IJCSE International Journal of Computer Sciences and Engineering Open Access

Technical Paper Vol.8, Issue.4, Apr 2020 E-ISSN: 2347-2693

Data Analysis: Finding the Most Effective Factors Causing Cancer Deaths

V. Naveen Babu[1*], T. Murali[2], Sk. Meer Hussian[3], S. Nava Chaitanya[4] and Madda.Varalakshmi[5]

[1,2,3,4,5] Dept of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur district, Andhra Pradesh , India.

DOI: https://doi.org/10.26438/ijcse/v8i4.9096 | Available online at: www.ijcseonline.org

Received: 3/Apr/2020, Accepted: 16/Apr/2020, Published: 30/Apr/2020

Abstract: The spreading of abnormal cells in the human body with much potential is a basic cause of disease cancer. The growth of abnormal cells may be affected by age group, being disease-oriented, or type of location in which people live and many factors. Because of the circumstances, there is no possibility of avoiding the growth of abnormal cells, but by taking corrective measures the growth can be slowed down to some extent. In addition to that, it will envisage the cancer causes which in turn can be used to create awareness among the people. In this fact, it is important to determine if someone has a high cancer risk by using biological test results which have been recorded. By working on these sample data, we can focus on finding the most influential factors that affect cancer. In this research, by applying a suitable Machine Learning algorithm on the data which have been collected using surveys, we are able to find the most important factors and mainly classification type of Machine Learning algorithms to be used for performance analysis.

Keywords: Data Analysis, Classification, Machine Learning, Cancer, XGBoost Algorithm

I. INTRODUCTION

The uncontrolled growth of incognito cells in the body causes cancer. Cancer mostly causes when the functioning of the human body fails and acts abnormally. Whenever cancer affects old cells in the body do not die and instead of it, the new cells will be born. All the extra cells may form the tumors.

Cancer is the second main cause of death in the world and is subject to an estimated 9.6 million deaths in 2018. In the world, about 1 out of 6 deaths are due to cancer. More or less 70% of deaths from cancer occur in low and middle stage countries.

Here, we consider cervical cancer which is one of the most causing cancer for women [2]. In this research, we predict cervical cancer diseases and identify the major factors that cause cervical cancer. With this result, we can take necessary precautionary measures to get rid of it and control them. We use machine learning techniques to find out the most causing attributes and also for predicting the occurrence of cervical cancer.

II. LITERATURE SURVEY

There is much research that has been done for the detection of many cancer types including cervical cancer based on data sets provided. The following specifies the literature survey of Cancer deaths:

- As per the study of Subramanian, Lung cancer causes 90% of deaths which is mainly caused due to smoking (90%-men, 75%-women) [1].
- According to the research done by Sreedevi et al, HPV infection prevalence is more (87.8%–96.67%) among women with cervical cancer and only less (9.9%–36.8%) among women with no cancer or other gynaecological morbidities [2].
- As per the study of Fernandes, a new strategy for predicting the outcome of the patient biopsy (a medical record) computed and also explored clinical findings from the embedding spaces [3]. In this research they used dimensionality reduction techniques like Joint dimensionality reduction and classification, Fully supervised embeddings and Deep supervised autoencoders.

- Roura et al. research concentrated only on tobacco smoking and how important quitting the habit of smoking is for causing cancer [4].
- As per the research done by Vineet Menon, the performance of accuracy calculated by K-Nearest Neighbor, Decision Tree and Random Forest algorithms is 95.3%, 85.11% and 87.90% are obtained respectively [5]. The algorithms used in the paper give the better result of the performance of accuracy to characterize cervical cancer

From the above studies, finding one factor that causes cancer is not a good solution even though the applied algorithm has better accuracy and the factor changes from one person to another. So we need to find all the factors from the given data set having more priority to cause cancer.

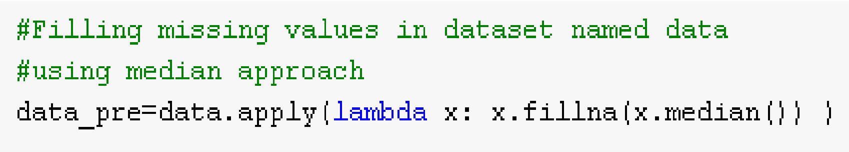
III. DATA SET

This dataset consists of 859 records with 36 attributes like Age, Number of sexual partners, First sexual intercourse (age), Number of pregnancies, Smokes, Smokes (years), Smokes (packs/year), Hormonal Contraceptives, Hormonal Contraceptives(years), IUD, IUD(years), STDs, STDs(number), STDs: condylomatosis, STDs: cervical condylomatosis, STDs: vaginal condylomatosis, STDs: vulvo-perineal condylomatosis, STDs: syphilis, STDs: pelvic inflammatory disease, STDs: genital herpes, STDs: molluscum contagiosum, STDs: AIDS, STDs: HIV, STDs: Hepatitis B, STDs: HPV, STDs: Number of diagnosis, STDs: Time since first diagnosis, STDs: Time since last diagnosis, Dx: Cancer, Dx: CIN, Dx: HPV, Dx, Hinselmann, Schiller, Cytology, Biopsy. 20% of the data is taken as training data and 80% of the data is taken for testing.

Abbreviations:

STDs	-	Sexually Transmitted Diseases
HPV	-	Human Papillomavirus
IUD	-	Intrauterine contraceptive device
HIV	-	Human immunodeficiency virus
DX	-	Diagnosis
CIN	-	Cervical intraepithelial neoplasia
AIDS	-	Acquired Immune Deficiency Syndrome

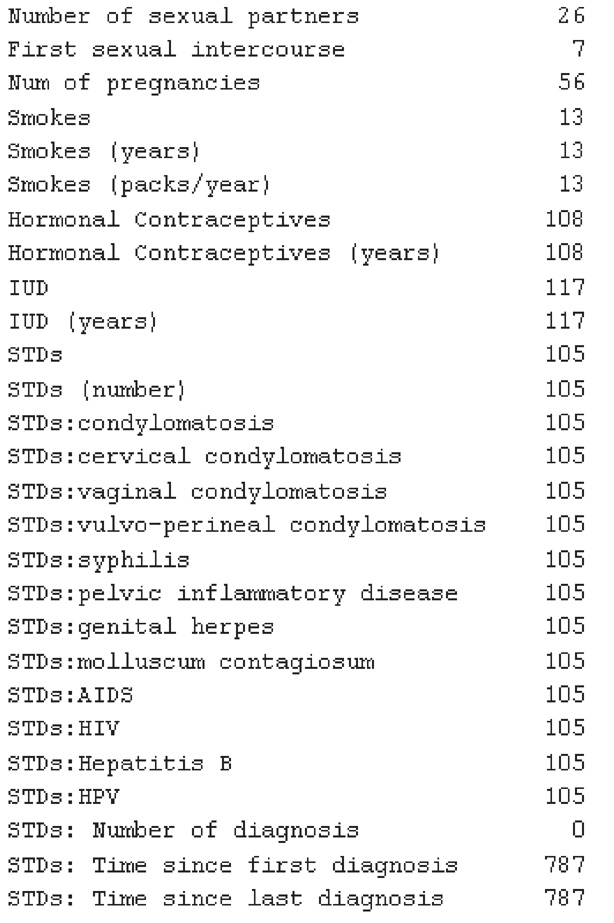
Data Preprocessing:- Before going to work on the sample data set, it has to be preprocessed such that we can get the good results [6], [9], [10]. We use modules such as "sklearn", "pandas" and "numpy" to preprocess and handle the data. The real-time data tends to be incomplete. So, some missing values occur and if we don't handle missing values, the analysis may get biased with the results falling far apart from the real and original values. In order to avoid such scenarios, it is always preferred to process the data prior to its usage in analysis using various techniques. Here we use a median approach for filling the missing values in the dataset.



```
#Filling missing values in dataset named data
#using median approach
data_pre=data.apply(lambda x: x.fillna(x.median()) )
```

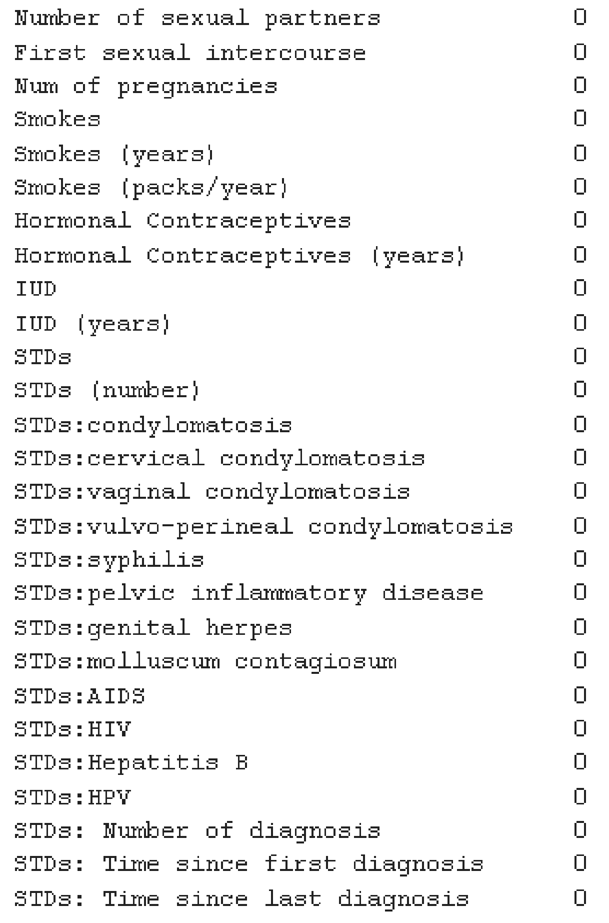
The following table indicates the attributes which have missing values in the dataset and the count of missing values:

BEFORE FILLING MISSING VALUES



```
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
```

AFTER FILLING MISSING VALUES



```
Number of sexual partners             0
First sexual intercourse              0
Num of pregnancies                    0
Smokes                                0
Smokes (years)                        0
Smokes (packs/year)                   0
Hormonal Contraceptives               0
Hormonal Contraceptives (years)       0
IUD                                   0
IUD (years)                           0
STDs                                  0
STDs (number)                         0
STDs:condylomatosis                   0
STDs:cervical condylomatosis          0
STDs:vaginal condylomatosis           0
STDs:vulvo-perineal condylomatosis    0
STDs:syphilis                         0
STDs:pelvic inflammatory disease      0
STDs:genital herpes                   0
STDs:molluscum contagiosum            0
STDs:AIDS                             0
STDs:HIV                              0
STDs:Hepatitis B                      0
STDs:HPV                              0
STDs: Number of diagnosis             0
STDs: Time since first diagnosis      0
STDs: Time since last diagnosis       0
```

IV. PROPOSED ALGORITHM:

As per the research done by T CHEN, XGBoost is suitable for solving real-world scale problems with a minimal amount of resources [7]. So, the **XGBoost** algorithm is chosen out of other algorithms.

4.1 XGBoost:

The abbreviation of the XGBoost is an eXtreme Gradient Boosting Algorithm. It is a decision tree-based algorithm. By applying this algorithm we predict better accuracy. For example, let us take cancer prediction. We can find whether the person is affected by cancer or not by taking some factors.

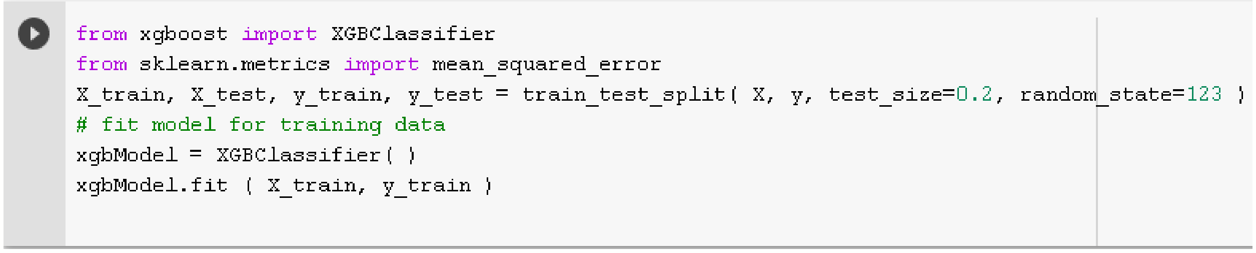
The two reasons to use XGBoost are also the two goals of the project:

1. Execution Speed.
2. Model Performance.

4.1.1 Applying XGBoost:

- First, we split the dataset into 20% of test data and 80% of train dataset.
- We train the model with the training data and then we test the performance using test data..
- The algorithm predicts the outcome for test data and returns the result

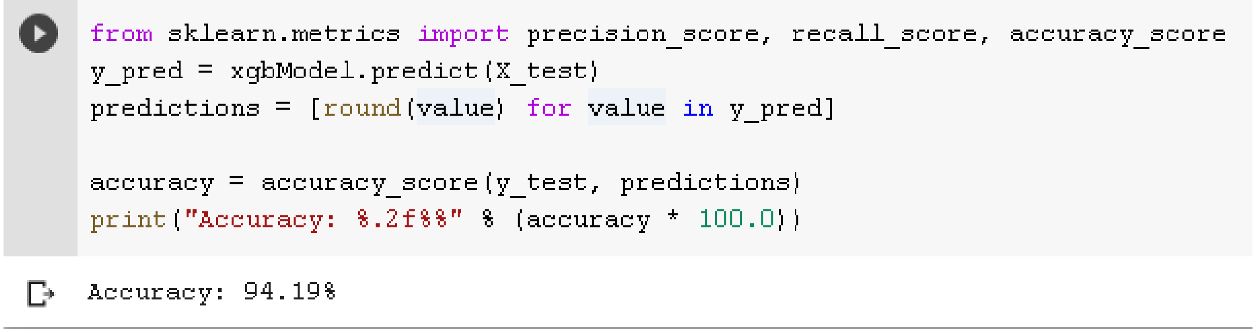
4.2 Xgboost code:



```
from xgboost import XGBClassifier
from sklearn.metrics import mean_squared_error
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.2, random_state=123 )
# fit model for training data
xgbModel = XGBClassifier( )
xgbModel.fit ( X_train, y_train )
```

4.3 Accuracy Calculation using an XGBoost:

After applying the XGBoost, the results are as follows:



```
from sklearn.metrics import precision_score, recall_score, accuracy_score
y_pred = xgbModel.predict(X_test)
predictions = [round(value) for value in y_pred]

accuracy = accuracy_score(y_test, predictions)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
Accuracy: 94.19%
```

4.4 Advantages:

- High speed and performance.
- The core of the algorithm is parallelizable.
- Parallel processing can be done by using the XGBoost algorithm.
- By using the XGBoost the missing values can be handled.

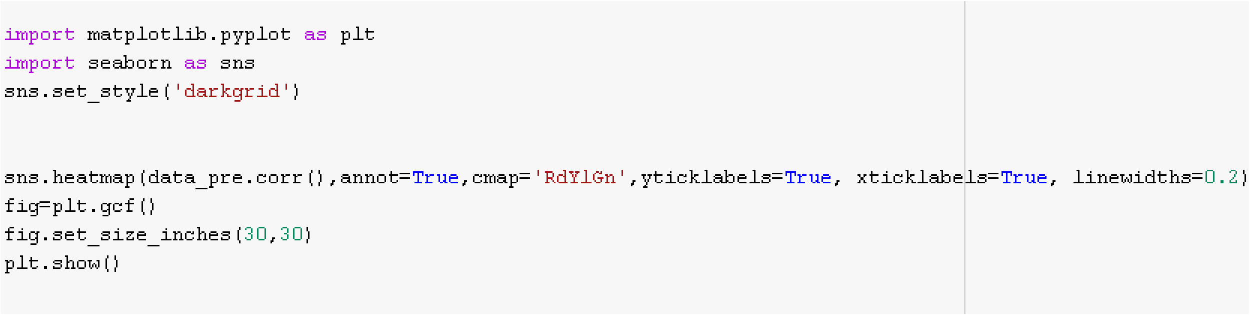
4.5 Disadvantages:

- Only work with numeric features.
- Leads to overfitting if hyperparameters are not tuned properly.

V. VISUALIZATION

The dependency between two variables can be shown using heatmap [8]. Heatmaps are used to show relationships between two variables, one plotted on each axis. By observing how cell colours change across each axis, you can observe if there are any patterns in value for one or both variables. Here, biopsy test results in cancer confirmation.

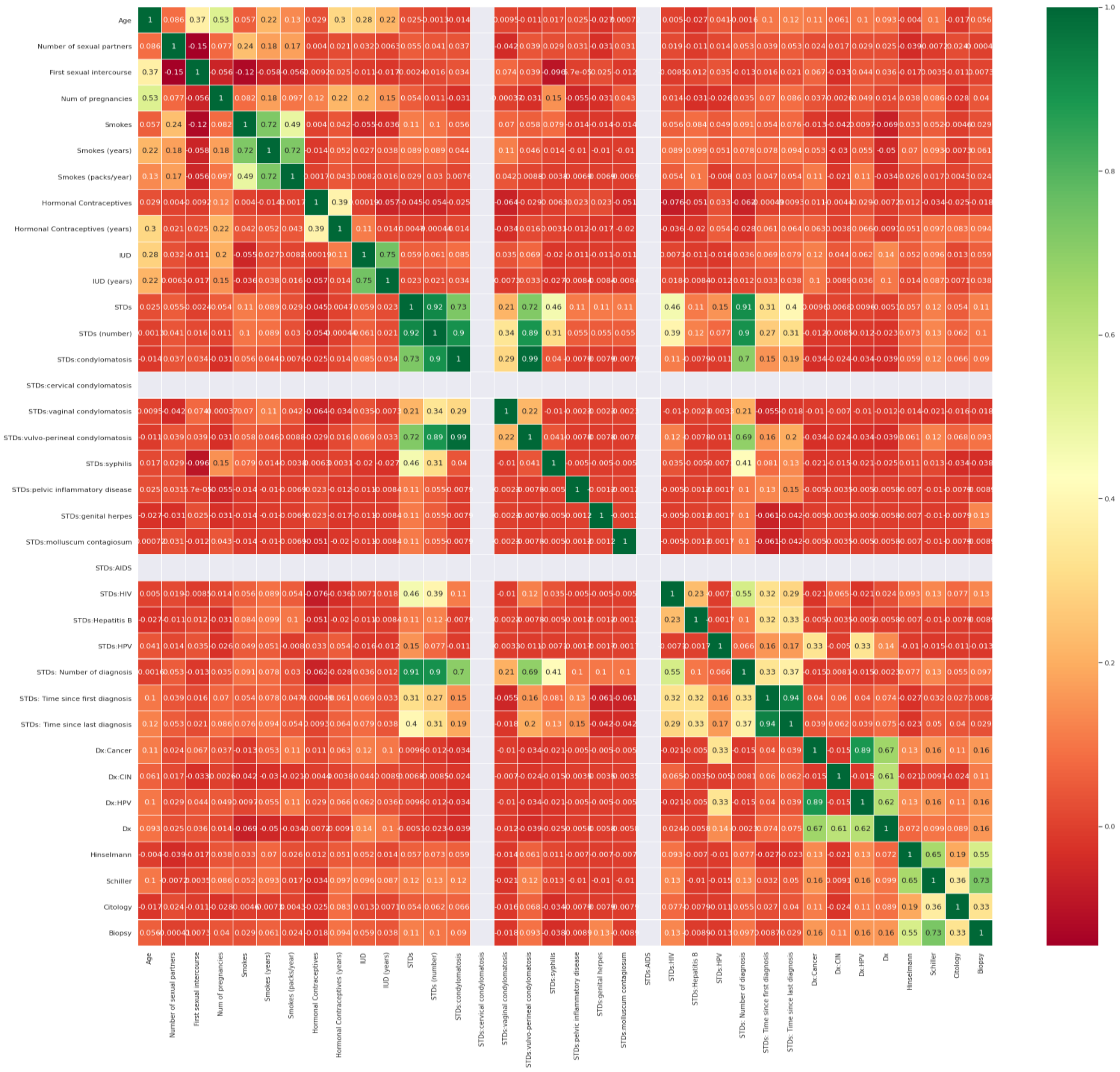
The following image represents the code for graphical representation of attributes dependency using heat map:



```
import matplotlib.pyplot as plt
import seaborn as sns
sns.set_style('darkgrid')


sns.heatmap(data_pre.corr(),annot=True,cmap='RdYlGn',yticklabels=True, xticklabels=True, linewidths=0.2)
fig=plt.gcf()
fig.set_size_inches(30,30)
plt.show()
```

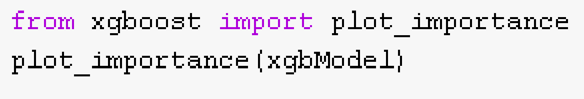
Attribute dependency can be visualized after applying the heatmap.



The above-coloured graph consists of each attribute dependency value on remaining attributes and the dependency values which are varying from 0 to 1. Here the colours are inbuilt taken starting from the lowest possible cases, from red, up to highest possible cases, to green and the diagonals are the fixed values. The diagonal represents self-dependency. Here each point represents the dependency of one attribute over another attribute.

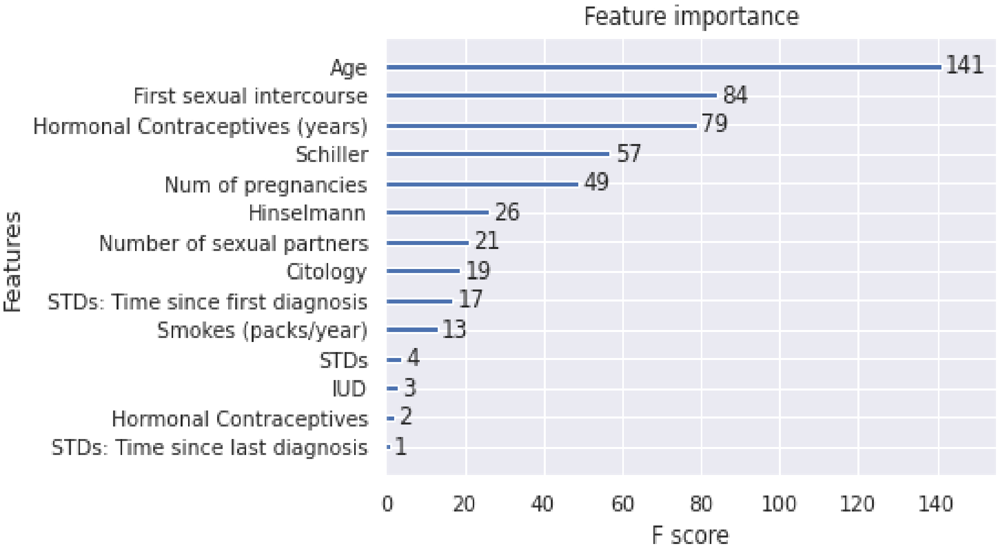
VI. FEATURE IMPORTANCE

Here we use the XGBoost algorithm for finding the most important features that affect the cervical cancer



```
from xgboost import plot_importance
plot_importance(xgbModel)
```

After applying this algorithm, it results in the most important features like Age, First sexual intercourse, Hormonal Contraceptives(years), Schiller, Number of pregnancies, Hinselmann, Number of sexual partners, Citology, STDs: Time since first diagnosis, Smokes(packs/year), STDs, IUD, Hormonal Contraceptives, STDs: Time since last diagnosis are showing much impact on the deaths of cancer patients. The feature importance picture shows the same.



.

VII. CONCLUSIONS

Based on the results obtained from the research, the most influential attributes that cause cervical cancer deaths are Age, First sexual intercourse, Hormonal Contraceptives(years), Schiller, Number of pregnancies, Hinselmann and Number of sexual partners. Hence, it is obvious, if these parameters can be controlled well, then the cervical cancer deaths can be controlled to some extent. Upon using the XGBoost algorithm, the obtained accuracy is 94.19%. Similarly, we can state all other types of cancer deaths related factors and identify using the same classification technique. Also, we can take the necessary measures in order to reduce the death rate of cancer patients.

REFERENCES

[1] J. Subramanian, R. Govindan, “Lung Cancer in Never Smokers”, Journal of Clinical Oncology, Vol.25(5), pp.561–570, 2007.

[2] A. Sreedevi, R. Javed, A. Dinesh, “Epidemiology of cervical cancer with special focus on India”, International Journal of Women’s Health, Vol.7, pp.405-414, 2015.

[3] K. Fernandes, D. Chicco, J.S. Cardoso, J Fernandes, “Supervised deep learning embeddings for the prediction of a cervical cancer diagnosis”, PeerJ Computer Science, Vol.4(8), p.e154, 2018.

[4] E. Roura, X. Castellsague, M. Pawlita, N. Travier, T. Waterboer, N. Margall et al., “Smoking as a major risk factor for cervical cancer and pre-cancer: Results from the EPIC cohort”, International Journal of Cancer, Vol.135(2), pp. 453–466, 2014.

[5] V. Menon, D. Parikh, “Machine learning applied to Cervical Cancer Data”, International Journal of Scientific & Engineering Research, Vol. 9, Issue.7, pp.46-50, July-2018.

[6] S. Jayaprakash, E. Balamurugan, “A Comprehensive Survey on Data Preprocessing Methods in Web Usage Mining”, International Journal of Computer Science and Information Technologies, Vol.6 (3), pp. 3170-3174, 2015.

[7] T Chen, C Guestrin, “ XGBoost: A Scalable Tree Boosting System”, Proceedings of the 22nd ACM
SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD ’16, August 13-17, 2016, San Francisco, CA, USAc©2016 ACM, ISBN 978-1-4503-4232-2/16/08.

[8] S. Zhao, Y Guo, Q. Sheng, Y. Shyr**,** “Advanced Heat Map and Clustering Analysis Using Heatmap3”, BioMed Research International, Vol. 2014, pp.1-6, 2014.

[9] M. Fernandes, “Data Mining: A Comparative Study of its Various Techniques and its Process”, International Journal of Scientific Research in Computer Science and Engineering, Vol.5, Issue.1, pp.19-23, February(2017).

[10] S Mahajan, "Convergence of IT and Data Mining with other technologies ", International Journal of Scientific Research in Computer Science and Engineering, Vol.01, Issue.4, pp.31-37, 2013.

Authors Profile:

Mr V Naveen Babu is currently pursuing Bachelor of Technology (B.Tech) in Computer Science and Engineering from Vasireddy Venkatadri Institute of Technology, Guntur. His research interest includes Machine Learning. He has presently 4 months of experience in the IT industry as an Intern. His main research focuses on finding the most effective factors causing cancer deaths.

Mr T Murali is currently pursuing Bachelor of Technology (B.Tech) in Computer Science and Engineering from Vasireddy Venkatadri Institute of Technology, Guntur. His research interest includes Machine Learning. His main research focuses on finding the most effective factors causing cancer deaths.

Mr Sk Meer Hussain is currently pursuing Bachelor of Technology (B.Tech) in Computer Science and Engineering from Vasireddy Venkatadri Institute of Technology, Guntur. His research interest includes Machine Learning. His main research focuses on finding the most effective factors causing cancer deaths.

Mr S Nava Chaitanya is currently pursuing Bachelor of Technology (B.Tech) in Computer Science and Engineering from Vasireddy Venkatadri Institute of Technology, Guntur. His research interest includes Machine Learning. His main research focuses on finding the most effective factors causing cancer deaths.

Mrs.M Varalakshmi completed her Bachelor of Technology from Gudlavalleru Engineering College, Andhra Pradesh in 2004 and Master of Technology from Nimra Institute of Science and Technology, Andhra Pradesh. She is currently working as an Assistant Professor in the Department of Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Andhra Pradesh since 2017. Totally she has more than 10 years of experience in teaching. Her main research work focuses on Data Analytics and Machine Learning.