Cervical Cancer Prediction Using Machine Learning Algorithms

[1]S. Manisha Sree, [2]Dr. S. Logeswari & [3]P. Sanmati

[1,3]PG – Scholars, [2]Professor, Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India.

Abstract

The cervical malignancy is a sickness emerging in the cervix, different strains of Human Papilloma Virus (HPV), an explicitly transmitted disease, assume job in causing most cervical malignant growth. Cervix is the lower some portion of uterus that associates vagina. In the vast majority of the ladies, the invulnerable framework forestalls it yet in few, theinfection goes on for a considerable length of time. The explanation of passing overall is because of compelling access to cervical screening strategies is an extraordinary test. The early recognizable proof of this infection can be effectively treated. The screening tests incorporate Pap test and HPV DNA test. The fair informational index is gotten and it comprise of the time of patient, smokers or not, hormonal contraceptives, Intra Uterine Device (IUD) utilizations, Sexually Transmitted Diseases (STD) nearness and so on. This task mostly center on the forecast of cervical malignant growth of HVP test and PAP test utilizing AI calculations such a Support Vector Machine (SVM), strategic calculation, direct relapse and choice tree.

Keywords: Support Vector Machine, Intra Uterine Device, Human Papilloma Virus, choice tree.

1. Introduction

Malignant growth is a procedure of development of cells wild. Any zone of body could be influenced with a malignant growth. Cells influences by a malignancy could spread to different territories as well. Cervical disease influences the cells at the cervix, which is the lower zone of uterus of a female. Any information mining-based research work is powerful whenever finished with the reasonable information. Cervical cancer is definitely not a standard illness, which discovers the individuals discussing wherever with everybody. A couple of infection like cold, cerebral pain, fever is well known and for them individuals do not feel modest discussing the causes. Cervical cancer is a kind of an ailment which individuals feel cumbersome to talk about openly. In such circumstance, the information gathered from overviews, structure fillings, surveys, interviews are not dependable and helpful as the essential hotspot for the exploration work.

1.1 Causes of cervical cancer

The major factors that cause the cervical cancer are

- Infection of Human Papillomavirus (HPV)
- Improper sexual practices
- Smoking and
- Long term usage of birth controlling pills (oral contraceptives)

1.2 Symptoms of cervical cancer

The Cervical cancer does not show any symptoms in the early stages, which makes it hard to detect it. However, when the cancer grows, their symptom gets stronger. The most common symptoms are

- Abnormal bleeding of vagina
- Increasing vaginal discharge
- Bleeding after going through menopause
- Increase in pain while having sex
- Pelvic pain

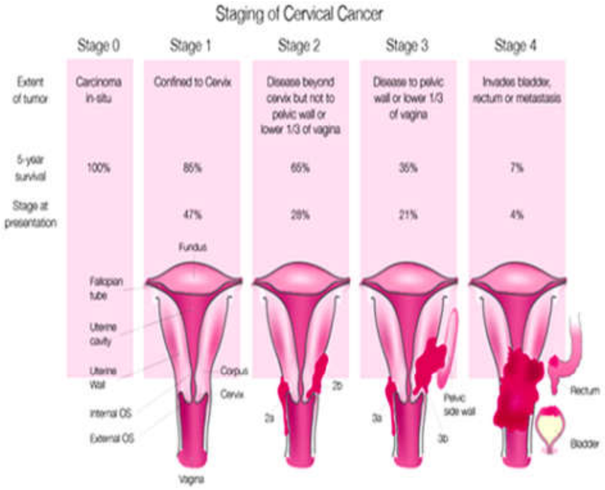
1.3 Objective

The main objective of this project is to find the best algorithm among several algorithms such as linear regression, support vector machine and logistic regression, based on the accuracy and performance efficiency for predicting the possibility of cervical cancer in women. These algorithms are compared based on recall or sensitivity, specificity, accuracy and precision score. An optimization algorithm – Gradient Boosting is applied on these algorithms, which increases the performance efficiency and accuracy.

1.4 Different stages of cervical cancer

The stages of the cancer are calculated based on the evaluation of the tumor and its size. There are totally four stages of cervical cancer.

- Stage-I
  - Stage-IA
    - Stage-IA1
    - Stage-IA2
    - Stage-IA3
  - Stage-IB
    - Stage-IB1
    - Stage-IB2
    - Stage-IB3
- Stage-II
  - Stage-IIA
    - Stage-IIA1
    - Stage-IIA2
  - Stage-IIB
- Stage-III
  - Stage-IIIA
  - Stage-IIIB
  - Stage-IIIC
    - Stage-IIIC1
    - Stage-IIIC2
- Stage-IV
  - Stage-IVA
  - Stage-IVB



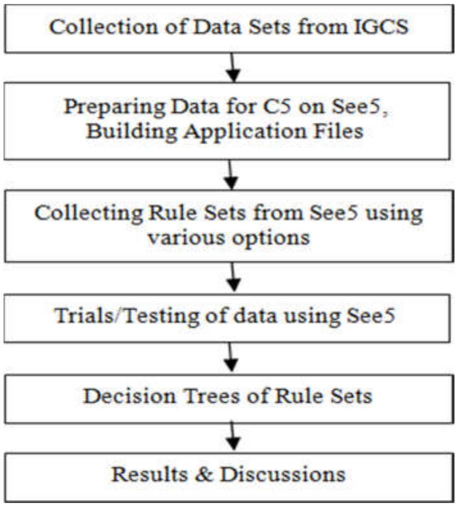

Collection of Data Sets from IGCS
Preparing Data for C5 on See5, Building Application Files
Collecting Rule Sets from See5 using various options
Trials/Testing of data using See5
Decision Trees of Rule Sets
Results & Discussions

Figure.1 Stages of cervical cancer

Figure.2 Performance life cycle

2. Methodologies

This project consists of several methods such as data collection, data preprocessing, model training, model testing, and result evaluation. In the testing phase, the accuracy, sensitivity, specificity and precision scores are calculated for each algorithm.

2.1 Data collection

The dataset is obtained from UCI repository. The dataset contains factors that play vital role in cervical cancer leading to biopsy test. The dataset used for this project contains 36 attributes. Each attribute contains 858 records.

2.2 Data preprocessing

The project requires certain fields, which determine the key factors for causing cervical cancer. The different credits used to recognize the nearness of cervical disease in the execution of this undertaking are as per the following

- Sexual accomplices
- Pregnancies
- Smokers
- Hormonal contraceptives
- IUD (Intrauterine contraceptive Device)

In this method, the null values are removed from the dataset and the unwanted attributes are dropped so that the model can train with better accuracy.

2.3 Training the models

The dataset is fed into the machine learning algorithms – Support Vector Machine, Linear Regression, Decision Tree and Logistic Regression. In these algorithms, the pattern of the attributes is identified and stored.

2.3.1 Decision Tree

Choice Trees [4] are a sort of Supervised Machine Learning where the information is being split based on certain conditions. Choice tree can be explained by two elements, in specific choice leaves and hubs. The leaves are the choices or the ultimate results. Choice Tree Analysis is a prescient demonstrating apparatus that has applications in various zones. Choice trees are developed through an algorithmic methodology that distinguishes approaches to part an informational collection dependent on various conditions. It is one of the most broadly utilized and reasonable techniques for directed learning [4]. Choice Trees are non-parametric regulated learning technique utilized for both order and relapse assignments. The objective is to develop a model that predicts the valuation of an objective variable by taking in upfront choice standards induced from the highlighted information. A choice tree is a tree-like diagram with hubs speaking to the edges and the qualities. The leaves speak to the genuine yield or class mark. They are utilized in non-straight dynamic [10] with straightforward direct choice surface. Choice trees group the models by arranging down the tree from the root to some leaf hub, with the leaf hub giving the arrangement to the model. Every hub in the tree goes about as an experiment for some trait, and each edge dropping from that hub compares to one of the potential responses to the experiment. This procedure is recursive in nature and is rehashed for each sub tree established at the new hubs.

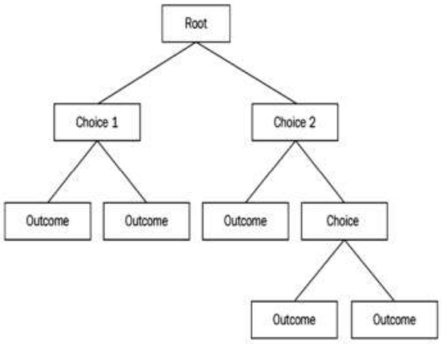


Figure.3 Choice tree structure

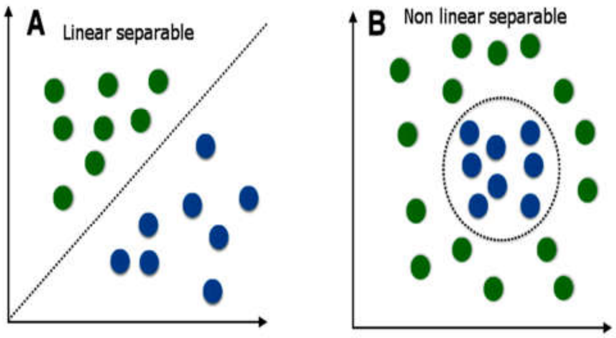
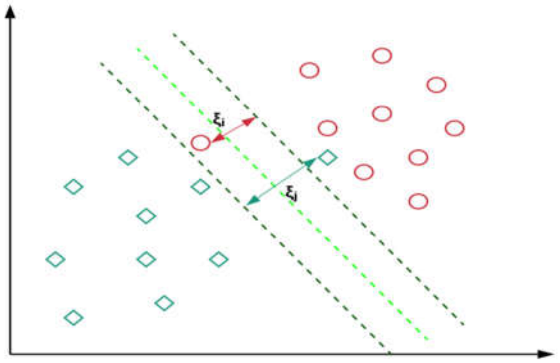


Figure.4 Types of support vector machine

2.3.2 Support vector machine

A Support Vector Machine (SVM) [1] orsupport vectornetworks is a discerning classifier officially characterized by hyper plane isolation. Ultimately, SVM model uses the given markedinformation (structured learning), the calculation of this model yields asupreme hyper plane, which orders new, and learned model. In two-dimensional spaces, the yielded hyper plane rifts the plane into two sections where each section holds a separate class. Bolster Vector Machine is a directed AI calculation, which is being utilized by the model for both arrangement and relapse. Despite that, it is for the most part utilized in categorization problems [5]. Right now, every datum is plotted in n-dimensional space as a point (wherenumber of highlights is ‘n’) with each element’s estimation being the estimation of specific expedite. At that point, accomplishthe directive by finding the hyper-plane that segregates the two types quite well. SVC works dependent on the accompanying classifiers.

- Linear SVM Classifier
- Non-Linear SVM Classifier



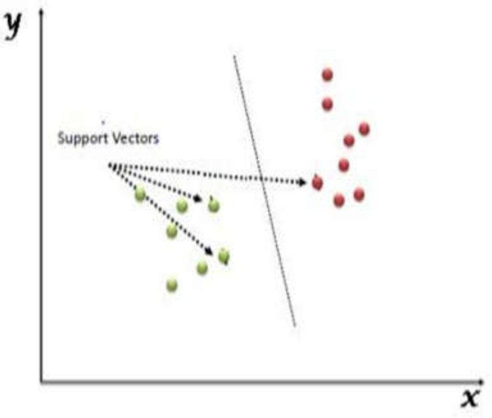


Figure.5 Structure of linear support vector machine

Figure.6 SVM Basic

2.3.3 Support vector linear classifier

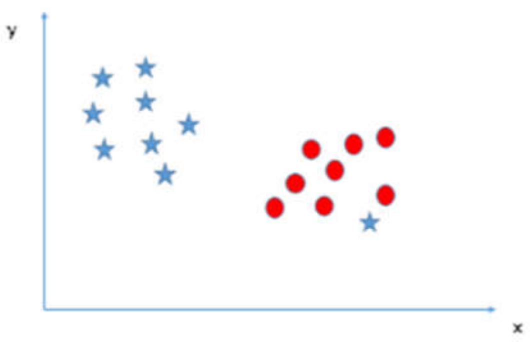
In the direct model, accept those preparation models plotted in space. The information focuses are relied upon to be isolated by a clear hole. It predicts a straight hyper plane partitioning two classes. The essential concentration while drawing the hyper plane is on expanding the good ways from hyper plane to the closest information purpose of either class. The drawn hyper plane called as a most extreme edge hyper plane.

2.3.4 Support vector non-linear classifier

In certainty, dataset is commonly scattered up somewhat. To face and solve this issue, segregation of data into different classes based on a conventional direct hyper plane, cannot be treated as a correctsolution. For this making Non-Linear Classifiers by applying the piece stunt to greatest edge hyper planes are utilized. In Non-Linear Support Vector Classification, focused informationgets strategized in a higher dimensional space. Figure.6. speaks to the essential chart

- Plot points (every datum thing) in N dimensional space.
- N speaks to Number of highlights.
- Hyper plane is line, which isolates the two classes.

Unfit to segregate the two classes by exploiting a straight line, as one of the stardeceits in the domain of other class (hoverclass) as an exclusion. One star at contrary end indicates an irregularity for star class.



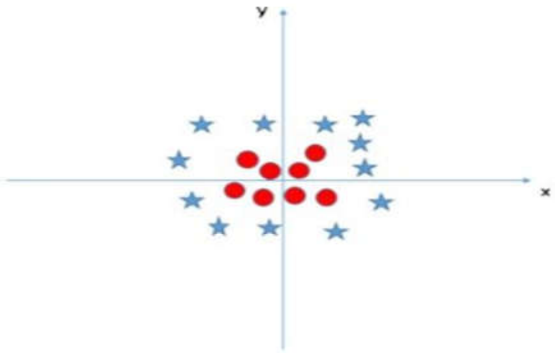


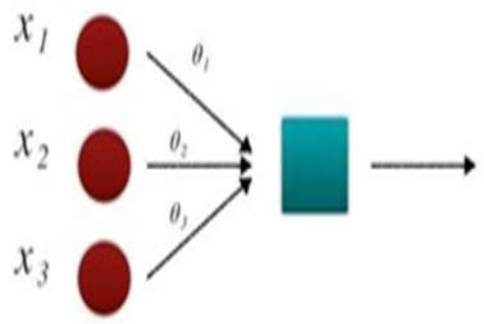
Figure.7 SVM Data focuses with diagram

Figure.8 Graph utilizing new element

Figure.7 speaks to the outliers has an element to overlook exceptions and identify the hyper-plane that has greatest advantage.Afterward, users can say, SVC is powerful to anomalies, it tells about the irregular cells present in the cervix. Figure.8 speaks to the above case can be overwhelmed by presenting new component Z. z=x^2+y^2. Plotting the focused information on hub x and z.

2.3.5 Calculated algorithm

Calculated relapse is a piece of a classification of statically models called speculation direct models. Strategic relapse permits forecast of a discrete result, for example, bunch participation, from set of factors that might be consistent, discrete, dichotomous, or a blend of any of these. Like all waninginvestigations, the strategic relapse is a discerninginvestigation Coordination deterioration is used to portray statistics and to simplify the linking between one ward matching variable and one increasingly ostensible, ordinal, short-term or part level free factor.



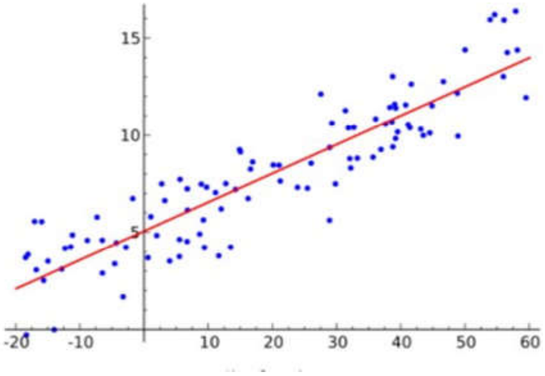


Figure.9 Logistic relapse model

Figure.10 Linear regression graph

It incorporates,

- Data assortments
- Analyzing information
- Data wrangling
- Train and test
- Accuracy check

2.3.6 Linear regression

Relapse is fundamentally a measurable methodology [8] to discover the connection between factors. In AI, this is utilized to anticipate the result of an occasion dependent on the connection between factors got from the informational collection. Straight relapse is one sort relapse utilized in Machine Learning. Basic direct relapse is a sort of waninginvestigation where the extent of autonomous aspects is one and there is a conservativelinking between the self-regulating(x) and reliant on(y) variable. The red line in the above figure is referred to as the finest fit straight line. In view of the given focused information, we try to plot a line that models focus the best. The line can be demonstrated based on the conservative condition demonstrated as follows.

3. Result evaluation

SVM was designed to solve the binary class and it gives better accuracy when compared to other Machine Learning algorithms. However, it takes more time to handle abnormal cells known as non – linear data points. The spectrum of the STDs was normalized using Kernel functions and Pap test can be easily identified using conventional supervised learning techniques.

Table.1 Comparison of machine learning algorithms

Algorithm	Decision tree	Logistic	Support Vector Machine	Linear
Accuracy in (%)	91	97	98	3.5

4. Conclusions

Cervical malignant growth is a typical ailment and its screening frequently includes tedious in clinical tests. In his point of view, AI can convey productive strategies to accelerate the determination methodology.

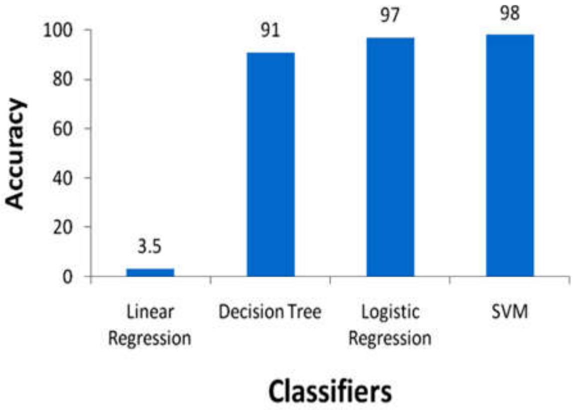


Figure.12 Comparison of machine learning algorithms

Numerous products apparatuses are accessible for examining the cervical disease, for example, SVM execution, choice tree, calculated calculation, straight relapse and so forth. Contrasting and different calculations SVMs are acceptable at finding the best direct separator. The portion stunt makes SVMs non-direct learning calculations. Picking a suitable bit is the key for good SVM and picking the correct portion work is not simple. Cervical patient, building SVMs on huge datasets. The developing assortment of cervical malignant growth patients (ladies) information and quickly propelling strategies for breaking down this information ready to distinguish best screening strategy for cervical disease patients that will be informatics for tolerant consideration. In future, this investigation can be utilized as a model to build up a social insurance for cervical disease patients and the future work is going to improve the effective calculation and to be performed.

References

1. Bora, Kangkana, et al, Automated classification of Pap smear images to detect cervical dysplasia, Computer methods and programs in biomedicine, 2017, 138, pp.31- 47.

2. Iliyasu, M.Abdullah, Chastine Fatichah, A Quantum Hybrid PSO Combined with Fuzzy k-NN Approach to Feature Selection and Cell Classification in Cervical Cancer Detection, Sensors, 2017, 17 (12), pp.2935.

3. Singh, Sanjay Kumar, Anjali Goyal, A Novel Approach to Segment Nucleus of Uterine Cervix Pap Smear Cells Using Watershed Segmentation, Advanced Informatics for Computing Research, Springer, Singapore, 2017, pp.164-174.

4. Kudva, Vidya, Keerthana Prasad, Shyamal Guruvare, Detection of specular reflection and segmentation of cervix region in uterine cervix images for cervical cancer screening, IRBM, 2017, 38 (05), pp.281-291.

5. Mukhopadhyay, Sabyasachi, et.al, Optical diagnosis of cervical cancer by intrinsic mode functions, Dynamics and Fluctuations in Biomedical Photonics XIV, 10063, International Society for Optics and Photonics, 2017.

6. Z.V.Mun, F.X.Osch, Ole.M.Jensen, Human Papillomavirus and Cervical Cancer, Lyon: International Agency for Research on Cancer, AJES, 2018, 07 (02).

7. V.Pushpalatha, S.Sathiamoorthy, M.Kamarasan Ranu Gorai, A Survey on Digital Image Processing, International Journal of Research in Engineering, Technology and Science, 2016, pp. 01- 07.

8. D.Selvathi, W.Rehan Sharmila, P.Shenbaga Sankari, Genetic Algorithms Techniques Based Computer Aided Diagnosis System for Cervical Cancer Detection Using Pap Smear Images", Classification in Bio Apps, Springer, Cham, 2018, pp.295-322.

9. Sudhir.B.Jagtap, B.G.Kodge, Census Data Mining and Data Analysis using WEKA, International Conference in Emerging Trends in Science, Technology and Management, 2013.

10. S.Archana, K.Elangovan, Survey of Classification Techniques in Data Mining, International Journal of Computer Science and Mobile Applications, 2014, 02(02).

11. A.Bharathi, E.Deepan kumar, Survey on Classification Techniques in Data Mining, International Journal on Recent and Innovation Trends in Computing and Communication, 2014, 02 (07).

12. Zan Huang, Hsinchun Chena, Chia-Jung Hsu, Wun-Hwa Chen, Soushan Wu, Credit rating analysis with support vector machines and neural networks: a market comparative study, Decision Support Systems (Elsevier), 2004, 37, pp.543–558.