PREDICTION ON ANTIBIOTIC RESISTANCE AND LENGTH OF STAY FOR ICU PATIENTS

Bilal Hussain, Hon Wing Eric Chan, Kyoun Huh, Sakina Patanwala

1. MOTIVATION AND BACKGROUND

Machine Learning has been revolutionary when it comes to the banking and education system, and now it is time to implement it in healthcare in order to potentially bring social benefits to the public, such as delivering better patient treatment, improving drug efficiency, helping clinical professionals and medical caregivers, and simplifying hospital operations. Hence, we decided on exploring and hopefully contributing towards something beneficial to the healthcare system that we, as future data scientists, can implement and improvise.

Antibiotic resistance imposes a heavy burden on healthcare because due to the time it takes for sample and culture results to return, clinicians rely on prior experience and static guidelines to prescribe antibiotics without accounting for patient-specific attributes, leading to several fatalities. Such experimental treatment often fails to account for the changes in antibiotic resistance with time and geographical locations. Hence, our project's first goal is to predict if an organism is resistive or sensitive towards numerous antibiotics, measured by interpreting a culture test. Length of Stay and mortality rate are insightful measures for medical professionals and are utilized better to manage minimal health care resources in the hospital. Assigning the right amount of healthcare resources to the right patient is challenging. Therefore, the second goal of our project is to predict a patient's length of stay and mortality rate using their demographics and first 24 hours vital signs.

2. PROBLEM STATEMENT

2.1 Antibiotic Effectiveness Prediction

This study aimed to analyze the patient and hospital-specific features to treat bacterial infections of patients in intensive care units (ICUs). It was a complex task as we did not have access to actual data from Providence Healthcare due to confidentiality constraints. Therefore, we decided to use the MIMIC-III database. MIMIC-III is a gigantic database laid out with good documentation. In the healthcare domain, machine learning models need to be highly accurate and explainable for every possible situation, and as of now, there are no end-to-end solutions available. Therefore, we had to consult medical professionals to select features carefully as feature selection is a mindful process and can be fatal, especially in the healthcare domain.

The task is challenging because the organism starts getting resistant to the antibiotic overtime after several doses, and there is no data for the patient for that part of the process. Also, we had no record of the patient's medical history, i.e., if the patient has used the same antibiotic before. The database lacks data about the patient after the specimen has been collected from their body, and that leads to us losing the chance of taking those vital features into account that can affect the organism-antibiotic resistance levels on a very dynamic basis.

2.2 Length of Hospital Stay and Mortality Rate Prediction

The mortality rate and length of stay prediction rates help hospital staff pay more nuanced attention to patients in probable danger without having prior health care knowledge about that patient. The goal is to build machine learning models to make predictions and make the probability of each task available in a webpage to allow hospital staff to retrieve the prediction and help them to make critical health care decisions.

One of the most challenging parts of this task is collecting around 8000 features and 23944 rows to help us make predictions for the length of stay and mortality rate. As data scientists, we do not inhibit enough healthcare domain knowledge to filter features and select meaningful features from this large dataset to build machine learning models. Another challenging part is the data cleaning. MIMIC-III hides or modifies some valuable information to protect the patient's confidentiality. Age is one feature that we want to use as a feature, but the age has been randomly shifted. We need to normalize the age and understand the modification on age to use it as a probable feature.

3. DATA SCIENCE PIPELINE, METHODOLOGY, AND EVALUATION

Each goal has a different workflow. We will describe the data science pipeline, methodology and evaluation by goals.

3.1 Antibiotic Effectiveness Prediction

Figure 1 shows the data science pipeline used for achieving the predictive models. Each of the labels in the pipeline are explained in detail below.

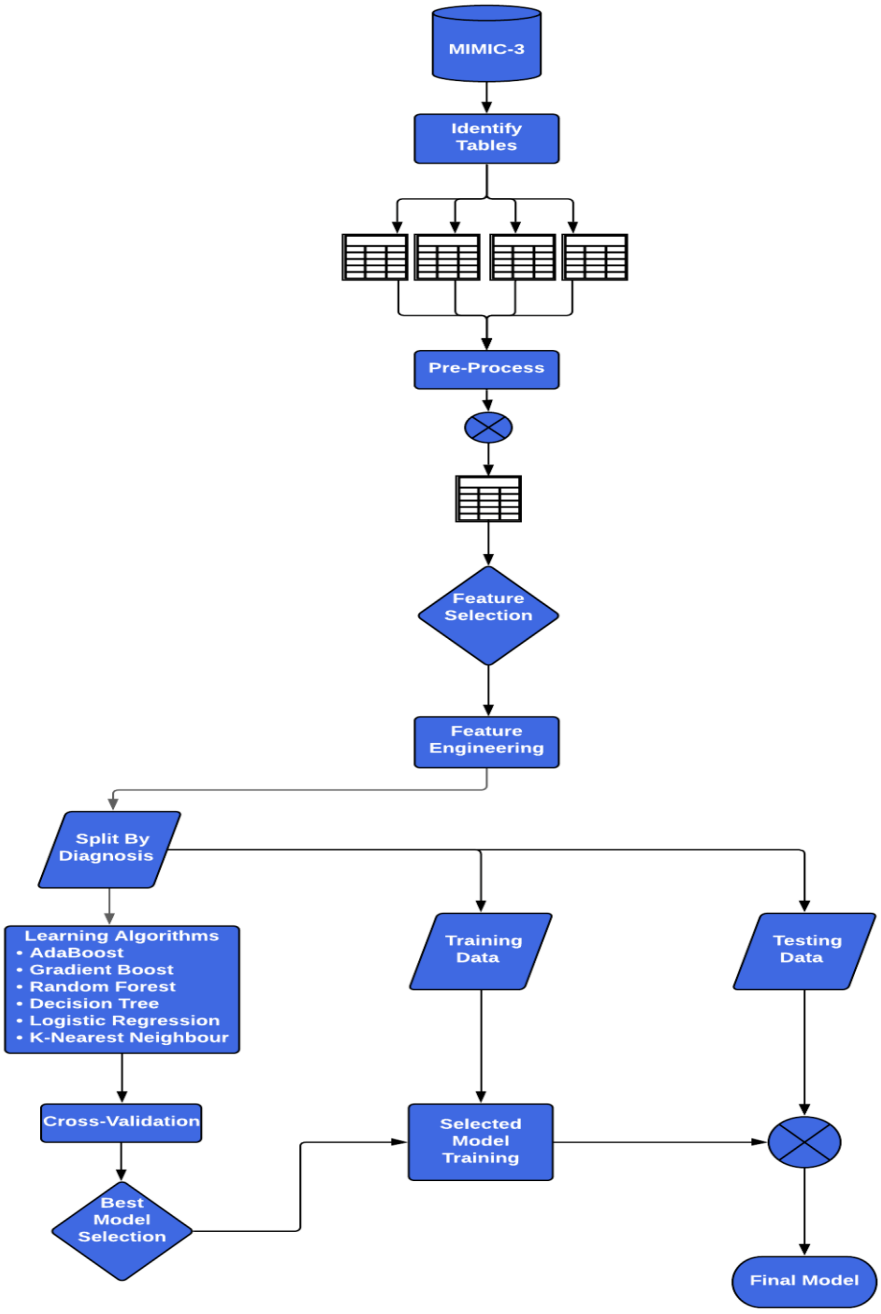

Figure 1. Data Science Pipeline for Antibiotic Effectiveness Prediction

3.1.1 DATA PREPROCESSING

i) Data Collection: We collected the patient's demographic information from the DEMOGRAPHICS table present in the MIMIC-III database for our goal. Since we concentrate on antibiotics and their effectiveness, we also used the table generated by the microbiology laboratory called MICROBIOLOGY EVENTS. To extract the different kinds of diagnoses, we used the DIAGNOSIS_ICD table, which had the icd9 codes, and the ADMISSIONS table for computing the total number of times a patient has been admitted.

ii) Data Cleaning: The DIAGNOSIS_ICD table had 6984 unique International Classification of Disease (ICD) codes with sparse data so that some diseases had only six patients' worth of data. In order to group the ICD codes better, we referred to the icd9-data website. We extracted the first three digits of every ICD code in order to put them into 19 different semi-broad diagnosis categories, namely 'endocrine, nutritional and metabolic diseases,' 'diseases of the blood-forming organs,' 'mental disorders,' 'diseases of the nervous system,' 'diseases of the circulatory system,' 'diseases of the genitourinary system,' 'symptoms, signs, and ill-defined conditions,' 'causes of injury and supplemental classification,' 'infectious and parasitic diseases,' 'diseases of the musculoskeletal system,' 'injury and poisoning,' 'neoplasms,' 'diseases of the respiratory system,' 'diseases of the skin and tissue,' 'diseases of the digestive system,' 'congenital anomalies,' 'certain conditions in the perinatal period' while excluding 'factors influencing health status and contact with health services and 'complications of pregnancy, childbirth, and the puerperium.' We had to drop two out of the nineteen types as the data for those two types was scanty.

In the original data, patients' age in their 90s was 300 and above, where 300 represented 90. We replaced 300 with 90 and calculated the other ages accordingly. For the missing age values, we imputed the rows with a mean age. Furthermore, we only kept rows where the culture was positive and dropped the remaining rows. Subsequently, we dropped the rows with missing antibiotic values. Our target column 'interpretation' for predicting antibiotic sensitivity contained three values, sensitive, resistant, and intermediate. We considered 'intermediate' as 'resistant' to exclude possible inappropriate use of antibiotics with less efficacy.

iii) Data Integration: We merged the MICROBIOLOGY EVENTS, DEMOGRAPHICS, and DIAGNOSIS-ICD tables using an inner join on ‘subject-id’ and ‘hadm-id’ to produce our final data frame. Further, we joined this produced table with an ADMISSIONS table using a left join to compute the total admission count for each patient. A schema demonstrating the same is shown using Figure 2 below:

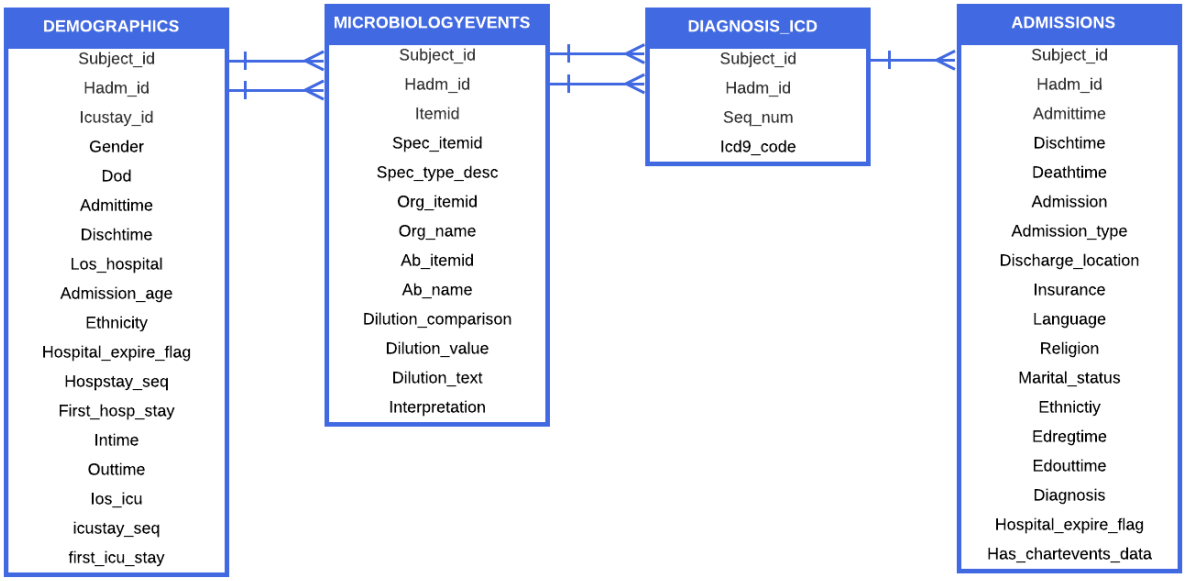

Figure 2. Table integration using different joins on columns

iv) <u>Data Analysis:</u> We plotted a few graphs using Dash on Plotly to understand better the data we are working with and how to utilize the data frame to move forward for model building.

Figure 3 represents the top 10 prevalent bacteria that mainly occur in patients and the count of each of these organisms. Below from Figure 3 see that Staphylococcus Aureus Coagulase Positive, Klebsiella Penumeniue, Pseudomonas Aeruginosa are most common in patients with a high interval between the hospital admit time and sample collection time; stipulating that these infections are most likely to be hospital-acquired. Past studies also suggested that bloodstream infections due to Pseudomonas aeruginosa are usually hospital-acquired.

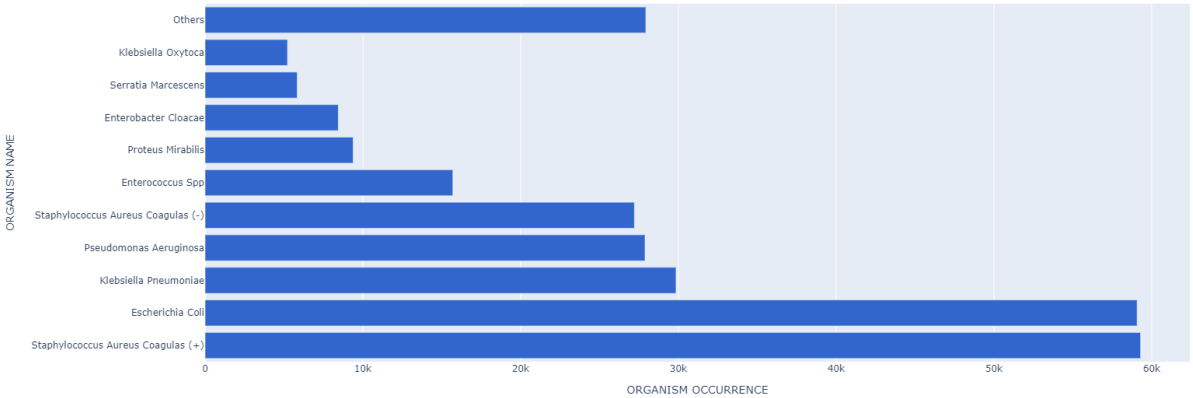

Figure 3. Total Organism Occurrence in Patients

Figure 4 shows the percentage of the most frequently collected specimens from patients. These specimens are sent for culture tests that interpret if the organism is sensitive or resistant to the antibiotic. We see that urine, sputum, and blood culture are three of the most common specimens collected over ten years.

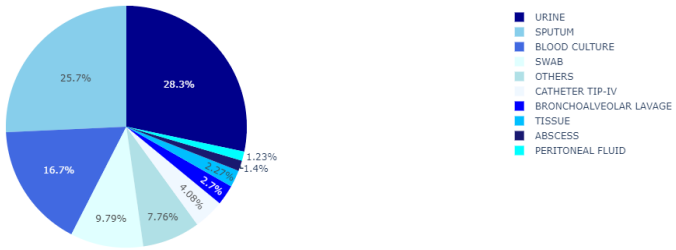

Figure 4. Percentage of top 10 collected specimens.

To examine if gender possibly can be one of the factors affecting the resistance of an organism, we plotted to see how several diagnoses are distributed amongst males and females. As we see from Figure 5, most diagnoses are nearly distributed amongst the two genders, but diagnoses like ‘Injury and Poisoning’ and ‘Diseases of the Musckolosecal System’ are highly found in males than females.

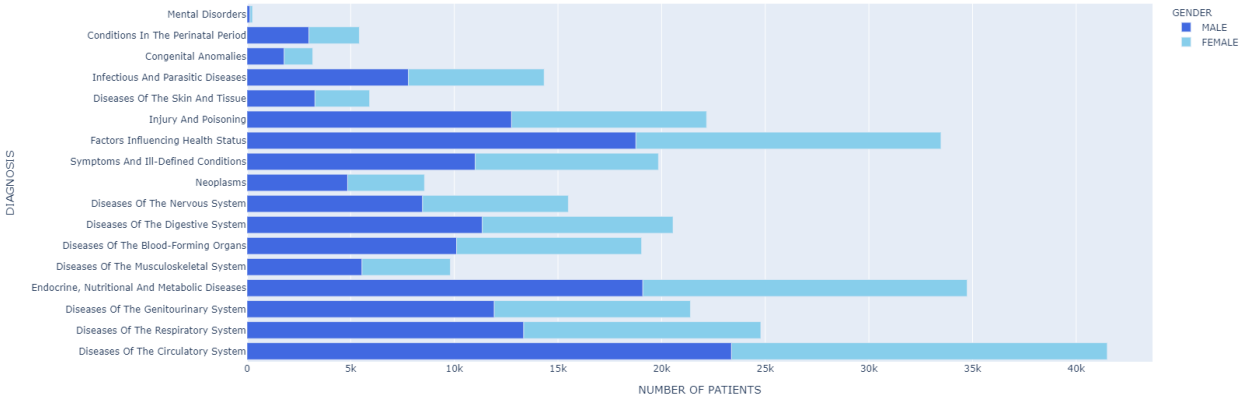

Figure 5. Count of patients distinguished by genders for categorized diagnosis.

The below sunburst plot in Figure 6 is used to demonstrate the relationship between organisms, antibiotics, and their interpretability. For every antibiotic, we see that it has more sensitive values than resistive, and that makes sense because most antibiotics are expected to be effective towards the organism present in the patient's body.

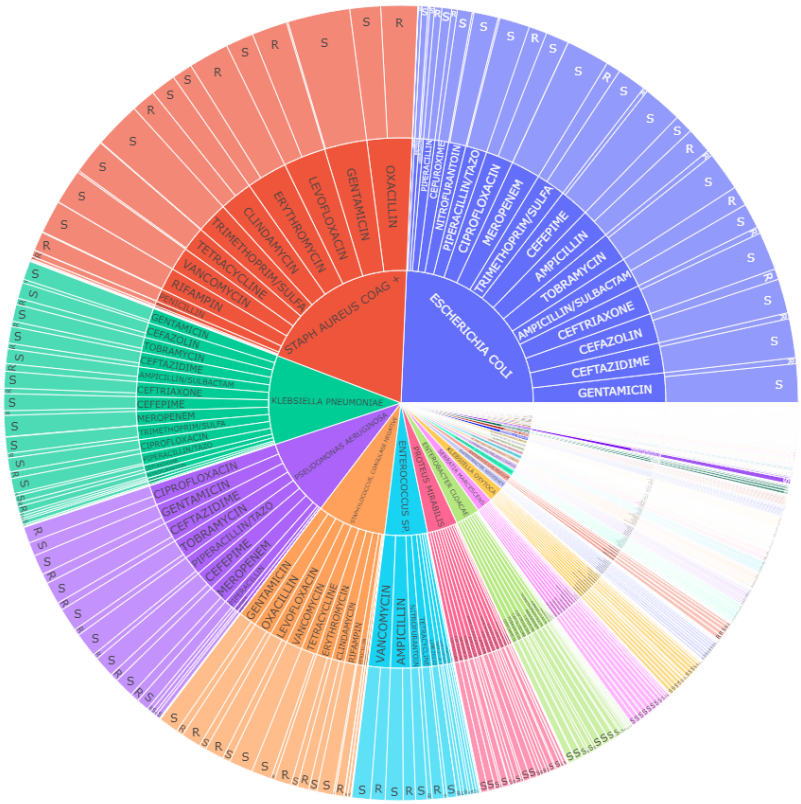

Figure 6. Interpretation patterns for antibiotic susceptibility against organisms.

3.1.2 FEATURE SELECTION

A set of 8 features (age, gender, ethnicity, diagnosis, antibiotic interpretation, antibiotic name, organism name, specimen type) were selected from the above-integrated tables, and two (sample collection interval and previous admissions) were computed. We preferred to select these features after in-depth advice from medical professionals, clinicians, medical caregivers, and a thorough research and medical literature review instead of relying on an algorithm for automated feature selection to have total liability and user acceptance for the results obtained from our predictive model.

i) <u>Derived Features:</u> We derived two custom features using the existing columns from the data frames based on the data analysis. We computed a feature named 'Sample Collection Interval' using the difference of the interval between the time the organism sample was collected and the patient's admission time in the hospital. We noticed that the number of times a patient is admitted to the hospital also plays a major role in the antibiotic-bacteria pair and computed a

'Previous Admissions' feature by counting the number of times a 'subject-id' was repeated for each patient.

ii) Feature Specification: The table below shows the details for each chosen and computed feature.

Feature Name	Feature Extraction	Feature Type	Feature Categorization	Feature Description
Age	DEMOGRAPHICS	Continuous	Age 0 -15, Age 15 - 30, Age 31 - 60, Age 60+	Age of the patient provided in years upon hospital admission time.
Gender	DEMOGRAPHICS	Categorical	Male: 5817 samples Female: 5096 samples	Gender of the patient upon hospital admission time.
Ethnicity	DEMOGRAPHICS	Categorical	White: 7852 samples Black: 931 samples Hispanic: 368 samples Asian: 275 samples Native: 10 samples Other: 316 samples Unknown: 1185 samples	Social group of the patient upon hospital admission type.
Diagnosis	DIAGNOSIS_ICD	Categorical	Endocrine, nutritional and metabolic diseases: 141227 samples Diseases of the blood-forming organs: 96283 samples Mental disorders: 976 samples Diseases of the nervous system: 71836 samples Diseases of the circulatory system: 168944 samples Diseases of the genitourinary system: 129345 samples Symptoms, signs, and ill-defined conditions: 104755 samples Causes of injury and supplemental classification: 110933 samples Infectious and parasitic diseases: 125231 samples	Diagnosis assigned to the patient upon hospital admission or even updated further on.

			diseases of the musculoskeletal system: 44179 samples Injury and poisoning: 137892 samples Neoplasms: 33305 samples Diseases of the respiratory system: 149911 samples Diseases of the skin and tissue: 51363 samples Diseases of the digestive system:107122 samples Congenital anomalies: 8244 samples Certain conditions in the perinatal period: 3936 samples	
Specimen Type	MICROBIOLOGY EVENTS	Categorical	Sputum: 459298 samples Urine: 372930 samples Blood culture: 193937 samples Swab: 151398 samples Catheter Tip-IV: 65052 samples Bronchoalveolar Lavage: 49500 samples Tissue: 34307 samples Abscess: 21757 samples Peritoneal Fluid: 20941 samples Spinal Fluid: 4206 samples Other: 112705 samples	Specimen is a sample which is tested for bacterial growth and is derived from a patient
Organism Name	MICROBIOLOGY EVENTS	Categorical	126 different types of organisms	The organism which grew when tested.
Antibiotic Name	MICROBIOLOGY EVENTS	Categorical	30 different types of antibiotics	Antibiotic that was tested against the given organism.
Antibiotic Interpretation	MICROBIOLOGY EVENTS	Categorial	Resistive: 500687 samples Sensitive: 985344 samples	Analysis of the antibiotic sensitivity, and indicates the results of the test.
Sample Collection	Calculated using ‘subject-id’ from ADMISSIONS	Float	1486031 total sample collection intervals with 0.1 being the lowest interval and 272.1 being the highest	Interval in days between hospital admission time and sample collection time.

Interval			interval	
Previous Admission	Calculated using 'admit-time' from DEMOGRAPHICS and 'chart-time' from MICROBIOLOGY EVENTS	Integer	1486031 total previous admissions with 42 being the highest and 1 being the lowest	Number of times patient was previously admitted before current admission.

3.1.3 FEATURE ENGINEERING

i) Feature Scaling: Standard Scalar was used to standardize the data such that it has a mean of 0 and a standard deviation of 1.

ii) Feature Transformation: We had categorial features with no ordinal relationship between them; therefore, we used one-hot encoding to convert them into numerical data for model training. We had ten columns in the original dataset; after encoding, we ended up with 181 numerical columns.

3.1.4 MODEL BUILDING

Before building our model, we split our existing data frames on the diagnosis column. We build 17 CSV files from the data frame after splitting each diagnosis type to help ease the perfect classification algorithm for each diagnosis. Also, to ease using our data product for medical professionals in the future, they can search for antibiotics based on the specific patient diagnosis rather than the actual data, which is about 1.4 Million rows.

After data preprocessing and careful feature selection, transformation, and scaling, we devised our binary classification algorithm to train our classifier for predicting the susceptibility of the bacteria to the administered antibiotic. The target label for each of our models is the Dissolution Test Interpretation which leads us to a binary classification model. The interpretation column has only two types of values: sensitive and resistive, which points towards an antibiotic's effectiveness towards an organism present in a patient. The data collected from the patient's time is admitted to the hospital until the specimen is collected and used as the prediction variables.

We made sure to use different models to find the best fitting and enhanced model for each type of diagnosis. We used tree-based ensemble learners like random forest and decision tree classifiers, non-parametric classifiers like k-nearest algorithm, binary output suitable classifiers like logistic regression, and boosting algorithms like adaptive and gradient boosting. However, one major drawback of choosing multiple models is increased time and structured complexity. Hence, we tried to greedily pick six models for cross-validation but ensured the prediction errors between them were low or uncorrelated.

To make sure our model was not overfitting the data, i.e., failing to generalize the pattern, we used Repeated Stratified k-fold with k = 10. Repeated Stratified improves the model performance. It simply repeats the cross-validation procedure multiple times and reports the mean result across all folds from the run.

The best performing model was identified, Gradient Boost in our case, through Repeated Stratified cross-validation. We randomly split the data of each type of diagnosis into 70% for training and 30% for testing. To check for overfitting and underfitting, we further divided the training set, keeping 15% for validation and 15% for training. We implemented a meta-learner with six different models to get better results, but we could not get a better performance than Gradient Boost pr Random Forest and decided to scratch the meta-learner and hopefully implement it in the future so it can successfully help us improve accuracy.

After selecting the best model for each diagnosis, we fine-tuned its parameters to enhance performance further. We used Gradient Boost with 200 n_estimators, learning_rate of 0.3 and max_depth of 10, and Random Forest with 200 n_estimators, entropy criterion, and balanced class_weight different datasets.

3.1.5 MODEL EVALUATION METRICS

i) Cross Validation: The following table contains the evaluation results of cross-validation of six different supervised models on all 17 diagnosis datasets.

DIAGNOSIS	SUPERVISED MODEL	ACCURACY	AUC-ROC SCORE	PRECISION	RECALL	F-1 SCORE
Mental Disorders	Gradient Boosting Classifier	0.8583	0.9025	0.8792	0.9172	0.8975
Mental Disorders	Random Forest Classifier	0.8839	0.9209	0.8923	0.9429	0.9165
Mental Disorders	AdaBoost Classifier	0.8433	0.8938	0.8688	0.9061	0.8867
Mental Disorders	Logistic Regression	0.8412	0.8861	0.8668	0.9056	0.8853
Mental Disorders	Decision Tree Classifier	0.8395	0.8195	0.8857	0.8773	0.8809
Mental Disorders	K-Neighbors Classifier	0.8395	0.8861	0.8637	0.9071	0.8844
Congenital Anomalies	Gradient Boosting Classifier	0.7992	0.8449	0.8168	0.9234	0.8668
Congenital Anomalies	Random Forest Classifier	0.8448	0.9084	0.8636	0.9271	0.8942

Congenital Anomalies	AdaBoost Classifier	0.8005	0.8477	0.8185	0.9227	0.8674
Congenital Anomalies	Logistic Regression	0.7866	0.8270	0.8056	0.9205	0.8592
Congenital Anomalies	Decision Tree Classifier	0.8143	0.7784	0.8710	0.8657	0.8683
Congenital Anomalies	K-Neighbors Classifier	0.7918	0.8312	0.8330	0.8828	0.8571
Injury And Poisoning	Gradient Boosting Classifier	0.8453	0.9160	0.8708	0.8984	0.8844
Injury And Poisoning	Random Forest Classifier	0.8167	0.8858	0.8439	0.8855	0.8642
Injury And Poisoning	AdaBoost Classifier	0.7669	0.8223	0.7916	0.8769	0.8320
Injury And Poisoning	Logistic Regression	0.7612	0.8201	0.7890	0.8701	0.8275
Injury And Poisoning	Decision Tree Classifier	0.7838	0.7632	0.8391	0.8312	0.8351
Injury And Poisoning	K-Neighbors Classifier	0.7825	0.8468	0.8299	0.8424	0.8361
Neoplasms	Gradient Boosting Classifier	0.8089	0.8625	0.8348	0.9011	0.8667
Neoplasms	Random Forest Classifier	0.8318	0.8963	0.8534	0.9128	0.8821
Neoplasms	AdaBoost Classifier	0.7851	0.8268	0.8090	0.9011	0.8526
Neoplasms	Logistic Regression	0.7808	0.8251	0.8089	0.8931	0.8489
Neoplasms	Decision Tree Classifier	0.7899	0.7591	0.8515	0.8422	0.8468
Neoplasms	K-Neighbors Classifier	0.7964	0.8439	0.8389	0.8723	0.8552
Diseases of the Circulatory System	Gradient Boosting Classifier	0.8447	0.9134	0.8704	0.9042	0.887
Diseases of the Circulatory System	Random Forest Classifier	0.8149	0.8803	0.8442	0.8894	0.8662
Diseases of the	AdaBoost	0.7756	0.8256	0.8006	0.8883	0.8422

Circulatory System	Classifier					
Diseases of the Circulatory System	Logistic Regression	0.7681	0.8229	0.7986	0.8772	0.836
Diseases of the Circulatory System	Decision Tree Classifier	0.7818	0.7556	0.8412	0.8334	0.8373
Diseases of the Circulatory System	K-Neighbors Classifier	0.7903	0.8495	0.8374	0.8548	0.846
Diseases of the Digestive System	Gradient Boosting Classifier	0.8453	0.9163	0.8701	0.8971	0.8834
Diseases of the Digestive System	Random Forest Classifier	0.8139	0.8829	0.8393	0.8846	.8614
Diseases of the Digestive System	AdaBoost Classifier	0.7620	0.8171	0.7843	0.8771	0.8281
Diseases of the Digestive System	Logistic Regression	0.7574	0.8155	0.7813	0.8732	0.8247
Diseases of the Digestive System	Decision Tree Classifier	0.7795	0.7608	0.8351	0.8257	0.8304
Diseases of the Digestive System	K-Neighbors Classifier	0.7790	0.8428	0.8245	0.8407	0.8325
Diseases of The Blood and Blood Forming Organs	Gradient Boosting Classifier	0.8444	0.9147	0.8701	0.8970	0.8833
Diseases of The Blood and Blood Forming Organs	Random Forest Classifier	0.8156	0.8832	0.8416	0.8861	0.8633
Diseases of The Blood and Blood Forming Organs	AdaBoost Classifier	0.7611	0.8194	0.7842	0.8779	0.8284
Diseases of The Blood and Blood Forming Organs	Logistic Regression	0.7546	0.8176	0.7810	0.8706	0.8234
Diseases of The Blood and Blood Forming Organs	Decision Tree Classifier	0.7780	0.7576	0.8344	0.8260	0.8302
Diseases of The Blood and Blood Forming Organs	K-Neighbors Classifier	0.7774	0.8427	0.8267	0.8365	0.8316
Diseases Of The	Gradient Boosting	0.8454	0.9158	0.8701	0.8988	0.8842

Respiratory System	Classifier					
Diseases Of The Respiratory System	Random Forest Classifier	0.8148	0.8851	0.8419	0.8839	0.8624
Diseases Of The Respiratory System	AdaBoost Classifier	0.7716	0.8272	0.7942	0.8801	0.8350
Diseases Of The Respiratory System	Logistic Regression	0.7660	0.8255	0.7933	0.8703	0.8300
Diseases Of The Respiratory System	Decision Tree Classifier	0.7830	0.7630	0.8379	0.8301	0.8340
Diseases Of The Respiratory System	K-Neighbors Classifier	0.7880	0.8533	0.8327	0.8474	0.8400
Diseases Of The Skin And Subcutaneous Tissue	Gradient Boosting Classifier	0.8472	0.9193	0.8711	0.8858	0.8784
Diseases Of The Skin And Subcutaneous Tissue	Random Forest Classifier	0.8205	0.8938	0.8414	0.8773	0.8590
Diseases Of The Skin And Subcutaneous Tissue	AdaBoost Classifier	0.7526	0.8181	0.7736	0.8524	0.8111
Diseases Of The Skin And Subcutaneous Tissue	Logistic Regression	0.7505	0.8167	0.7756	0.8438	0.8082
Diseases Of The Skin And Subcutaneous Tissue	Decision Tree Classifier	0.7843	0.7740	0.8309	0.8209	0.8259
Diseases Of The Skin And Subcutaneous Tissue	K-Neighbors Classifier	0.7704	0.8435	0.8174	0.8132	0.8153
Diseases Of The Musculoskeletal System And Connective Tissue	Gradient Boosting Classifier	0.8502	0.9155	0.8781	0.9032	0.8905
Diseases Of The Musculoskeletal System And Connective Tissue	Random Forest Classifier	0.8238	0.8880	0.8481	0.8997	0.8732
Diseases Of The Musculoskeletal System And Connective Tissue	AdaBoost Classifier	0.7671	0.8217	0.7914	0.8888	0.8372

Diseases Of The Musculoskeletal System And Connective Tissue	Logistic Regression	0.7612	0.8190	0.7887	0.8820	0.8328
Diseases Of The Musculoskeletal System And Connective Tissue	Decision Tree Classifier	0.7838	0.7571	0.8423	0.8358	0.8390
Diseases Of The Musculoskeletal System And Connective Tissue	K-Neighbors Classifier	0.7829	0.8411	0.8298	0.8529	0.8412
Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders	Gradient Boosting Classifier	0.8449	0.9132	0.8709	0.9028	0.8866
Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders	Random Forest Classifier	0.8141	0.8802	0.8433	0.8881	0.8651
Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders	AdaBoost Classifier	0.7701	0.8210	0.7941	0.8875	0.8382
Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders	Logistic Regression	0.7632	0.8189	0.7923	0.8771	0.8326
Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders	Decision Tree Classifier	0.7810	0.7559	0.8400	0.8323	0.8361
Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders	K-Neighbors Classifier	0.7858	0.8458	0.8338	0.8504	0.8420
Diseases Of The Genitourinary System	Gradient Boosting Classifier	0.8452	0.9133	0.8705	0.9023	0.8861
Diseases Of The	Random Forest	0.8148	0.8799	0.8435	0.8870	0.8647

Genitourinary System	Classifier					
Diseases Of The Genitourinary System	AdaBoost Classifier	0.7651	0.8186	0.7890	0.8846	0.8341
Diseases Of The Genitourinary System	Logistic Regression	0.7669	0.8194	0.7897	0.8870	0.8355
Diseases Of The Genitourinary System	Decision Tree Classifier	0.7806	0.7570	0.8394	0.8300	0.8347
Diseases Of The Genitourinary System	K-Neighbors Classifier	0.7830	0.8439	0.8322	0.8453	0.8387
Certain Conditions Originating In The Perinatal Period	Gradient Boosting Classifier	0.8775	0.9133	0.9106	0.9232	0.9168
Certain Conditions Originating In The Perinatal Period	Random Forest Classifier	0.8830	0.9370	0.9057	0.9378	0.9214
Certain Conditions Originating In The Perinatal Period	AdaBoost Classifier	0.8932	0.9381	0.9139	0.9429	0.9281
Certain Conditions Originating In The Perinatal Period	Logistic Regression	0.8995	0.9362	0.9262	0.9373	0.9317
Certain Conditions Originating In The Perinatal Period	Decision Tree Classifier	0.8581	0.8228	0.9050	0.9006	0.9027
Certain Conditions Originating In The Perinatal Period	K-Neighbors Classifier	0.8876	0.9283	0.9202	0.9269	0.9234
Symptoms, Signs, And Ill-Defined Conditions	Gradient Boosting Classifier	0.8455	0.9161	0.8707	0.8969	0.8836
Symptoms, Signs, And Ill-Defined Conditions	Random Forest Classifier	0.8167	0.8852	0.8420	0.8858	0.8634
Symptoms, Signs, And Ill-Defined Conditions	AdaBoost Classifier	0.7654	0.8240	0.7883	0.8766	0.8301
Symptoms, Signs, And Ill-Defined Conditions	Logistic Regression	0.7596	0.8230	0.7860	0.8690	0.8254

Symptoms, Signs, And Ill-Defined Conditions	Decision Tree Classifier	0.7824	0.7635	0.8375	0.8278	0.8326
Symptoms, Signs, And Ill-Defined Conditions	K-Neighbors Classifier	0.7781	0.8443	0.8263	0.8365	0.8314
Factors Influencing Health Status	Gradient Boosting Classifier	0.8471	0.9153	0.8739	0.9020	0.8877
Factors Influencing Health Status	Random Forest Classifier	0.8173	0.8853	0.8452	0.8906	0.8673
Factors Influencing Health Status	AdaBoost Classifier	0.7702	0.8248	0.7959	0.8837	0.8375
Factors Influencing Health Status	Logistic Regression	0.7651	0.8227	0.7935	0.8780	0.8336
Factors Influencing Health Status	Decision Tree Classifier	0.7842	0.7597	0.8422	0.8343	0.8382
Factors Influencing Health Status	K-Neighbors Classifier	0.7873	0.8492	0.8359	0.8492	0.8425
Diseases Of The Nervous System And Sense Organs	Gradient Boosting Classifier	0.8539	0.9170	0.8814	0.9118	0.8964
Diseases Of The Nervous System And Sense Organs	Random Forest Classifier	0.8242	0.8862	0.8509	0.9049	0.8771
Diseases Of The Nervous System And Sense Organs	AdaBoost Classifier	0.7739	0.8203	0.7995	0.8993	0.8465
Diseases Of The Nervous System And Sense Organs	Logistic Regression	0.7642	0.8142	0.7926	0.8937	0.8401
Diseases Of The Nervous System And Sense Organs	Decision Tree Classifier	0.7855	0.7523	0.8487	0.8403	0.8445
Diseases Of The Nervous System And Sense Organs	K-Neighbors Classifier	0.7884	0.8424	0.8365	0.8634	0.8497
Infectious And Parasitic Diseases	Gradient Boosting Classifier	0.8419	0.9142	0.8655	0.8937	0.8794

Infectious And Parasitic Diseases	Random Forest Classifier	0.8111	0.8820	0.8377	0.8770	0.8569
Infectious And Parasitic Diseases	AdaBoost Classifier	0.7593	0.8201	0.7812	0.8705	0.8234
Infectious And Parasitic Diseases	Logistic Regression	0.7559	0.8184	0.7796	0.8663	0.8207
Infectious And Parasitic Diseases	Decision Tree Classifier	0.7793	0.7625	0.8323	0.8237	0.8279
Infectious And Parasitic Diseases	K-Neighbors Classifier	0.7761	0.8432	0.8244	0.8294	0.8269

ii) <u>Model Performance for Different Diagnosis over 70:30 split:</u> The following table contains the Gradient Boost, Random Forest, and Logistic Regression results as they were the top three models for most diagnosis data frames with the highest accuracy after being fine-tuned and trained on 70% of the data for all the diagnosis data frames.

Diagnosis	Supervised Model	Accuracy	AUC-ROC score	Precision	Recall	F-1 score
Mental Disorders	Random Forest Classifier	88.40%	0.9278	0.8868	0.9495	0.9171
Congenital Anomalies	Random Forest Classifier	83.27%	0.9002	0.8531	0.9223	0.8863
Injury and Poisoning	Gradient Boost	83.98%	0.9097	0.8651	0.8965	0.8805
Neoplasms	Random Forest Classifier	82.92%	0.8920	0.8496	0.8496	0.8806
Diseases Of The Circulatory System	Gradient Boost	83.58%	0.9037	0.8662	0.8944	0.8801
Diseases Of The Digestive System	Gradient Boost	84.00%	0.9099	0.8676	0.8911	0.8792
Diseases of The Blood and Blood Forming Organs	Gradient Boost	83.65%	0.9082	0.8649	0.8901	0.8773
Diseases Of The	Gradient Boost	83.78%	0.9108	0.8641	0.8934	0.8785

Respiratory System						
Diseases Of The Skin And Subcutaneous Tissue	Gradient Boost	83.60%	0.9090	0.8597	0.8805	0.8700
Diseases Of The Musculoskeletal System And Connective Tissue	Gradient Boost	84.08%	0.9072	0.8729	0.8940	0.8833
Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders	Gradient Boost	83.58%	0.9063	0.8646	0.8956	0.8798
Diseases Of The Genitourinary System	Gradient Boost	83.79%	0.9054	0.8643	0.8982	0.8809
Certain Conditions Originating In The Perinatal Period	Logistic Regression	89.33%	0.9323	0.9291	0.9248	0.9269
Symptoms, Signs, And Ill-Defined Conditions	Gradient Boosting	83.94%	0.9095	0.8655	0.8931	0.8791
Factors Influencing Health Status	Gradient Boosting	84.12%	0.9093	0.8709	0.8959	0.8832
Diseases Of The Nervous System And Sense Organs	Gradient Boosting	84.84%	0.9118	0.8813	0.9029	0.8920
Infectious And Parasitic Diseases	Gradient Boosting	83.73%	0.9086	0.8615	0.8909	0.8759

iii) <u>Confusion matrix:</u> Below are the confusion matrices for each of our diagnoses models showing the performance measurement for machine learning classification models on the testing data for which the true values are known.

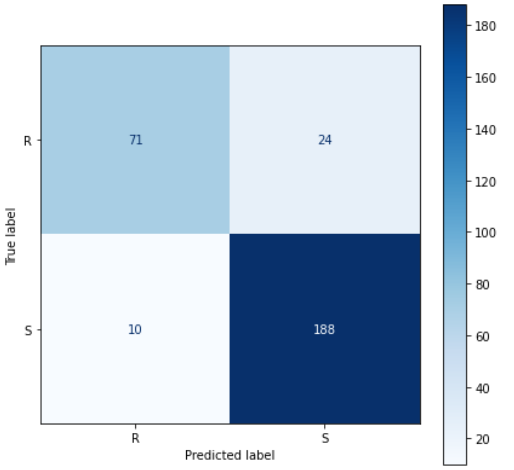

Figure 7a. Confusion matrix for Mental Disorders using Random Forest Classifier:

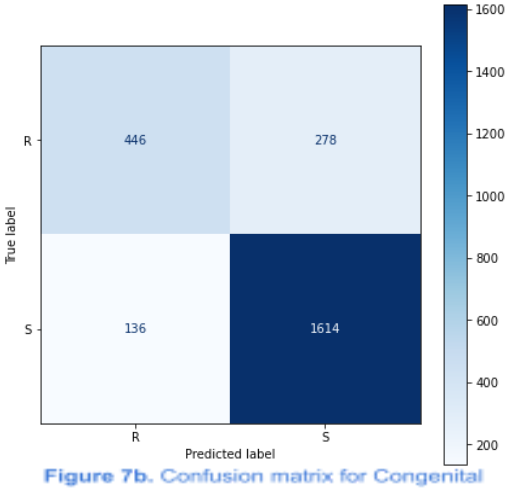

Figure 7b. Confusion matrix for Congenital Anomalies using Random Forest Classifier

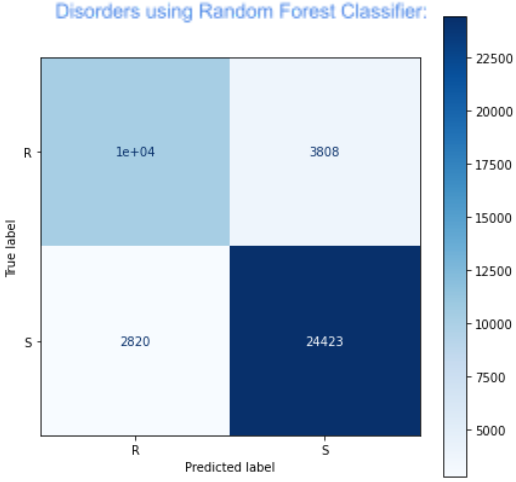

Figure 7c. Confusion Matrix for Injury and Poisoning using Gradient Boosting Classifier

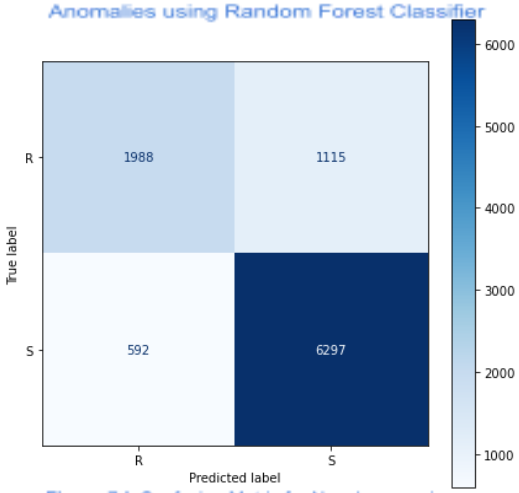

Figure 7d. Confusion Matrix for Neoplasms using Random Forest Classifier:

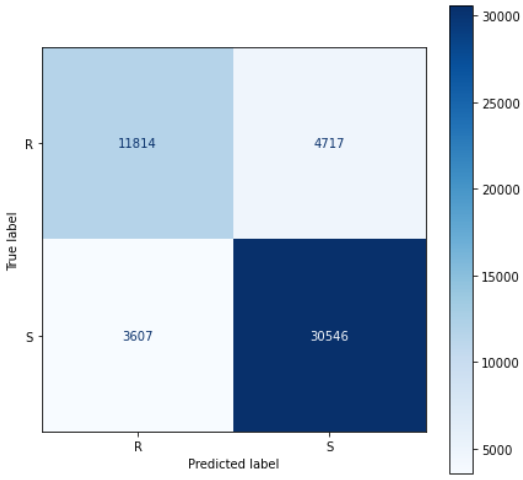

Figure 7e. Confusion Matrix for Diseases of the Circulatory System using Gradient Boosting Classifier

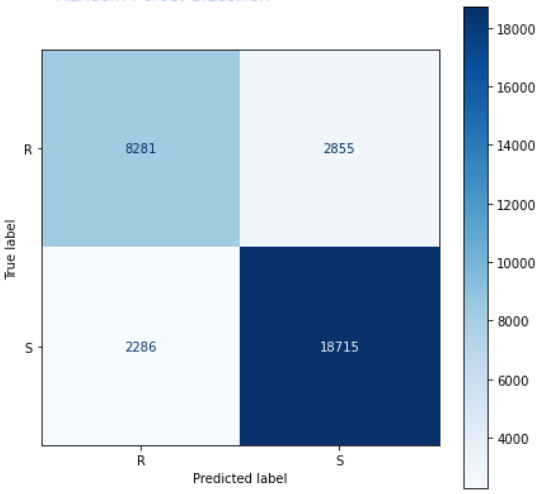

Figure 7f. Confusion Matrix for Diseases of the Digestive System using Gradient Boosting Classifier

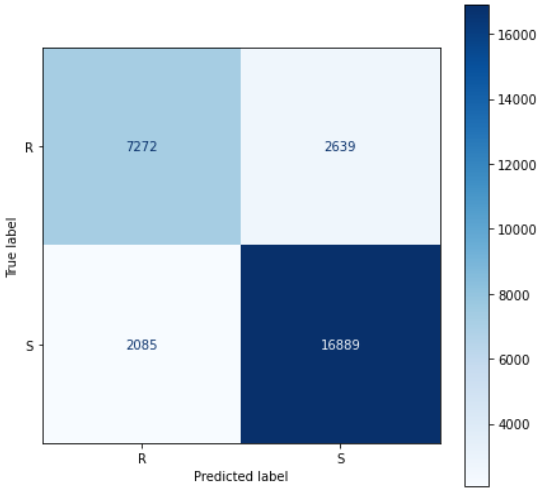

Figure 7g .Confusion Matrix for Diseases of The Blood and Blood Forming Organs using Gradient Boosting Classifier

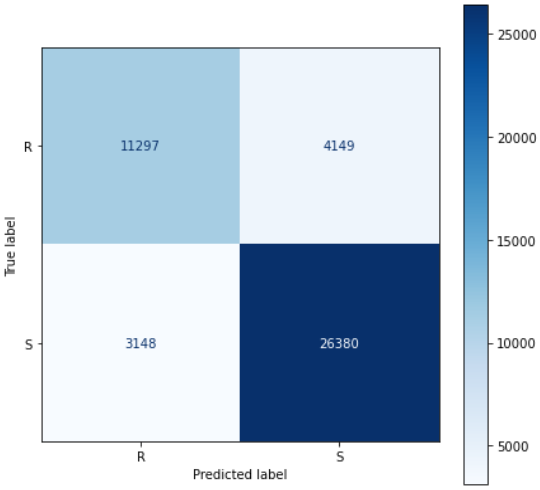

Figure 7h .Confusion Matrix for Diseases Of The Respiratory System using Gradient Boosting Classifier

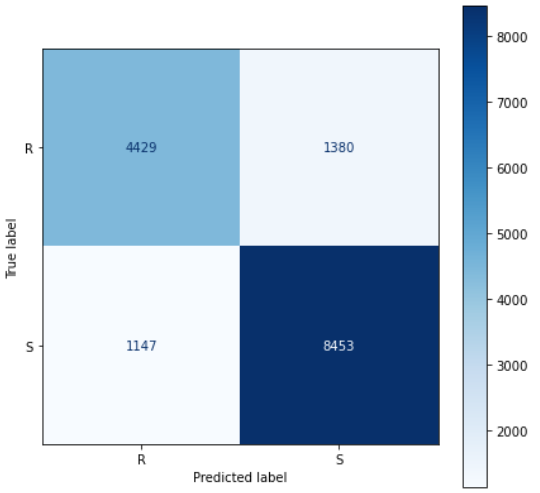

Figure 7i .Confusion Matrix for Diseases Of The Skin And Subcutaneous Tissue using Gradient Boosting Classifier

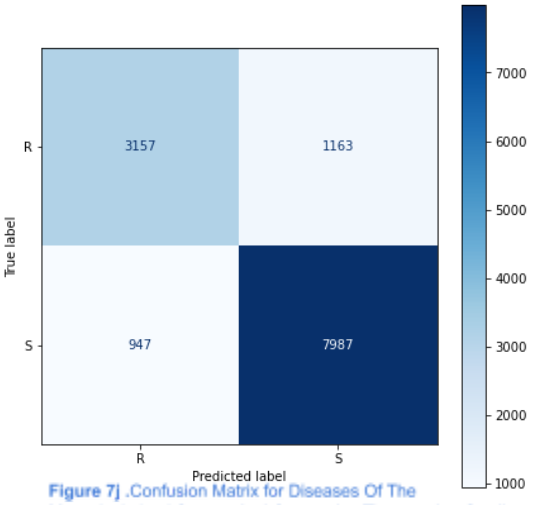

Figure 7j .Confusion Matrix for Diseases Of The Musculoskeletal System And Connective Tissue using Gradient Boosting Classifier

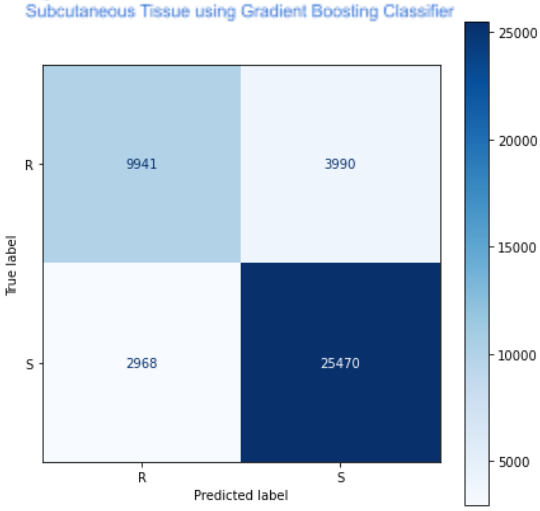

Figure 7k .Confusion Matrix for Endocrine, Nutritional And Metabolic Diseases, And Immunity Disorders using Gradient Boosting Classifier:

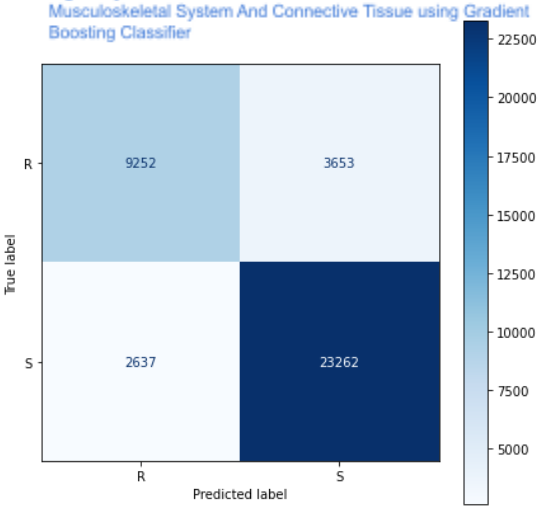

Figure 7l .Confusion Matrix for Diseases Of The Genitourinary System using Gradient Boosting Classifier

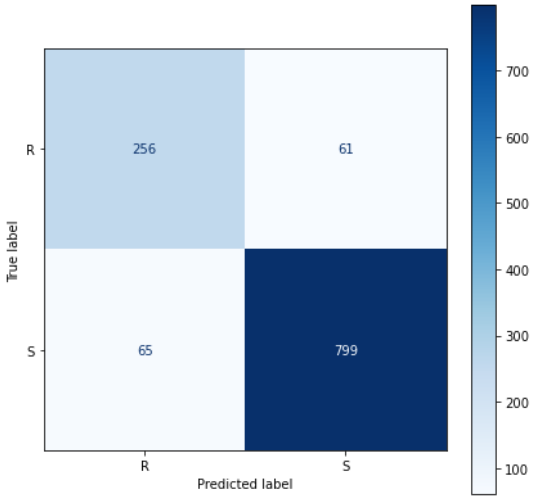

Figure 7m .Confusion Matrix for Certain Conditions Originating In The Perinatal Period using Logistic Regression

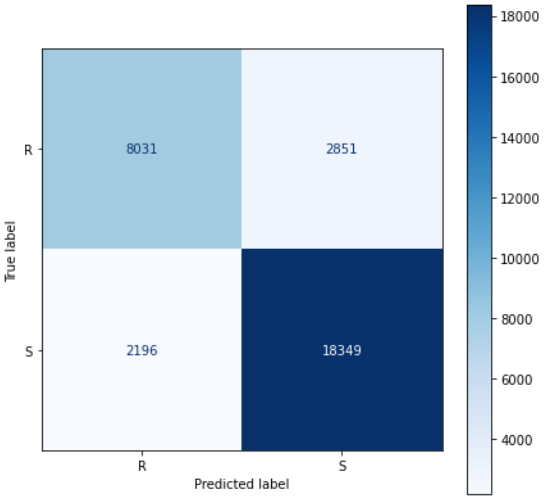

Figure 7n .Confusion Matrix for Symptoms, Signs, And Ill-Defined Conditions using Gradient Boosting Classifier

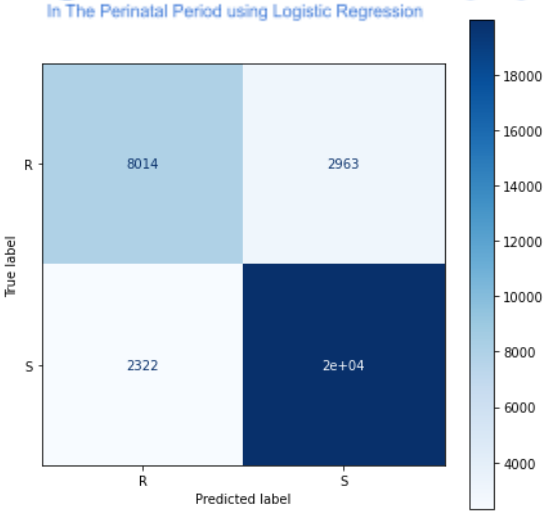

Figure 7o .Confusion Matrix for Factors Influencing Health Status using Gradient Boosting Classifier

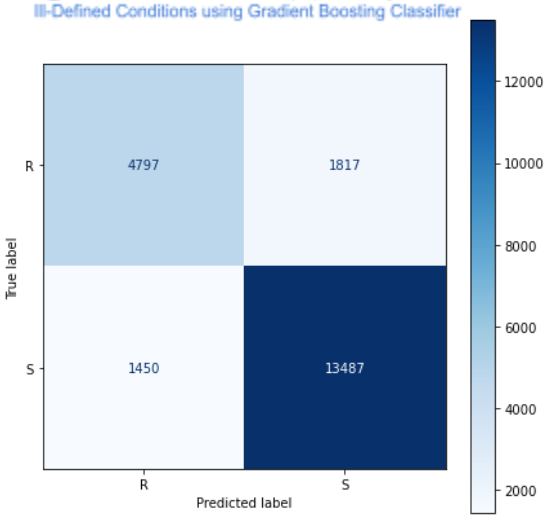

Figure 7p .Confusion Matrix for Diseases Of The Nervous System And Sense Organs using Gradient Boosting Classifiers

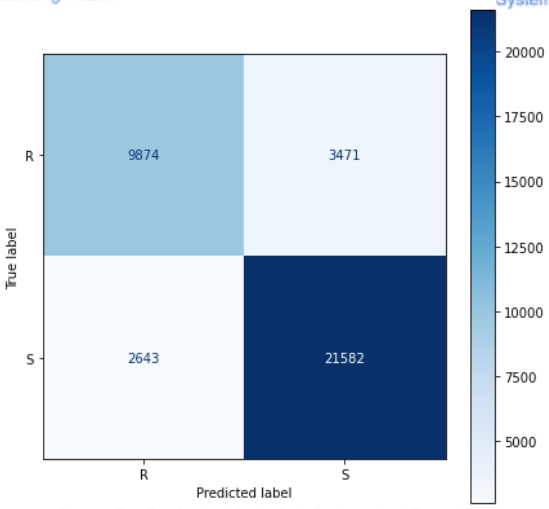

Figure 7q .Confusion Matrix forInfectious And Parasitic Diseases using Gradient Boosting Classifier

iv) <u>ROC curve:</u> Below are the ROC Curves for machine learning models. The receiver operating characteristic curve was plotted after training the model on 70% of the training dataset and 30% of the testing dataset. The ROC curves show the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Our classifiers give curves closer to the top-left corner and indicate top performance. Figures below show the ROC Curve for Mental Disorders, Injury and Poisoning, Diseases of The Blood and Blood Forming Organs, and Certain Conditions Originating In The Perinatal Period.

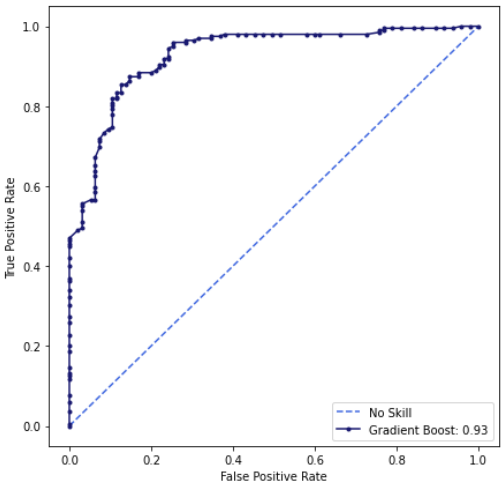

Figure 8a .ROC Curve for Mental Disorders using Random Forest Classifier

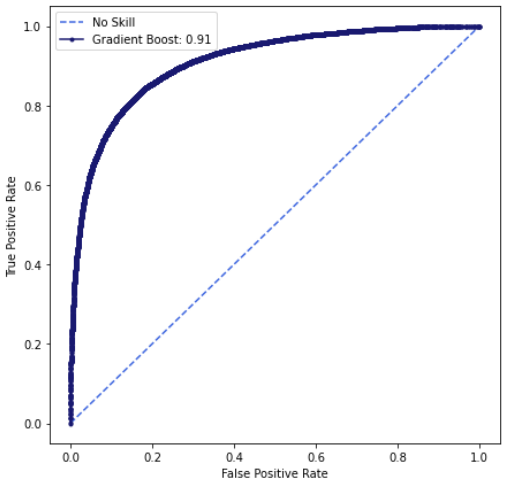

Figure 8b. ROC Curve for Injury and Poisoning using Gradient Boosting Classifier

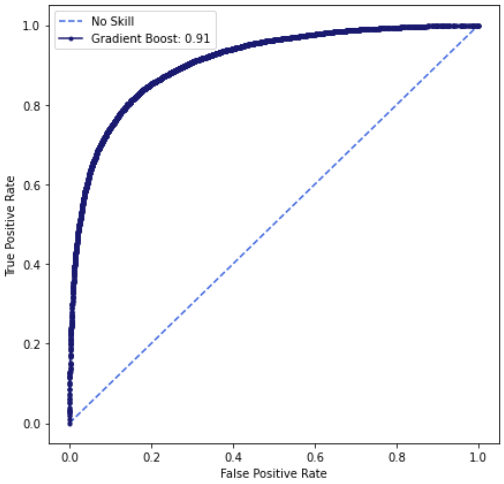

Figure 8c. ROC Curve for Diseases of The Blood and Blood Forming Organs using Gradient Boost Classifier

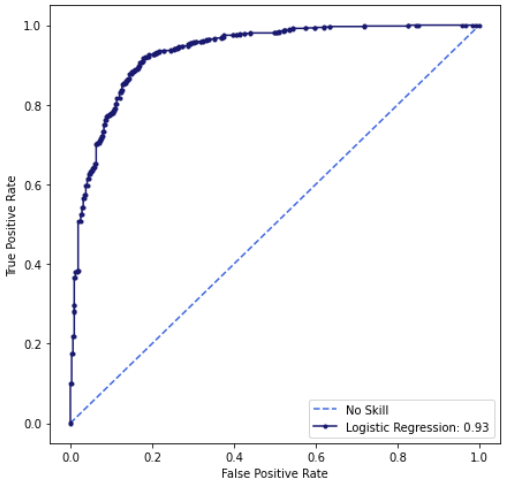

Figure 8d. ROC Curve for Certain Conditions Originating In The Perinatal Period using Logistic Regression

v) Comparing actual labels with predicted labels: In the figure 9 below, we have plotted a line chart that shows the points where we fail to match the antibiotic resistance 3 out of 27 times. The red line shows predicted testing data that matches the actual testing data shown by the blue line with an 89% accuracy. The y-axis has 0 and 1, 0 being resistive and 1 being sensitive, while the x-axis has antibiotic names.

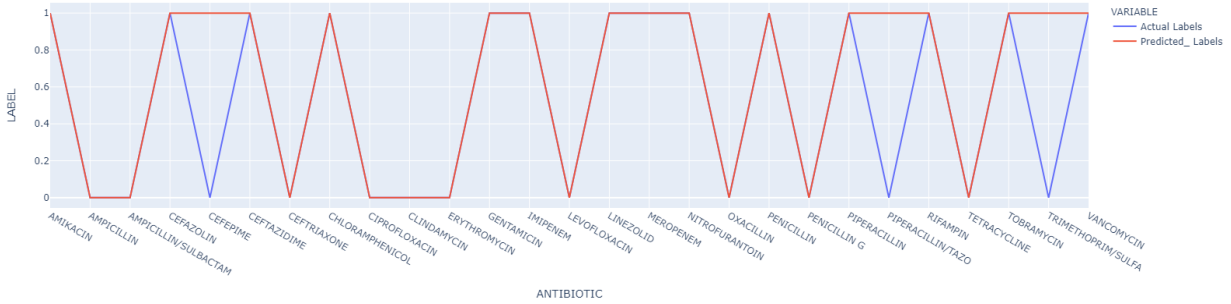

Figure 9. Model classification plot showing actual vs predicted labels

3.2 Length of Hospital Stay and Mortality Rate Prediction

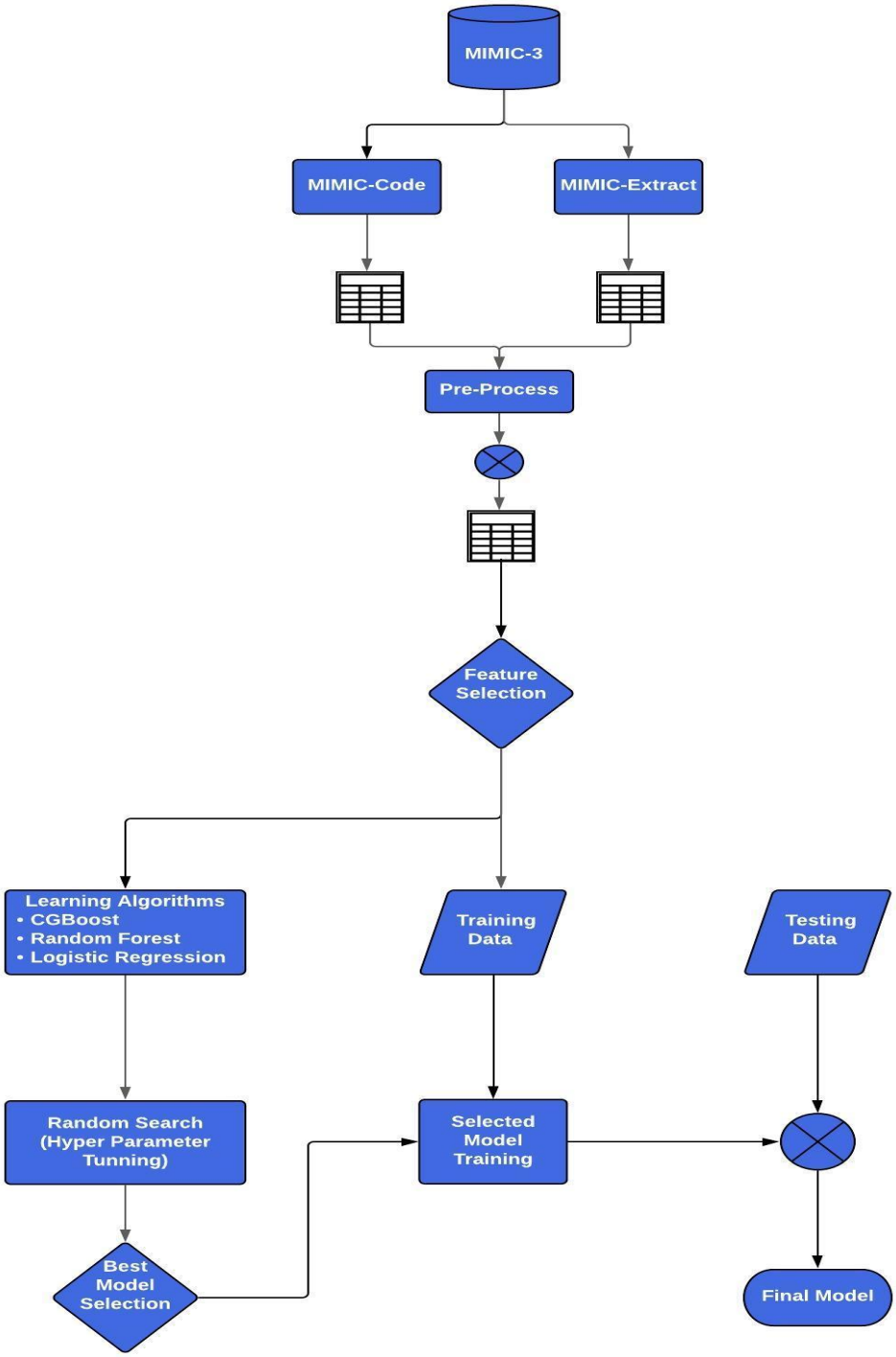

Figure 10. Data Science Pipeline for Mortality Rate Prediction

3.2.1 Data Preprocessing

i) Data Collection: While MIMIC-III's availability as an open dataset has attracted many research studies with machine learning approaches, it is still challenging and risky to work on MIMIC-III data with minimal healthcare domain knowledge. The complexity of the patient's Electronic Health Record (EHR) data stored in the CHARTEVENTS table with more than 300 million rows makes it more challenging to handle the dataset. Since the prediction on ICU patient's mortality rate and length of stay has been common interests among many previous pieces of research, we tried to refer to them to reduce chances of error and redundant efforts on existing work. We were able to find a recent study that introduces an open-source pipeline called 'MIMIC-Extract' to extract static demographic information and vital lab data from MIMIC-II I v1.4. We partly adopted the pipeline as the foundation for extracting 24 hours of ICU patients' demographic information and time-varying vital features. Besides, since MIMIC-III is officially available on cloud platforms such as Google Cloud Platform (GCP), we utilized BigQuery to retrieve data more efficiently than using a local PostgreSQL DB. We adapted SQL scripts shared by the MIT Lab for Computational Physiology's research community and collected severity of illness scores for the patients.

ii) Data Cleaning: We observed some of the data features contain outliers that may affect the prediction of this task. Therefore, we used the 1.5 IQR rule to filter outliers for 'weight_first,' 'height_first,' 'weight_diff' and 'BMI' features. The number of values available in 'height_first' and 'BMI' features is a lot less than other features and the primary cohort. After we considered the importance of 'height_first' and 'BMI' and the amount of data needed to be dropped, we removed 'height_first' and 'BMI' because half of the data rows need to be removed to use those two features. After that, we merge data frames to store height and weight data and severity scores from MIMIC-III and vital signs data from MIMIC-Extract by 'icustay_id.' After obtaining an aggregated data frame, we created four label classes to simplify the predictions in this part. They are 'Mortality in Hospital,' 'Mortality in ICU,' 'Length of Stay Longer than 3 days,' and 'Length of Stay Longer than 7 days'.

3.2.2 Feature Selection

After we aggregate all data tables into one data frame, we have 7514 columns that may be helpful to make predictions. We used the Chi-Squared test to select eight best categorical features and ANOVA to select 40 best numerical features. After testing many best features, we found that the best combination uses the top eight categorical features, and the 40 best numerical features return the most reliable result.

3.2.3 Feature Engineering

i) Feature Transformation: We used one hot encoding to convert categorical features for model building. We had 3 categorical columns in the original dataset, after encoding we ended up with 10 binary columns.

3.2.4 Model Building

After feature selection of each task for mortality and length of stay prediction, we split the data frame that contains all selected features to 70% training, 10% validation, and 20% testing. We plotted the distribution of the label classes and observed around 1(False label): 10(True label) imbalance labels in 'Mortality in Hospital', 'Mortality in ICU' and "Length of Stay Longer than seven days'. We used Synthetic Minority Oversampling Technique(SMOTE) to oversample the minority class in the training dataset. Even though the class labels for 'Length of Stay Longer than three days' are balanced, we attempted to apply SMOTE for this task and improved the validation set for all tasks. F-1 score improved by approximately 0.1 - 0.3, and the AUC-ROC score improved by approximately 0.01 - 0.03.

In the machine learning model training part, we used XGBoost, Random Forest, and Logistic Regression to make predictions for different tasks. We defined a reasonable range of hyperparameter values and used a random search to pick 30 random hyperparameters combinations for model training. The validation dataset helped us find the best combination by the highest AUC-ROC score. After that, we combined the training and validation dataset and utilized 80% training data for training the current model with the best hyperparameter combination. Next, we stored the best model for each task in one pickle file and will employ the best model to make a probability prediction after transferring the prediction with patient ICU stay demographic data to the cloud database.

3.2.5 Model Evaluation

i) Model Evaluation Metrics: The following table contains results of the evaluation metrics for all four tasks and their corresponding best result. The test dataset class label is highly imbalanced. Therefore, we mainly focused on the AUC-ROC score and F-1 score to evaluate our results and looked into the confusion matrix and ROC Curve below to analyze the result deeply after being fine-tuned and trained on 80% of the data.

Task	Model	Accuracy	AUC-ROC score	Precision	Recall	F-1 score
Mortality in Hospital	Logistic Regression	0.7880	0.8289	0.4356	0.7053	0.4219
Mortality in ICU	Logistic Regression	0.8173	0.8689	0.4466	0.7659	0.3894
Length of Stay > 3 days	XGBoost	0.6951	0.7452	0.6934	0.5345	0.5958
Length of Stay > 7 days	Random Forest	0.8928	0.7848	0.2199	0.3434	0.3018

ii) Confusion matrix: The following confusion matrices for each task. We used them to analyze our result in each label class.

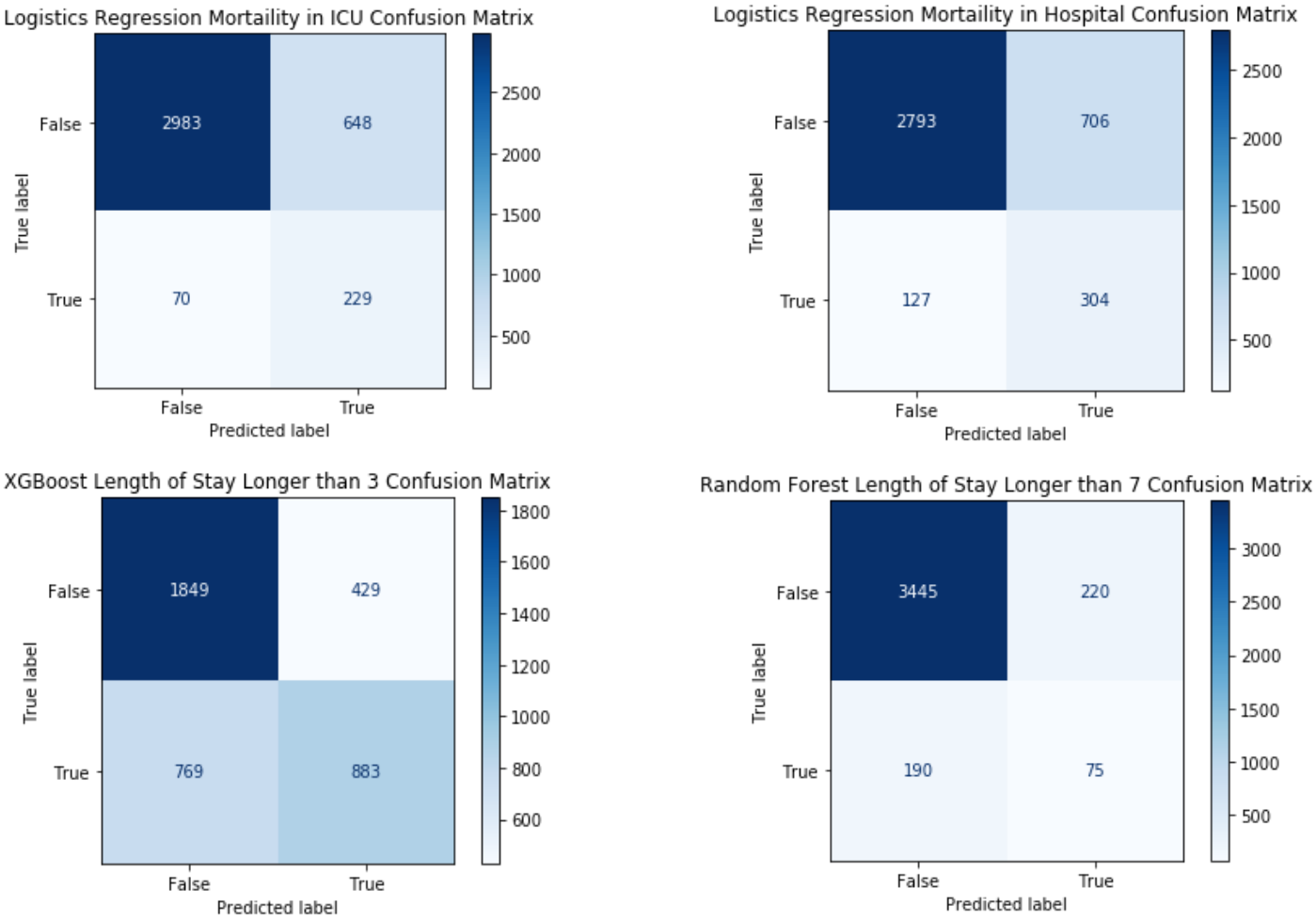

Figure 10. Confusion Matrices for Mortality Rate and Length of Stay Prediction

iii) ROC curve: We used the following ROC curve for each task to analyze our result, especially sensitivity and specificity (1-FPR).

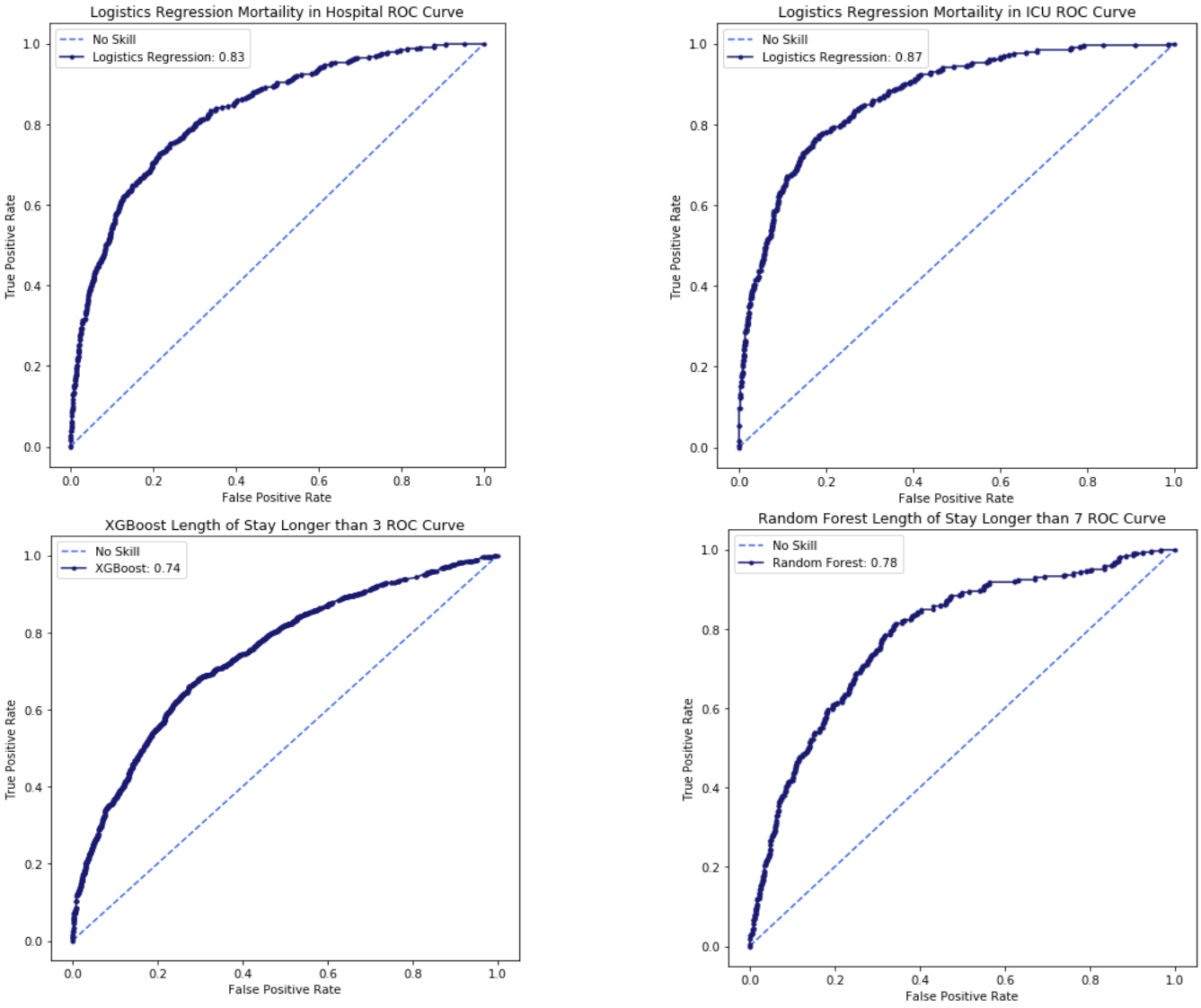

Figure 11. ROC Curves for Mortality Rate and Length of Stay Prediction

iv) Feature Importance: We have plotted several feature importance scores for each model to examine input features based on how valuable each input feature predicts a target variable. We contemplated the distribution of importance scores and the reason for a particular feature being essential. For instance, we observed that the 'glascow coma scale' is vital in most tasks. We researched the importance of the 'glascow coma scale' and detected that it is reasonable to use this feature. Below are the feature importance plots for each task:

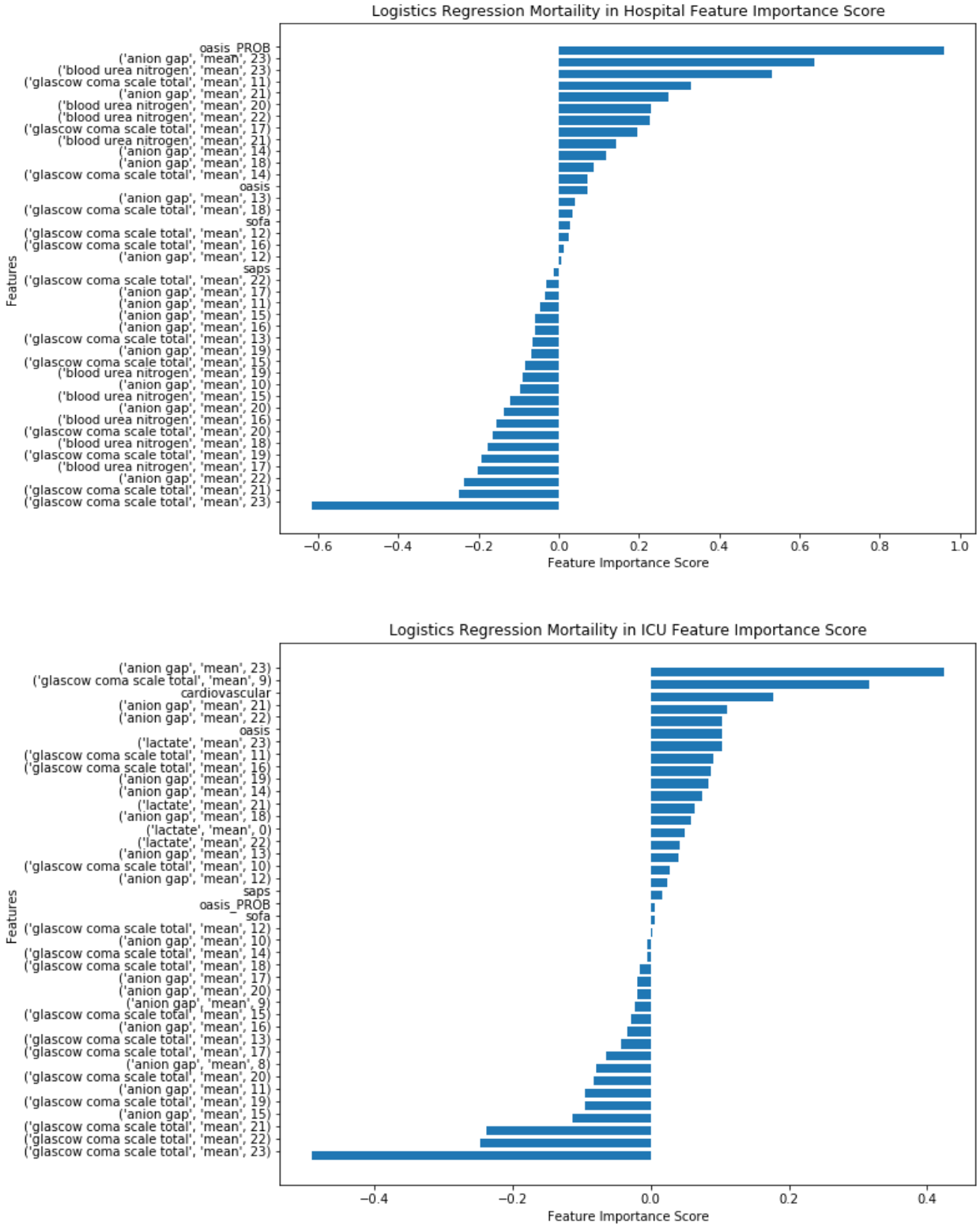

Figure 12a. Feature Importance Score for Mortality Rate in Hospital and ICU

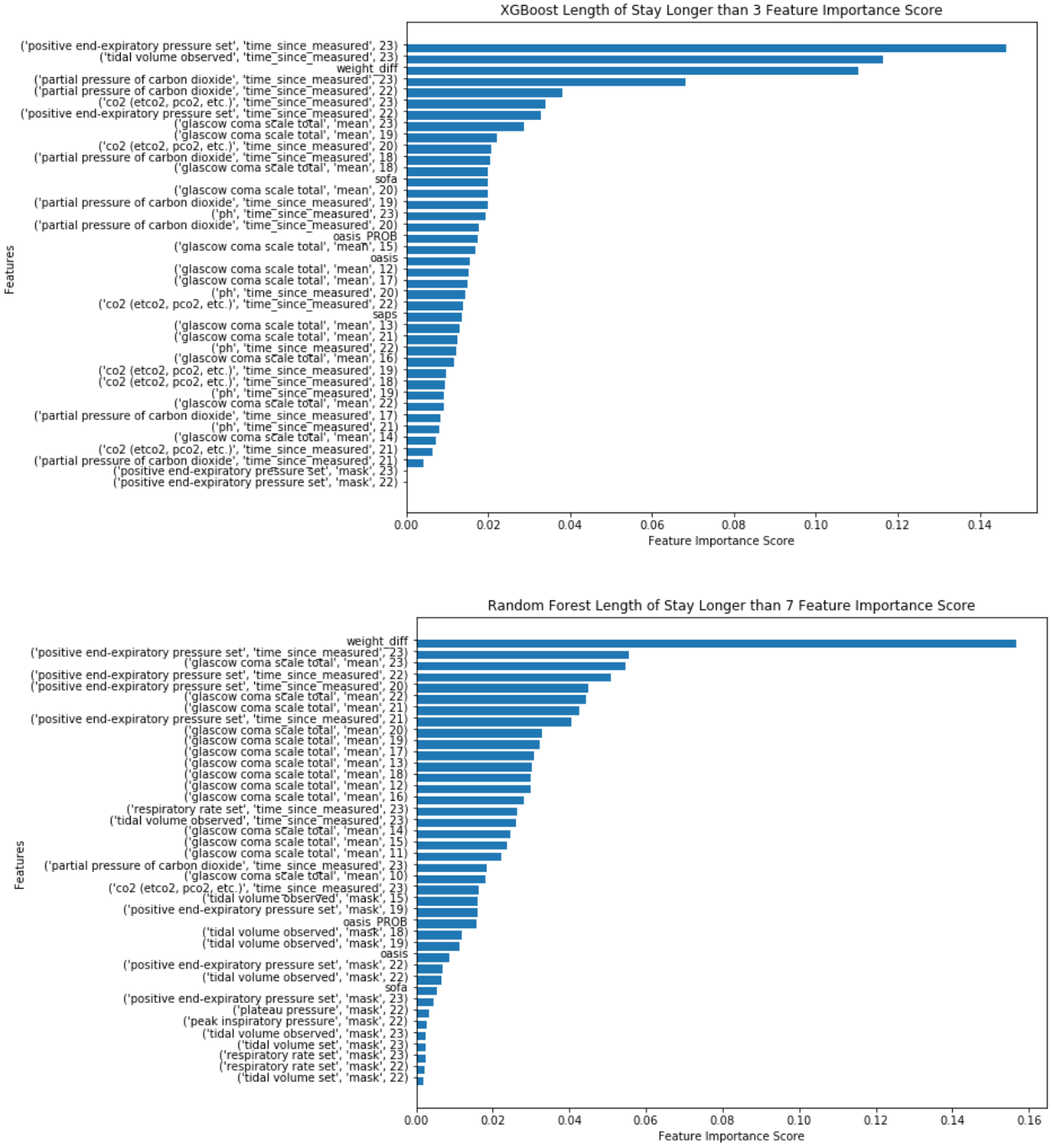

Figure 12b. Feature Importance Score for Length of Stay Prediction for 3 and 7 days.

4 Data Product

4.1 Purpose and Deployment

We designed our data product with two main goals: to perform as a general dashboard with processed and saved data, and the other for delivering real-time prediction. We also deployed both applications on production using Docker and registered the Docker registry images on a Github registry space to facilitate the release of products. At the same time, we utilized AWS EC2 to host the docker containers for enhanced availability. The architecture of our data product is shown below in Figure 13.

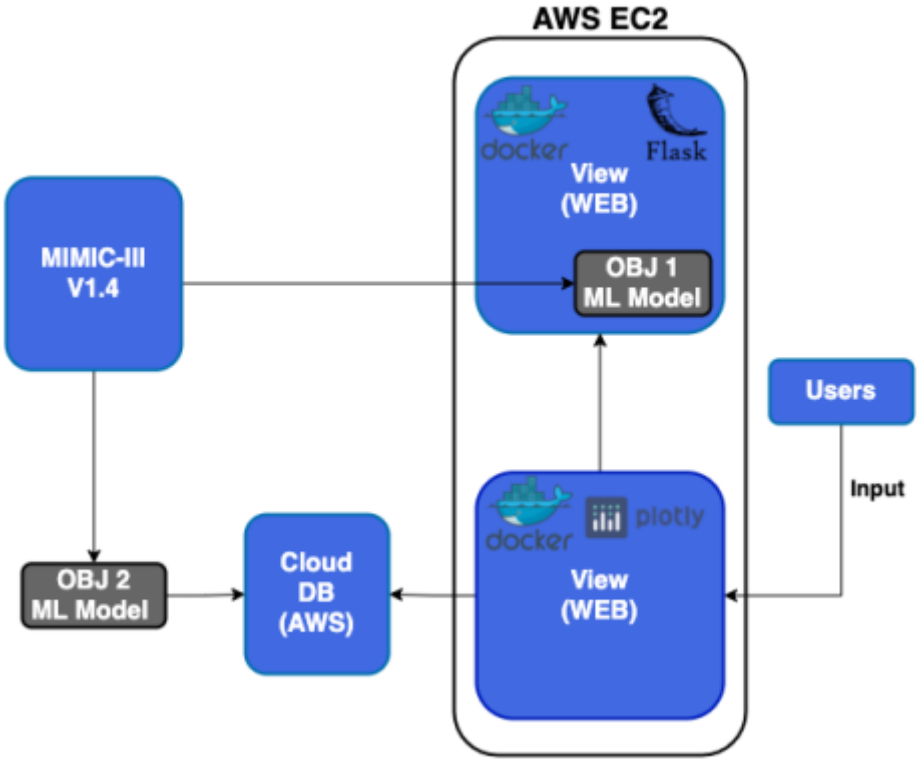

Figure 13. Overview of Data Product Architecture

4.2 Data Analysis Dashboard

We built a dash application offering data visuals and a function of retrieving a batched result for the mortality rate and length of stay prediction. The 'Antibiotic Effectiveness Dashboard' tab displays important findings on antibiotic effectiveness shown in Figure 14. In the 'Mortality Rate and LOS Prediction' tab, a user can query the probability of mortality and length of stay with demographic information for each patient based on 'icustay_id.' For scalability, we utilized a cloud database (AWS RDS) to store the batched dataset.

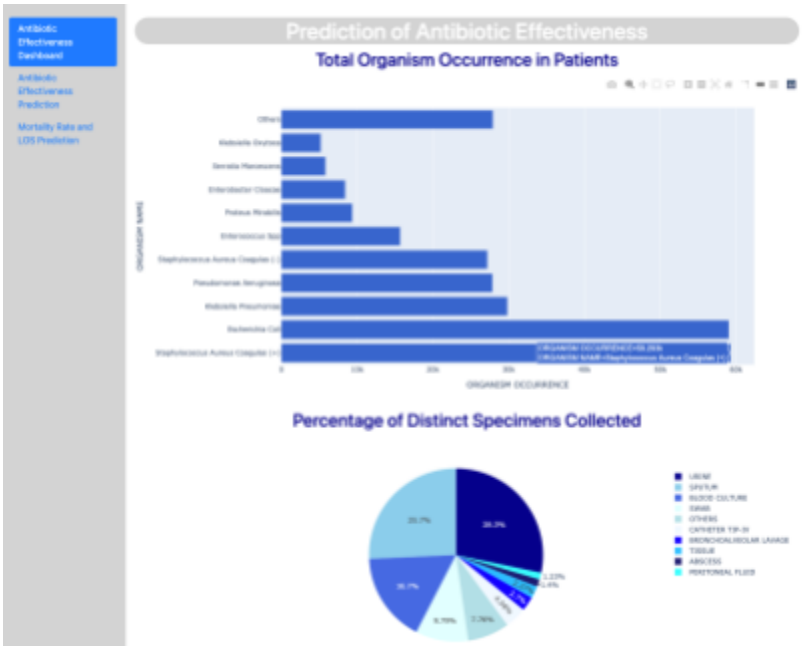

Figure 14. Antibiotic Effectiveness DashBoard Page

4.3 Real-time Machine Learning Antibiotic Effectiveness Inference Application

For Antibiotic Effectiveness prediction, we built a containerized Flask web application offering real-time inference function to potential healthcare personnel. It takes nine features, including diagnosis, antibiotic name, specimen type by a user, and generates a result as 'Sensitive' or 'Resistant.' We stacked 17 different models into pickle files differentiated by diagnosis name during the prediction process. This page can be accessed through the 'Antibiotic Effectiveness Prediction' tab in the dashboard and is shown in Figure 15.

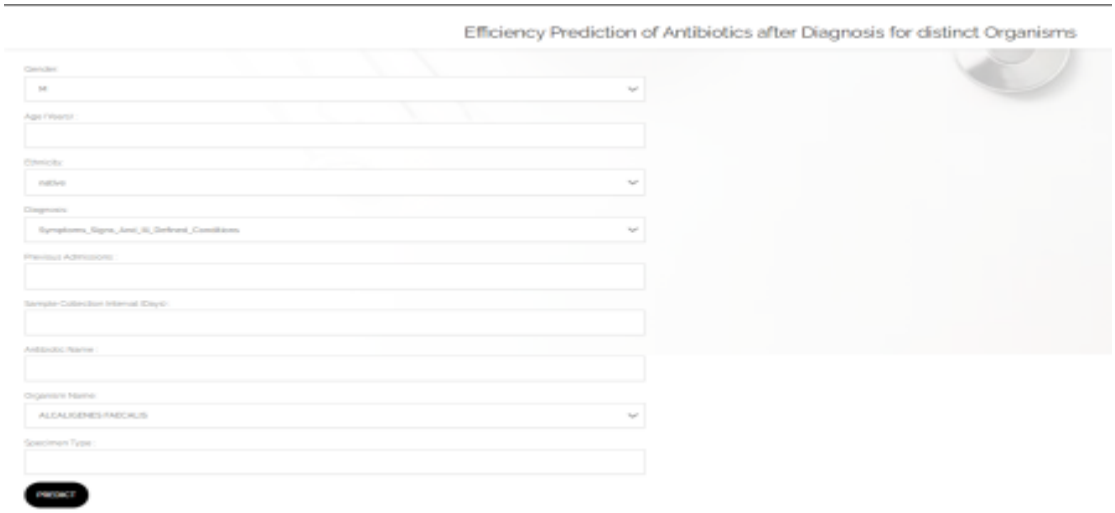

Figure 15. Efficiency Prediction of Antibiotics after Diagnosis for distinct Organisms Page

4.4 Mortality and Length of Stay Prediction Dashboard

We built a Dash web application for mortality and length of stay prediction to visualize the prediction and demographic data. Users can search a particular ICU stay by ICU stay ID. The prediction probability of a specific task is shown in Figure 16 in the gauge chart. When the probability is below 50%, the gauge bar has a blue color. When the probability is above or equal to 50%, the gauge bar will turn red and alert the caregiver.

Figure 16. Mortality Rate and LOS Prediction Page

5. Lessons Learnt

The current predictive model still has limitations in everyday evolving healthcare; new antibiotics are being produced, making variable selection complicated as the antibiotic-resistance behavior depends on many other factors that are not even a part of the MIMIC-III database. Our model is based on data from 2001 to 2011, a decade old and only from one hospital. The dataset does not contain various other scenarios, including specimen contamination or organisms getting more potent due to bacteria carriers from ventilation and other modes. Furthermore, there is no information on if the patient has taken that particular antibiotic before, and maybe that is why the organism is already resistant to it or is if the patient even takes the antibiotics after they are prescribed. In the future, by incorporating data provided by Providence Healthcare, we plan to successfully implement the meta learner and improve the accuracy of future predictions and reduce diagnostic error. We aim to use the predictive power of our model to reduce incorrect antibiotic prescription and consumption.

This project allowed us to experience handling complex, large and real-world datasets in the medical area. We implemented different algorithms to pick useful features and analyze the importance of features to improve machine learning models and evaluate those models, respectively. Moreover, we learned how imbalance class labels affect the evaluation metrics, and using SMOTE may improve the result, especially the F-1 score and ROC-AUC score. However, the main lesson from the project is understanding the importance of having a proper approach towards handling critical data. We also gained particle experience and a deeper understanding of the following tools: Sci-kit Learn, NumPy, Pandas, Matplotlib, Data Prep,SMOTE, SciPy, XGboost, Docker, Big Query, AWS RDS, Dash, Plotly, and Flask.

6. Summary

In this project, we have used the MIMIC-III openly available dataset developed by the MIT Lab for Computational Physiology to predict if an organism is resistant or sensitive towards an antibiotic measured by interpreting a culture test as well as to predict a patient's length of stay and mortality rate using their demographics and first 24-hour vital signs. We have specifically selected several features from the dataset after in-depth advice from medical professionals, medical literature review, and algorithms in some instances to train numerous supervised learning algorithms. We have designed a general dashboard with processed and saved data and a Flask application for delivering real-time prediction and have deployed both applications on production using Docker, which is registered in the Docker registry images on a Github registry space to facilitate the release of products. Simultaneously, we have utilized an AWS EC2 to host the docker containers for enhanced availability. Our web application allows a caregiver to enter patient-specific details to know the antibiotic effectiveness. A medical professional can also enter a patient's ICU Stay ID to determine if the patient has a high mortality rate or length of stay in ICU through our system based on vital signs to help them make critical healthcare decisions.

7. References

Luzzaro, F. et al. Prevalence and epidemiology of microbial pathogens causing bloodstream infections: results of the oasis multicenter study. Diagn. microbiology infectious disease 69, 363–369 (2011).

Shirly Wang, Matthew B. A. McDermott, Geeticka Chauhan, Michael C. Hughes, Tristan Naumann,
and Marzyeh Ghassemi. MIMIC-Extract: A Data Extraction, Preprocessing, and Representation Pipeline for MIMIC-III. arXiv:1907.08322.

Johnson, Alistair EW, David J. Stone, Leo A. Celi, and Tom J. Pollard. "The MIMIC Code Repository: enabling reproducibility in critical care research." Journal of the American Medical Informatics Association (2017): ocx084.

Ghosh, D., Sharma, S., Hasan, E., Ashraf, S., Singh, V., Tewari, D., . . . Sengupta, D. (2019, January 01). Machine learning based prediction of antibiotic sensitivity in patients with critical illness. Retrieved April 19, 2021, from https://www.medrxiv.org/content/10.1101/19007153v1

James Wiggins and Alistair Johnson | on 20 AUG 2019 |
Perform biomedical informatics without a database using MIMIC-III data and Amazon Athena
https://aws.amazon.com/blogs/big-data/perform-biomedical-informatics-without-a-database-using-mimic-iii-data-and-amazon-athena/

Hou, N., Li, M., He, L. *et al.* Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J Transl Med* 18, 462 (2020). https://doi.org/10.1186/s12967-020-02620-5

Supervised learning¶. (n.d.). Retrieved April 19, 2021, from https://scikit-learn.org/stable/supervised_learning.html

Engineer, K., & Scientist, D. (2019, September 19). Machine learning mastery. Retrieved April 19, 2021, from https://machinelearningmastery.com/
Chen Q, Zhang L, Ge S, He W, Zeng M. Prognosis predictive value of the Oxford Acute Severity of Illness Score for sepsis: a retrospective cohort study. PeerJ. 2019 Jun 10;7:e7083. doi: 10.7717/peerj.7083. PMID: 31218129; PMCID: PMC6563807.

Le Gall JR, Lemeshow S, Saulnier F. A new Simplified Acute Physiology Score (SAPS II) based on a European/North American multicenter study. JAMA. 1993 Dec 22-29;270(24):2957-63. doi: 10.1001/jama.270.24.2957. Erratum in: JAMA 1994 May 4;271(17):1321. PMID: 8254858.

Jones AE, Trzeciak S, Kline JA. The Sequential Organ Failure Assessment score for predicting outcome in patients with severe sepsis and evidence of hypoperfusion at the time of emergency department presentation. *Crit Care Med*. 2009;37(5):1649-1654. doi:10.1097/CCM.0b013e31819def97

Jason Brownlee (November 27, 2019). How to Choose a Feature Selection Method For Machine Learning from
https://machinelearningmastery.com/feature-selection-with-real-and-categorical-data/

Dash Plotly¶. (n.d.). Retrieved April 19, 2021, from https://plotly.com/dash/