Volume 9, Issue 3, March – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR211

Heart Disease Detection Using AI

Narannagari Chaathurya (Student)
Artificial Intelligence and Machine Learning
Sphoorthy Engineering College (JNTUH)
Hyderabad, India

Sikharam Abhinav (Student)
Artificial Intelligence and Machine Learning
Sphoorthy Engineering College (JNTUH)
Hyderabad, India

Battu Sri Vamshidhar (Student)
Artificial Intelligence and Machine Learning
Sphoorthy Engineering College (JNTUH)
Hyderabad, India

Kandula Revathi (Assistant professor)
Artificial Intelligence and Machine Learning
Sphoorthy Engineering College (JNTUH)
Hyderabad, India

Abstract:- Over the past few decades, cardiovascular disease has emerged as the primary cause of death worldwide in both industrialized and developing nations. Early detection of heart problems and continued clinical monitoring can reduce death rates. However, because it takes more time and experience, it is not possible to accurately detect heart disorders in all cases and to have a specialist talk with a patient for 24 hours. We demonstrate how machine learning can be used to estimate an individual's risk of developing heart disease. This study presents data processing, which includes converting categorical columns and working with categorical variables. We outline the three primary stages of developing an application: gathering datasets, running logistic regression, and assessing the properties of the dataset. The random forest classifier technique is developed to diagnose cardiac problems more precisely. Data analysis is needed for this application since it is considered noteworthy. The random forest classifier algorithm, which improves the accuracy of research diagnosis, is next covered, along with the experiments and findings.

Keywords:- *Artificial Intelligence; Early Detection; Machine Learning; Heart Disease Detection; Data Analysis.*

I. INTRODUCTION

This paper discusses the relevance of Python programming language in healthcare applications, specifically in the development of dynamic and scalable solutions for heart disease detection, and the significance of machine learning in cardiac disease diagnosis and prediction. Python's versatility and rich ecosystem of libraries indeed make it a popular choice for such tasks.

In the context of heart diseases, machine learning models can analyse various medical data to predict the likelihood of a patient having a heart condition. By leveraging libraries like Pandas for data manipulation, Matplotlib for data visualization, developers can build robust predictive models. These models can process diverse data sources, including patient demographics, medical history, and diagnostic test results, to provide valuable insights to healthcare professionals.

The use of Python in healthcare extends beyond just predicting heart diseases; it also facilitates the development of applications for managing patient records, analysing medical imaging data, and even assisting in surgical procedures. The language's ease of use and extensive community support contribute to its widespread adoption in the medical field.

Overall, the combination of machine learning techniques and Python programming offers promising opportunities for improving healthcare outcomes, particularly in the early detection and management of heart diseases.

II. PROBLEM STATEMENT

It's evident that the healthcare sector is increasingly relying on data-driven insights to improve patient care and optimize healthcare delivery. Python, with its versatility and extensive libraries, plays a crucial role in extracting valuable insights from healthcare data.

Regarding heart disease, Python can aid in analysing various factors such as cholesterol levels, patient demographics, and medical history to predict and diagnose conditions like coronary artery disease (CAD). As mentioned, CAD often goes undetected in its early stages, making predictive analytics especially valuable for early intervention.

Moreover, Python's capabilities extend to ensuring compliance with regulations like HIPAA, which are paramount in handling sensitive healthcare records. With built-in tools for software-defined security, Python helps healthcare projects adhere to strict data protection standards.

Machine learning algorithms further enhance healthcare analytics by enabling the development of tracking and health monitoring applications. Python's ease of use and robust libraries make it an ideal choice for building these applications, ultimately leading to better patient outcomes.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR211

Python's prominence in the healthcare sector stems from its ability to handle complex data analysis tasks, ensure data security, and facilitate the development of innovative healthcare solutions, including those aimed at detecting and managing heart disease.

III. LITERATURE REVIEW

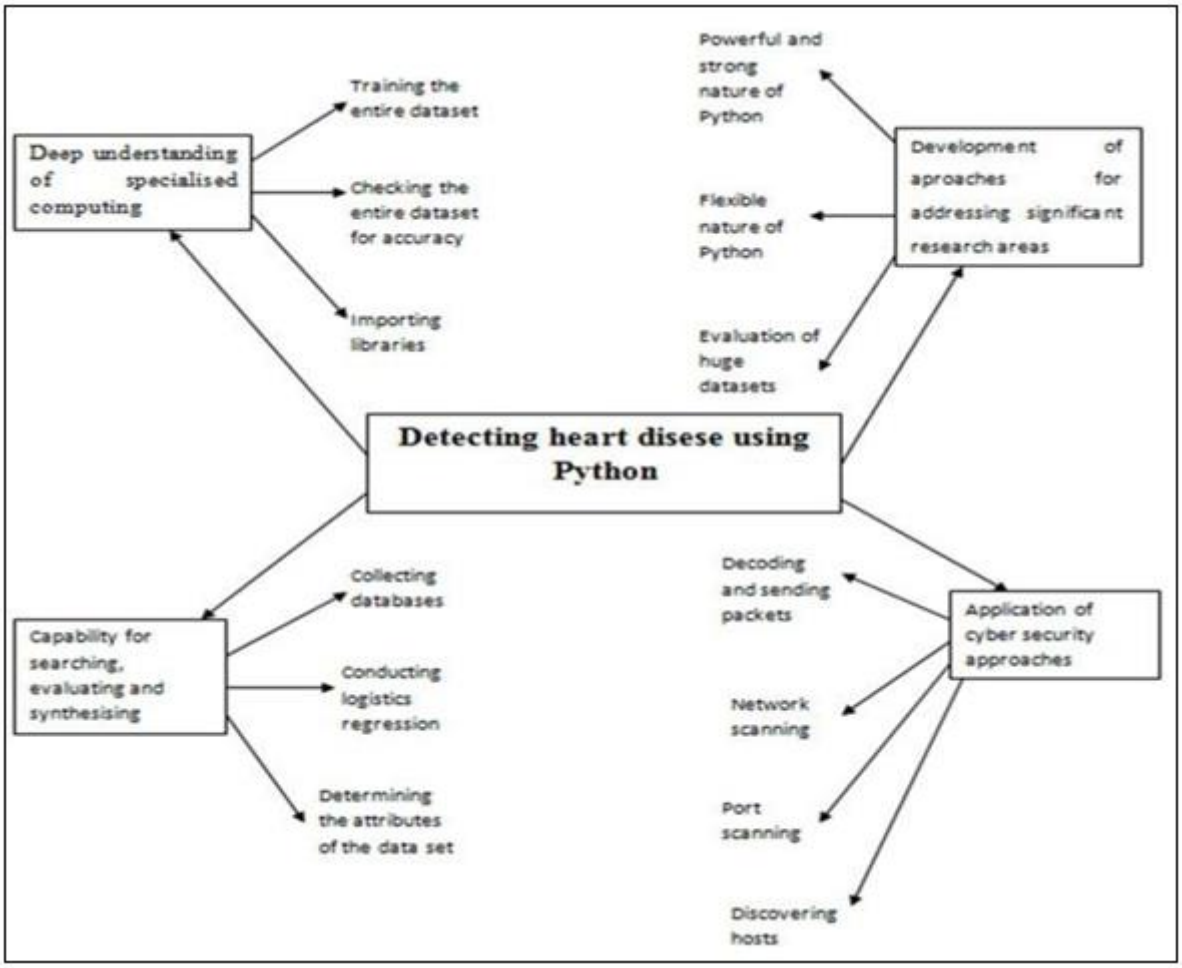

Fig. 1. Theoretical Framework.

It's important to understand the risk factors associated with heart disease, as well as the symptoms that may indicate a heart problem. Diabetes, obesity, unhealthy diet, overweight, excessive alcohol use, and physical inactivity are all significant contributors to heart disease risk. And while chest pain is a common symptom and often a warning sign of cardiovascular issues, it's important to note that symptoms like nausea, indigestion, heartburn, or stomach pain can also sometimes be associated with heart problems, particularly in women. It's always crucial to pay attention to any unusual symptoms and consult a healthcare professional if one has concerns about heart health.

Additionally, employing machine learning techniques can assist in diagnosing and predicting heart disease based on the relevant features in the dataset and patient information. Using a correlation matrix can help identify relationships between different variables related to heart disease, while histograms can provide insights of distribution of these variables within the dataset. The dataset comprises of several factors, such as, age, sex, cp, trestbps, Chol, fbs, and others. Libraries such as NumPy, pandas, matplotlib and scikit-learn were used.

Machine learning classifiers such as K Neighbors Classifier, Random Forest Classifier, Logistic Regression, and Decision Tree Classifier can be applied to predict heart disease based on input features. These algorithms can learn from historical data to classify new instances into different categories, such as presence or absence of heart disease. Hybrid methods, involve combining multiple algorithms or techniques to improve the accuracy and robustness of the predictive models. This could include integrating logistic regression, K-nearest neighbor, and neural networks to develop more comprehensive heart disease diagnostic algorithms. It's an exciting and important area of research and application, as artificial intelligence and machine learning can potentially enhance medical diagnosis and treatment by leveraging large datasets and advanced algorithms to extract meaningful insights.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR211

IV. METHODOLOGY

Developing heart disease detection using AL involves a systematic methodology to ensure the model's effectiveness, reliability, and ethical considerations. Here's a step-by-step methodology:

A. Objectives

Critically evaluate the methods used to acquire and pre-process the heart disease data. Assess the quality of the data sources, including their reliability, completeness, and representativeness.

Evaluating the effectiveness of feature selection methods in identifying informative features and reducing dimensionality while preserving predictive power.

Critically assessing the reliability and generalization capability of the models.

B. Data Collection and Exploration:

Gather relevant datasets containing features related to heart disease, such as demographic information, medical history, and diagnostic test results. Explore the data to understand its characteristics, identify potential biases or missing values, and gain insights into feature distributions and relationships.

C. *Data Pre-processing:*

- Handle missing numbers, outliers, and inconsistent data to make the data cleaner.
- Encode categorical variables and perform feature scaling or normalization as necessary.
- Split the data into training, validation, and testing sets to facilitate model training and evaluation.

D. Feature Selection and Engineering:

- Select informative features that are likely to be predictive of heart disease based on domain knowledge and data exploration.
- Engineer new features or transformations to enhance the predictive power of the model.

E. Model Selection:

- Choose appropriate ML algorithms for heart disease detection based on the nature of the problem, data characteristics, and computational resources.
- Consider a variety of algorithms, such as logistic regression, decision trees, random forest, support vector machines, neural networks, or ensemble methods.

F. *Model Training and Evaluation:*

- Train the selected ML models using the training data and validate their performance using the validation set.
- Evaluate the models using appropriate evaluation metrics, such as accuracy, precision, recall, F1-score, area under the ROC curve, or confusion matrix.
- Perform cross-validation to assess the models' generalization performance and robustness.

G. Model Interpretation and Validation:

- Interpret the trained models to understand their decision-making process and identify important features for heart disease detection.
- Validate the models with domain experts to ensure their clinical relevance and interpretability.

V. PROPOSED SYSTEM

A. Description of Suitable Libraries:

Python's popularity in the programming community is well-deserved, especially in the realm of data science and machine learning. Let's delve into the purposes of the libraries you mentioned:

- *Numpy:*

This library is fundamental for numerical computing in Python. Large, multi-dimensional arrays and matrices are supported, and a number of mathematical operations are available for effective manipulation of these arrays. The foundation of many other libraries in the ecosystem of scientific computing is NumPy.

- *Pandas:*

Pandas is a powerful data manipulation and analysis library. It provides data structures like Data Frame and Series that make it easy to work with structured data, perform data cleaning, manipulation, and analysis tasks. pandas is particularly useful for handling tabular data, such as CSV files or SQL database tables.

- *Scikit-Learn:*

Scikit-learn is one of the most popular machine learning libraries in Python. It offers a wide range of supervised and unsupervised learning algorithms, along with tools for model selection, evaluation, and pre-processing. scikit-learn's user-friendly interface makes it accessible for both beginners and experts in machine learning.

- *Matplotlib:*

Matplotlib is a plotting library for creating static, interactive, and animated visualizations in Python. It provides a MATLAB-like interface for creating plots and charts, making it easy to generate a wide variety of graphical representations of data. matplotlib is often used in combination with NumPy and pandas for visualizing data and analysis results.

- *Seaborn:*

For Python statistical graphics plotting, Seaborn is an incredible visualization library. In order to enhance the visual appeal of statistical graphs, it offers lovely default styles and color schemes. The data structures from Pandas are strongly interwoven with the matplotlib library upon which it is developed.

These libraries, when combined, form a powerful toolkit for data scientists and machine learning practitioners, enabling them to efficiently explore, manipulate, analyse, and model data, as well as visualize their findings. Their versatility and extensive documentation make them

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR211

indispensable assets in the field of data science and artificial intelligence.

B. Description of Suitable ML Algorithms:

Using Machine Learning algorithms like Random Forest, k-Nearest Neighbors, Decision tree Classifier, utilized to classify heart disease risk. Correlation matrix analysis helps identify the relationships between different variables and heart disease indicators.

➢ *Decision Tree:*

Decision tree algorithms are employed in heart disease detection to create predictive models based on splitting data into hierarchical decision nodes. Decision trees are intuitive models that recursively split the data based on features, aiming to create homogeneous subsets that are more predictive of the target variable—in this case, the presence or absence of heart disease.

➢ *K-Nearest Neighbor:*

KNN (k-Nearest Neighbors) is utilized in heart disease detection by classifying patients based on the majority class of their nearest neighbors. It is a non-parametric, lazy learning algorithm that classifies data points based on the majority class among their nearest neighbors.

➢ *Logistic Regression:*

Logistic regression is a type of regression analysis used for predicting the probability of a binary outcome (such as the presence or absence of heart disease) based on one or more predictor variables**.** It is employed in heart disease detection to model the probability of patients having heart disease based on their characteristics.

➢ *Random Forest:*

The random forest algorithm is utilized in heart disease detection to build an ensemble of decision trees, where each tree is trained on a random subset of the data and votes on the final classification, providing robustness and accuracy in prediction. It's often used for classification tasks, including heart disease detection.

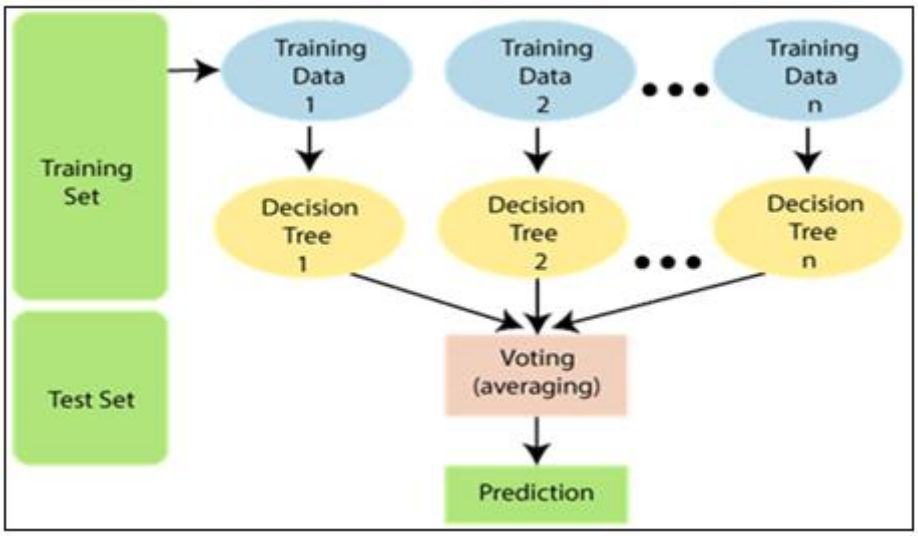

Fig. 2. Random Forest Algorithm.

VI. DATA DESCRIPTION

The following URL contains the dataset used to identify and analyse heart diseases:
https://archive.ics.uci.edu/ml/datasets/heart+disease

Age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, and num are among the variables that are employed. In this project, a total of 14 factors that can improve the accuracy of heart disease detection are taken into consideration. This dataset contains 14 columns of sample data from 302 patients. For the purpose of detecting heart disease, every trait is equally important. The data is split between 70% and 30% of the 303 samples. A total of 212 samples i.e., 70% of the data, are used to train the model. The remaining 91 samples i.e., 30% of the data, are used to test the model.

One crucial step in the pre-processing of data is dividing the dataset into train and test sets. This allows us to enhance the predictability of our model and increase its performance. It is possible to think of it this way: if we use a training set to train our model and an entirely different test dataset to test it on, our model will be unable to recognize the correlations between the features.Consequently, the model's performance will suffer if we use two different datasets for training and testing. Therefore, it is crucial to divide a dataset into a train set and a test set. This makes it simple for us to assess how well our model is performing. For example, if the model performs well on training data but poorly on the test dataset, it's possible that the model is overfitted.

The train_test_split function from scikit-learn can be used to split the dataset.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR211

- *Training Data:*

The first set of data used to build a machine learning model—from which the model builds and improves its rules—is referred to as training data. The quality of this data has a significant impact on how the model is developed going forward, establishing a strong standard for all applications that use the same training set in the future.

- *Testing Data:*

The model needs to be tested using the test dataset after it has been trained using the training dataset. Using the new or unknown dataset, this dataset assesses the model's performance and guarantees that it can generalize successfully. A different subset of the original data that is unrelated to the training dataset is the test dataset.

As is customary, the patient's age is recorded in years, their sex is recorded as 0/1, with 1 being a male patient and 0 representing a female patient. The type of chest pain (CP) is recorded as 1, 2, 3, and 4. One for typical angina pain, two for atypical angina, three for non-anginal pain, and four for asymtomatic. Upon admission to the hospital, the patient's resting blood pressure (trestbps) is measured in millimeter-Hg. Chol is measured in milligrams per deciliter of serum cholesterol. The fasting blood sugar, or fbs, is entered as 0/1 and should be greater than 120 mg/dl. 1 denotes true and 0 falsehood. Resting electrocardiogram (resting ECG) findings are entered as 0/1/2, where 0 indicates normal, 1 indicates ST-T wave abnormalities (T wave inversions and/or ST elevation or depression of > 0.05 mV), and 2 indicates probable or definitive left ventricular hypertrophy according to Estes' criteria. The highest heart rate attained is thalach. Exang is the angina exercise that accepts input values of 0/1. 0 means no and 1 means yes. Oldpeak, as opposed to rest, is the ST depression brought on by exertion.

Slope, which accepts 1/2/3 as input, is the slope of the peak exercise ST segment. Three for downsloping, two for flat, and one for upsloping. Approximately for the number of large vessels (0–3), colored by flourosopy. The thalasemia indicator, or thal, has three values: 3 for normal, 6 for fixed defects, and 7 for reversible defects. The diagnosis of cardiac illness, num or target, accepts an input of 0/1. A diameter narrowing of 0 means less than 50%, whereas a narrowing of 1 means more than 50%.

VII. CONCLUSION

The discussion revolves around utilizing Python for heart disease prediction and detection. Python is highlighted as an object-oriented, high-level programming language with quick development cycles and robust building options. This language is deemed beneficial for accurately predicting the pathway of heart disease due to its attributes. The healthcare industry, particularly the heart care sector, is emphasized as actively generating data from various facilities and patients, leveraging effective data strategies. Additionally, doctors are depicted as utilizing superior predictive models for treatments, thereby enhancing the overall healthcare delivery system. The prediction model for heart disease is specifically mentioned as being employed by clinicians and institutions to improve patient outcomes through scalable and dynamic applications. Chapter two further delves into the application of Python for detecting the presence of heart diseases, utilizing a dataset containing patient data such as age, sex, cholesterol levels (Chol), and other relevant factors.

Individual libraries such as Matplotlib, NumPy, Pandas, warnings, and others are imported for use in the heart disease detection application. Python, being a robust language, is noted for its computational capabilities, facilitating the extraction of valuable insights from patient information related to heart diseases. Additionally, it is emphasized that Python complies with HIPAA regulations, ensuring the safety of medical information. Machine learning is underscored as crucial for predicting threats like heart disease. Specifically, the Random Forest algorithm is chosen for developing the heart disease detection methodology. Furthermore, it's mentioned that the ML model, particularly Random Forest classification, plays a significant role in achieving accuracy and determining results using training data. The selection of Random Forest is based on the specific dataset, as well as its comparison with the decision tree algorithm. Data analysis is highlighted, particularly the handling of categorical variables by breaking them into dummy columns with binary values (1s and 0s). The output of the application includes medical parameters such as age, gender, blood pressure, cholesterol, and obesity, which are used for prediction and software development requirements. Machine learning application using Python is described as a subset of the Artificial Intelligence model, with Python libraries being essential for making predictions. The Scikit-learn (SKLEARN) library is specifically mentioned as commonly used in machine learning prediction tasks. Random Forest is identified as the preferred algorithm for predicting heart disease due to its simplicity and ability to produce precise results.

REFERENCES

[1]. L. Loku, B. Fetaji, A. Krstev, M. Fetaji, Z. Zdravev, Using python programming for assessing and solving health management issues, South East Eur. J. Sustain.Dev. 4 (1) (2020). https://eprints.ugd.edu.mk/27485/

[2]. P. Guleria, M. Sood, Intelligent learning analytics in healthcare sector using machine learning, in: Machine Learning with Health Care Perspective, Springer,Cham, 2020. https://link.springer.com/chapter/10.1007/978-3-030-40850-3_3

[3]. Spencer R., Thabtah F., Abdelhamid N., Thompson M. Exploring feature selection and classification methods for predicting heart disease. *Digital Health* . 2020. https://journals.sagepub.com/doi/10.1177/2055207620914777

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24MAR211

[4]. Javeed A., Zhou S., Yongjian L., Qasim I., Noor A., Nour R. An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection. *IEEE Access* . 2019. https://ieeexplore.ieee.org/document/8894128

[5]. Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G.Y.H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach. *Cardiovasc. Diabetol.* 2022. https://link.springer.com/article/10.1186/s12933-022-01672-9

[6]. Ouf, S.; ElSeddawy, A.I.B. A proposed paradigm for intelligent heart disease prediction system using data mining techniques. *J. Southwest Jiaotong Univ.* 2021. http://jsju.org/index.php/journal/article/view/949

[7]. A. Zahariev, M. Zveryakov, S. Prodanov, G. Zaharieva, P. Angelov, S. Zarkova, M. Petrova, Debt management evaluation through support vector machines: on the example of Italy and Greece, Entrepreneurship Sustain. Issues 7 (3) (2020) 1–12. https://www.sciencedirect.com/science/article/abs/pii/S0300483X17303451

[8]. Bhunia, P.K.; Debnath, A.; Mondal, P.; D E, M.; Ganguly, K.; Rakshit, P. Heart Disease Prediction using Machine Learning. *Int. J. Eng. Res. Technol.* 2021. https://scholar.google.com/scholar_lookup?title=Heart+Disease+Prediction+using+Machine+Learning&author=Bhunia,+P.K.&author=Debnath,+A.&author=Mondal,+P.&author=D+E,+M.&author=Ganguly,+K.&author=Rakshit,+P.&publication_year=2021&journal=Int.+J.+Eng.+Res.+Technol.&volume=9

[9]. Hassan, C.A.U.; Iqbal, J.; Irfan, R.; Hussain, S.; Algarni, A.D.; Bukhari, S.S.H.; Alturki, N.; Ullah, S.S. Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers. *Sensors* 2022. https://www.mdpi.com/1424-8220/22/19/7227

[10]. Subahi, A.F.; Khalaf, O.I.; Alotaibi, Y.; Natarajan, R.; Mahadev, N.; Ramesh, T. Modified Self-Adaptive Bayesian Algorithm for Smart Heart Disease Prediction in IoT System. *Sustainability* 2022. https://www.mdpi.com/2071-1050/14/21/14208

[11]. P. Mathur, Overview of machine learning in healthcare, in: Machine Learning Applications using Python, A Press, Berkeley, CA, 2019. https://link.springer.com/chapter/10.1007/978-1-4842-3787-8_1

[12]. A. Navlani, Understanding random forests classifier in python, DataCamp (2018) Available at: https://www.datacamp.com/community/tutorials/random-forestsclassifier-python, [Accessed on 5th March, 2021]. https://www.datacamp.com/tutorial/random-forests-classifier-python

[13]. D. Pedrozo, F. Barajas, A. Estupiñán, K.L. Cristiano, D.A. Triana, Data analysis for a set of university student lists using the k-Nearest Neighbors machine learning method, J. Phys. Conf. Ser. 1514 (1) (2020) 1–8. https://iopscience.iop.org/article/10.1088/1742-6596/1514/1/012011/meta

[14]. A. Anees, I. Hussain, A novel method to identify initial values of chaotic maps in cybersecurity, Symmetry 11 (2) (2019) 140. https://www.mdpi.com/2073-8994/11/2/140

[15]. Y. Fan, J. Li, D. Zhang, J. Pi, J. Song, G. Zhao, Supporting sustainable maintenance of substations under cyber-threats: An evaluation method of cybersecurity risk for power CPS, Sustainability 11 (4) (2019) 1–30. https://www.mdpi.com/2071-1050/11/4/982