Cervical Cancer Risk Classification

A.P.K.C. Prabodhani
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
chandi.20@cse.mrt.ac.lk

W.A.H. Ranasingha
Department of Computer Science and Engineering
University of Moratuwa
Moratuwa, Sri Lanka
hasini.20@cse.mrt.ac.lk

Abstract— Cervical cancer is the fourth most common ma-lignant disease in women worldwide. In most cases, cervical cancer symptoms are not noticeable at its early stages. There Are a lot of factors that increase the risk of developing cervical cancer like human papillomavirus, sexually transmitted diseases, and smoking.Identifying those factors and building a classification model to classify whether the cases are cervical cancer or not is a challenging research. This study aims at using cervical cancer risk factors to build classification models using five classifiers with the synthetic minority oversampling technique (SMOTE) and two feature reduction techniques, recursive feature elimination.

INTRODUCTION

Cervical cancer is the most common cancer among women in developing countries. Cervical cancer data has been studied from different researchers in the last few years. The percentage of cervical cancer cases in developing countries is 80%. Each year cervical cancer kills about 300,000 women worldwide. AGE Fifty percent of cervical cancer diagnoses occur in women ages 35 - 54, and about 20% occur in women over 65 years of age. The median age of diagnosis is 48 years. About 15% of women develop cervical cancer between the ages of 20 - 30. Cervical cancer is extremely rare in women younger than age 20. [2]

Machine learning techniques are widely used to solve real world problems.[1] They play an important role in the medical field and disease diagnosis. There are various techniques used to classify cervical cases. In this paper, we apply five classifiers such as GNB, KNN, DT, LR and RF. Then select what are the suitable classifiers for analysis of this dataset based on accuracy and recall score. After that achieve some optimization techniques to improve the performance of the model. Performance measure using mainly recall score because it is most important to predict as a cancer patient correctly.

In this paper organized as follows. There are three main steps in the data mining, preprocessing, classification process and the decision-making with analysis classification process and the decision-making with analysis

DATA DESCRIPTION

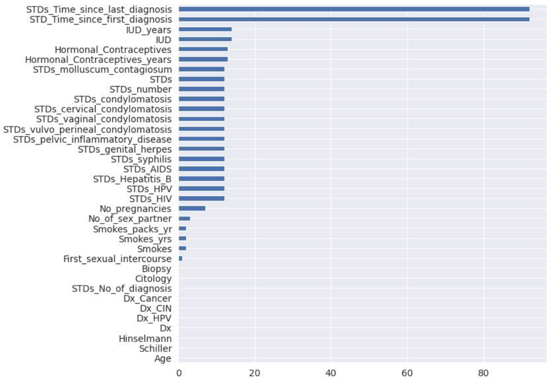
The dataset comprises demographic information, habits, and historic medical records of 858 patients. Several patients decided not to answer some of the questions because of privacy concerns (missing values). This dataset focuses on the prediction of indicators/diagnosis of cervical cancer. The features cover demographic information, habits, and historic medical records

Cervical cancer data involves 858 samples and 36 features as well as four classes (Hinselmann, Schiller, Cytology And Biopsy) has been published in . In this dataset includes 3622 missing values. This paper focuses on studying Understanding the cancer risk factors leads to Biopsy and how much they are affected by cervical cancer.

METHODOLOGY

Handling Missing Values

This section concentrates on the methodology of this survey, which can be described into three main parts. First, the preprocessing experiments, this involves missing values treatment using standard measurements like the mean for numerical values and mode for categorical attributes.



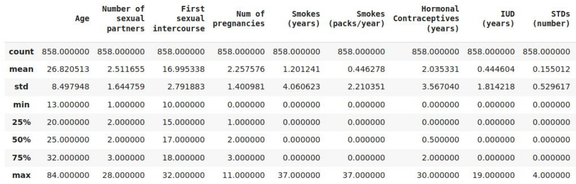
Several patients decided to not answer some questions due to personality. As a result, 13% of total questions were missed.

There are two features with 92% of missing values which are STDs: Time since first diagnosis and STDs: Time since last diagnosis, so they have been omitted.

Secondly, five classifiers including the GNB, KNN, DT, LR and RF are applied to figure out the appropriate classifier for this data.Imbalanced data have been studied by applying three techniques, over-sampling, under sampling and both together.
Eventually, for sharpening the results and looking at the main risk factors of cervical cancer, feature selection methods are applied like wrapper methods. Wrapper methods such as Sequential Feature Selector, both Forward and Backward version are used.

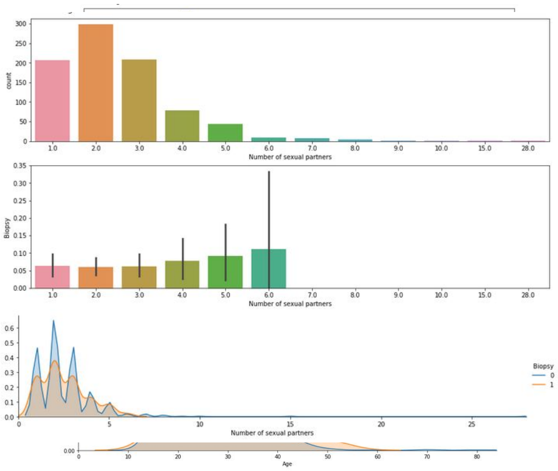
Descriptive analysis

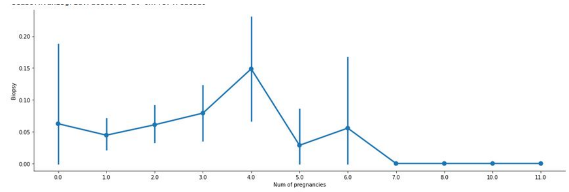
In here using this analysis we can get great insight into the shape of each attribute. Count, Mean, Standard deviation, Minimum value, 25th Percentile, 50th Percentile (Median), 75th Percentile and Maximum value are eight descriptive statistics that we used in this analysis.



	Age	Number of sexual partners	First sexual intercourse	Num of pregnancies	Smokes (years)	Smokes (packs/year)	Hormonal Contraceptives (years)	IUD (years)	STDs (number)
count	858.000000	858.000000	858.000000	858.000000	858.000000	858.000000	858.000000	858.000000	858.000000
mean	26.820513	2.511655	16.995338	2.257576	1.201241	0.446278	2.035331	0.444604	0.155012
std	8.497948	1.644759	2.791883	1.400981	4.060623	2.210351	3.567040	1.814218	0.529617
min	13.000000	1.000000	10.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	20.000000	2.000000	15.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	25.000000	2.000000	17.000000	2.000000	0.000000	0.000000	0.500000	0.000000	0.000000
75%	32.000000	3.000000	18.000000	3.000000	0.000000	0.000000	2.000000	0.000000	0.000000
max	84.000000	28.000000	32.000000	11.000000	37.000000	37.000000	30.000000	19.000000	4.000000

Age, number of sesxual parttnesr and number of pregnancy is highly correlated with biopsy.





Inference :

Most of the patients are in the age group 20 -40.We have just grouped the overall features as

Sexual habits attributes-

- Predominant of the patients had 0 -5 sexual partners.

- Most of them had their first sexual intercourse between 15 - 20 years.
- The larger group of patients had 1 -3 pregnancies overall in their life.

Smoking habits attributes-

- Relatively larger proportion of the patients are non-smokers (around 700) and only a very few (around 100) are smokers.

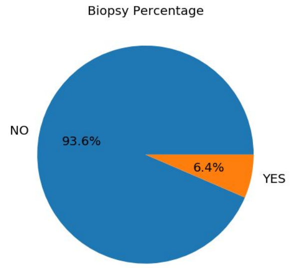
Birth control attributes

- 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 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 whereas IUD needs an doctor supervision and the patient needs to be in hospital.
- 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.

Sexually Transmitted Diseases attributes :

- The countplot above depicts that only a very very few people are affected by any one of the STDs.
- So clearly there's an imbalance here and hence we can suggest that the STD attributes may have a significant role while building the models.

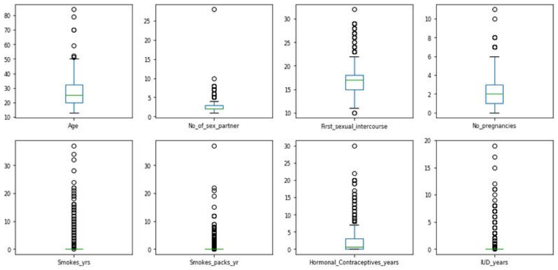
On classification problems we need to know how balanced the class values are. This is a highly imbalanced problem.



So only 6.4% of records have positive biopsy results. Since there is an imbalance in data, which needs to be taken care of in the model building section.

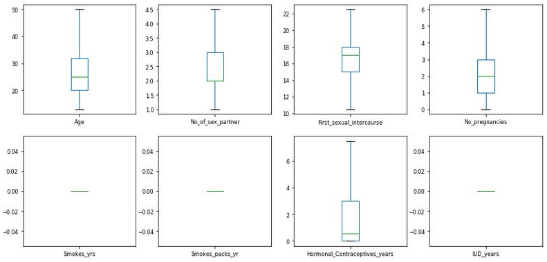
Feature Engineering

To improve the performance of the model, prepare the proper input dataset and be compatible with the machine learning algorithm requirements used in feature engineering in this study. There are many techniques that are used for feature engineering. In this study used handling outliers. [3]



The above graph implies that the data contains outliers. In this study, we used the IQR method to remove outliers. The interquartile range (IQR), also called the midspread or middle 50%**,** or technically H-spread, is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles, IQR = $Q3 - Q1$. [5]

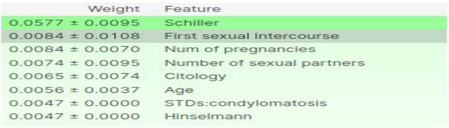
Bad data, wrong calculation these can be identified as outliers and should be dropped to correct them. After the removal of the outliers it may be displayed like below. [3]



While building models ,though outliers removal may have positive impact like getting higher accuracy and other metrics. Once again since it's an medical dataset, it's not recommended to just cap or remove outliers. for example : there are some females who are aged 70+ which comes out as extreme values, generally we should not be capping them to the upper whisker value (around 50) as it would alter the information provided by the data. Hence we are here building models with the original values as such.

FEATURE IMPORTANCE

Feature importance is a very classic and popular method.The importance of a feature is the increase in the prediction error of the model after we permuted the feature’s values, which breaks the relationship between the feature and the true outcome.Feature importance applied after the model has been fitted.There are many ways you can calculate feature importance. In this model, we will use 'Permutation importance' by the eli5 library.



Weight	Feature
0.0577 ± 0.0095	Schiller
0.0084 ± 0.0108	First sexual intercourse
0.0084 ± 0.0070	Num of pregnancies
0.0074 ± 0.0095	Number of sexual partners
0.0065 ± 0.0074	Citology
0.0056 ± 0.0037	Age
0.0047 ± 0.0000	STDs:condylomatosis
0.0047 ± 0.0000	Hinselmann

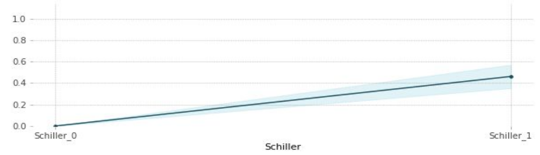
In the plot, the most important feature is at the top and most less important features are at the bottom.

According to this plot most affect features for cervical cancer are schiller, First Sexual intercouse, Number of pregnancies.

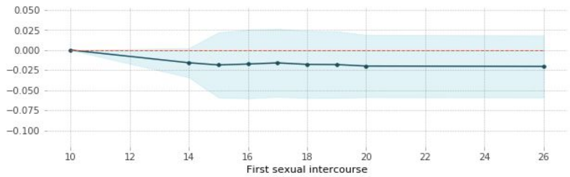
How a single feature affect our prediction

Partial Dependence Plots or PDP is also a very popular method. PDP is calculated after the model is fitted. We then use a single row from test data to predict the outcome. Instead of predicting one prediction, we repeatedly alter one variable of the row to make a series of predictions. For example, for our cervical cancer model, we take a row from test data and repeatedly alter a single variable value like age, and then make a series of predictions. And we do these for multiple rows, then plot average predicted the outcome on vertical axes.

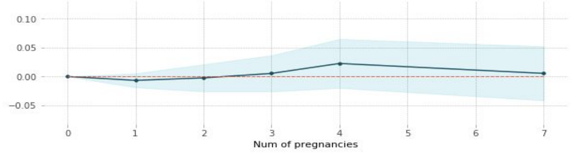
PDP for schiller



PDP for first sexual intercourse

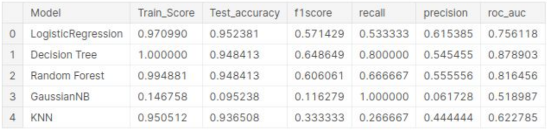


PDP for number of pregnancies



FINAL MODEL AND OPTIMIZATION

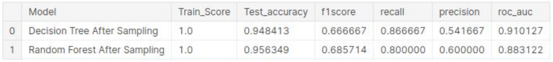
First of all, we have to take care of the data imbalance of the target variable and train model data using Logistic Regression, Decision Tree, Random Forest, GaussianNB and KNN. Below show the results accordingly.



	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
0	LogisticRegression	0.970990	0.952381	0.571429	0.533333	0.615385	0.756118
1	Decision Tree	1.000000	0.948413	0.648649	0.800000	0.545455	0.878903
2	Random Forest	0.994881	0.948413	0.606061	0.666667	0.555556	0.816456
3	GaussianNB	0.146758	0.095238	0.116279	1.000000	0.061728	0.518987
4	KNN	0.950512	0.936508	0.333333	0.266667	0.444444	0.622785

Specifically as this is 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. In this study why recall should be given higher importance because we have to predict actual cancer patients as cancer patients accurately.

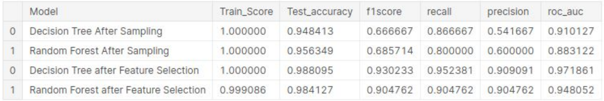
After that, we have used an oversampling technique which is called SMOTE to overcome the data imbalance. In this case 0 and 1 have the same proportion. Below show the results.



	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
0	Decision Tree After Sampling	1.0	0.948413	0.666667	0.866667	0.541667	0.910127
1	Random Forest After Sampling	1.0	0.956349	0.685714	0.800000	0.600000	0.883122

Feature Selection

We have used Recursive Feature Elimination technique for feature selection. Recursive Feature Elimination (RFE) as its title suggests recursively removes features, builds a model using the remaining attributes and calculates model accuracy. RFE is able to work out the combination of attributes that contribute to the prediction on the target variable. Used RFE on Decision Tree and Random Forest separately and found the best features for both the models individually. The features are chosen based on recall score. As our results show, obviously that recall score got better after the feature selection.

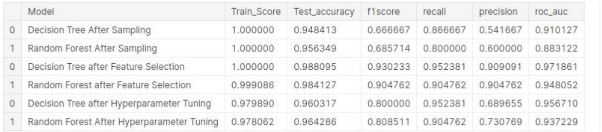


	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
0	Decision Tree After Sampling	1.000000	0.948413	0.666667	0.866667	0.541667	0.910127
1	Random Forest After Sampling	1.000000	0.956349	0.685714	0.800000	0.600000	0.883122
0	Decision Tree after Feature Selection	1.000000	0.988095	0.930233	0.952381	0.909091	0.971861
1	Random Forest after Feature Selection	0.999086	0.984127	0.904762	0.904762	0.904762	0.948052

Accuracy can be further improved by tuning the hyper parameters of the model.

Hyper Parameter Tuning

Here we used Grid Search Cross Validation for Decision Trees and Randomized Search Cross Validation for Random Forest for choosing the best parameter values. The recall score has improved a lot after hyperparameter tuning.

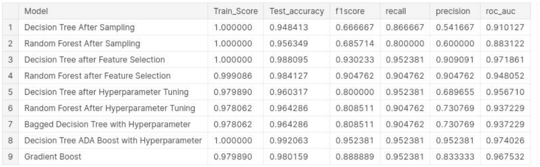


	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
0	Decision Tree After Sampling	1.000000	0.948413	0.666667	0.866667	0.541667	0.910127
1	Random Forest After Sampling	1.000000	0.956349	0.685714	0.800000	0.600000	0.883122
0	Decision Tree after Feature Selection	1.000000	0.988095	0.930233	0.952381	0.909091	0.971861
1	Random Forest after Feature Selection	0.999086	0.984127	0.904762	0.904762	0.904762	0.948052
0	Decision Tree after Hyperparameter Tuning	0.979890	0.960317	0.800000	0.952381	0.689655	0.956710
1	Random Forest After Hyperparameter Tuning	0.978062	0.964286	0.808511	0.904762	0.730769	0.937229

After that finally check out can improve furthermore using ensembling methods.

Ensembling

Here we used three ensembling techniques. They are Bagging, AdaBoost and Gradient Boost. According to our results there is no improvement when using Bagging technique. Adaboost may not be the optimum way as there is overfitting. And Gradient boost gave the best recall score of 95.2%. Finally after applying all optimization techniques, we were able to increase recall score from 53.3 % to 95.2 %.



	Model	Train_Score	Test_accuracy	f1score	recall	precision	roc_auc
1	Decision Tree After Sampling	1.000000	0.948413	0.666667	0.866667	0.541667	0.910127
2	Random Forest After Sampling	1.000000	0.956349	0.685714	0.800000	0.600000	0.883122
3	Decision Tree after Feature Selection	1.000000	0.988095	0.930233	0.952381	0.909091	0.971861
4	Random Forest after Feature Selection	0.999086	0.984127	0.904762	0.904762	0.904762	0.948052
5	Decision Tree after Hyperparameter Tuning	0.979890	0.960317	0.800000	0.952381	0.689655	0.956710
6	Random Forest After Hyperparameter Tuning	0.978062	0.964286	0.808511	0.904762	0.730769	0.937229
7	Bagged Decision Tree with Hyperparameter	0.978062	0.964286	0.808511	0.904762	0.730769	0.937229
8	Decision Tree ADA Boost with Hyperparameter	1.000000	0.992063	0.952381	0.952381	0.952381	0.974026
9	Gradient Boost	0.979890	0.980159	0.888889	0.952381	0.833333	0.967532

So, finally we have chosen the 'Gradient boosting' model as it gave the best accuracy score, roc score, f1 score and recall score compared to other models. So, the Gradient Boost model looks superior while considering the overall evaluation metrics. Therefore, we have finalized it as the final model.

As well as this dataset has some limitations also. The dataset had a lot missing values, class-imbalance, lack of adequate records which would adversely affect the model prediction. We overcame these limitations by best practices of industry like imputing the null values with algorithms and overcoming class imbalance using sampling techniques.

CONCLUSION

This article presents the comparison between different machine learning classifiers with respect to the best predictive model for Cervical Cancer Dataset. It is an unbalanced dataset. First analyze data using this unbalanced data. After that respectively used feature selection, hyper parameter tuning and ensembling techniques to improve performance of the model. In assembly we use three techniques also. Among these techniques finally we choose the Gradient boosting as our best model for analysis. It has a Train_score, Test_accuracy, f1_score, recall, precision and roc_auc of 0.979890, 0.980159, 0.888889, 0.952381, 0.833333, 0.967532 respectively. This model is used to record the patient's past record of habits and actions and can predict the presence of cancer cells. The proposed method is faster and accurate which has a huge scope in medical assistance and hence can be used to save many lives.

REFERENCES

[1] Abdoh, S., Abo Rizka, M. and Maghraby, F., 2018. Cervical Cancer Diagnosis Using Random Forest Classifier With SMOTE and Feature Reduction Techniques. IEEE Access, 6, pp.59475-59485

[2] Alwesabi, Y., Won, D. and Choudhury, A., "*Classification Of Cervical Cancer Dataset*"

[3] Medium. 2006. Fundamental Techniques Of Feature Engineering For Machine Learning.

[4] Rençberoğlu, E., 2019. *Fundamental Techniques Of Feature Engineering For Machine Learning*. [online] Medium.

[5] Sharma, N., 2018. Ways To Detect And Remove The Outliers. [online] Medium.