Model Explainability of Cervical Cancer Dataset

Yusuf Furkan Yücesoy,

Eskişehir Osmangazi University,TR

Abstract

The problem of black box in machine learning models is a danger for models used in risky areas. For example, the decision-making mechanism of a model developed in the field of health needs to be transparent. In this paper, the data set of a risky subject such as cervical cancer was studied and analyzed how to select the performance metric of models trained with imbalanced datasets. Cervical cancer dataset has been published in 2017 by, which involves 858 samples and 32 features as well as four targets. These attributes include demographic information, habits like smoking and historic medical records [1]. In addition, model explainability by various approaches is emphasized.

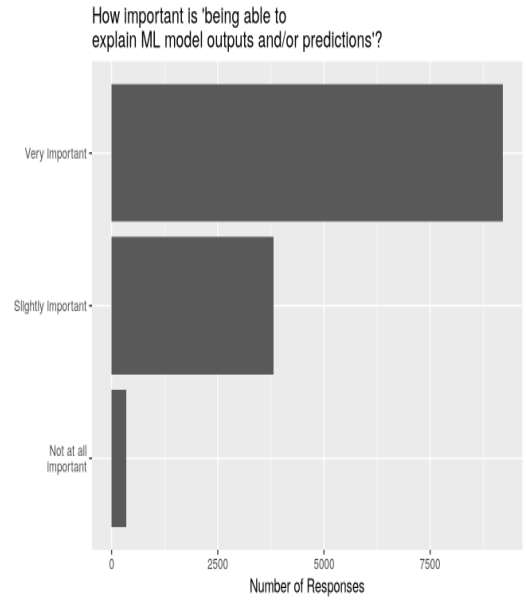
Keywords

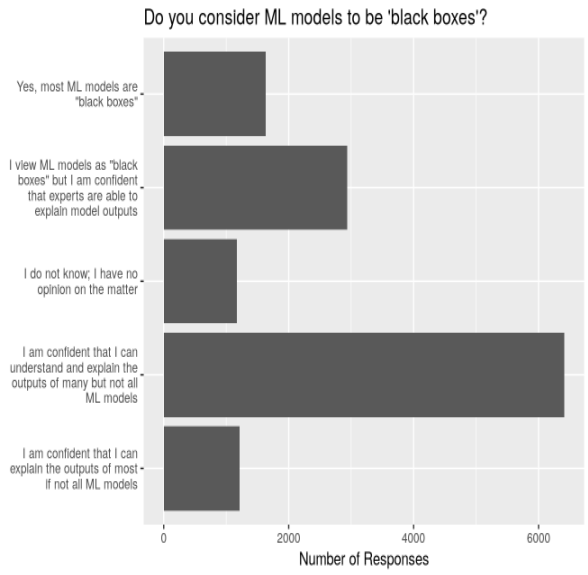
Cervical cancer, model explainability, imbalanced data, black box problem, data analysis, feature importance, model behavior

1.Introduction

Today, machine learning is used to predict factors such as whether a person can repay their loan or whether they have certain diseases. We can give endless examples of the use of machine learning in the modern world. However, we have some disadvantages as well as the advantages of machine learning. One of the major disadvantages is the 'lack of identifiable machine learning model'.

In the 2018 Kaggle survey, Kaggle asked data scientist how they view about Machine learning model explainability [2]. Here is the result of that questions:





According to the survey result, we find that

- Model explainability is very important.
- Many people have skills to explain the outcome of many models but not all.
- Unfortunately, there are also many people who think that expert can only explain the outcome of a model and many of them think machine learning model as a black box.

Since machine learning is a part of life, we want to know why our model makes such a prediction. This will create the reliability and overall use of the model. In addition, our Model can select bias from training data. This machine learning model can turn our model into a racist model that discriminates against some groups. We can better solve the bias problem by explaining the model definition. Model explainability helps us in scenarios where a single error can cause major damage. In particular, the accuracy score of models trained with imbalanced dataset can mislead us. If it is working with an unstable dataset, other metrics should also be checked.

The aim of the paper is to work on a vital issue such as cervical cancer by working with an imbalanced dataset to ensure the explainability of the medically used model for predicting cervical cancer. Cervical cancer is the most common cancer among women in developing countries, the WHO report [3]. In the United States, there are 129,001 new cases in 2015 despite the provided healthcare facilities, where 273,000 deaths in 2002 worldwide. Cervical cancer dataset has been published in 2017 by, which involves 858 samples and 32 features as well as four targets. These attributes include demographic information, habits like smoking and historic medical records.

2. Data Description

The Cervical Cancer Dataset is obtained from UCI Repository and involves 858 samples and 32 features as well as 4 classes (Hinselmann, Schiller, Cytology and Biopsy) has been published in [1]. This paper focuses on studying the Biopsy target as it recommended by the literature review. In addition, the dataset studied was imbalanced according to the Biopsy target because only 55 of the patients were diagnosed with cancer.

Attribute	**Type**	**Attribute**	**Type**	**Attribute**	**Type**
Age	Integer	STDs	Bool	STDs:HIV	Bool
Number of sexual partners	Integer	STDs (number)	Integer	STDs:Hepatitis B	Bool
First sexual intercourse (age)	Integer	STDs:condylomatosis	Bool	STDs: Number of diagnosis	Integer

Number of pregnancies	Integer	STDs:cervical condylomatosis	Bool	STDs: Time since first diagnosis	Integer
Smokes	Bool	STDs:vaginal condylomatosis	Bool	STDs: Time since last diagnosis	Integer
Smokes (years)	Real	STDs:vulvo-perineal condylomatosis	Bool	Dx:Cancer	Bool
Smokes (packs/year)	Real	STDs:syphilis	Bool	Dx:CIN	Bool
Hormonal Contraceptives	Bool	STDs:pelvic inflammatory disease	Bool	Dx:HPV	Bool
Hormonal Contraceptives (years)	Real	STDs:genital herpes	Bool	Dx	Bool
IUD	Bool	STDs:molluscum contagiosum	Bool	Biopsy (target)	Bool
IUD (years)	Real	STDs:AIDS	Bool		

Table 1: Attributes and their types

3. Methodology

The approach of the paper is to work on a vital issue such as cervical cancer by working with an imbalanced dataset to ensure the explainability of the medically used model for predicting cervical cancer. Besides, the competence of accuracy metric in models developed with imbalanced datasets is examined. In this study, random forest classifier as an ML algorithm is used for the test. Furthermore, Partial Dependence Plots determine the effect of each feature on success. Also, ”Permutation Importance” function is used to analyze the importance of attributes in the model. Finally, SHAP Values (an acronym from SHapley Additive exPlanations) break down a prediction to show the impact of each feature. It explains why a model made a certain prediction.

Thus, the usability of the features used in the prediction of a risky condition such as cervical cancer on disease detection will be analyzed.

Some metrics used in the experiment are shown in Table 2.

Term	**Formula**	**Definition**
Accuracy	(TP + TN)/(P+N)	Rate of the correct prediction for both healthy and not healthy patients
Sensitivity=recall= true positive rate	TP/(TP+FN)	The percentage of sick people who are correctly identified as having the disease.

Table 2: Basic notations [3]

.4. Experiment and Results

A. Test of Accuracy in Random Forest Classifier Algorithm Without Preprocessing on Dataset

In order to work with Random Forest Classifier, our data set is divided into two groups as test and train (test: %25, train: %75). The results obtained from the training and test results are given in the Table 3.

Accuracy score:	0.9627906976744186

Table 3: Accuracy

Our data set, which has an imbalanced distribution with no pre-processed values, gives a performance score of 95%. However, we have no idea what our model currently predicts a patient as cancer or not cancer. In the literature, this is called the black box problem. Currently, models are evaluated using accuracy score.

In order to test to what extent the accuracy score is a parameter that can provide confidence in the model, a randomly selected (UID) attribute and target value, Cancer attribute, is allocated as 25-75% test and train. The accuracy score obtained as a result of training and test is given in the Table 4.

Accuracy score:	0.9488372093023256

Table 4: Accuracy

It was observed that the model can make predictions with approximately the same accuracy, even when training with a single attribute. Here, it was revealed that the accuracy score parameter did not fully reflect the success of the model. The reason for the problem is that the model cannot actually perform the learning process correctly. Because the dataset being studied is a dataset with imbalance in the number of data. This imbalance is shown in Table 5.

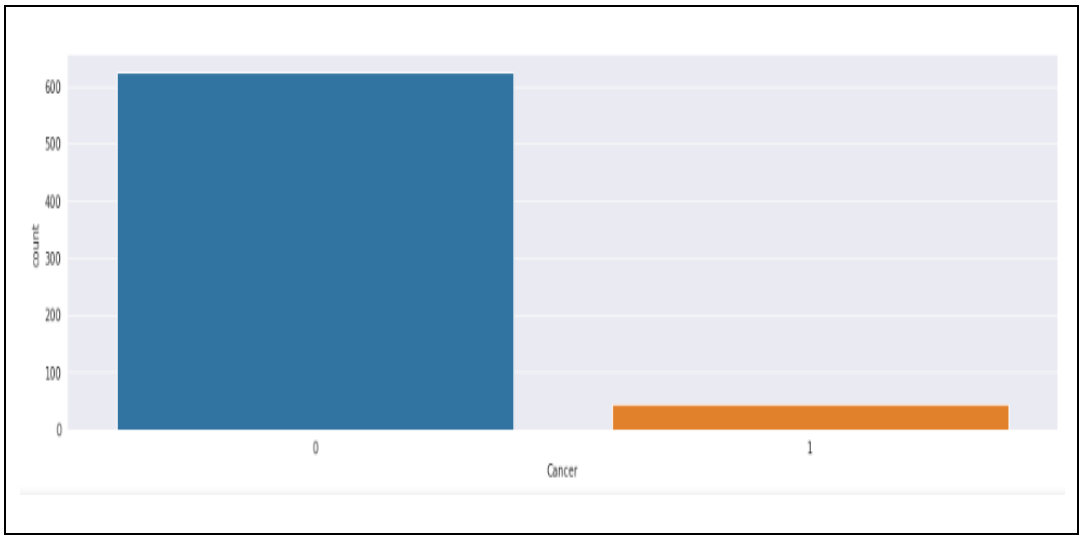


Table 5: Imbalanced dataset, 0: not cancer, 1: cancer

In this context, the other metric recall (sensitivity) and cancer patients with the correct rate of the case was investigated. As a result of the recall (sensitivity) metric, it was found that the model with a very high performance score was in fact very low in the correct rate of cancer. The recall metric is shown in Table 6.

Recall (Sensitivity)	0.38461538461538464

Table 6: Recall

As can be seen from the complexity matrix [Table 7], only 5 of the 13 cancer patients in the test set were diagnosed with cancer by the model. 8 cancer patients were defined as not cancer. This is actually an indication that the model has made the wrong decision.

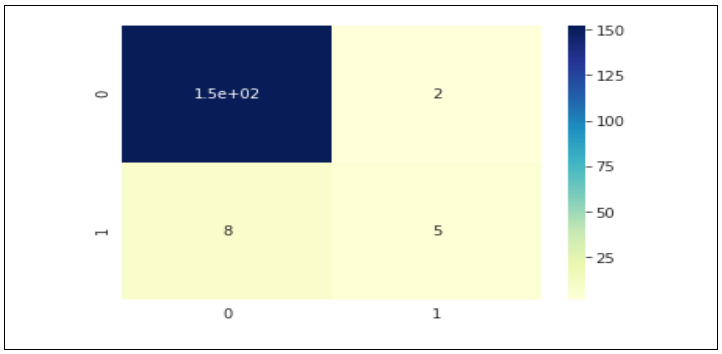


Table 7: Confusion matrix

The aim of this first experiment is to show the inadequacy of the accuracy score parameter in terms of model reliability due to the fact that most of the real life data is imbalanced.

B. Data Preprocessing

In line with the observation on the dataset, it was observed that there were missing values in the dataset. The number of missing values found in the "STDs: Time since first diagnosis" and "STDs: Time since last diagnosis" is remarkably high given that the total number of samples is 858. [Table 8]

Age	0
Number of sexual partners	26
First sexual intercourse	7
Num of pregnancies	56
Smokes	13
Smokes (years)	13
Smokes (packs/year)	13
Hormonal Contraceptives	108
Hormonal Contraceptives (years)	108
IUD	117

IUD (years)	117
STDs	105
STDs (number)	105
STDs:condylomatosis	105
STDs:cervical condylomatosis	105
STDs:vaginal condylomatosis	105
STDs:vulvo-perineal condylomatosis	105
STDs:syphilis	105
STDs:pelvic inflammatory disease	105
STDs:genital herpes	105
STDs:molluscum contagiosum	105
STDs:AIDS	105
STDs:HIV	105
STDs:Hepatitis B	105
STDs:HPV	105
STDs: Number of diagnosis	0
STDs: Time since first diagnosis	787
STDs: Time since last diagnosis	787
Dx:Cancer	0
Dx:CIN	0
Dx:HPV	0
Dx	0
Hinselmann	0
Schiller	0
Citology	0
Biopsy	0

Table 8: Number of missing values for each attribute

Because there are too many missing values in the "STDs: Time since first diagnosis" and "STDs: Time since last diagnosis" attributes in our dataset, it is removed from the dataset to avoid breaking the realistic approach. In addition, each sample with missing value was extracted from dataset. There are 668 samples with 34 attributes that do not contain missing values in the dataset after the missing value subtraction operations.

C. SMOTETomek Approach for Imbalanced Dataset

By resampling, we can make unbalanced data sets more balanced. The first method to do this is to increase the minority class data by various methods to obtain classes with an equal number of data (oversampling). The other method is to obtain a balanced data set (under-sampling) by subtracting the data from the weighted class from the data set.
In this study, the data set was balanced with SMOTETomek (synthetic minority oversampling technique + Tomek Link) approach, which is a combination of oversampling and under-sampling [4].

SMOTETomek technique; It is a technique in the imbalanced_learn library that uses the SMOTE oversample method to generate synthetic data by interpolation and the Tomek Link method, which clears overlapping samples. Therefore, SMOTETomek technique can solve the problem of

imbalanced data set without disturbing the structure of the dataset. The number of cancer patients before and after the SMOTETomek method was applied to the Train dataset is given in the Table 9.

	Number of patients with cancer in train datadet	Number of patiens with not cancer in train dataset
Before applying the SMOTETomek method	33	468
After applying the SMOTETomek method	465	465

Table 9:Dataset before and after SMOTETomek

The accuracy score and recall values of the data set balanced by SMOTETomek method are given in the Table 11. After the balancing study, our Recall (sensitivity) metric increased significantly.

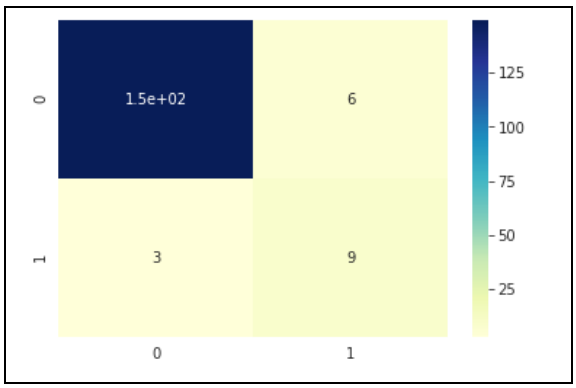


Table 10: Confusion matrix after applying SMOTETomek

Accuracy score:	0.9461077844311377
Recall:	0.75

Table 11:Accuracy score and recall value after SMOTETomek method

D. Importance of Attributes

There are many attributes in the cervical cancer dataset used. In the approach of ensuring the model's explainability, knowing which of these features is more effective in predicting cancer and which is less effective helps to understand the working principle of the model. In the features columns, it randomly shuffles each column without touching the target and other feature column and calculates how it affects the accuracy of the prediction of new shuffled data. There are many ways you can calculate feature importance. In this paper, used ‘Permutation importance' by the eli5 library [5] [6]. The Permutation Importance method presented by Eli5 provides the ability to calculate attribute significance for black-box problems by measuring how the score decreases when an attribute is not available.

Weight	Feature
0.0647 ± 0.0418	Schiller
0.0072 ± 0.0048	STDs:syphilis
0.0060 ± 0.0000	STDs: Number of diagnosis
0.0048 ± 0.0048	STDs
0.0036 ± 0.0059	Num of pregnancies
0.0036 ± 0.0096	Age
0.0012 ± 0.0048	Hinselmann
0.0012 ± 0.0048	First sexual intercourse
0.0012 ± 0.0048	IUD
0.0000 ± 0.0076	Number of sexual partners
0 ± 0.0000	IUD (years)
0 ± 0.0000	Dx:HPV
0 ± 0.0000	STDs:condylomatosis
0 ± 0.0000	STDs:vaginal condylomatosis
0 ± 0.0000	Smokes (packs/year)
0 ± 0.0000	STDs:molluscum contagiosum
0 ± 0.0000	Dx:CIN
0 ± 0.0000	STDs:genital herpes
0 ± 0.0000	STDs (number)
0 ± 0.0000	STDs:cervical condylomatosis
… 13 more …	

Table 12: Importance of attributes

This study analyzes the extent to which attributes affect the decision-making process of the model when estimating As the table moves from the bottom up, the effect on the decision mechanism is increased. For example, the most important feature for the model was found to be Schiller test.[Table 12]

E. Analysis of the Effect of a Single Attribute on Prediction

In the previous study, the most important features of the model have been identified, but examining the effect of each attribute on decision-making is one of the approaches that clarify the black-box problem. For example, the attribute 'Age' was found to be an important feature for decision-making, but we do not know that the chance of cancer increases or decreases with age. As in this example, knowing how the values of a single attribute affect the prediction of the model in itself is one of the factors that increase model identifiable. In this study, a technique called "Partial Dependence Plots" is used to observe how a single attribute affects model estimation. The working principle is based on making a series of estimates by repeatedly changing a variable of the line instead of making a single estimate. A few examples obtained during the study are given in the Table 13.

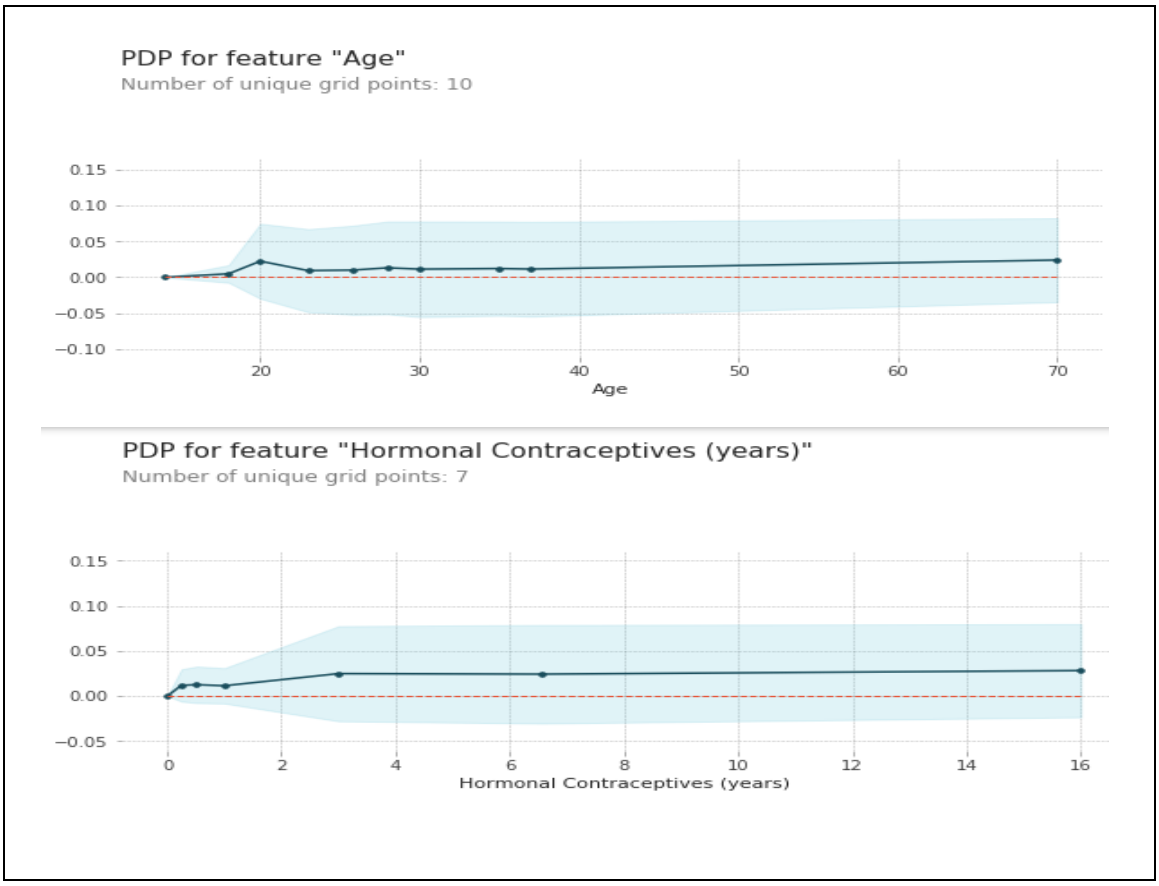


Table 13: Partial Dependence plot samples of "Age" and "Hormonal Contraceptives (years)"

The findings show that the most risky age is 20 years and the risk of cancer increases with increasing age. Also, the findings show that the number of hormonal contraceptives used annually increases the risk of cancer.[Table 13]

F. Analysis of Prediction

In this study, it is analyzed which model makes an estimation based on which attributes and how much. SHAP values are used for these analyzes. SHAP values are one of the techniques used to interpret the estimation of the model [6].
For example; In the cancer prediction model studied, it clarifies how each attribute together contributes to prediction if the model predicts that someone is cancer based on certain attributes. Contributing here means how an attribute affects the prediction of cancer done.
In this study, two samples randomly selected from the dataset are used as tests and it is estimated whether there is cancer or not.

Sample1

Prediction result for Sample1:	array([[0.726, 0.274]])

Table 14:Prediction result for Sample1

The model estimated that Sample1 was not cancer with a 72.6% probability, with a 27.4% chance of cancer. SHAP values are given in the Table 15 in order to analyze why the estimation results in this way.

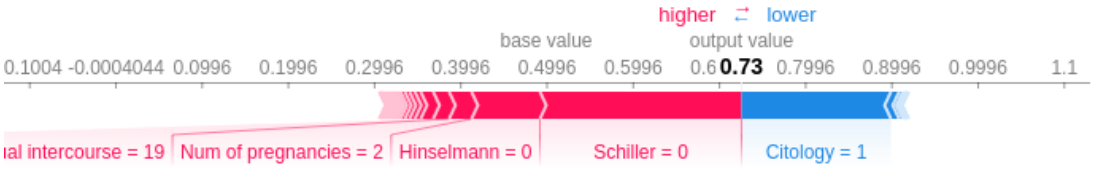


Table 15: SHAP values for sample1 prediction result

***Attribute values that lead to incremental estimates are pink, attribute values that reduce the estimate are blue. The visual dimensions show the magnitude of the effect of the attribute*.**

- Schiller = 0 has the greatest effect on the estimation result of the model. In addition, Citology = 1 has a huge impact on the reduction of the result. This situation is interpreted as Schiller = 0 increases the likelihood of not having cancer, and Citology = 1 decreases the probability of not having cancer.

Sample2

Prediction result for Sample2:	array([[0.092, 0.908]])

Table 16:Prediction result for Sample2

The model estimated that Sample2 was not cancer with 9.2% probability and cancer with 90.8% probability. The SHAP values are below to analyze why the prediction yields this result. [Table17]

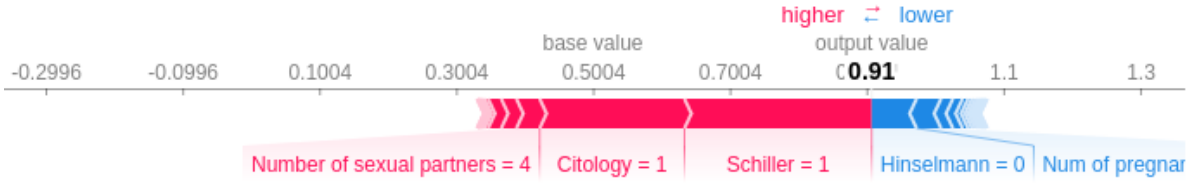


Table 17: SHAP values for sample2 prediction result

- In this analysis, it was observed that Number of sexual partners = 4, Schiller = 1 and Citology = 1 were influenced to increase the outcome of the model. This is plausible with the consistency of the previous analysis.
- In addition, Hinselmann = 0 and Number of pregnancies = 1 have the effect of reducing the output of the model, that is, the rate of cancer.

5. Conclusion

In this study, the solution of the black box problem and the necessity of model identification were emphasized and experiments were conducted in this context. For a critical prediction model such as cervical cancer, the behavioral analysis of the model has been performed by following different techniques and approaches in terms of model explainability and comprehensibility. With the analysis made, the decision mechanism of the model is clarified; which features affect how this decision mechanism. In particular, a sector that does not accept error, such as health, was selected and studied with real data. In critical areas, “Accuracy Score” parameter was found to be insufficient and may mislead the user about the reliability of the model which is trained by imbalanced dataset.

References

[1] "UCI-Machine Learning Repository," 03 March 2017. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+%28Risk+Factors%29. [Accessed 21 November 2019].

[2] G. Keleher, "2018 Kaggle ML & DS Survey-Importance of Interpretability," 13 November 2018. [Online]. Available: https://www.kaggle.com/graeme16161/importance-of-interpretability.

[3] Sharma, P., and Pattanshetty ,S.M., "A Study on Risk Factors of Cervical Cancer Among Patients Attending A Tertiary Care Hospital: A Case-Control Study," *Clinical Epidemiology and Global Health (In press),* 2017.

[4] "imbalanced-learn," [Online]. Available: https://imbalanced-learn.readthedocs.io/en/stable/generated/imblearn.combine.SMOTETomek.html#imblearn.combine.SMOTETomek. [Accessed 30 November 2019].

[5] A. Prakash, "medium.com," 7 July 2019. [Online]. Available: https://medium.com/towards-artificial-intelligence/how-to-use-scikit-learn-eli5-library-to-compute-permutation-importance-9af131ece387. [Accessed 1 December 2019].

[6] "Kaggle-SHAP values," [Online]. Available: https://www.kaggle.com/dansbecker/shap-values. [Accessed 15 December 2019].