EAST WEST UNIVERSITY

Project Report

Submitted by

Student's Name	Student ID
Fahmida Afrose Dipti	2020-1-60-025
Fahim Arefin	2020-1-60-052
Sadia Rahman Ani	2020-2-60-172

Department of Computer Science and Engineering

Course Title: Artificial Intelligence
Course Code: CSE366
Section: 02

Submitted To

Redwan Ahmed Rizvee
Lecturer, Department of Computer Science and Engineering
East West University

Submission Date

December 27, 2022

❖ Problem Statement

This data set is all about covid-19. In this data set showcases various features of a patient such as USMER MEDICAL_UNIT, SEX, PATIENT_TYPE, DATE_DIED, INTUBED, PNEUMONIA, AGE, PREGNANT, DIABETES, COPD, ASTHMA, INMSUPR, HIPERTENSION, OTHER_DISEASE, CARDIOVASCULAR, OBESITY, RENAL_CHRONIC, TOBACCO etc. And finally it showcases the predict patients who are at high risk of death from covid.

❖ Content

This dataset contains an enormous number of anonymized patient-related information including pre-conditions. The raw dataset consists of 21 unique features and 1,048,576 unique patients. In the Boolean features, 1 means "yes" and 2 means "no". values as 97 and 99 are missing data.

- sex: 1 for female and 2 for male.
- age: of the patient.
- classification: covid test findings. Values 1-3 mean that the patient was diagnosed with covid in different degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.
- patient type: type of care the patient received in the unit. 1 for returned home and 2 for hospitalization.
- pneumonia: whether the patient already have air sacs inflammation or not.
- pregnancy: whether the patient is pregnant or not.
- diabetes: whether the patient has diabetes or not.
- copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.
- asthma: whether the patient has asthma or not.
- inmsupr: whether the patient is immunosuppressed or not.
- hypertension: whether the patient has hypertension or not.
- cardiovascular: whether the patient has heart or blood vessels related disease.
- renal chronic: whether the patient has chronic renal disease or not.
- other disease: whether the patient has other disease or not.
- obesity: whether the patient is obese or not.

- tobacco: whether the patient is a tobacco user.
- usmr: Indicates whether the patient treated medical units of the first, second or third level.
- medical unit: type of institution of the National Health System that provided the care.
- intubed: whether the patient was connected to the ventilator.
- icu: Indicates whether the patient had been admitted to an Intensive Care Unit.
- date died: If the patient died indicate the date of death, and 9999-99-99 otherwise.

❖ Training dataset:

Columns=18

Rows=820121

Among the columns the first one is the USMER and the last one is the DEATH column. In between

there are all the relevant features

MEDICAL_UNIT,SEX,PATIENT_TYPE,PNEUMONIA,AGE,PREGNANT,DIABETES,COPD,ASTHMA,INMSUPR,HIPERTENSION,CARDIOVASCULAR,OBESITY,RENAL_CHRONIC,CLASIFFICATION_FIANL

❖ Test dataset:

Columns=18

Rows=205031

The dataset is similar

❖ Solution Approach:

Data Preprocessing:

- **Get rid of missing values:** Several steps have been taken to handle the null values. Such as:
  - We dropped the INTUBED, PREGNANT,ICU columns since they have too many missing values.
- Replaced the DATE_DIED column to DEATH column.
- Counted the values of the Death column It got 1 appears (74714) times and 2 appears (950438) time.
- Dropped
  the values that doesn't have a positive correlation with DEATH column.
- After that we replaced the values 97 and 98 to 2 in the pregnant column
- Then we determine the X,Y data.
- Finally we have split the training data 80% into Test data and 20 % into Training data.

❖ Algorithms:

- **LOGISTIC REGRESSION:** Firstly, we have chosen Logistic regression. Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

  In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

  Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems.

- **Decision Tree:** Secondly, we have chosen the Decision tree algorithm. Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. We can represent any boolean function on discrete attributes using the decision tree.

- **Gaussian Naive Bayes:** Thirdly we have chosen Gaussian Naive Bayes. In Gaussian Naive Bayes, continuous values associated with each feature are assumed to be distributed according to a Gaussian distribution. A Gaussian distribution is also called Normal distribution. When plotted, it gives a bell-shaped curve which is symmetric about the mean of the feature.

- **Bernoulli Naive Bayes:** Fourthly we have chosen Bernoulli Naive Bayes. In the multivariate Bernoulli event model, features are independent Booleans (binary variables) describing inputs. Like the multinomial model, this model is popular for document classification tasks, where binary term occurrence (i.e. a word occurs in a document or not) features are used rather than term frequencies (i.e. frequency of a word in the document).

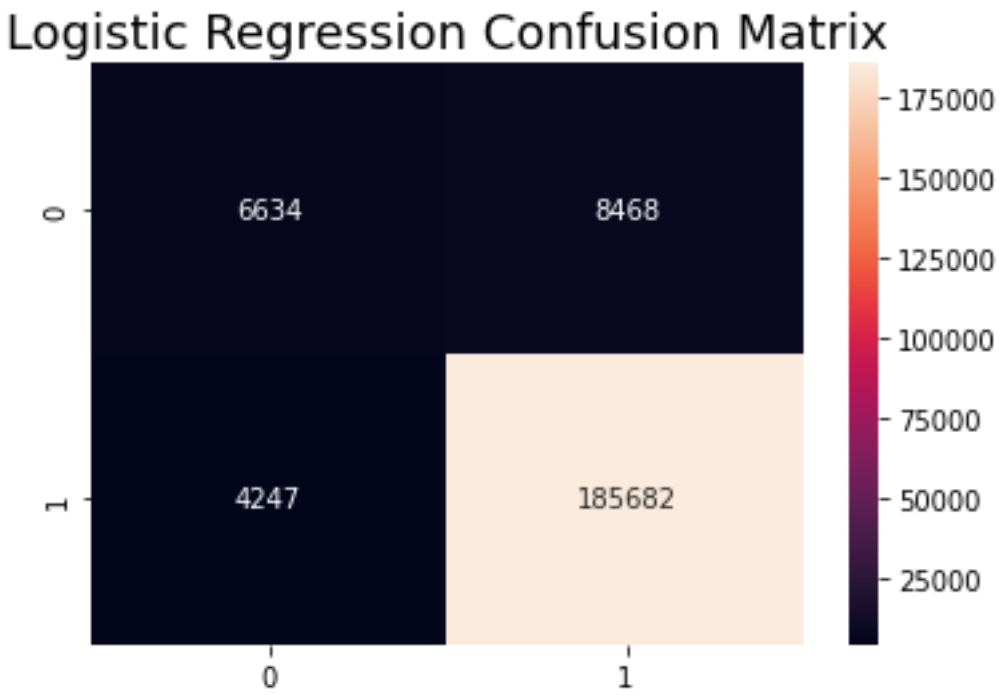
➢ Performance Analysis:

Logistic Regression:

Logistic Regression Accuracy: 0.937984987636016

Logistic Regression F1 Score: [0.51064157 0.96689483]

Logistic Regression Confusion Matrix

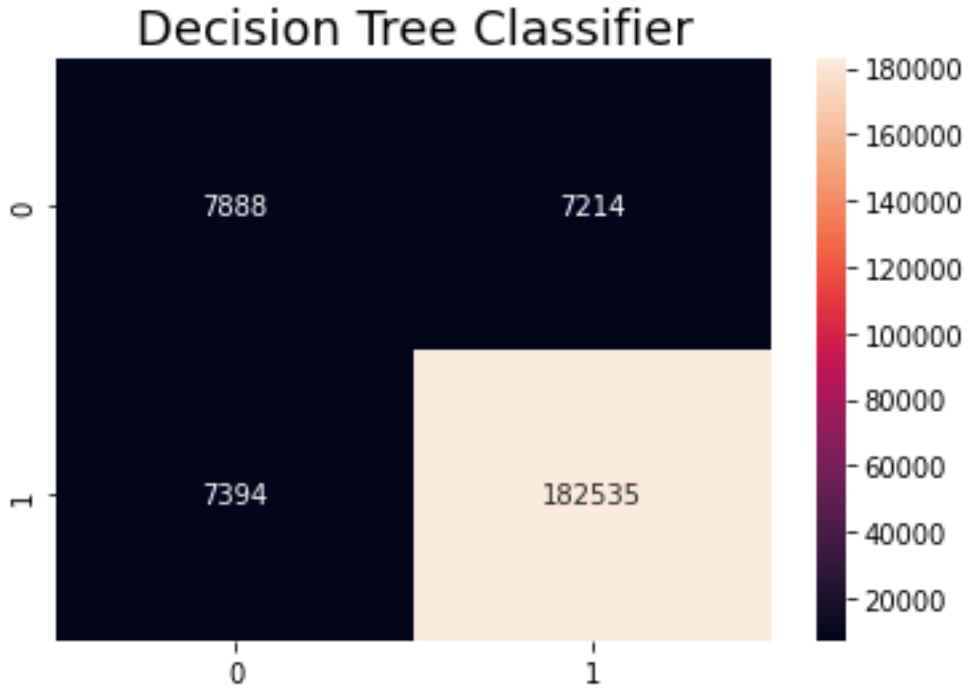


Decision Tree:

Decision tree Accuracy: 0.9287522374665295

Decision Tree F1 Score: [0.51922064 0.96152529]

Decision Tree Heat Map:

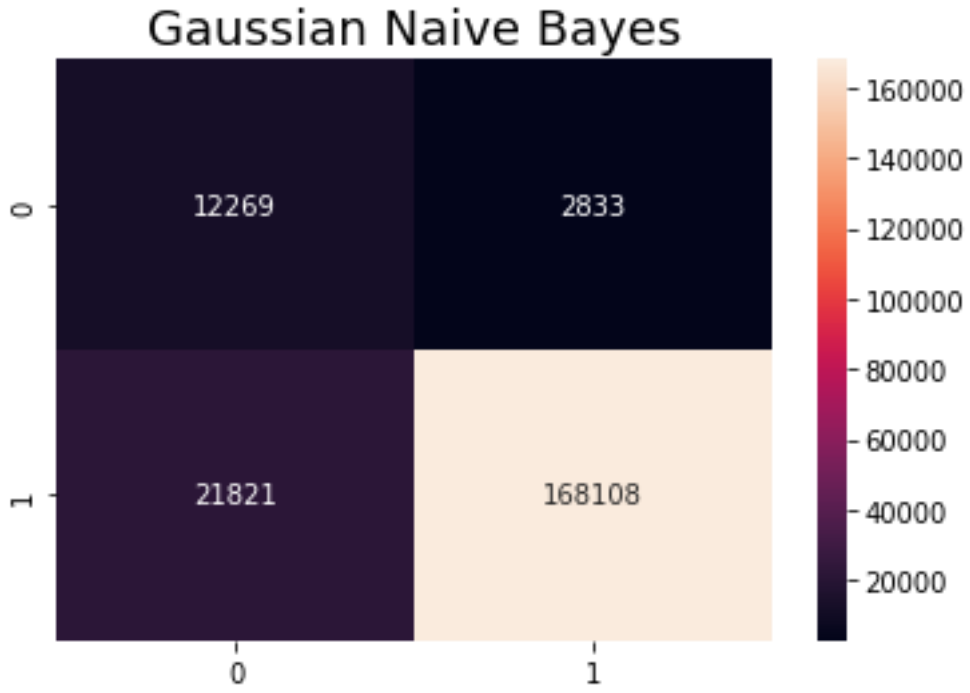


Gaussian Naive Bayes:

Gaussian Naive Bayes Accuracy: 0.8797547687910608

Gaussian Naive Bayes F1 Score: [0.49882095 0.93168177]

Gaussian Naïve Bayes Heat Map:

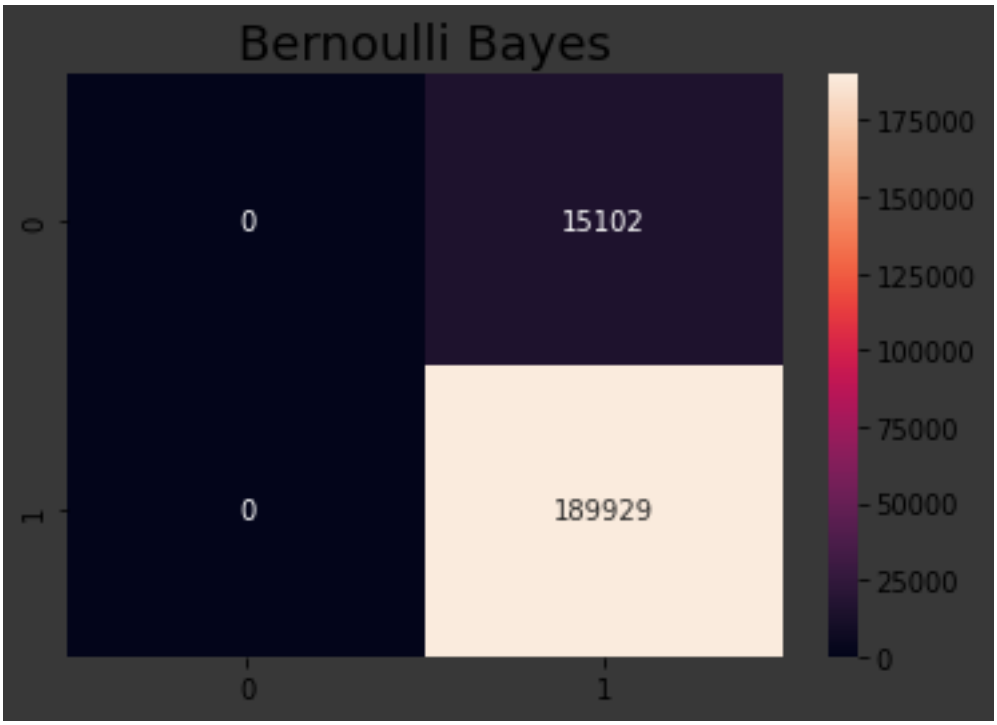


Bernoulli Naive Bayes:

Bernoulli Naive Bayes Accuracy: 0.9263428457160137

Bernoulli Naive Bayes F1 Score: [0. 0.96176322]

Bernoulli Naïve Bayes Heat Map:



➢ Conclusion:

These are just the basic forms of classification algorithms. There are many other classification algorithms which generate even better results. We have tried several ways to ensure the best results. Most importantly, we have learned several things along the way. We have learned about data cleaning, data processing and using different machine learning models. We have also learned about Kaggle contests.