have Allergy if Shivering is present and if shivering is not there, then Malaria. If a person with no itching, does not have chills, then he might be affected with Diabetes if skin rash is not there, but he feels the symptoms of skin rash, he probably has Fungal Infection, which is our main prognosis of interest. Training error of 0.6552632 shows that around 34% of the training data has been explained by the model, while testing error of 0.6815789 shows that around 32% of testing data has been explained by the model Seeing the decision tree, Chills and Shivering seems to be non associated with the Fungal Infection.

As you can notice, most of the diseases in the decision tree are related to skin, this is because we have chosen the variables which are mostly correlated with skin diseases.

Boosting

To validate our previous model in order to make better predictions, we will be applying Boosting over our model. Boosting is used to create a collection of predictors. In this technique, learners are learned sequentially with early learners fitting simple models to the data and then analyzing data for errors. Consecutive trees (random sample) are fit and at every step, the goal is to improve the accuracy from the prior tree.

We also calculated the testing and training errors for the boosting model




```
##                                   var   rel.inf
## itching                         itching 45.055338
## chills                           chills 29.663946
## skin_rash                     skin_rash 23.004072
## nodal_skin_eruptions nodal_skin_eruptions  1.207991
## shivering                     shivering  1.068652
```

Figure 2. Relative Influence Plot

Training Error	0.6447368
Test Error	0.6657895

Results

After seeing the summary of the Boosting model, the top most variable Itching is the most important variable, Chills and Skin rash being important and Shivering and Nodal Skin eruption being the least important variables for the response variable. Relative influences of itching(45.006) and chills(29.75) were highest among all for Fungal Disease. Skin rash also has a significant relative influence of 22.61.

Bagging

In order to move further in validating our model, we will apply bagging method which is also known as bootstrap aggregation. Bagging is used when the goal is to reduce the variance of a decision tree classifier. Here, the objective is to create several subsets of data from training

sample chosen randomly with replacement. Each collection of subset data is used to train their decision trees.

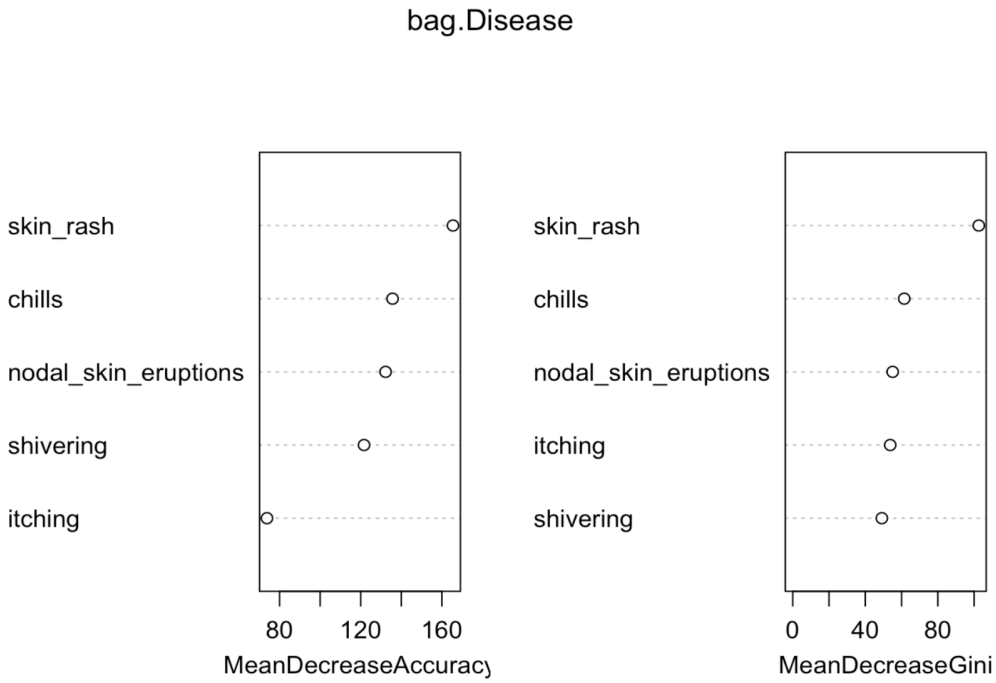


Figure 3. Variable Importance Plot

Training Error	0.6429825
Test Error	0.6675439

Results

After performing bagging, we notice that the mean decrease accuracy of skin rash is the highest(approx.160) shows that it's the most important variable followed by chills, nodal skin eruption, shivering and itching. The mean decrease accuracy expresses how much accuracy the model losses by excluding each variable. The more the accuracy suffers, the more important the variable is for the successful classification. The variables are presented from descending importance.

The mean decrease in Gini coefficient is a measure of how each variable contributes to the homogeneity of the nodes and leaves in the resulting random forest. The higher the value of mean decrease accuracy or mean decrease Gini score, the higher the importance of the variable

in the model. So, we can say that skin rash has the highest importance in the model followed by chills, nodal skin eruption, itching and shivering.

F. Summary

After the careful evaluation of the results of all the machine learning methods we applied to our model in order to make a prediction, we found out that we were able to predict disease based on symptoms as predictors. Based on symptoms we have chosen, Fungal Infection was significantly correlated to skin rash, chills, nodal skin eruptions, shivering, and itching. These results can be of significant use in future clinical science and can help medical professionals and clinical industries around the globe along with advancement of diagnosis and treatment of patients. Due to limitations of our programming tool R Studio, we had to choose only 5 variables(unique symptoms), but for future analysis, we will be using various other symptoms to predict more diseases.

Team Member Contributions

a. Introduction - Irving
b. Related Work - Irving
c. Methods - Irving
d. Data & Experiment Setup - Pranjal
e. Results - Pranjal
f. Discussion - Pranjal
g. Presentation - Pranjal & Irving