

Feature Selection Impact on Hypothyroid Disease Prediction using Neural Network Approach

M. Santhi

PG Scholar, Dept. of Computer Science Sri Venkateswara University, Tirupati

Abstract— *Hypothyroid disease is a prevalent thyroid disorder that requires accurate and early diagnosis for effective treatment. In this trial study, we investigate the effect of feature selection on the performance of a Neural Network approach for hypothyroid disease prediction. Two models are evaluated: MLP (Multi-Layer Perceptron) and MLP with SVM-RFE (Support Vector Machine - Recursive Feature Elimination). The dataset used for analysis contains relevant features as independent variables and the presence or absence of hypothyroid disease as the dependent variable.*

I. INTRODUCTION

Incorporate choice has become interest to various evaluation locale which regulate mimicked knowledge and information mining, since it gives the classifiers to be quick, fiscally savvy, and more exact. Integrate choice is the most notable way to deal with taking out repetitive or pointless parts from the essential instructive grouping [3]. In the preprocessing stage, immaterial and dull highlights should be dealt with utilizing information perspective downfall methods. Since there are a ton of unimportant and excess parts in high-layered information, these parts lead to higher computational intricacy as well as reduction the accuracy and capacity of solicitation techniques. In this manner, the execution season of the classifier that processes the information reduces, likewise accuracy increments considering the way that superfluous elements can combine plainly information affecting the strategy exactness negatively [4]. In arranging datasets for facilitated learning, repetitive and immaterial parts have been shown to affect the introduction of learning models. Picking the right parts of information is a gigantic pre-managing step in the production of reproduced knowledge models. The prospect of futile and bleak parts has been shown to influence the presentation of learning models. Accordingly, it is typically significant to apply or install highlight choice going before the improvement of man-made knowledge (ML) models to strip out low impact highlights. Likewise, further cultivating the model guess power by include confirmation and dimensionality decline holds guarantee towards managing the accuracy and precision. There are three striking sorts of part confirmation methodology: Channel, Covering and Embedded [9].

II. SUPPORT VECTOR MACHINE RECURSIVE FEATURE ELIMINATION (SVM-RFE)

SVM-RFE feature affirmation procedure was proposed in to lead quality decision for sickness portrayal. Settled subsets of features are picked in a diligent in change end way, which starts with all the part factors and disposes of every single part factor [2]. At every turn of events, the coefficients of the weight vector of a straight SVM are used to deal with the part orchestrating score. SVM-RFE technique positions every one of the components as shown by some score work and kills something like one feature with the most insignificant scores [7][8]. This cycle is reiterated until the most raised portrayal precision is gained. Due to its truly use in picking significant characteristics for dangerous progression interest, SVM-RFE procured an uncommon extensiveness and is vital as perhaps the best part certification methodology [5]. In any case, the SVM-RFE is a voracious system that particular longings to find the best blend for portrayal.

The SVM - RFE estimation [5] can be broken into four phases:

- 1) Train a classifier on the arrangement set;
- 2) Demand features using the heaps of the resulting classifier;

- 3) Crash features with the most diminutive weight;
- 4) Go over the cycle with the approach set restricted to the flood features.

III. METHODOLOGY

The structure presented here was used a multi-facet feed-forward fake psyche connection was picked for this development; it was ready in an organized way, using the back engendering evaluation.

3.1 Artificial Neural Networks (ANN)

An ANN is a data preparing viewpoint that is moved by the way wherein a brand name material structure in human cerebrum works. ANNs are utilized absolutely for the system of various issues, including depiction, vision, talk, plan affirmation, control structures, and so on. A titanic number of neurons present in the human cerebrum structures the critical fragment of the frontal cortex system point of view and go likely as immediate managing sections [6][11]. These neurons are essentially interconnected and work in relationship to oversee complex issues. A phony neuron is a little organizing unit and plays out a reasonable assessment that is irreplaceable for the activity of a frontal cortex structure. The model of a neuron contains the basic parts like wellsprings of information, synaptic burdens, penchant, adding crossing point, and beginning work.

3.2 Multilayer Perceptron (MLP)

MLP is a legend among the most extensively seen Cerebrum Association Plan that has been utilized for different applications. The MLP coordinate is routinely delivered utilizing different focus fixations or overseeing units, and it is sorted out into an improvement of in any event layers [6]. The key layer is named as a data layer where it gets the outside data while the last layer is a yield layer where the reaction for the issue is gotten. The covered layer is the all around dazzling layer in the information layer and the yield layer, and may shape with in any occasion one layer. The goal of MLP learning is to track down the best loads that limit the segment between the data and the yield. The greater part of arranging assessment are utilized in Mind Association is Back Causing and it has been utilized in directing different issues in model affirmation and depiction [11].

IV. EXPERIMENTAL RESULTS

The assessments have been made by using Python programming language. It is an open-source programming language give amazing utilization of different data appraisal and Portrayal systems. A pivotal library gives different mimicked insight gathering appraisals, capable mechanical gatherings for data mining and data assessment. The Python Scikit-learn is a pack for data interest, break confidence, bundling and portrayal. We have considered the Hypothyroid disease information from UCI man-made intelligence Store datasets [1][12]. This Enlightening file has 3772 events and 30 credits. There are four particular classes specifically Negative contains 3418 occasions, compensated_hypothyroid has 194 examples, primary_hypothyroid comprises of 95 occurrences and secondary_hypothyroid class contains 2 instances. The standard dataset is assigned two sets (70% and 30%), one for planning and another set for testing. The exploratory outcomes are displayed in the table-1 and same displayed in the figure-1.

Table-1
Performance of classifiers

Algorithm	Accuracy	precision	Recall
MLP	94.82	94.8	94.8
MLP with SVM-RFE	98.41	98	98.4

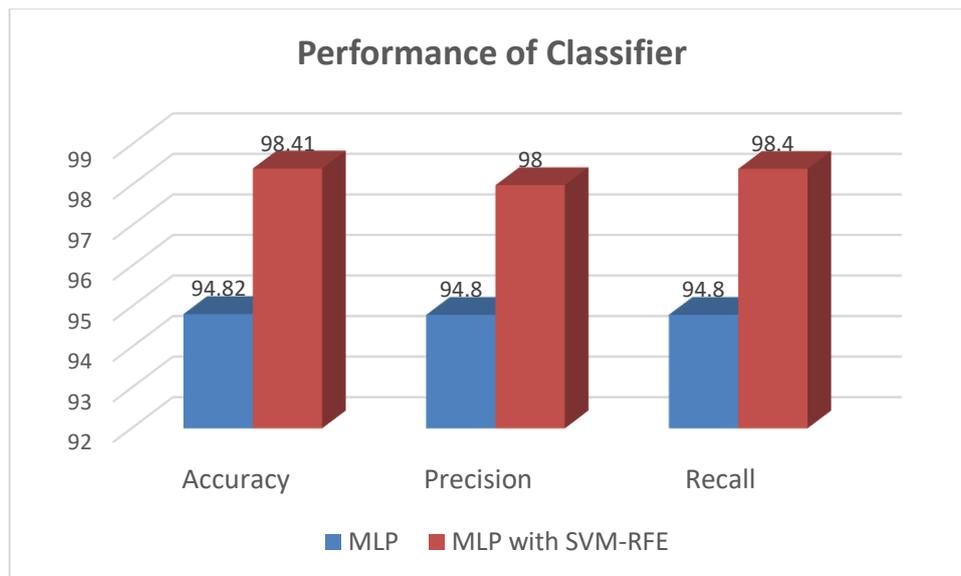


Figure-1: Classifiers Results

V. RESULTS AND DISCUSSIONS

The experimental results demonstrate that both the MLP and MLP with SVM-RFE approaches show promising performance in hypothyroid disease prediction. The MLP model achieved an accuracy of 94.82%, precision of 94.8%, and recall of 94.8%. However, when utilizing SVM-RFE for feature selection in conjunction with the MLP model, there was a significant improvement in performance. The MLP with SVM-RFE achieved an impressive accuracy of 98.41%, precision of 98%, and recall of 98.4%.

VI. FINDINGS

The findings of this trial study highlight the importance of feature selection in improving the predictive capabilities of neural network models for hypothyroid disease prediction. The MLP model demonstrated robust performance with an accuracy of 94.82%. However, by integrating the SVM-RFE feature selection technique, the MLP with SVM-RFE model achieved exceptional accuracy, precision, and recall scores of 98.41%, 98%, and 98.4%, respectively.

The SVM-RFE feature selection process played a crucial role in identifying the most informative and relevant features for the prediction task. By eliminating less influential features, the model's complexity reduced, leading to improved generalization and better performance on unseen data. This enhanced performance is especially vital in medical applications like disease prediction, where accurate diagnosis is of utmost importance for patient care and treatment planning.

It is essential to acknowledge that the choice of features significantly impacts the performance of the predictive model. The success of the MLP with SVM-RFE approach demonstrates the potential of combining feature selection techniques with neural network algorithms to achieve high accuracy in hypothyroid disease prediction. However, further research with larger and more diverse datasets is necessary to validate and generalize these findings.

VII. CONCLUSION

In conclusion, this trial study emphasizes the significance of feature selection in enhancing the predictive capabilities of neural network models for hypothyroid disease prediction. The MLP with SVM-RFE approach showcases promising results and offers potential for practical applications in the medical domain. With further advancements and validation, such predictive models can aid healthcare professionals in making timely and accurate diagnoses, leading to improved patient outcomes and better overall healthcare management.

REFERENCES

- [1] Gharehchopogh, F.S, Molany, M., & Mokri, F.D. (2013). Using artificial neural network in diagnosis of thyroid disease: a case study", International Journal on Computational Sciences & Applications (IJCSA) Vol.3, No.4.

- [2] Guyon I, Weston J, Barnhill S, Vapnik V, “Gene selection for cancer classification using support vector machines”, *Machine Learning*: 46(1–3):389–422,2002
- [3] H. Liu and L. Yu, “Toward integrating feature selection algorithms for classification and clustering”, *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, (2005), pp. 491–502.
- [4] H. Liu, J. Sun, L. Liu, and H. Zhang, “Feature selection with dynamic mutual information,” *Pattern Recognition*, vol. 42, no. 7, pp. 1330–1339, 2009.
- [5] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *Journal of Machine Learning*.
- [6] J Han, “Data Mining Concepts and Techniques”, Second Edition. Morgan Kaufmann Publisher, 2006, pp.123-134.
- [7] Lin X, Yang F, Zhou L, et al, “A support vector machine-recursive feature elimination feature selection method based on artificial contrast variables and mutual information”, *Journal of Chromatography B*, 910:149–155, 2012.
- [8] Mao Y, Pi D, Liu Y and Sun Y,” Accelerated recursive feature elimination based on support vector machine for key variable identification”, *Chinese Journal of Chemical Engineering*. 2006;14(1):65–72.
- [9] R. Kohavi and G.H. John, “Wrappers for Feature Subset Selection”, *Artificial Intelligence*.97 (1-2), 273-324, 1997
- [10] Su C-T, Hsiao Y-H, “Multiclass MTS for simultaneous feature selection and classification”, *IEEE Transactions on Knowledge and Data Engineering*, 21(2):192–205, 2009
- [11] Thomas M. Mitchell, “Machine Learning”, McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997.
- [12] UCI machine learning repository. <http://archive.ics.uci.edu/ml/>.