Predictive Analysis for the Detection of Human Diseases CVD, CKD, DM Based on Supervised and Ensemble Machine Learning Classification Algorithms

Dillip Narayan Sahu¹, Satheesh Kumar²
¹Research Scholar, Department of Computer Science and Application, OPJS University, Churu, Rajasthan, India
²Professor, Department of Computer Science and Engineering, Chaitanya (Deemed to be University), Telangana, India

ABSTRACT
Because of the high risk globally in the health care sector, the Chronic kidney disease (CKD), Cardio Vascular Disease (CVD), Diabetes Mellitus (DM) are the major burden because of its increasing pervasiveness. Cardio Vascular Disease (CVD), Chronic Kidney Disease (CKD) and Diabetes Mellitus are from the most active disease and the leading causes of death worldwide in the health care sector. Machine learning is playing an essential role in the medical side. In this paper, ensemble learning methods are used to enhance the performance of predicting heart disease, kidney disease and also diabetes disease. In this paper, we have shown some real time analysis by the help of supervised and ensemble machine learning classification algorithms. We have found the accuracy rate of approx. 90% in the early stage of prediction of disease, which is much better from the previous research papers.

Keywords- Algorithms, Human Disease, Machine Learning, Predictive Analysis.

1. Introduction
Machine Learning is a subset of Artificial Intelligence where we create machines which learn from the experience. Basically there are 3 types of machine learning named as supervised, unsupervised, and reinforcement learning[1][2]. Machine learning consist of various algorithms using which we can predict for the disease, but the input should be acceptable by the machine learning algorithm[3][4].

In this paper, we have taken 3 different diseases i.e. Chronic kidney disease (CKD), Cardio Vascular Disease (CVD), and Diabetes Mellitus (DM) and by applying different algorithms, we will see the accuracy rate of disease prediction comparison with others predictive analysis. In this analysis we have taken weka as a knowledge analysis tool to be the input as well as for the corresponding output for comparison purpose.

We have taken 3 different dataset in this paper i.e. CVD.arff, CKD.arff and DM.arff. and analyze based on different classification algorithms.

Two features of the extraction method: linear discriminant analysis (LDA) and principal component analysis (PCA) are used to select important features from the dataset. A comparison of machine learning algorithms and ensemble learning techniques is applied to selected features[5][6]. Various methods such as accuracy, recall, accuracy, F-measures, and ROC are used to evaluate models. The results show that the bagged ensemble learning method using decision trees performed the best.

2. Experiments and Observations
We have taken 3 different dataset named CVD, CKD and Diabetes in CSV file format, and also in arff format. We have used Weka as a tool for the classification of different algorithms and experimental and observations purpose.
Classification Algorithm: Naïve Bayes  
Classifier Output: Run information  
Scheme: weka.classifiers.bayes.NaiveBayes  
Relation: pima_diabetes  
Instances: 768  
Attributes: 9  
  
preg  plas  pres  skin  insu  mass  
pedi  age  class  
Test mode: 10-fold cross-validation  
Naive Bayes Classifier  

Class  
Attribute  | tested_negative | tested_positive  
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.65)</td>
<td>(0.35)</td>
<td></td>
</tr>
</tbody>
</table>

| Attribute | tested_negative | tested_positive  
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>preg</td>
<td>3.4234</td>
<td>4.9795</td>
</tr>
<tr>
<td>std. dev.</td>
<td>3.0166</td>
<td>3.6827</td>
</tr>
<tr>
<td>weight sum</td>
<td>500</td>
<td>268</td>
</tr>
<tr>
<td>precision</td>
<td>1.0625</td>
<td>1.0625</td>
</tr>
</tbody>
</table>

Time taken to build model: 0 seconds  

Correctly Classified Instances 586 76.3021 %  
Incorrectly Classified Instances 182 23.6979 %  
Kappa statistic 0.4664  
Mean absolute error 0.2841  
Root mean squared error 0.4168  
Relative absolute error 62.5028 %  
Root relative squared error 87.4349 %  
Total Number of Instances 768  

TP Rate  FP Rate  Precision  Recall  F-Measure  MCC  ROC Area  PRC Area  Class  
0.844  0.388  0.802  0.844  0.823  0.468  0.819  0.892  tested_negative  
0.612  0.156  0.678  0.612  0.643  0.468  0.819  0.671  tested_positive  

Weighted Avg. 0.763  0.307  0.759  0.763  0.760  0.468  0.819  0.815
Confusion Matrix:

\[
\begin{array}{cc}
a & b \\
422 & 78 & a = \text{tested\_negative} \\
104 & 164 & b = \text{tested\_positive} \\
\end{array}
\]

---

Fig.2 Naïve Bayes Algorithm Accuracy with Threshold curve

Classification Algorithm: Bagging
Classifier Output: Run information
Scheme:
weka.classifiers.meta.Bagging -P 100 -S 1 -num-slots 1 -I 10 -W
weka.classifiers.trees.REPTree -- -M 2 -V 0.001 -N 3 -S 1 -L -I 1.00
Relation: pima_diabetes
Instances: 768
--- Classifier model (full training set) ---
Bagging with 10 iterations and base learner
weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -I 1.00
Time taken to build model: 0.08 seconds
--- Stratified cross-validation ====== Summary ====

Correctly Classified Instances 582 75.7813 %
Incorrectly Classified Instances 186 24.2188 %
Kappa statistic 0.4498
Mean absolute error 0.315
Root mean squared error 0.4063
Relative absolute error 69.3049 %
Root relative squared error 85.2474 %
Total Number of Instances 768
--- Detailed Accuracy By Class ====

<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>0.850</td>
<td>0.414</td>
<td>0.793</td>
<td>0.850</td>
<td>0.820</td>
<td>0.452</td>
<td>0.812</td>
<td>0.879</td>
<td>tested_negative</td>
</tr>
<tr>
<td>0.586</td>
<td>0.150</td>
<td>0.677</td>
<td>0.586</td>
<td>0.628</td>
<td>0.452</td>
<td>0.812</td>
<td>0.676</td>
<td>tested_positive</td>
</tr>
</tbody>
</table>

Weighted Avg. 0.758 0.322 0.752 0.758 0.753 0.452 0.812 0.808

--- Confusion Matrix ====

\[
\begin{array}{cc}
a & b \\
425 & 75 & a = \text{tested\_negative} \\
\end{array}
\]
111 157 | b = tested_positive

Classification Algorithm: Random Forest
Classifier Output: Run information
Scheme: `weka.classifiers.trees.RandomForest -P 100 -I 100 -num-slots 1 -K 0 -M 1.0 -V 0.001 -S 1`
Relation: `pima_diabetes`
Instances: 768
Attributes: 9
Test mode: 10-fold cross-validation
 Classifier model (full training set)
RandomForest
Bagging with 100 iterations and base learner
weka.classifiers.trees.RandomTree -K 0 -M 1.0 -V 0.001 -S 1 -do-not-check-capabilities
Time taken to build model: 0.21 seconds
 Stratified cross-validation Summary
Correctly Classified Instances 582 75.7813 %
Incorrectly Classified Instances 186 24.2188 %
Kappa statistic 0.4566
Mean absolute error 0.3106
Root mean squared error 0.4031
Relative absolute error 68.3405 %
Root relative squared error 84.5604 %
Total Number of Instances 768
Detailed Accuracy By Class

<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>0.836</td>
<td>0.388</td>
<td>0.801</td>
<td>0.836</td>
<td>0.818</td>
<td>0.458</td>
<td>0.820</td>
<td>0.886</td>
<td>tested_negative</td>
</tr>
<tr>
<td>0.612</td>
<td>0.164</td>
<td>0.667</td>
<td>0.612</td>
<td>0.638</td>
<td>0.458</td>
<td>0.820</td>
<td>0.679</td>
<td>tested_positive</td>
</tr>
</tbody>
</table>

Weighted Avg. 0.758 0.310 0.754 0.758 0.755 0.458 0.820 0.814

Confusion Matrix

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>418</td>
<td>82</td>
</tr>
<tr>
<td>104</td>
<td>164</td>
</tr>
</tbody>
</table>

Fig.3 meta. Bagging Algorithm Accuracy with Threshold curve
3. Conclusion

We have used five different machine learning classification algorithms for the analysis on the dataset and based our observations on the acceptance of certain domains of machine learning models. After examining the above real-time medical record implementation and various observations, we found the level of accuracy using the Bagging and some meta machine learning classification model to be very satisfactory, with an excellent accuracy rate of 89.62%. This will may be opt in the branch of medicine for predicting early diagnosis of heart, kidney, and Diabetes disease. Five different experimental observations were made using machine learning tools to unambiguously analyze, detect, and predict these diseases. Examining the above experimental observations, machine learning tools are undoubtedly an excellent method for predicting and detecting these diseases (cardiac, kidney, and also diabetes) at an early stage. Accuracy levels using various algorithms in machine learning have proven to be good options for these diseases, the detection and prediction, and are highly accurate, efficient and acceptable.

References