Journal of Ambient Intelligence and Humanized Computing (2023) 14:10347–10362
https://doi.org/10.1007/s12652-022-03694-3

ORIGINAL RESEARCH

Applicability of classifier to discovery knowledge for future prediction modelling

Ritu Chauhan[1] · Eiad Yafi[2,3]

Received: 28 July 2020 / Accepted: 10 January 2022 / Published online: 1 February 2022

Abstract
The immense growth of new technological interventions has forced researchers and scientists around the globe to adopt the widely anticipated technology of Machine Learning (ML) and Artificial Intelligence (AI). ML and AI have generously prospected itself from the past decade in the discovery of knowledge from databases. Several ML and AI based adoptive technologies have emerged in varied application domains, and are thus widely opening a new era of knowledge in decision making. Moreover, ML and AI are techniques that can improve the treatment and diagnosis of diseases. In the current study, we have designed and deployed a "PROCLAVE". The tool was designed in varied layers of structure, where each layer plays a significant role in determining the patterns. We have applied several libraries for the processing of a prototype to develop a visualization interface. The tool forecasts health vulnerability, makes a comparison among variable classifiers and visualize the results for end users. Moreover, the proposed architecture is based on the concepts of conceptualization and visualization to detect the overall dashboard. Furthermore, the current approach was synthesized and populated with a database that allows the end users to select the variable features and relatively determine the interactive patterns for the number of cases. The database was collected from the National Institute of Health Stroke (NIHS) in the United States. Data was gathered for stroke patients who were diagnosed with stroke from 1950 to 2015. The study was based on several attributes which included causes of death, sex, race, Hispanic origin and others to discover unknown patterns for future decision making.

Keywords Machine learning (ML) · Artificial intelligence (AI) · Decision tree · Random forest · Support vector machines (SVM)

1 Introduction

Digital automation has widely aligned varied application domains, but has faced problems including the analysis and knowledge discovery from large scale databases. Moreover, it requires an imputative technology to handle the burden of data and further determines varied patterns for future decision making. In the past, several traditional statistical techniques were utilized to recognize hidden information, but they tend to have a higher error rate due to increase in burden of data. Hence, Machine Learning was designed to explore valid, novel and understandable patterns to identify future knowledge discovery (Ioannis et al. 2017; Quinlan 1986, Nápoles et al. 2014)

Machine learning (ML) has widely grown in capacity from the past decade and has consistently predicted patterns for future decision making. In the current scenario, algorithmic tools are playing a vital role in diagnosing diseases by utilizing several techniques, which include ML, statistics, artificial intelligence (AI), database sets, pattern recognition and visualization for future prediction models (Liu et al. 2019; Vaka et al. 2020; Shi et al. 2018; Barbat et al. 2019; Breiman 2001; Peiffer-Smadja et al. 2020; Sammut and Webb 2011; Prajwala 2015). Moreover, information technology has widely changed the era of medical data-bases and benefited

Ritu Chauhan and Eiad Yafi contributed equally to this work.

✉ Ritu Chauhan
rchauhan@amity.edu

✉ Eiad Yafi
eiad.yafi@uts.edu.au

[1] Amity Institute of Biotechnology, Amity University, Street, Noida, Uttar Pradesh 100190, India

[2] School of Computer Science, University of Technology Sydney, 15 Broadway, Sydney 2007, NSW, Australia

[3] Faculty of Information and Communication Technologies, Institute of Business, Karuhun Bidau, Dili, Timor-Leste

the diagnosis of diseases (Xu et al. 2017; Choi et al. 2017; Menad et al. 2019; Breslow and Aha 1997). In the past, several anticipated data mining techniques were developed to detect hidden patterns and knowledge from healthcare databases. This approach was synthesized to encompass real-world datasets of tuberculosis (TB) patients, where 700 data records were synthesized to diagnose and classify the disease. The learning was based on K means clustering on varied classifiers to determine a better technique for prediction of the data. This technique assisted healthcare practitioners in the analyses of TB's prognosis and diagnosis based on different categories. The above study contributed with a precision of 99.7% for Support Vector Machines (SVM) as compared to the current Neural Network (NN) classifiers by Asha et al. (2012).

However, applying ML techniques on healthcare databases is a trivial task where the challenge is to generate patterns and make real-time clinical decisions (Ioannis et al. 2017). Predictive techniques begin with gener-ating hypothetical models in medical research and the results are made to arbitrarily fit into a hypothesis (Pouriyeh et al. 2017; Joloudari et al. 2019; Otoom et al. 2015). Nevertheless, medical data mining benefit patients as the quality of service will be based on automatic generative decision making, where the errors are reduced and the quality of decision making is significantly improved (Moloud et al. 2018, Lavanya and Rani 2011, Nápoles et al. 2014; Tanwar et al. 2021; Choi et al. 2017; Beck et al. 2020).

In addition, a new adoptive technology has been implemented in ML to generously fetch large amount of data for future analysis. In general, deep learning had outrageously discussed ML to solve patterns for complex data analysis. It has a tendency to deliver a prediction which can predict models with the capability of a superhuman interface. Recent studies have applied deep learning technology in varied application domains to solve varied research questions and generate prediction models with higher accuracy (Chen and Lin 2014; Tanwar et al. 2021; Thomas 2020; Gárate-Escamila et al. 2020; Rong et al. 2020). However, these models usually require higher computational power, where the comprehended technology is vastly studied among researchers and scientists to be improved through several possibilities.

Another study was conducted on the survivors of stroke, whereby data was classified and potential hazard was determined for its occurrence (Pouriyeh et al. 2017). Distribution of hazard factors related to stroke, their relative significance and relationship from the fundamental conduct chance factor dataset were distinguished. From the openly accessible collection across the national information, the last information of stroke in the Assembled States including 397 factors (19) were dissected (Kaur et al. 2015). A mining calculation including C4.5 and a direct relapse with the M5s strategy were connected to build an applicable model of post-stroke impedance utilizing Weka programming, resulting in high accuracy. As effectively comprehended by the researcher's relative significance and pertinence of the 70 factors, confusions were exhibited in Infographics or graphical examinations. It was discovered that 55% of patients after stroke became handicapped. Exercise, business and living fulfillment were moderately critical components identified with the inabilities in the patients. Modifiable behavioral factors emphatically identified with the incapacity were to incorporate exercise and great rest OR 0.37, $P < 0.01$). Data mining was resolved to identify data of inability after stroke from an expansive populace informational collection, that was picked by the variable traits. This finding may possibly be profitable for clinicians and scientists by initiating trainings related to understanding stroke. This strategy can be summed up to other health conditions for further discoveries.

In the current study, we have designed and deployed a "PROCLAVE" tool in python language. The tool was designed in varied layered structure where each layer plays a significant role in determining patterns. The tool was deployed in windows 10 operating system for developing and implementing the entire script. Several packages were configured to develop a visual and analytical interface for significant and rich configuration. We applied several libraries for processing of the prototype to develop a visualization interface. The tool was applicable to forecast the health vulnerability for comparison among variable classifiers and to visualize the results for end users. Moreover, the proposed architecture was based upon conceptualization and visualization to detect the overall dashboard.

PROCLAVE was also applied to determine the significant patterns which can be easily and readily interpreted by the end users. Each layer works on corresponding interfaces where ML plays as interactive module in the second layer, which can proportionally relate data with applicable classifiers. Additionally, varied distribution patterns were determined to correlate the study with applicable features in the databases.

Furthermore, the current approach was synthesized and populated with the database where it allows the end users to select the variable features and relatively determine the interactive patterns for the number of cases. The database was collected from the National Institute of Health Stroke (NIHS) in United States. Data was gathered for stroke patients diagnosed with stroke for the year 1950 to 2015 (ASA 2015; CDC 2015). The study was based on several attributes which included causes of death, sex, race, Hispanic origin and more to discover unknown patterns for future decision making. The observed data was expressed as continuous variables where each attribute tends to relate to each other. Furthermore, data was classified using appropriate classifiers which included decision tree (Patil et al. 2010; Franco-Arcega et al. 2011; Lysaght et al. 2019), random forest (Li et al. 2010;

Robnik-Šikonja 2004; Mosavi et al. 2019; Sharma et al. 2020; El Saghir et al. 2014) and support vector machine (SVM) (Chang et al. 2010). The applicability of each classifier was determined using Python 3.3 to measure the accuracy and regulate the classifier capable of assessing the database for future discovery of knowledge.

Overall, the paper is outlined as follows: Introduction is discussed in Sect. 1, the proposed mod-el of application is presented in Sects. 2, 3 discusses the materials and methods, while results are presented and discussed in Sect. 4, and finally conclusion and discussion are shared in Sect. 5.

2 The proposed "PROCLAVE" model

The proposed architecture is based upon conceptualization and visualization concept to detect the overall dashboard. The framework is based on three layers where first layer is the pre-processing, where the focus is to determine a relative correlated pattern by removing missing values and inconsistencies among the data elements. Moreover, the more deterministic visualization patterns are retrieved at the end of layer. Further, the second layer is termed as data classifier layer which can be conceptualize as the best fit deterministic model to measure the accuracy among each classifier using perpetual modelling and knowledge discovery. Hereby, knowledge discovery is an important deterministic measure which is required to determine the efficacy of model generated. The third layer is based on the visualization layer where each component is based upon user based interactive protocol, each attribute can be easily accessed with the user input and applicable classifier.

2.1 Architecture of PROCLAVE

The entire architecture is based on user interactive mode where the overall procedure is involved with pre-processing data which is gathered through pre-processing layer, furthermore data is applied to measure the best performed model by measuring the accuracy of each classifier applied. Additionally, a special interactive drop-down menu is created to identify and select a particular classifier and attributes which can relatively generate the accuracy. Figure 1 discusses the overall architecture involved for PROCLAVE tool to determine the significant patterns from the database.

2.2 Prototype of PROCLAVE

The proposed PROCLAVE Tool was designed and developed in Python language on the interface of IDE Jupyter lab (McKinney et al. 2010). The windows 10 operating system was utilized for developing and imple-menting the entire script. Several packages were configured to develop a visual and analytical inter-face for a significant configuration rich interface. Several libraries were applied for processing the prototype. The Scikit-learn package (Pedregosa et al. 2011) was for application of ML techniques, Pandas for pre-processing purpose, Plotly to develop a visualization interface and the entire script was run in Jupyter Dash Enterprise (2020). Table 1 represents the pseudo code applied to develop a rich web interface.

2.3 Layer based interface

PROCLAVE was applied to determine the significant patterns which can be easily and readily interpreted by the end users. Each layer works on a corresponding interface where ML plays as an interactive module.

2.3.1 Data pre-processing layer

The pre-processing layer was focused to determine a relative correlated pattern by removing missing values and inconsistencies among the data elements.

2.3.1.1 Data cleaning/ selection Data cleaning is a well-known term in the process of knowledge discovery. It is a rationalized term which utilized for removing and discovering inconsistencies within the database. However, quality of the data is the first step in data processing, so all impor-

Fig. 1 The proposed "PROCLAVE" model

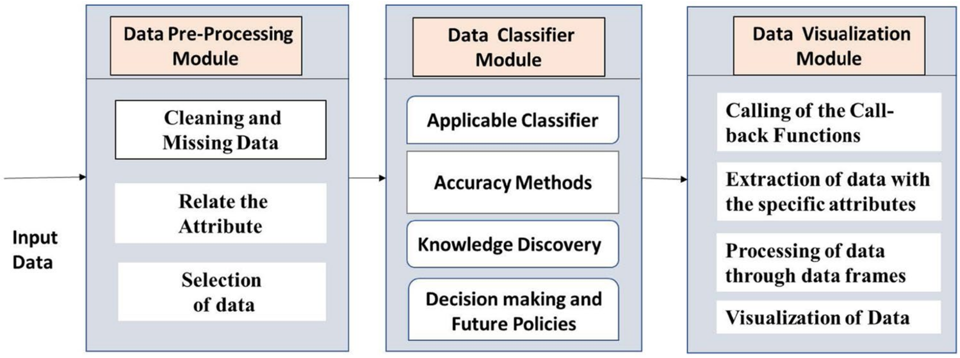

Input Data
Data Pre-Processing Module
Cleaning and Missing Data
Relate the Attribute
Selection of data
Data Classifier Module
Applicable Classifier
Accuracy Methods
Knowledge Discovery
Decision making and Future Policies
Data Visualization Module
Calling of the Call-back Functions
Extraction of data with the specific attributes
Processing of data through data frames
Visualization of Data

Table 1 The pseudo code of the analytics underlying the dashboard

Pseudo-code for functionalities of the analytics dashboard

```
1: Check acquired dataset for missing values and inconsistencies
2: Determine target classes for different diagnosis
3: Select features through selection algorithms
4: Split the dataset into training and test modules with a random sampling
5: if selected classifier is decision tree
6:   Train the model using the training set
7:    Test and cross-validate the model
8:    Send results to dashboard callback methods
9: else if selected classifier is Random Forest
10:    Repeat steps 6 till 8
11: else // the classifier will be Support Vector Machine
12:    Repeat steps 6 till 8
13: end if
```

tant measures should be acquired to maintain the databases. Additionally, database retrieved from vivid resources can be represented in different formats, hence analysis is a trivial process in the same process. In the current study, data was utilized from integrative resources, and while understanding the raw data, an attempt was made to overcome the missing values and inconsistencies among the data. This measure was to design a well conceptualized model which can retrieve hidden effective patterns from healthcare databases. In order to retrieve the data for a model and its proper representation, the database should be accurate to avoid redundancy, missing values and inconsistencies among its data

2.3.1.2 Data classifier module (applicable classifier) The current approach relatively determines and acquire deterministic classifier such as Random For-est, decision tree and SVM. These classifiers could be explored for databases to discover knowledge from which policies may be derived. In a model, each set of features are studied for each classifier to determine the best one to be utilized for future prediction modelling. In the current study's ap-proach, we utilized 3 sets of classifiers (Random Tree, Random Forest and decision tree) to evalu-ate the database, where both test sets and training datasets were explored for viability to achieve significance. A classifier was compared on each set to compare a specific database and its features that impact the global outcome of the study. The Random Forest works on the concept of inter-linked tree like structure where the random vectors concept is utilized with each sub linked tree. It is usually discussed as $\{F(x,\Theta k)k = 1,2,\ldots.\}$, where the $\{\Theta k\}$ is distributed random vectorization, which is identical in nature. Each tree contributes a kind of unity vote which work as an input for the class X. The random forest has a casting nature to en-semble or generate the name of decision trees. Each classifier is further forming a single decision tree. However, SVM is a popular ensemble method used for building predictive models. It can be used for both classifications and regression problems. However, SVM is also termed as a discriminative classifier as its approach is based on separation of hyperplanes. SVM works with labeled training data as the input and hyperplane as the output, thus can distinctly and appropriately classify the data points in the particular space. Similarly, decision tree works on the concept of multi covariates to develop a prediction-based result. The decision tree builds a tree based on the classification of root and internal nodes with no children as leaf nodes. The construction of the tree was based on an inverted form which is a non-complicated structure, to discover patterns. Furthermore, if the size of the data is larger than the amount the tree, the model utilizes the technique of valida-tion and training datasets. As a result, the data size increases, the decision tree utilizes its algorithm to optimize the outcome.

2.3.1.3 Decision trees Decision trees tend to be the best-known applicable technique in data mining, where its application is based on statistical principles such as linear regression and logistic regression. Decision trees share its wide role in the field of cognitive science, which include neural networks to determine appropriate predictive results. The first decision tree was formulated on the basis of mimicking the human method of solving equations and utilizing the same principle, a designed protocol was formulated for the above method. The decision tree modelling is one of the simple ways to utilize multivalent technique to analyze the dataset. They anticipate a wide statistical technique such as multiple line regressions to discover hidden knowledge for future decision making by Esposito et al. (1997). Decision trees are explained as an analysis for partition inner space tool data.

The decision tree works on the concept of a tree, where each tree consists of nodes that form a rooted tree. It also involves nodes with an out coming edge, which refers to as an internal node or a test node. All other nodes are called leaves, terminals or decision nodes. In the decision tree, each internal node divides the instance space into two or more subspaces according to a specific attribute function of the input attribute value. The formation of the tree depends on the majority of selection among the attributes, where one attribute is selected at a time and the nodes with sub-nodes will be considered depending on the other values. For numeric attributes, tree formation depends on the type of method to be used, which attributes are selected within a range of value. Each branch of the tree is assigned to one class, that represents the most appropriate attribute value. Similarly, the leaf denotes the probability value for the attribute. This defines the probability of the target attribute of the dataset to be analyzed. Results of the tree are always classified through an analysis from the root of the tree to its leaves and branches of the test. Every final choice decision from the tree framing calculations takes after the essential rule, i.e, Idea Taking in framework (CLS) (Esposito et al. 1997; Quinlan 1986). The CLS method mimic human process method of data analyzing process. The decryption method is known for understanding the statistical value of tree formation and its analysis. The method works by explaining the two classes and inducing a rule to differentiate in between the two classes based on different attributes' availability. For example, if the training dataset is X with classes (C1, C2, C3.... Cn) the decision tree is built by repeatedly differentiating the query dataset using the splitting principle, until all the records in the partition conclude into a same category. The decision tree for X contains a decision node classifying the test, and one division for each possible outcome is defined by the sub nodes.

As the model of trees are generated, they produce many experimental explanations for the engineering and sciences fields, as well as for the applied areas including business and data mining. Tree modeling shows many advantages as the produced results connect very well in the visual and statistical terms. Choice trees would not be difficult to create, recognize and utilize. The tree shows many functional features in a basic and orderly way, including the display of information. Tree modelling techniques tend to build a tree with varied sets of rules; however, creation of a decision tree utilizes varied measurements simultaneously. These are qualitative and quantitative measurements, positive and negative characteristics, and others, to depict better prediction results.

In a decision tree, the formation of a model depends on data which can easily adapted to the variables, which include offsets, unbalanced effects, nested effects, noise of the datasets, etc. (Vaka et al. 2020; Liu et al. 2019). Moreover, building of the decision tree consists of three main stages, which include the Structure, clipping phase and Pruning phases. The structure phase modulates the datasets in correspondence with the partition criterion to decrease the noise until a ceasing standard is not met. The next phase is Clipping, in which the tree built in the earlier phase may not be able to give a suitable result in the finest possible set of rules, due to over-fitting. The Pruning period dispenses with a portion of more level divisions. Furthermore, the sub partition will enhance the phase's probability of analysis.

2.3.1.4 Random forest A random forest is a collection of non-pruned decision trees, capable of vastly utilizing huge datasets. A random forest tree cab is built of ten to hundred nodes in a single decision tree (Li et al. 2010; Barbat et al. 2019; Robnik-Šikonja 2004). The technique can work efficiently with large datasets with a minimum error rate. Moreover, a random forest can be utilized when dealing with noise in the datasets, as compared to the decision tree. The random forest techniques tend to be very competitive to other methods and can be applied differently with vector machines for future decision making.

Random forest selects the number of trees that needs to be built while choosing the structure size of datasets and its attributes. While working with this technique, there is an option of selecting the number of trees to be built. The random forest adds considerable robustness to various associated problems, while working with the basic tree method through the addition of randomness to the process. Randomness also offers more information about computational analysis, since the aspect can be used for statistical and computational analysis. This tree can also work with regression and classification. Random tree proves to have a better conclusion because of no overfitting and can be used for engineering and other field-related queries. The most important feature of Random Tree is its ability to easily identify the most important feature from the example datasets, to conclude a progressive result. The Random forest implementation works on the following concepts:

Step 1: The number of training cases be "x" and consider the number of attributes included in the classifier be "y".

Step 2: Number of input variables which are used to make prediction at the node of a decision tree be "a". Assume that a is always smaller than y.

Step 3: Select any dataset for making the tree with chosen attributes.

Step 4: To every node of the tree select random attributes on which to search for the best tree formation.

Step 5: Calculate the tree nodes formation depends on these selected variables in the training set

Step 6: Every decision tree is completely produced, not pruned. The tree is fully retained.

Step 7: observation of the data set.

2.3.1.5 Support vector machine (SVM) SVM is a popular ensemble method used for building predictive models. It can be used for both classifications and regression problems. However, SVM is also termed as discriminative classifier as its approach is based on the separation of hyperplanes. SVM works with labeled training data as the input and hyperplane as the output, so that it can distinctly and appropriately classify the data points in the particular space.

In other words, hyperplanes can be discussed to generalize the datasets or classified as a line, to de-termine which data falls on either side of the plane. For example, if the dataset has 2 features, the plane is 1 dimensional, and the planes are discussed with respect to the features present. We can elaborately discuss that SVM are quite effective for high dimensional datasets and can be effectively applied for the n dimensional space. Additionally, they widely utilize the subset of datasets for training, thus decision making and the memory capability are efficient. However, SVM has the overfitting problem, which occurs when the number of relative features exceed as a result of higher sample number.

2.3.1.6 Measure quality The fundamental concept of performance is to measure the quality of classifiers on viable datasets. All aspects of a classifier cannot be gradually covered on the single platform, where a single classi-fier cannot be utilized in all varied databases. Consequently, performance of a classifier is based on the application domain and can be specific to the databases. However, in the past, several studies were based on measuring applicability of the classifier, which may include area under the curve (ROC), F-measure, Kappa Statistics and other quality measures to classify the data. In the current study, utilized the number of True positive divided by certain instances to classify and retrieve a practical overall analysis of classifiers in the application domain. The principal analysis retrieved at each end of the classifier determines the applicability and utilization of each classifier for variable datasets. However, the foremost challenge is to gain insights of each classifier which is well suited for the viable datasets. The statistical techniques were utilized to measure the accuracy and deter-mine significant applicable features. The *F* measure can be discussed as:

Precision can be discussed after repeated measuring of the degree each time the conditions show the same value or results. This means that the value is truly precise in nature. However, it is the percentage of actual results that will be considered for determining the actual values

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive (Actual Results)}} \tag{1}$$

Recall can be discussed as the total results correctly classified by our algoritm:

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative (Predicted Results)}} \tag{2}$$

F-measure can be calculated as mean of both precision and recall,

$$F = 2 * \frac{\text{Precision.Recall}}{\text{Precision+Recall}} \tag{3}$$

Accuracy can be discussed as weighted average value of precision and recall. It takes into consideration both false positive and false values. However, accuracy works significantly with even distribution of class. It can be discussed as below:

$$Accuracy = \frac{\text{true positive} + \text{true negative}}{\text{true positive} + \text{true negative} + \text{false positive} + \text{false negative}} \tag{4}$$

2.3.1.7 Knowledge discovery and decision making The machine learning approach of data mining has extensively given a broad vision to scientists and researchers around the globe to generate patterns and discover hidden information from large scale or big databases. Hence, a major objective is to discover knowledge which can generate information for future operational goals. Data mining is widely discussed as supervised and unsupervised algorithms to discover patterns for future application domains. The supervised learning algorithm retrieves information in the context and with class, which is also known as classification and regression. Unsupervised learning-based algorithms are not based on the class but measures data with similarity or clusters, such as clustering and others. Furthermore, knowledge discovery and discovering meaningful information in similar terms can be utilized for decision making. Knowledge discovered from applicable machine learning algorithms can be used for the decision-making process.

2.3.2 Data visualization layer:

The proposed visualization layer is populated with cleaned and preprocessed data, applicable with a measured classifier to attain the best accurate model among the applied database. The dashboard architecture is based on call-back where several drop-down menus have been created to assure a significant interaction among the user and applied database.

2.3.2.1 Policy making The knowledge from research outcomes is a prerequisite for new research innovations or planning for future beneficial outcomes. Knowledge can be gained via newly implemented technologies, including AI, ML, IoT (Internet of Things) or other newly intervening technologies which has changed the global arena of future predictive modelling. In past statistical model approaches, de-terministic technology was to anticipate and generate new models, but in the new digital era with wide application domains, these approaches have leveraged itself as a boom for new modelling technologies. As a result, new tools and technologies have surfaced for future decision making and competing policies to reform every application domain. The knowledge gained through viable high-tech tools can be utilized to develop future policies and decision making. Hence, ML-based tech-nology can be used as a front step technology to gain insights of data and determine future polices for decision making.

3 Materials and methods

The proposed PROCLAVE Tool was designed and developed in Python language on the interface of IDE Jupyter lab (McKinney et al. 2010). The windows 10 operating system was utilized for developing and imple-menting the entire script. Several packages were configured to develop a visual and analytical interface for a significant configuration rich interface. Several libraries were applied for processing of the prototype, Scikit-learn package was used by Pedregosa et al. (2011) for application of ML techniques, Pandas for pre-processing purpose, Plotly to develop a visualization interface and the entire script was run in Jupyter Dash. The tool was to forecast the health vulnerability among variable classifiers and to visualize the results for end users.

Moreover, an analytical study was conducted for the patients with a stroke diagnosis while correlat-ing the factors which are at potential risk for developing the prognosis of the disease. The datasets for the analysis were collected from the National Institute of Health Stroke (NIHS). The NIHS database had the list of all patients diagnosed as a stroke case in the United States, whereas the yearly data was gathered from 1950 to 2015 for age-adjusted death rates for selected causes of death, sex, race, and of Hispanic origin. The diagnosis of each case was upheld with demographic/socioeconomic factors and correlated to factors influencing the study of research. Furthermore, the data was extracted using data mining to discover hidden and unknown information for future decision making. In current study's approach, decision tree, Random Tree and Random forest were utilized to determine the best technique for measuring the error rate of each classifier and determining the proposed outcomes.

The overall exploration of data represents the varied attributes which can be a root cause or potential risk for the stroke diagnosis. Figure 2 represents the total distribution of age group and race in the database. Here, the data clearly indicates the equal number of cases in each group to relatively encompass the diagnosis at each age.

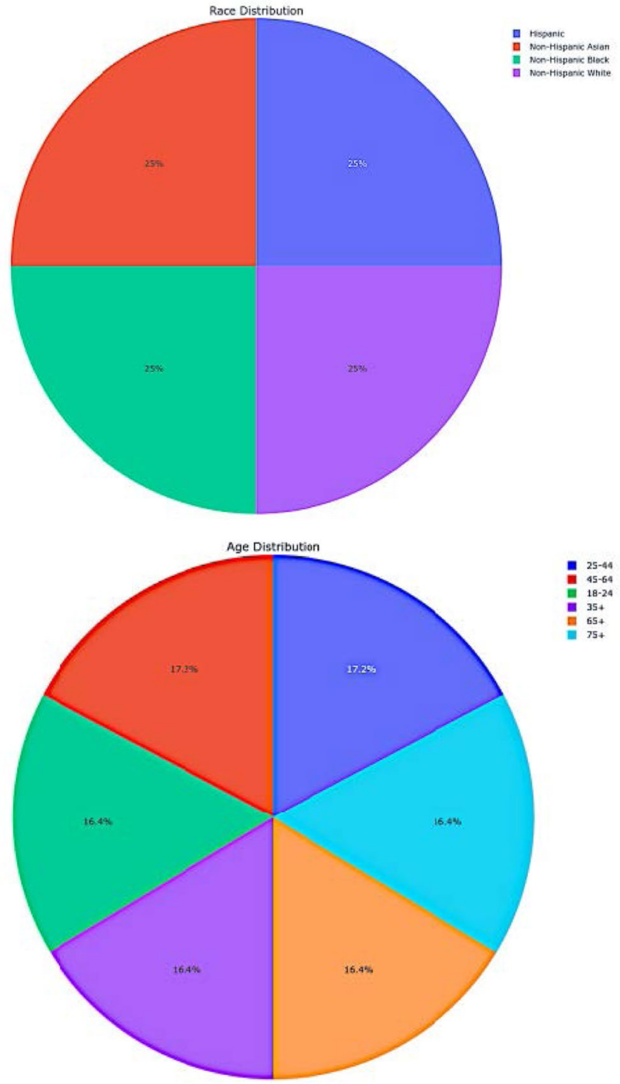


Fig. 2 Representation of the varsity of age and race

The overall number of cases based on gender and its correlation to race was explored. In Fig. 3, the same number of male and female cases were reported for the study were analyzed. Hence, race was categorized into Hispanic, non-Hispanic Asian, non-Hispanic Black and non-Hispanic white. The results were retrieved using a probability distribution using Python.

There is awareness that the real-world healthcare datasets are of extreme variation, however to synthesize the variation between varied parameters, an overall distribution was conducted between age categories and race. In Fig. 4. almost equal number of cases between each age group and correlated with race was found. Hence, the frequency seems to be similar in this case, and actual parameter can be one of the factors for retrieval or prognosis of the disease.

Fig. 3 Overall distribution of race with gender

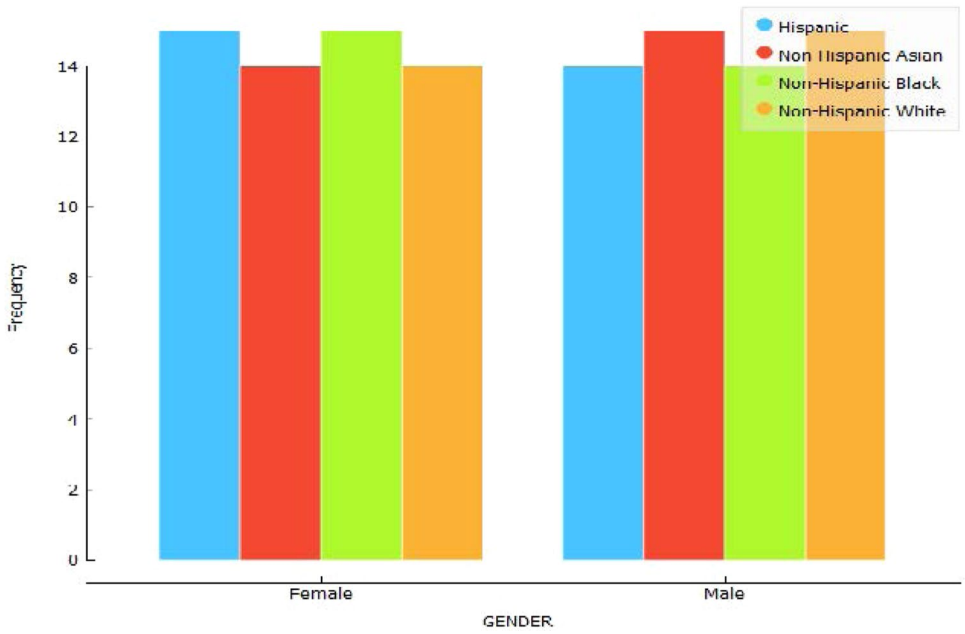
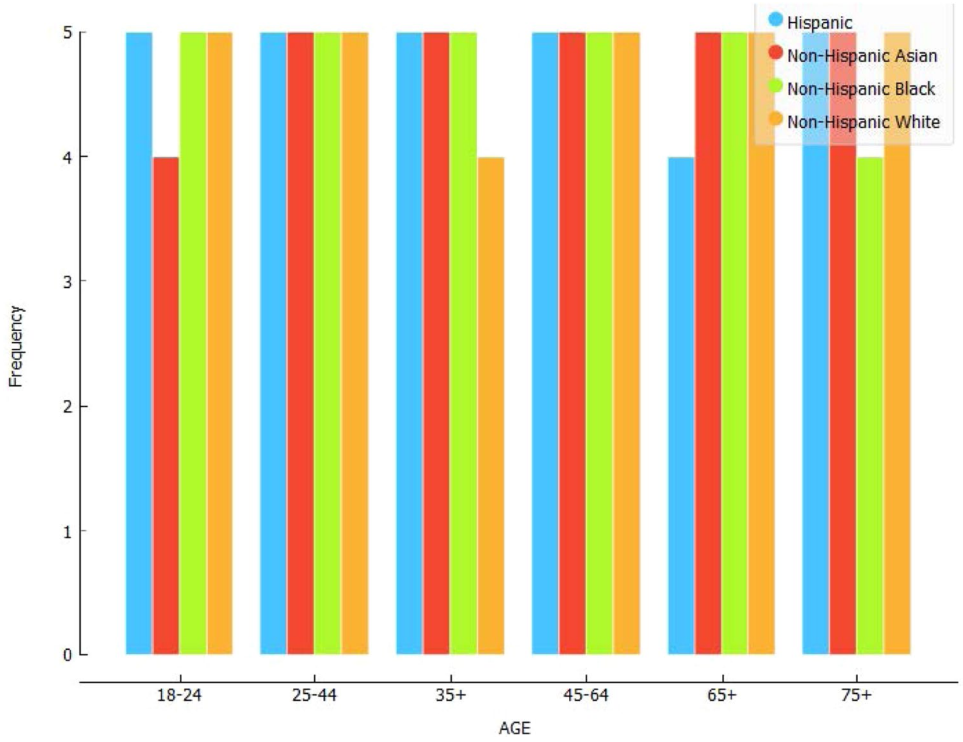


Fig. 4 Representative distribution of age with race



Furthermore, data was analyzed with respect to the year in which cases have gained its number as compared to each race which is exploited. Figure 5 represents the overall cases per year from 2013 to 2016, which was relatively synchronized with race in the population.

Figure 6 represents the age-based calculation with proportion by year. The data relatively represented that each

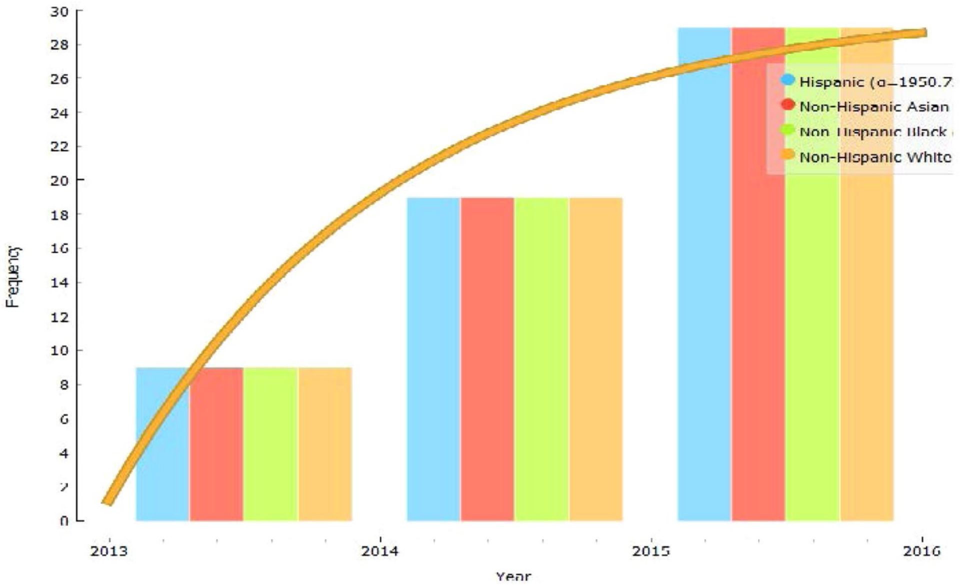


Fig. 5 Overall distribution year wise

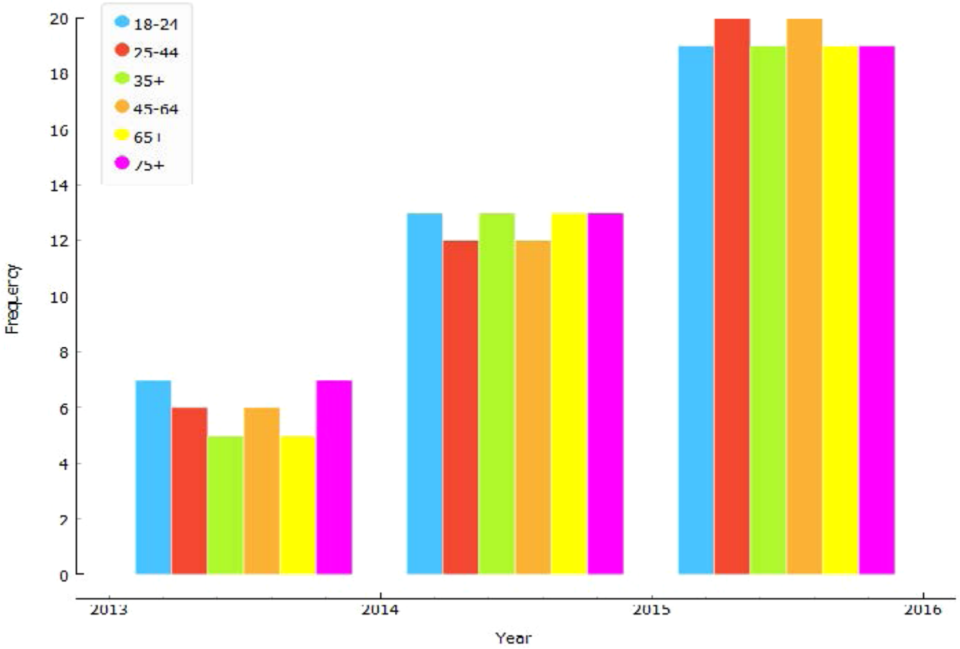
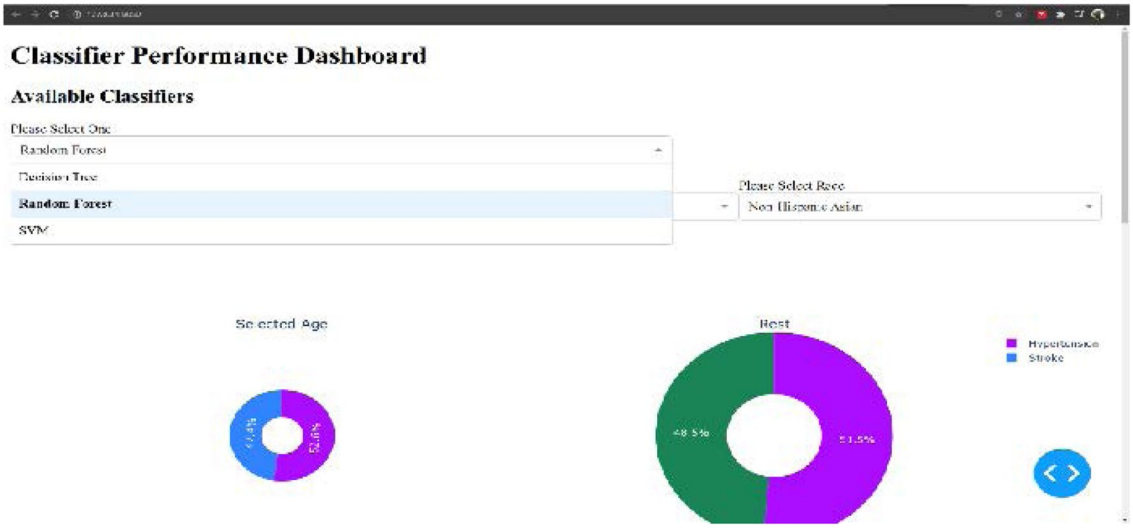


Fig. 6 Age wise calculation with proportion of year wise



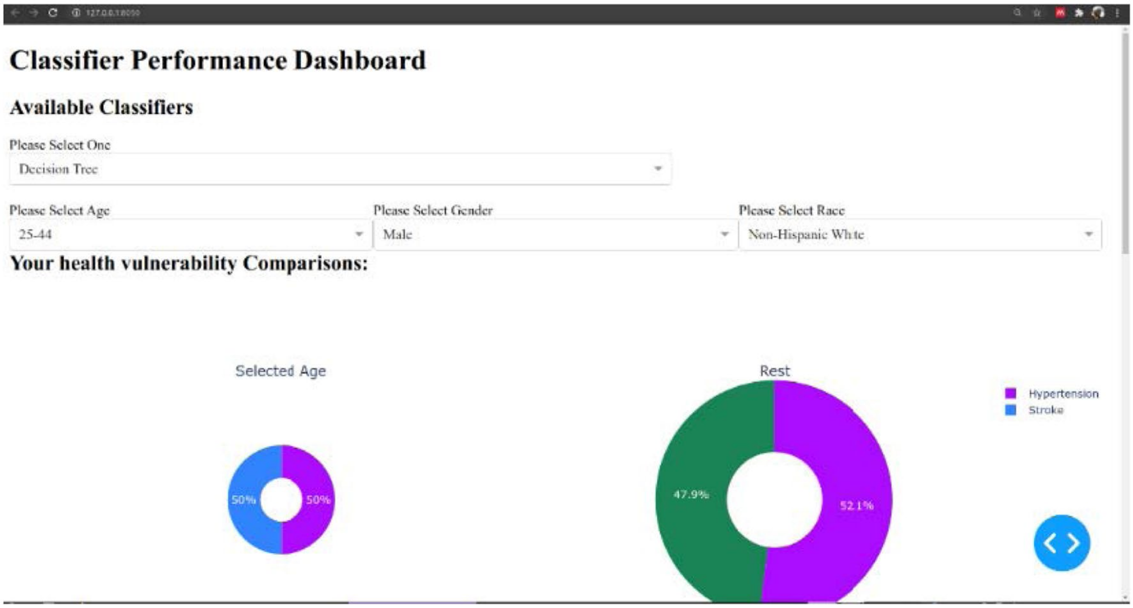
Classifier Performance Dashboard

Available Classifiers

Please Select One: Random Forest

Please Select Race: Non Hispanic Asian

Fig. 7 User derived visual analytical web interface for classifier performance



Classifier Performance Dashboard

Available Classifiers

Please Select One: Decision Tree

Please Select Age: 25-44

Please Select Gender: Male

Please Select Race: Non-Hispanic White

Your health vulnerability Comparisons:

Fig. 8 Represent the decision tree

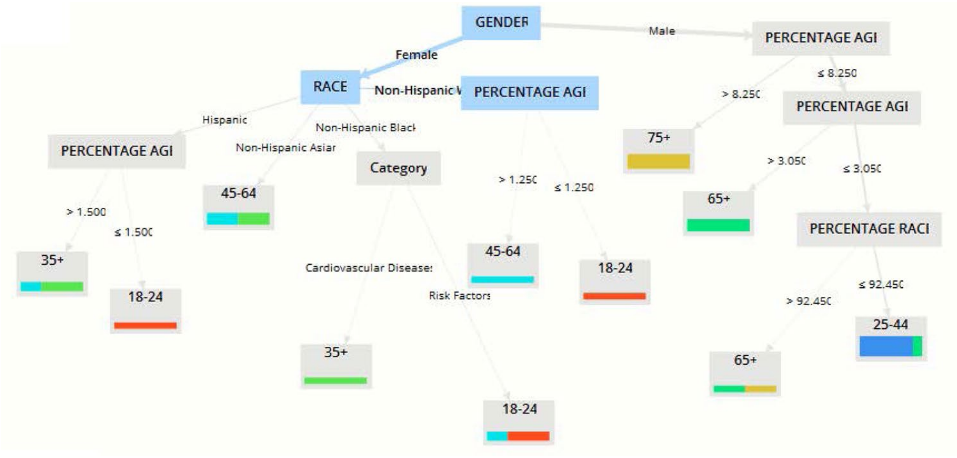


Fig. 9 Decision tree

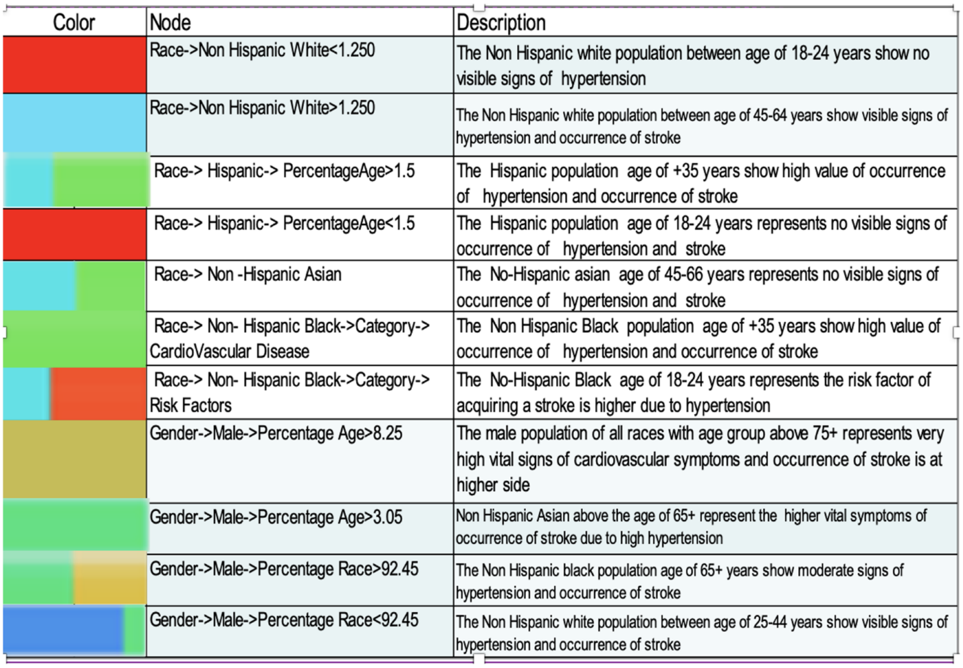
age group had embracingly epitomized the higher rate in each year in accordance to the population studied.

4 Experiment and results

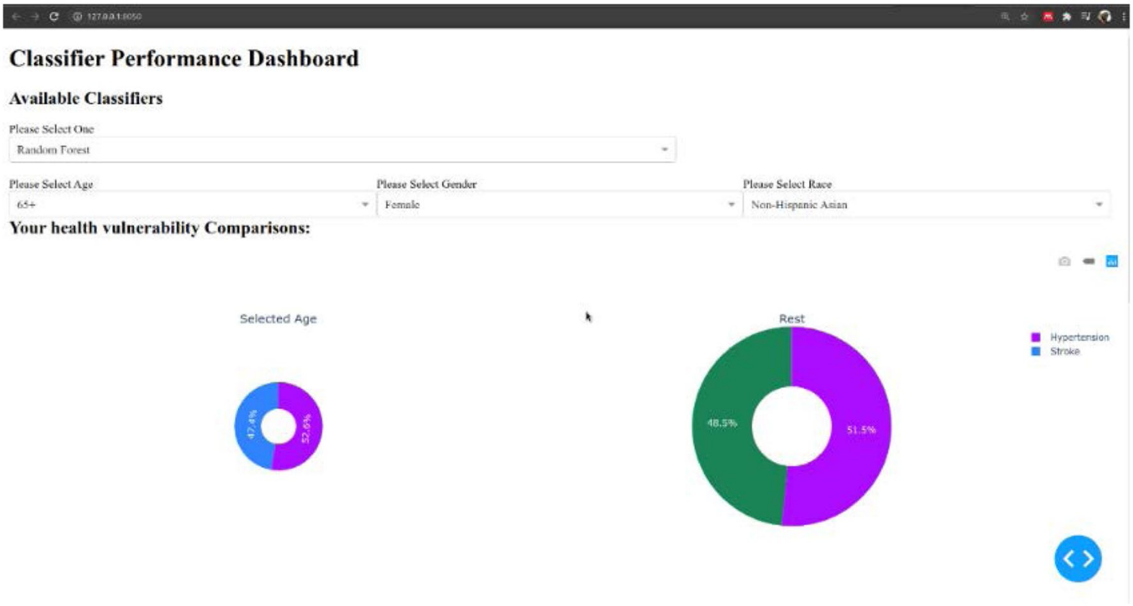
The PROCLAVE tool was developed to provide an overall user centric outcome, which is implemented on an online web interface to visualise the attributes with an applicable classifier. The entire functionality tool was based on the applicable classifier, with hosted datasets to extract and visualize the features. The hosted data was free from the missing values and inconsistencies among to design a well conceptualized model. The model was able to retrieve hidden effective patterns from healthcare databases. Figure 7 represents a classifier performance dashboard where the classifier was implemented with hosted datasets as a visual analytical interface. .

The proposed approach was hypothesized to develop a visual analytical interface which can detect the varied patterns with an applicable classifier. The upper corner of the PROCLAVE dashboard represents the availability of a classifier where the bottom tool visualizes the represented features which can be selected from the menu. The tool can retrieve the trends as required, and the number of cases were classified according to the classifier reported with race and age. The plot retrieved can effectively generate knowledge from the datasets. Figure 8. represents the applicability of the tool to visualize the results from variable features. .

Fig. 10 Color code representation of node



Color	Node	Description
🟥	Race->Non Hispanic White<1.250	The Non Hispanic white population between age of 18-24 years show no visible signs of hypertension
🟦	Race->Non Hispanic White>1.250	The Non Hispanic white population between age of 45-64 years show visible signs of hypertension and occurrence of stroke
🟦🟩	Race-> Hispanic-> PercentageAge>1.5	The Hispanic population age of +35 years show high value of occurrence of hypertension and occurrence of stroke
🟥	Race-> Hispanic-> PercentageAge<1.5	The Hispanic population age of 18-24 years represents no visible signs of occurrence of hypertension and stroke
🟦🟩	Race-> Non -Hispanic Asian	The No-Hispanic asian age of 45-66 years represents no visible signs of occurrence of hypertension and stroke
🟩	Race-> Non- Hispanic Black->Category-> CardioVascular Disease	The Non Hispanic Black population age of +35 years show high value of occurrence of hypertension and occurrence of stroke
🟦🟥	Race-> Non- Hispanic Black->Category-> Risk Factors	The No-Hispanic Black age of 18-24 years represents the risk factor of acquiring a stroke is higher due to hypertension
🟨	Gender->Male->Percentage Age>8.25	The male population of all races with age group above 75+ represents very high vital signs of cardiovascular symptoms and occurrence of stroke is at higher side
🟩	Gender->Male->Percentage Age>3.05	Non Hispanic Asian above the age of 65+ represent the higher vital symptoms of occurrence of stroke due to high hypertension
🟩🟨	Gender->Male->Percentage Race>92.45	The Non Hispanic black population age of 65+ years show moderate signs of hypertension and occurrence of stroke
🟦🟩	Gender->Male->Percentage Race<92.45	The Non Hispanic white population between age of 25-44 years show visible signs of hypertension and occurrence of stroke



Classifier Performance Dashboard

Available Classifiers

Please Select One: Random Forest

Please Select Age: 65+

Please Select Gender: Female

Please Select Race: Non-Hispanic Asian

Your health vulnerability Comparisons:

Fig. 11 Further information on random forest

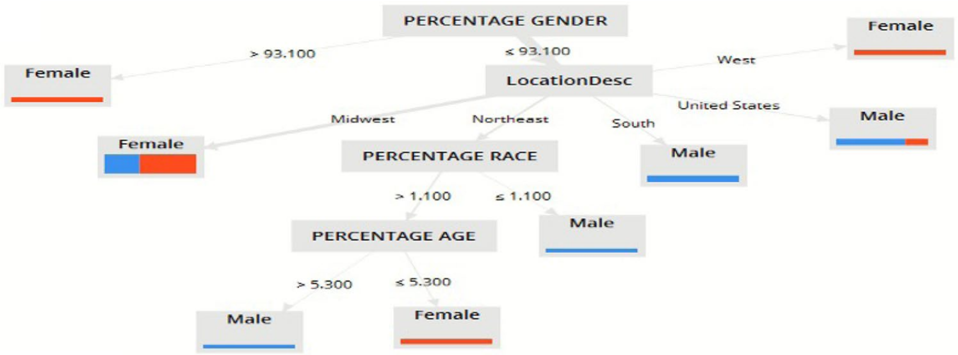
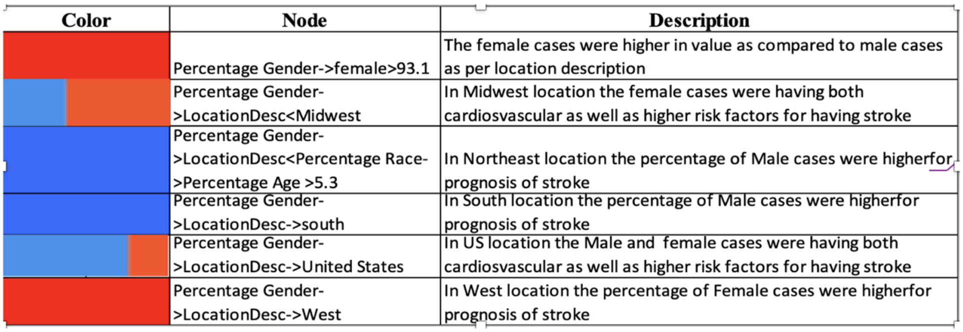


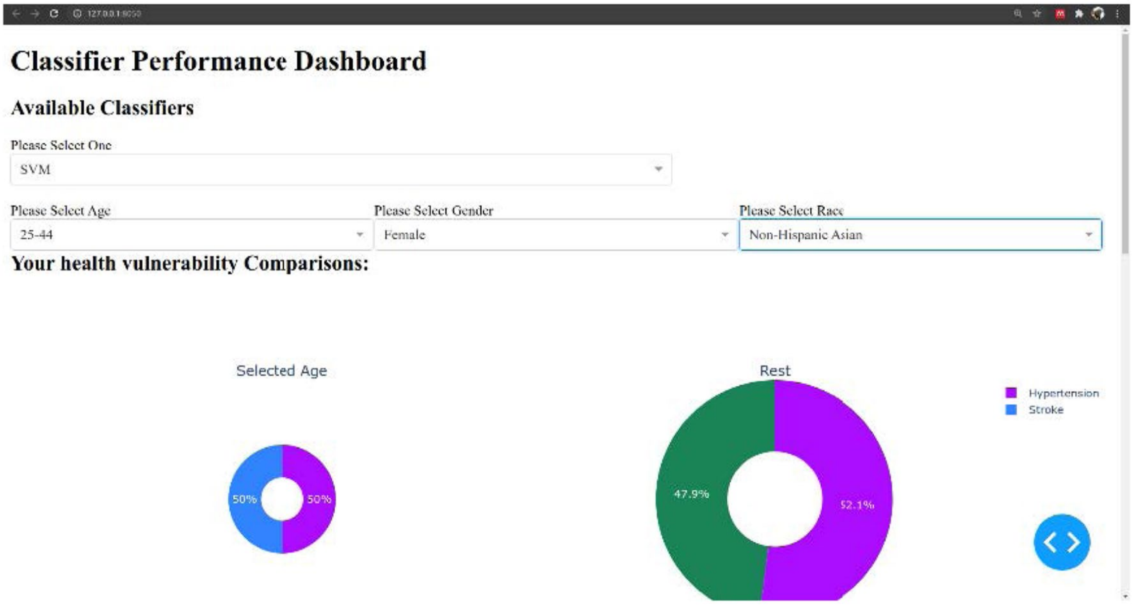
Fig. 12 Random forest tree

Fig. 13 Color code representation of tree



Color	Node	Description
🟥	Percentage Gender->female>93.1	The female cases were higher in value as compared to male cases as per location description
🟦🟥	Percentage Gender->LocationDesc<Midwest	In Midwest location the female cases were having both cardiosvascular as well as higher risk factors for having stroke
🟦	Percentage Gender->LocationDesc<Percentage Race->Percentage Age >5.3	In Northeast location the percentage of Male cases were higherfor prognosis of stroke
🟦	Percentage Gender->LocationDesc->south	In South location the percentage of Male cases were higherfor prognosis of stroke
🟦🟥	Percentage Gender->LocationDesc->United States	In US location the Male and female cases were having both cardiosvascular as well as higher risk factors for having stroke
🟥	Percentage Gender->LocationDesc->West	In West location the percentage of Female cases were higherfor prognosis of stroke

The results were analyzed for attributes related to stroke, corresponding to age category and race, while the remaining basis of gender was to study the incidence of stroke. The age of the patients was categorized in retrospect to 18–24, 25–44, 45–66, 35+,65+ and 75+. The race or ethnicity was included for Hispanic, Non-Hispanic Asian, Non-Hispanic Black, Non-Hispanic White. The decision tree signified that the higher percentage of stroke incidence occurred in males as compared to that of fe-males. The incidence value for males was 4.750 and females was that of 0.100, which also denoted the equal chances of getting the disease with reference to the race. Figure 9 represents the decision tree which signified



Classifier Performance Dashboard

Available Classifiers

Please Select One: SVM

Please Select Age: 25-44

Please Select Gender: Female

Please Select Race: Non-Hispanic Asian

Your health vulnerability Comparisons:

Fig. 14 PROCLAVE dashboard for SVM

Fig. 15 SVM tree

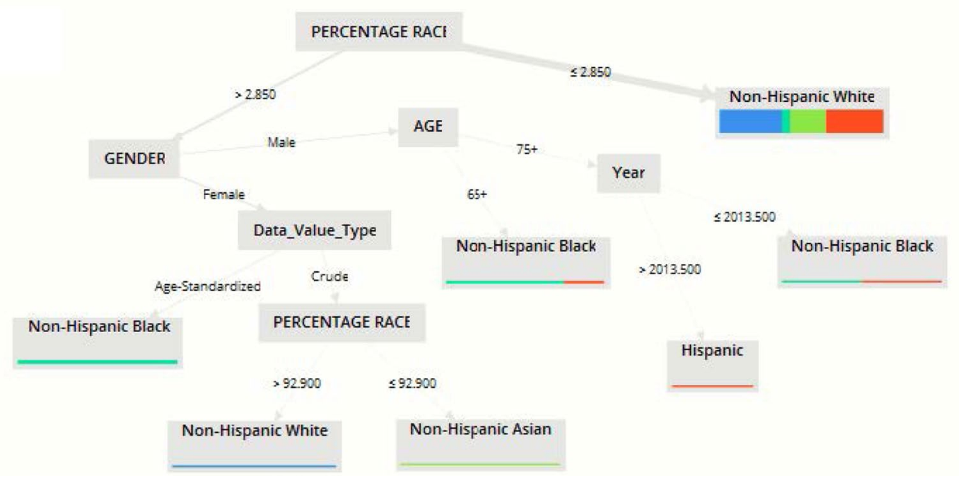


Fig. 16 Color node representation

Color	Node	Description
🟩	Percentage Race->Gender->Non Hispanic Black	The Non Hispanic Black female of age 35+has the higher prognosis of having stroke as the risk factors are higher
🟦	Percentage Race->Gender->Non Hispanic white	The Non Hispanic white population of female are giving vissible increase in Hypertension level at the age group of 35+
🟩	Percentage Race->Gender->Non Hispanic Asian	The Non Hispanic Asian population of female are having very less percentage of prognosis of cardio vascular disease
🟩🟧	Percentage Race->Gender-> Age ->Non Hispanic Black	The Non Hispanic Black male of age 65+has the higher prognosis of having stroke as well as they have hyterstension on higher level
🟧	Percentage Race->Gender-> Age ->Year ->Hispanic	In year 2014-2015 the Hispanic male population of age 75+ represent higher prognosis of cardiovascular disease
🟩🟧	Percentage Race->Gender-> Age ->Year -> Non-Hispanic Black	The Non Hispanic Black population of age 75+ represent higher prognosis in year <2013 of cardiovascular disease
🟦🟩🟥	Percentage Race->Non Hispanic white	The Non-Hispanic white population has lower number of cases for prognosis of Stroke as compared to other population

Table 2 The comparison of accuracy results

Accuracy measure			
Classifier	F1	Precision	Recall
Decision iree	0.96	0.93	1
Random forest	0.87	0.821	0.68
SVM	0.83	0.69	0.37

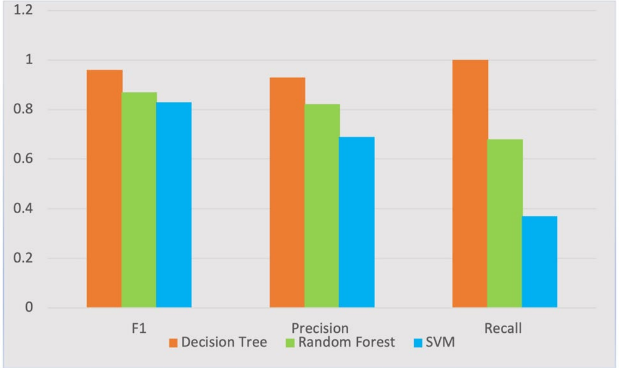


Fig. 17 Graphical representation of three models

that highest rate of incidence of stroke was found in the age group percentage of + 75 years for males and 65 years for race percentage in males. The least occurrence of stroke was in females in the age group of 18–24 for both race and age percentages. For the age group of 35+ and 45 years in both males and females, the rate depended on the race category. Further, the colors in the nodes are represented as relative to the selected root and the node of the edges correspond to the proportion of instances in the corresponding root with respect to all the instances in the training data. Figure 10 represents the description of each node with color associated with the target.

The designed and implemented model was based on an interactive interface where each information was displayed and can be utilized as an interactive format. The PROCLAVE represents the trends and distribution of each selected feature, with the applicable classifier, where it discovers results to obtain more information. Figure 11 represents the interactive tool of the selected random forest and Figs. 12, 13 represents the formulated tree.

In Fig. 12, a representative tree was designed utilizing the random forest, and it was observed that according to the region-wise prevalence of stroke, both males and females showed a variation where western females were most affected, while in the southern region, males were most affected, in the northeast region males were more impacted compare to females. These observations explained the fact that males were found to be highly suffering from stroke. Figure 13 depicts each color node description.

Furthermore, the study was gathered for SVM to relatively determine relevant patterns for compara-tive study. Figure 14 represents a PROCLAVE added feature for SVM which can be chosen from the drop-down menu.

In Figs. 15, 16, SVM can be effectively utilized to determine the variation in the rate of occurrence of stroke among the different races in the United States. .

The next section of the study represents a comparative study where the predicted results suggested that the decision tree had a better accuracy model as compared to the Random forest and that SVM as the accuracy model was achieved with F measure, Precision and Recall. Table 2. represents the accuracy measure among the classifiers.

Figure 17. Graphical Representation of the three models was done by observing the graphical analysis. Decision Tree yielded better accuracy rates compared to Random Forest and SVM for all the selected attributes.

5 Conclusion and discussion

AI and ML have widely grown in capacity from the past decade and had consistently predicted patterns for future decision making. In the current scenario, algorithmic tools are playing a vital role in diagnosing disease by utilizing several techniques, which include ML, statistics, AI, data-base sets, pattern recognition and visualization for future prediction models. Moreover, IT has widely changed the era of medical databases and generally benefited the diagnosis of disease.

Conversely, ML and AI tends to be key fundamental technologies that play a major role in the healthcare domain. Several new learning systems were established to deal and align with forthcom-ing healthcare technologies. In fact, healthcare systems have adopted a new-fangled technology from the past decade, which include imaging, video, EHR (Electronic Health Records) and others. This has profusely changed the learning era and has forced new AI and ML based technologies to design tools which can predict results for future diagnosis.

In the current study, we designed and deployed a "PROCLAVE" tool in python language. The tool was designed in varied layered structures, where each layer plays a significant role in determining the patterns. The tool was deployed in windows 10 operating system for developing and implementing the entire script. Several packages were configured to develop a visual and analytical inter-face for significant configuration rich interface. Several libraries were applied for processing the pro-totype to develop a visualization interface. The tool was applicable to forecast the health vulnerabil-ity among variable classifiers and to visualize the results for end users. Moreover, the proposed ar-chitecture was based on conceptualization and visualization concepts to detect the overall dash-board.

Moreover, PROCLAVE was applied to determine the significant patterns which can be easily and readily interpreted by the end users. Each layer works on a corresponding interface where ML plays as interactive module in the second layer to proportionally relate data with applicable classifiers. Additionally, varied distribution patterns were determined to correlate the study with applicable fea-tures in the databases.

Furthermore, the current approach was synthesized and populated with databases that allowed the end users to select the variable features and relatively determine the interactive patterns for a number of cases. The database was collected from the National Institute of Health Stroke (NIHS) in the United States on stroke patients diagnosed from 1950 to2015. The study was based on causes of death, sex, race, and Hispanic origin and other attributes to discover unknown patterns for future decision making. The observed data was expressed as continuous variables where each attribute tends to relate each other. Furthermore, data was classified using appropriate classifiers which in-cluded the decision tree and SVM. The applicability of each classifier was deter-mined using Python 3.3 to measure accuracy and regulate the classifier which can assess the data-base for future knowledge discovery.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

American Stroke Association (ASA) (2015) What is a stroke. https://www.stroke.org/en/about-stroke

Asha T, Natarajan S, Murthy K (2012) Data mining techniques in the diagnosis of tuberculosis. Underst Tuberc-Glob Exp Innov Approaches Diagn 16:333–353

Barbat MM, Wesche C, Werhli AV, Mata MM (2019) An adaptive machine learning approach to improve automatic iceberg detection from sar images. ISPRS J Photogramm Remote Sens 156:247–259

Beck BR, Shin B, Choi Y, Park S, Kang K (2020) Predicting commercially available antiviral drugs that may act on the novel coronavirus (sars-cov-2) through a drug-target interaction deep learning model. Comput Struct Biotechnol J 18:784–790

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Breslow LA, Aha DW (1997) Simplifying decision trees: a survey. Knowl Eng Rev 12(01):1–40

Centers for Disease Control and Prevention (CDC) (2015) Stroke facts. https://www.cdc.gov/stroke/facts.htm

Chang F, Guo C-Y, Lin X-R, Lu C-J (2010) Tree decomposition for large-scale svm problems. J Mach Learn Res 11:2935–2972

Chen X-W, Lin X (2014) Big data deep learning: challenges and perspectives. IEEE Access 2:514–525

Choi S, Lee J, Kang M-G, Min H, Chang Y-S, Yoon S (2017) Large-scale machine learning of media outlets for understanding public reactions to nation-wide viral infection outbreaks. Methods 129:50–59

El Saghir NS, Assi HA, Jaber SM, Khoury KE, Nachef Z, Mikdashi HF, El-Asmar NS, Eid TA (2014) Outcome of breast cancer patients treated outside of clinical trials. J Cancer 5(6):491

Enterprise (2020) Dash user guide. https://dash.plotly.com/

Esposito F, Malerba D, Semeraro G, Kay J (1997) A comparative analysis of methods for pruning decision trees. IEEE Trans Pattern Anal Mach Intell 19(5):476–491

Franco-Arcega A, Carrasco-Ochoa JA, Sánchez-Díaz G, Martínez-Trinidad JF (2011) Decision tree induction using a fast splitting attribute selection for large datasets. Expert Syst Appl 38(11):14290–14300

Gárate-Escamila AK, El Hassani AH, Andrès E (2020) Classification models for heart disease prediction using feature selection and PCA. Inf Med Unlocked 19:100330

Ioannis K, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I (2017) Machine learning and data mining methods in diabetes research. Comput Struct Biotechnol J 15:104–116

Joloudari JH, Saadatfar H, Dehzangi A, Shamshirband S (2019) Computer-aided decision-making for predicting liver disease using pso-based optimized svm with feature selection. Inf Med Unlocked 17:100255

Kaur D, Bedi R, Gupta SK (2015) Review of decision tree data mining algorithms: ID3 and C4. 5. In: Proceedings of international conference on Information Technology and Computer Science, pp 11–12

Lavanya D, Rani KU (2011) Performance evaluation of decision tree classifiers on medical datasets. Int J Comput Appl 26:1–4

Li HB, Wang W, Ding HW, Dong J (2010) Trees weighting random forest method for classifying high-dimensional noisy data. In: 2010 IEEE 7th International Conference on E-Business Engineering, IEEE, pp 160–163

Liu T, Fan W, Wu C (2019) A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artif Intell Med 101:101723

Lysaght T, Lim HY, Xafis V, Ngiam KY (2019) Ai-assisted decision-making in healthcare. Asian Bioeth Rev 11(3):299–314

McKinney W et al (2010) Data structures for statistical computing in python. Proc Python Sci Conf 445:51–56

Menad NA, Hemmati-Sarapardeh A, Varamesh A, Shamshirband S (2019) Predicting solubility of CO2 in brine by advanced machine learning systems: application to carbon capture and sequestration. J CO2 Util 33:83–95

Moloud A, Yen NY, Hung JC-S (2018) Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees. J Med Biol Eng 38:953–965

Mosavi A, Salimi M, Faizollahzadeh Ardabili S, Rabczuk T, Shamshirband S, Varkonyi-Koczy AR (2019) State of the art of machine learning models in energy systems, a systematic review. Energies 12(7):1301

Nápoles G, Grau I, Bello R, Grau R (2014) Two-steps learning of fuzzy cognitive maps for prediction and knowledge discovery on the hiv-1 drug resistance. Expert Syst Appl 41(3):821–830

Otoom AF, Abdallah EE, Kilani Y, Kefaye A, Ashour M (2015) Effective diagnosis and monitoring of heart disease. Int J Softw Eng Appl 9(1):143–156

Patil DD, Wadhai V, Gokhale J (2010) Evaluation of decision tree pruning algorithms for complexity and classification accuracy. Int J Comput Appl 11(2):23–30

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Peiffer-Smadja N, Rawson TM, Ahmad R, Buchard A, Georgiou P, Lescure F-X, Birgand G, Holmes AH (2020) Machine learning for clinical decision support in infectious diseases: a narrative review of current applications. Clin Microbiol Infect 26(5):584–595

Pouriyeh S, Vahid S, Sannino G, De Pietro G, Arabnia H, Gutierrez J (2017) A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease. In: 2017 IEEE Symposium on Computers and Communications (ISCC), IEEE, pp 204–207

Prajwala T (2015) A comparative study on decision tree and random forest using r tool. Int J Adv Res Comput Commun Eng 4(1):196–199

Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106

Robnik-Šikonja M (2004) Improving random forests. European conference on machine learning. Springer, Berlin, pp 359–370

Rong G, Mendez A, Assi EB, Zhao B, Sawan M (2020) Artificial intelligence in healthcare: review and prediction case studies. Engineering 6(3):291–301

Sammut C, Webb GI (2011) Encyclopedia of machine learning. Springer Science and Business Media, Berlin

Sharma P, Choudhary K, Gupta K, Chawla R, Gupta D, Sharma A (2020) Artificial plant optimization algorithm to detect heart rate and presence of heart disease using machine learning. Artif Intell Med 102:101752

Shi Y, Liu H, Wang Y, Cai M, Xu W (2018) Theory and application of audio-based assessment of cough. J Sens. https://doi.org/10.1155/2018/9845321

Tanwar G, Chauhan R, Yafi E (2021) Artycul: a privacy-preserving ml-driven framework to determine the popularity of a cultural exhibit on display. Sensors 21(4):1527

Thomas M (2020) Researchers want your voice to train coronavirus-detecting ai'. https://thenextweb.com/news/researchers-want-your-voice-to-train-coronavirus-detecting-ai

Vaka AR, Soni B, Reddy S (2020) Breast cancer detection by leveraging machine learning. ICT Express 6(4):320–324

Xu S, Zhang Z, Wang D, Hu J, Duan X, Zhu T (2017) Cardiovascular risk prediction method based on cfs subset evaluation and random forest classification framework. In: 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), IEEE, pp 228–232

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.