Decision Tree approach in Machine Learning for Prediction of Cervical Cancer Stages using WEKA

Sunny Sharma¹, Sandeep Gupta²
¹,²Department of computer Science, Hindu College, Amritsar, Punjab

Abstract— Around the world cervical cancer or malignancy is the main motivation of cancer or tumor death in ladies. It impacts the cervix in the female regenerative framework which prompts death. The decision tree machine learning approach recognizes the phases of cervical disease. Decision tree arrange the phases of the cervical tumor in progressive basic leadership framework approach which manage the oncologist to take decision on phases of cervical disease, which safes human life. The proposed philosophy utilizes the examination information acquired from http://www.igcs.org and drives the prediction towards the phases of cervical cancer utilizing the tool Weka.

Keywords—Cervical Cancer-prediction, Weka, Machine Learning, C5, See5, SIPINA, Decision Tree

I. INTRODUCTION

The body is comprised of heaps of living cells. Ordinary body cells develop, parcel into new cells, and pass away in a systematic way. Tumor starts when cells in a part of the body start to grow out of control. Cancer cell development is differing from typical cell development. Rather than dying or pass away, disease cells proceed to develop and frame new, uncommon cells. Cells get to be tumor cells due to harm to DNA structure. We can say cancer or malignancy is one of the disorder in which the cells are apportioned and duplicated in uncontrolled way. Cervical malignancy is a standout amongst the most influencing tumors in ladies overall now nowadays. Its rate of event is around 85% in low and middle income nations or in low financial gatherings of nations and around 15% in higher wage nations or in developed countries. The primary issue with cervical growth is that it can't be recognized as it doesn't demonstrate any indications until the last stages typically [23] [25].

The machine learning is the strategy in which decision limits are investigated. The Rule base mining procedure and Decision trees assumes basic part for basic leadership and additionally supportive in machine learning. Another real favorable position of utilizing decision tree methodology is the white box nature of this methodology. It gets to be less demanding to investigate and fathom the course of basic leadership process.

II. DECISION TREES

Decision Tree portrays the various leveled choice methodology of the issue; with root node and other leaf or inner nodes. It has active outgoing edges instead of inward approaching edges. Internal nodes are called testing nodes and leaf nodes are called terminal nodes and in addition decision nodes here in test information sets the leaf nodes are the phases of cervical cancer/tumor. The capacity of Decision tree is that it can delineate the choice among various traits [26] [27].

III. INTRODUCTION TO WEKA AND OTHER MINING TOOLS

There are few data mining approaches or tools which have crucial problem to handle large databases but power of Weka is that it can deal with large databases easily and ascertain the hidden information in large databases. Weka is a workbench that contains an accumulation of perception apparatuses and
calculations for information examination and prescient demonstrating, together with graphical-UlS for simple access to these capacities. The unique adaptation was essentially outlined as an tool for dissecting information from agricultural domains, however the later completely Java-based variant (Weka 3) for which advancement began in 1997, it is presently utilized as a part of various application ranges, specifically for instructive purposes and research. Favorable circumstances of Weka included as:

- Free accessibility under the GNU-General Public License.
- Versatility or portability, since it is completely executed in the Java programming dialect and consequently keeps running on any modern cutting edge processing or computing stage.
- A far reaching gathering of information pre-processing and displaying modeling procedures.
- Convenience because of its graphical-UlS.

Weka underpins a few standard information mining undertakings, all the more particularly, information pre-processing, grouping or clustering, order or classification, regression, representation, and highlight choice for feature selection. The greater part of Weka's methods are predicated on the suspicion that the information is accessible as one level document or connection, where every information point is depicted by a settled number of properties. Weka gives access to SQL databases utilizing Java Database Connectivity and can prepare the outcome returned by a database question. It is not fit for multi-social information mining, but rather there is separate programming for changing over an accumulation of connected database tables into a solitary table that is reasonable for handling utilizing Weka. Another imperative range that is as of now not secured by the calculations incorporated into the Weka appropriation is arrangement displaying [21]

There are few other data mining tools are present like See5, SIPINA etc. They have data mining capability as well as machine learning capabilities. SIPINA especially intended to decision trees induction or to do classification, using supervised learning. SIPINA is incorporated with dedicated classification trees algorithms like GID3, ID3, ASSISTANT 86, C4.5, Improved CHAID, One Vs All Decision Tree etc. it also has some other mining capability through Rule Induction, Neural Network, Discriminant Analysis, Decision List etc. we can use any one of them.

Experiment is done on cervical cancer data bank using weka. Small sample of data set of 237 contains seven stages and 12 attributes. [17]

VI. ALGORITHM IMPLEMENTATION USING WEKA

Weka incorporated with the dedicated various classifiers techniques like Bayes, Functions, Lazy, Meta, Rules and Trees etc. It contains various algorithms like Naive Bayes, SMO, LWL, Iterative Classifier Optimizer, Decision Table, Decision Stump, J48 etc.

The other data mining Tools like See5 and SIPINA use Quinlan’s C5.0 algorithm is widely used for classification process. Algorithm primarily focuses on constructing a decision with the identification of most important attributes from the supplied/identified data-set. There are few options present over See5 tool like, Boosting is to generate several classifiers (decision trees or rule-sets) instead of one. [18], [19] other option like Winnowing is a mechanism to separate the useful attributes from useless attributes. It provides option to select among the predictors and have an edge to create a suitable decision-tree. However, it’s time intensive task and primarily suitable for bigger application domain. [18], [20]. In Advanced pruning technique a massive tree is first allowed to grow to fit the data closely after that it’s pruned i.e. error causing segments are removed. Every sub-tree undergoes pruning then replacement by a leaf or sub branch is decided and then a global stage evaluates performance of the tree as a unit. [18], [20].
Here the experiment is done through weka and produced the classification of the cervical stages the various algorithms are applied on the data set which can be explained as bayes based algorithm Naive Bayes (33.3333%), functions based algorithm SMO (35.865 %), lazy based algorithm LWL (34.5992), meta based algorithm Iterative Classifier Optimizer (37.9747%), rules based algorithm Decision Table (29.1139%), trees based algorithm Decision Stump (37.9747%). The overall accuracies achieved by these algorithms are described with them in bracket with them. It analysed that Iterative Classifier Optimizer and Decision Stump outperform among them but Decision Table performed least among them.

VII. STAGES OF CERVICAL CANCER

1) Stage 0-Carcinoma in Situ:
Stage 0 is carcinoma in situ i.e origin of abnormal cells in the inner-most lining of the cervix. They become cancer & affect nearby customary tissue.

2) Stage I
At this stage presence of cancer is in cervix only.
- Stage I-A: with the help of microscope cancer can be seen in cervix tissues. Further detail is expressed as.
  - Stage I-A (1): cancer depth is not more than 3 millimeters and it is not greater than 7 millimeters wide in tissues.
  - Stage I-A (2): cancer depth is greater than 3 but not greater than 5 millimeters, but not more than 7 millimeters wide.
- Stage I-B detail is expressed as.
  - Stage I-B (1): the depth is greater than 5 millimeters & greater than 7 millimeters wide.
  - Stage I-B (2): the not more than 4 centimeter wider cancer can be seen without the help of microscope.

3) Stage II
At this stage cancer has spread beyond the cervix but not towards the pelvic wall or towards the inferior third of the vagina. With deep penetration of cancer further detail can be expressed as.
- Stage II-A: Cancer increased from cervix towards the superior two third of the vagina, but not towards tissues of the uterus.
  - Stage II-A (1): without the help of microscope tumor of not more than 4 centimeters can be seen.
  - Stage II-A (2): without the help of microscope tumor of more than 4 centimeters can be seen.
- Stage IIB: Cancer stretched from cervix towards the tissues of the uterus.

4) Stage III
At this stage cancer stretched towards the inferior third of the vagina, Pelvic wall can cause kidney troubles. With deep penetration of cancer further detail can be expressed as.
- Stage III-A: Cancer has stretched towards the inferior third of the vagina but not towards pelvic wall.
- Stage III-B: Cancer has stretched towards the pelvic wall, tumor has become huge enough to chunk the ureters which increase the size of kidneys or can stop kidneys working.

5) Stage IV
At this stage cancer stretched towards the bladder/rectum, or other parts of the body.
- Stage IV-A: Cancer stretched towards nearly connected organs like rectum/bladder.
- Stage IV-B: Cancer stretched towards other parts of the body like liver/lungs/bones/distant lymph nodes [3]
Common types of treatments for cervical cancer are as follow: Surgery/Radiation therapy/Chemotherapy & combination of them

VIII. LITERATURE SURVEY

In 2012 M. Singh et al. use the decision trees to predict the protein classes through see5 tool and express the importance of decision trees in bioinformatics. In 2014 S. Sharma et al. describe the importance of evolutionary multiobjective optimization algorithms in bioinformatics [15]. In 2009 A. Satija et al. revealed the statistics of cervical cancer in India. It is primarily caused by human papillomavirus (HPV) infection with the vaccination much progress has been made in the prevention and control of cervical cancer [2]. In 2011 C. Balleyguier et al. proposed the guidelines for staging & follow up of patients which suffered from uterine cervical cancer & provides the radiologists with a framework & expressed the importance towards adequate patient preparation, protocol optimization and MRI reporting expertise are essential to achieve high diagnosis accuracy [11]. In 2009 the C. Todd et al. proposed computer assisted algorithm for the classification of cervical cancer using digitized histology images of biopsies. Texture analysis of the nuclei structure is very important for the classification of cervical cancer histology [6]. In 2009 G. Jayalalitha et al. discussed technique of grading cervical cancer images according to the cell formation of tissues. They expressed the use of Box Counting Method (DB) and Harfa Programme software to detect the fractional dimension and calculate the variation of intensity and texture complexity of cancer cell images [4]. In 2010 S. Allwin et al. proposes approach which use textural properties to classify the various malignancies in cervical cyto images [10]. In 2010 M. Ross et al. explore the focus towards the use of histology images for the classification of cervical cancer [8]. In 2016 S. Sharma et al. proposed methodology which classifies the stages of cervical cancer using See5 and SIPINA. The methodology used for the research experiment is expressed as [5] [7] [9].

IX. RESULTS AND DISCUSSIONS

The contribution of features like NodePET, ClinDiameter, MRIVol, UterineBody, Status, RelPrimary, Relpelvic, RelAbdo, RelSupraclav, RelDistant is obtained and shown in the table. Which describe that RelAbdo, RelDistant, Status, Histology & RelPrimary plays vital role in decision making and to identify the stages of cervical cancer. It depicted in the Figure 1.
The suggested attribute for decision tree by Weka is MRIVol and the suggested value for the decision node is 48.7 which expressed in decision tree. The total 237 data set cases are used and the contribution to identify the cervical cancer stages using MRI attribute is expressed in Figure 2.

The detailed accuracy achieved using Decisionn Stump classifier under tress is expressed in the Figure3. The error generated during the process is also expressed with the average accuracy is also expressed as in Figure 3.
Figure 3: Classified Instances with Error for Cervical Cancer Stages

The various analysis like TP Rate, FP Rate, Precision, Recall, F-Measure, MCC, Roc Area, PRC Area for prediction of stages with average weighted values are expressed in Figure 4.

<table>
<thead>
<tr>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.328</td>
<td>0.004</td>
<td>Stage-1a</td>
</tr>
<tr>
<td>0.493</td>
<td>0.298</td>
<td>0.405</td>
<td>0.493</td>
<td>0.444</td>
<td>0.185</td>
<td>0.573</td>
<td>0.339</td>
<td>Stage-1b</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.520</td>
<td>0.153</td>
<td>Stage-2a</td>
</tr>
<tr>
<td>0.700</td>
<td>0.618</td>
<td>0.366</td>
<td>0.700</td>
<td>0.481</td>
<td>0.081</td>
<td>0.509</td>
<td>0.339</td>
<td>Stage-2b</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.384</td>
<td>0.033</td>
<td>Stage-3a</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.658</td>
<td>0.255</td>
<td>Stage-3b</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.233</td>
<td>0.004</td>
<td>Stage-4a</td>
</tr>
<tr>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>?</td>
<td>?</td>
<td>Stage-4b</td>
</tr>
<tr>
<td>0.380</td>
<td>0.295</td>
<td>0.241</td>
<td>0.380</td>
<td>0.292</td>
<td>0.081</td>
<td>0.550</td>
<td>0.283</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: Contribution of various features in cervical cancer stages

The cost benefit analysis using decision stump classifier for true positive cases to predict cervical cancer stages is expressed in Figure 5 and Figure 6. Which expressed maximum accuracy of 85.65% but average accuracy to predict cancer stages as 37.98%.
Figure 5: Cost Benefit Analysis using Decision stump

Figure 6: Cost Benefit Analysis using Decision stump

The Correlation for Attribute Evaluation is also done using Rankers method which expressed the correlation Attribute selection using 10 fold cross-validations (stratified) as
The confusion matrix for Decision stump classifier on experimental data sets is expressed in Figure 7.

\[
\begin{array}{cccccccc}
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
3 & 0 & 35 & 0 & 0 & 0 & 0 & 0 \\
18 & 0 & 16 & 0 & 0 & 0 & 0 & 0 \\
24 & 0 & 56 & 0 & 0 & 0 & 0 & 0 \\
3 & 0 & 6 & 0 & 0 & 0 & 0 & 0 \\
4 & 0 & 39 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
\end{array}
\]

\text{<-- classified as}

\text{a = Stage-1a}

\text{b = Stage-1b}

\text{c = Stage-2a}

\text{d = Stage-2b}

\text{e = Stage-3a}

\text{f = Stage-3b}

\text{g = Stage-4a}

\text{h = Stage-4b}

\text{Figure 7: Confusion Matrix}

Based on this scoring as well as on confusion matrix the accuracy on Weka is determined and produced the classification of the cervical stages. The various algorithms are applied on the data set like based on bayes technique algorithm is Naive Bayes (33.333%), functions based algorithm SMO (35.865%), lazy based algorithm LWL (34.5992%), meta based algorithm Iterative Classifier Optimizer (37.9747%), rules based algorithm Decision Table (29.1139%), trees based algorithm Decision Stump (37.9747%). The overall accuracies achieved by these algorithms are described with them in bracket with them. It describes that Iterative Classifier Optimizer and Decision Stump outperform among them but Decision Table performed least among them. The accuracy chart is shown in Figure 8 and Decision tree obtained from Weka classifier Decision Stump is shown in Figure 8.

\text{Figure 8: Shows the Maximum accuracies obtained}
X. CONCLUSION

For treatment of cancer patients or in diagnostic process large amount of clinical information is processed. Oncologist can use decision tree based computerized technique to diagnose their patient. The proposed methodology classifies the stages of cervical cancer using Weka.

REFERENCES

Figure 8: Decision Tree obtained from Decision Stump using Weka