# 

# INFULNZA DISEASE PREDECTION

# Abstract

Influenza is a disease with a high level of contagiousness that impacts millions of individuals globally annually. The devastating effects of influenza outbreaks are well-documented, with the Spanish Flu being one of the most significant outbreaks in history. To combat the spread of influenza, the use of Machine Learning techniques in medical research has gained popularity, with applications in early disease diagnosis, pathology, and disease classification.

In this study, we extracted a dataset of Human Surveillance records from the Influenza Research database. Initially, we extracted 30531 records covering the period of 2006 to 2020 for data pre-processing. After thorough pre-processing, we selected 18581 records for further analysis.

The next step involved the classification and analysis of the data using four different machine learning methods: Support Vector Machines, K-nearest Neighbors, Artificial Neural Networks, and Random Forest. The performance of these methods was evaluated in terms of sensitivity, specificity, and accuracy.

The experimental results demonstrated that Random Forest was one of the most effective machine learning techniques for the early detection of influenza. This finding has significant implications for the medical field, as it can aid in the development of more accurate and efficient influenza detection systems.

Overall, the study highlights the potential of Machine Learning techniques in medical research and their effectiveness in the early detection of influenza. By analysing large datasets and employing sophisticated algorithms, we can make significant strides in the prevention and control of infectious diseases like influenza.

# Objective

Through this project, we wish to further enhance upon the use of Machine learning techniques and algorithms in the field of medicine, exhibiting the accuracy and ease of access of such amenities and the utility of these techniques in this day and age.

With the human-constrained data procured through inter-connected resources and the availability of past research done by various publications, we have combined and amplified the methodology of our research, providing efficient and much more precise results.

Considering the type of data provided and the nature of the research, utmost care has been taken to not let outliers and unnecessary outputs deviate the desired results, inflicting upon the learning methods itself and creating significant problems for the research.

Taking into consideration that much of the work has been done through Python language and past resources provided in this paper, the project is easy to understand and implementable for further research upon this subject

# Introduction

Influenza is an acute respiratory disease caused by the influenza A, B or C virus. Although type D virus does exist, it mainly affects cattle with frequent spill over to other species. It often emerges as outbreaks and epidemics worldwide, mainly during the winter season. Respiratory secretions of infected individuals contain a substantial amount of influenza virus particles, thus making it possible for the infection to spread through the dispersion of large droplets generated during sneezing and coughing. “The mean duration of influenza virus shedding in immunocompetent adult patients is ~5 days but may continue for up to 10 days or more, particularly in children, elderly adults, patients with chronic illnesses, and immunocompromised hosts”. The onset of influenza is usually sudden and characterized by the appearance of high-grade fever, along with symptoms such as myalgia, headache, and malaise. These rapid manifestations are accompanied by symptoms of respiratory tract illnesses such as non-productive cough, sore throat, and nasal discharge. Following a certain duration of illness, influenza has the potential to affect various organs including the lungs, brain, and heart to a greater extent than the respiratory tract, potentially leading to hospitalization.

The greatest number of deaths inflicted on the human population by the virus is in the year 2016, when 3985 lives were lost, comprising almost equal Male (1987) and Female (1998) deaths.

The virus spreads as yearly outbreaks and in the past has affected at least 2 million people every outbreak and caused >200,000 deaths. Notable records of the virus exist such as “Spanish flu”, “Asian Influenza”, and” Hong-Kong Influenza”.

Machine learning techniques in the field of healthcare have reached the point where they are being used to analyse pictorial data in which circumstance, they use pattern recognition to look for certain indications. Within this project, techniques used for analysis are mainly implemented in the form of classifying algorithms.

Many kinds of datasets were available for such a research subject, some of which showed weekly updates to cases and symptoms. For this project, the dataset used was chosen to fine tune the model to as much accuracy as possible and remain within the boundaries of such researches for synergetic and ethical purposes.

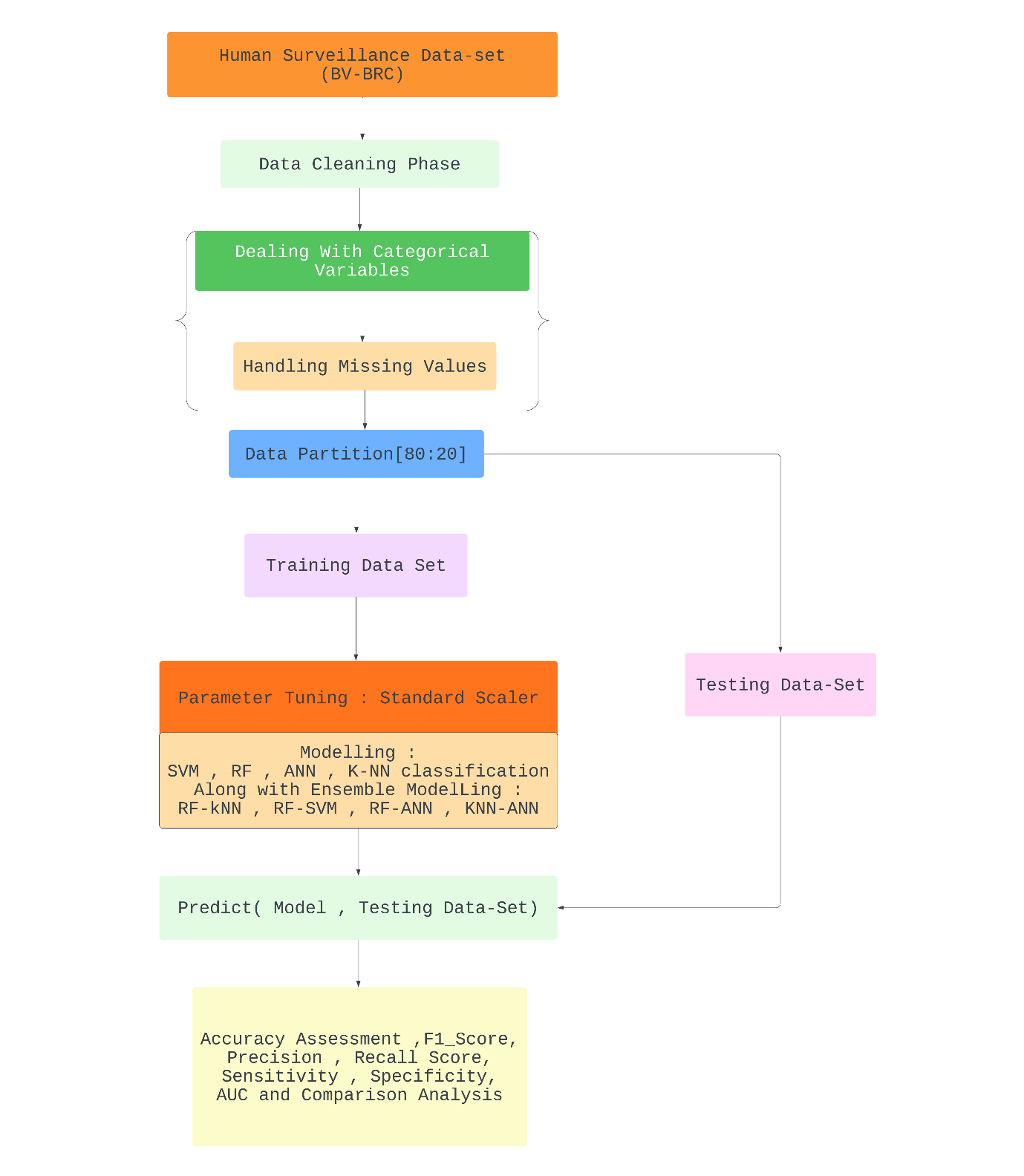
Machine learning researches have reached to a point where such impactful diseases can be detected early on, playing a significant role in efficient classification and medication. This project uses various methods such as Machine Learning techniques along with Deep Learning networks on a dataset through which the model is trained and validated. By providing a sufficient accuracy rate, this project is helping to further integrate Machine Learning models in medication systems and improvising them. The dataset has been procured from a trustable source, which among many other great sources out there, provides a well-defined dataset with all the necessary information required for the model along with a good sample size, helping us to further give precision to our model.

# 

STUDY AREA: DATASET

For the objective of this research, the data chosen comprises the symptoms the patient exhibits along with the gender, age and collection year of the patient. The data pre-processing done is to escalate and expand upon such features. Furthermore, the techniques used in this regard such as encoding will help the model to understand the data better. The data is obtained from the databases of BV-BRC, and contains data of 30,531 patients identified over the years and countries.

In previous researches done by various authors, the use of singular algorithms was appropriated to carry-out the research and calculate the results and conclusions. In this research paper, we have emphasised on the use of combined learning algorithms instead of using the learning models one at a time in order to obtain more accurate results.



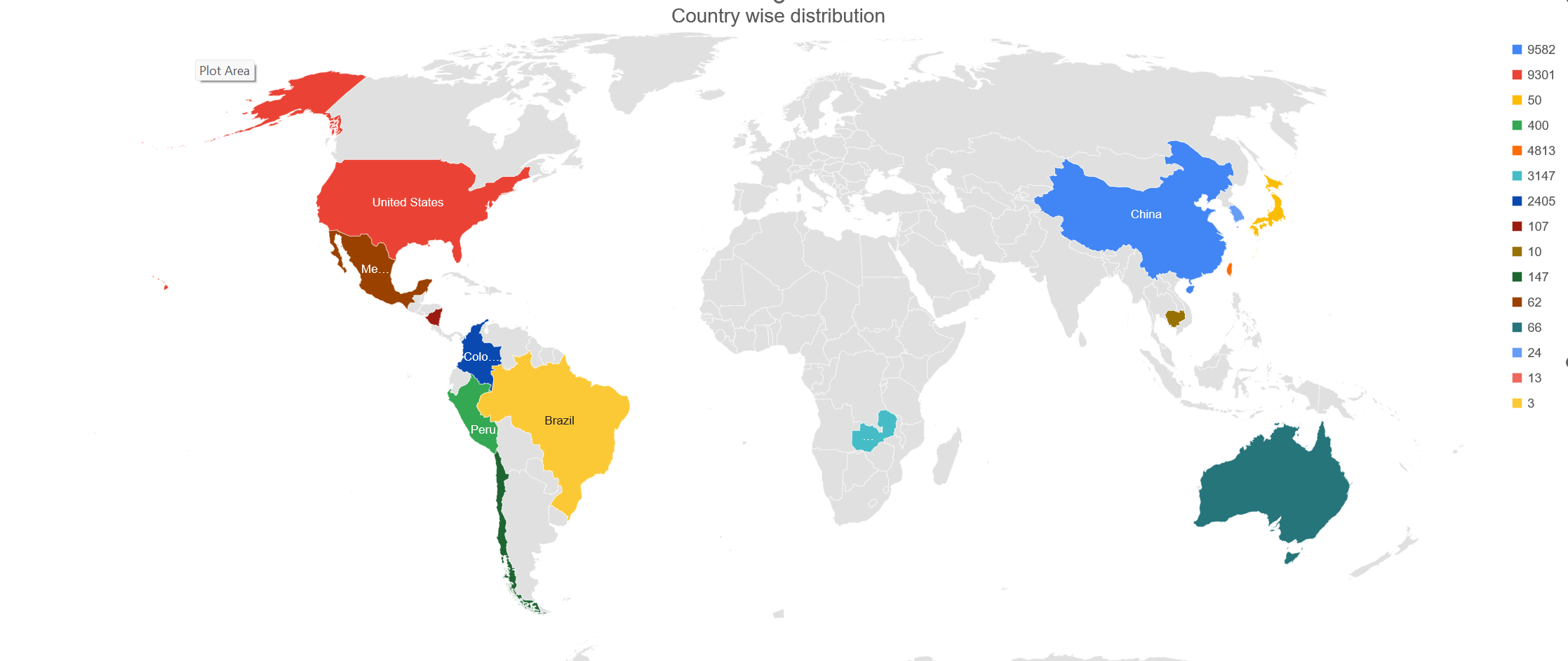
* 1. Data used

Within the data used in this paper, the patient’s information such as collection year, age of patient and symptoms shown by the patient were studied and passed into the models which were then evaluated and concluded upon.

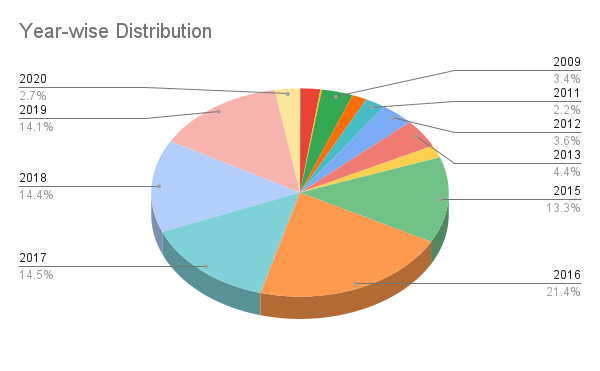
The dataset also contains the type of pathogen the patient is infected with, the demographic information of the patient, the strain the patient is infected with, and many other ineffective information.

The dataset containing appropriate structure and variables was procured from Bacterial and Viral Bioinformatics Resource Centre. For this project, all obtained data remains constrained to humans only, in consideration of feasible and ethical reasons.

Some attributes such as patient’s name, email-id and similar were removed in the beginning since they contain irrelevant information to the purpose of the project and the models used



Country wise distribution of Records in Dataset



(Left) Year-Wise Distribution

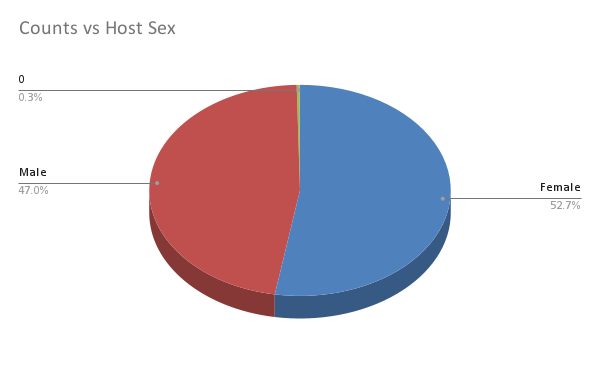


Figure 4:(Right) Gender-Wise Distribution Of ‘Positive’ Class

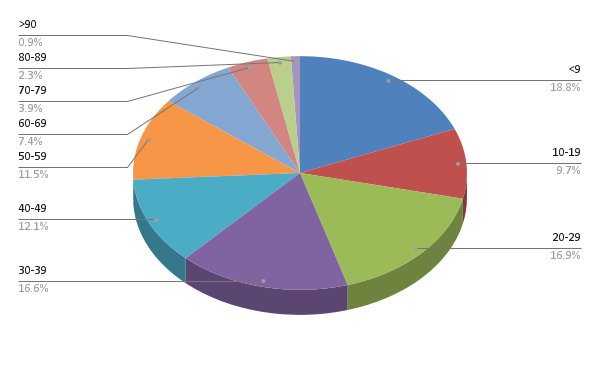


Figure 5: Age Distribution Chart

* 1. Data Pre-processing

Initially, the data consisted of 30,531 rows and 98 columns, most of the columns having little to no instances of data. Various data mining techniques were used to pre-process and transform the data so that the learning models are able to comprehend the data without giving out too many bugs and errors. The data was cleaned of necessary outliers and impurities, and later recompiled with changed positions of certain data points or rows or columns and different, easier-to-understand labels. All blank columns were firstly taken out.

Next, the “symptoms” column contained data points in the form of key-value pairs, which were then separated into rows and columns, respectively. Furthermore, all the Yes/No/Blank data points were converted into binary data, with Blank and No being considered as 0, and yes being considered as 1. The data points in the gender column were converted to binary data, with Males getting initialised as 0 and Females getting initialised as 1. Then, only rows having all instances filled were taken into consideration, the rest were removed. Any row or column value having its instance as null was removed and disregarded. Similarly, all columns having more than 95% data as null values were removed.

A certain number was assigned to the string values in the “chronic condition” column in consideration of how many conditions were in many instances of the data were combined together in one large string. These numbers assigned were then aggregated so as to compile some conditions and form a model-understandable arithmetic value. After having finalised the data pre-processing, 18582 rows and 23 columns remain, which only contain a patient’s fully-procured relevant data with the collection year, symptoms, gender and age, all columns necessary for processing.

At the start of our analysis, we observed that there were 697 distinct combinations of medical diseases present in the data. However, we needed to identify the most relevant factors to ensure our analysis was accurate and meaningful. Thus, we identified the four most commonly repeating factors in the dataset. Based on these four factors, we defined a power set that divided the factors into 30 unique combinations.

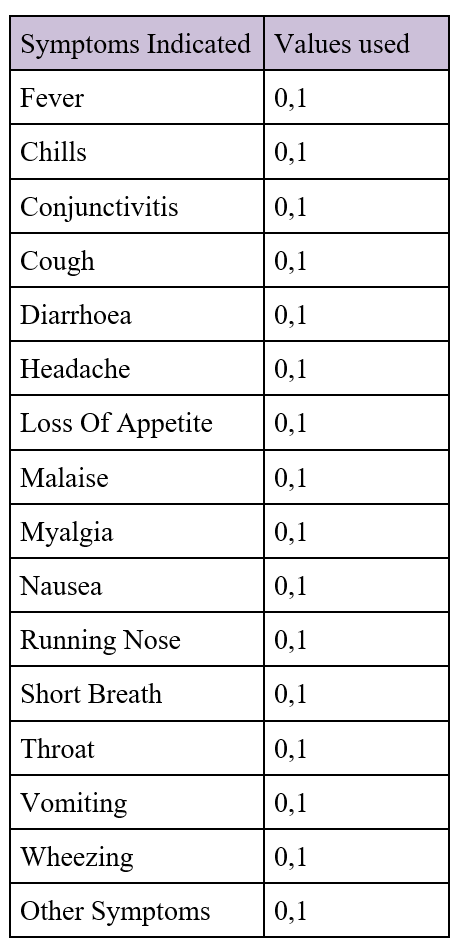
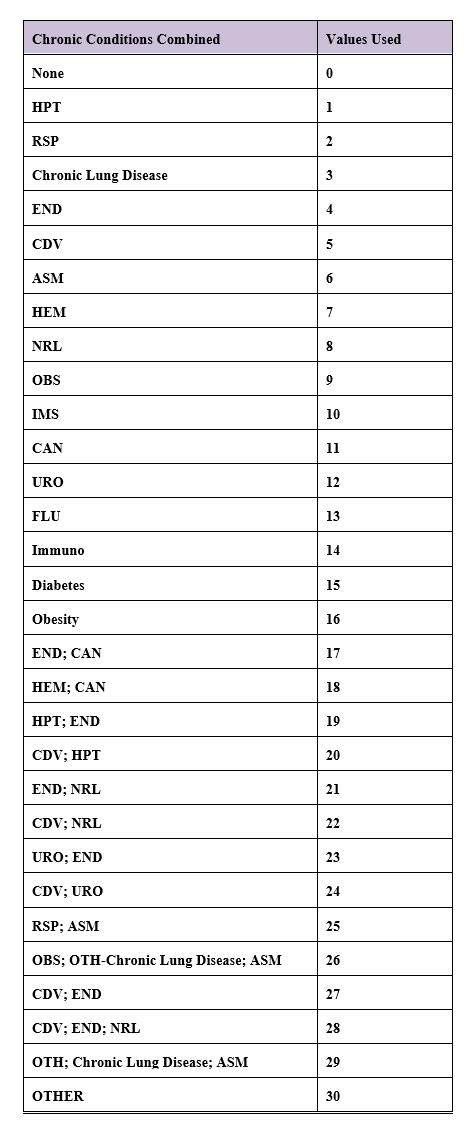


Table 1: Medical Condition

Table 2: Symptom Distribution

To better understand the medical conditions, present in the data, we used Table 2 to define the value substitution in the dataset. This allowed us to transform the medical conditions into a

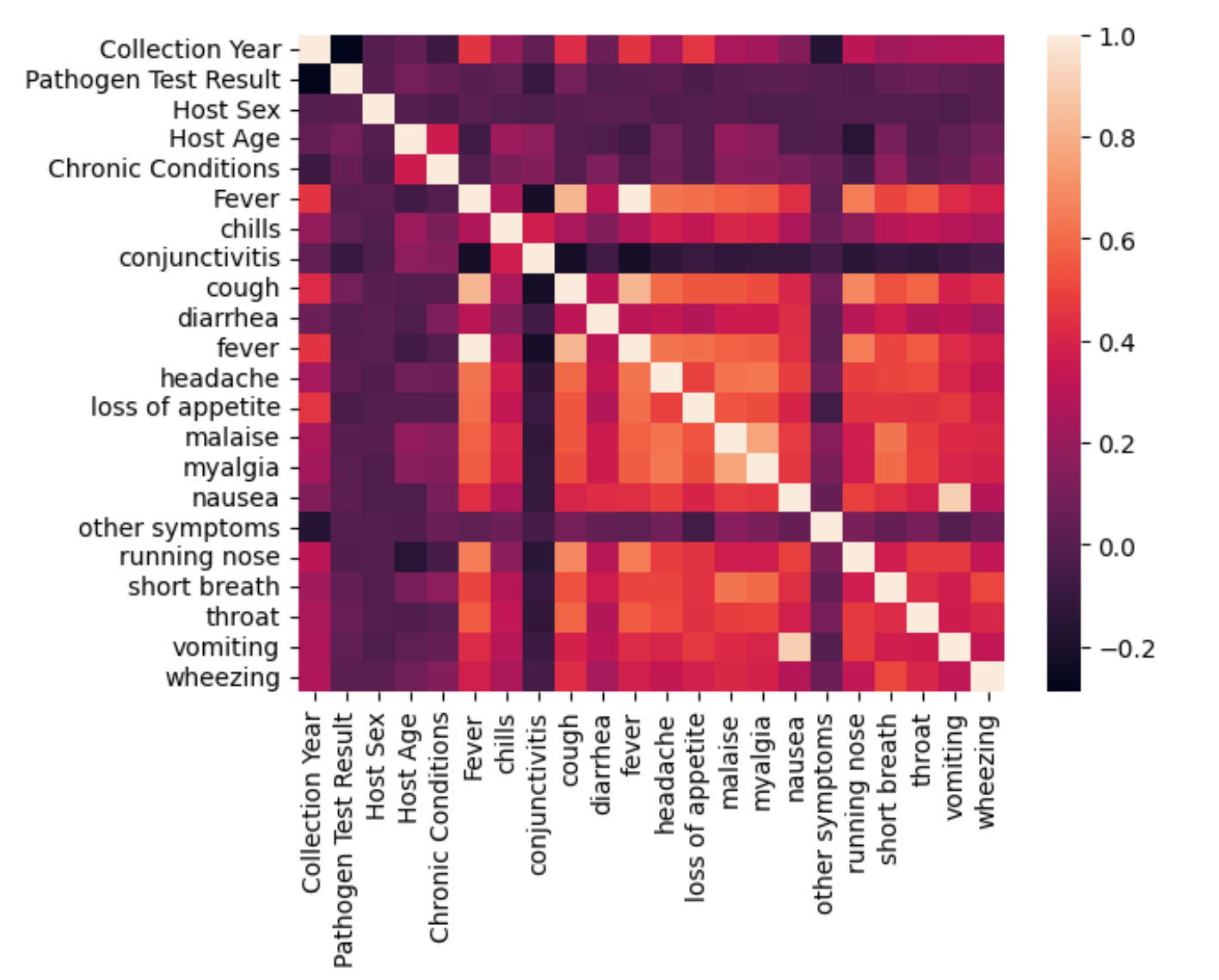


Figure I: Heat map of Correlation between each Variable in after cleaning

more manageable form that could be used for further analysis. By identifying the most relevant factors and defining a power set, we were able to streamline the data and ensure our analysis was focused on the most important medical conditions related to early detection of influenza.

3.4 Training and testing of data

The data was divided into sets of 70% or 80% training and rest for testing, to be comprehended by the learning models. The data was divided using the train\_test\_split command available in Python through the use of the Scikit-Learn package. Hence, the models were trained with either 13,007 training samples or 14,865 training samples. Training sample divided as ‘Positive’ (54.31%) and ‘Negative’ (45.69%).

Formula defined for our analysis:

Pathogen Test Result ~ Gender + (all the 25 Symptoms) \* + Medical conditions

*\*For the Symptoms attributes refer Table 2*

3.5 Classification Algorithm

3.5.1 SVM

Support vector machines (SVMs) are a category of supervised learning algorithms utilized for classifying, regressing, and identifying outliers in data. To address different machine learning tasks, there exist specialized types of SVMs such as support vector regression (SVR), which is an expansion of the support vector classification (SVC) technique. SVMs are different from other classification algorithms because of the way they choose the decision boundary that maximises the distance from the nearest data points of all the features. The superiority of the linear SVM algorithm over other algorithms, such as k-nearest neighbours, stems from its ability to identify the optimal line to classify data points. The algorithm selects the line that maximizes the distance between the line and the closest data points, allowing it to capture intricate relationships within the data without necessitating extensive pre-processing or transformation.

Support Vector Machines (SVM) is a machine learning algorithm that involves tuning several parameters, depending on the type of kernel used. These parameters include the regularization parameter (C), kernel type, degree, and gamma. For our study, we focused on the linear kernel at the initial level. Therefore, we did not require the values of gamma and degree since these parameters are used for higher-level kernels.

To set the kernel to linear, we used the method "svm Linear" in the train () method. The default value of C for classification was set to 1. This allowed us to tune the SVM algorithm to suit our specific needs in predicting early detection of influenza using the BV-BRC dataset. By focusing on the linear kernel and tuning the regularization parameter, we were able to optimize the performance of the SVM algorithm for our dataset.

* The data was first split into two sets of testing and training data, with 30% of the data taken up for testing. After setting up the standard scaler, parts of training and testing data were passed into it. These transformed parts, along with the other parts of the training set were used to train the model, which was then fitted upon the test set of the data

3.5.4 RANDOM FOREST

A random forest is a powerful machine learning technique used for solving classification and regression problems. It employs ensemble learning, which involves combining multiple decision trees to provide a robust and accurate solution to complex problems. The random forest algorithm predicts the outcome by aggregating the predictions of many decision trees. The average or mean output of the individual trees is used to establish the final prediction. Increasing the number of trees in the forest enhances the precision of the outcome.

A decision tree is a basic component of the random forest algorithm and has three main parts: decision nodes, leaf nodes, and a root node. The algorithm divides the training data set into branches, which are further divided into additional branches until a leaf node is reached. The leaf node represents the end of the decision-making process and cannot be further divided. The decision nodes in the tree represent attributes or features used for predicting the outcome. These nodes serve as links to the leaves, where the final prediction is made. The data was split into training and testing data, with 20% of the data being used for testing. Then, the classifier object was set up taking 100 trees into consideration. Then, this model was trained on the training set.

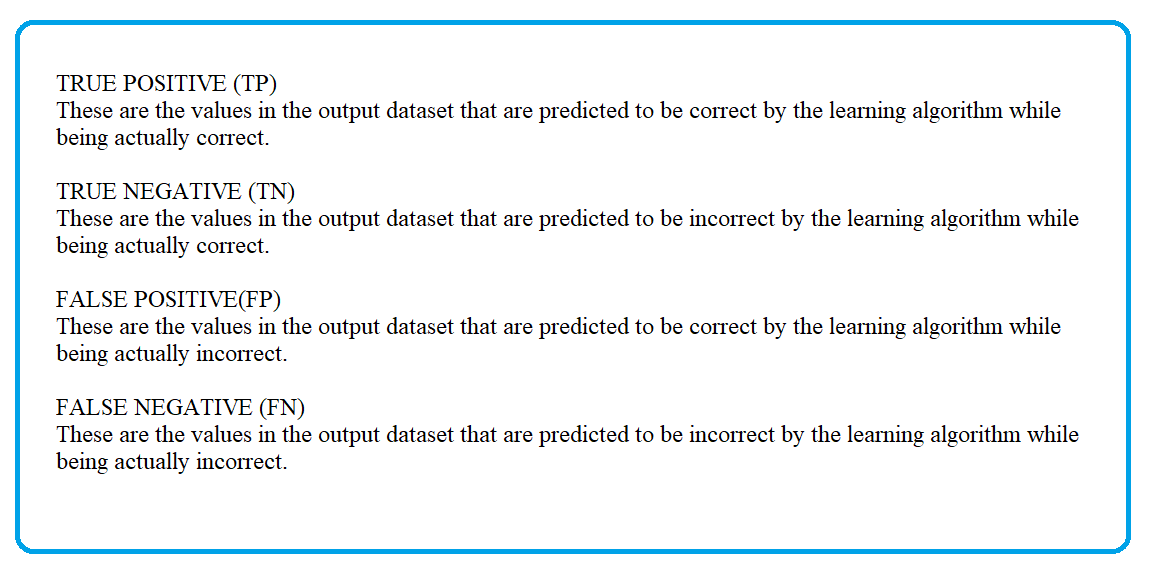
Further, to obtain better results and expand on our research, two techniques were combined and taken at a time.

* The ‘Pathogen Test Result' values of the two algorithms were taken along with their metrics such as F1 Score, precision and recall score. These attributes were combined with the use of AND logic, that is, if either of the algorithms gave 0 as output, the final result was taken as 0. Only when both algorithms gave 1 as output, the final result was considered as 1. This logic is applied in order to further verify the accuracy of both of the algorithms and bring out a much more significant output.

3.6 Accuracy Assessment and Comparisons

The evaluation parameters used in this study were based on the work of Zhu et al. [30]. These parameters were selected to ensure that the performance of the different machine learning models could be compared fairly and accurately. The specific evaluation parameters used included sensitivity, specificity, and accuracy. Sensitivity is the proportion of true positives among all actual positive cases. Specificity is the proportion of true negatives among all actual negative cases. Accuracy is a statistical metric that quantifies the performance of a model in correctly predicting the target variable. It is commonly used as a evaluation metric in classification tasks. However, accuracy may not be the best measure of performance for imbalanced datasets, where one class may dominate the other. In such cases, other metrics such as precision, recall, and F1-score may provide a more comprehensive evaluation of the model's performance. These parameters were calculated for each of the four machine learning techniques used in the study:

“Support Vector Machines, K-Nearest Neighbors, Artificial Neural Networks, and Random Forest”. The evaluation results were then used to compare the performance of the different techniques and identify the most effective method for early detection of influenza.



ACCURACY

It is the measure of how close the output results are to the desired results, i.e., the number of correct predictions made by the algorithm. It is calculated by dividing the total number of true results by the total number of all results.

Accuracy = (TP+TN)/(TP+FP+TN+FN)

# Research findings / Results

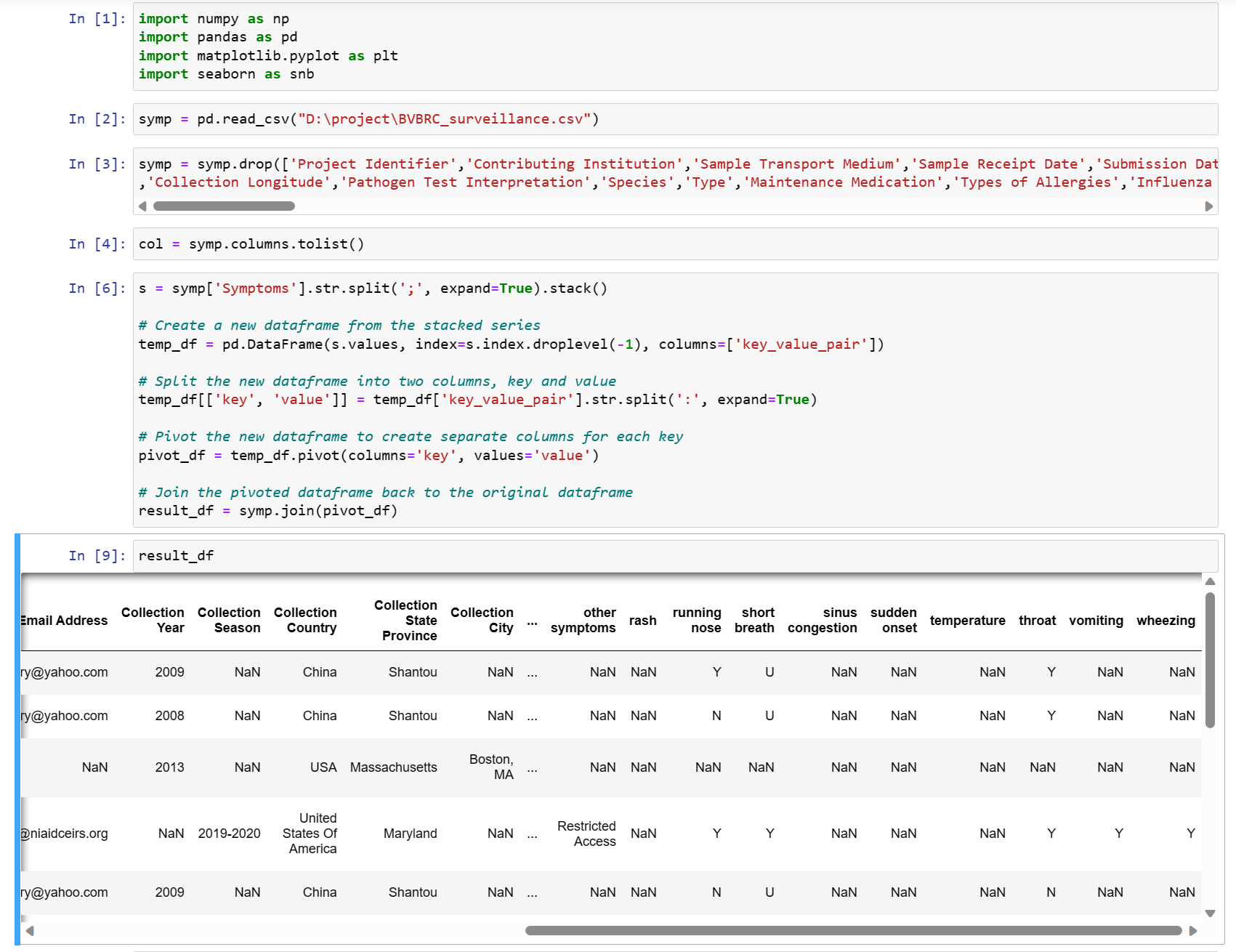
The analysis involved applying different learning techniques on the dataset and the results were evaluated. This helped in presenting the data in a meaningful and organized manner. Based on the dataset and the learning techniques used, the results of the project are as follows:

1. Applying the Linear Support Vector Machine algorithm, the True Positive Rate turns out to be 63.70, while Specificity turns out to be 78.95. The accuracy for this algorithm is 70.66. On its own, it is a pretty low performing algorithm.
2. Applying the Random Forest algorithm, the True Positive Rate turns out to be 84.09, while Specificity turns out to be 90.95. The accuracy for this algorithm is 87.20. This is the better performing algorithm than SVM.

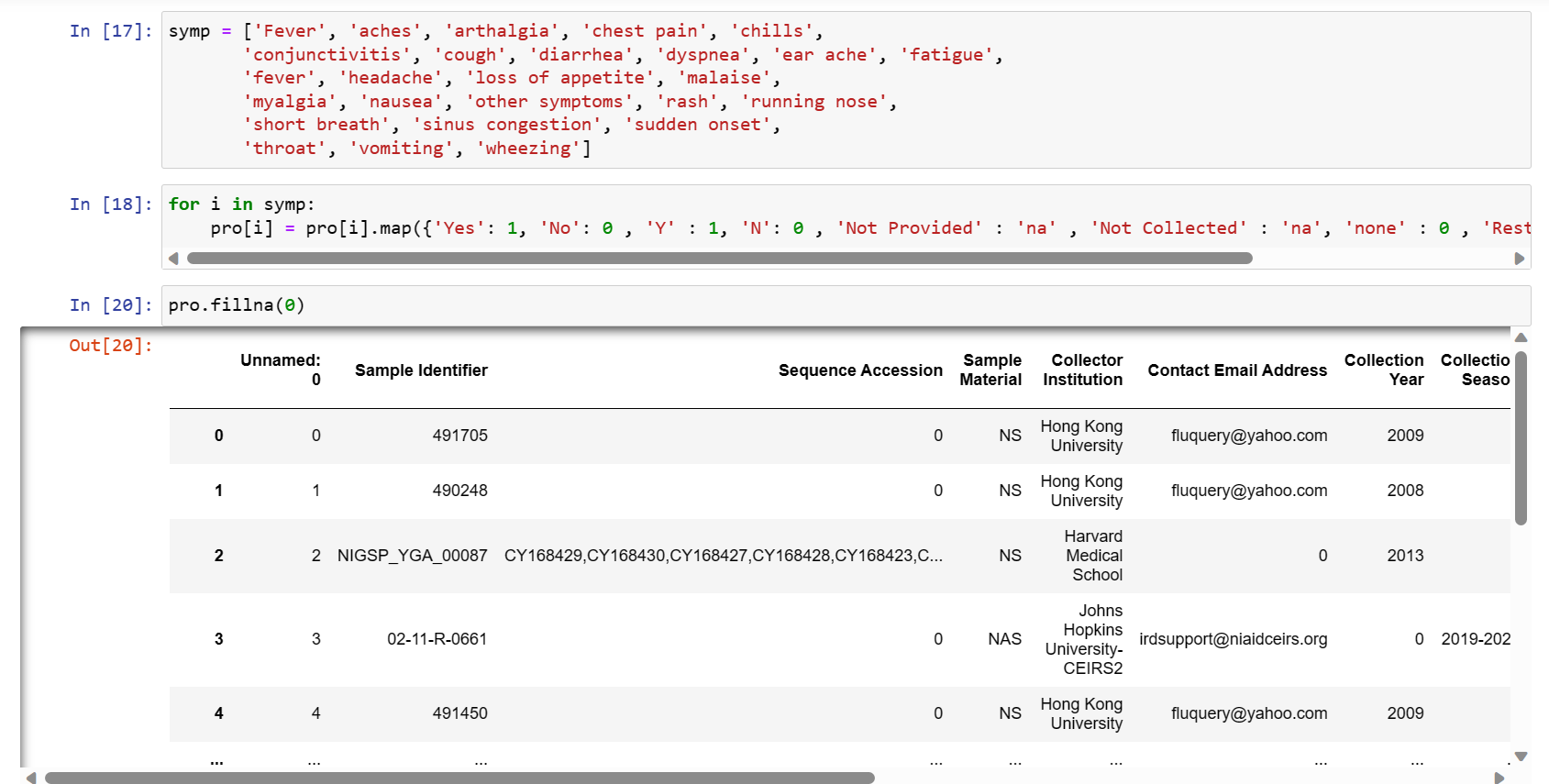
In order to obtain much higher accuracy and more significant results, these techniques were combined with each other, providing these results:

Code and Outputs

*The following code was implemented to divide the Symptom attribute into individual symptom attributes:*

******

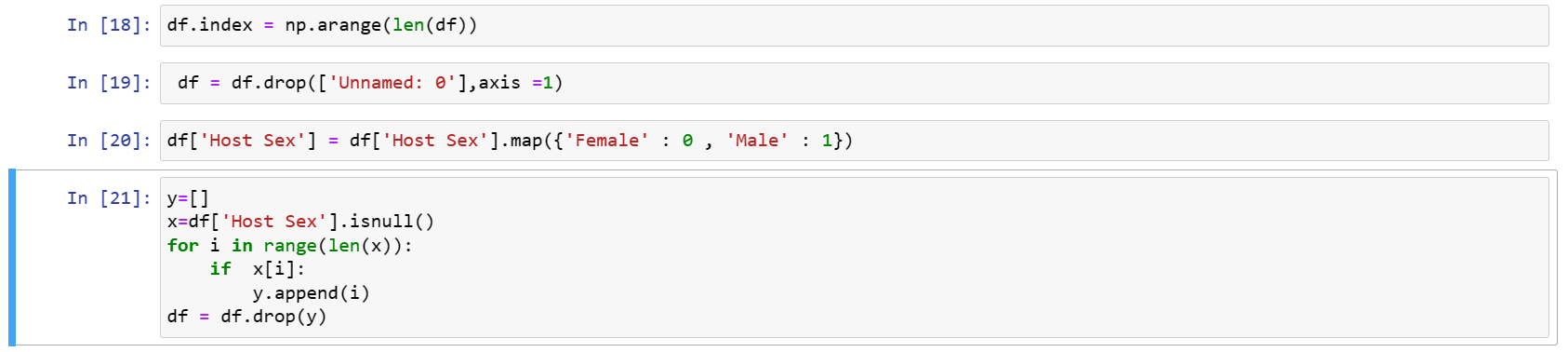
*Code to divide the Symptom attribute to each individual symptom attribute (Handling the values ‘Yes’ and ‘No’).*



*Code to map Positive and Negative value of Target variable with 1 and 0 respectively. Handling Age variable disrupted values.*

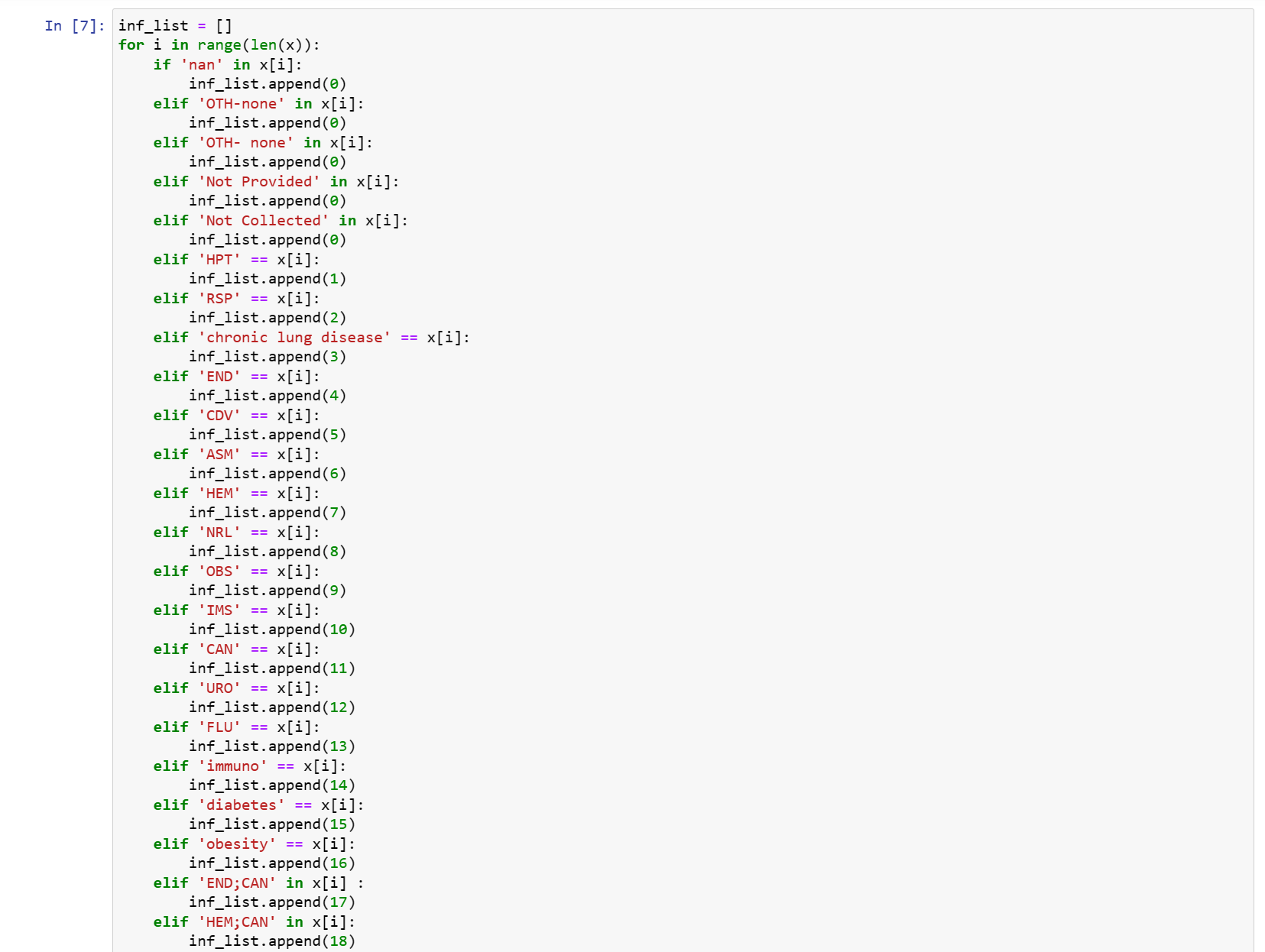


*Code to map Male and Female value of Host Sex with 1 and 0 respectively.*



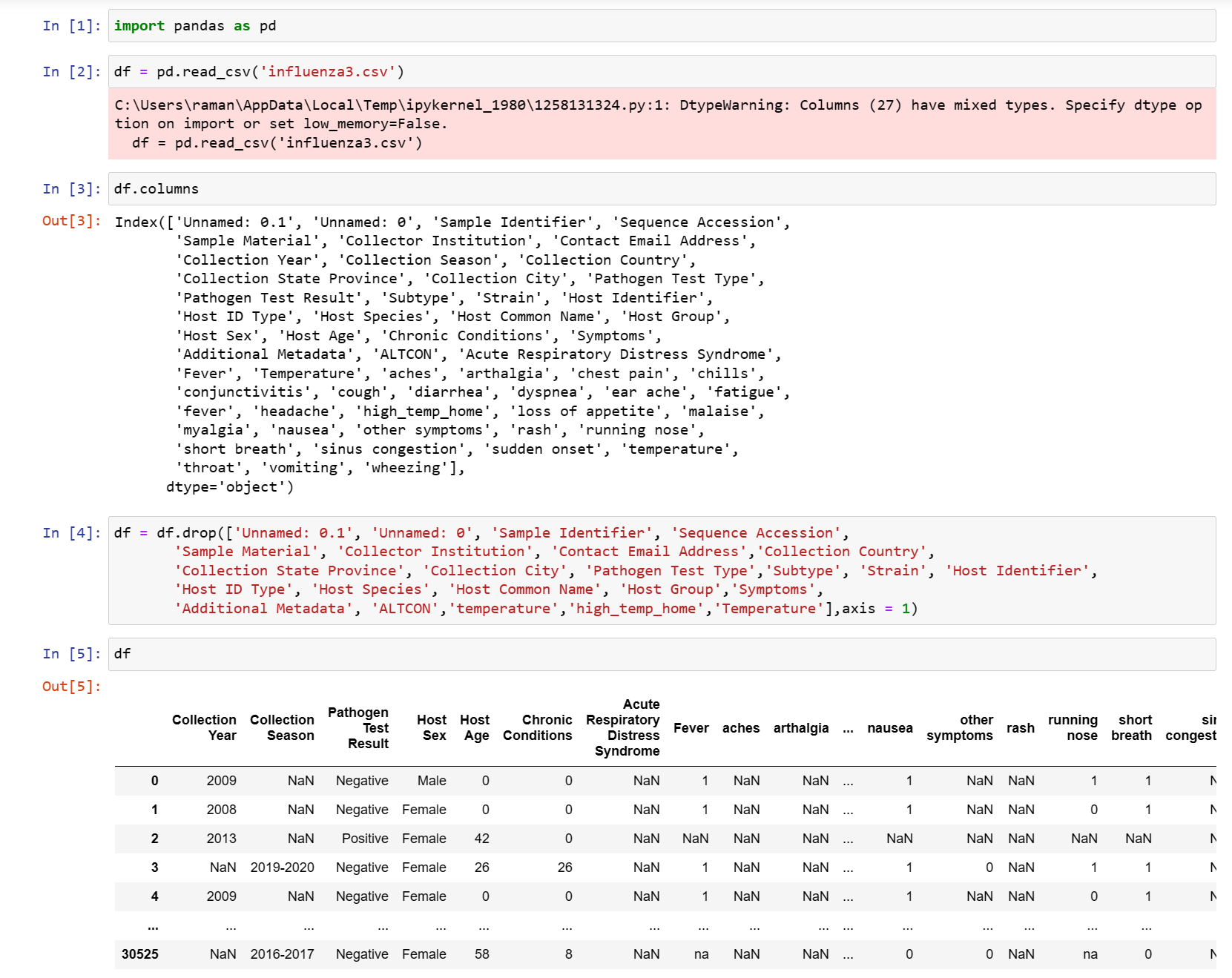
*Pre-processing Chronic Condition variable of Data set.*

******

******

******

*Code for removing unnecessary variable****.***

******

*Code to merge Collection Year variable and Collection Season variable.*

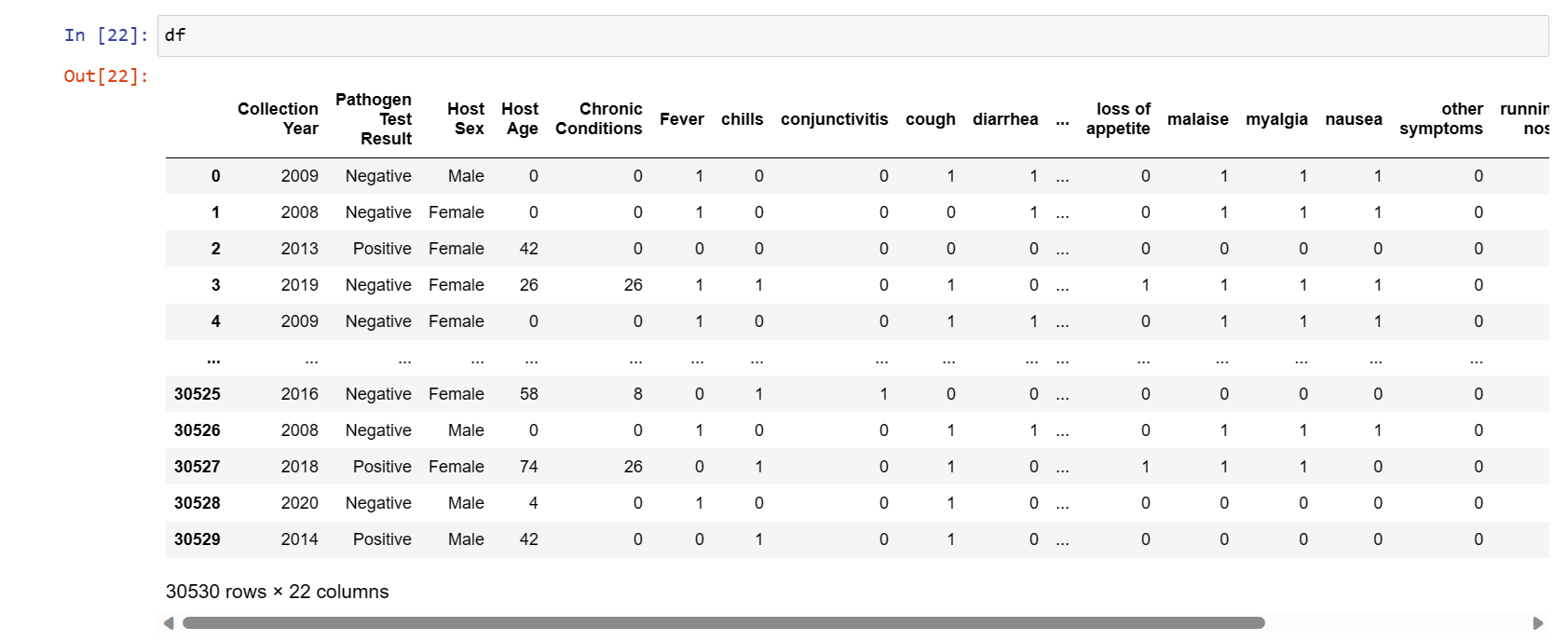
******

******

*Code for removing null value rows, symptoms having 95% of null values and filling null values with 0 in remaing values****.***

******

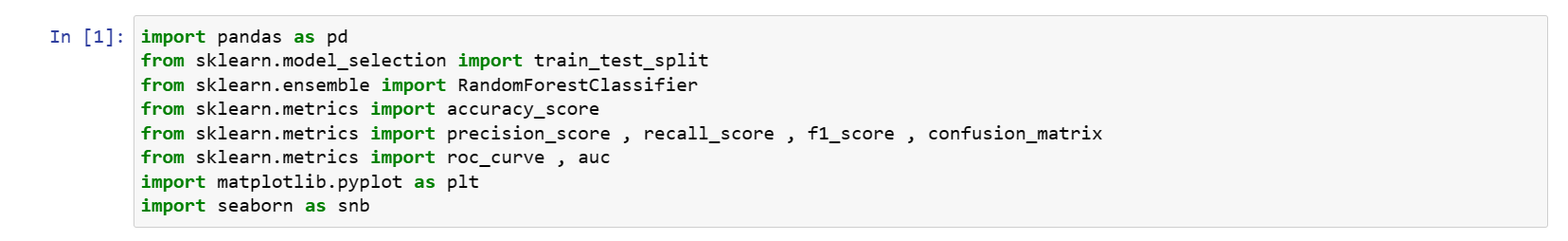
******

******

*Code for splitting data set for training and testing.*

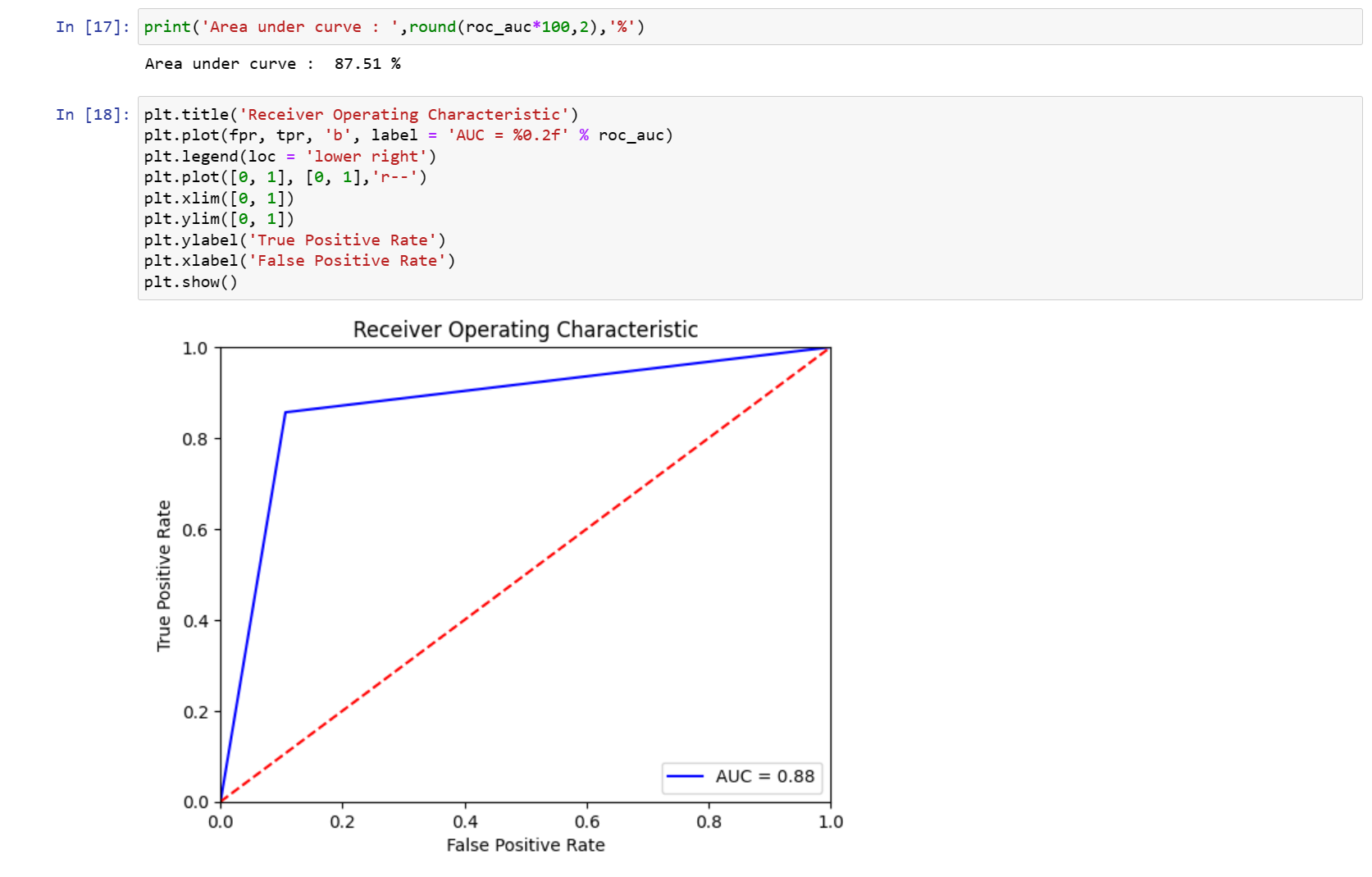
******

*Code for implementing RandomForest Model.*

******

******

******

******