Prediction of heart disease using calibration of hyper parameter optimization

Guided By:
Dr.N.Balaganesh

Anto Nidhish M (Reg. No. : 201904019)
Hariharan C (Reg. No. : 201904048)

Problem statement

Coronary artery disease prediction is considered to be one of the most challenging tasks in the health care industries. Hyperparameters are set by the machine learning engineer before training and they play a major role in performance of the model. Selecting the best hyperparameter boosts the performance of the Machine Learning model.

Objectives

- To predict heart disease efficiently
- To find the best features for classification
- To optimize hyper parameters to achieve better results

Outcomes

- Feature selection techniques select the best features for classification and drops irrelevant features.
- Various Hyper Parameter Optimization Techniques are used to find best combination of hyper parameters.

Literature Survey - 1:

Name : Implementation of Machine Learning Model to Predict Heart Failure Disease
Author Names : Fahd Saleh Alotaibi
Published In: (IJACSA) International Journal of Advanced Computer Science and Applications
Year : 2019

Firstly, to increase the size of the dataset, Random Number has been generated for each column. The missing values has been imputed using k-nearest neighbor method. Outlier detection is done using rapid miner's operator using distance method. The pre-processed dataset has been trained and tested using five ML models: Decision Tree, Naïve Bayes, Random Forest, Logistic Regression and SVM.

PROS

KNN imputation preserves the relationships between variables, which can be important for some types of analysis. RapidMiner enables rapid development of data cleaning pipelines, which can save time and effort.

CONS

KNN imputation may not work well for highly correlated variables, as it can introduce noise into the imputed values. RapidMiner may not handle complex data cleaning scenarios that require custom code or more advanced data cleaning techniques.

Literature Survey - 2:

Name : Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques
Author Names : Senthilkumar Mohan, Chandrasegar Thirumalai, Gautam Srivastava
Published In: Elseiver
Year : 2021

This paper propose a novel method that aims at finding significant features by applying machine learning techniques resulting in improving the accuracy in the prediction of cardiovascular disease. The prediction model is introduced with different combinations of features and several known classification techniques.

PROS

Data preprocessing is done to remove data with missing values.
Several classification algorithms are compared which provides insight on best working model.

CONS

The HRFLM algorithm may not perform well if the random forest model is overfitting the data, as this will result in poor predictions for the linear regression model.
Both random forest and linear regression models have inability to handle missing data or outliers.

Literature Survey - 3:

Name : Coronary artery disease detection using computational intelligence methods
Author Names : Roohallah Alizadehsani, Mohammad Hossein Zangooei, Mohammad Javad Hosseini, Jafar Habibi, Abbas Khosravi, Mohamad Roshanzamir, Fahime Khozeimehe , Nizal Sarrafzadeganf , Saeid Nahavandi
Published In: Elsevier
Year : 2016

Firstly, the feature selection was done using “Weights by SVM” using 10-fold cross validation. Then the classification was done using Support Vector Machine (SVM) with a set of kernels. Apriori algorithm is used to obtain rules for Association rule mining.

PROS

Weighting can help SVMs to achieve better accuracy, sensitivity, specificity, and F1-score, especially in imbalanced datasets, by giving more importance to the minority class.
The approach can be used with a variety of kernel functions, including linear, polynomial, Gaussian, and others, which can capture different types of non-linearity and complexity in the data.

CONS

Weighting may not always improve the performance of SVMs, especially when the data is not highly imbalanced or when the weighting scheme is not properly calibrated.
The approach may require more computational resources and time than using a single kernel function, especially when dealing with a large dataset or a complex combination of kernels.

Literature Survey - 4:

Name : Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques
Author Names : Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava
Published In: IEEE Access
Year : 2019

Firstly, the records with missing values are removed from the dataset. ring of datasets is done on the basis of the variables and criteria of Decision Tree (DT) features. Classifiers such as Decision Tree, Language Model, SVM, Random Forest, Naïve Bayes, Neural Network and K-Nearest Neighbor are used and the best model is identified.

PROS

RBFN can approximate complex and nonlinear functions efficiently, making it an effective tool for solving problems that cannot be solved using linear models and it is quicker.

CADSS can automate many decision-making processes, reducing the time and effort required to make complex decisions

CONS

RBFN can easily overfit the training data if the number of hidden nodes is not optimized.

CADSS depends on the quality and accuracy of input data, which can be affected by errors, bias, and incomplete data.

Literature Survey - 5:

Name : Effective diagnosis of heart disease through neural networks ensembles
Author Names : Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur
Published In: Elsevier
Year : 2009

The proposed methodology is implemented with the SAS base software 9.1.3. Data partition component was used to partition the input data into train and validation data sets. Variable selection component was used in reducing the number of inputs by setting the status of the input variables that are not related to the target as rejected. Neural networks block component was used to classify the feature space. Three independent neural networks models were used to construct this component. The Levenberg–Marquardt (LM), scaled conjugate gradient (SCG) and Pola–Ribiere conjugate gradient (CGP) algorithms are used. A tangent Sigmoid transfer function was used for both hidden and output layers.

Literature Survey - 6:

Name : Classifier identification using deep learning and machine learning algorithms for the detection of valvular heart diseases
Author Names : Tanmay Sinha Roy,Joyanta Kumar Roy, Nirupama Mandal
Published In: Elseiver
Year : 2019

This study aims to find the best classifiers for different valvular heart problems using popular CNN-based deep learning models and machine learning algorithms written in Python 3.8. the CNN-based Xception network model for the first time has been proposed for valvular heart sound analysis.

PROS

Xception has fewer parameters and requires less computation compared to other deep neural network architectures, such as VGG and ResNet.

CONS

Xception may suffer from overfitting when the dataset is small or noisy.

Xception is mainly designed for image recognition tasks and may not be suitable for other types of data.

Literature Survey - 7:

Name : Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization
Author Names : Jia Wu* | Xiu-Yun Chen | Hao Zhang | Li-Dong Xiong | Hang Lei | Si-Hao Deng
Published In: Elseiver
Year : 2018

This paper considers building the relationship between the performance of the machine learning models and their hyperparameters by Gaussian processes. In this way, the hyperparameter tuning problem can be abstracted as an optimization problem and Bayesian optimization is used to solve the problem. Bayesian optimization is based on the Bayesian theorem. It sets a prior over the optimization function and gathers the information from the previous sample to update the posterior of the optimization function.

PROS

Bayesian HPO provides a systematic and efficient approach for hyperparameter tuning that can save a lot of time and resources compared to manual tuning or grid search.

Bayesian HPO can handle non-convex and non-smooth search spaces and can identify complex and non-linear relationships between hyperparameters.

CONS

Bayesian HPO requires the specification of prior distributions over hyperparameters, which can be difficult and subjective.

The effectiveness of Bayesian HPO can depend on the choice of the acquisition function and the optimization algorithm used to search the hyperparameter space.

Literature Survey – 8 :

Name : Heart Disease Prediction Using Hybrid Genetic Fuzzy Model
Author Names : T. Santhanam1 and E. P. Ephzibah
Published In: Elseiver
Year : 2021

The objective of the work is to diagnose heart disease using computing techniques like genetic algorithm and fuzzy logic. In this paper a hybrid genetic-fuzzy heart disease diagnosis system is designed. The genetic algorithm is used for a stochastic search that provides the optimal solution to the feature selection problem. The relevant features selected from the dataset help the diagnosing system to develop a classification model using fuzzy inference system.

PROS

The use of a hybrid genetic-fuzzy model allows for a more robust and accurate prediction of heart disease compared to using only one type of model.
Stochastic search algorithms are useful for problems with inherent random noise or deterministic problems which can be solved by injected randomness.

CONS

Stochastic search is a random process, and the results may not be consistent across different runs of the algorithm.
The performance of a fuzzy classifier heavily depends on the quality of input features and the selection of fuzzy sets.
The fuzzy rules are difficult to design and can be time-consuming, especially for large datasets.

Literature Survey – 11 :

Name : Heart disease prediction using hyper parameter optimization (HPO) tuning
Author Names : R. Valarmathi, T. Sheela
Published In: Elseiver
Year : 2021

The objective of the work is to improving Random forest classifier and XG Boost classifier model using three Hyper Parameter Optimization (HPO) techniques Grid Search, Randomized Search and Genetic programming (TPOT Classifier) for efficient heart disease risk prediction. The performance of the models is evaluated with the publicly available datasets Cleveland Heart disease Dataset and Z-Alizadeh Sani dataset.

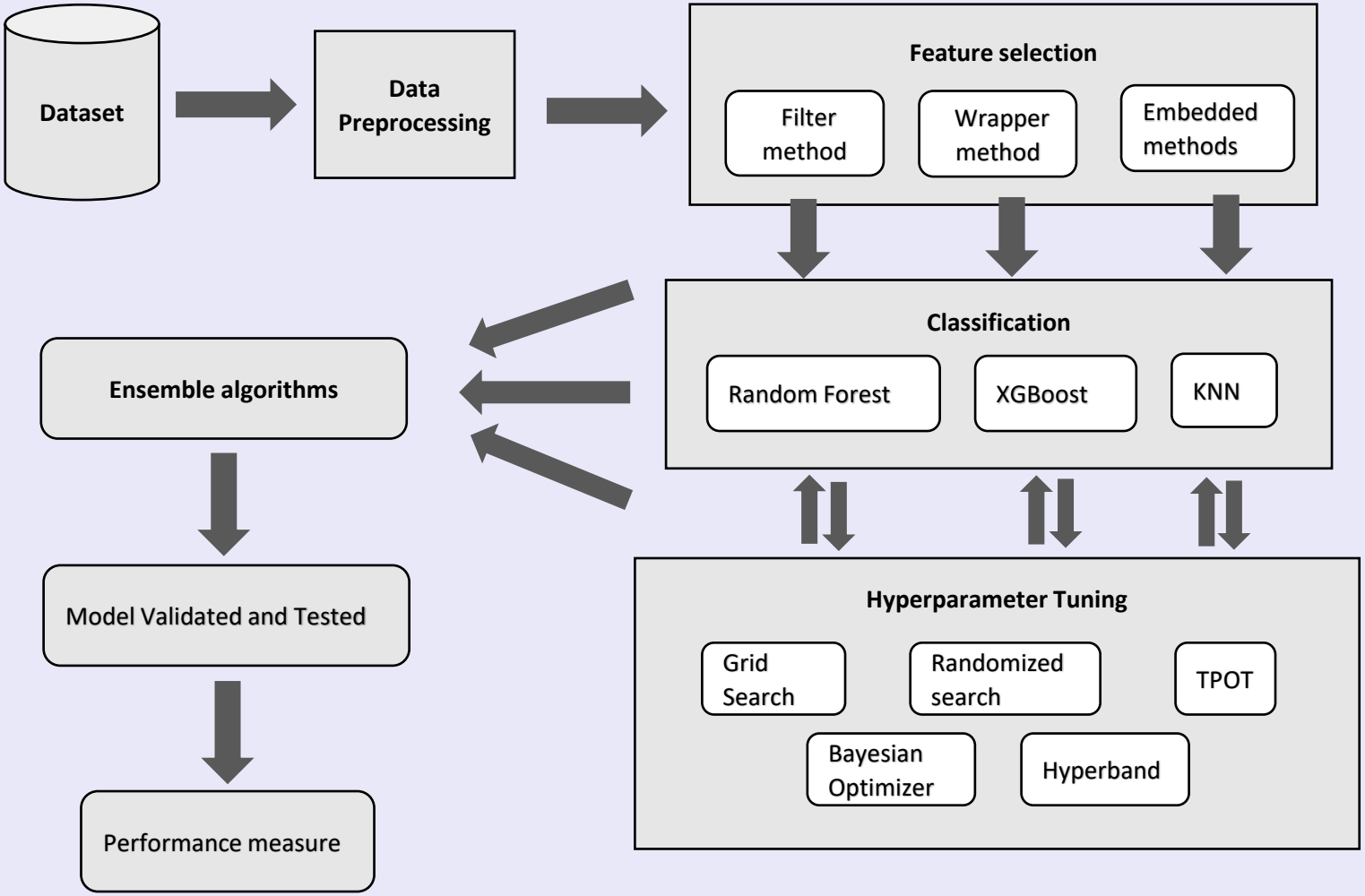
PROS

The use of Sequential feature selection picks the best feature subsets for the corresponding classification models.
Various Hyperparameter optimization techniques have been used to improve the performance of the model.

CONS

Usage of single feature technique cannot be always good for picking up best feature subset for a classification model
Hyperparameter may get stuck at local optima if the search space does not contain the global optimum value for the respective hyperparameters

System Design



Modules:

I. **Data Preprocessing**
   I. Removing duplicate entries
   II. Balancing Dataset

II. **Feature selection**
   I. Filter method (Mutual Information, ANOVA, Chi- square)
   II. Wrapper method (Sequential Forward selection, Sequential Backward elimination, Boruta feature selection)
   III. Embedded method (LASSO technique, Decision Tree, Genetic algorithm)

III. **Model building**
   I. Random Forest Classifier
   II. Extreme Gradient Boosting (XGBoost)
   III. K Nearest Neighbors

IV. **Hyper Parameter Optimization**
   I. Grid Search
   II. Randomized Search
   III. Tree-based Pipeline Optimization Tool (TPOT)
   IV. Bayesian Optimization
   V. Hyperband

Module I :

Data Preprocessing

Removing duplicate entries & Balancing Dataset:

- The rows with duplicate entries are dropped.
- The provided dataset is balanced using **Synthetic Minority Oversampling Technique (SMOTE).** It aims to balance class distribution by randomly increasing minority class examples by replicating them.

Module 2:

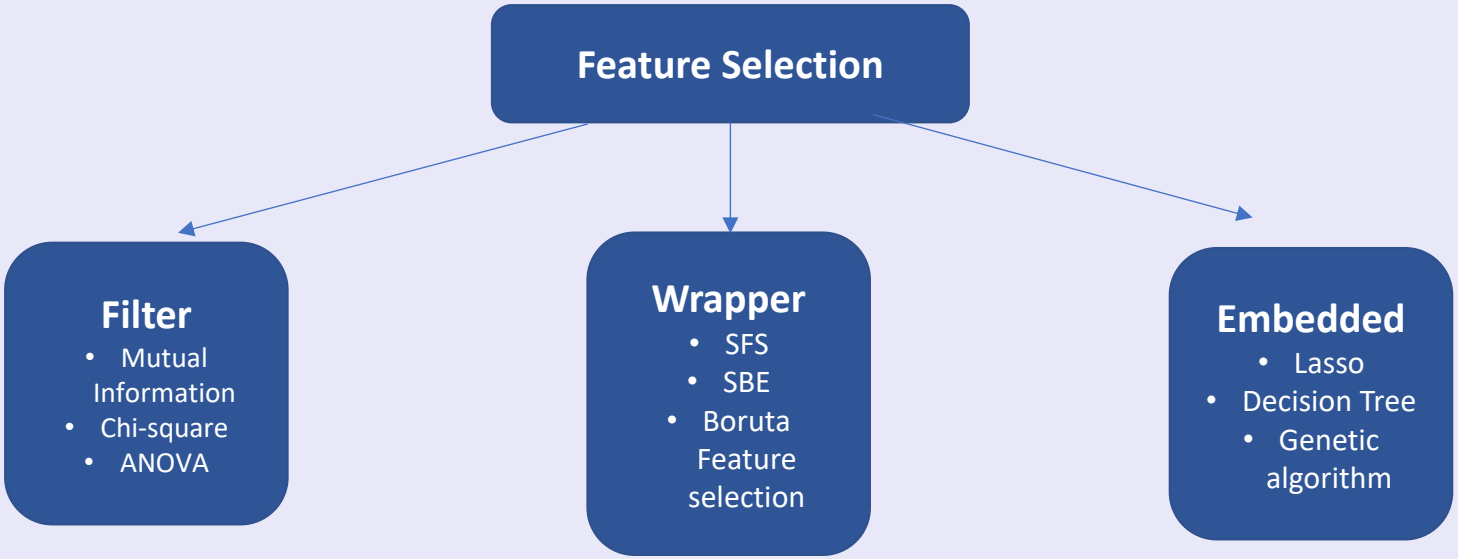
Feature Selection

The irrelevant features decreases the performance of the classification model. To select the best features, 9 different feature selection techniques from 3 different methods. Some of the benefits of feature selection are;

•It helps in avoiding the curse of dimensionality.

•It helps in the simplification of the model so that it can be easily interpreted by the researchers.

•It reduces the training time.



Module 2:

1. Mutual Information:

- Mutual information measures how much information is communicated, on average, in one random variable about another.
- Mutual information is a measure of dependence or "*mutual dependence*" between two random variables. As such, the measure is symmetrical, meaning that ***I(X ; Y) = I(Y ; X)*.**

2. Chi-square test:

We calculate Chi-square between each feature and the target and select the desired number of features with best Chi-square scores. Chi- square score is given by :

$$x^2 = \frac{(Observed\ frequency\ -\ expected\ frequency)^2}{expected\ frequency}$$

where –
Observed frequency = No. of observations of class
Expected frequency = No. of expected observations of class if there was no relationship between the feature and the target.

3. ANOVA test:

- ANOVA - "analysis of variance" and is a parametric statistical hypothesis test for determining whether the means from two or more samples of data (often three or more) come from the same distribution or not.
- Each of the features of the data will be ranked according to the F-statistic component, and the features with the higher scores can be selected as the optimal set of components from the data available.

4. Sequential Forward Selection:

- First, the best single feature is selected using some criterion function. Then, pairs of features are formed using one of the remaining features and this best feature, and the best pair is selected.
- Next, triplets of features are formed using one of the remaining features and these two best features, and the best triplet is selected.
- This procedure continues until a predefined number of features are selected.

5. Sequential Backward Elimination:

- First, the criterion function is computed for all n features. Then, each feature is deleted one at time, the criterion function is computed for all subsets with n-1 features, and the worst feature is discarded.
- This procedure continues until a predefined number of features are left.

6. Boruta Feature Selection:

- Firstly, it adds randomness to the given data set by creating shuffled copies of all features which are called **Shadow Features**.
- Then, it trains a **Random Forest classifier** on this extended data set and applies a feature importance measure such as **Mean Decrease Accuracy**, and evaluates the importance of each feature.
- At every iteration, Boruta Algorithm checks whether a real feature has a higher importance.
- Finally, the Boruta Algorithm stops either when all features gets confirmed or rejected or it reaches a specified limit of random forest.

7. LASSO L1 regularization:

- Least Absolute Shrinkage and Selection Operator, is a statistical formula whose main purpose is the feature selection and regularization of data models.
- It turns out that the Lasso regularization has the ability to set some coefficients to zero. This means that Lasso can be used for variable selection in machine learning.
- If the coefficients that multiply some features are 0, we can safely remove those features from the data. The remaining are the important features in the data.

8. Decision Tree:

- Decision tree is built using the training data, and the importance of each feature is calculated based on its contribution to the decision tree.
- Feature importance will be calculated based on the information gain.

9. Genetic algorithm:

- Genetic algorithm is based on the principle of natural evolution.
- The algorithm tries to 'mimic' the concept of human evolution by modifying a set of individuals called a population, followed by a random selection of parents from this population to carry out reproduction in the form of mutation and crossover.
- This process continues till the stopping criterion is met. In the end, it gives the best individual/solution.

Module 3:

Model Building

1. Random Forest Classifier:

- A random forest is an ensemble classifier that fits a number of decision tree classifiers on several sub-samples of a dataset and uses averaging to improve accuracy and to control overfitting.
- It takes prediction from each decision tree and based on the majority votes of predictions, it predicts the final output.

2. K Nearest Neighbor:

- KNN algorithm is based on the principle that data points that are close to each other in a feature space are likely to belong to the same class or have similar output values.

Module 3:

Model Building

3. Extreme Gradient Boosting (XGBoost):

- XGBoost is an implementation of Gradient Boosted decision trees. Here, decision trees are created in a sequential form.
- Weights are assigned to all independent variables and are fed as inputs to decision tree which predicts the result.
- The weights of variables that are predicted wrong are increased and is fed to another decision tree.

4. Ensemble

- Stacking and Voting based Ensemble models are build using the three classifiers.
- Voting ensemble classifiers take predictions of all classifiers and provides output based on majority votes.
- Stacking ensembles train multiple models to solve similar problems, and based on their combined output, it builds a new model with improved performance.

Module 4:

Hyper Parameter Optimization

1. Grid Search

- It is an exhaustive search that is performed in a specific parameter values of a model, also called as an estimator.
- Grid search brute-forces all possible combinations to find the best value of parameter.

2. Randomized Search

- Random Search is similar to Grid Search but it chooses random combination of hyperparameters rather than brute forcing.
- Here, we don't specify a set of possible values for every hyperparameter.

3. Tree-Based Pipeline Optimization Tool (TPOT)

- TPOT is an AutoML tool designed for efficient construction of optimal pipelines through genetic algorithm.
- The goal of TPOT is to automate the building of ML pipelines by combining a flexible expression tree representation of pipelines with stochastic search algorithms such as genetic programming.

4. Bayesian Optimization:

- Bayesian methods differ from random or grid search in that they use past evaluation results to choose the next values to evaluate.
- The concept is limit expensive evaluations of the objective function by choosing the next input values based on those that have done well in the past.
- Bayesian Optimization is efficient with all types of Hyperparameters (continuous or categorical).

5. Hyperband feature selection:

- Generate small-sized subsets and allocate budgets to each hyper-parameter combination based on its performance.
- Hyperband is essentially just a grid search over the optimal allocation strategy.
- So at each individual trial the set of hyper parameters is chosen randomly.
- Since, it enable parallelization, it converges to the best hyperparameter faster.

Dataset Description:

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, and Long Beach V. It contains 76 attributes, including the predicted attribute, but all published experiments refer to using a subset of 14 of them. The "target" field refers to the presence of heart disease in the patient. It is integer valued 0 = no disease and 1 = disease.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	52	1	0	125	212	0	1	168	0	1.0	2	2	3	0
1	53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
2	70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
3	61	1	0	148	203	0	1	161	0	0.0	2	1	3	0
4	62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
5	58	0	0	100	248	0	0	122	0	1.0	1	0	2	1
6	58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
7	55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
8	46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
9	54	1	0	122	286	0	0	116	1	3.2	1	2	2	0

Dataset Description:

1. age: age in years
2. sex: sex (1 = male; 0 = female)
3. cp: chest pain type
-- Value 0: typical angina
-- Value 1: atypical angina
-- Value 2: non-anginal pain
-- Value 3: asymptomatic
4. trestbps: resting blood pressure (in mm Hg on admission to the hospital)
5. chol: serum cholestoral in mg/dl
6. restecg: resting electrocardiographic results
-- Value 0: normal
-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
7. thalach: maximum heart rate achieved
8. exang: exercise induced angina (1 = yes; 0 = no)
9. oldpeak = ST depression induced by exercise relative to rest
10. slope: the slope of the peak exercise ST segment
-- Value 0: upsloping
-- Value 1: flat
-- Value 2: downsloping
11. ca: number of major vessels (0-3) colored by flourosopy
12. thal: 0 = normal; 1 = fixed defect; 2 = reversable defect
and the label
13. fbs: (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
14. condition: 0 = no disease, 1 = disease

Previous review comments:

We have been asked to expand the feature selection module

Action taken: We have used 9 different feature selection techniques, 3 from each categories and We have tried to compare how a model behaves based on the feature subsets selected by different feature selection techniques

Implementation

Random Forest Classifier:

Filter method:

Technique	Mutual information	ANOVA test	Chi - Square
Accuracy	91.8027	91.8027	**93.4426**
No.of features	11	13	**10**

Features: **'thalach', 'oldpeak', 'ca', 'cp', 'exang', 'chol', 'age', 'trestbps', 'slope', 'sex'**

Wrapper method:

Technique	Seq. Frwd Selection	Seq.Bkwd Elimination	Boruta FS
Accuracy	**90.1639**	88.5246	88.5246
No.of features	**10**	9	8

Features: **'age', 'sex', 'cp', 'trestbps', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'**

Embedded method:

Technique	Lasso	Decision tree	Genetic algorithm
Accuracy	85.2459	83.6065	**88.5246**
No.of features	7	7	**9**

Features: **'age', 'sex', 'cp', 'trestbps', 'restecg', 'exang', 'slope', 'ca', 'thal'**

Random Forest Classifier:

Chi – square features: **'thalach', 'oldpeak', 'ca', 'cp', 'exang', 'chol', 'age', 'trestbps', 'slop 'sex'**

SFS features: **'age', 'sex', 'cp', 'trestbps', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'**

Genetic algorithm features: **'age', 'sex', 'cp', 'trestbps', 'restecg', 'exang', 'slope', 'ca', 'th**

Random forest hyperparameters:

n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features , criterion

	Chi-sq	SFS	Genetic algorithm
Grid search	**95.0819**	**91.8033**	90.1639
Randomized search	93.4426	**91.8033**	91.8033
TPOT	**95.0819**	90.1639	**93.4426**
Bayesian Optimization	93.4426	**91.8033**	90.1639
hyperband	90.1639	90.1639	90.1639

Extreme Gradient boost Classifier:

Filter method:

Mutual information	ANOVA test	Chi - Square
90.1639	88.5246	88.8246
9	13	10

Features: **'cp', 'thal', 'ca', 'slope', 'oldpeak', 'exang', thalach', 'chol', 'sex'**

Wrapper method:

Sequential Frwd Selection	Sequential Bkwd Elimination	Boruta Feature selection
91.803	88.5246	**91.8033**
5	9	**6**

Features: **'cp', 'thal', 'ca', 'slope', 'exang', 'sex'**

Embedded method:

Lasso	Decision tree	Genetic algorithm
78.6885	**80.3279**	80.327
7	**3**	5

Features: **'cp', 'ca', 'thal'**

Extreme Gradient boost Classifier:

Mutual Information features: **'cp', 'thal', 'ca', 'slope', 'oldpeak', 'exang', thalach', 'chol', 'sex'**

Boruta FS features: **'cp', 'thal', 'ca', 'slope', 'exang', 'sex'**

Model based features: **'cp', 'ca', 'thal'**

XGBoost hyperparameters:

n_estimators, max_depth, Learning rate, subsample, min_child_weight

	Mutual Information	Boruta FS	Decision tree
Grid search	**93.4426**	**93.4426**	**93.4426**
Randomized search	88.5245	**93.4426**	**93.4426**
TPOT	90.1639	91.8033	90.8033
Bayesian Optimization	88.5246	**93.4426**	**93.4426**
hyperband	90.1639	90.1639	88.6344

K Nearest Neighbor Classifier:

Filter method:

Mutual information	ANOVA test	Chi - Square
91.8023	85.2459	85.2459
5	8	5

Features: **'thal', 'cp', 'ca', 'oldpeak', 'slope'**

Wrapper method:

Sequential Frwd Selection	Sequential Bkwd Elimination	Boruta Feature selection
75.4098	67.2131	Not applicable
8	8	-

Features: **'age', 'sex', 'cp', 'fbs', 'exang', 'slope', ca', 'thal'**

Embedded method:

Lasso	Decision tree	Genetic algorithm
91.803	83.6065	**91.8033**
6	7	**7**

Features: **'sex', 'cp', 'fbs', 'oldpeak', 'slope', 'ca', 'thal'**

K Nearest Neighbors Classifier:

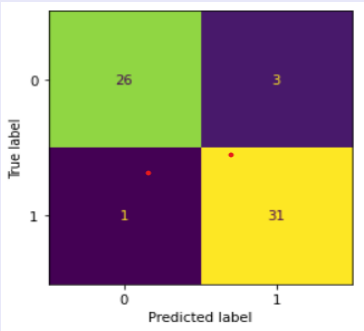
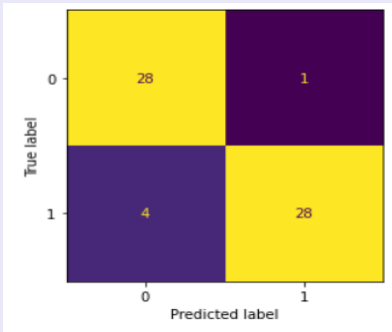
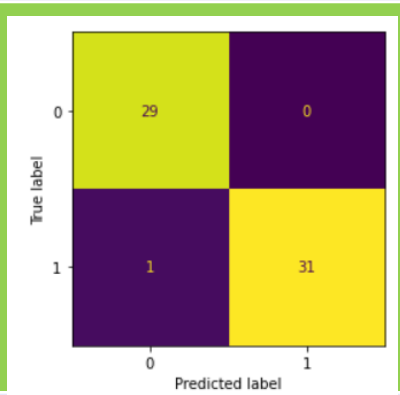
Mutual information features: **: 'thal', 'cp', 'ca', 'oldpeak', 'slope'**

SFS features: **: 'age', 'sex', 'cp', 'fbs', 'exang', 'slope', ca', 'thal'**

Genetic algorithm features: **'sex', 'cp', 'fbs', 'oldpeak', 'slope', 'ca', 'thal'**

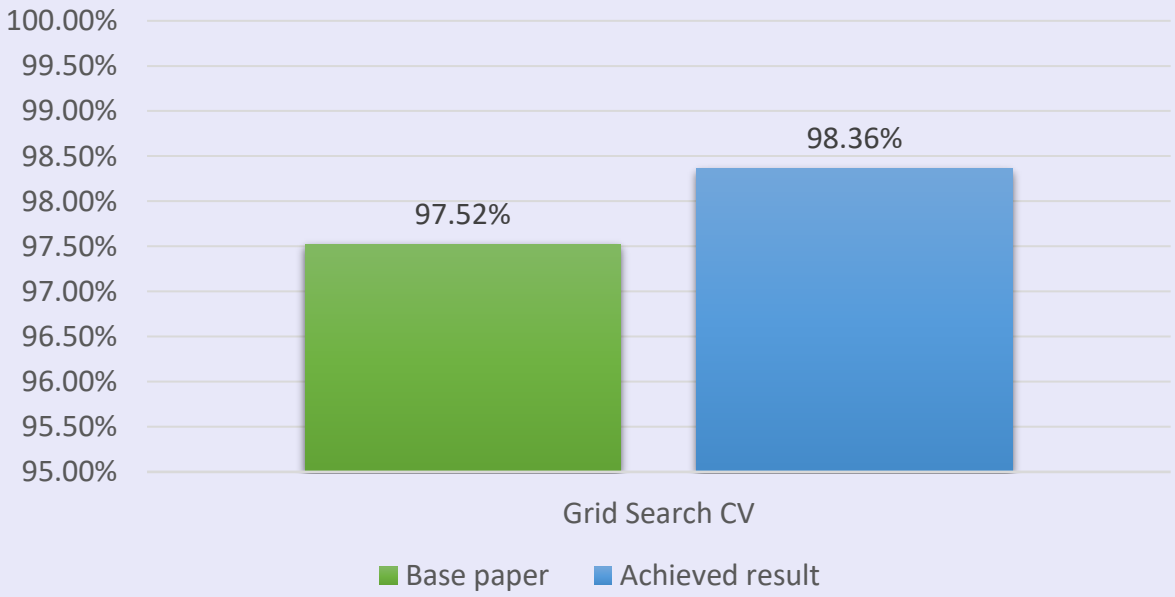
	Mutual Information	SFS	Genetic algorithm
Grid search	**91.8033**	80.3277	**91.8233**
Randomized search	**91.8033**	**80.3279**	90.1639
TPOT	90.1639	**80.3279**	90.1639
Bayesian Optimization	90.1639	78.6885	90.1639
hyperband	83.6344	78.6885	83.6344

Classifier	Feature selection technique	Hyperparameter tuning algorithm	Accuracy
Random forest	Chi-square	Grid search, Bayesian optimization	95.0819
XGBoost	Boruta feature selection	Grid, random, TPOT	93.4426
KNN	Genetic algorithm	Grid search	91.8233

Classifier	Accuracy	Recall	
Voting classifier	93.44%	96.88%	
Stacked Classifier	91.80%	87.55%	
Decision tree	98.36%	96.88%	

Our base paper achieved maximum accuracy of 97.52% using Sequential Forward Selection with Random Forest Classifier and Hyperparameter optimization using TPOT classifier.

Our proposed model achieved accuracy of 98.36% when ensembling Random Forest, XGBoost and KNN using decision tree.



Metrics	Formula
Accuracy	(TP + TN) / (TP + TN + FP + FN)
Recall	TP / (TP + FN)
Precision	TP / (TP + FP)
F1 Score	2 (Precision * Recall) / (Precision + Recall)

Social impacts :

1. Early detection and prevention: By using heart disease prediction models, individuals at high risk of developing heart disease can be identified early, allowing for appropriate interventions and preventive measures.

1.Improved access to healthcare: Heart disease prediction models can be used to target at-risk populations and ensure that they have access to appropriate healthcare services.

2.Increased awareness and education: The use of heart disease prediction models can increase public awareness of heart disease risk factors and encourage individuals to adopt healthier lifestyles.

3. Reduced healthcare costs: Early detection and prevention of heart disease can help to reduce healthcare costs by avoiding costly interventions such as surgery, hospitalizations, and medications.

Economical aspects:

1.Cost savings: Early detection and intervention may also reduce the need for ongoing medical treatment, leading to long-term cost savings.

2.Return on investment (ROI): The ROI of a heart disease prediction system can be measured by the amount of money saved through reduced healthcare costs and improved health outcomes, as well as the potential for increased revenue through improved patient outcomes and satisfaction.

3.Accessibility: The cost of the heart disease prediction system should be considered in the context of accessibility for patients. A system that is too expensive may not be accessible to low-income or uninsured patients, limiting its impact on population health.

4.Data privacy and security: The cost of ensuring data privacy and security must also be considered in the development and implementation of a heart disease prediction system.

Works completed :

Different feature selection techniques have been tested with various classification models.

Various hyperparameter tuning techniques have been applied to those classification models.

References

[1] Wu. Jia, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, Si-Hao Deng, Hyperparameter optimization for machine learning models based on Bayesian optimization, J. Electron. Sci. Technol. 17 (1) (2019).

[2] Debabrata Swain, Preeti Ballal, Vishal Dolase, Banchhanidhi Dash, Jayasri Santhappan, An efficient heart disease prediction system using machine learning. Advances in Intelligent Systems and Computing, Springer Nature Singapore Pvt Ltd, 2020.

[3] R. Das, I. Turkoglu, A. Sengur, Effective diagnosis of heart disease through neural networks ensembles, Expert Syst. Appl. 36 (4) (2009) 7675–7680.

[4] Youness Khourdifi, Mohamed Bahaj, Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization, Int. J. Intell. Eng. Syst. 12 (1) (2018) 242–252.

[5] Roohallah Alizadehsani, Jafar Habibi, Mohammad Javad Hosseini, Hoda Mashayekhi, Reihane Boghrati, Asma Ghandeharioun, Behdad Bahadorian, Zahra Alizadeh Sani, A data mining approach for diagnosis of coronary artery disease, Comput. Methods Programs Biomed. 111 (1) (2013) 52–61, https://doi.org/10.1016/j.cmpb.2013.03.004.

[6] R.S. Olson, J.H. Moore, TPOT: A Tree-Based Pipeline Optimization Tool for Automating Machine Learning, in: F. Hutter, L. Kotthoff, J. Vanschoren (Eds.), Automated Machine Learning. The Springer Series on Challenges in Machine Learning, Springer, Cham, 2019, https://doi.org/10.1007/978-3-030-05318-5_8.

THANK YOU