Torin White - 657467127

Junha Liu - 658846751

Binh Nguyen - 668379209

CS-412 (Introduction to Machine Learning)

Predictive ML Modeling for COVID-19 Patients

Based on Synthetic Data: Synthea

Introduction

As we delve into the domain of healthcare data science, our team is dedicated to the task of anticipating patient results. This endeavor comes with its fair share of challenges, such as handling sensitive information about patients and navigating the intricate process of gaining approval from Institutional Review Boards (IRBs). To overcome these obstacles, we have found a valuable solution in Synthea[1]. By utilizing this platform, we can generate synthetic, but representative health data that serves as a secure and ethically sound data source for our research on predictive modeling. Moreover, as Synthea provides data in the HL7 standard, FHIR schema it may be used as a testing ground for developing new models to iterate and later deploy in healthcare practice Our project has an overarching vision: to spearhead advancements in healthcare analytics by predicting disease diagnoses based on patient history and symptoms. With the assistance of Synthea's synthetic datasets, our objective is to refine and validate models that can predict patient outcomes meticulously. Accurate models may become a game-changer in the world of healthcare, where many rare and deadly diseases manifest before human-detectable symptoms and patterns emerge, with diagnoses coming too late for intervention, especially in the case of cancer, dementia, and Alzheimer’s.

Initial Approach

Our group focused on establishing a dementia prediction model, utilizing Synthea’s synthetic patient data and the well-known study “Synthea synthetic patient data for lung cancer risk prediction machine learning”[2]. However, the challenge lies in the low incidence of critical diseases and their correlated symptoms within Synthea datasets. With only a limited number of occurrences, it was deemed unfeasible to produce the number of patient records required with our group’s storage and computational resources. As a result, we've shifted our focus to COVID-19 cases—a critical illness with high incidence. This strategic pivot allows us to explore predictive modeling for critical diseases within a more prevalent context. Our goal is to leverage data science methodologies to enhance predictive analytics to build a model that can predict the likelihood of a patient dying of COVID-19 given diagnoses with COVID-19, their reported symptoms, and their prior health history.

Data Processing

Leveraging Synthea, we curated a dataset comprising mock data for 100,000 patients; using 88160 patient records indicating positive incidence of COVID-19, strategically focusing on COVID-related features. The dataset includes critical patient information encapsulated in features such as 'AGE,' 'RACE,' and 'GENDER,' forming a comprehensive patient physical record. To enhance the relevance of our predictive model, we integrated COVID-related symptoms, considering factors like 'Body mass index 30+ - obesity,' 'Cough,' 'Fever,' and others.

In addition to reported symptoms related to COVID-19 and general patient observations, the data includes cases with:

Negative cases of COVID-19: *Patients who, as predicted by the model, were suspected to have COVID-19 and sought diagnoses but ultimately tested negative.*

Positive cases of COVID-19: *Patients who, as predicted by the model, were suspected to have COVID-19 based on symptoms and tested positive.*

Of the positive cases the dataset contains:

Deceased Patients by COVID-19: *Patients who, as predicted by the model, succumbed to COVID-19.*

ICU-Admitted Patients after Infection: *Individuals forecasted to be admitted to the Intensive Care Unit (ICU) after contracting COVID-19.*

Patients Opting for Home Isolation: *Patients predicted to choose home isolation as a response to their COVID-19 diagnosis.*

Recorded Features **features are binary recorded*

'AGE', 'RACE' 'GENDER', 'Body mass index 30+ - obesity', 'Body mass index 40+ - severely obese', 'Chill', 'Cough', 'Diarrhea symptom', 'Dyspnea', 'Fatigue', 'Fever', 'Headache', 'Hemoptysis', 'Joint pain', 'Loss of taste', 'Muscle pain', 'Nasal congestion', 'Nausea', 'Passive conjunctival congestion', 'Respiratory distress', 'Sore throat symptom', 'Sputum finding', 'Vomiting symptom', 'Wheezing'

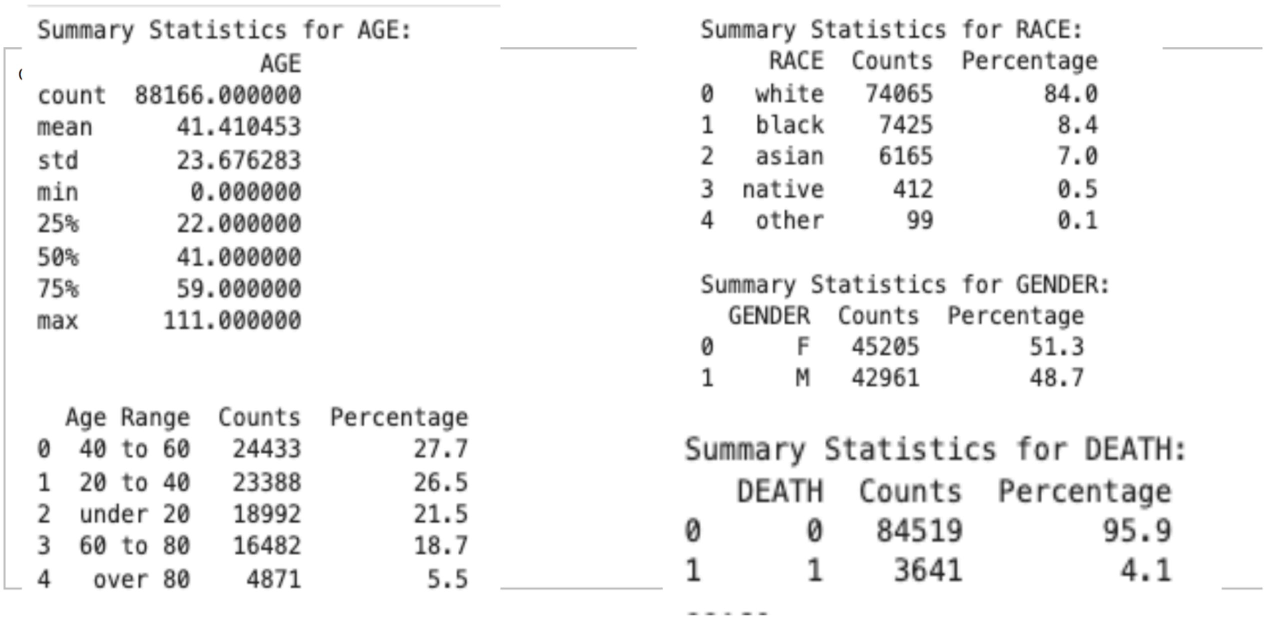
Data Preparation for Prediction Models

As we used synthetic data rather than real data, the data provided by Synthea was already clean and no missing data were present. However, many data transformations were required to convert the data from the FHIR JSON schema to the pandas data frame appropriate to apply machine learning models. Conversion of data from JSON format to CSV was provided by Synthea, resulting in approximately ten CSV files each containing a subset of patient data such as “PATIENT” including unique ID, patient name, and details; “OBSERVATIONS” including symptoms experienced by the patient and recorded by the doctor; “CONDITIONS” including the distinct diseases and conditions diagnosed by the doctor. Before the data could be applied to our model, we had to perform several joins, using patient-unique ID across the datasets. Then, we calculated age based on recorded birthdate and discretized AGE into 4 distinctive age ranges: 20-39, 40-59, 60- 79, and 80 and above. We also converted AGE and GENDER to binary variables. Finally, we recorded which patients with COVID-19 had deceased due to the disease. With these conversions, we maintain the dataset consistency and interpretability throughout the features, in which all indicating values are binary. By doing so, we optimize its compatibility for our models as it will become more transferable.

Prediction Models

Having prepared a dataset with all features and with a high contributing factor to our target. Our primary focus shifts towards building a robust predictive model to achieve a high accuracy rate in predicting deceased patients. The refined dataset incorporates critical patient information, demographic features, and specific COVID-related symptoms, forming a comprehensive foundation for accurate predictions. The prediction target is a fatality rate (DEATH). After reviewing the US census data[3], we found that the Synthea dataset was fairly representative of the US population, except for two RACE features, where 'white' is overrepresented in the Synthea dataset and 'black' is underrepresented.

Data Summary



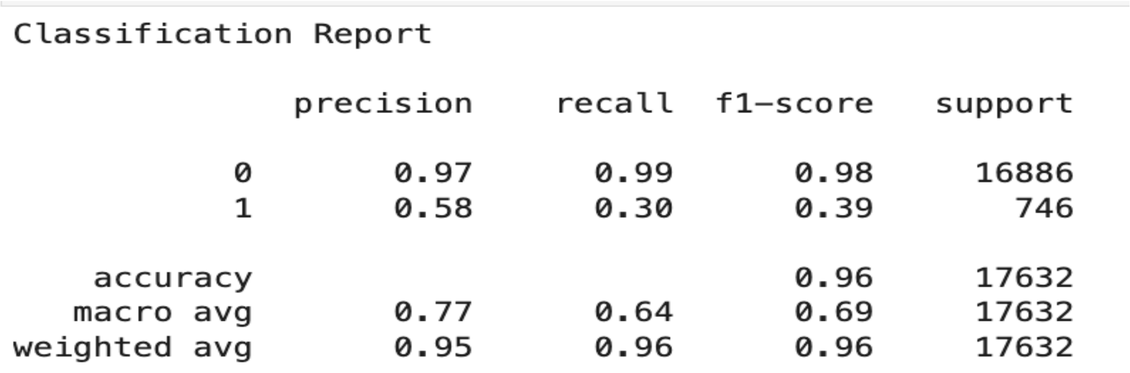
```
Summary Statistics for AGE:
                 AGE
count  88166.000000
mean      41.410453
std       23.676283
min        0.000000
25%       22.000000
50%       41.000000
75%       59.000000
max      111.000000


  Age Range  Counts  Percentage
0  40 to 60   24433        27.7
1  20 to 40   23388        26.5
2  under 20   18992        21.5
3  60 to 80   16482        18.7
4   over 80    4871         5.5
```

```
Summary Statistics for RACE:
     RACE  Counts  Percentage
0   white   74065        84.0
1   black    7425         8.4
2   asian    6165         7.0
3  native     412         0.5
4   other      99         0.1

Summary Statistics for GENDER:
  GENDER  Counts  Percentage
0      F   45205        51.3
1      M   42961        48.7

Summary Statistics for DEATH:
   DEATH  Counts  Percentage
0      0   84519        95.9
1      1    3641         4.1
```

Figure 1: summarized features (after AGE conversion)

Random Forest

Considerably dealing with a high-features dataset, the RF (Random-Forest) model is suitable for diversifying its training data set into each “decision tree” for the given feature data, recursively splitting its entries to predict the weights of each feature. However, due to its ability to minimize model overfitting. Its prediction for features’ weight is somewhat acceptably representable due to a higher generalization. However, for the given data set and the model’s intended purposes as a classification model, our RF model produces a fairly predictable outcome with a high accuracy rate. However, although the overall accuracy is high, the precision and recall are quite low in the cases of patient death, which is the most important criterion. In this case, as well, we would prefer the recall to be higher so we can identify all the cases where patients are at a higher risk of death, regardless of whether they die or not so that preventative measures or increased monitoring may be implemented.

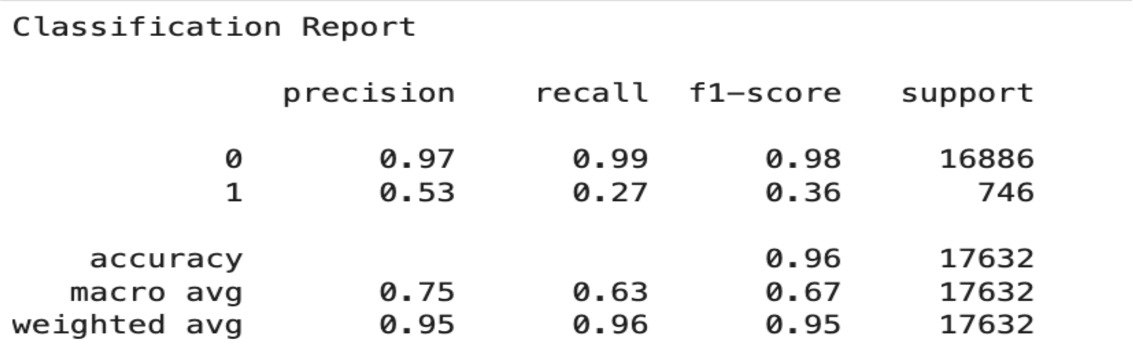


```
Classification Report

              precision    recall  f1-score   support

           0       0.97      0.99      0.98     16886
           1       0.58      0.30      0.39       746

    accuracy                           0.96     17632
   macro avg       0.77      0.64      0.69     17632
weighted avg       0.95      0.96      0.96     17632
```

Figure 2: Classification Report of Random Forest Model

KNN

On the other hand, we implement the KNN classification model to answer the same question, we chose KNN because of its simplicity because KNN doesn’t necessarily build a training model, eliminating the need for retraining when new data are added. However, this means higher memory demand for each K iteration (finding optimal K). During the process, we found the most optimal K neighbor to be 5, based on the accuracy rate and cross-validation result. Although the overall accuracy was comparable to RandomForest(though a bit less), the precision and recall for the cases where patients died were less than the Random Forest model with only a .53 precision score and .27 recall.



```
Classification Report

              precision    recall  f1-score   support

           0       0.97      0.99      0.98     16886
           1       0.53      0.27      0.36       746

    accuracy                           0.96     17632
   macro avg       0.75      0.63      0.67     17632
weighted avg       0.95      0.96      0.95     17632
```

Figure 3: Classification Report of KNN Model

Result Analysis

Using Random forest, we were able to determine the importance of each feature, contributing to the target outcome.

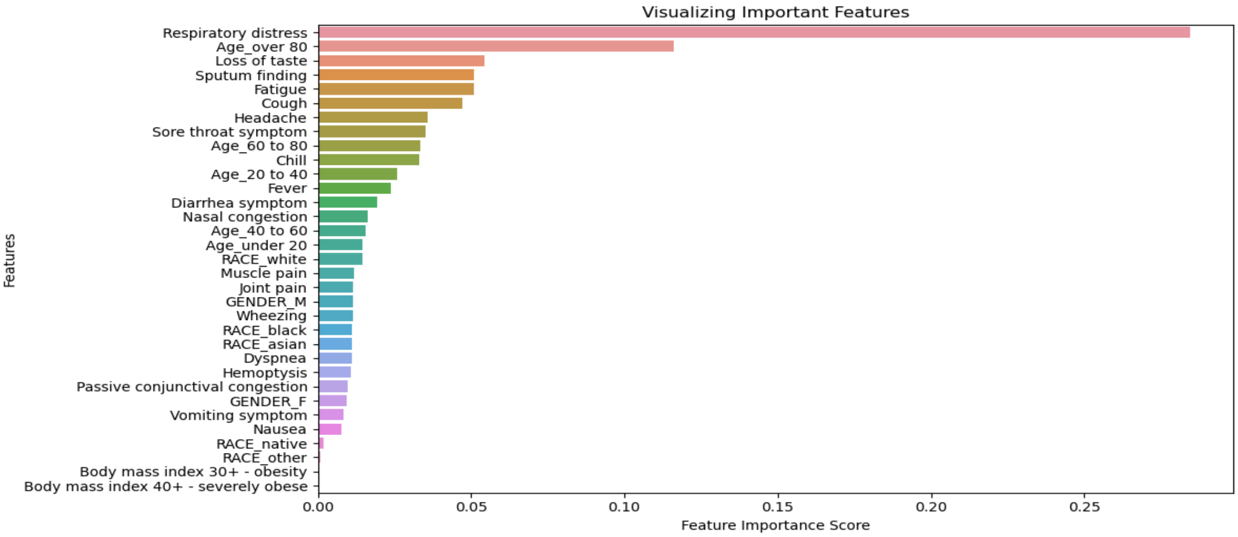
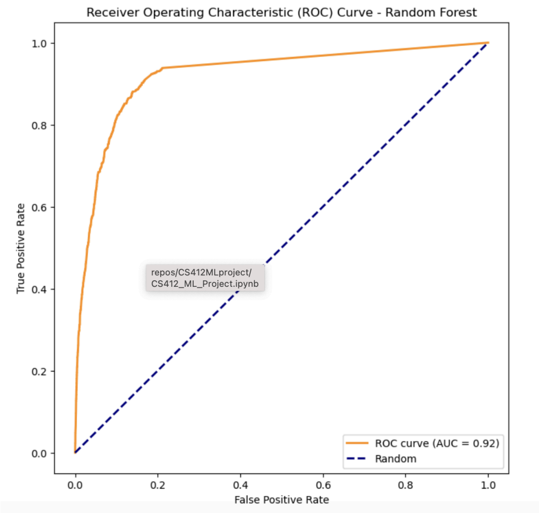
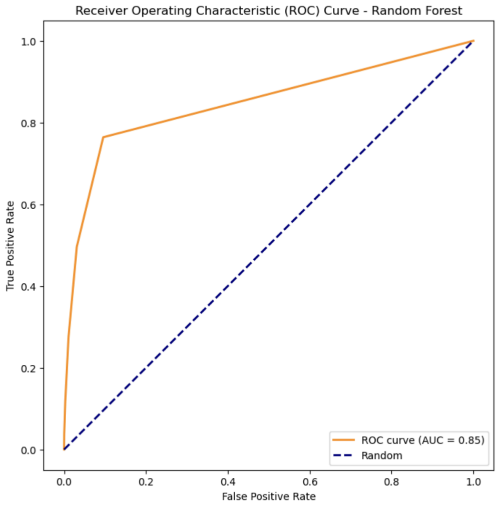


Figure 2: weight of features.

Respiratory distress or symptoms have the highest importance score which is accurately predicted, and comparable to real-life data for Covid-19. Also, loss of taste, fatigue, and headache are signs and symptoms that patients are most likely to experience when contracting the virus. Moreover, gender has limited influence on the target outcome, which is also understandably expected.

However, there are some noticeable discrepancies displayed in Fig. 2. For example, body mass index (BMI) is often a crucial factor in determining patients' health status, yet the data in Fig.2 doesn’t align with this expectation. At a closer look at the dataset, we saw that there were only several hundred patients with BMI >=30 and three with BMI >=40. This is not representative of the US population and is likely to blame for the discrepancy of the BMI features compared to real-life data which studies have shown obesity to be a large factor in the risk of dying from COVID-19[4].





One of the reasons for these discrepancies, shown in our findings, is the lack of entries for the target outcome, the data are neat yet only 4% of them are accountable to deceased patients. This introduced a lack of diversity, not in terms of the number of features, but usable entries.

One of the improvable factors for better prediction would be to increase the side of our test data which will prominently improve the accuracy rate of our models.

References

1. Synthea. https://synthetichealth.github.io/synthea/#home

2. Chen, AJ, 2022, "Synthea synthetic patient data for lung cancer risk prediction machine learning", https://doi.org/10.7910/DVN/GD5XWE , Harvard Dataverse, V3, UNF:6:y/jZNqbZHRiY9uqwt4W84Q== [fileUNF]

3. 2020 Census Results. https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-results.html

4. Sawadogo W, Tsegaye M, Gizaw A, Adera T. Overweight and obesity as risk factors for COVID-19-associated hospitalisations and death: systematic review and meta-analysis. BMJ Nutr Prev Health. 2022 Jan 19;5(1):10-18. doi: 10.1136/bmjnph-2021-000375. PMID: 35814718; PMCID: PMC8783972.