ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

Cervical Cancer Cell Prediction using Machine Learning Classification Algorithms

Prianka R R*[1], Prof. Celine Kavida A[2], Bibin M R[3]

[1]*Assistant Professor, Department of CSE, RMK College of Engineering and Tech, Puduvoyal, Chennai, India*
[2]*Associate Professor, Department of Physics, Vel Tech Multi Tech Dr. RR Dr.SR Engg College, Avadi, Chennai*
[3]*Assistant Professor, Department of ECE, Vel Tech Multi Tech Dr. RR Dr.SR Engg College, Avadi, Chennai.*

Abstract— In cancer identification, extrapolation of cervical cancer in patients shows a vital role. To save people from the cancer field of cancer detection, machine learning can play a big role in saving lives. In this paper, to make the detection process a portion faster and accurate machine learning techniques such as Decision Stump, C4.5 and Averaged One Dependence Estimators (AODE) for novel NCBI cervical cancer data set are made. Classification and Regression Tree (CART) is a simple decision tree algorithm that is used to create a decision tree of a given set. Here, a top-down Greedy search is used in order to check each attribute at every tree node. For building a Decision Stump algorithm, a decision tree which consists of nodes and an arc that connects nodes with the Entropy concept is used. The extension of the basic Decision Stump algorithm is the C4.5 algorithm on selecting the optimal split it recursively visits each decision node. The process gets continued until there is no further split is possible. In this way, the prediction is possible for the given data set. Bayesian methods are those that explicitly apply Bayes' Theorem for problems such as classification algorithms acts as a quick method for the creation of a statistical predictive model. AODE are based on the Bayesian theorem which is commonly used to solve prediction problems for ease usage in the medical field. In this research Decision Stump, C4.5, and AODE are implemented with help of the training set. The basic designs are used to predict whether a feminine is having cervical cancer or not.

Keywords — *Prediction; Decision Stump; C4.5; AODE; Cervical Cancer.*

1. Introduction

Cancer is one of the most vital medical dangers caused by the swelling cell. Uncharacteristic development of tumor cells has always been a huge challenge to present day's technology. Chemotherapy using Gamma rays and Laser endoscopy has frequent risks occupied in them. These systems have a high probability of destroying healthy cells while treating the tumor cells. Cervical cancer is cancer that arises from the cervix [10]. It is caused due to the unusual growth of cells that can spread or invade other parts of the body. Usually, symptoms will not be visible and the progressed cervical cancer symptoms may include a typical vaginal exploiting, pelvic pain, pain during sexual association, loss of enthusiasm, weight loss, fatigue, pain, back pain, leg pain, swollen legs, heavy vaginal bleeding and bone fractures, leakage of urine or feces from the vagina.

The common sign of cervical cancer is bleeding after douching or pelvic exam. Human Papilloma Virus (HPV) causes 92% of infection among people and those infections can be prevented by HPV vaccines [1]. Cervical cancer screening with Pap smear or acetic caustic can make out the precancerous changes which can be treated to the knack development of cancer. Diagnosis is usually done by screening followed by a biopsy. Medical imaging is the user decides whether cancer has spread or not. World's cervical cancer is the fourth widespread cause of cancer cause of death in females [3].

Data mining is the process of extracting knowledge from large volumes of raw data. It is user ascertain knowledge out of facts and presenting it in an easily understandable to humans. As a result, the mining [1] technique has become accepted research. For medical researchers to spot and expect the outcome condition using the data set. In a decision tree, one internal node (the root) which is immediately connected to the terminal nodes (its leaves) determines and decision stump makes a prediction based on the value of just a single input feature. Attributes and continuous attributes are used to predict the disease in medical field. This paper analyzes the performance of learning techniques Decision Stump operator is used for generating a decision tree with only one single split. C4.5 and AODE algorithms are used in predicting the occurrence of cervical cancer. The block used for analyzing NCBI (National Center for Biotechnology Information) is termed as a microarray set. This research helps the physician to take a decision on the prognosis of cervical cell patients. At the end of the analysis, we identify the class proves better performance than Decision Stump and C4.5 algorithm.

2. Related Works

To analyze medical data sets Machine learning algorithms were designed to use from the very beginning. Several indispensable tools being provided by the machine learning algorithms are used for intelligent data analysis, [1].

DOI: **10.30726/esij/v8.i1.2021.81006**

The digital revolution in the last few years offered an economical and accessible means to gather and store the Monitoring and other data collection devices are equipped in the Modem hospitals; here data is gathered shared in large information systems. At present for medical data, Machine learning technology is well suited in particular there is a lot of work done in medical diagnose with little specialized diagnostic problems [2]. The classifier can then be used, either to aid the physician in diagnosing new patients in such a way to advance diagnostic speed, accuracy, constancy, or to train.

The Naive Bayesian classifier is the same as AODE algorithm works, where there are many input values, it can make sense to only use dependence estimators in those cases where interdependence is proven or at least suspected. The most widely used approach for this purpose is a decision tree. It was found to be efficient in other disciplines such as data mining, machine learning, and pattern cognition. In many real-world applications, Decision trees are also implemented.

The most popular construction approach is the Top-down construction of decision trees [4]. Machine learning has been of great benefit in many medical applications and can be used as a classifier in the early uncovering of the cancerous cells present in the cervix region the various existing techniques for the prediction of cervical cancer using medical data and points.

C4.5 is a Decision Stump uses the gain ratio as splitting criteria and the progression of this Decision Stump is C4.5. The splitting ceases when a number of instances to be spitted is below a certain threshold. Error based sniping is executed after the developing use. C4.5 is skilled in handling numeric attributes. It can keep a fit set that includes missing values corrected gain ratio criteria [8]. Data mining emerges in healthcare medicine pacts with knowledge models also gives importance about several diseases. All festivities involved in the healthcare industry can greatly benefit using the data mining applications Naive Bayesian is nothing but a simple probabilistic sanctifier, which is based on an assumption about the shared dependency of attributes. It was originally involved in predicting the heart rent later the thirteen attributes are reduced to 10 attributes see classifiers like Naive Bayes, Decision Tree, and Aging algorithm are used to predict the diagnosis of patients with the equivalent accuracy as achieved before. 10 doubling cross method was used to assess the unbiased estimate of computer science prediction models in our analysis [8].

3. Classification Methods

3.1 Decision Stump Algorithm

Decision Stump is a simple decision tree algorithm used to create one root tree of given values using a top-down Greedy search check for each attribute at every tree node. For building each Decision Stump's, the decision tree consists of nodes and arcs or sweeps connect nodes. In the Decision tree, the whole data hunted to create a tree that builds the fastest tree. Here, understandable prediction rules all created m the training data, and it builds a short tree with ID3, which alone using to test enough attributes until all data is classified. Are, finding leaf nodes enables test data to be pruned, finding several tests.

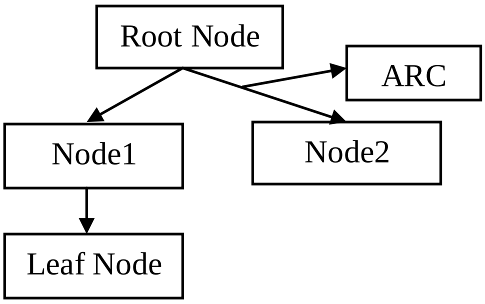


Fig. 1: Structure of Decision Tree Algorithm

Step 1: A leaf node corresponds to the expected number of the output attribute when the input attributes are described in the path from the root node to that leaf node.

Step 2: We can predict the output attribute in a "good" decision tree with its paths from the root node to that node.

Step 3: Entropy will determine how informative an input data is about the output data for a subset of the training data.

ID3: Machine learning algorithm Split (node, {examples}):

- A the best attribute for splitting the examples
- Decision attribute for this node A
- For each value of node A create a new child node
- Split training (e} to child nodes
- For each child node | subset:

If subset is pure: STOP else split(childnode,(subset})

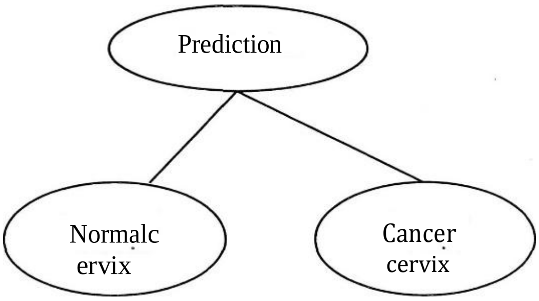


Fig. 2: Attributes of decision

Algorithm:

CART (Examples, Target-attribute, Attributes)

CART is also a predictive model that helps to find a variable based on other labeled variables. The attribute is various forms of the listed attribute that may be tested by the learned decision tree. Returns a decision tree - correctly classifies example.

DOI: 10.30726/esij/v8.i1.2021.81006

Create a Root node

If all examples are positive, then return single node tree Root, with label = +
If all are negative, returns single node tree Root, with a label = -
If Attributes are empty, return the single node tree Root with Common value of Target attribute.
Otherwise Begin
A ß the attribute that best* classifiers Examples
The decision attribute for Root, each possible value, vi, of A,
Add a new branch below Root, matching to the test A = vi
Let examples vi be the subset, that have value vi for A
If vi is empty

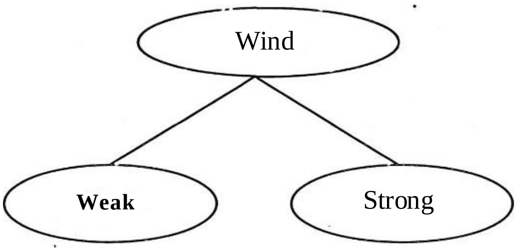


Fig. 3: Decision on Wind for example

Add a leaf node with a label most.
Common value on value of Target attribute in Examples
Else add sub-tree below this new branch
Decision Stump(Examples, argetattribute ,Attributes(A})End
Return Root

B. Entropy

Entropy is computed for each attribute in the database. It measures the impurity of training objects.

For a set S,
Entropy(S)= -p+log^2(P+)-P-log^2(p-)
Where,
P+ →is the proportion of positive examples
p→is the proportion of negative examples
S→data sample
The formula for entropy is,
Entropy(S)=$\sum_{i=1}^{n} p_i log^2 p_i$
P, → is the portion of S , that belonging to i

When entropy is 0 (zero) and when all members of 'S' belong to the same class and the system provides a prediction using the Decision Stump algorithm. The decision tree provides in this system is very clear and the rules provided by the system's easy.

The limitation of Decision Stump is (i) only one attribute at a time is tested for making a decision and is overly sensitive to features with large numbers of values. (ii) If the tiny part of testing, data might be over fitted or over classified. Classify continuous data may be computationally expensive as trees, and we must generate to see where to break to continue the system

3.2 C4.5 Algorithm

C4.5 is the extension of the ID3 algorithm. We include avoidance over-fitting the data reduced cv pruning, handling continuous attributes, and handling data; missing attribute values. C4.5 uses entropy and informed gain for tree splitting. In testing phase, we used training with known results of C4.5 algorithm to obtain the ruleset. In the testing phase, the classifications of various rules are applied to the whole preprocessed data.

In 1993 C4.5 was introduced by Ross Lsuinlan to overcome the limitation of ID3.C4.5used“Information gain to calculate,
GainRatio(piT) = Gain(p_2T) / splitInfo(p_2T)
WhereSplitInfo(p_1Text)=$\sum_p$ '$(j/p$)*log(p’ j/p)
p’(j/p) - is the proportion of elements present at the positionP1takingthevalueofjthtest.C4.5 also manages the cases of attributes with value› continuous intervals.

The pseudo code for Regression Tree

Input: Import DecisionTreeRegressor from sklearn.tree
Output: From sklearn.Tree import RecisionTreeRegressor test set

```
                                        # Instantiate dt
dt = DecisionTreeRegressor(max_depth=8,
min_samples_leaf=0.13,
random_state=3)
# Fit dt to the training set
            dt.fit(U_train, v_train)
# Predict test set labels
            v_pred = dt.predict(U_test)
# Compute mse
             mse = MSE(v_test, v_pred)
# Compute rmse_lr
R            mse = mse**(1/2)
# Print rmse_dt
Print('Regression Tree test set RMSE:
{:.2f}'.format(rmse_dt))
```

Classification Tree

Classify by that model which has minimum execution. Use classification done by the model which has maximum

Input: use a seed value for reusability
SEED = 1 .Import DecisionTreeClassifier from sklearntree from sklearn.tree import DecisionTreeClassifier

Output: documentation for Decision tree or try dt.get_params()
dt = DecisionTree Classifier (max_depth=6,
random_state=SEED)
# Fit dt to the training set

DOI: **10.30726/esij/v8.i1.2021.81006**

ISSN 2394-7187(Online)
ISSN 2394 - 7179 (Print)

```
dt.fit(U_train, v_train)
# Predict test set labels
v_pred = dt.predict(U_test)
# Import accuracy_score
From sklearn.metrics import accuracy_score
# Predict test set labels
v_pred = dt.predict(U_test)
# Compute test set accuracy
acc = accuracy_score(v_pred, v_test)
print("Test set accuracy: {:.2f}".format(acc))
```

Possibility to use continuous data.
Ability to use attributes with different weights.
Pruning the tree after being created.
Pessimistic prediction error.
Sub-tree raising.

4. Naive AODE Algorithm

The performance of ODE is well with a large number of training or input data items. However, all pairs of input values are measured in a combinatorial manner, it is difficult to use the AODE algorithm with high dimensional vectors (multiple input values for each data). So, it can use dependence estimators and interdependence at least suspected. For each discrete attribute, the possibility that the attribute X will take on the particular x when class C is modeled by a single real number between 0 and l.

AODE seeks the probability of each classy with specified set of features *u*1, ... un, P(v | u1, ... *u*n).

$$\hat{P}(v\,|u1\ldots un)=\frac{\sum_{i:1\le i\le n\wedge F(ui)\ge w}P\hat{}(v,ui)\prod_{j=1}^{n}P\hat{}(uj|v,ui)}{\sum v'\in V\sum_{i:1\le i\le n\wedge F(ui)\ge w}P\hat{}(v',ui)\prod_{j=1}^{n}\widehat{P}\hat{}(uj|v',ui)}$$

where $\hat{P}$ denotes an estimate of $\hat{P}$, F is the regularity with which the argument appears in the sample data and m is a user specified minimum frequency with a term must appear in order to use in the outer summation. The m is usually set at 1.

To estimate P (v | u1, ... un). By the definition of conditional probability mass function of U given as $u_1 \ldots u_n$ as follows.

$$P(v\ |u1\ldots un)=\frac{P(v,u1,\ldots un)}{P(u1,\ldots un)}$$

For any 1≤i≤n,

$$P(v,u1,\ldots un)=\ \ P(v,ui)P(u1,\ldots un|v,ui)$$

Under an assumption that u_1,... un are independent given v and u_i, it follows that

$$P(v,u1,\ldots un)=P(v,ui)\prod_{j=1}^{n}P(uj|v,ui)$$

This method defines One Dependence Estimator (ODE), a variant of the naive Bayes classifier that constructs the above independence assumption weaker and potentially less harmful than the naive Bayes' assumptions.

5. Methodology

Pronouncement trees are a very effective method of supervised learning. It aims to partition a dataset into groups as homogeneous as possible way of the variables to be predicted. It takes as input a set of confidential data and outputs a tree that resembles a positioning diagram where each end node (leaf) is a decision (a class) of each nontrivial node (Internal) represents a test. Each greenery represents the decision to the class data verifying all tests path from the root to the leaf.

5.1 Pruning

We are necessary to prune the tree in such a way to reduce the prediction error rate. Pruning is a way's technique in machine learning to resize trees by deleting sections of the tree that provide little power to classify instances. The dual goal of pruning is to reduce the problem will predictive accuracy by the drop of overfitting the removals of the section the classifier based on noisy or erroneous data.

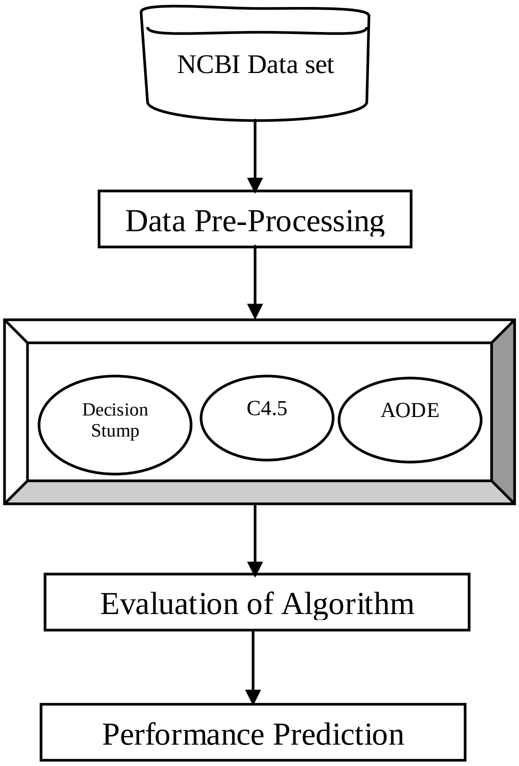


Fig. 4: Prediction Model

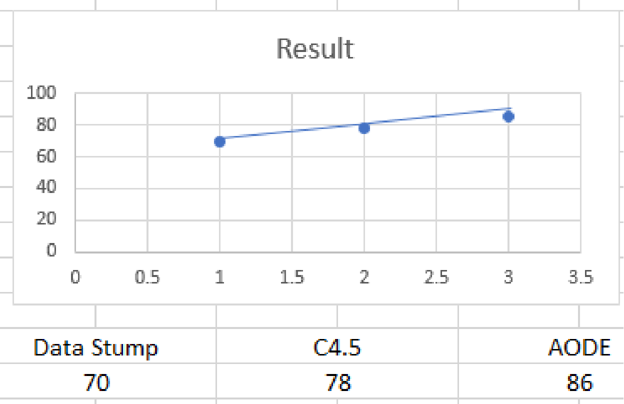
I Level: The Phase sampling process is with the database will be done with the NCBI data set.

DOI: **10.30726/esij/v8.i1.2021.81006**

II Level: Database will be used with implementation of Decision Stump algorithm and finding out accuracy result.
III Level: We will implement the concept with C algorithm and then NB algorithm to enhance the completion of prediction.
IV Level: Results from various scenarios will compared and an analysis will be fetched in with comparison with basic Data Mining Algorithms. The optimization of result is determined using AODE algorithm.

6. Experiment and Analysis

In this study, the models are evaluated based on performance measures of accuracy, sensitivity specificity. Expected outcomes to get out for ten-fold cross-validation to each model and averaged from the test each fold. Our whole experiments were done in MATLAB with Comparing accuracy, sensitivity, and specificity for the classifier was measured by using our cervical cancer NCE data set. The same training sets and test sets are utilized effectively in these experiments. We achieved results with NB (81%) (Naive Bayesian), Decision Stump (80%), C4.5 (78%), and the preferred method is Naive Bayesian.



Data Stump	C4.5	AODE
70	78	86

Fig. 5: Result Analyses

7. Conclusion

This analysis shows a feminine is having cervical cancer or not prediction to explain, compare, and assess the performances of unrelated machine learning techniques. Unequivocally, we need several trend concerns types of machine learning methods to be the type of training combined the kinds of endpoint guesses. The overall concert of these methods in forecasting shows predisposition or results. In Artificial Neural grind (ANNs) still, predominate. A growing alternate machine learning approach is being in that they are applied to many types of tumors to forecast three unlike kinds of outcomes. It is explicit machine learning methods that generally prerequisite the Concert or predictive precision of most prognoses; when compared to traditional statistical or exported systems. The act of AODE classifier and C4.5 analysis on NCBI data set in prediction cancer made. The common occurring of AODE is a high-level accuracy result than other classifiers. We consider that if the supremacy of studies continues to prove, it is likely that machine learning classifier develops into a much more public place and in many hospital usages.

References

[1] R.Vidhya, G.M.Nasira. " Predicting Cervical Cancer using Machine Learning Techniques- An Analysis" Global Journal of Pure and Applied Mathematics • ISSN 0973-1768 Volume 12, Number 3 (2016).

[2] Walboomers, Jan MM, et al. "Human papillomavirus is a necessary cause of invasive cervical cancer worldwide" The .Journal of pathology 189 1 (1999): 12-19.

[3] Clifford, G. M., et a1. "Human papillomavirus types in invasive cervical cancer worldwide: ameta-analysis" British journal of cancer 88.1 (2003): 63-73.

[4] Peto, Julian, et al."The cervical cancer epidemic that screening has prevented in the UK" The Lancet 364.9430 (2004): 249-256.

[5] Kononenko, Igor. "Machine learning for medical diagnosis: history, state of the art and perspective" Artificial Intelligence in medicine 23.1 (2001): 89-109.

[6] Quinlan, J. Ross. "Induction of decision trees" Machine learning I.I (1986): 81-106.

[7] Kononenko, Igor. "Inductive and Bayesian learning in medical diagnosis"Applied Artificial Intelligence an International Journal 7.4 11993): 317-337.

[8] Buntine, Wray. "Learning rules using Bayes" Proceedings of the sixth international workshop on Machine learning. 2014.

[9] Quinlan, J. Ross. C4. 5: prograirs for machine leaming. Elsevier, 2014.

[10] Dey, Monali, and SiddharthSwarajRautaray. "Study and Analysis of Data mining Algorithms for Healthcare Decision Support System" planning 5 (2014).

DOI: 10.30726/esij/v8.i1.2021.81006