School of Computer Science and Engineering

Impact of Lifestyle and Diet in Combating SARS-2 COVID-19 in India

ITA5007 –Data Mining Project Report

Submitted to:

Dr Punitha.K

Submitted by:

ABIRAMI.R (20MCA1024)

SWETHA.S (20MCA1008)

MRIDUL NATHANI (20MCA1029)

In partial fulfilment for the award of the degree of

Masters
In
Computer Applications

June 2021

TABLE OF CONTENTS

1. Abstract:

The world is experiencing an unprecedented challenge due to the coronavirus disease (COVID-19) pandemic. Whether there is an association between lifestyle behaviors, diet and the acquisition of COVID-19 remains unclear. The aim of this project is to analyze the role played by lifestyle behavior and diet followed among the general population in India in controlling the widespread outbreak of coronavirus (covid-19) pandemic. This project also identifies the likeability of a person getting the novel coronavirus based on the data identified through the lifestyle and dietary factors of an individual.

Impact of Lifestyle and Diet in Combating SARS-2 COVID-19 in India

2. Introduction:

A novel Coronavirus was identified in December 2019. A pandemic severely affected around 220 countries worldwide and WHO had announced it as a pandemic on 30th January 2020 stating that it can cause severe loss to humankind. Many variants and mutants started developing in due time causing several deaths due to SARS-2 covid19. Although everything seems new and unidentified India was initially able to control the covid cases well despite having a wide population.

The world is experiencing an unprecedented challenge due to the corona virus disease (COVID-19) pandemic. Whether there is an association between lifestyle behaviors, diet and the acquisition of COVID-19 remains unclear. The aim of this project is to analyze the role played by lifestyle behavior and diet followed among the general population in India in controlling the widespread outbreak of coronavirus (covid-19) pandemic.

3. Feasibility study:

Technical Feasibility:

We have done some research work and found out that it is feasible to implement this project on a large-scale basis. To make this project, we just require a rapid-miner and a survey dataset with essential column names. An account for registering with rapid-miner and to work with the interface.

Economic Feasibility:

This project is also economically feasible as the only cost involved in this project is developing it. Rapid Miner is a data mining tool that helps in model deployments. Alternative tools like IBM Watson studio or R Studio can also be used.

Operational Feasibility:

This project can be executed with the following System requirements based on the tool chosen. Rapid-Miner requires the following system:

Recommended:

- Quad-core
- 3GHz or faster processor
- 16GB RAM
- >100GB free disk space

Operating System:

- Windows 7, Windows 8, Windows 8.1, Windows 10 (64-bit *highly* recommended)
- Linux (64-bit only)
- macOS X 10.10 - 10.15

Java platform

- 64-bit *highly* recommended
- OpenJDK Java 8

4. About the dataset:

An online survey was conducted among the general population in India with distinguishing parameters ranging from physical exercise, diet intake, etc. to identify the relationship between lifestyle and diet in combating Covid-19. Vaccines were not introduced during the time of the survey. Hence this will have a conditional impact on the results produced.

The dataset was strictly taken abiding by the rules of data privacy.

The dataset contains the names of the following columns, such as Age, Gender, Height, Weight, Lifestyle habits such as Smoking, Alcohol, Tobacco, Chronic illness if any such as Heart Disease, Respiratory, Auto-immune Disease, etc., Sleep schedule, Food Frequency, Exercise Regularity, Bathing Frequency, Yoga practices, Intake of Supplements, Home Remedies, Intake of medicinal Decoction, Practiced Medicinal type, Vegan or not, Frequency of Non-veg intake, Fruit's intake, Use of Indian herbs, History of covid-19 and its recovery rate. The dataset contains around 226 responses from various age groups. An online review was conducted to collect RAW Data form actual users.

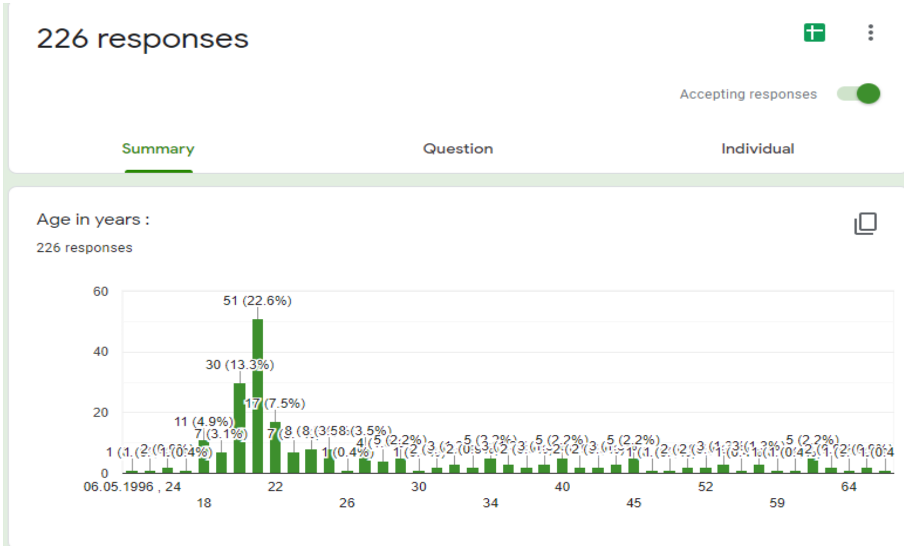


Figure 1 Age range of the responses

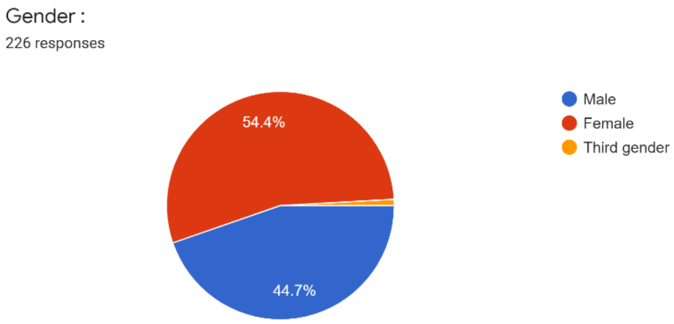


Figure 2 Gender of the responses

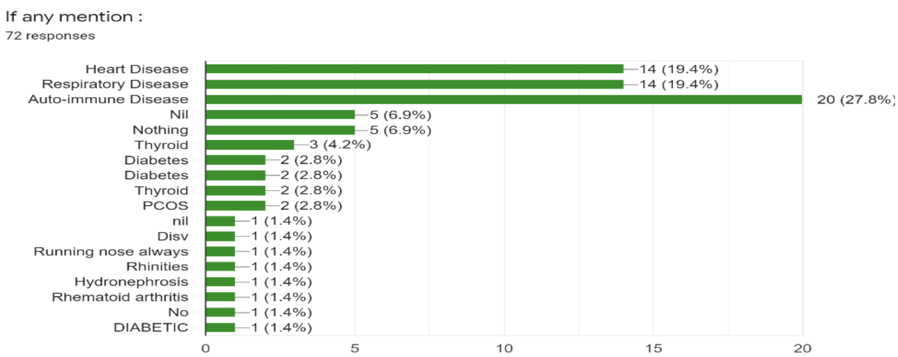


Figure 3 Survey Analysis

Link for Google Form: https://forms.gle/J42gPGDMXCA6rme66

5. Design and flow of models:

The design and process flow focus on one data mining approach that is clearly explained below.

5.1 Data Mining Approach:

Process flow of machine learning approach is as follows: -

- Planning the strategy.
- Collection of the dataset.
- Data preprocessing.
- Visualizing the preprocessed dataset.
- Splitting the dataset into train and test data.
- Implementing three data mining classifiers.
- Gathering prediction results.

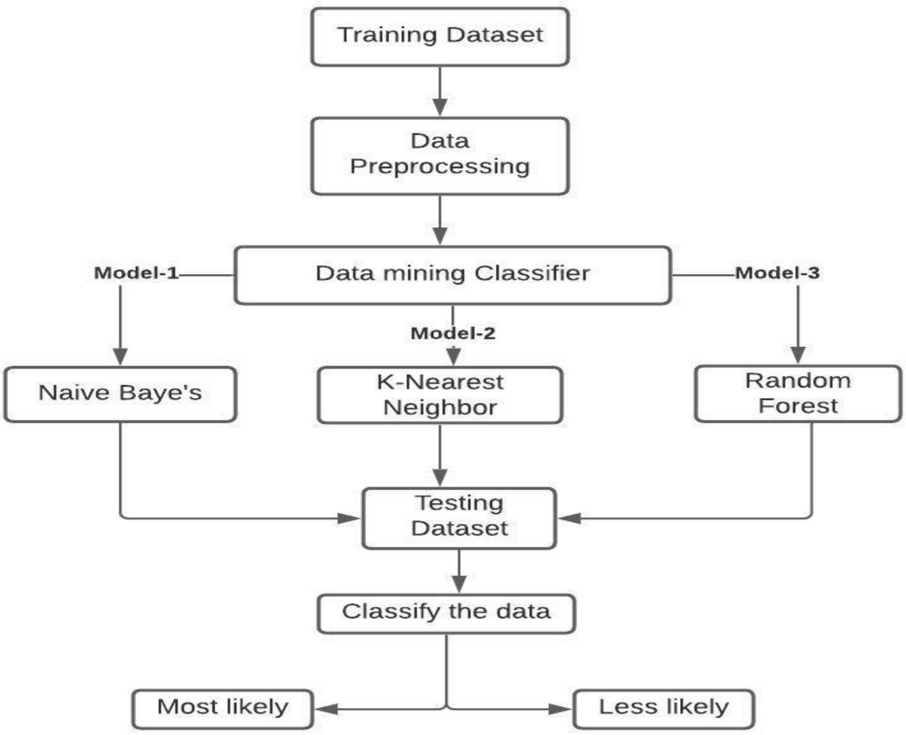


Figure 4 Flowchart of the process flow

6. Pre-Processing of Data

- Pre-processing: cleaning
- Duplicate removal, supplying missing values.
- Transforming the RAW data and transform it in a useful and efficient format.
- It involves cleaning Missing Data, Noisy Data etc.

Filter (24 / 24 attributes): Search for Attributes

Name	Type	Missing	Least	Most
Medicinal type	Polynominal	0	Traditio [...] ine (102)	Western medicine
Traditional medicine type	Polynominal	69	Siddha;H [...] pathy (1)	Ayurveda (40)
Vegetarian	Polynominal	0	Yes (104)	No (121)
Consumption rate of Non-veg	Polynominal	58	Rarely (27)	Sometimes (86)
Fruits Intake	Polynominal	0	Rarely (30)	Sometimes (114)
Indian Herbs	Polynominal	0	No (30)	Yes (156)
Bathing Frequency	Polynominal	2	Alternate days (33)	Once a day (145)
Histroy of covid-19	Polynominal	0	Yes (58)	No (167)

Figure 5 Cleanup for Missing Values

Index	Nominal value	Absolute count	Fraction
2	Respiratory Disease	11	0.153
3	Heart Disease	10	0.139
4	Nil	5	0.069
5	Nothing	5	0.069
6	Thyroid	3	0.042
7	Diabetes	2	0.028

Figure 6 Updated values

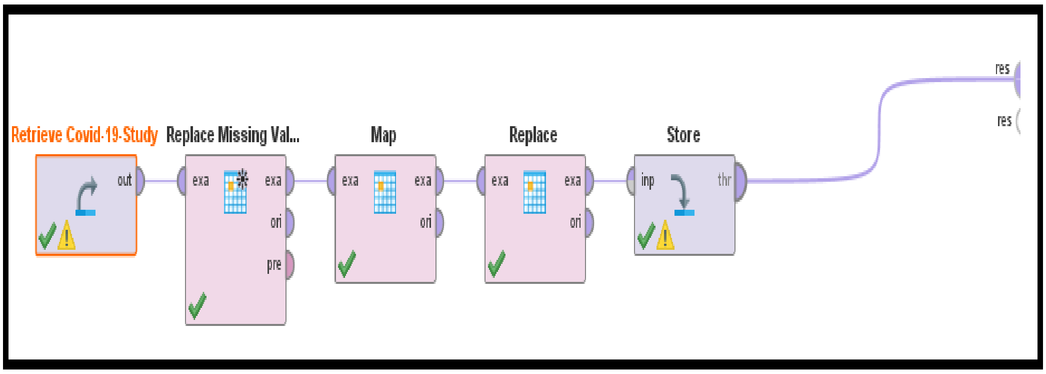
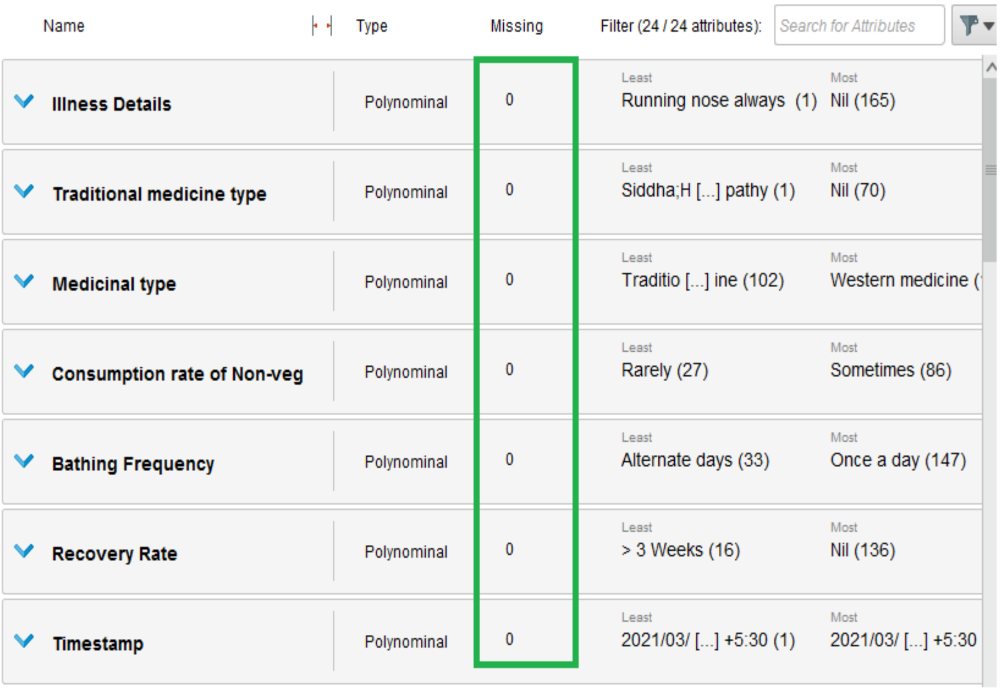


Figure 7 Rapid-Miner Cleanup Procedure



Filter (24 / 24 attributes): Search for Attributes

Name	Type	Missing	Least	Most
Illness Details	Polynominal	0	Running nose always (1)	Nil (165)
Traditional medicine type	Polynominal	0	Siddha;H [...] pathy (1)	Nil (70)
Medicinal type	Polynominal	0	Traditio [...] ine (102)	Western medicine (
Consumption rate of Non-veg	Polynominal	0	Rarely (27)	Sometimes (86)
Bathing Frequency	Polynominal	0	Alternate days (33)	Once a day (147)
Recovery Rate	Polynominal	0	> 3 Weeks (16)	Nil (136)
Timestamp	Polynominal	0	2021/03/ [...] +5:30 (1)	2021/03/ [...] +5:30

Figure 8 Clean Data with No Missing Values

7. Risk Analysis:

The major limitation involved in this project is the vaccines and its effects could not be incorporated since the survey was issued before the introduction of vaccines in India. This can have a significant impact in the prediction of likeability. Also, the dataset taken under consideration has only limited features and it can't be guaranteed that the prediction of the chances of contracting the virus is only dependent on the lifestyle factors. Even if a person follows an ideal healthy lifestyle, there are still more chances for the person to experience the worst symptoms like inability to breathe, stomachache and headache in some cases. More features should be taken into consideration with Standard Medical reports which include Blood Tests and Prenatal Tests that determine the immune system levels along with the history of Vaccinations for Better Prediction.

8. Module List:

A detailed description of all the modules in the two approaches is explained below.

For data modelling we are using an efficient tool that is “Rapid Miner”. Rapid Miner Studio is a powerful data mining tool for rapidly building predictive models. The all-in-one tool features hundreds of data preparation and machine learning algorithms to support all your data mining projects.

8.1 Data Mining Classifiers:

Data mining Classifier has the ability to automatically categorize the data given into one or more classes. In our dataset, categories can be referred to as Most likely to get covid and Less likely to get covid. We have implemented classifiers like Naive Bayes (NB), K-nearest Neighbors (KNN), and Random Forest (RF).

8.1.1 Naive Bayes:

Assume a probability model on generation of data.

$$\text{predictedclass } c = \max_{c_j} p(c_j \mid d) = \max_{c_j} \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Apply Bayes theorem to find most likely class as:

$$c = \max_{c_j} \frac{p(c_j)}{p(d)} \prod_{i=1}^{n} p(a_i \mid c_j)$$

- ✓ Naïve Bayes Assume attributes conditionally independent given class value
- ✓ Easy to learn probabilities.
- ✓ Useful in some domains e.g., text
- ✓ Handles both discrete and continuous data.

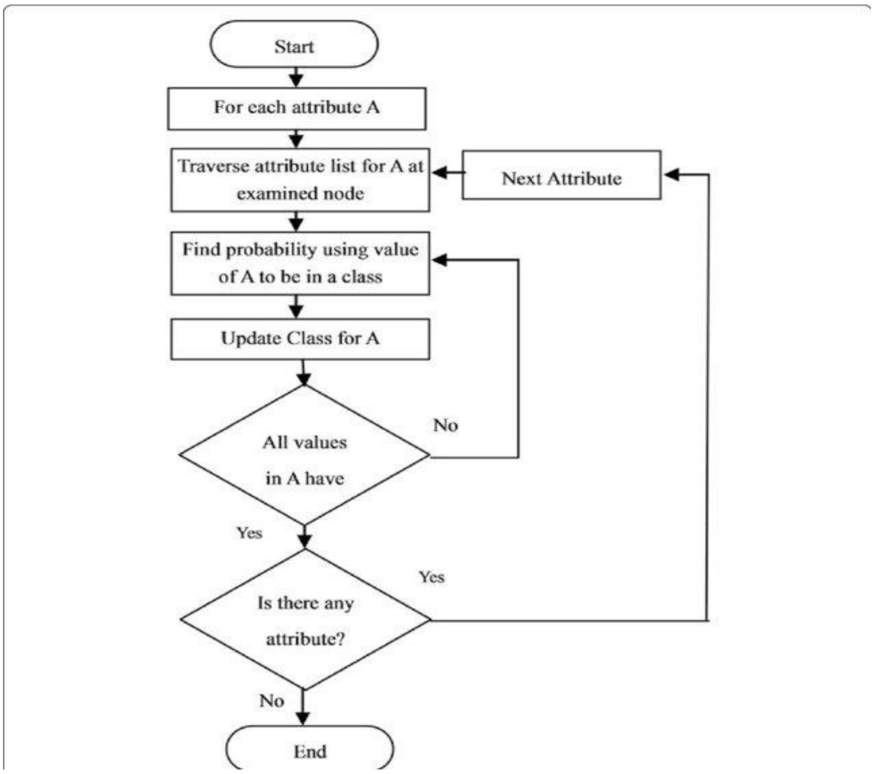


Figure 9 Process Flow of Naive Bayes

8.1.2 K-Nearest Neighbors:

Define proximity between instances, find neighbors of new instance and assign majority class Case based reasoning: when attributes are more complicated than real-valued.

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

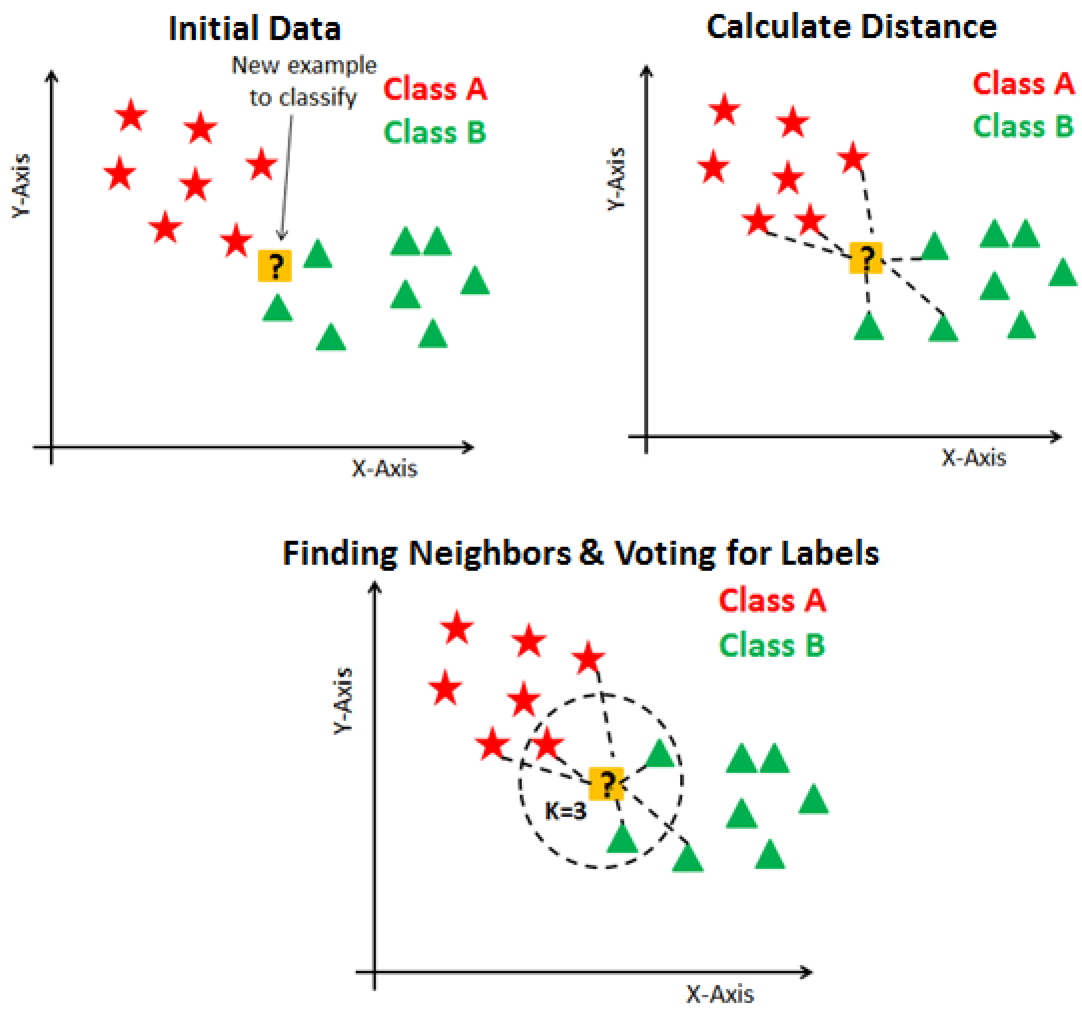


Figure 10 Visualization of KNN

8.1.3 Random Forest:

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become our model's prediction.

Random forest is a supervised learning algorithm. The "forest" it builds, is an ensemble of decision trees, usually trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result.

Put simply: random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

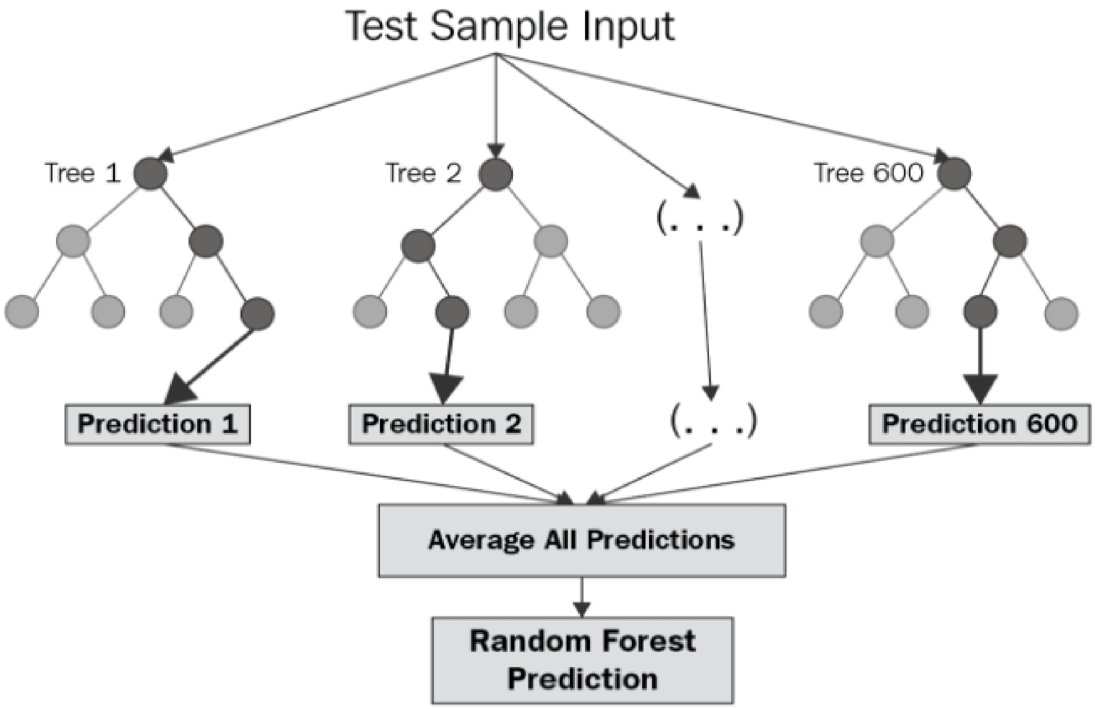


Figure 11 Process Flow of Random Forest

9. Implementation :

9.1 Process:

9.1.1Naïve Bayes Model:

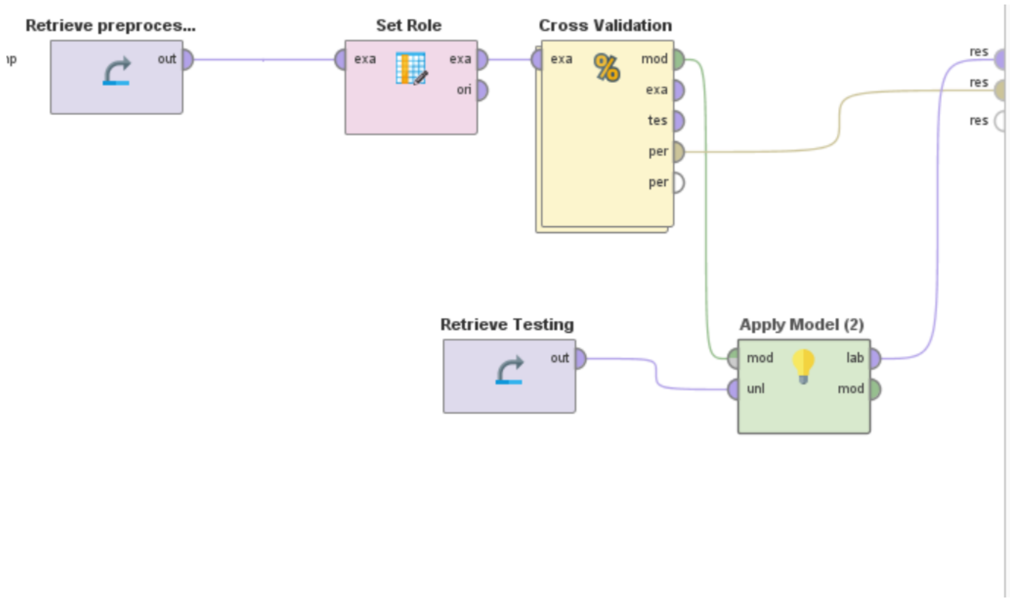


Figure 12 : Naive Bayes Process

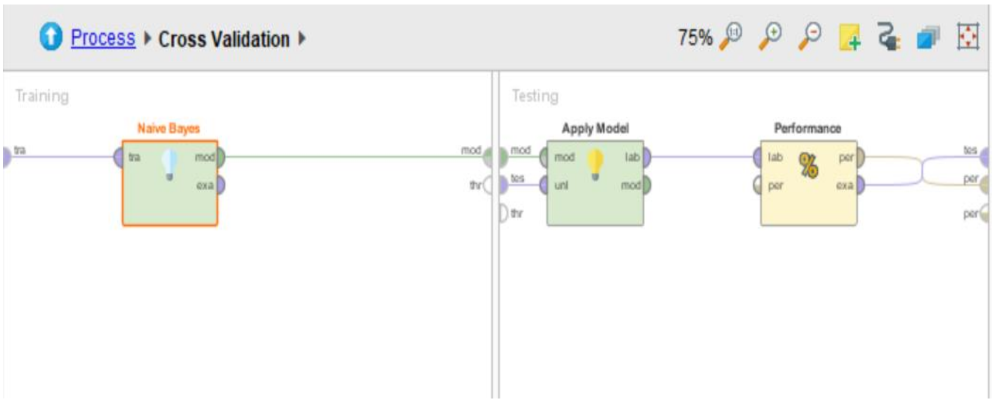


Figure 13 Naïve Bayes Cross Validation Process

Results:

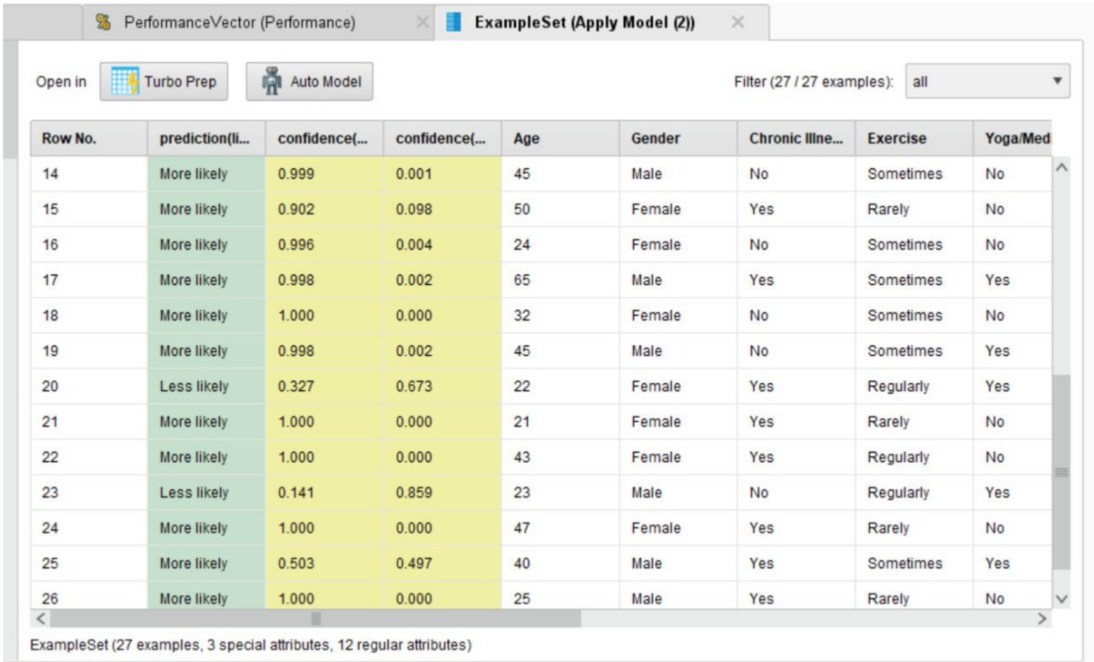
PerformanceVector (Performance) | ExampleSet (Apply Model (2))

Open in Turbo Prep | Auto Model — Filter (27 / 27 examples): all

Row No.	prediction(li...	confidence(...	confidence(...	Age	Gender	Chronic Illne...	Exercise	Yoga/Med
1	More likely	0.995	0.005	38	Female	No	Sometimes	Yes
2	More likely	1.000	0.000	21	Female	No	Regularly	No
3	More likely	0.999	0.001	31	Female	Yes	Rarely	Yes
4	More likely	0.995	0.005	24	Male	No	Regularly	Yes
5	More likely	1.000	0.000	27	Female	No	Sometimes	No
6	More likely	1.000	0.000	20	Female	No	Sometimes	No
7	More likely	0.969	0.031	18	Female	No	Sometimes	No
8	More likely	0.971	0.029	20	Male	Yes	Sometimes	Yes
9	More likely	0.996	0.004	22	Male	No	Rarely	Yes
10	More likely	0.999	0.001	18	Female	No	Regularly	No
11	More likely	0.997	0.003	20	Female	No	Sometimes	No
12	More likely	0.999	0.001	25	Female	No	Rarely	No
13	More likely	1.000	0.000	20	Male	No	Sometimes	No

ExampleSet (27 examples, 3 special attributes, 12 regular attributes)

Figure 14 Test Prediction Using Naive Bayes



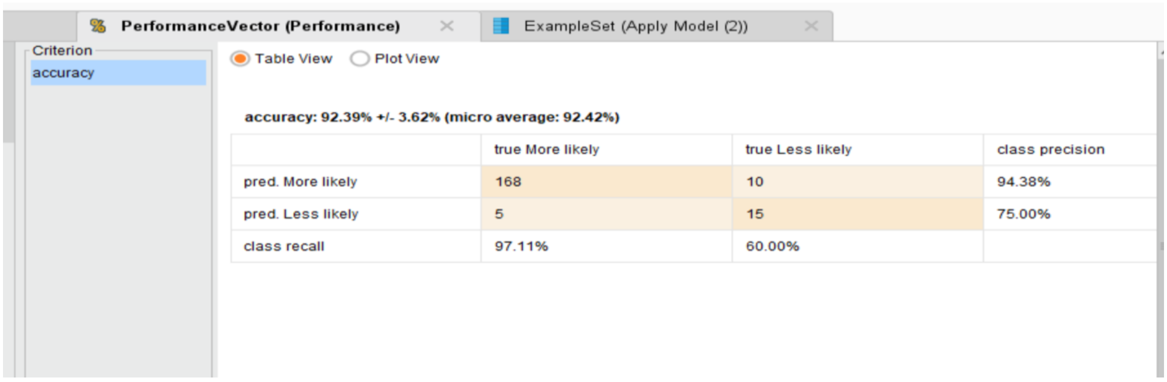
PerformanceVector (Performance) | ExampleSet (Apply Model (2))

Open in: Turbo Prep | Auto Model — Filter (27 / 27 examples): all

Row No.	prediction(li...	confidence(...	confidence(...	Age	Gender	Chronic Illne...	Exercise	Yoga/Med
14	More likely	0.999	0.001	45	Male	No	Sometimes	No
15	More likely	0.902	0.098	50	Female	Yes	Rarely	No
16	More likely	0.996	0.004	24	Female	No	Sometimes	No
17	More likely	0.998	0.002	65	Male	Yes	Sometimes	Yes
18	More likely	1.000	0.000	32	Female	No	Sometimes	No
19	More likely	0.998	0.002	45	Male	No	Sometimes	Yes
20	Less likely	0.327	0.673	22	Female	Yes	Regularly	Yes
21	More likely	1.000	0.000	21	Female	Yes	Rarely	No
22	More likely	1.000	0.000	43	Female	Yes	Regularly	No
23	Less likely	0.141	0.859	23	Male	No	Regularly	Yes
24	More likely	1.000	0.000	47	Female	Yes	Rarely	No
25	More likely	0.503	0.497	40	Male	Yes	Sometimes	Yes
26	More likely	1.000	0.000	25	Male	Yes	Rarely	No

ExampleSet (27 examples, 3 special attributes, 12 regular attributes)

Figure 15 Prediction Using Naive Bayes (2)



PerformanceVector (Performance) | ExampleSet (Apply Model (2))

Criterion: accuracy

Table View | Plot View

accuracy: 92.39% +/- 3.62% (micro average: 92.42%)

	true More likely	true Less likely	class precision
pred. More likely	168	10	94.38%
pred. Less likely	5	15	75.00%
class recall	97.11%	60.00%	

Figure 16 Performance table Using Naïve Bayes

9.1.2 KNN Model:

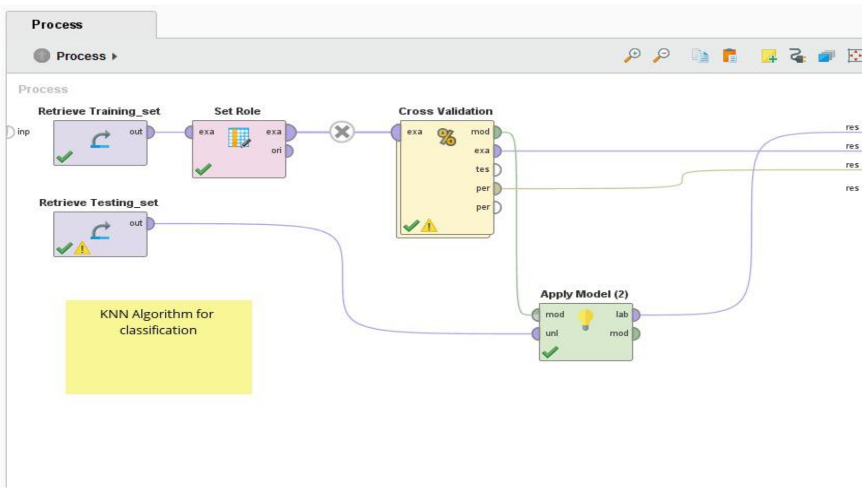


Figure 17 : K-NN Model Process

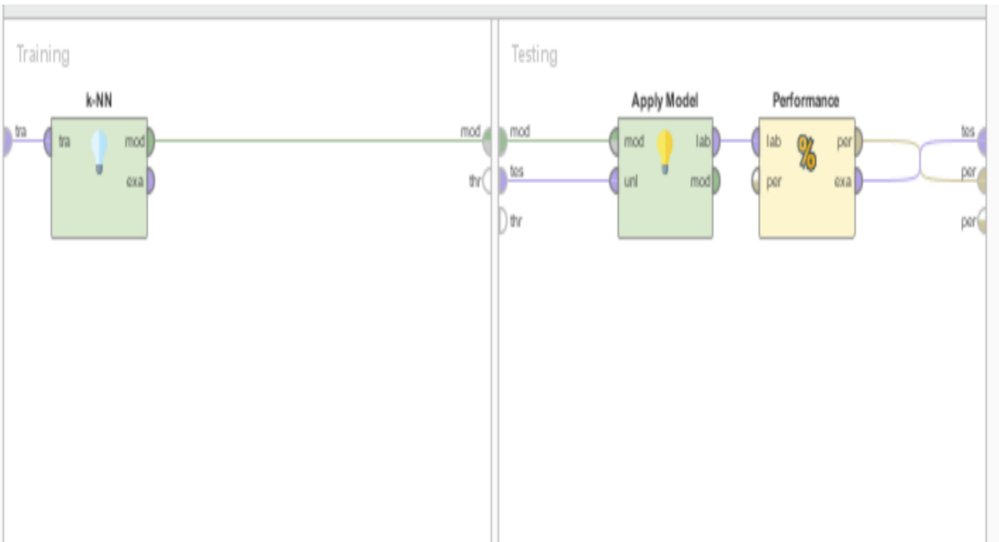
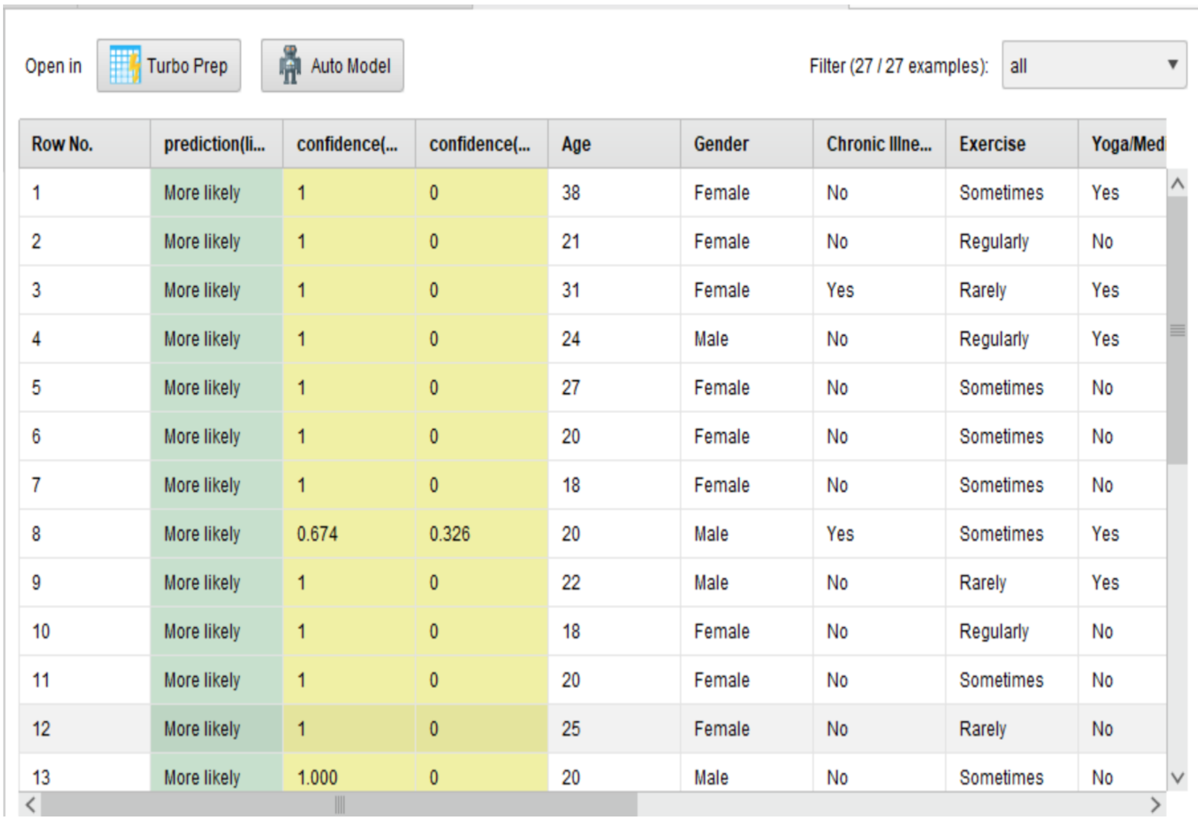


Figure 18 : K-NN Cross Validation Process

Results:



Row No.	prediction(li...	confidence(...	confidence(...	Age	Gender	Chronic Illne...	Exercise	Yoga/Med
1	More likely	1	0	38	Female	No	Sometimes	Yes
2	More likely	1	0	21	Female	No	Regularly	No
3	More likely	1	0	31	Female	Yes	Rarely	Yes
4	More likely	1	0	24	Male	No	Regularly	Yes
5	More likely	1	0	27	Female	No	Sometimes	No
6	More likely	1	0	20	Female	No	Sometimes	No
7	More likely	1	0	18	Female	No	Sometimes	No
8	More likely	0.674	0.326	20	Male	Yes	Sometimes	Yes
9	More likely	1	0	22	Male	No	Rarely	Yes
10	More likely	1	0	18	Female	No	Regularly	No
11	More likely	1	0	20	Female	No	Sometimes	No
12	More likely	1	0	25	Female	No	Rarely	No
13	More likely	1.000	0	20	Male	No	Sometimes	No

Figure 19 : Test Prediction Using K-NN

Open in Turbo Prep Auto Model Filter (27 / 27 examples): all

Row No.	prediction(li...	confidence(...	confidence(...	Age	Gender	Chronic Illne...	Exercise	Yoga/Med
15	More likely	1	0	50	Female	Yes	Rarely	No
16	More likely	0.651	0.349	24	Female	No	Sometimes	No
17	More likely	1	0	65	Male	Yes	Sometimes	Yes
18	More likely	1	0	32	Female	No	Sometimes	No
19	More likely	1	0	45	Male	No	Sometimes	Yes
20	More likely	0.645	0.355	22	Female	Yes	Regularly	Yes
21	More likely	1	0	21	Female	Yes	Rarely	No
22	More likely	1	0	43	Female	Yes	Regularly	No
23	Less likely	0.273	0.727	23	Male	No	Regularly	Yes
24	More likely	1	0	47	Female	Yes	Rarely	No
25	Less likely	0.291	0.709	40	Male	Yes	Sometimes	Yes
26	More likely	1	0	25	Male	Yes	Rarely	No
27	More likely	1	0	57	Male	Yes	Sometimes	No

Figure 20 : Test Prediction Using K-NN (2)

PerformanceVector (Performance) ExampleSet (Set Role) ExampleSet (Apply Model (2))

Criterion: accuracy

Table View Plot View

accuracy: 90.00% +/- 4.71% (micro average: 89.95%)

	true More likely	true Less likely	class precision
pred. More likely	167	13	92.78%
pred. Less likely	7	12	63.16%
class recall	95.98%	48.00%	

Figure 21 Performance table Using KNN

9.1.3 Random Forest Model:

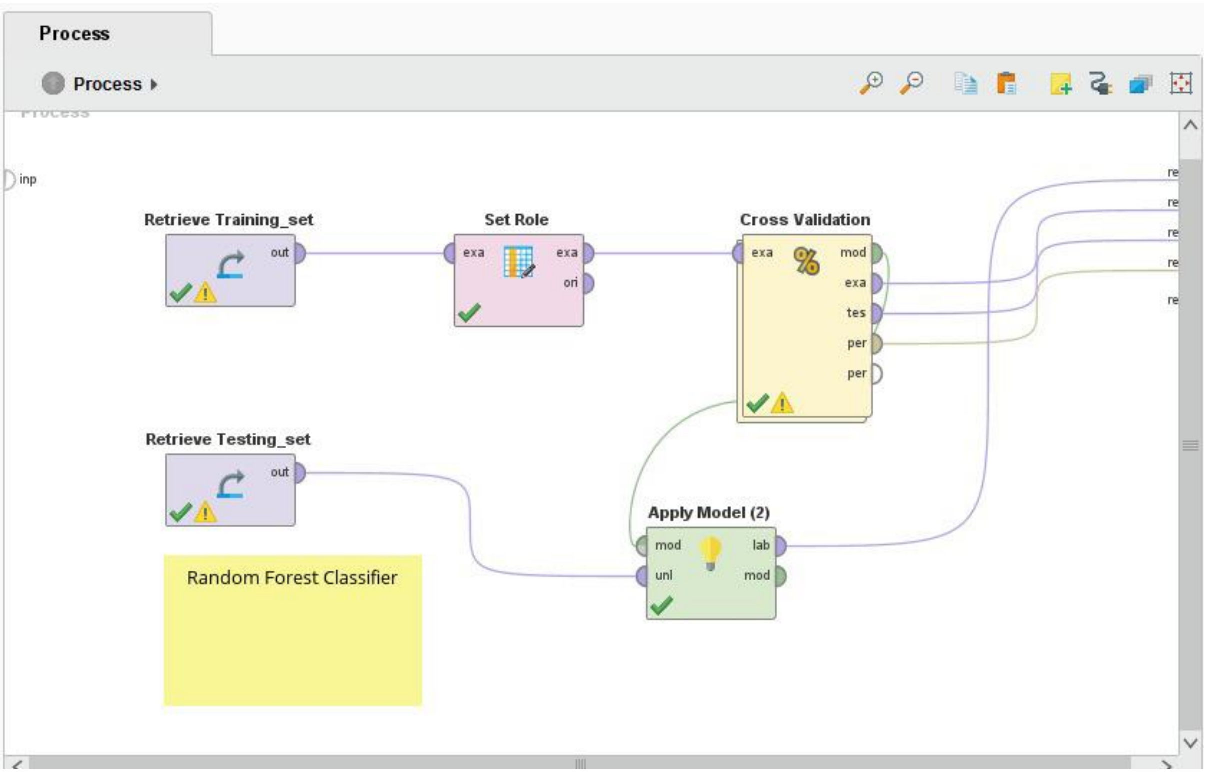


Figure 22: Random Forest Process

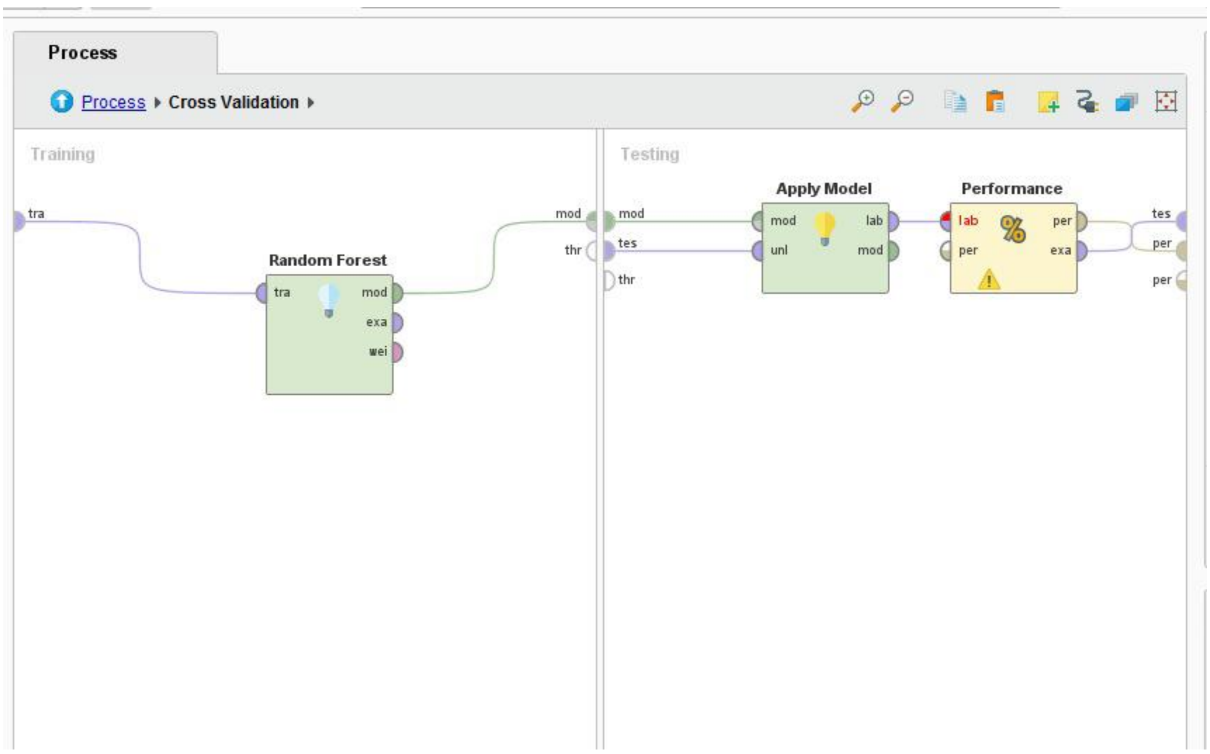


Figure 23 : Random Forest Cross Validation Process

Results:

Open in Turbo Prep Auto Model Filter (27 / 27 examples): all

Row No.	prediction(li...	confidence(...	confidence(...	Age	Gender	Chronic Illne...	Exercise	Yoga/Medita...	Supplements	Home reme
1	More likely	0.755	0.245	38	Female	No	Sometimes	Yes	No	Sometimes
2	More likely	0.977	0.023	21	Female	No	Regularly	No	No	Rarely
3	More likely	0.953	0.047	31	Female	Yes	Rarely	Yes	No	Sometimes
4	More likely	0.980	0.020	24	Male	No	Regularly	Yes	No	Rarely
5	More likely	0.990	0.010	27	Female	No	Sometimes	No	Yes	Sometimes
6	More likely	1	0	20	Female	No	Sometimes	No	No	Sometimes
7	More likely	0.940	0.060	18	Female	No	Sometimes	No	Yes	Rarely
8	More likely	0.920	0.080	20	Male	Yes	Sometimes	Yes	No	Regularly
9	More likely	0.957	0.043	22	Male	No	Rarely	Yes	No	Rarely
10	More likely	0.968	0.032	18	Female	No	Regularly	No	No	Sometimes
11	More likely	0.953	0.047	20	Female	No	Sometimes	No	No	Regularly
12	More likely	1	0	25	Female	No	Rarely	No	No	Sometimes
13	More likely	1	0	20	Male	No	Sometimes	No	No	Sometimes

Figure 24 : Test Prediction Using Random Forest

Row No.	prediction(li...	confidence(...	confidence(...	Age	Gender	Chronic Illne...	Exercise	Yoga/Medita...	Supplements	Home reme
15	More likely	0.856	0.144	50	Female	Yes	Rarely	No	Yes	Regularly
16	More likely	0.960	0.040	24	Female	No	Sometimes	No	Yes	Sometimes
17	More likely	0.941	0.059	65	Male	Yes	Sometimes	Yes	Yes	Sometimes
18	More likely	0.902	0.098	32	Female	No	Sometimes	No	Yes	Sometimes
19	More likely	0.993	0.007	45	Male	No	Sometimes	Yes	No	Sometimes
20	More likely	0.621	0.379	22	Female	Yes	Regularly	Yes	Yes	Regularly
21	More likely	1	0	21	Female	Yes	Rarely	No	No	Sometimes
22	More likely	0.978	0.022	43	Female	Yes	Regularly	No	Yes	Sometimes
23	Less likely	0.342	0.658	23	Male	No	Regularly	Yes	Yes	Sometimes
24	More likely	0.930	0.070	47	Female	Yes	Rarely	No	Yes	Sometimes
25	More likely	0.690	0.310	40	Male	Yes	Sometimes	Yes	Yes	Regularly
26	More likely	0.983	0.017	25	Male	Yes	Rarely	No	No	Rarely
27	More likely	0.968	0.032	57	Male	Yes	Sometimes	No	No	Regularly

Figure 25 : Test Prediction Using Random Forest (2)

Table View Plot View

accuracy: 92.50% +/- 5.40% (micro average: 92.46%)

	true More likely	true Less likely	class precision
pred. More likely	169	10	94.41%
pred. Less likely	5	15	75.00%
class recall	97.13%	60.00%	

Figure 26 : Performance table Using Random Forest

9.2 Visualization:

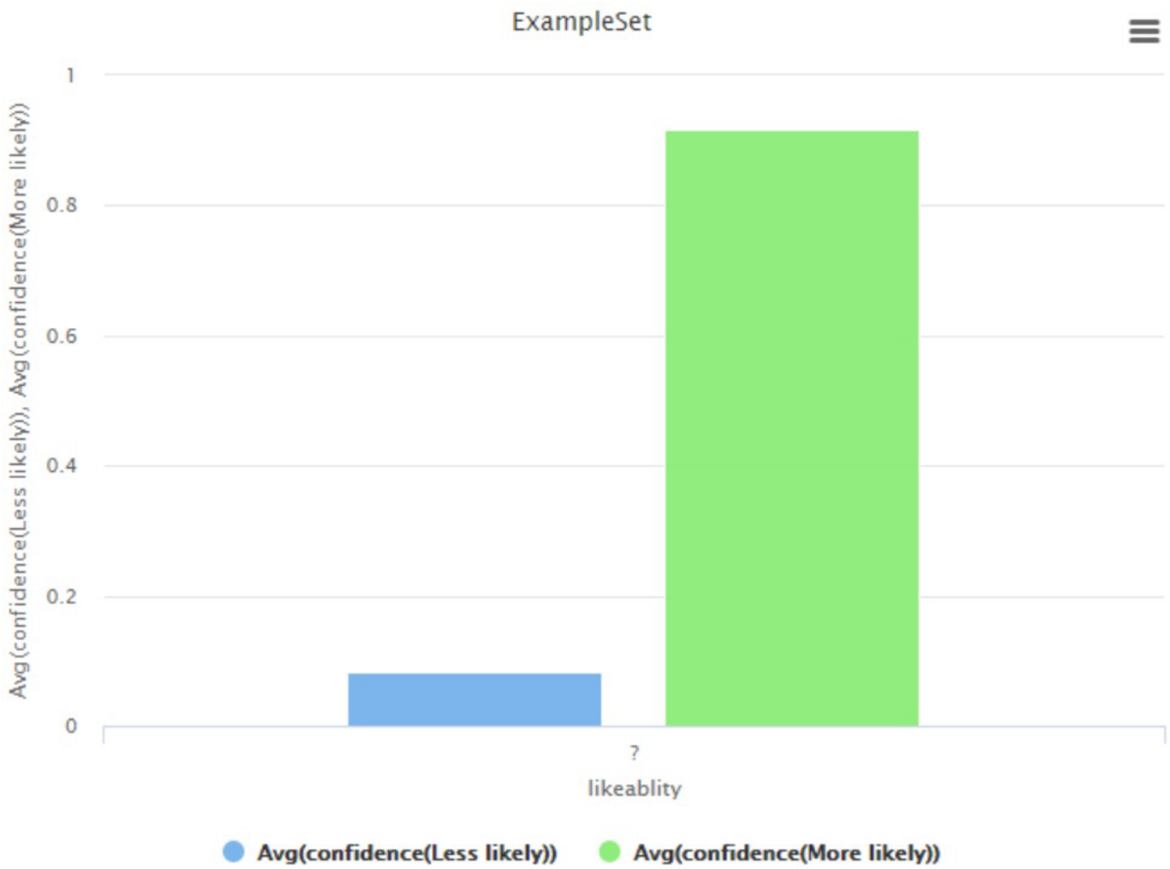


Figure 27 Data Visualization of the prediction arrived

9.3 Comparison Models

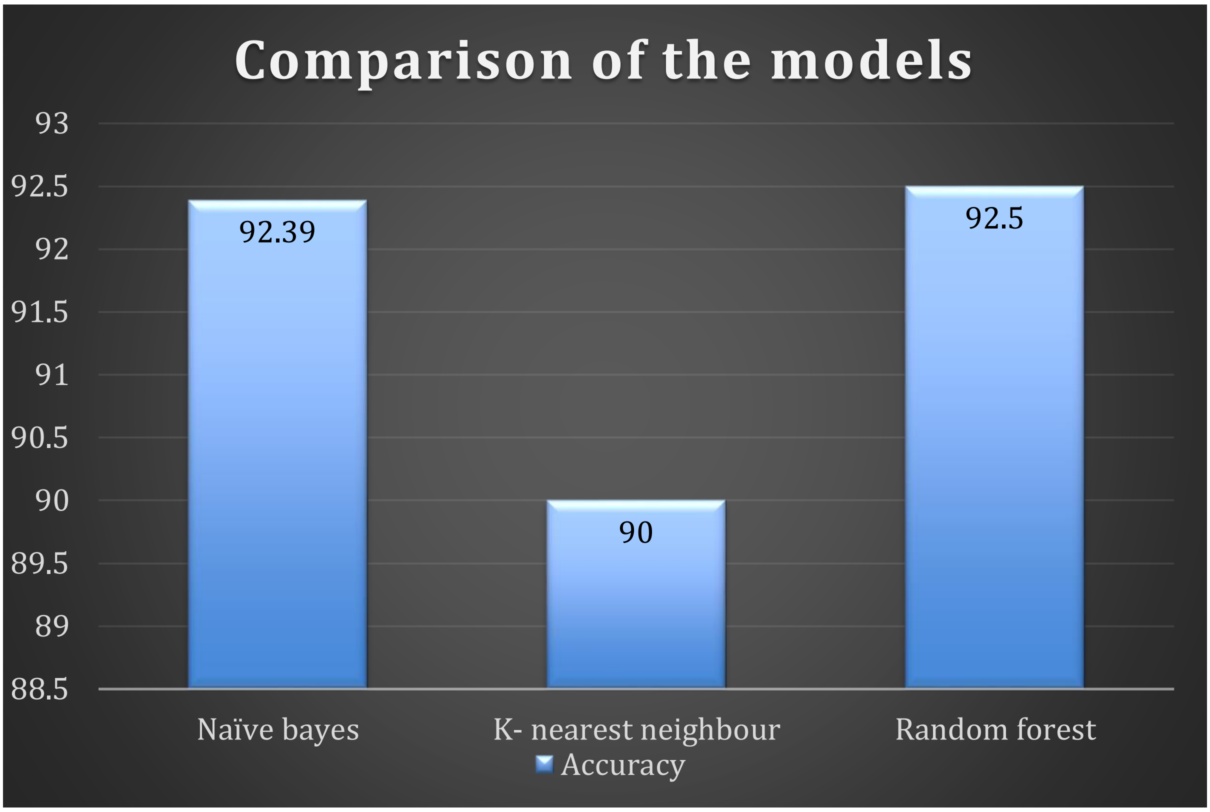


Figure 28 : Comparison of Models

10. Conclusion:

This Data Mining project was mainly focused on COVID 19 prediction. All the results and snapshots have good accuracy. Impact of Lifestyle and Diet in Combating SARS-2 COVID-19 in India. The world is experiencing an unprecedented challenge due to the corona virus disease (COVID-19) pandemic. Whether there is an association between lifestyle behaviors, diet and the acquisition of COVID-19 remains unclear. An online survey was conducted among the general population in India with distinguishing parameters ranging from physical exercise, diet intake etc., Data mining techniques and various predictive models (KNN, Decision Tree etc.,) were deployed in evaluating the result.

11. References:

- S. Tabik;A. Gómez-Ríos;J. L. Martín-Rodríguez;I. Sevillano-García;M. Rey-Area;D. Charte;E. Guirado;J. L. Suárez;J. Luengo;M. A. Valero-González;P. García-Villanova;E. Olmedo-Sánchez;F. Herrera. COVIDGR Dataset and COVID-SDNet Methodology for Predicting COVID-19 Based on Chest X-Ray Images. IEEE,2021.

- Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model. Vishan Kumar Gupta;Avdhesh Gupta;Dinesh Kumar;Anjali. IEEE,2021.

- Analysis and predictions of spread, recovery, and death caused by COVID-19 in India Rajani Kumari;Sandeep Kumar;Ramesh Chandra Poonia;Vijander Singh;Linesh Raja;Vaibhav Bhatnagar;Pankaj Agarwal. IEEE,2021.

- Explainable Machine Learning for Early Assessment of COVID-19 Risk Prediction in Emergency Departments. Elena Casiraghi;Dario Malchiodi;Gabriella Trucco. IEEE,2021.