iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** VI **Month of publication:** June 2021

DOI: https://doi.org/10.22214/ijraset.2021.35882

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue VI Jun 2021- Available at www.ijraset.com

The Machine Learning Algorithm for Prediction of Risk Factors of Cervical Cancer

Mr. Mahesh Patil[1], Dr. Mrs. S.V. Deshmukh[2]
[1]*Student,* [2] *MCA-SEM VI, Faculty Research Guide, Department of Computer Application,YM Institute of Management Karad.*

Abstract: ***Cervical cancer is caused by the Human Papilloma Virus (HPV), the most common infection of the reproductive tract. Almost all cervical cancer cases (99 %) are linked to infection with high-risk human papillomaviruses. The peak time for infection is shortly after becoming sexually active, and most individuals with healthy immune systems will clear the virus within a few years. Almost all sexually active individuals will become infected with HPV at some point in their lives and some may repeatedly be infected. Cervical cancer is the fourth most commonly occurring cancer in women and the eighth most commonly occurring cancer overall. In 2018, an estimated 5,70,000 womens were diagnosed with cervical cancer worldwide and about 3,11,000 womens died from the disease. In India, cervical cancer contributes to approximately 6-29% of all cancers in women***

Keywords: Cervical Cancer, Machine Learning, KNN, Random Forest, Decision Tree, Logistics Regression

I. INTRODUCTION

Cervical cancer is one of the most preventable and successfully treatable forms of cancer. UICC supports the draft WHO Global Strategy towards the elimination of cervical cancer. Fighting cancer together. Cervical cancer is a type of cancer that occurs in the cells of the cervix — the lower part of the uterus that connects to the vagina. Various strains of the human papillomavirus (HPV), a sexually transmitted infection, play a role in causing most cervical cancer. If not diagnosed and treated, cervical cancer can spread to other parts of the body and become deadly. Cervical cancer is often curable if it's diagnosed at an early stage. When cervical cancer is not curable, it's often possible to slow its progression, prolong lifespan and relieve any associated symptoms, such as pain and vaginal bleeding. This is known as palliative care. The machine learning methods have been adopted in many areas of medical science. However researchers rae always looking for ways to optimize and improve these methods . To effectively evaluate the performance of the proposed method , the Risk factors of cervical cancer data set from kaggle is used. A comparative study is then conducted with some recent scholarly works and other well known machine algorithm including K-nearest neighbor (KNN), logistic regression(LR), Decision tree, Random forest.

II. MACHINE LEARNING APPROACH

A. Model Selection

This is the most exciting phase in Applying Machine Learning to any Dataset. It is also known as Algorithm selection for Predicting the best results. Usually Data Scientists use different kinds of Machine Learning algorithms to the large data sets. But, at high level all those different algorithms can be classified in two groups: supervised learning and unsupervised learning. Without wasting much time, I would just give a brief overview about these two types of learnings. Supervised learning: Supervised learning is a type of system in which both input and desired output data are provided. Input and output data are labelled for classification to provide a learning basis for future data processing. Supervised learning problems can be further grouped into Regression and Classification problems.

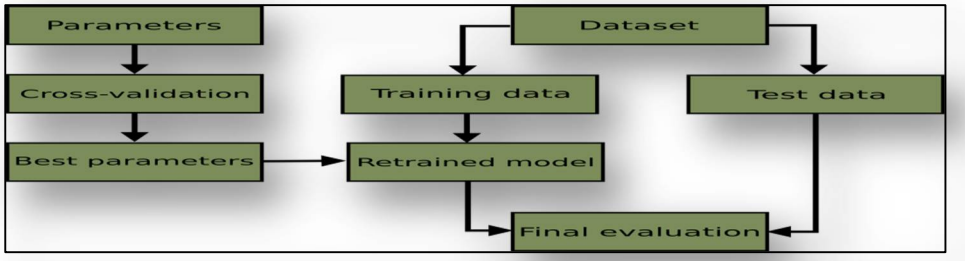
B. Confusion Matrix

A confusion matrix contains information of actual and predicted classification done by a classification model. The performance of such system is commonly evaluated using the data in the matrix.

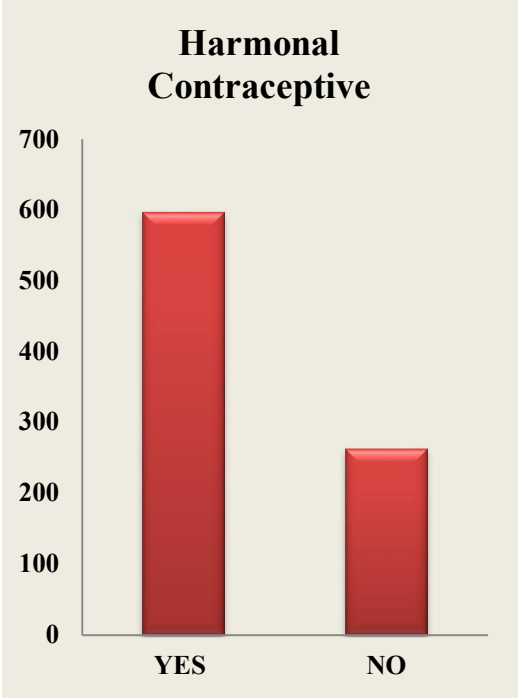
C. Measures of Performance Evaluation

1) Accuracy = (TP+TN)/(TN+TP+FN+FP) i.e. It is the % of correct classification by the classifier.
2) Recall = TP/(TP+FN) i.e. Proportion of correct positive classification (True positives) from cases that are actually positive.
3) Precision = TP/(TP+FP) i.e. Proportion of correct positive classification (True positive) from cases that are actually positive.
4) F1-Score = 2TP/(2TP+FP+FN) i.e. F1 Score is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account.

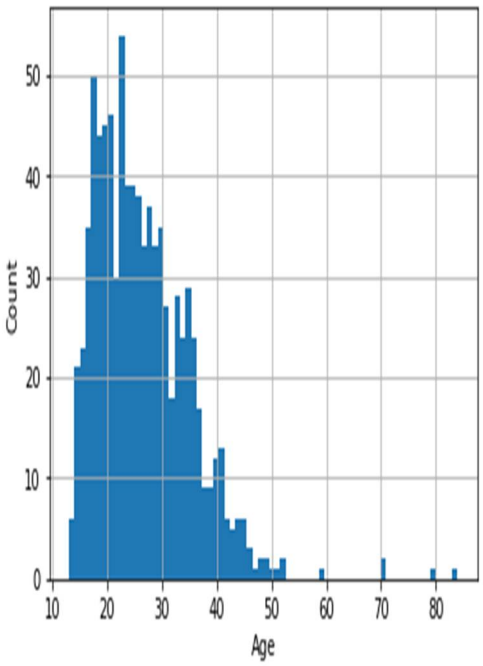
The methodology and performance of every model is as follows:



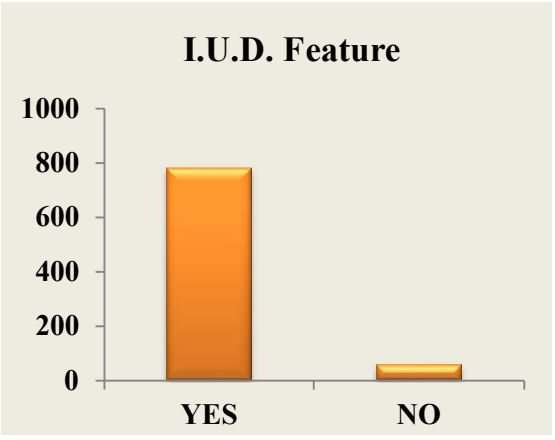
III. EXPLORATORY DATA ANALYSIS

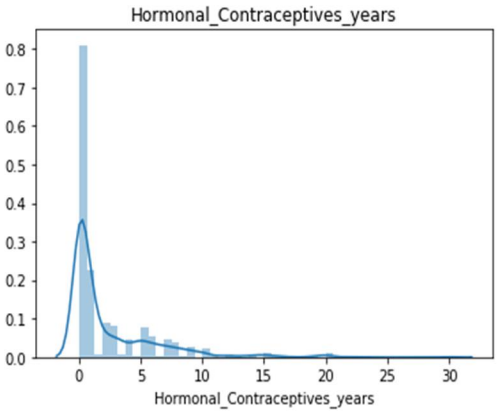


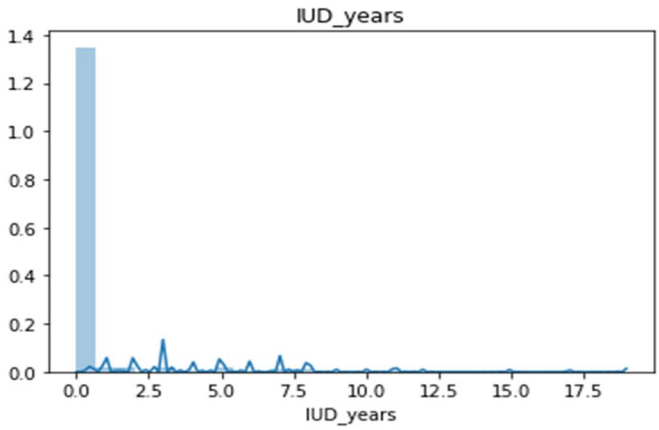
Hormonal Contraceptive Habit



Mean Age Of women facing cervical cancer

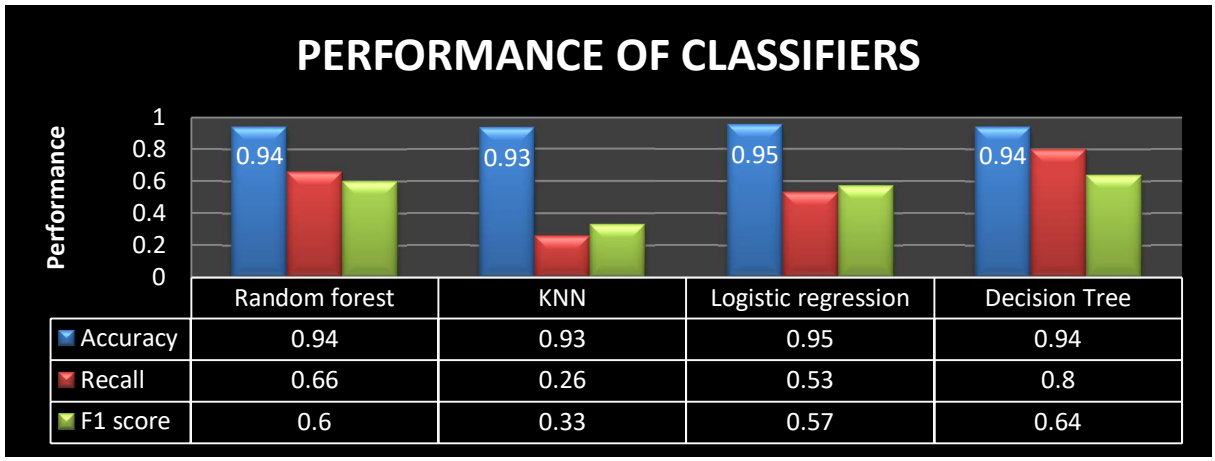






The Performance of the above Classifier is as follows:

Performance / Classifier measure	Accuracy	Recall	F1-Score
KNN	0.93	0.26	0.33
Decision tree	0.94	0.80	0.64
Logistic Regression	0.95	0.53	0.57
Random Forest	0.94	0.66	0.60



IV. CONCLUSION

A. Mean age of women facing the cervical cancer 26.

B. Most of the patients have used Hormonal contraceptives methods like pills and medications for birth controls where only a few of them have opted for intrauterine devices (IUDs). The reason for this is may be that hormonal contraceptives are readily available in shops (needs prescription) and one can take those at their home on their own with some sort of guidance where as IUD needs an doctor supervision and the patient needs to be in hospital.

C. Generally most patients have used birth control methods only for less than 2 years while very few of them have used more than 2 years.

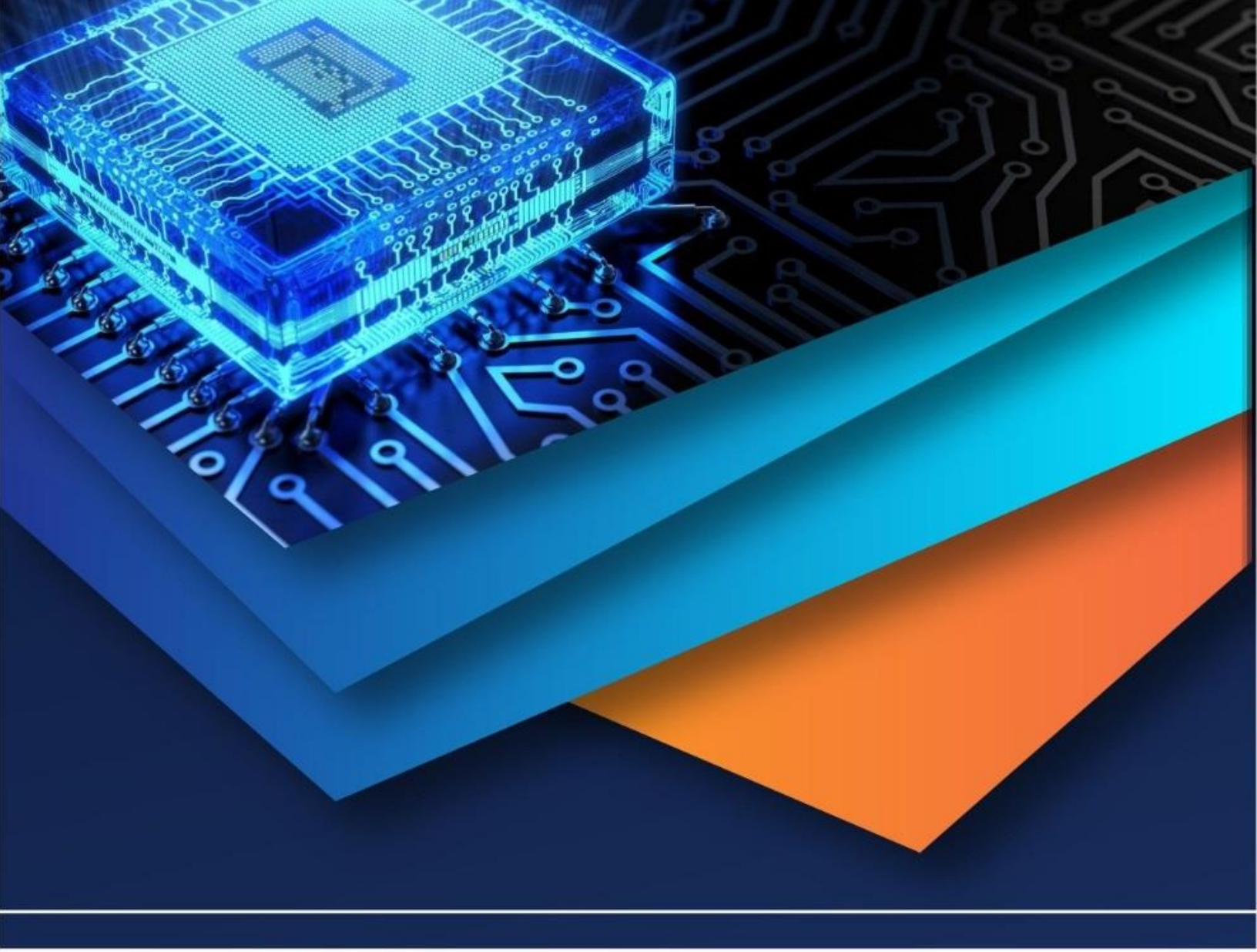
D. Specifically as this is an sensitive medical data, recall score needs to be given higher importance and hence we are choosing both "Decision Tree" and "Random Forest" models as our base model because of their higher recall and roc_auc scores.

E. Why recall should be given higher importance is that we have to predict actual cancer patients as cancer patient accurately.

F. Predicting a cancer patient as a healthy (non-cancer) is very dangerous and if predicted wrongly it may cause chaos to the life of a patient.

REFERENCES

[1] The data set used for study is : https://www.kaggle.com/loveall/cervical-cancer-risk-classification
[2] https://christophm.github.io/interpretable-ml-book/references.html
[3] https://www.merriam-webster.com/dictionary/algorithm.
[4] Aamodt, Agnar, and Enric Plaza. "Case-based reasoning: Foundational issues, methodological variations, and system approaches." AI communications 7.1 (1994): 39-59.
[5] Breiman, Leo.“Random Forests.” Machine Learning 45 (1). Springer: 5-32 (2001).



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY