TABLE OF CONTENTS

ABSTRACT

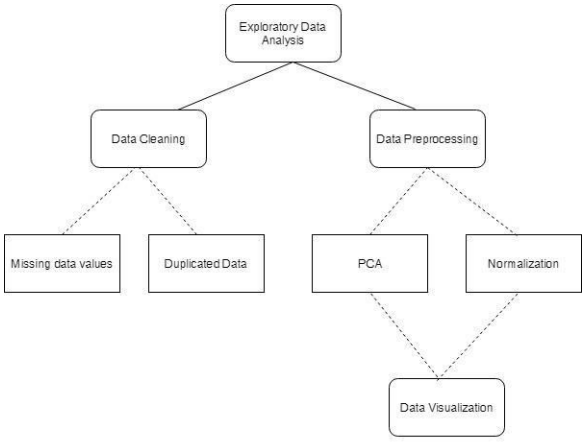
Big data presents large quantities of data in various formats and in order to extract maximum useful information from the dataset, a number of data integration techniques like data fusion can be used. Data is normalized by setting the mean to 0 and variance to 1. By doing so, we attain a linear distribution of components from the dataset when we perform PCA (Principal Component Analysis). PCA is performed in order to reduce the dimensionality of the dataset and create PCA components which would orthogonally portray a relationship between all the other components of the dataset so that when there is a new patient entry in the dataset, it can easily categorize the outcome of the new entry based on the PCA component. In order to have better accuracy of our analysis, a model is trained to predict the outcome of our analysis. After comparing the accuracies of different algorithms, a model consisting of common components with similar characteristics between the datasets is trained and existing machine learning algorithms are applied to predict if a patient having chronic kidney disease symptoms suffers from heart disease as well by categorizing the outcome into four different stages. After applying multiple algorithms, the algorithm with the most reduced error-rate, and the best accuracy to perform the data analysis to associate chronic kidney disease and heart disease are selected.

INTRODUCTION

Before beginning with the data analysis process, it is really important to clean the data that we will be using for our process. There are numerous ways to perform to clean the data and transform it with the intent of making it more appropriate to work with and make more accurate.

Data Wrangling is the process of mapping data in one format to another format in order to make it more valuable for a variety of downstream purposes such as analytics. This may include data visualization, data aggregation, training a model, and many such uses. These data transformations are usually applied to different entities of the dataset to create desired wrangling output which could be used for further analysis of our dataset. In order to clean the data, various techniques such as filling the missing values, handling incorrect data types and duplicated data etc can be used. After cleaning the dataset, data analysis by applying different functions using various python libraries and the result of our analysis is displayed using different data visualizations methods.

A majority of datasets consist of a lot of components and rescaling the data becomes a priority. One single entry in the dataset will have projections from all the components in the dataset and predicting if a patient has the disease or not becomes difficult. In order to improve the outcome of our analysis, we normalize the data before beginning with data analysis. StandardScaler preprocessing module is used to normalized the features, which sets the 'mean=0' and 'variance=1' for all the features. If you normalize the features, it would provide a fair comparison between the explained variance in the dataset. Principal Component Analysis is a dimensionality reduction technique that uses orthogonal transformation to convert a set of observations of possibly correlated variables into principal components. The overall performance of the PCA is dominated by high variance features. Therefore, features should be normalized.



ARCHITECTURE DIAGRAM

The PCA components can only be numeric. So the common way is to encode the categorical values by using one-hot encoding or any other encoding method to transform the values into 0 and 1. PCA looks for the correlation between the features and reduces the dimensionality. The same process is repeated for both the datasets and the outcome of our analysis is portrayed using visualizations.

PCA is an unsupervised machine learning method and hence we use this method only for exploratory data analysis purpose in order to gather insights about the dataset. Next, a supervised machine learning algorithm is used to train a model for predictive analysis. As usual, the dataset is processed first and a deep neural network is made where the components with common characteristics between the datasets are given as input. The Deep Neural Network (DNN) is trained, tested and the result is displayed with accuracy as the metric. Compare the result from the DNN to another ML algorithm.

Literature Survey

2.1 A reduced complexity data fusion algorithm using belief propagation for location tracking in heterogeneous observations.

Author	Technique used	Advantages	Disadvantages	Application	Purpose
Yih-Shyh Chiou and Fuan Tsai	Bayesian Filtering	Less computational complexity And High Accuracy	Incoming information may not always be highly reliable	Location Tracking system	To reduce the computational load of the traditional data fusion algorithm with heterogeneous observations for Location tracking

2.2 A data fusion approach for track monitoring from multiple in-service trains

Author	Technique used	Advantages	Disadvantages	Application	Purpose
A data fusion approach for track monitoring from multiple in-service trains	Kalman Filter	Handles Asynchronous data, Increases overall reliability of inspection.	High detection cost, works with high cost sensors	Rail track monitoring system	Enabling data-driven rail-infrastructure monitoring from multiple in-service trains.

2.3 A fault tolerant architecture for data fusion: A real application of kalman filters for mobile robot localisation

Author	Technique used	Advantages	Disadvantages	Application	Purpose
Kaci Bader, Benjamin Lussier, Walter Schon	Kalman filter data fusion	Focusing both hardware and software faults, No false positives	// Time consuming, Accuracy is less(62.32% hardwar e faults, 90% Software faults)	Vehicle Localization	Identification and recovery of hardware and software faults (fault tolerant architecture)

2.4 Double sample data fusion method based on combination rules

Author	Technique used	Advantages	Disadvantages	Application	Purpose
Jianbin Xiong	Double sample data fusion method	9.45% of accuracy improvement when compare to KNN combination method	Reliability calculation method is inadequate	Petrochemical rotating machinery equipment	Fault dignosis

2.5 Object based analysis and fusion of optical and SAR satellite data for dwelling detection in refugee camps

Author	Technique used	Advantages	Disadvantages	Application	Purpose
Kristin Sprohnle, Eva-Maria Fuchs and Patrick Aravena Pelizari	Feature based data fusion	High accuracy (nofalse positives), Less time, Works with complex structure	Automatic detection need to be improved	Refugee Camps	Dwelling detection

2.6 A data fusion based hybrid sensory system for older people's daily activity and daily routine recognition.

Author	Technique used	Advantages	Disadvantages	Application	Purpose
Yan Wang,Shuang Cang and Hongnian Yu	Feature level fusion and SVM method	Increased accuracy of 98.32%	Works for only single structure(not for multiple)	Health Care	Human activity recognition

2.7 Inference from literature survey

The above working papers of literature survey show that the reinforcement learning methods which are used, although functional, are not completely effective. Each study has a drawback or a high cost function which cannot be discarded without affecting picture quality. Each study, however, does overcome a significant drawback found in previous studies but does not eliminate the problem altogether.

Post survey the important techniques that were being studied were:

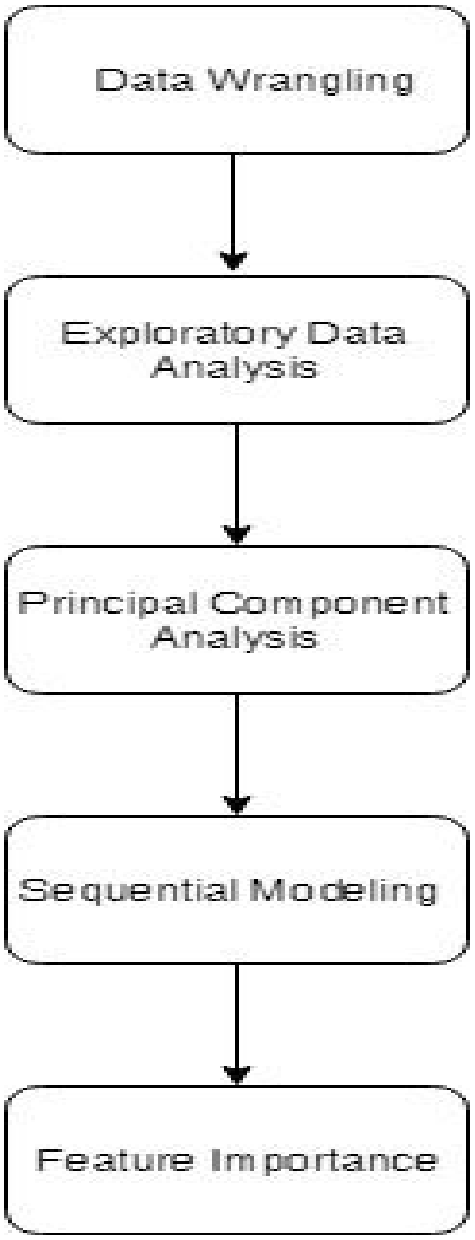
1. Feature based Data fusion: This approach enables us to take advantages of the strengths and limitations of various modalities in a unified analytic framework and demonstrates that data fusion techniques but the features chosen to perform the analysis are subject to performance issues and change in its properties based on various changes that occur at every step of our analysis. This thus proves to be inefficient for best results.
2. Bayesian Data Analysis: It improves the knowledge of the system by adding observations and easily combining information from diverse sources but the problem with this technique is that it requires detailed knowledge of physical system and the complete measurement process and hence cannot be used.
3. Kalman Filter: It uses the previously used data to predict the future data and its memory is low i.e. it can store nothing else but the previous value. Kalman filter is not using the previous estimate but prediction. Error covariance is calculated and the value is sent to the prediction process to determine the efficiency of the filter. Difficulty of Kalman Filter to achieve a more optimal result is only because it is a complex and considerate algorithm. If the sample rate could be increased, the Kalman filter would get a better output.

So for our project we are going to use Kalman filter. The accuracy is to be increased using k-means algorithm.

MODULES DESCRIPTION

The main modules used here are :

1. Data Wrangling
2. Exploratory Data Analysis
3. Principal Component Analysis
4. Sequential Modeling
5. Feature Importance



1. Data Wrangling

- Handling missing values
- Correcting data formats
- Incorrect Data types
- Duplicated data
- We convert all the '?' in the dataset with 'NaN' and then use the 'dropna()' function to drop all the entries with 'NaN'

2. Exploratory Data Analysis

- Make separate lists for categorical and continuous values.
- Convert all categorical values into numerical values.
- Reconstruct data frame for effective modeling
- Perform Recursive Feature Elimination
- Split dataframe into train and test set

3. Principal Component Analysis

- We set mean=0 and variance=1 to attain a linear distribution
- All the features in the dataset project itself to a two or three-dimensional plane
- It is a dimensionality reduction technique
- It is an unsupervised machine learning algorithm
- All of the components of our dataset project itself to derive a new

principal component which can be the defining component to represent all of the remaining components

- We plan to use t-Distributed Stochastic Neighbour Embedding which gives us better results for dimensionality reduction
- We drop the classification and target column and then perform PCA. The remaining columns are then compared with the target column and we plot the distribution.

- After performing PCA, each new entry of the patient as represented by a 'point' in the 2D plane can be clearly classified after normalizing the data.
- The principal components are always orthogonal to each other.

4. Sequential Modeling

- We use the sequential model for our analysis as we need to fetch the output of the previous node to the next node.
- We add layers to our model.
- We apply activation function to determine the output of the neural network. It maps the resulting values in between 0 to 1 or -1 to 1.
- We then compile the model to calculate loss by using binary_crossentropy() function.
- We then plot our results and display the accuracy of our analysis.

5. Feature Importance

- Feature importance is the increase in model error when the feature's information is destroyed.
- Feature importance provides a highly compressed, global insight into the model's behavior.
- We measure the importance of a feature by calculating the increase in the model's prediction error after permuting the feature.
- A feature is "important" if shuffling its values increases the model error, because in this case the model relied on the feature for the prediction.
- A feature is "unimportant" if shuffling its values leaves the model error unchanged, because in this case the model ignored the feature for the

prediction.

- A correlation matrix is a table showing correlation coefficients between sets of variables. This allows us to see which pairs have the highest correlation.
- From the correlation matrix, we can observe which input features contribute most during the training process of various classifiers.

ARCHITECTURE DIAGRAM

ALGORITHMS USED

1. **Backpropagation** : Used to train the data by using MinMaxScaler. It tries to calculate the gradient value that is needed for calculation of weights in the network. Backward propagation of errors occur and this data is fed to the model.

- import MinMaxScaler
- from keras.models import Sequential
- from keras.layers import Dense
- model = Sequential()
- model.add(Dense(*no_of_nodes, activation function)*
- model.compile(loss='binary_crossentropy', optimizer, metrics=['accuracy'])
- model.fit
- plt.plot()

2. **Logistic Regression**

- Logistic Regression is a predictive analysis used to describe data and to explain the relationship between one dependent binary variable and one more independent variable.

- lr = LogisticRegression()
- lr.fit(*selected_x_train, selected_y_train)*
- printf(lr.score)

3. **Random Forest Classifier**

- Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

- from sklearn.ensemble import RandomForestClassifier
- rf = RandomForestClassifier(n_estimators = 1000, random_state = 1)

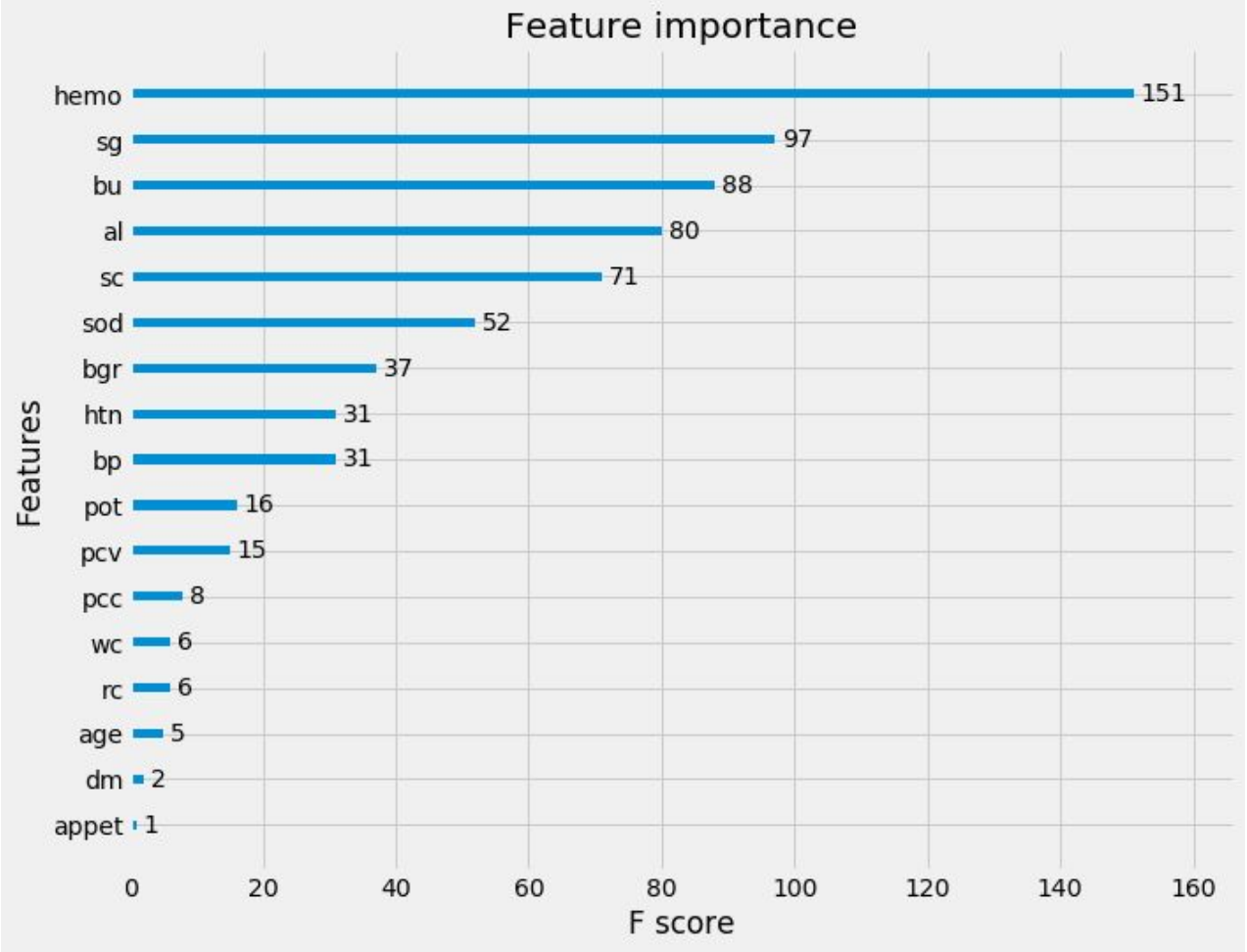
- rf.fit(x_train.T, y_train.T)
- print("Random Forest Algorithm Accuracy Score : {:.2f}%".format(rf.score(x_test.T,y_test.T)*100))

4. **Feature Importance**

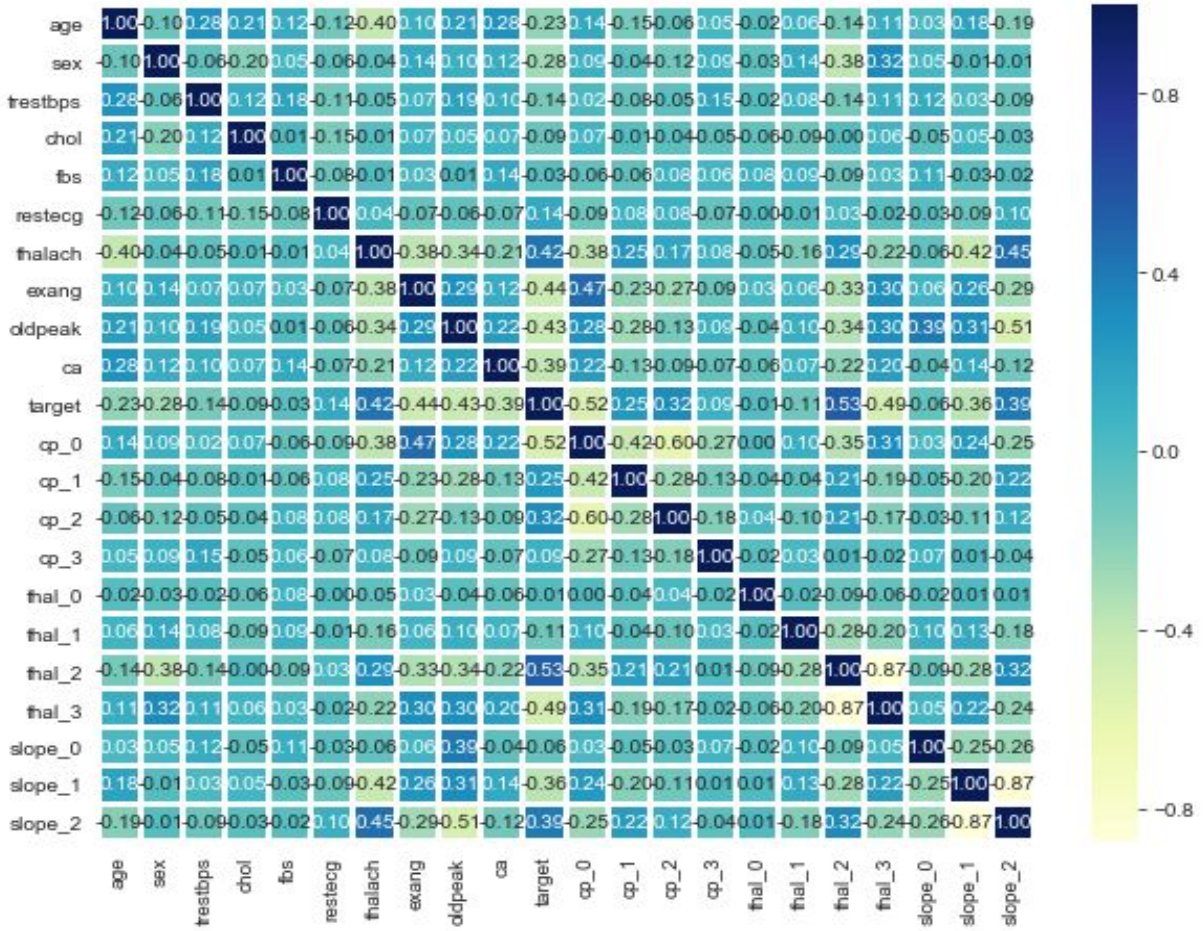
- We measure the importance of a feature by calculating the increase in the model's prediction error after permuting the feature.
- Extreme Gradient Boost Classifier is used for this algorithm.

- xgb_cl = xgb.XGBClassifier()
- X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
- random_state=21, stratify=y)
- xgb_cl.fit(X_train, y_train)
- figsize(10,8)
- plt.style.use('fivethirtyeight')
- xgb.plot_importance(xgb_cl)

APPENDIX



CHRONIC KIDNEY DISEASE



O/P FOR HEART DISEASE

REFERENCES

1. Florian Endel, Harald Piringer : Data Wrangling: making data useful again.

2. Dr. Rama Kishore, Taranjit Kaur : Backpropagation Algorithm: An Artificial Neural Network Approach for Pattern Recognition .

3. Cristinel Constantin : Principal Component Analysis - A powerful tool in computing marketing information.

4. Y. Chiou and F. Tsai, "A Reduced-Complexity Data-Fusion Algorithm Using Belief Propagation for Location Tracking in Heterogeneous Observations," in IEEE Transactions on Cybernetics, vol. 44, no. 6, pp. 922-935, June 2014.

5. George Lederman, Siheng Chen, James H. Garrett, Jelena Kovacevic, Hae Young Noh, Jacobo Bielak, "A data fusion approach for track monitoring from multiple in-service trains,"Mechanical Systems and Signal Processing,Volume 95, 2017.

6. B. Kaci, L. Benjamin and S. Walter, "A Fault Tolerant Architecture for Data Fusion Targeting Hardware and Software Faults," 2014 IEEE 20th Pacific Rim International Symposium on Dependable Computing, Singapore, 2014

7. J. Xiong, Q. Zhang, Z. Peng, G. Sun and Y. Cai, "Double Sample Data Fusion Method Based on Combination Rules," in IEEE Access, vol. 4, pp. 7487-7499, 2016.

8. K. Spröhnle, E. Fuchs and P. Aravena Pelizari, "Object-Based Analysis and Fusion of Optical and SAR Satellite Data for Dwelling Detection in Refugee Camps," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 10, no. 5, pp. 1780-1791, May 2017.

9. Y. Wang, S. Cang and H. Yu, "A Data Fusion-Based Hybrid Sensory System for Older People's Daily Activity and Daily Routine Recognition," in IEEE Sensors Journal, vol. 18, no. 16, pp. 6874-6888, 15 Aug.15, 2018.