Disease Prediction Using Machine Learning

Arkaprabha Kar

November 2021

Abstract

The report provides clarified details of the machine learning processes that produces a general outcome for data value limitation. The machine learning topic here works with the disease prediction and analysis based on specified classifier functions. The model data values are processed in accordance with the machine learning fundamentals. The increased usage of data processing and machine learning structures in the allotted topic gives a basic analytical overview. The data set and the visualization processes are correctly displayed in the report. The data analysis and the result discussion are also discussed thoroughly in the report.

1 INTRODUCTION

The report serves as a basic context for the machine learning outcomes in the coding sequence. The file decoding of the data sets is yet to be improved based on the correct value. The data set improvisation in relation to the provided values are connected and then implemented. Here the data set chosen for the machine learning process is applied and executed. The topic is disease prediction with the help of machine learning.

The topic explores the machine learning processes with the help of the major machine learning approaches. The machine learning outcomes that mainly process the data values in terms of different operations executed is of great importance. The value models obtained from the data set file execution are mainly categorical or non-categorical in nature. The procedures are to be implemented in accordance with the correct measured values for the data provided. The data values are to be clarified and checked according to machine learning algorithms like Random Forest and Naive Bayes.

The report discusses the machine learning tactics using the data prediction and analysis from the data set allocated. The methodology discusses the data operations in the machine learning process along with data analysis and results. The machine learning fundamentals in respect to the chosen topic is discussed thoroughly.

2 METHODOLOGY

The data values to be collected from the data set file is to be undergone in several operations in terms of the machine learning procedures. The disease prediction analysis and the data value calculation in machine learning improve the correct data value consistency. The data set file applied in machine learning is the disease prediction that consists of several data values in the manual headers.

The methodology process in the report consists of the data execution process using coordinate values such as X and Y. The data set files consist of data training and data testing. Here, the disease prediction values encounter the cleaning process and read and library importing process. The value declaration in the respective coordinate headers can be used to allocate the correct testing and training models. According to paper reviewing (1), every value present in the disease prediction data set file catches and records the underlying process accuracy through different methods mentioned in the report. The correct accuracy in the respective data value is set according to the working procedures in the analysis section in data prediction.

The methods applied in the correction and execution of the data value sets for the disease prediction outcomes are almost negligible. The data clearance and value processing follows the supervised and unsupervised learning for the data set control. Each and every data process is rechecked and evaluated to obtain the accurate model data value. Though, the data set file checks each record according to the operational command executed in the code interface.

The code declaration and the graphical data assurance is also part of the machine learning process that approves the data accuracy of the given file. According to resource researching (2), the data analysis catches the methods used in the data set prediction and model outcome generation with the help of certain operations. The analysis reflects the machine testing procedures in the correct format for data prediction.

3 DATA ANALYSIS

Data set

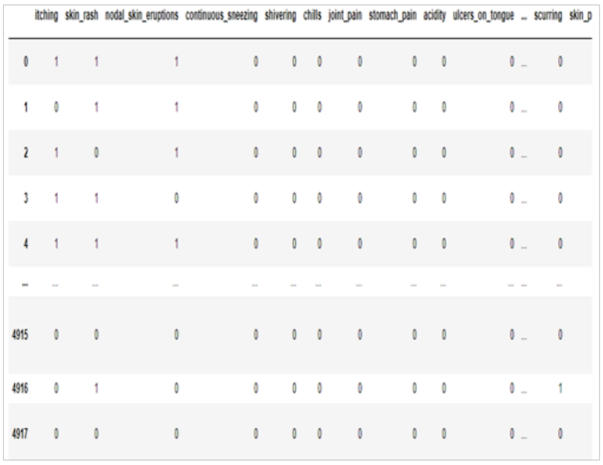
The data set file that is the disease prediction file is obtained. Then the data set is stored in the admin folder, where the python coding will be carried out. The data set values are imported and then washed out according to the variable segregation. The disease prediction data set file is collected from the Kaggle website. The training and the testing files are separate with different column headers. Both the test and train files have overall 133 columns. Out of 133 columns, 132 of these columns are symptoms that a person experiences and the last column is the prognosis disease.

Data pre-processing

The data values are created using the library import values for holding the data functions. The data set files consisting of the training and testing files are basically labeled in the code editor. Each of the data values cleared is then sent to the iteration check flow to manage the value rotation rate in disease prediction outcomes. The processing of each value considers the most recurring value input under the coordinate axes. According to data mined (3), the libraries used for importing in the Jupyter notebook uses pandas, sea born, Gauss Classifier and Random Forest Classifier. The libraries distinguish the classification methods by accessing the function from the resource data values.

Data processing

The data processing part is the vital segment in the data analysis because the predicted output of the disease-causing factors are to be operated using a round value fraction. The X and Y values are segregated according to the training and testing file operations conducted for value prediction. Generating a full-fledged calculation measure based on data set file reading, library importing, segregation and testing operations are properly followed.



	itching	skin_rash	nodal_skin_eruptions	continuous_sneezing	shivering	chills	joint_pain	stomach_pain	acidity	ulcers_on_tongue	...	scurring	skin_p
0	1	1	1	0	0	0	0	0	0	0	...	0	
1	0	1	1	0	0	0	0	0	0	0	...	0	
2	1	0	1	0	0	0	0	0	0	0	...	0	
3	1	1	0	0	0	0	0	0	0	0	...	0	
4	1	1	1	0	0	0	0	0	0	0	...	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	
4915	0	0	0	0	0	0	0	0	0	0	...	0	
4916	0	1	0	0	0	0	0	0	0	0	...	1	
4917	0	0	0	0	0	0	0	0	0	0	...	0	

Figure 1: **Data set framing in Jupyter notebook**
(Source: Created by the learner)

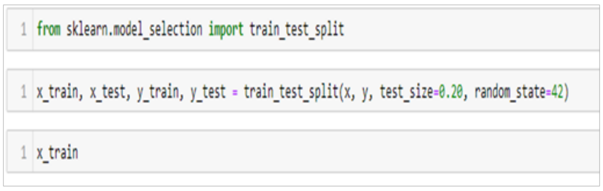
The figure mentioned above describes the data set framing based on the testing and training file operations. According to paper reviewed (4), the data prediction follows the random forest classification in the correct disease column calculation.

Methods

The methods followed while prediction generation in the machine learning process uses the Random Forest classification. The random forest classification is mainly used in the prediction value generation, but other operations are also used in the process. The operations are mainly Naive Bayes and SVM model prediction. The methods clearly follow the understanding of the machine learning structure in the data visualization and prediction. The supervised and unsupervised class types are mainly followed because the Random Forest classification is itself a supervised class type.

4 RESULT AND DISCUSSION

The random forest has been identified as one of the popular machine learning algorithms interrelated with the "supervised learning technique". Improvisation of the predictive accuracy has been done by this specified algorithm with the influence of taking the average of the provided data set. The efficiency regarding the training time also has brought better preference over this random forest algorithm in a specific manner.
As per reports reviewed (5), maintenance of accuracy during wide range data missing also can be retrieved in a decent way as per the impact of the random forest algorithm. The implementation of the test and train values are conducted by executing the data values segregation in terms of the machine language processing functions.



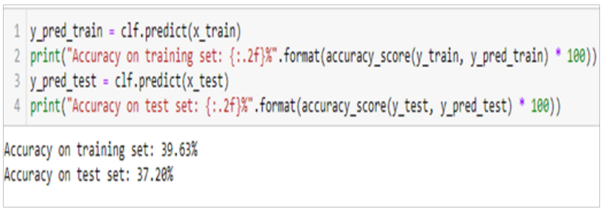
```
from sklearn.model_selection import train_test_split
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.20, random_state=42)
```

```
x_train
```

Figure 2: **Splitting of data values**
(Source: Created by the learner)

The figure mentioned above displays the train test splitting process for the X and Y data values in terms of testing and training formats. Several subsets of a given data set with a specific number of decision trees are generally classified with this specific classifier called "Random Forest". It also has been identified that greater numbers of trees can provide better level accuracy, and specific level barriers towards different problems related to over-fitting also can get resolved.



```
y_pred_train = clf.predict(x_train)
print("Accuracy on training set: {:.2f}%".format(accuracy_score(y_train, y_pred_train) * 100))
y_pred_test = clf.predict(x_test)
print("Accuracy on test set: {:.2f}%".format(accuracy_score(y_test, y_pred_test) * 100))
```

```
Accuracy on training set: 39.63%
Accuracy on test set: 37.20%
```

Figure 3: **Accuracy of the testing and training set values**
(Source: Created by the learner)

The figure mentioned above displays the accuracy values of the testing, and the training data set values in accordance with the random forest classification style. The standard value errors are dropped down, and then the null values are checked in the data set file. The value declaration in pursuit of the regression outcome declaration and value separation does not work equally as the train, and test values differ in the accuracy rate. As the value process is conducted through a random classifier method, the data coherency in the attributed value slows down while value gradation.
The result obtained through running various machine learning codes in the data set prediction encompasses the correct value analysis for the coordinate access. The machine learning merits includes such as data automaton process, time consumption and error identification pattern. The run time errors are hidden as the value predicted using the various model generators are assessed normally due to the train and test classifications for the inner and outer data. According to data resources (6), the data analysis is collected, and the final accuracy data values are calculated as per the operational functions mentioned.

5 CONCLUSION

It can be concluded from the above discussion that machine learning has brought an organized platform for the purpose of disease prediction. The utilization of Artificial intelligence also has been under study as per taking utmost advantage of machine learning with all elements of technological transformation. Enhancement of the healthcare infrastructure also can be taken in a constant interval with the impact of machine learning. Random forest classification also has emerged with sufficient outcomes against the implementation of machine learning in the region of disease prediction.

The python programming language with the provided data set has come up with relevant and effective outcomes in a decent way. Random forest algorithm with the programming language has brought effective scope for big data analysis of a wide number of patient's data sets. The classification method duly explores the vast analysis segment from the coordinate data procession. The value classification and the prediction of the generated outcomes in the disease prediction data file are continuous in nature. The machine learning evaluation in the data set processing can relate the fundamental concepts of the data machine activities. The usefulness of the machine learning and data processing activities are proving to be a resourceful structure in the upcoming future.

REFERENCES

[1] Mohan, S., Thirumalai, C. and Srivastava, G., 2019. Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, pp.81542-81554.

[2] Ramalingam, V.V., Dandapath, A. and Raja, M.K., 2018. Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering Technology, 7(2.8), pp.684-687.

[3] Wu, C.C., Yeh, W.C., Hsu, W.D., Islam, M.M., Nguyen, P.A.A., Poly, T.N., Wang, Y.C., Yang, H.C. and Li, Y.C.J., 2019. Prediction of fatty liver disease using machine learning algorithms. Computer methods and programs in biomedicine, 170, pp.23-29.

[4] Vinitha, S., Sweetlin, S., Vinusha, H. and Sajini, S., 2018. Disease prediction using machine learning over big data. Computer Science Engineering: An International Journal (CSEIJ), 8(1), pp.1-8.

[5] Khourdifi, Y. and Bahaj, M., 2019. Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization. International Journal of Intelligent Engineering and Systems, 12(1), pp.242-252.

[6] Battineni, G., Sagaro, G.G., Chinatalapudi, N. and Amenta, F., 2020. Applications of machine learning predictive models in the chronic disease diagnosis. Journal of personalized medicine, 10(2), p.21.