I.J. Mathematical Sciences and Computing, **2019, 1, 53-64**
Published Online January 2019 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijmsc.2019.01.05

Available online at ***http://www.mecs-press.net/ijmsc***

Machine Learning Applied to Cervical Cancer Data

Dhwaani Parikh [a], Vineet Menon [b]

RMIT University, 124 La Trobe St, Melbourne VIC 3000

Received: 11 July 2018; Accepted: 16 October 2018; Published: 08 January 2019

Abstract

Cervical Cancer is one of the main reason of deaths in countries having a low capita income. It becomes quite complicated while examining a patient on basis of the result obtained from various doctor's preferred test for any automated system to determine if the patient is positive with the cancer. There were 898 new cases of cervical cancer diagnosed in Australia in 2014. The risk of a woman being diagnosed by age 85 is 1 in 167. We will try to use machine learning algorithms and determine if the patient has cancer based on numerous factors available in the dataset. Predicting the presence of cervical cancer can help the diagnosis process to start at an earlier stage.

Index Terms: Cancer, Cervical cancer, Decision tree Classifier, herpes virus, K-nearest neighbor, Machine learning, Random forest.

1. Introduction

Cervical Cancer is cancer arising from the cervix. It arises due to the abnormal growth of cells and spreads to other parts of the body. It is fatal most of the time. HPV causes most of the cases (90 %). The data is cleaned, and outliers were taken care. Smoking is also considered as one of the main causes for cervical cancer. Long-term use of Oral contraceptive pills can also cause cancer. Also having multiple pregnancies can cause cervical cancer. Usually it is very difficult to identify cancer at early stages. The early stages of cancer are completely free of symptoms. It is only during the later stages of cancer that symptoms appear. We can use machine learning techniques to predict if a person has cancer or not.

Around the globe, around a quarter of million people die owing to cervical cancer. Screening and different deterministic tests confuse the available Computed Aided Diagnosis (CAD) to treat the patient correctly for the cancer. Several factors for cancer include age, number of sexual partners, age of first sexual intercourse,

* Corresponding author.
E-mail address: vineet.menon@outlook.com

number of pregnancies, smoking habits, hormonal, STD's detected in the patient and any diagnosis performed for cancer and diseases like HPV (Human papillomavirus) and CIN (Cervical intraepithelial neoplasia). For cervical cancer, the screening methods include cytology, Schiller, Hinselmann and the standard biopsy test. Each test is carried out to detect the presence of cervical cancer. Different models are tried on the dataset. Fine tuning of the parameters is done. The performance of Different models are compared. And finally, the best models are recommended that can be used to predict cancer[7,6].

2. Data Exploration

The data has 36 columns out of which 4 columns are the target columns of the four tests conducted which determine if the patient has cancer or not. The remaining 32 columns are the several factors that may lead to the cervical cancer. These factors play a significant role to identify the cancer in patient. We have made several data visualizations to check the instances of these factors.

After looking the data, we found several question mark signs (?) present in the columns. For modelling purpose, the question mark's need to be replaced with some meaningful values. We decided to replace all the question marks with median value of the column[11,14].

Target column: In the given dataset, there are four target columns present, which include the results from Hinselmann test, Schiller test, Cytology test and Biopsy test. All these columns have the values 0 or 1. We are combining the two possible outcomes into one column by adding the outcomes from each column for one patient. By doing this step, if all the tests are positive for the patient, maximum value in the target column will be 4.The target column will have values 0,1,2,3,4. We deleted the previous four target variables for deploying the machine learning algorithm on this one column with 5 possible outcomes. When we tried to implement an algorithm, the results were very bad, and model failed because presence of 0 values were high in the outcome.

For making better predictions we replaced the value 1,2,3,4 by 1 meaning that the patient has cancer and 0 meaning that the patient does not have cancer[15,8]

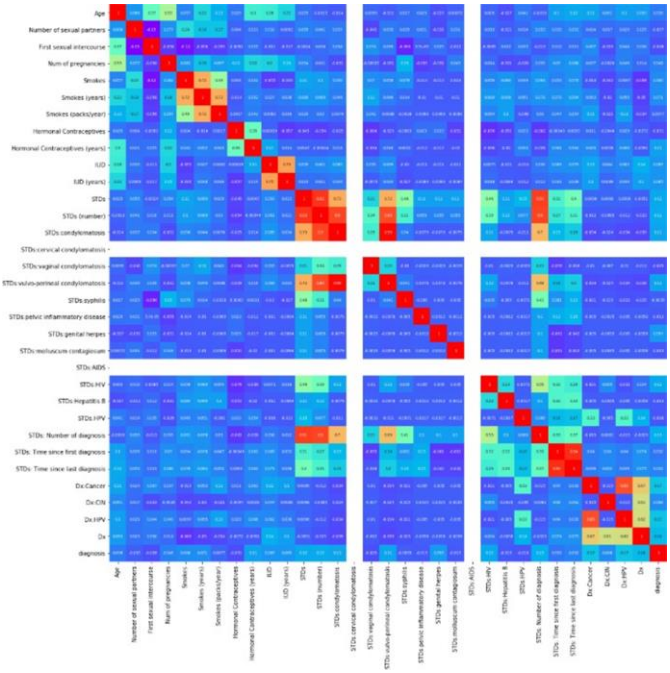

Fig.1. Correlation Plot of all the Columns.

The below heatmap, displays all the factors responsible for cervical cancer and their correlation with each other. The last column 'diagnosis' represents the target column. We have created this target column from the available four target columns for modelling purpose.[9,10]

3. Methodology

In order to build the model, three different classifier techniques are used. Decision tree, K-nearest neighbor and random forest. During the initial analysis it was found that the dataset is biased hence k-fold cross-validation is done. For k-nearest neighbor dataset is split 50-50. Feature selection is done, and the selected features are used for prediction. For decision tree and random forest, the data is split 25-75. Parameter tuning is done to get the best predictions with optimal evaluation scores.[3,5] We also used Hill climbing Algorithm for feature selection. The algorithm is a mathematical optimization that use heuristic search in the felid of Artificial Intelligence. Following are steps how the algorithm does a feature selection.

Step 1: Make initial state as current state after evaluating the initial state
Step 2: Run the loop till there are no features present which can be applied to current state.

a) Select a feature that has not been yet applied to the current state and apply it to produce a new state.
b) Perform these to evaluate new state

i. If the current state is a goal state, then stop and return success.
ii. If it is better than the current state, then make it current state and proceed further.
iii. If it is not better than the current state, then continue in the loop until a solution is found.

Step 3: Exit

Confusion matrix, classification error rate, Recall, F1score are used to evaluate different models. AUC curve is used to evaluate the model and is used to select the best model by parameter tuning.

3.1. K-Nearest Neighbor classifier

For the nearest neighbor K-fold cross validation is used. Ap- propriate k value is selected based on the formula sqrt(N)/2 which gives us 10.3. The sample size of the training dataset is chosen to be 429. After running several models the value of K is chosen as 5. Euclidean distance method is used to calculate the distance between two values. The result for the k-fold cross validation technique is shown below.

```
[fold 0] score: 0.8023255814
[fold 1] score: 0.9302325581
[fold 2] score: 0.8720930233
[fold 3] score: 0.9186046512
[fold 4] score: 0.9186046512
[fold 5] score: 0.9418604651
[fold 6] score: 0.9418604651
[fold 7] score: 0.8604651163
[fold 8] score: 0.8823529412
[fold 9] score: 0.9529411765
```

Fig.2. K-fold Cross Validation

Lower value of k is biased. Higher values of k is less biased but it can show variance. Since k=5 which is neither less nor more. Or the k-nearest neighbour the dataset is split 50-50. Feature selection is done which help us to select the features that will improve the prediction. The 25 features selected for the predictions are

Table 1. Features selected for K-Nearest Neighbor Classifier

Features selected for K-Nearest Neighbour Classifier	
Age	STDs (number)
Number of sexual partners	STDs:condylomatosis
First sexual intercourse	STDs:cervical condylomatosis
Num of pregnancies	STDs:vaginal condylomatosis
Smokes	STDs:vulvo-perineal condylomatosis
Smokes (years)	STDs:syphilis
Smokes (packs/year)	STDs:pelvic inflammatory disease
Hormonal Contraceptives	STDs:genital herpes
Hormonal Contraceptives (years)	STDs:molluscum contagiosum
IUD	STDs:AIDS
IUD (years)	STDs:HIV
STDs	STDs:Hepatitis B
	STDs:HPV

3.2. Decision tree Classifier

For the decision tree classifier, the dataset is split 25%. Presort function is used for fast implication of the algorithm. To minimize the size of the tree the minimum split for sample is set at 110. Class weight is used as balanced so that it automatically adjusts the weight inversely proportional to class frequencies in the input data. Feature selection is done for decision tree classifier. The following features are selected for the prediction.

Table 2. Features selected for Decision tree Classifier

Features selected for Decision tree Classifier	
STDs:AIDS	Dx:CIN
STDs:herpes	STDs:vaginalcondylomatosis
STDs:cervicalcondylomatosis	Dx
STDs:HPV	Dx:HPV
STDs: Time since first diagnosis	STDs:AIDS
Smokes	STDs:vulvo-perineal condylomatosis
First sexual intercourse	STDs:syphilis
IUD (years)	STDs:syphilis
	STDs:contagiosum

3.3. Random Forest Algorithm

The random forest algorithm uses the training data set and generates multiple level decision tree. For the

decision tree the data is split 25-75 for training and testing data. The depth of the tree is limited to 10 to make the tree less complex. After running the algorithm several times, the maximum sample split is decided to be 75. As per the inverse of frequency of input data class weight is again used as 'balanced' for automatic adjustment. For the random forest after we do the feature selection only 11 features are selected for prediction.

Table 2. Random Forest Algorithm

Features selected for Random Forest	
STDs:HPV	Smokes
STDs:molluscum contagiosum	First sexual intercourse
STDs: Time since first diagnosis	IUD (years)
IUD	Dx:CIN
Dx	STDs:vaginal condylomatosis
	Dx:HPV

4. Evaluation

4.1. Confusion Matrix

It is a performance measurement for machine learning classification problem where output can be two or more classes. It is a table with 4 different combinations of predicted and actual values.

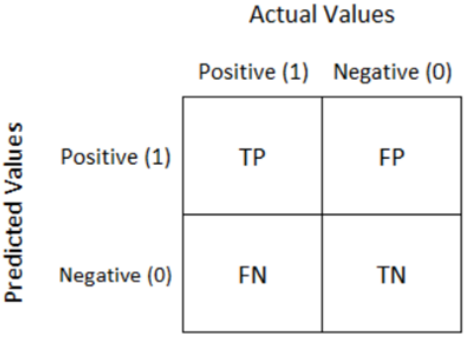

Fig.3. Confusion Matrix.

a) K-nearest neighbor classifier:

```
[[163   0]
 [  8   1]]
```

Fig.4. Confusion Matrix of K-nearest Neighbor Classifier.

According to the above confusion matrix, [1]

- True positive count is 163.
- False negative count is 0.
- False positive count is 8.
- True negative count is 1.

b) Decision tree classifier:

```
[[181  24]
 [  8   2]]
```

Fig.5. Confusion Matrix of Decision tree Classifier.

According to the above confusion matrix, [1]

- True positive count is 183.
- False negative count is 24.
- False positive count is 8.
- True negative count is 2.

c) Random Forest Algorithm:

```
[[187  18]
 [  8   2]]
```

Fig.6. Confusion Matrix Random Forest Algorithm.

According to the above confusion matrix, [1]

- True positive count is 187.
- False negative count is 18.
- False positive count is 8.
- True negative count is 2.

4.2. Accuracy

Accuracy is one of the metric used for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

Accuracy =(TP + TN)/ Total (1)

K-nearest neighbour: Accuracy of this algorithm is **95.3%.**
Decision Tree Classifier: Accuracy of this algorithm is **85.11%.**
Random-Forest: Accuracy of this algorithm is **87.90%**

4.3. Classification Error Rate

Classification error rate is one of the metric used for evaluating classification models. Informally, error rate is defined as predictions our model got it wrong.

CER = 1-accuracy (2)

K-nearest neighbour: The classification error rate is found out to be **4.7%.**
Decision Tree Classifier: The classification error rate is found out to be **14.89%.**
Random-Forest: The classification error rate is found out to be **12.1%**

4.4. Precision, Recall and F1 Score

Out of all the classes, how much we predicted correctly. It should be high as possible. it is difficult to compare two models with low precision and high recall or vice versa. So to make them comparable, we use F-Score. F-score helps to measure Recall and Precision at the same time. It uses Harmonic Mean in place of Arithmetic Mean by punishing the extreme values more[1,2]

Precision = TP/TP+FP (3)

Recall = TP/TP+FN (4)

F1 Score = 2*(Recall * Precision) / (Recall + Precision) (5)

a) K-nearest Neighbor:

	precision	recall	f1-score	support
0	0.95	1.00	0.98	163
1	1.00	0.11	0.20	9
avg / total	0.96	0.95	0.94	172

Fig.7. Precision Recall and F1 Score of K-nearest Neighbor Classifier.

b) Decision tree Classifier:

	precision	recall	f1-score	support
0	0.96	0.88	0.92	205
1	0.08	0.20	0.11	10
avg / total	0.92	0.85	0.88	215

Fig.8. Precision Recall and F1 Score of Decision tree Classifier.

c) Random Forest Algorithm:

	precision	recall	f1-score	support
0	0.96	0.91	0.94	205
1	0.10	0.20	0.13	10
avg / total	0.92	0.88	0.90	215

Fig.9. Precision Recall and F1 Score of Random Forest.

4.5. AUC/ROC

Area Under Curve(AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem[3,5]. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example.

a) K-nearest neighbor:

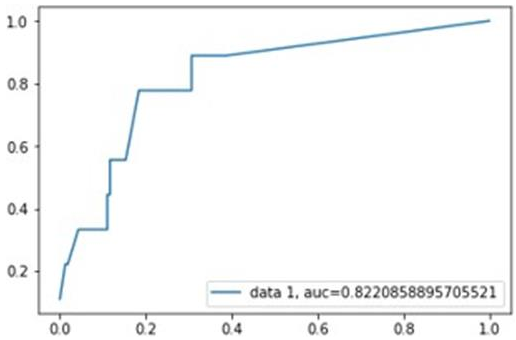

Fig.10. plot of ACU K-nearest Neighbor Classifier.

The figure shows the AUC chart and the AUC value is good which is 0.8220. This model can be used for prediction

b) Decision tree classifier:

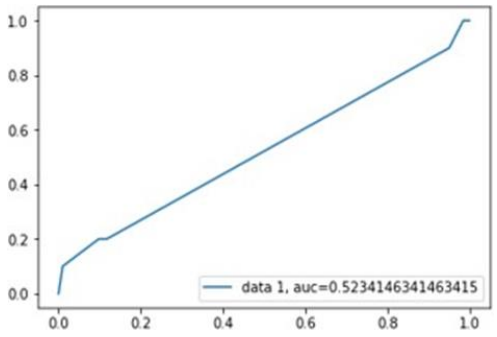

Fig.11. plot of ACU Decision tree Classifier.

The AUC value for Decision tree is 0.52. It is very less compared to K-nearest neighbor method. But this AUC value is acceptable since it is more than 0.5.

c) Random Forest Algorithm:

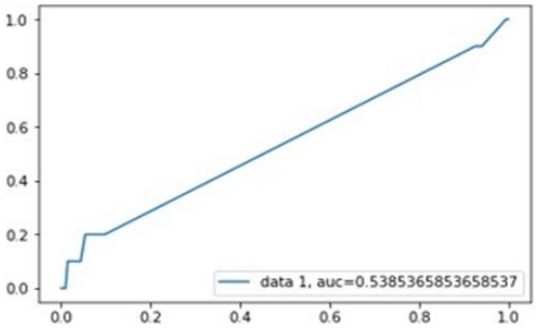

Fig.12. plot of ACU Random forest.

The AUC curve for random forest is similar as decision tree with an AUC value of 0.538536

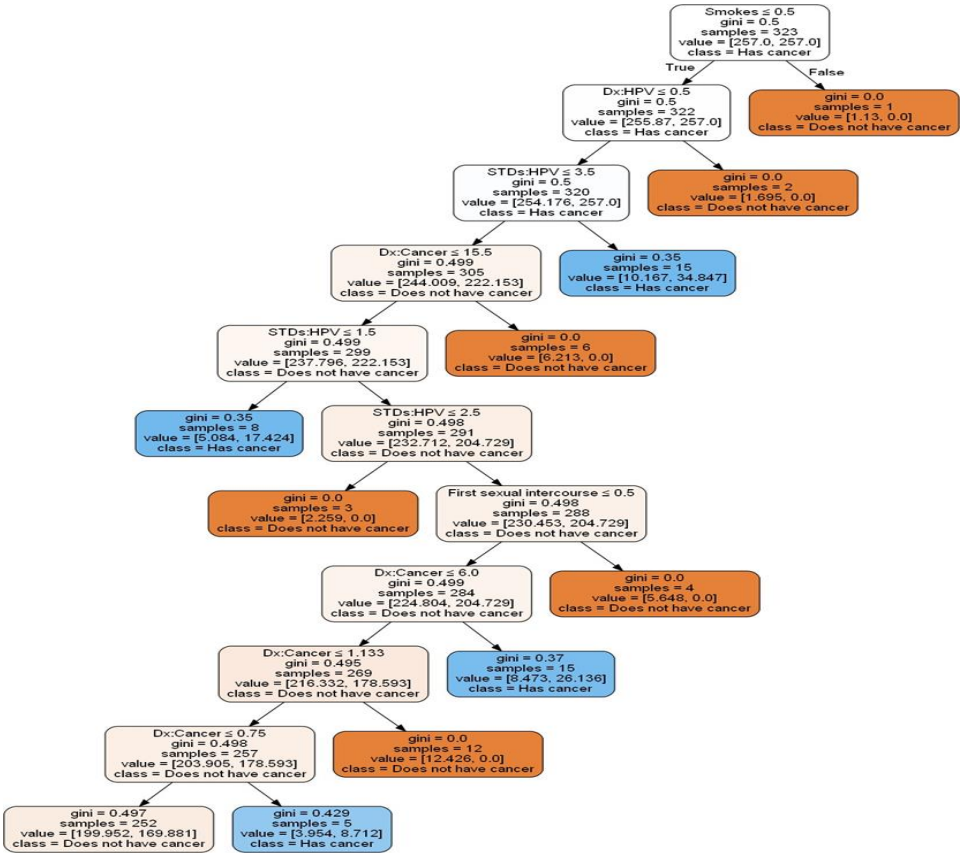

Fig.13. Random Forest tree for Prediction.

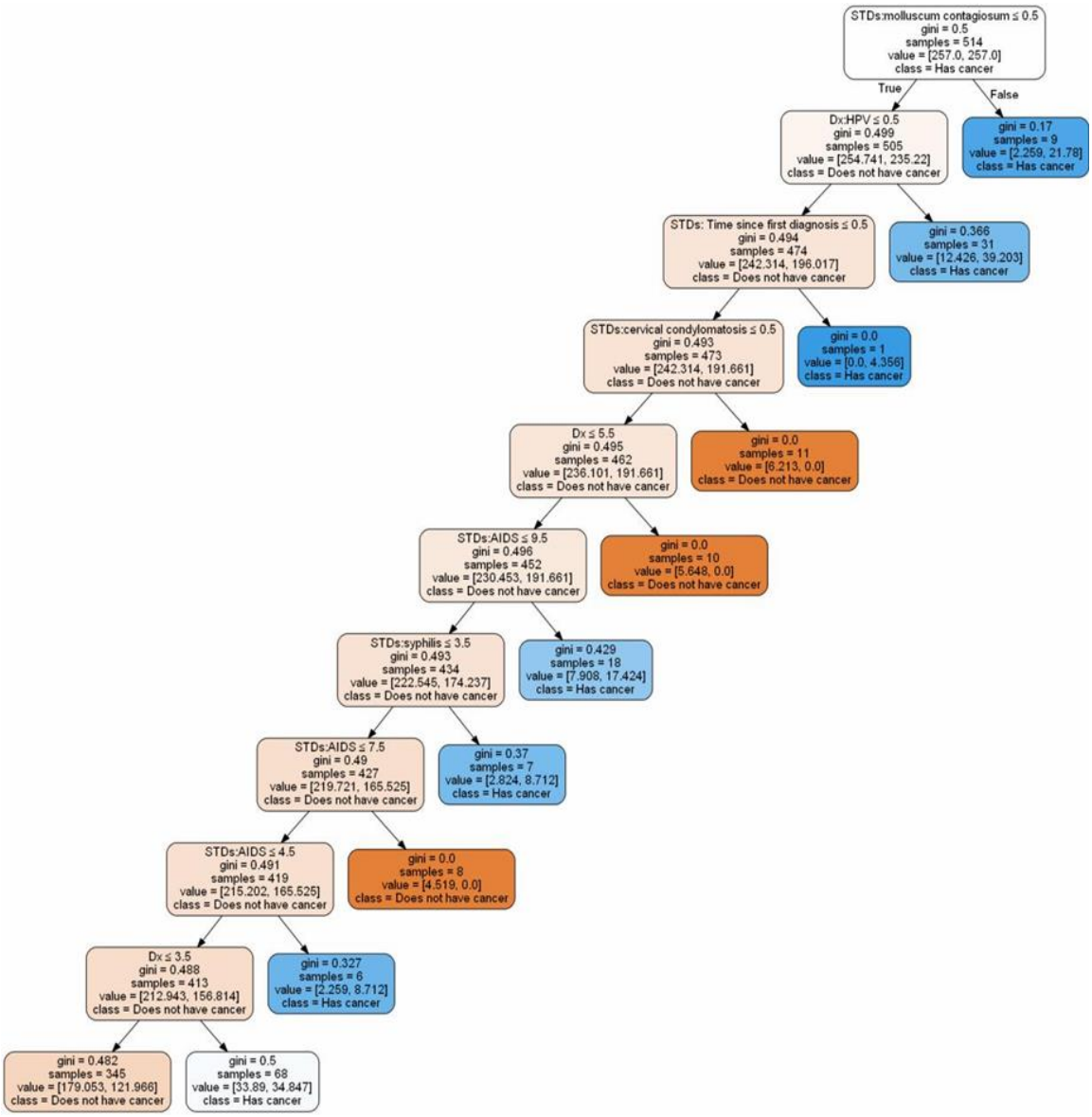

Fig.14. Decision tree Classifier tree for Prediction.

5. Discussion

After completing the analysis, it is found that all the models are good, that is k-nearest neighbor, decision tree and random forest. K-nearest neighbor seems to be the best model with higher accuracy of the model, Higher AUC which is 0.822 as compared 0.52(Decision tree) and 0.532(random forest) which is very low. Precision recall and f1 score is also high for nearest-neighbor model. The f1 score for nearest-neighbor is 0.94 which is high when compared to decision tree (0.88) and random forest (0.90).

From the confusion matrix we can also note that false negative is zero which means that a patient with

cancer will have pre- diction that he does not have cancer. It will be a bad scenario when a person having cancer will be informed that he does not have cancer. By the time symptoms starts showing it could be too late. So, it is important to choose a model that has very low false negative rates in some cases such as ours. Also, the k-nearest neighbor model has the best accuracy and AUC value which is one of the strengths of the model. So, k-nearest neighbor will be used for prediction. The dataset is biased it has a lot of zero values i.e. the patient does not have cancer for the target variable. If the dataset was little less-biased, we would have got better models and more accurate predictions. Technique used for feature selection has some disadvantages. It is difficult to know if the hill found is the highest possible hill. The local maxima are a state that is best for its neighboring states, but it might not be better than the states further away.

6. Conclusion

- From the initial analysis it was clear that the data is biased
- In order to address the problem necessary measures were taken while modelling.
- Fine tuning is done on all three models to get the best accuracy.
- The best 3 models from each of them is selected and performance is compared
- It is found that all 3 models are good.
- But the k-nearest-neighbour model has better accuracy, precision, recall and better AUC value.
- The research also showed that herpes virus was able to fight cancer cells. This observation was made based on the available data and more scientific analysis needs to be carried out in order to verify this findings.

References

[1] R), I & R), I 2018, "Improve Your Model Perfor- mance using Cross Validation (in Python / R)", Analytics Vidhya, viewed 14 May, 2018, https://www.analyticsvidhya.com/blog/2018

[2] Anon 2018, "Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures - Exsilio Blog", Exsilio Blog

[3] Anon 2018, S3.amazonaws.com, viewed 26May,2018, https://s3.amazonaws.com/MLMastery/MachineLearningAlgo-rithms.png? =8cjo2uzaybyqe5a1fipt

[4] Anon 2018, " Performance Measures - Exsilio Blog", Rmt Blog, viewed 26 May, 2018,<https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>.

[5] Anon 2018, " Google ML", Blog, viewed 26 May, 2018,< https://www.guru99.com/r-generalized-linear-model.html>.

[6] Mishra, A. (2018, February 24). Metrics to Evaluate your Machine Learning Algorithm. Retrieved from https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

[7] Narkhede, Sarang. "Understanding Confusion Matrix – Towards Data Science." Towards Data Science, Towards Data Science, 9 Feb. 2018, towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62.

[8] Australia, Cancer Council. "Cervical Cancer." CancerCouncilAustralia, www.cancer.org.au/about-cancer/types-of-cancer/cervical-cancer.html.

[9] Benjamin, S.M., Humberto, V.H., Barry, C.A. Use of Dagum Distribution for Modeling Tropospheric Ozone levels. J. Environ. Stat. 2013, 5(5).

[10] Dagum, C. A New Model of Personal Income Distribution: Specification and Estimation. Economie Appliqu ́ee ; 1977, 30,413–437.

[11] Domma, F., Giordano S., Zenga, M. The Fisher information matrix on a type II doubly censored sample from a Dagum distribution. Appl. Mathe. Sci.2013,7, 3715–3729.

[12] Feigl, P. and Zelen, M. Estimation of exponential probabilities with concomitant information. Biometrics. 1965, 21, 826–838.
[13] Proschan, F. Theoretical explanation of observed decreasing failure rate. Technometrics.1963, 5, 375–383.
[14] Naqash, S., Ahmad, S.P., Ahmed, A. Bayesian Analysis of Dagum Distribution. JRSS.2017, 10 (1), 123- 136.
[15] Kleiber, C., Kotz, S. Statistical Size Distributions in Economics and Actuarial Sciences. New York: Wiley. Meth and Appl. 2003, 18, 205–220.

Authors' Profiles

Dhwaani Parikh Interned at S Kant Healthcare limited under the R&D department. Currently Working at Four Care Hospital has an Analyst. Has a Bachelor's degree in Pharmacy. Always on a look out of Technology that can assist in the medical field. She considers Joseph L. Goldstein has her idol.

Vineet Menon Pursuing his Masters in Analytics. Has a Bachelor's degree in Electronics and telecommunications. Technology enthusiast. Steve Jobs is his idol. Impressed by the work done by Google in regards Digital well-being.

How to cite this paper: Dhwaani Parikh, Vineet Menon,"Machine Learning Applied to Cervical Cancer Data", International Journal of Mathematical Sciences and Computing(IJMSC), Vol.5, No.1, pp.53-64, 2019.DOI: 10.5815/ijmsc.2019.01.05