A Data Enhancement Approach to Indication of Health Status Using Linear Regression and Random Forest

[1]Mrs.R.Leela Jyothi , [2]Dr .Ch.Rajendra Babu, [3]M.Vamsi Priya

[1]Assistant Professor, Dept of CSE, SRKR Engineering College (A), Chinnaamiram, Bhimavaram, West Godavari District, Andhra Pradesh, India
[2]Professor, Dept of CSE, Andhra Loyola Institute of Engineering And Technology, Vijayawada, NTR District, Andhra Pradesh, India
[3]Assistant Professor. Dept of CSE (Data Science), CMR Technical Campus, Medchal Road, Kandlakoya, Hyderabad, Telangana, India
Mail Id's:- [1]rudraraju.leela92@gmail.com, [2]chikkalarajendra@gmail.com, [3]vamsipriya.ds@cmrtc.ac.in

ABSTRACT: The fastest-growing field in health informatics and computer science is Machine Learning (ML), which enhances the performance of remote healthcare systems. The development of algorithms that are capable of learning over time, making progress over time, and making predictions is the goal of machine learning. The healthcare industry faces difficulties in critical areas like electronic record management, in light of the need to move toward personalized healthcare and reduced costs in healthcare, data integration and computer-aided disease diagnosis and prediction are essential. This analysis describes the Data Enhancement Method to Health Status Indication Using Linear Regression and Random Forest. Adults aged between forty and seventy make up the data set. Weak, normal, overweight, and other health statuses can be predicted using our model's use of BMI. Using the user's weight and height, the BMI is determined. Compared to other models, the Linear Regression and Random Forest model's health status identification has a high accuracy rate of 98% and a less identification time of 13.4 milliseconds.

KEYWORDS: BMI, electronic record management, Machine Learning, Health Status, Linear Regression and Random Forest.

Doi: 10.31838/ecb/2023.12.8.810

I. INTRODUCTION

Health is described as "a condition of total physical, mental, and social well-being, instead of only the absence of disease or disability, as the World Health Organization describes it (2015). Health is a multi-dimensional concept. Quality of life is closely linked to physical health, mental well-being, and cognitive capacity for older people [1].

There are three commonly used indicators of physical health: the presence of illness, functional capability, and subjective health assessment. Through supporting individuals in predicting and obtaining an accurate understanding of the building's health, scientific manner and formulating corresponding protective measures based on the building health status, the building health status recognition has the potential to reduce the likelihood of building collapse accidents. Both socially and economically, it is essential and provides major advantages [2].

The condition of obesity is determined by its indexes, which range from zero to five (Very weak, weak, average weight, overweight, obesity, and extremely obesity,) [3]. The most precise and reliable method for determining an individual's level of obesity is the BMI (Body Mass Index). The individual's weight and height can be used to calculate BMI. The excessive consumption of calories and insufficient calorie burning as a result of inactivity is one of the main contributors to obesity [4]. The Body Mass Index (BMI) can be used as a measure to determine whether or not our weight is healthy, if it is unhealthy, it can indicate related diseases including diabetes and heart disease.

Time series data analysis is seen as a challenge that helps extend a building's lifespan by determining its health status has been proposed by some developed countries [5]. Identification of health status research is a relatively recent field, but country's economy, scientific and technological advancements have accelerated the development of health status identification research technology in recent years. As a result, numerous efficient approaches have been proposed. There are two types of current building health status identification methods: the first is a classification technology-based method for recognizing a building's health status, and the second is a time series-based method for recognizing a building's health status. First, an approach that takes structures into consideration. There is some internal temporal connection to the health issue. Future building health state is predicted based on the current building health status. However, a significant amount of historical information regarding the building's health status is required for the method. The effectiveness of determining a building's health condition is very poor if this requirement is not achieved, the second approach to consider the challenge of classifying a building's health condition as a multi-class problem, indicating that there are different states of a building's health, utilizing artificial neural networks most frequently [6].

The methodology that utilizes machine learning is considered for improving the results of health status identification, to evaluate the effectiveness and superiority of developing health status identification and specific illustrations are provided. Industrial organizations deal with a lot of data, which needs to be identified using Machine Learning. Organizations are able to work more effectively and by gaining insights from these data, they can gain an advantage over their competitors [7]. With the use of machine learning algorithms, creative prediction models have been effectively used in a range of industries. Techniques and applications of machine learning are used in everyday activities like searching, watching advertisements, and YouTube. Multidisciplinary health care informatics has associated with technological advancements and data handling difficulties. The scientific field of medical or health informatics is related with the optimal use, storage, retrieval of medical data and information, as well as providing knowledge for problem solving and decision making [8]. The field of health technology has undergone significant development over time, including advancements in information collection, treatment, communication, and research. Described method is made public so that anyone can use it to obtain accurate results in accordance with medical regulations. Those in the age range of less than or equal to 70 years are the primary focus of described study. There are five sections in this analysis. The related works are presented in Section II. The model for identifying health status is described in Section III. In this Section IV, the experimental results are explained, and in Section V, the analysis is completed..

II. LITERATURE SURVEY

Al Hanai T., Ghassemi M. M., and J. R. Glass et al. [9] proposed an automated depression detection method using audio and text. Interviews were conducted with both the real person and the virtual agents. These models are supposed to learn through a series of questions and answers without having to perform based on a specific topic. Based on the responses of the individuals, depression was identified using LSTM (Long Short-Term Memory) neural network models. A. Sau and Bhakta I. et al. [10] an efficient prediction model based on various variables, including age, literacy, place of residence, work status, previous experiences with anxiety and depression, and etc, was developed using machine learning algorithms to support in the diagnosis of depression and anxiety in older patients. The Waikato Environment for Knowledge Analysis (WEKA) tool was used to run ten machine learning classifiers during the research. K Star, Na¨ıve Bayes (NB), Random Forest (RF), Logistic Regression (LR), Bayesian Network and the Multiple Layer Perceptron (MLP) were all parts of it. J48, minimal sequential optimization, Random subspace, and Random subspace were also included.

Gehrmann, S., Dernoncourt, F, Li, Y., Carlson, E.T., Wu, J.T., Welt, J., Foote, J., Jr., Moseley, E.T., Grant, D.W., Tyler, P.D. et. al. [11] presents Comparing rule-based and deep learning models for patient phenotyping. Binary classification is a natural choice because main objective was to predict the presence of a disease by analyzing the available medical information. For binary classification tasks, the performance metrics most commonly implemented are recall and precision, but they may not accurately reflect the model's performance. As a result, presenting the F1 score is recommended, which provides a more accurate representation of the model's performance in accurately identifying the disease in its entirety and in preventing false positives.

D. LaFreniere, F. Zulkernine, D. Barber and K. Martin et al. [12] propose a neural network-based model for predicting hypertension. The CPCSSN (Canadian Primary Care Sentinel Surveillance Network) data set, the researchers utilized significant samples of 1 patient, 85,371 patients, and 93,656 controls in this study. During the training network, the following variables are considered: gender, birth year, BMI, diastolic and systolic Blood pressure (BP), Low and high density lipoprotein (LDL) levels, triglycerides, cholesterol, urine albumin creatinine ratio, and microalbumin. It's attractive to see that the authors also discussed that crucial sample size is when creating a prediction model based on machine learning. Todor Ivascu, Ovidiu Aritoni, et. al. [13] presents an Intelligent System based on Intelligent Agents and Fuzzy Logic that makes it easier to monitor a patient's health status remotely in real time. Patients can use the described system to stay in the comfort of their own homes and carry out activities of daily living while the system "silently monitors" their health condition, identifies irregularities, and notifies medical staff in the event of an emergency. This agent identified abnormalities by utilizing the warning score system for each controlled vital sign to access external services and notify physicians or emergency services of significant changes in the patient's health. Two experts examined and compared the patient's health status classification results.

Dongmei Chen, Max Q.-H. Meng, et. al. [14] the patient's status changes, such as health, sub-health, and abnormalities are represented by a density ratio made from the training density and testing density propose a new framework for monitoring the physiological state of the patient. Without requiring density estimation, they estimate the parameters of the density ratio by utilizing a Least Squares algorithm. They conduct the pilot experiments, verifying the usability and effectiveness of the described framework using physiological monitoring data (11901 beats) from the Physionet database. The methodology is successful in identifying the patient status, according to the results.

Sohn, S. Savova, G.K. Mayo et. al. [15] presnts a method, to improve the accuracy is determining a person's smoking status, the cTAKES is used in connection with a hybrid machine learning and rule-based methodology. One of the following classes can be used to describe this status: current smoker, smoker, nonsmoker, and unidentified status. The first step, to determine whether or not the patient is a smoker by combining three models. The second model, determines whether the patient is a smoker or a nonsmoker if the patient's smoking status is known, while the third model determines whether the patient has smoked in the past, is a smoker or does so currently. The reported micro-average accuracy for each of the three evaluation methods is 0.967. One of the reported performance measures is F1, which shows the way the model is performing.

III. DATA ENHANCEMENT APPROACH TO INDICATION OF HEALTH STATUS

Figure 1 shows the block design of the Data Enhancement Method to indication of health status using linear regression and Random Forest. They gathered the heights, weights, genders, and indexes of 500 selected randomly women and men using Kaggle. BMI is calculated using height and weight, and then the index is evaluated. After determining a person's BMI using their height and weight, the index is calculated. For instance, A person with a height of 149 centimeters and a weight of 61 kilograms, would have an index value of 3(a lot of weight). A individual with a height and weight of 174 cm and 96 kg has an index value of 4, which indicates obesity. From a 500-person dataset, 198 people have a high health status, 130 have an obesity level, 69 have a normal level, 68 have an overweight level, 22 have a low health status, and 13 have an extremely low health status.

Any machine learning project's performance is based on how well the data it uses for analysis is. Typically, at this stage of machine learning, human input is crucial. 80% of machine learning efforts, according to an often cited estimate, are information, and at this stage, data exploration is utilized to learn more about the information and its complexities. An appropriate algorithm will be chosen based on the specific machine learning task, and the algorithm will use a model to represent the data.

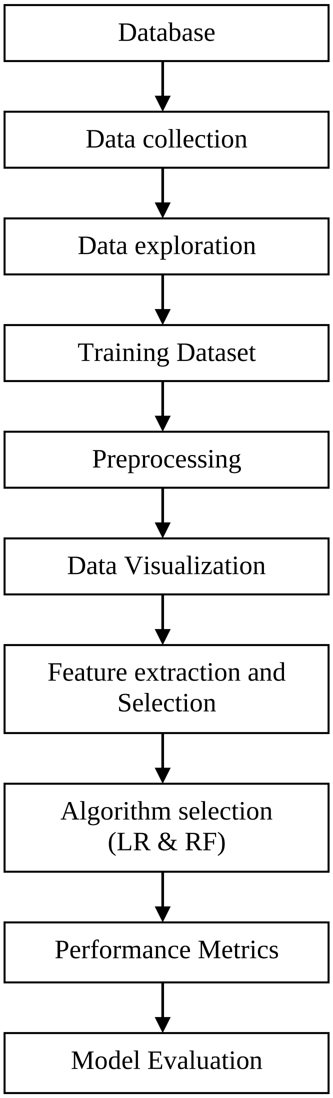

Fig. 1: ARCHITECTURE OF INDICATION OF HEALTH STATUS APPROACH

On computers, which are unable to understand human language and can only interpret organized data, the application of machine learning techniques. So that the machine learning system can handle the unstructured clinical notes, they need to be transformed into structured data. The medical notes dataset's characteristics include clinical notes words and terminologies, must be quantified using Natural language processing (NLP) techniques during this transformation process. Preprocessing is a procedure wherein clinical notes must first be cleaned up and prepared in order to undergo this transformation.

The next stage involves choosing a subset of the dataset's characteristics, feature extraction and selection preserves the dataset's information while improving learning generalization. The notes contain words and medical terminologies are evaluated through the selection and feature extraction process. The most popular techniques are Term Frequency-Inverse Document Frequency (TF-IDF) and Bag Of Words (BOW). Irrespective of the dataset's additional notes, BOW is maintaining a record of the amount of times each words appears in the generated clinical texts.

The machine learning model's accuracy and performance are affected by high-dimensional features, which can also lead to overfitting. Decrease the amount of features and choose those that offer more relevant data for the instructional process. In feature selection approaches, when selecting a subset of features that combined algorithms will enhance the effectiveness or accuracy of a model.

The following phase involves choosing the optimized model that produces the most accurate results using a methodology. Logistic Regression and Random Forest are the two algorithms are used in this process. The type of decision tree is the random forest. For our dataset's model training and testing, they used the random forest machine learning approach. The detection of health status using BMI has been performed using this technique to resolve classification and regression model issues. In order to address regression issues with descfribed dataset, they used a machine learning approach for linear regression. Gender, height, and weight are the three independent variables in descfribed model, while BMI and health status are the two dependent variables.

The BMI calculation, the algorithm compares values to described predetermined ranges, estimates the relevant health condition by matching the connected ranges with the corresponding index values. Since every machine learning model produces a biased answer to the learning issue, in order to determine effectively the algorithm learned from its previous experience, it is essential to evaluate model performance. Based on the model is utilized, a test dataset can be used to evaluate the model's accuracy.

IV. EXPERIMENTAL RESULT

They obtained the heights using Kaggle, weights, genders, and indexes of 500 random individuals, including men and women. Testing is conducted using specific application examples are in the same testing environment, in order to compare the effects of the machine learning algorithms (RF+LR) on health status identification, the Neural Network health status identification method and the Support Vector Machine health status identification method are chosen. The conditions for their test environment are still exactly the same as they were during the test. Accuracy rate and identification Time are two parameters used in this study for performance analysis.

Fig. 2 represents the accuracy rate as a graphical representation. The accuracy of neural networks and support vector machines for determining a building's health status is less than 90%, generating impossible to accurately differentiate between buildings. The RF+LR accuracy rate for identifying health status using machine learning algorithms. By effectively identifying the health state of different models, the success rates of health status identification may create the perfect health status identification model. Here the health index value is calculated by BMI using height and weight of person. The health index value 1, 2, 3, 4 and 5 denotes the health status type with extremely obesity, obesity, overweight, Normal, Weak respectively.

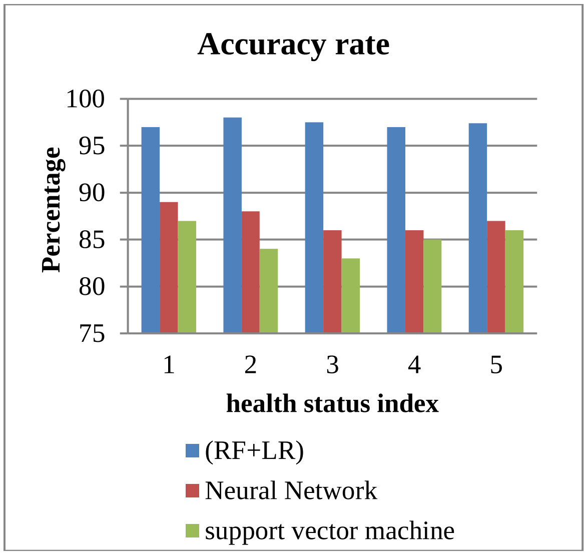

Fig. 2: PERFORMANCE IN TERMS OF ACCURACY RATE

Fig. 3 shows the graphical representation of Identification Time in milliseconds. As can be shown, the machine learning algorithm (RF+LR) has a 13.4 millisecond health recognition time; the support vector machine takes an average of 32.4 milliseconds to recognize a building's health, while the neural network's average building health recognition time is 26.2 milliseconds. Therefore health status identification time is less for machine learning algorithm (RF+LR).

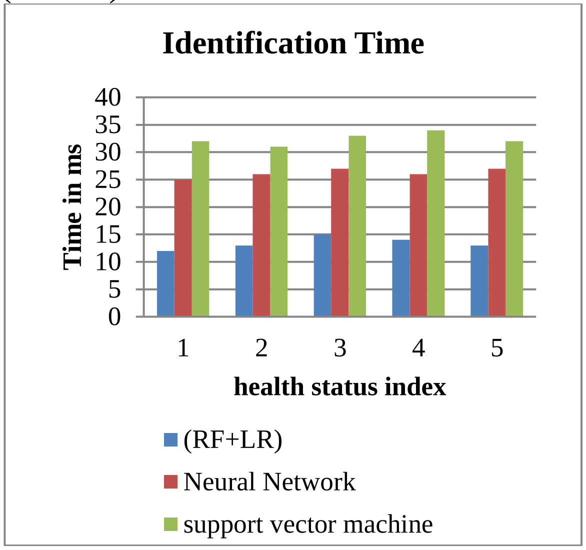

Fig. 3: PERFORMANCE IN TERMS OF IDENTIFICATION TIME

From a dataset of 500 people, table 1 shows the total number of people with a specific health conditions, including the extremely obesity with 198 people, obesity with 130 people, typical with 69 people, overweight with 68 people, weak with 22 people, and extremely weak with 13 people. Fig. 4 represents the Linear Regression and Random Forest model-based pie chart for identifying health status.

Table 1: PERSONS WITH HEALTH STATUS TYPE

Extreme Obesity	198
Obesity	130
Normal	69
overweight	68
Weak	22
Extreme Weak	13

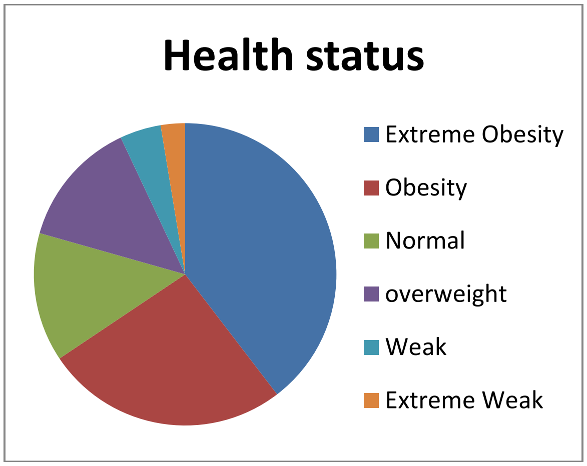

Fig. 4: HEALTH STATUS IDENTIFICATION

Therefore from results it is clear that, the high Accuracy rate and less Identification Time is achieved by described Health Status identification using Linear Regression and Random Forest model.

V. CONCLUSION

In this study, describes the Data Enhancement Method to Health Status Indication Using Linear Regression and Random Forest. Recognizing a person's health status is important for extending a person's or patient's life. Preprocessing, Data collection and exploration, preprocessing, feature selection and extraction, method selection, and model evaluation are some of the components that constitute this process. Using a data set of 500 individuals (without a single variable being missing), they evaluated the accuracy of the BMI calculation using the Random Forest and Linear Regression models. The notes contain words and medical terminologies are quantified through the feature extraction and selection process. The most popular techniques are Term frequency-inverse document frequency (TF-IDF) and Bag of words (BOW). Accuracy rate and identification Time are two parameters used in this study for performance analysis. They compare the effectiveness of the model that is presented with that of the support vector machine and which neural network health status identification technique is most effective. From results it is clear that, the high Accuracy rate and less Identification Time is achieved by described Health Status identification using Linear Regression and Random Forest model.

VI. REFERENCES

[1] Oritsetimevin Arueyinzho, Korede Sanyaolu, "Digital Health Promotion For Fitness Enthusiasts In Africa", 2022 IEEE International Conference on Digital Health (ICDH), Year: 2022

[2] Mengshan Jia, Ziyuan Qi, Deqing Xue, Chang'an Zhu, "Health status Evaluation of Bearing Based on DE-ELM", 2022 IEEE 2nd International Conference on Data Science and Computer Application (ICDSCA), Year: 2022

[3] Tong Cui, Yuxuan Chen, Jiahao Wang, Haoran Deng, Yuchen Huang, "Estimation of Obesity Levels Based on Decision Trees", 2021 International Symposium on Artificial Intelligence and its Application on Media (ISAIAM), Year: 2021

[4] Cheng Chen, Peng Ning, Yibin Zeng, Xizhe Bao, "Research on the Correlation between Body Mass Index and Physical Health Index of Medical College Students", 2021 International Conference on Information Technology and Contemporary Sports (TCS), Year: 2021

[5] Ou Li, Bailin Liu, "Evaluation Model of Health status of Complex Equipment Based on Degradation Degree", 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA), Year: 2020

[6] Zhe Wang, Feng Tang, Zhu Liang Yu, "Design and Implementation of a Health status Reporting System Based on Spring Boot", 2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE), Year: 2020

[7] Gregory Yauney, Aman Rana, Lawrence C. Wong, Perikumar Javia, Ali Muftu, Pratik Shah, "Automated Process Incorporating Machine learning Segmentation and Correlation of Oral Diseases with Systemic health", 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Year: 2019

[8] Mian Yan, Ting Qu, Congdong Li, Suxiu Xu, "Impacts of Health information technology on Health care quality in hospital-related settings: A systematic review", 2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC), Year: 2018

[9] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, "Detecting depression with audio/text sequence modeling of interviews." in Interspeech, 2018, pp. 1716–1720.

[10] A. Sau and I. Bhakta, "Predicting anxiety and depression in elderly patients using machine learning technology," Healthcare Technology Letters, vol. 4, no. 6, pp. 238–243, 2017

[11] Gehrmann, S., Dernoncourt, F, Li, Y., Carlson, E.T., Wu, J.T., Welt, J., Foote, J., Jr., Moseley, E.T., Grant, D.W., Tyler, P.D. "Comparing rule-based and deep learning models for patient phenotyping". arXiv 2017, arXiv:1703.08705.

[12] D. LaFreniere, F. Zulkernine, D. Barber and K. Martin, "Using machine learning to predict hypertension from a clinical dataset," 2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens, 2016, pp. 1-7.https://doi.org/10.1109/SSCI.2016.7849886.

[13] Todor Ivascu, Ovidiu Aritoni, "Real-time health status monitoring system based on a fuzzy agent model", 2015 E-Health and Bioengineering Conference (EHB), Year: 2015

[14] Dongmei Chen, Max Q.-H. Meng, "Health status detection for patients in physiological monitoring", 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Year: 2011

[15] Sohn, S. Savova, G.K. Mayo "clinic smoking status classification system: Extensions and improvements", In Proceedings of the AMIA Annual Symposium Proceedings, San Francisco, CA, USA, 14–18 November 2009; American Medical Informatics Association: Bethesda, MD, USA, 2009; Volume 2009, p. 619.