Final Project Report – Disease Prediction

Rhichard Koh (100842848)

Durham College

COSC – 21000 – 01 – Intro to Machine Learning

Dr. Ruba Al Omari

December 11, 2022

Introduction

The problem that I will be tackling in my final project is disease prediction based on symptoms given. The motivation for this work is to help physicians diagnose people and it can be an aid for people that live in places where they do not have access to immediate healthcare. The benefits of running a machine learning model over this dataset rather than hard coding each disease and their symptoms are that it is more efficient to program, easier to maintain, and very easy to update with new diseases.

Related Work

Comparing different supervised machine learning algorithms for disease prediction (Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A., 2019).

In this paper they compared logistic regression, support vector machine, decision tree, random forest, naïve bayes, k nearest neighbor, and an artificial neural network with each other to see which model performs the best to predict diseases. They used various datasets; I believe to compare the models with different biases in the dataset. They found the most applicable algorithms are SVM and naïve bayes. However, the most accurate performer is random forest.

Disease Prediction by Machine Learning Over Big Data From Healthcare Communities (Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L., 2017)

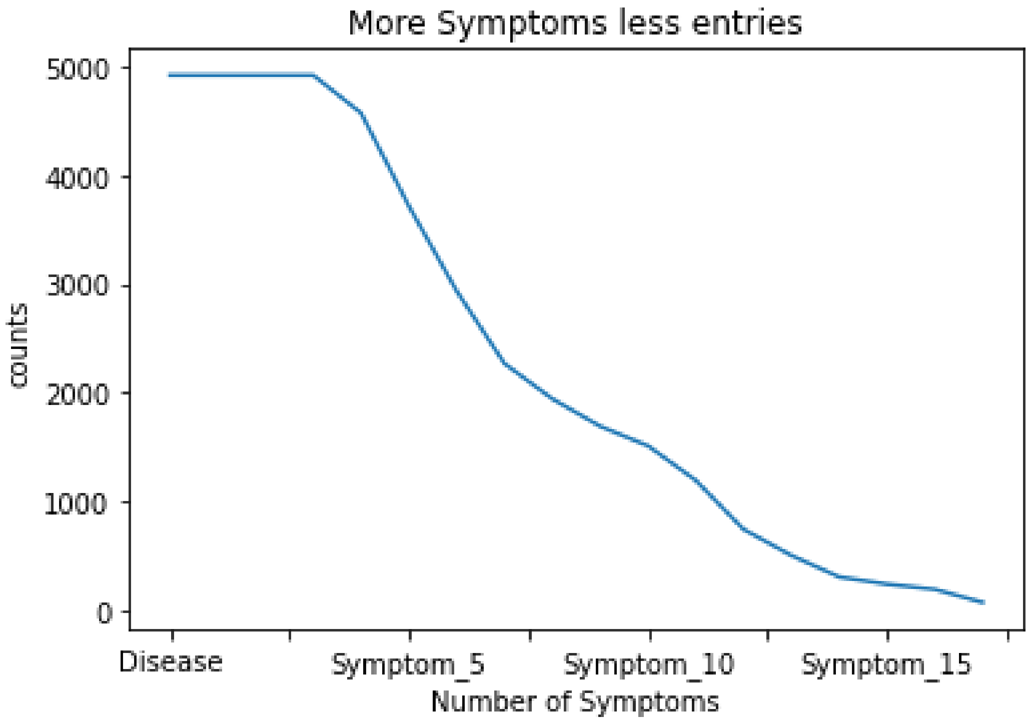
In this paper they created a new disease predictor using a modified convolutional neural network. It uses both structured and unstructured data straight from the hospital. It is stated that their accuracy score is 94.8% and a faster convergence speed than other CNN based disease predictors. The evaluation metrics they used were accuracy score, precision, recall, and f1 score.

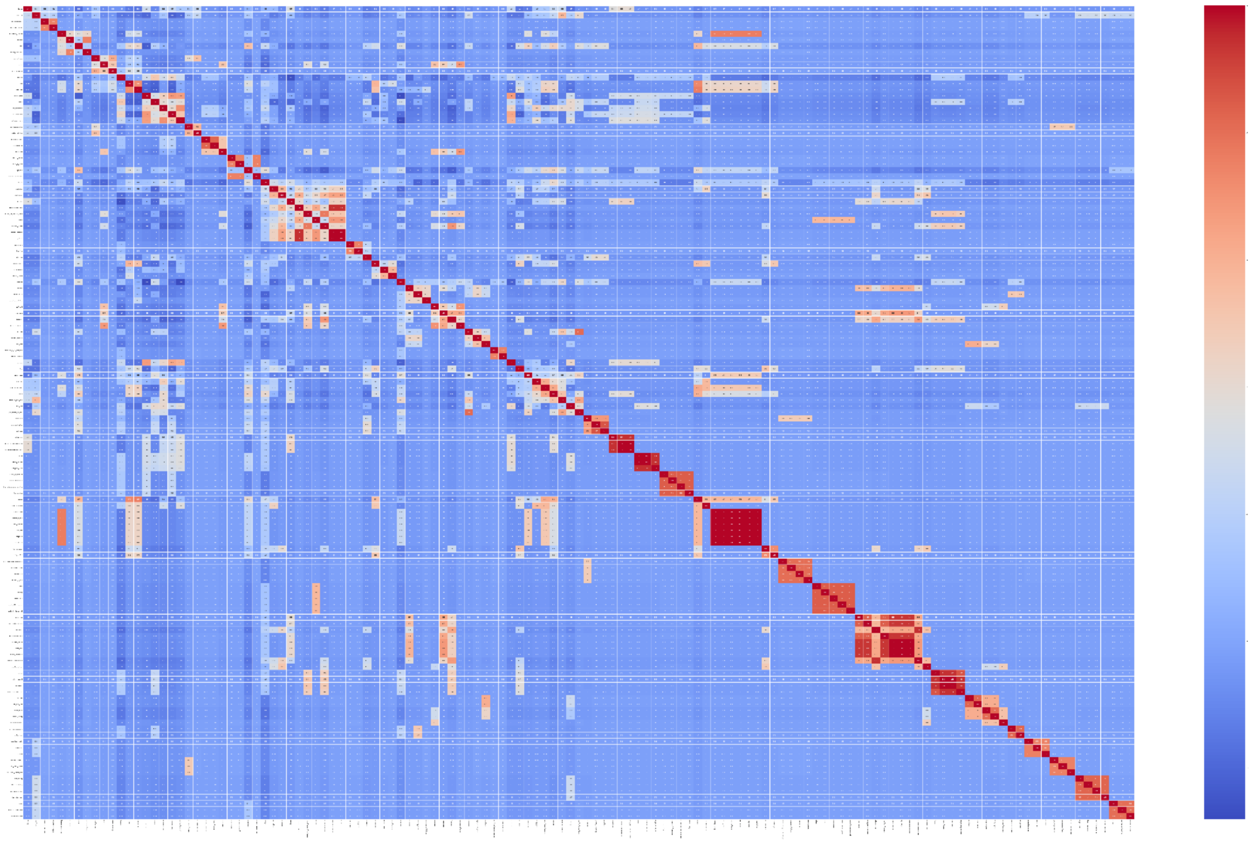
Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques (Mohan, S., Thirumalai, C., & Srivastava, G., 2019)

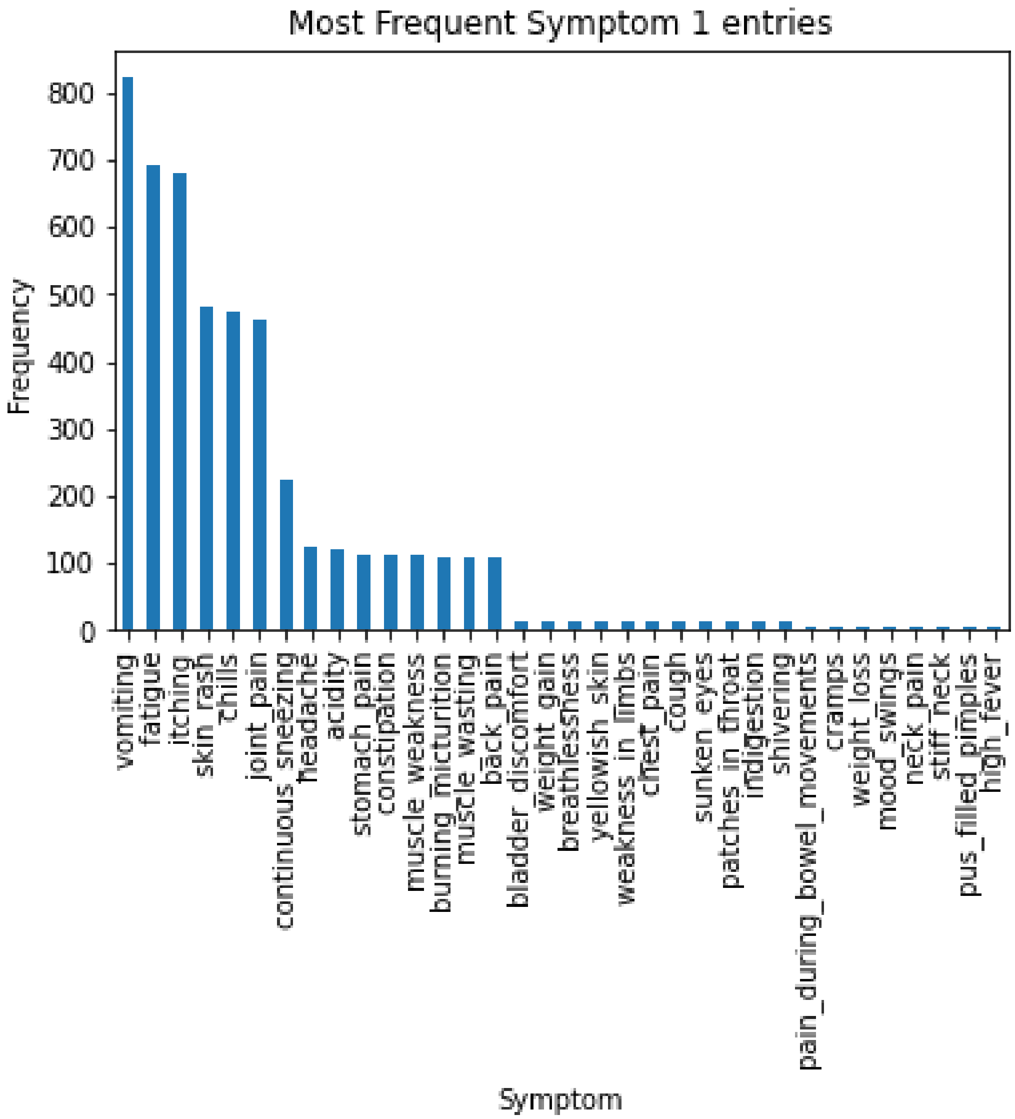
In this paper they went through various machine learning models to try and match them together to predict heart diseases. They used the UCI Cleveland dataset and performed data mining algorithms to get the important features. They used PCA to reduce the feature size. They then fit it to their hybrid models. They found the best hybrid model to be a combination of random forest and linear method. It performed with an accuracy of 88.7%

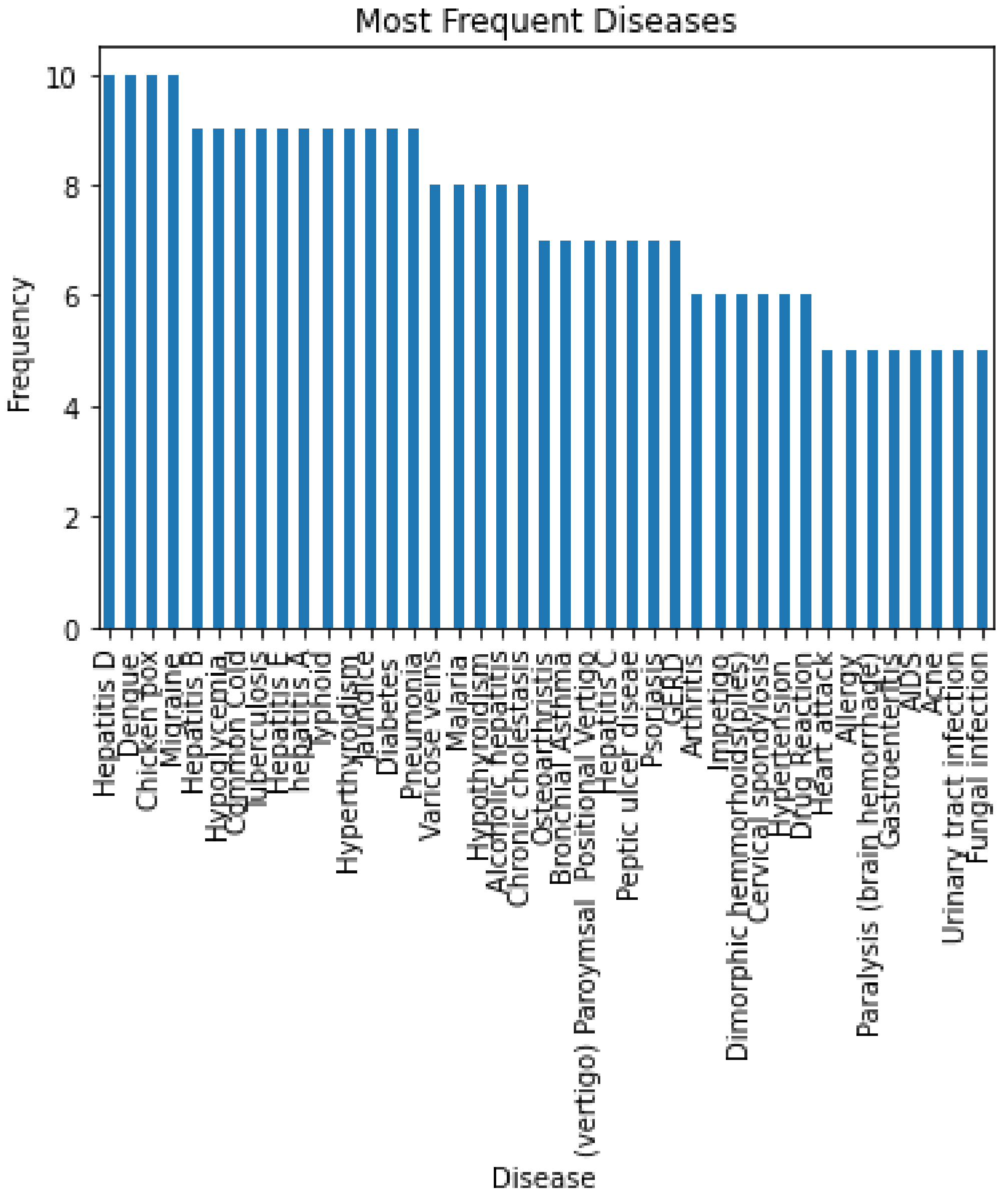
Exploratory Data Analysis

The dataset given provides a column for the disease name and columns for the names of the symptoms the disease has. It was collected through a team collecting information on the web and adding it into the dataset.


More Symptoms less entries
5000
4000
3000
2000
1000
0
counts
Disease
Symptom_5
Symptom_10
Symptom_15
Number of Symptoms




Most Frequent Symptom 1 entries
Frequency
800
700
600
500
400
300
200
100
0
vomiting
fatigue
itching
skin_rash
chills
joint_pain
continuous_sneezing
headache
acidity
stomach_pain
constipation
muscle_weakness
burning_micturition
muscle_wasting
back_pain
bladder_discomfort
weight_gain
breathlessness
yellowish_skin
weakness_in_limbs
chest_pain
cough
sunken_eyes
patches_in_throat
indigestion
shivering
pain_during_bowel_movements
cramps
weight_loss
mood_swings
neck_pain
stiff_neck
pus_filled_pimples
high_fever
Symptom

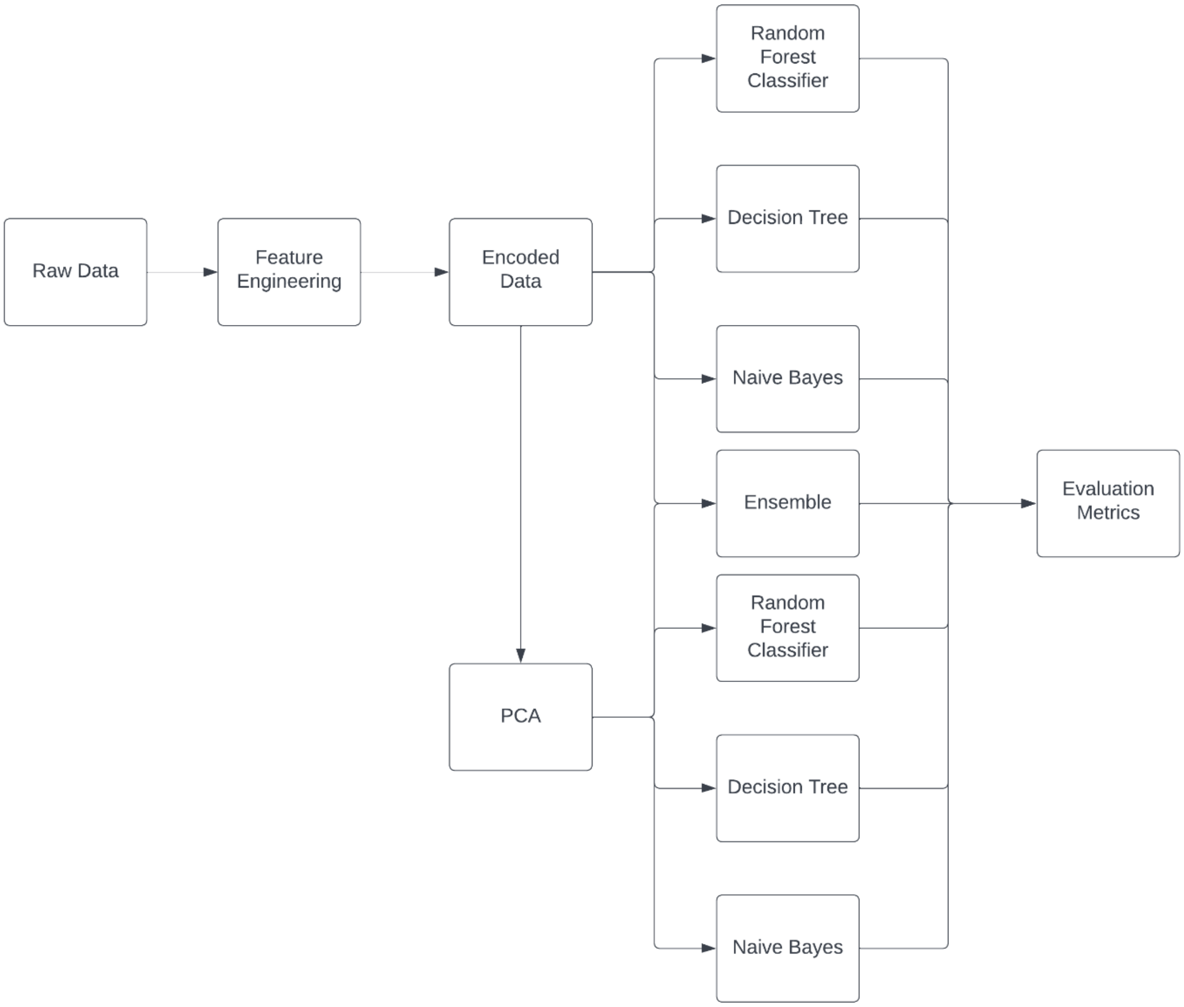

Most Frequent Diseases
Frequency
0
2
4
6
8
10
Disease
Hepatitis D
Dengue
Chicken pox
Migraine
Hepatitis B
Hypoglycemia
Common Cold
Tuberculosis
Hepatitis E
hepatitis A
Typhoid
Hyperthyroidism
Jaundice
Diabetes
Pneumonia
Varicose veins
Malaria
Hypothyroidism
Alcoholic hepatitis
Chronic cholestasis
Osteoarthristis
Bronchial Asthma
(vertigo) Paroymsal Positional Vertigo
Hepatitis C
Peptic ulcer diseae
Psoriasis
GERD
Arthritis
Impetigo
Dimorphic hemmorhoids(piles)
Cervical spondylosis
Hypertension
Drug Reaction
Heart attack
Allergy
Paralysis (brain hemorrhage)
Gastroenteritis
AIDS
Acne
Urinary tract infection
Fungal infection

Experiment Design

The three algorithms I will be using are: Random Forest Classifier, Naïve Bayes, and Decision Tree. I chose these algorithms because they are the best performing multiclass classification models. The evaluation metrics I will be using are accuracy score, confusion matrix, and f1 score. Accuracy score to see how accurate the model can be, confusion matrix to see which predictions it got incorrectly, and f1 score to see the precision and recall of the model. The baseline I will be comparing my model to are some of the models on Kaggle which has a baseline score of 98% accuracy score.

What I did first was data preprocessing and feature engineering. The data only gave the target column and the symptoms that came with that disease with different diseases their symptoms ranged from 4 – 17 symptoms. If the disease only had 4 symptoms the rest of the columns from 5-17 were NaN values. I transformed the dataset and engineered the features by converting the columns to unique symptom names and the disease target of course. Then if the disease has that symptom, I encoded 1 as true and 0 as false. I then encoded the target column from 1-41 depending on the disease's unique name.

When different people get the same diseases, they most likely get the same symptoms. Therefore, I have gotten so many duplicate rows, I dropped all the duplicate rows. I was left with a little over 300 rows coming from almost 5000 rows. After I dropped the duplicates, I plotted the correlation of each symptom with one another to get a visual to see which symptoms are correlated with one another. I then split the data to 20% test and 80% train. I then fit the train data to the models I have chosen which are: Random Forest, Decision Tree, and Naïve Bayes. I checked the cross validation of each model then I checked their accuracy score, plotted their predictions on a confusion matrix, and checked their f1 scores. To fine tune my models and to prevent overfitting I used a soft voting ensemble method classifier combining all of my previous models. I checked the ensemble's cross validation score, accuracy score, f1 score, and its predictions using the confusion matrix. I wanted to see if Principal Component Analysis could improve my model's scores. I preformed PCA on the encoded data set and managed to reduce 131 features down to 46. I then split the PCA dataset to 20% test 80% train and fit them back into my previous models and checked their cross-validation score, accuracy score, f1 score, and their predictions through the confusion matrix.



Experiment Results

After completing and observing the evaluation metrics of all three algorithms the best performing algorithm is the Random Forest Classifier with PCA. It performed with a 98% cross validation score, 100% accuracy score, and 100% f1 score. I believe that PCA helped boost the accuracy score of the random forest classifier because it helps eliminate the 'noise' features of the dataset. There were many symptoms in the dataset that did not correlate with one another therefore PCA removed some of those useless features. It also helped reduce overfitting on the dataset. Random forest without PCA performed with 96% cross validation score, 95% accuracy score, and 96% f1 score. PCA helped increase its cross-validation score by 2%, accuracy score by 5% and f1 score by 4%. Decision tree was the worst performing algorithm. Its cross-validation score was 60%, accuracy score was 57%, and f1 score was 59%. With PCA decision tree's cross validation score became 78%, accuracy score became 87%, and f1 score became 85%. Naïve bayes without PCA performed as well as Random Forest with PCA and scored 100% on accuracy, cross validation, and f1. However, it has the risk of overfitting because of its many 131 features. With PCA, Naïve Bayes' performance decreased to 83% cross validation score, 92% accuracy score, and 89% f1 score. Another method to avoid overfitting is using an ensemble method. I used a soft voting ensemble that consisted of the three algorithms I have chosen. It performed with a cross validation score of 99.6%, accuracy score of 98% and an f1 score of 99%. With PCA, the ensemble model's performance decreased to 92% cross validation, 96% accuracy, and 96% f1. PCA does not always improve the scores because it depends on the dataset and model used in the experiment. I believe the random forest performed the best because it is a bagging ensemble of 500 decision trees. With PCA to reduce overfitting, I believe it is the best model overall.

Conclusion

The main take away from my work is that the Random Forest Classifier is a very powerful algorithm that can be used in many different applications of machine learning. Machine learning could be used to change the way people have access to medical advisors. It could help medical professionals diagnose people faster. That would lead to medical services in less fortunate countries being less expensive. That PCA is also not going to improve all algorithms and datasets. For the case of random forest and this dataset, PCA managed to improve the results of cross validation, accuracy score, and f1 score. But it diminished the results of naïve bayes.

Potential ethical implications of this project are if the model predicts a disease incorrectly causing the user to diagnose the patient, it could risk very harmful consequences. Worst case scenario, the patient could die. The important question on who would be accountable for this terrible mistake? The designer of the model or the user that did not know better. There is also a risk of privacy, for example, when using a patient’s sensitive data. Is it ethical to store their private data to improve our model? Or should we discard their data after every use? If we discarded every patient’s data, the model would never get updated.

Bibliography

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC medical informatics and decision making*, *19*(1), 1-16.

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. *Ieee Access*, *5*, 8869-8879.

Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE access*, *7*, 81542-81554.