Prediction of Heart Disease using Forest Algorithm over Linear Regression Algorithm using Machine Learning With Improved Accuracy

K N S Shanmukha Raj, K Thinakaran

Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602105.

Abstract

**Aim:** To perform Predicting heart disease using the Forest algorithm and comparing its feature extraction precision with the Linear Regression Algorithm for improving the accuracy of the prediction. **Methods and Materials:** In the proposed work, Predicting heart disease was carried out using machine learning algorithms such as Linear Regression (n=10)and Forest algorithm(n=10). Here the pretest power analysis was carried out with 80% and the sample size for the two groups are 20. **Results:** From The implemented experiment, the Forest algorithm accuracy is 90.32% and the Linear Regression Algorithm 77.21%. There is a statistical 2-tailed significant difference in accuracy for two algorithms is 0.001 (p<0.05) **Conclusion:** This study concludes that the Forest algorithm on patients healthcare analysis is significantly better than the Linear Regression Algorithm.

**Keywords:**
Forest Algorithm, Linear Regression Algorithm, Predicting Heart Disease, Machine Learning, Supervised Classification, Novel Dimensionality Reduction.

INTRODUCTION

Prediction of heart disease is a challenging task in medical field so as to treat the patients in advance this research is useful. According to a medical survey, the population of elderly people is increasing rapidly and they need continuous care for their treatments (Ali et al. 2021). Prediction of heart disease helps chronically ill patients to progress their medical treatment and to save their life in advance (Jmaiel et al. 2020). Here the system uses 14 parameters for predicting heart disease that includes blood pressure, cholesterol, chest pain and heart rate. These parameters are used to improve an accuracy level for predicting heart disease and to predict if a person is affected with Heart disease or not by utilizing novel principal component analysis (Swain, Pattnaik, and Gupta 2020). Novel Principal component analysis decreases the data set by investigating the presentation measurements like accuracy, Precision and F-score (Kim and Geem 2015)). The enormous arrangement of factors into a more modest one that actually contains the greater part of the data in the huge set (Swain, Pattnaik, and Gupta 2020). This project application is to demonstrate better outcomes and help the area specialists and surprisingly the individual connected with the clinical field to get ready for a superior and early determination for the patient. This framework performs practically well even without retraining. Moreover, the test results show that the framework predicts coronary illness with ~100% exactness by utilizing neural organizations.

In this research of heart disease prediction using supervised classification, number of papers published in the research gate are 1140 and the number of papers published in google scholar are 1201 (Shahid and Rahman 2014). Machine learning is creating advancements attributable to its algorithms used for creating predictions using supervised classification with Principal Component Analysis using Supervised Classification. In calculating effort by any of the higher than mentioned techniques, requirement of massive data-sets with most info concerning a research. This analysis makes use of machine learning to predict the hassle of knowing all the certain parameters. Structural features were a new approach used on 14 parameters for predicting heart disease that includes blood pressure, cholesterol, chest pain and heart rate. These parameters are used to improve an accuracy level for predict-
ing heart disease to predict if a person is affected with Heart disease or not (Borra et al. 2019). This proposed method has an accuracy of 85.00% by using Decision tree with the mentioned approach (Bansal et al. 2018). The data set is pre-processed with Feature Scaling and Missing Values. The classifiers like Forest Algorithm, linear Regression are fitted to the raw data set. Using the proposed system the accuracy is 85.00% and 80.00% (Giudici 2005). The best article since it compares various techniques to compute the heart disease prediction with Supervised Classification(Shahid and Rahman 2014; IEEE Staff 2019)). Our team has extensive knowledge and research experience that has translate into high quality publications (Chellapa et al. 2020; Lavanya, Kannan, and Arivalagan 2021; Raj R, D, and S 2020; Shilpa-Jain et al. 2021; S, R, and P 2021; Ramadoss, Padmanaban, and Subramanian 2022; Wu et al. 2020; Kalidoss, Umapathy, and Rani Thirunavukkarasu 2021; Kaja et al. 2020; Antink et al. 2020; Paul et al. 2020; Malaikolundhan et al. 2020)(Chellapa et al. 2020; Lavanya, Kannan, and Arivalagan 2021; Raj R, D, and S 2020; Shilpa-Jain et al. 2021; S, R, and P 2021; Ramadoss, Padmanaban, and Subramanian 2022; Wu et al. 2020; Kalidoss, Umapathy, and Rani Thirunavukkarasu 2021; Kaja et al. 2020; Antink et al. 2020; Paul et al. 2020; Malaikolundhan et al. 2020).

Therefore in the existing system the accuracy percentage is very low in predicting heart disease with health care analysis((Dinesh Peter, Fernandes, and Thomaz 2019)). It may be difficult to predict heart disease accurately and precisely. So it is important to add more values to the data set and trained data set to predict accurately. As an author, the machine learning based heart disease prediction solution has been carried out. This research is to predict heart disease using Forest algorithm and Linear Regression algorithms as Supervised Classification.

**MATERIALS AND METHODS**

Experiment was conducted in Artificial Intelligence Laboratory, Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences. In this research two groups were taken, one group refers to Forest Algorithm and other group refers to Linear Regression Algorithm. Proposed system has a total of 20 samples iterated for two algorithms. The application of algorithms for Forest Algorithm and Linear Regression Algorithm with the Sample size is 304 sample data for each algorithm and pretest power (GPower) obtained is 80% and alpha value is 0.143 (Bansal et al. 2018).

Dataset was collected from Kaggle((Baig 2020). Total 304 patient's data was collected. Whole the Data set is divided into two sets: Testing Set and Training Set(Kim and Geem 2015)). Training set contains 243 patients data and Test set contains 60 patients data. Research on prediction of heart disease is conducted in Spyder Anaconda with Windows 10.1 system(Swain, Pattnaik, and Gupta 2020). Proposed system uses a two groups forest algorithm over a Linear Regression ((Baig 2020). Where these algorithms are fitted into a dataset which is then tested and trained for the process of prediction of heart disease with novel principal component analysis using Supervised Classification.

**Forest Algorithm**

Forest algorithm is a notable AI estimation that has a spot with controlled learning technique. It tends to be utilised for both Classification and Regression issues in Machine learning. Overall thought of the bagging method is that a blend of learning models expands the overall result. One major benefit of random forest is that it tends to be utilised for both classification and regression problems (Bansal et al. 2018).

**Pseudocode:**

**Step1:** For each tree in forest, Select a bootstrap sample from S

where S (i) denotes bootstrap.

**Step2:** Then learn a decision-tree using a modified decision-tree learning algorithm.

**Step3:** Algorithm is modified as follows: at each node of tree, instead of examining all possible feature-splits.

**Step4:** Randomly select some subset features $f \subseteq F$.

**Step5:** where F is a set of features. Node then splits on best feature in $f$ rather than F.

**Step6:** In practice $f$ is much, much smaller than F. Deciding on which feature to split is oftentimes the most computationally expensive aspect of decision tree learning. By narrowing the set of features, Drastically speed up learning of the tree.
**Precondition:** A preparation set $S := (x_1, y_1),..., (x_n, y_n)$, highlights $F$, and number of trees in forest $B$.

1. function RandomForest $(S, F)$
2. $H \leftarrow \emptyset$
3. for $i \in 1,..., B$ do
4. $S(i) \leftarrow$ A bootstrap sample from $S$
5. $h_i \leftarrow$ Randomised Tree Learn $(S(i), F)$
6. $H \leftarrow H \cup \{h_i\}$
7. end for
8. return $H$
9. end function
10. function Randomised Tree Learn $(S, F)$
11. At each node:
12. $f \leftarrow$ tiny subset of $F$
13. Split on best element in $f$
14. return learned tree
15. end function

**Linear Regression**

Linear Regression is a machine Learning in AI algorithm dependent on administered learning. Regression models target prediction. Relapse models an objective expectation esteem dependent on autonomous factors. It is for the most part utilized for discovering the connection among factors and anticipating. Different regression models contrast depending on the sort of connection among reliant and autonomous factors (variables) they are thinking about and the quantity of free factors being used.

**Pseudocode:**

**Step 1:** Start
**Step 2:** Read the number of data ($n$)
**Step 3:** For $i = 1$ to $n$:
   - Read $X_i$ and $Y_i$
**Step 4:** Initialize:
   - $\text{sumX} = 0$
   - $\text{sumX}^2 = 0$
   - $\text{sumY} = 0$
   - $\text{sumXY} = 0$
**Step 5:** Calculate required sum
   For $i = 1$ to $n$:
   - $\text{sumX} = \text{sumX} + X_i$
   - $\text{sumX}^2 = \text{sumX}^2 + X_i^2$
   - $\text{sumY} = \text{sumY} + Y_i$
   - $\text{sumXY} = \text{sumXY} + X_iY_i$
**Step 6:** Calculate Required Constant $a$ and $b$ of $y = a + bx$:
   
   $b = (n \times \text{sumXY} - \text{sumX} \times \text{sumY}) / (n \times \text{sumX}^2 - \text{sumX} \times \text{sumX})$

**Step 7:** Display value of $a$ and $b$
**Step 8:** Stop

Software tool used to evaluate Forest algorithm and Linear Regression is Spyder Anaconda. Hardware configuration was an i5 processor with 8GB RAM and operating Windows 10.

**Statistical Analysis**

Analysis was done using IBM SPSS version 21. It is a statistical software tool used for information analysis. Independent variable is BMI, smoking status, HbA1C, systolic BP, age, sex, nationality and dependent variable is accuracy. Independent sample T-test analysis is carried out in this research work. For both Proposed and existing frameworks 15 cycles were finished with a limit of 304 Samples and for every emphasis the anticipated exactness was noted for dissecting precision with the values obtained from the iterations Independent Sample T-Test was performed and graphs were also plotted (Guo et al. 2021). For the Forest Algorithm algorithm, it was observed that if number of iterations increased, then error rate decreased and accuracy increased using novel dimensionality reduction technique.

**RESULTS**

In Table 1, it was observed that Forest Algorithm is significantly better than Linear Regression. In kaggle dataset, it is observed that accuracy and performance of Forest Algorithm are significantly better than Linear Regression.

<table>
<thead>
<tr>
<th>Execution</th>
<th>Forest algorithm</th>
<th>Linear Regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>90.32</td>
<td>77.21</td>
</tr>
<tr>
<td>2</td>
<td>91.02</td>
<td>78.01</td>
</tr>
<tr>
<td>3</td>
<td>89.02</td>
<td>76.03</td>
</tr>
<tr>
<td>4</td>
<td>91.03</td>
<td>78.02</td>
</tr>
<tr>
<td>5</td>
<td>89.04</td>
<td>76.23</td>
</tr>
<tr>
<td>6</td>
<td>91.33</td>
<td>76.54</td>
</tr>
<tr>
<td>7</td>
<td>89.43</td>
<td>78.21</td>
</tr>
<tr>
<td>8</td>
<td>91.32</td>
<td>76.84</td>
</tr>
<tr>
<td>9</td>
<td>89.62</td>
<td>78.35</td>
</tr>
<tr>
<td>10</td>
<td>89.62</td>
<td>76.72</td>
</tr>
</tbody>
</table>

Table 1

Comparison of prediction accuracy between Forest algorithm and Decision Tree. Forest algorithm is 90.32% compared to Linear Regression Algorithm 85.00%.
In Table 2, Forest Algorithm achieved a mean of 90.1340, standard deviation of 0.97244 and standard error mean of 0.30751. Linear Regression mean is 77.2160, standard deviation is 0.86757 and standard error mean is 0.27435. In Table 3, 2-tailed significance values smaller than 0.001 (p<0.05) showed that our hypothesis holds good.

<table>
<thead>
<tr>
<th>Accuracy Groups</th>
<th>N</th>
<th>Means</th>
<th>Std deviation</th>
<th>Std error means</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random Forest</td>
<td>10</td>
<td>90.1340</td>
<td>0.97244</td>
<td>0.30751</td>
</tr>
<tr>
<td>Linear Regression</td>
<td>10</td>
<td>77.2160</td>
<td>0.86757</td>
<td>0.27435</td>
</tr>
</tbody>
</table>

Figure 1 represents mean accuracy of the software effort estimation for Forest Algorithm and Linear Regression. Forest Algorithm achieved better performance than Linear regression with novel dimensionality reduction technique.

**DISCUSSION**

In this study we observed that predicting heart disease is better using Forest Algorithm than Linear Regression algorithm using Supervised Classification. The accuracy value of Forest Algorithm obtained is 90.32% and Linear regression obtained is 77.21%.

In this correction survey, respondents (Rajesh and Malathi 2021) reported greater preferences in accuracy than respondents (Kavitha et al. 2019). In this adjustment study, researchers (Yang 2009) detailed more prominent inclinations in precision than respondents of parameters (Goel, n.d.). In this adjustment study, respondents (Esme et al. 2018) detailed more noteworthy inclinations in precision than respondents.

<table>
<thead>
<tr>
<th>Levene's Test for Equality of Variances</th>
<th>t-test for Equality of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>sig</td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
</tr>
<tr>
<td>Equal variances assumed</td>
<td>0.648</td>
</tr>
<tr>
<td>Equal variances not assumed</td>
<td>31.346</td>
</tr>
</tbody>
</table>
(Thippanna and Vutkur, n.d.). Hence algorithms produce effective output with novel Dimensionality Reduction techniques.

Existing system accuracy for Forest Algorithm and Linear regression are 90.32% and 77.00% respectively. An endeavour is made to carry out dimensionality decrease through principal component analysis for the coronary illness dataset. This analysis makes use of machine learning to predict the hassle if knowing all certain parameters like age, sex, nationality, BMI, smoking status, HbA1C, systolic BP (Bansal et al. 2018). Zhao used a Forest Algorithm tree method to get a better accuracy of 90.32% in predicting heart disease with supervised learning (Guo et al. 2021; Zhao et al. 2020). The Forest Algorithm has higher intricacy than Linear regression calculation (Jmaiel et al. 2020). Identification and selection of significant attributes for better diagnosis of heart disease are very challenging tasks for future research (Ko et al. 2021). Estimation of software development in heart disease prediction is also a challenging factor. Factors affecting the algorithms are sample size of dataset and test size of dataset (Guo et al. 2021). In light of above discovery, proposed calculation was picked to work on the precision (Guo et al. 2021; Zhao et al. 2020). It is important to add more values to the data set and trained data set to predict accurately.

Limitation is that Forest Algorithm uses a higher time complexity than the Linear Regression, to improve the results further it can use a Machine Learning approach, for that this work requires a lot of learning data, so this becomes the further scope (Yan et al. 2021). The factor that affects this system is incomplete data and lack of technical support in hospitals and in the medical industry (Jmaiel et al. 2020).

In order to improve results further it can utilise Deep Learning approach, for that this work requires part of learning information, so this turns into the future scope. In this technique to notice the strength of the calculations by expanding the clump size and fluctuating the number of hidden layers. By extracting and implementing them, it is to enhance the performance of the versions. Moreover, it would be good to do some modifications to both the reference set and the rules in order to make our program more general.

CONCLUSION

The accuracy outcome of the Forest Algorithm 90.32% is better than Linear Regression Algorithm 77.21% in predicting heart disease.

DECLARATIONS

Conflict of Interests
No irreconcilable situation in this original copy

Authors Contribution
Author KNS was involved in data collection, data analysis, and manuscript writing. Author KT was involved in conceptualization, data validation, and critical review of Manuscript.

Acknowledgements
The creators might want to offer their thanks towards Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences (Formerly known as Saveetha University) for giving the important framework to effectively do this work.

Funding
We thank the accompanying associations for offering monetary help that empowered us to finish the review.
1. VBLP Techn Solutions..
2. Saveetha University.
3. Saveetha Institute of Medical and Technical Sciences.
4. Saveetha School of Engineering

REFERENCES
5. Borra, Surekha, Nilanjan Dey, Siddhartha Bhattacharyya, and Mohamed Salim Bouhlel. 2019. Intelli-


