1. Introduction

Coronavirus (COVID-19) has continued to be a global threat to public health. As a matter of fact, it needs unreserved effort to monitor the prevalence of the virus. However, applying an effective prediction of the prevalence is thought to be the fundamental requirement to effectively control the spreading rate. The first case of coronavirus (COVID-19) outbreak was reported on December 31, 2019 in Wuhan, the central province of China, and then, it has been conveyed to the global pandemic owning severe new type of threat to human health and life in a continues method. According to the World Health Organization (2020), more than 15 million COVID-19 cases, 600,000 deaths, and 9 million recoveries have been currently reported.

Data Science rapidly evolving using Python (Pandas and scikit-learn) which are popular open-source Python packages providing fast and high performance in manipulating different data structures. Python is considered as one of the most important and most popular software spreading in data science nowadays to predict different patterns of data sets and generate the optimum model. Predictive analytics is the process of analyzing historical data to estimate the future results. They have quickly emerged as a popular choice from different tools for researchers and scientists to solve real-world problems.

1. Objectives

This research aims to find the best techniques suitable for the research main purpose which is predicting whose new Positive COVID-19 will be deceased or discharged regards the patient medical history which will be discussed in depth latter in next sections, After searching in different papers for the best Algorithms and models can serve this research aims finding that Decision Trees Model, Bernoulli Naive Bayes, and Logistic Regression model are the most common models should be used in the research problem according to the dataset used for this study which is presented Next section providing information about the data source. The fourth section is the descriptive analysis for the respondents and research variables presented, the fifth section the pre-processing stage of the data and preparing it to the models requirements. The sixth section presents splitting dataset to Train and Test datasets, The seventh section presents the model selection and the algorithm used for each model. The eighth section displays the data analysis and the models’ evaluation. Finally, a conclusion is driven with the main findings of this chapter.

1. Data Source
2. Descriptive Statistics
3. Pre-processing

Data preparation is a fundamental stage of data analysis. While a lot of low-quality information is available in various data source, many researchers are interested in how to transform the data into cleaned forms which can be used for high-profit purposes serving their research’s. This goal generates an urgent need for data analysis aimed at cleaning the raw data. comprises those techniques concerned with analyzing raw data to yield quality data, mainly including data collecting which happened in previous section, data transformation will be illustrated in eature engineering, data cleaning, and other concepts out of the scoop of the research problem.

* 1. Exploring & Assessing

This an important step in machine learning process, Data exploration and assessing is about efficiently extracting knowledge from data even if we do not know exactly what we are looking for this main information are necessary to know about the data is How many missing values? Are there duplicates in the observations? Checking data types, How many columns and rows in the dataset? Checking the consistency of values in each variable, how many unique vales in each variable? Calculate descriptive statistics for numerical variables, checking the outliers in ever numerical variable and as much as information can be collected about the data can give a lot of ideas that can be applied on the data and how to deal with it.

a lot of these informations are covered in descriptive statistics section.

* 1. Cleaning

Finding that there is 230 unique observations about positive COVID-19 cases in the study dataset with 3 records have some missing values, even though the set isn’t large as much as needed to get better results but in this case this data was in the important variables about medical history of this 3 patients it’s recommended to exclude this 3 records till not affect the models and it’s accuracy

* 1. Features selection

selecting the related features from the dataset and remove the irrelevant ones. Irrelevant features can negatively impact the performance of a machine learning model. Selected features working as inputs to the model is:

[ 'Gender', 'Non-Invasive Positive Pressure Ventilation (NIPPV) ', 'Mechanical Ventilation', 'ECMO', 'HFNC', 'Received tracheostomy’, ‘Prone-positioning ventilation', 'Vasopressors', 'TPE', 'Remdesivir', 'Hydroxychloroquine', 'Antibacterial agent', 'Corticosteroid', 'Interleukin-6 receptor antagonist', 'Renal Replacement Therapy (Dialysis) in ICU', 'Chronic Cardiovascular-Lung Disease', 'Chronic Liver Kidney Disease', 'Diabetes', 'Hypertension', 'Any other Risk Factors or Comorbidities?']

Some of them selected because of their relation to the desired output and some selected be it’s importance from researchers point of view or some logical assumptions and some restrictions of the models that will process the data can’t be implemented on it like logistic regression and Bernoulli Naieve bayes mainly implements on Binary variables like selected feature and th desired output

* 1. Feature Engineering

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Yes** | **1** | **Deceased** | **0** | **Male** | **1** |
| **No** | **0** | **Discharged** | **1** | **Female** | **0** |

is the process to set up the data for better model performance. There is many techniques can be used in this state but the most common transformations is standardization, normalization, One hot encoding for categorical variables or encoding which will be used in the study dataset by converting the whole features and the respondents to (0, 1) according to the following table :

1. Split using Cross-validation

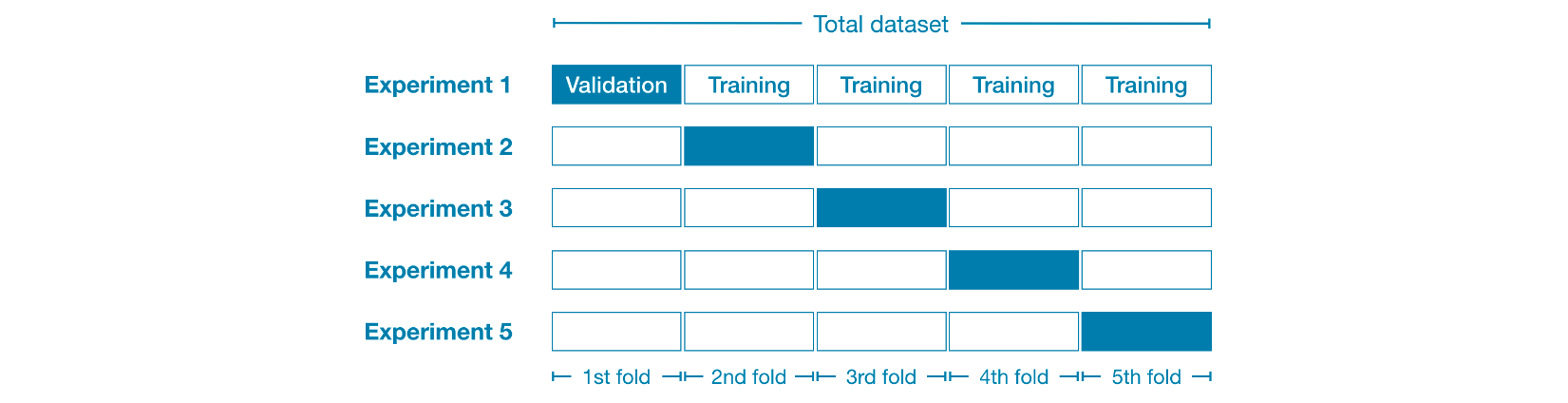
Cross-validation is a statistical method used to estimate the skill of machine learning models.

It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem because it is easy to understand, easy to implement, and results in skill estimates that generally have a lower bias than other methods.

Cross-validation gives a more accurate measure of model quality, which is especially important if you are making a lot of modeling decisions. However, it can take longer to run, because it estimates multiple models (one for each fold).

So, given these tradeoffs, when should using these approach?

* For small datasets, where extra computational burden isn't a big deal, you should run cross-validation.
* For larger datasets, a single validation set is sufficient. the code will run faster, and may have enough data that there's little need to re-use some of it for holdout.

For example, it’s begin by dividing the data into 5 pieces, each 20% of the full dataset. In this case it’s said the data broken into 5 "folds".

Then, runing one experiment for each fold:

In Experiment 1, we use the first fold as a validation (or holdout) set and everything else as training data. This gives us a measure of model quality based on a 20% holdout set.

In Experiment 2, we hold out data from the second fold (and use everything except the second fold for training the model). The holdout set is then used to get a second estimate of model quality.

We repeat this process, using every fold once as the holdout set. Putting this together, 100% of the data is used as holdout at some point, and we end up with a measure of model quality that is based on all of the rows in the dataset (even if we don't use all rows simultaneously).

1. Models and Algorithm Selection

The Models and algorithms are the key pieces that allow the machine to learn from input data and improve from experience, searching for best algorithms that fits to our input and output data, leaded to the best method founded fits to that state there are the Decision Tree, Bernoulli Naive Bayes , and Logistic Regression model

* 1. Decision Tree

A decision tree is a flowchart-like tree structure where an internal node represents feature(or attribute), the branch represents a decision rule, and each leaf node represents the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret. Diagram

Description automatically generated

How does the Decision Tree algorithm work?

The basic idea behind any decision tree algorithm is as follows:

Diagram

Description automatically generatedSelect the best attribute using Attribute Selection Measures(ASM) to split the records. Make that attribute a decision node and breaks the dataset into smaller subsets. Starts tree building by repeating this process recursively for each child until one of the condition will match: All the tuples belong to the same attribute value. There are no more remaining attributes. There are no more instances.

The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. A tree can be seen as a piecewise constant approximation.

It’s obvious in this sample of comparison between the predicted values from test data set and Their desired output almostly they are the same which means this model fits to the problem

After running the model using cross validation technique with 5 CV the result is :

Accuracy scores:

[0.82608696 0.86956522 0.97777778 0.86666667 0.91111111]

average accuracy : 0.8902415458937198

precision scores:

[0.84210526 0.83333333 1. 0.875 0.91304348]

average precision : 0.8926964149504195

recall scores:

[0.66666667 0.82608696 0.91304348 0.91304348 0.91304348]

average recall : 0.846376811594203

* 1. Bernoulli Naive Bayes

implementing the naive Bayes training and classification algorithms for data that is distributed according to multivariate Bernoulli distributions, i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable. Therefore, this class requires samples to be represented as binary-valued feature vectors; if handed any other kind of data, a Bernoulli instance may binarize its input (depending on the binarize parameter). The decision rule for Bernoulli naive Bayes is based on

Graphical user interface, text, application

Description automatically generated

After running the model using cross validation technique with 5 CV the result is :

Accuracy scores:

[0.80434783 0.93478261 0.93333333 0.88888889 0.95555556]

average accuracy : 0.9033816425120772

precision scores:

[0.89473684 0.91666667 0.91666667 0.95 0.92 ]

average precision : 0.9196140350877192

recall scores:

[0.70833333 0.95652174 0.95652174 0.82608696 1. ]

average recall : 0.8894927536231885

which is higher than the decision tree evaluations

* 1. Logistic Regression model

Logistic Regression is one of the most simple and commonly used Machine Learning algorithms for two-class classification. It is easy to implement and can be used as the baseline for any binary classification problem. Its basic fundamental concepts are also constructive in deep learning. Logistic regression describes and estimates the relationship between one dependent binary variable and independent variables.

Logistic regression is a statistical method for predicting binary classes. The outcome or target variable is dichotomous in nature. Dichotomous means there are only two possible classes. For example, it can be used for cancer detection problems, and like he research problem, It computes the probability of an event occurrence.

It is a special case of linear regression where the target variable is categorical in nature. It uses a log of odds as the dependent variable. Logistic Regression predicts the probability of occurrence of a binary event utilizing a logit function.

Linear Regression Equation:

Where, y is dependent variable and x1, x2 ... and Xn are explanatory variables.

Sigmoid Function:

Appling the Sigmoid function on linear regression:

After running the model using cross validation technique with 5 CV the result is :

Accuracy scores:

[0.80434783 0.89130435 0.95555556 0.91111111 0.91111111]

average accuracy : 0.8946859903381643

precision scores:

[0.85714286 0.875 0.95652174 0.95238095 0.91304348]

average precision : 0.9108178053830228

recall scores:

[0.75 0.91304348 0.95652174 0.86956522 0.91304348]

average recall : 0.8804347826086957

as this evaluations good but not better than the previous model

1. Model Evaluation

The most common two types of metrics that helps in evaluating classification models is Accuracy and Precision-recall which used in the previous section

Accuracy is how close a measured value is to the actual (true) value Accuracy = num of right prediction / total num of predictions

Average precision computes the average precision value for recall value over 0 to 1 is a popular metric in measuring the accuracy of object detectors like Faster R-CNN, SSD, etc. It sounds complicated but actually pretty simple the next illustration for it with a mathematical definitions. But before that, here is quick recap on precision, recall.

Precision measures how accurate is the predictions. i.e. the percentage of your predictions are correct.

Recall measures how good you find all the positives. For example, we can find 80% of the possible positive cases in our top K predictions. Here are their mathematical definitions:

Text, letter

Description automatically generated

1. Conclusion

This table showing the scores for every used algorithm

|  |  |  |  |
| --- | --- | --- | --- |
| Measure | Decision Tree | Bernoulli Naive Bayes | Logistic Regression |
| Accuracy | 89% | 90% | 89% |
| Precision | 89% | 92% | 91% |
| Recall | 85% | 89% | 88% |

It seems that they all could fit in this problem but the best model obviously the Bernoulli Naive Bayes has the highest score so it’s the recommended model for this problem with accuracy 90% in average