IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Application and Comparison of Several ML Algorithms and their Integration Models in Regression Problems

A.GANESH KUMAR [#1], N. SIVAKUMAR REDDY[#2]

[#1] Assistant Professor, Department of MCA,
Sanketika Vidhya Parishad Engineering College, P.M. Palem,
Visakhapatnam, Andhra Pradesh.

[#2] MCA Student, Department of MCA,
Sanketika Vidhya Parishad Engineering College, P.M. Palem,
Visakhapatnam, Andhra Pradesh.

ABSTRACT

As machine learning technology advances quickly, more and more people pay attention to regression problems, which aid in extracting the pattern from vast amounts of data in order to produce a prediction impact. People now heavily rely on data prediction in their daily lives. The technique is currently widely employed in a variety of industries, including forecasting the weather, diagnosing illnesses, and the financial markets. As a result, one of the most active areas of research in the field of machine learning in recent years has been the study of machine learning algorithms in regression situations. This research mainly uses three popular machine learning algorithms: neural network, extreme learning machine, and support vector machine in order to more thoroughly explore the application effect of machine learning method in regression issue. Next, by contrasting the results of the benefits and drawbacks of each machine learning technique are examined by applying the single model and integrated model of these algorithms to regression situations. Here, we may choose any assignment that has to do with health, the weather, or an illness and compare it to the other difficulties to see which would provide the most accuracy.

KEYWORDS:

Financial, Weather, Machine Learning, Regression Problems, diagnosing Illness.

1. INTRODUCTION

According to the World Health Organization (WHO), cancer is the second biggest cause of death, with a projected 9.6 million deaths globally in 2018—mostly in poorer nations1. Additionally, 1.5 million of the 422 million adults who have diabetes die as a result of their condition. Even worse, cardiovascular disorders account for 17.5 million annual fatalities (CVDs). In addition, it was predicted that 214,360 Chinese women had died of breast cancer by the year 2008, and that the figure might rise to 2.5 million by the year 2021 [2].

Patients and their family both suffer as a result of this grave condition [3]. Therefore, it is crucial to determine the actual causes of "such a significant number of fatalities". Although the WHO claims that many cancer cases are discovered too late, accurate and early identification would guarantee the long-term survival of more than 30% of these people [4]. As a result, it is crucial that we develop a successful strategy for illness early detection in order to enhance our society's access to healthcare. In general, machine learning (ML) approaches have been widely used for medical diagnosis because they can efficiently extract meaningful knowledge from massive, complex, diverse, and hierarchical time series clinical data [5-8].

Additionally, by using ML approaches, pathologists and doctors may prevent potential medical mistakes brought on by inexperience, exhaustion, stress, and other factors, and medical data can be analysed quickly and in greater depth [9]–[11]. To the best of our knowledge, classification problems have been linked to the difficulty of medical diagnosis. Previous research has shown that a variety of classification techniques, including neural networks, Naive Bayes, KNN, and SVM, have been used for medical diagnosis, and the majority of these classification models produced excellent results.

To the best of our knowledge, the difficulty of medical diagnosis has been attributed to classification problems, as previous research showed that a variety of classification methods, including neural networks, Naive Bayes, KNN, and SVM, have been used for medical diagnosis and that the majority of these classification models achieved excellent performance. However, these cutting-edge classification models were primarily concerned with classification accuracy, ignoring the unbalanced nature of the initial data input.

2. LITERATURE SURVEY

Literature survey is that the most vital step in software development process. Before developing the new application or model, it's necessary to work out the time factor, economy and company strength. Once all these factors are confirmed and got an approval then we can start building the application.

MOTIVATION

1) **Diabetes Prediction using Machine Learning Algorithms**

Authors: Aishwarya Mujumdar

Numerous people suffer from diabetes mellitus, one of the most serious diseases. Age, obesity, inactivity, genetic diabetes, a poor diet, high blood pressure, and other factors can all contribute to diabetes mellitus.

Diabetes increases a person's chance of developing several illnesses, including heart disease, renal disease, stroke, vision problems, nerve damage, etc. A variety of tests are currently used in hospitals to get the data needed to diagnose diabetes, and depending on that diagnosis, the proper therapy is given. The healthcare sector greatly benefits from big data analytics. Databases in the healthcare sector are very vast. Using big data analytics, one may examine enormous datasets and uncover hidden patterns and information to learn from the data and forecast results appropriately.

2) Predicting Diabetes Mellitus with ML Techniques.

Authors: Quan Zou

Hyperglycemia is a chronic condition associated with diabetes mellitus. It might lead to a lot of difficulties. According to current increases in morbidity, the number of diabetic patients worldwide is expected to reach 642 million by 2040, or one out of every 10 persons. Without a doubt, this worrying number requires a lot of attention. Machine learning has been used in many facets of medical health thanks to its quick progress. In order to predict diabetes mellitus, we

employed decision trees, random forests, and neural networks in this study. The hospital physical examination statistics in Luzhou, China, make up the dataset. There are 14 qualities in it. Five-fold cross validation was utilised in this work to evaluate the models. For the universal to be verified.

3) Diabetes Prediction Using ML Algorithms
Authors: Jana S

Various services and resources inside distributed systems need to be secured against unauthorised usage. The most popular technique for confirming a distant client's identity is remote authentication.

This study examines a methodical procedure for client authentication using a password, a smart card, and a biometric. The transition from two-factor authentication to three-factor authentication is suggested using a general and safe architecture. In distributed systems, the conversion not only greatly raises information assurance at a minimal cost but also safeguards client privacy. Additionally, we feel that our framework is of independent importance since it preserves a number of the two-factor authentication's practice-friendly characteristics.

3. EXISTING SYSTEM AND ITS LIMITATIONS

Nearly all machine learning algorithms used today have advantages in certain stages and limitations in others. When dealing with the right dataset, no machine learning system is capable of precisely recognising all the elements. As a result, several apps that used ML algorithms to accomplish the task failed to attain accuracy.

LIMITATION OF PRIMITIVE SYSTEM

The following are the limitations of the existing system.

1) In the existing days there is integrated approach for performing the task.

2) There is no concept to achieve all the aspects on appropriate dataset.

3) There is no mechanism which can compare multiple factors in single algorithm and achieve accuracy in all aspects.

4. PROPOSED SYSTEM AND ITS ADVANTAGES

This suggested work primarily uses the neural network, extreme learning machine, and support vector machine as three popular machine learning algorithms in order to more thoroughly explore the application effect of machine learning algorithm in regression issue. The merits and disadvantages of each machine learning algorithm are then investigated by contrasting the results of the single model and integrated model of various machine learning algorithms when applied to regression issues. Here, we may choose any work that is associated with medicine, the weather, or an illness, compare the many issues, and chooses which will provide the highest accuracy.

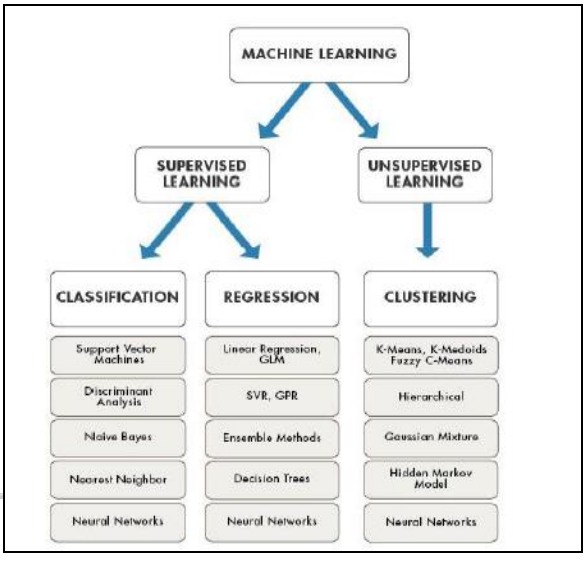


Figure 1. Represents the Several ML Algorithms

The following are the advantages of the proposed system. They are as follows:

1) In the proposed system we try to compare several algorithms and its accuracy.
2) Here we take one sample dataset as example and perform the ML algorithms to check accuracy.
3) In this proposed mechanism we can achieve high level of accuracy by comparing several algorithms on common dataset.

5. IMPLEMENTATION PHASE

Implementation is the stage where the theoretical design is converted into programmatically manner. In this stage we will divide the application into a number of modules and then coded for deployment. The front end of the application takes Google Collaboratory and as a Back-End Data base we took UCI Heart Patients Records as dataset. Here we are using Python as Programming Language to Implement the current application. The application is divided mainly into following 5 modules. They are as follows:

1) Load Dataset Module
2) Visualize the Data
3) Data Pre-Processing
4) Train the Model Using Several ML Algorithms
5) Find the Performance of ML Algorithms

1) Load Dataset Module

Here the data set is visualized into number of attributes like

1. Number of Pregnancies
2. Glucose Level
3. BloodPressure
4. SkinThickness
5. Insulin
6. BMI
7. DiabetesPedigreeFunction
8. Age
9. Outcome

Here the first 8 are main attributes to check the diabetes is present or not and 9th attribute is Outcome.

2) Visualize the Module

In this module we try to load the dataset which is downloaded from a pre-defined location:

https://www.kaggle.com/uciml/pima-indians-diabetes-database

The data set is mainly contains all the information about several female patients who are suffered with diabetes disease symptoms. This data is collected, tested and then formed as input for the ML applications.

3) Data Pre-Processing Module

Here the data pre-processing is performed and if there is any in complete data present in the dataset those incomplete records need to be removed from the dataset and only which are complete and accurate must only be kept in the input dataset. Here we try to identify each and every record contains information about all the 9 attributes or not and if any record present missing attributes then such a record need to be removed and cleaned from that dataset.

4) Train the Model Using Several ML Algorithms

Here we try to apply classification algorithms like :

1) XGBoost,
2) Support Vector Machine(SVM),
3) KNN,
4) Random Forest Algorithm

Finally find out which algorithm is best suited for early prediction of diabetes disease from a set of patients records.

5) Find the Performance of ML Algorithms

In this module we try to classify each and every algorithm and then find out the accurate algorithm from a set of classification algorithms. Finally we try to find out the best algorithm based on accuracy. Here we get random forest and Xgboost as good accuracy and one among four is random forest stood at first position.

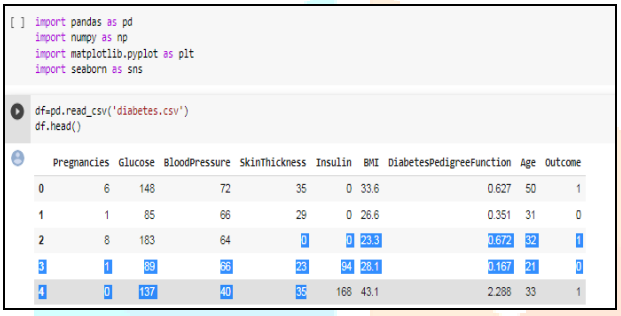
6. EXPERIMENTAL RESULTS

In this section we try to design our current model using Python as programming language and we used google collab as working environment for executing the application.Now we can check the performance of our proposed application as follows:

LOAD DATASET

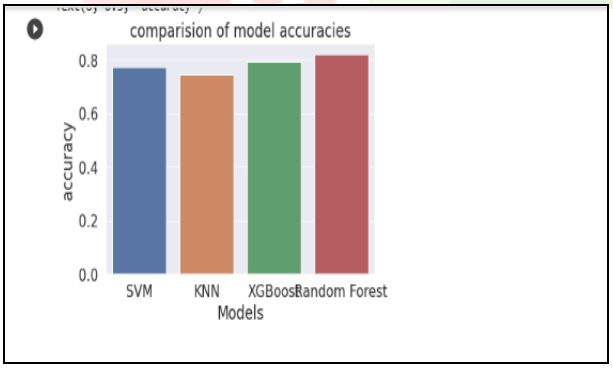
The above window clearly represent how the dataset is loaded.

Import Libraries



From the above window the necessary libraries are imported.

Comparison of Several ML Algorithms



From the above window we can clearly see the Random Forest gives more accuracy compared with other ML Algorithms for disease prediction.

7. CONCLUSION

The three most popular machine learning algorithms—neural networks, extreme learning machines, and support vector machines—are mostly used in this project's suggested application to get over the drawbacks or effects of machine learning methods in regression problems. The benefits and drawbacks of each machine learning method are then examined by contrasting the results of the single model and integrated model applications of various machine learning algorithms to regression situations. Here, we may choose any work that is linked to medicine, the weather, or an illness, compare the many issues, and choose which will provide the highest accuracy. Finally, our experimental findings demonstrate that, among several algorithms, Random forest is the best.

8. REFERENCES

[1] Mujumdar, A., & Vaidehi, V. (2019). Diabetes Prediction using Machine Learning Algorithms. *Procedia Computer Science*, *165*, 292-299. https://doi.org/10.1016/j.procs.2020.01.047

[2] Zou Q, Qu K, Luo Y, Yin D, Ju Y and Tang H (2018) Predicting Diabetes Mellitus With Machine Learning Techniques. *Front. Genet.* 9:515. doi: 10.3389/fgene.2018.00515

[3] J. S, B. N, S. P, S. K. K and V. Mani Nageshwar, "Diabetes Prediction Using Machine Learning Algorithms," 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), 2022, pp. 46-51, doi: 10.1109/ICACCS54159.2022.9785073.

[4] Jiang, J., Li, X., Zhao, C., Guan, Y., & Yu, Q.. Learning and inference in knowledge-based probabilistic model for medical diagnosis, *Knowledge-Based Systems*, 139, 58-68,2017.

[5] Kovalchuk, S V., Krotov, E., Smirnov, P., Nasonov. D A., & Yakovlev, A N. Distributed data-driven platform for urgent decision making in cardiological ambulance control, *Future Generation Computer Systems*, 79, 144-154, 2018.

[6] Piri, S., Delen, D., & Liu, T. A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets, Decision Support Systems, 106, 15-29, 2018.

[7] Eshtay, M., Hm Faris., & N, Obeid. Improving Extreme Learning Machine by Competitive Swarm Optimization and its application for medical diagnosis problems, *Expert Systems with Applications*, 104, 134-152, 2018.

[8] Nagarajan, R., & M, Upreti.(2017). An ensemble predictive modeling framework for breast cancer classification, Methods, 131, 128-134.

[9] Chen, H., Yang, B., Liu, J., & Liu, D. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications, 38(07), 9014-9022, 2011.

[10] Kazemi, Y., & S.A, Mirroshandel. A novel method for predicting kidney stone type using ensemble learning, Artificial Intelligence in Medicine, 84, 117-126, 2018.

[11] Wang, H., Zheng, B., Sang, W Y., & Ko, HS. A support vector machine-based ensemble algorithm for breast cancer diagnosis, European Journal of Operational Research, 267(02), 687-699, 2018.

[12] Liu, Y., Yu, X., Huang, J X., & An, A. Combining integrated sampling with SVM ensembles for learning from imbalanced datasets. Information Processing & Management, 47(4),617-631, 2011.

About the Authors

A.GANESH KUMAR is currently working

as an Assistant Professor in Department of MCA at Sanketika Vidhya Parishad Engineering College, P.M. Palem, Visakhapatnam, Andhra Pradesh. He has more than 12 years of teaching experience. His research interest includes Cloud Computing and Programming.

N. SIVAKUMAR REDDY is currently

pursuing his 2 years MCA in Department of MCA at Sanketika Vidhya Parishad Engineering College, P.M. Palem, Visakhapatnam, Andhra Pradesh.His area of interest includes Python, Java, C, C++, Devops.