Ilkogretim Online - Elementary Education Online, 2021; Vol 20 (Issue 5): pp. 7391-7400
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.05.838

A Novel Machine Learning Approaches for Heart Disease Dataset

Arulanantham Zechariah Jebakumar, Lecturer, Prince Sultan Military College of Health Sciences, Dhahran PO Box: 33048, Dammam – 31448, Kingdom of Saudi Arabia, zechariah@psmchs.edu.sa
Dr. R. Ravanan, Joint Director of Collegiate Education, Chennai region, Chennai-15, Tamilnadu, India

Abstract- CVDs are concertedly contributed by hypertension, diabetes, overweight and unhealthy lifestyles. Around 17.9 million people die every year due to heart diseases accounting for 31% of all the deaths in the world. It is important for early and accurate detection of heart diseases This work focuses on the optimal solution for producing pattern by using deductive learning algorithms for heart disease dataset. The RBF Kernel has high correlation coefficient compare with other models which is 0.69. The Linear Kernel has the correlation coefficient value is 0.68. The Puk kernel produces the low correlation coefficient compare with others. The Linear kernel has very low Mean absolute error, Root Mean Squared Error, Relative absolute error and Root squared error which are 0.28, 0.36, 57.16% and 72.90% respectively. This model is comparatively good for other models.

Keywords: Gaussian Processes, Mean Squared Error, Correlation Coefficient, RRSE.

I. I INTRODUCTION

In this section presents introduction of this research work. Heart disease proves to be the leading cause of death for both men and women.[1-4] This affects the human life very badly. Patients with a high risk of heart diseases demonstrate raised blood pressure, glucose, and lipids along with overweight, and obesity. Tobacco use, unhealthy diet, excessive alcohol intake, and inadequate physical activity are leading reasons for heart diseases.[5] Lifestyle also plays an important factor in heart diseases along with physiological factors.[6]Identifying such people and ensuring they are given appropriate treatment could prevent premature deaths. [7]The diagnosis of heart disease in most cases depends on a complex combination and huge volume of clinical and pathological data. Machine learning has been shown to be effective assisting in making decisions and predictions from the large quantity of data produced by the health care industry.[8]

In this paper presents section 2 of this paper explains the detail on the related works. In section 3 presents the materials and methods adopted and section 4 presents the details of the experiments and discussions. Finally section 5 concludes the paper by sharing our inferences and future plans.

II. RELATED WORKS

In this section presents focuses the related works of this research work. The UCI dataset for comparison of different classifiers such as Multilayer perceptron ,Naive Bayes, KNN etc. and validated that SVM with boosting hyper parameters outperformed others.[9] Various kernel implementations with certain rule based classifiers.[10] It concludes that the RBF kernel is best for infinite data and Hyper parameter tuning can be added to make the model more effective. [11] New selection features and methods can be adopted to get broader perception of performance.[12] The accuracy and precision statistics for different algorithms such as Support Vector machines, KNN, Decision Trees, and Neural networks being most popular.[13] The UK Biobank dataset observed that rather than complex models, information gain was better by consideration of different risk factors. a south African dataset consisting of 462 instances for analyzing algorithms such as Naive Bayes, SVM , and decision trees.[14] the accuracy of decision trees in the prediction of heart diseases with the help of a dataset consisting of 573 instances. More number of attributes and hyper parameters can result in better performance classification.[15] The machine learning techniques providing the accuracy of 88.7% in prediction of cardiovascular diseases with a hybrid random forest and linear model.[16] Association rules, clustering and other data mining algorithms prove to be useful to mine huge amounts of unstructured data. [17]. Naive bayes obtained good accuracy results however specificity and sensitivity results can be improved with more instances.[18] The traditional machine learning algorithms that aim in improving the accuracy of heart disease prediction. [19]

III. MATERIALS AND METHODS

In this section presents the materials and methods of this research work. This research work focuses exploratory data analysis and using Weka 3.8.3. The dataset used in this work is UCI Heart Disease dataset. It has 76 features (attributes) from 303 patients. This work uses the dataset consisting of 270 patients with 14 features set.

Table 1: Meta Data Description

S.No	Attribute	Description of the Attribute	Type of the Attribute	Range

1	age	Age in years	Continuous	Minimum = 29
				Maximum = 77

2	sex	0=Female ; 1=Male	Binary	Male=183
				Female=87

3	chest pain type (cp)	Type of the chest pain	Categorical	Asymtomatic=129 Non Angina=79 Atypical Angina=42 Typical Angina=20

4	resting blood pressure (restbps)	in mm Hg on admission to the hospital	Continuous	Minimum=94
				Maximum=200
				Mean= 131.34
				StdDeviation=17.86

5	serum cholestoral (chol)	serum cholestoral in mg/dl	Continuous	Minimum=126
				Maximum=564
				Mean= 249.66
				StdDeviation=51.69

6	fasting blood sugar (fbs)	0=false;
		1=true
			Binary	False=230
				True=40

7	Resting ECG (restecg)	(fbs>120 mg/dl) 0=Normal ; 1=Having ST-T wave abnormality; 2=Showing probable or define left ventricular hypertrophy	Categorical	Normal=131 ST-T Wave Abnormality=2 Left Ventriclar Hypertrophy=137

8	maximum heart rate achieved (thalach)	maximum heart rate reached	Continuous	Minimum=71
				Maximum=202
				Mean= 149.68
				StdDeviation=23.17

9	exercise induced angina (exang)	0=No;
		1=Yes
			Binary	No=181
				Yes=89

10	oldpeak	ST depreve to restse relatission induced by exercise relative to rest	Continuous	Minimum=0
				Maximum=6.2
				Mean= 1.05
				StdDeviation=1.145

11	slope	the slope of the peak exercise ST segment
		0=upsloping;
		1=Flat;
			Categorical	Flat=122
				Upsloping=130
				Downsloping=18

		2=Downsloping		
12	ca	number of major vessels(0-3) colored by flourosopy 0=Typical Angina 1=Atypical Angina 2=Non Anginal Pain 3=Asymptomatic	Categorical	Typical Angina=160 Atypical Angina=58 Non AnginalPain=33 Asymptomatic=19

13	thal	1=normal;
		2=fixed defect;
		3=reversible defect
			Categorical	Normal=152
				Fixed Defect=14
				Reversible Defect=104

14	diagnosis of heart disease (Target or Class)	angiographic disease status 0=Disease; 1=No Disease	Binary	Disease=151 No Disease=119

The GaussianProcesses classifiers implemented in below kernels:

- Linear Kernel : K(x,y)=<x,y>
- Puk Kernel
- RBF Kernel: K(x,y)=exp(-0.01 * (x-y)^2)

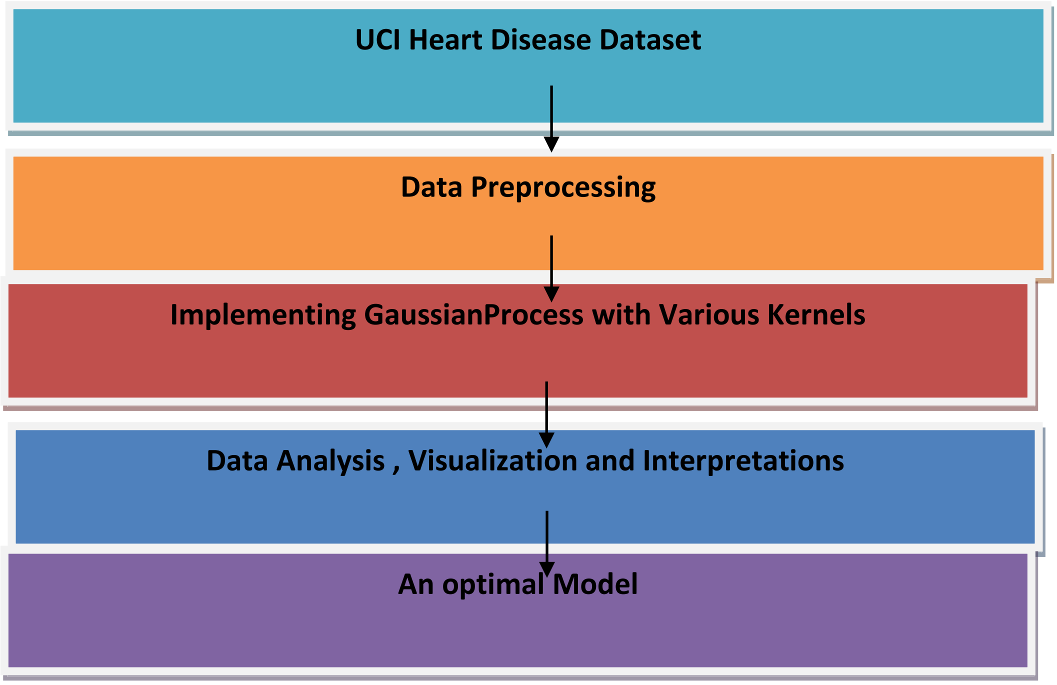

UCI Heart Disease Dataset
Data Preprocessing
Implementing GaussianProcess with Various Kernels
Data Analysis , Visualization and Interpretations
An optimal Model

Figure 1: Architecture of Proposed System

The above diagram represents the proposed architecture of this research work. The Gaussian processes apply with various kernels for this research work.

IV. RESULTS AND DISCUSSIONS

In this section focuses the results and discussions of this research work. This project covers exploratory data analysis like data visualization and implementing K means clustering approaches by using Weka 3.8.3.

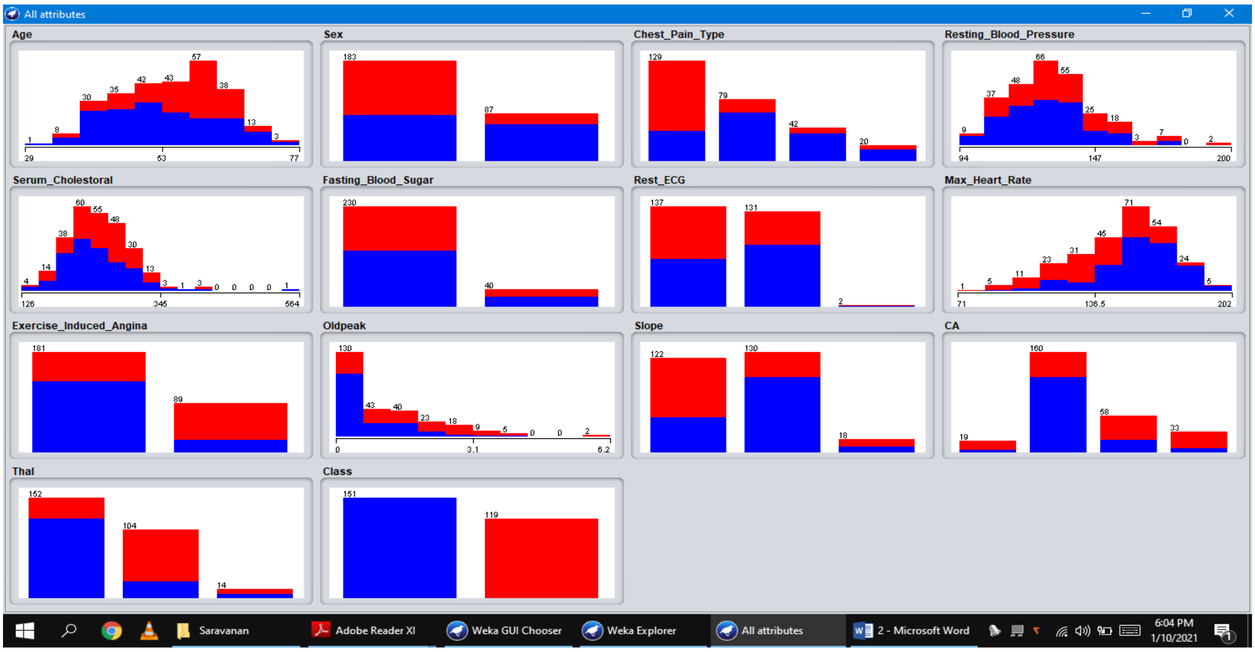
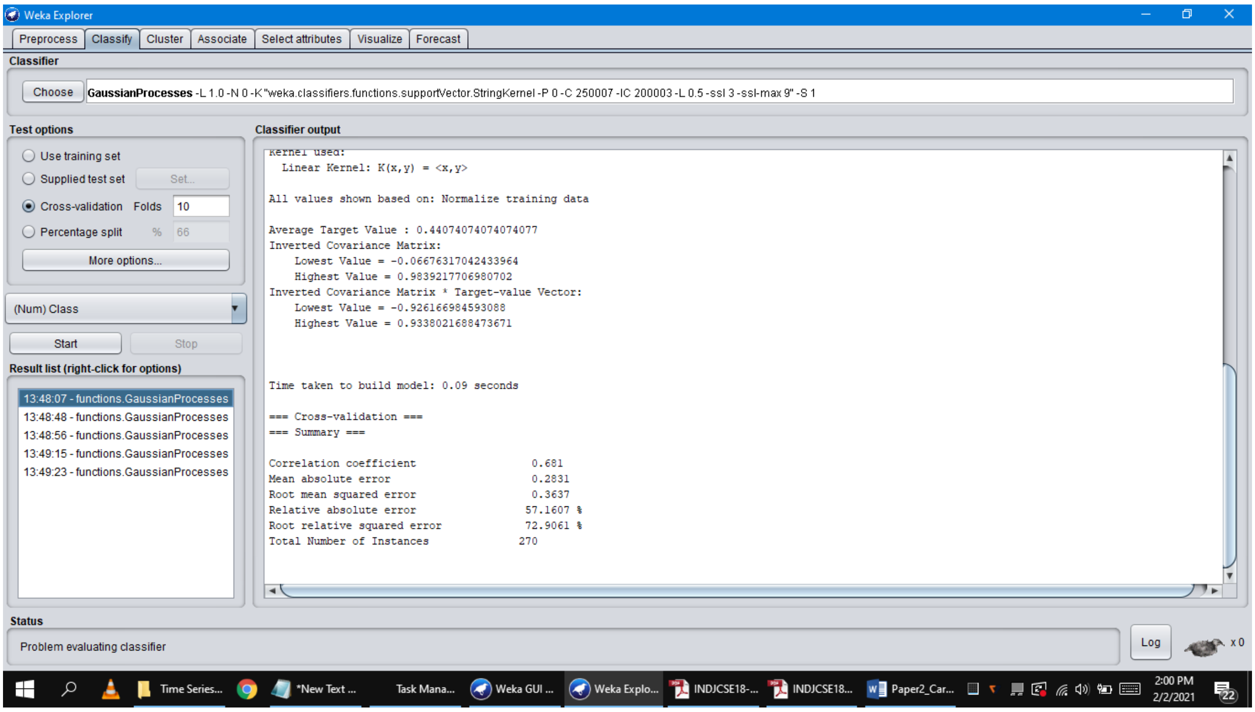


Figure 2: Data Visualization for all Attributes



Classifier

GaussianProcesses -L 1.0 -N 0 -K "weka.classifiers.functions.supportVector.StringKernel -P 0 -C 250007 -IC 200003 -L 0.5 -ssl 3 -ssl-max 9" -S 1

Test options

○ Use training set

○ Supplied test set

● Cross-validation Folds 10

○ Percentage split % 66

(Num) Class

Result list (right-click for options)

13:48:07 - functions.GaussianProcesses
13:48:48 - functions.GaussianProcesses
13:48:56 - functions.GaussianProcesses
13:49:15 - functions.GaussianProcesses
13:49:23 - functions.GaussianProcesses

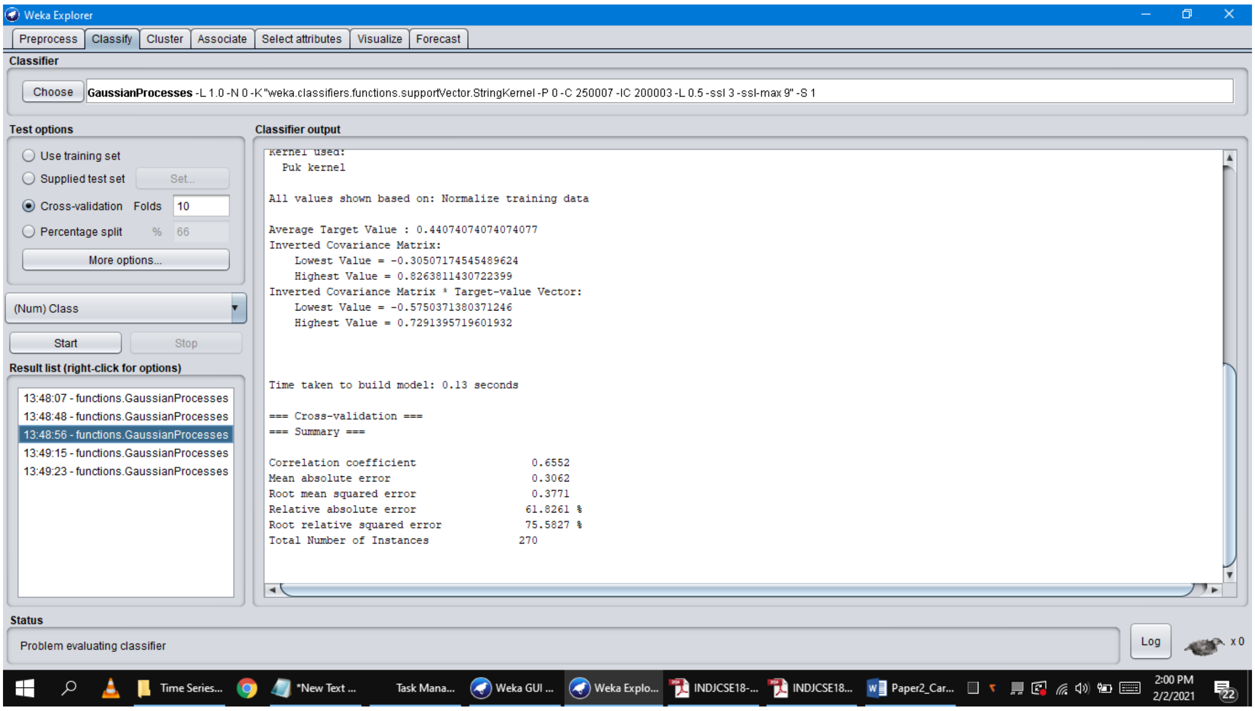
Classifier output

```
Kernel used:
  Linear Kernel: K(x,y) = <x,y>

All values shown based on: Normalize training data

Average Target Value : 0.44074074074074077
Inverted Covariance Matrix:
    Lowest Value = -0.06676317042433964
    Highest Value = 0.9839217706980702
Inverted Covariance Matrix * Target-value Vector:
    Lowest Value = -0.926166984593088
    Highest Value = 0.9338021688473671



Time taken to build model: 0.09 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.681
Mean absolute error                      0.2831
Root mean squared error                  0.3637
Relative absolute error                 57.1607 %
Root relative squared error             72.9061 %
Total Number of Instances              270
```

Status

Problem evaluating classifier

Figure 3: Linear Kernel



Classifier

Choose **GaussianProcesses** -L 1.0 -N 0 -K "weka.classifiers.functions.supportVector.StringKernel -P 0 -C 250007 -IC 200003 -L 0.5 -ssl 3 -ssl-max 9" -S 1

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Num) Class

Result list (right-click for options)

13:48:07 - functions.GaussianProcesses
13:48:48 - functions.GaussianProcesses
13:48:56 - functions.GaussianProcesses
13:49:15 - functions.GaussianProcesses
13:49:23 - functions.GaussianProcesses

Classifier output

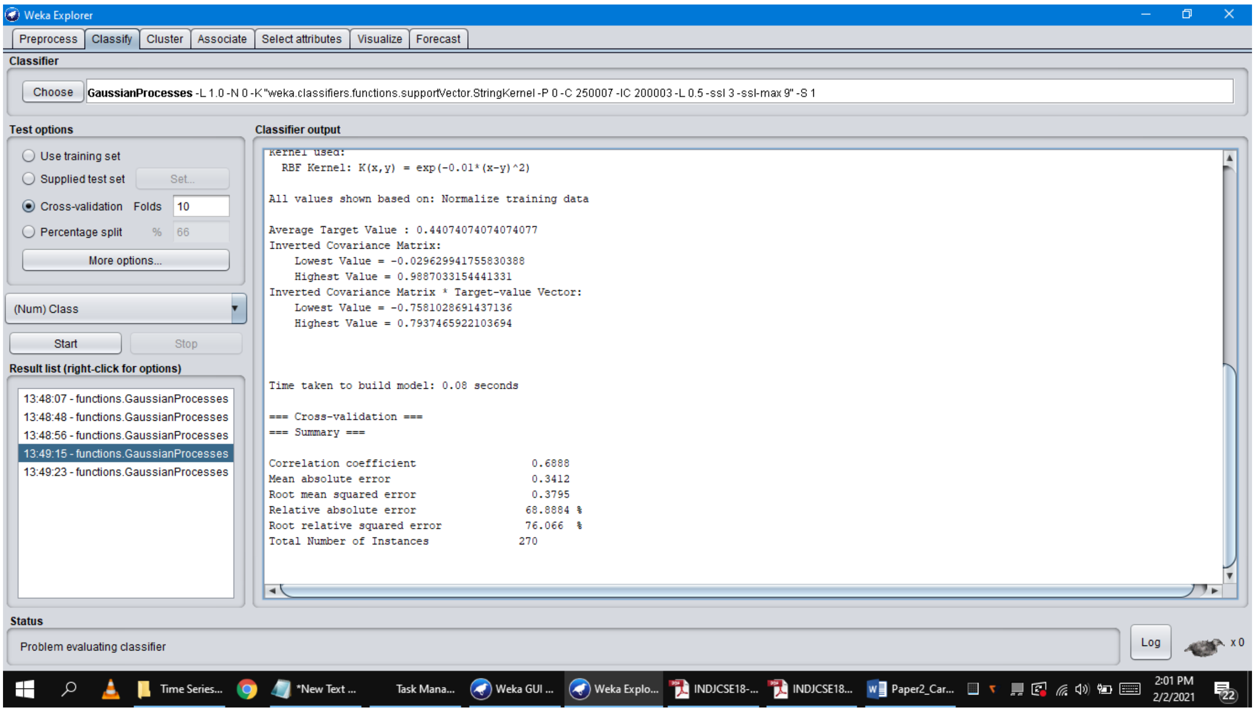
```
Kernel used:
  Puk kernel

All values shown based on: Normalize training data

Average Target Value : 0.44074074074074077
Inverted Covariance Matrix:
    Lowest Value = -0.30507174545489624
    Highest Value = 0.8263811430722399
Inverted Covariance Matrix * Target-value Vector:
    Lowest Value = -0.5750371380371246
    Highest Value = 0.7291395719601932



Time taken to build model: 0.13 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.6552
Mean absolute error                      0.3062
Root mean squared error                  0.3771
Relative absolute error                 61.8261 %
Root relative squared error             75.5827 %
Total Number of Instances              270
```

Status

Problem evaluating classifier

Time Series... *New Text ... Task Mana... Weka GUI ... Weka Explo... INDJCSE18-... INDJCSE18... Paper2_Car... 2:00 PM 2/2/2021

Figure 4: Puk Kernel



Classifier

Choose **GaussianProcesses** -L 1.0 -N 0 -K "weka.classifiers.functions.supportVector.StringKernel -P 0 -C 250007 -IC 200003 -L 0.5 -ssl 3 -ssl-max 9" -S 1

Test options

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Num) Class

Result list (right-click for options)

13:48:07 - functions.GaussianProcesses
13:48:48 - functions.GaussianProcesses
13:48:56 - functions.GaussianProcesses
13:49:15 - functions.GaussianProcesses
13:49:23 - functions.GaussianProcesses

Classifier output

```
Kernel used:
  RBF Kernel: K(x,y) = exp(-0.01*(x-y)^2)

All values shown based on: Normalize training data

Average Target Value : 0.44074074074074077
Inverted Covariance Matrix:
    Lowest Value = -0.029629941755830388
    Highest Value = 0.9887033154441331
Inverted Covariance Matrix * Target-value Vector:
    Lowest Value = -0.7581028691437136
    Highest Value = 0.7937465922103694



Time taken to build model: 0.08 seconds

=== Cross-validation ===
=== Summary ===

Correlation coefficient                  0.6888
Mean absolute error                      0.3412
Root mean squared error                  0.3795
Relative absolute error                 68.8884 %
Root relative squared error             76.066  %
Total Number of Instances              270
```

Status

Problem evaluating classifier

Time Series... *New Text ... Task Mana... Weka GUI ... Weka Explo... INDJCSE18-... INDJCSE18... Paper2_Car... 2:01 PM 2/2/2021

Figure 5: RBF Kernel

Table 2: Various measurements of ML approaches

S.No	Kernels	CCE	MAE	RMSE	RAE	RRSE	Time taken to build the model(In seconds)
1	Linear	0.68	0.28	0.36	57.16%	72.90%	0.09
2	Puk	0.66	0.30	0.38	61.83%	75.58%	0.13

3	**RBF Kernel**	**0.69**	**0.34**	**0.8**	**68.89%**	**76.06%**	**0.08**

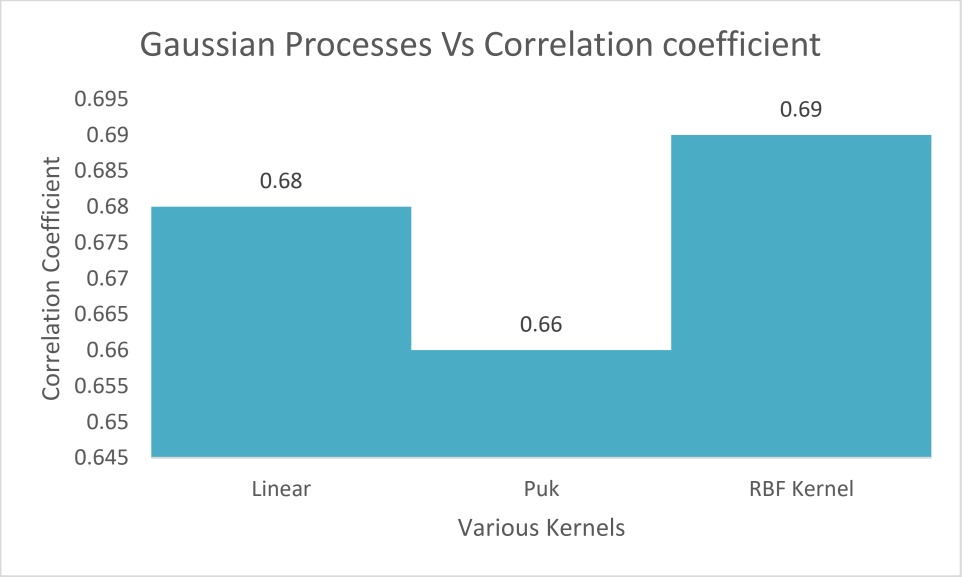


Figure 6: Gaussian Processes (with various kernels) Vs Correlation Coefficient

The above diagram represents that the liner kernel has 0.68 correlation coefficient, Puk kernel produces 0.66 correlation coefficient, RBF kernel has 0.69 correlation coefficient.

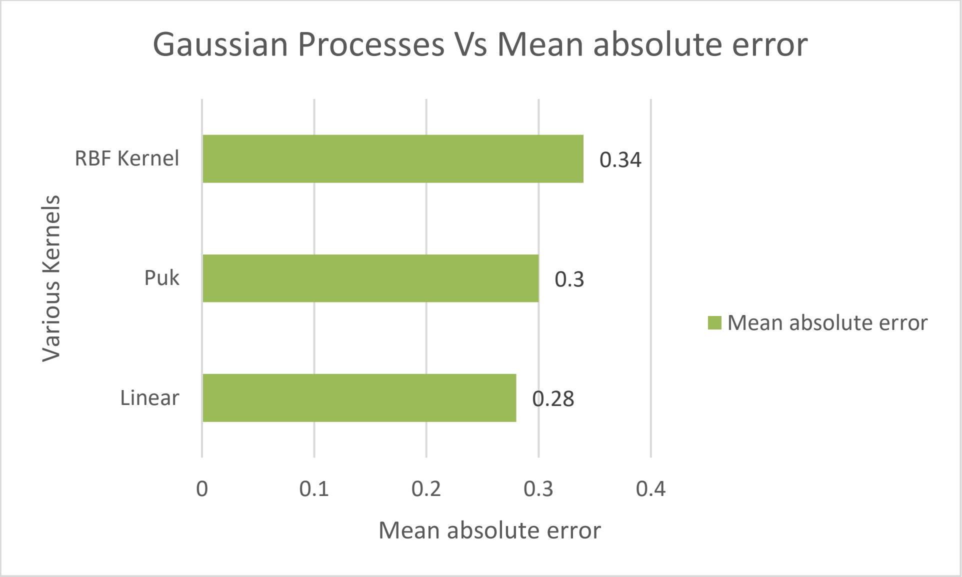


Figure 7: Gaussian Processes (with various kernels) Vs Mean Absolute Error

The above diagram represents that the liner kernel has 0.28 mean absolute error, Puk kernel produces 0.3 mean absolute error, RBF kernel has 0.34 mean absolute error for Gaussian processes in Machine Learning approach.

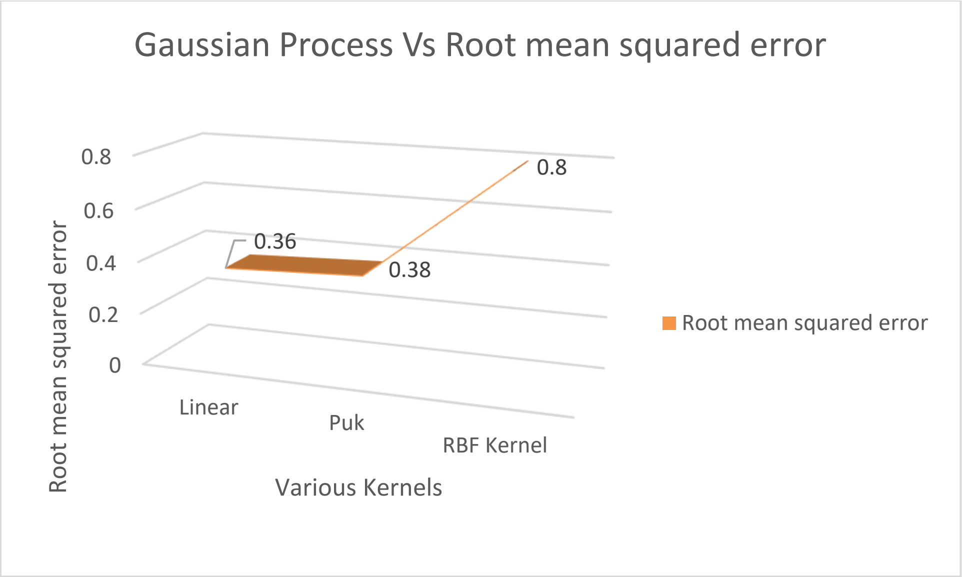


Figure 8: Gaussian Processes (with various kernels) Vs Root Mean Squared Error

The above diagram represents that the liner kernel has 0.36 Root mean squared error, Puk kernel produces 0.38 Root mean squared error, RBF kernel has 0.8 Root mean squared error for Gaussian processes in Machine Learning approach.

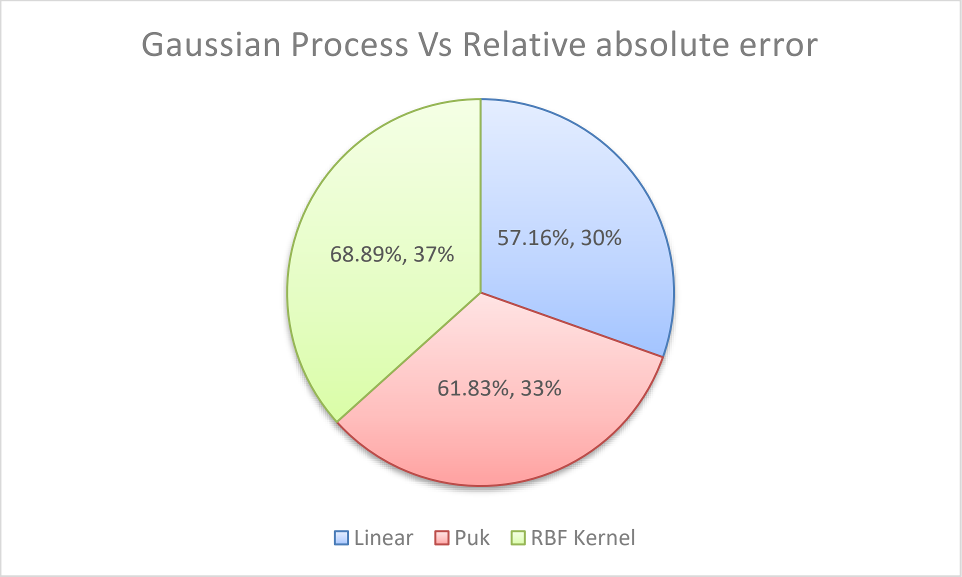


Figure 9: Gaussian Processes (with various kernels) Vs Relative Absolute Error

The above diagram represents that the liner kernel has 57.16% Relative absolute error, Puk kernel produces 61.83% Relative absolute error, RBF kernel has 68.89% Relative absolute error for Gaussian processes in Machine Learning approach.

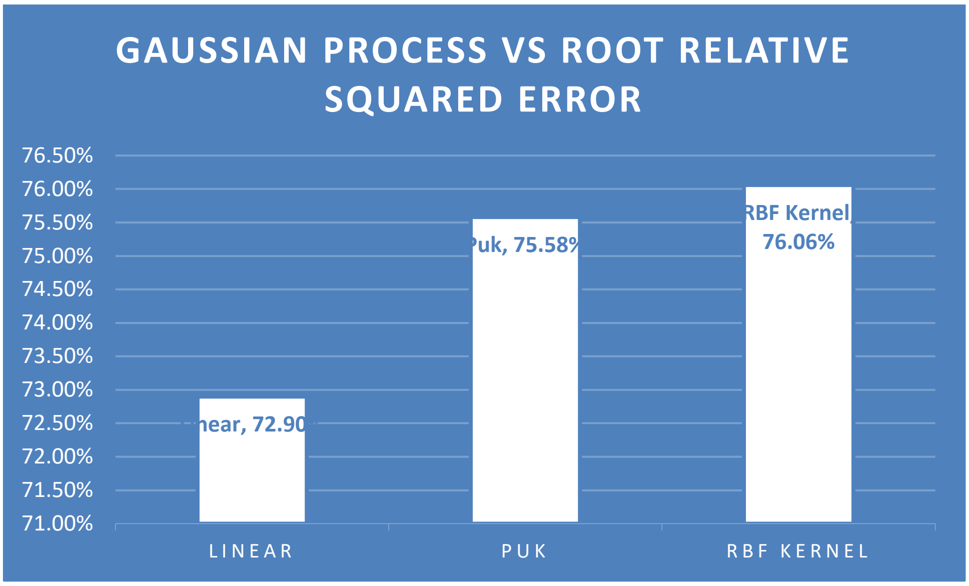
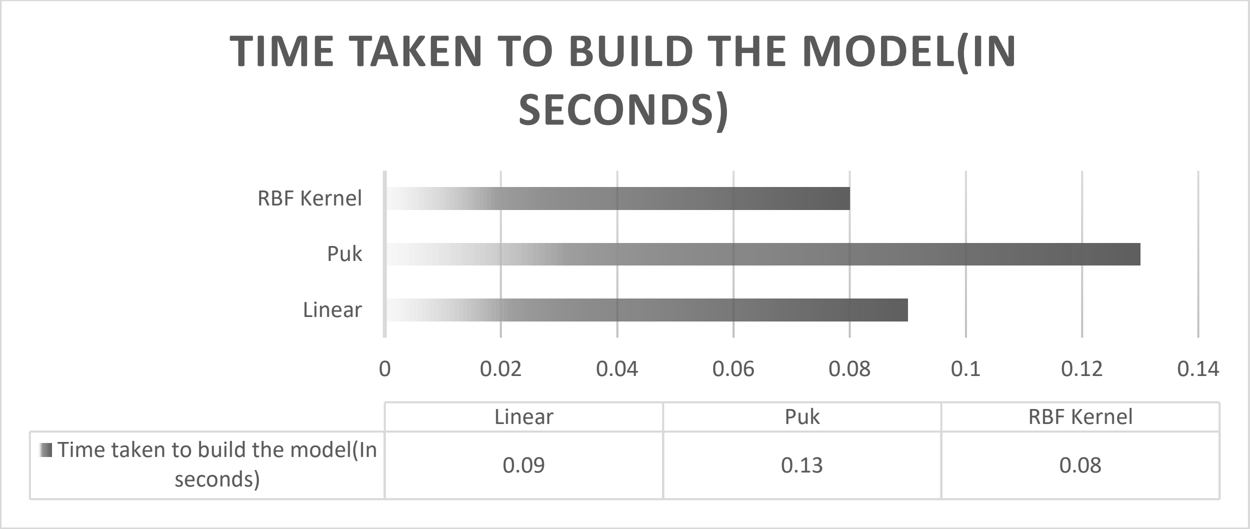


Figure 10: Gaussian Processes (with various kernels) Vs Root Relative Squared Error

The above diagram represents that the liner kernel has 72.90% Root relative squared error, Puk kernel produces 75.58% Root relative squared error, RBF kernel has 76.06% Root relative squared error for Gaussian processes in Machine Learning approach.



	Linear	Puk	RBF Kernel
▮ Time taken to build the model(In seconds)	0.09	0.13	0.08

Figure 11: Gaussian Processes (with various kernels) Vs Time(In Seconds)

The above diagram represents that the liner kernel has taken the time to build the model is 0.09 seconds., Puk kernel has taken the time to build the model is 0.13 seconds, RBF kernel has taken the time to build the model is 0.08 seconds for Gaussian processes in Machine Learning approach.

The RBF Kernel has high correlation coefficient compare with other models which is 0.69. The Linear Kernel has the correlation coefficient value is 0.68. the Puk kernel produces the low correlation coefficient compare with others. The Linear kernel has very low Mean absolute error, Root Mean Squared Error, Relative absolute error and Root squared error which are 0.28, 0.36, 57.16% and 72.90% respectively. This model is comparatively good for other models.

V. CONCLUSION

Finally this work concludes that the RBF Kernel has high correlation coefficient compare with other models which is 0.69. The Linear Kernel has the correlation coefficient value is 0.68. the Puk kernel produces the low correlation coefficient compare with others. The Linear kernel has very low Mean absolute error, Root Mean Squared Error, Relative absolute error and Root squared error which are 0.28, 0.36, 57.16% and 72.90% respectively. This model is comparatively good for other models.

REFERENCES

[1]Cömert, Z., A. F. Kocamaz, 2017. Comparison of machine learning techniques for fetal heart rate classification. Acta Phys. Pol. A 132.3 (2017): 451-454.

[2]Patel, Jaymin, Tejalupadhyay, Samir, Patel,Samir, 2016. Heart Disease Prediction using Machine learning and Data Mining Technique. International Journal of Computing Science and Communication10.090592/IJCSC.2016.018.

[3]S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, J. Gutierrez, 2017. A comprehensive investigation and comparison of Machine Learning Techniques in the domain of heart disease. IEEE Symposium on Computers and Communications (ISCC), Heraklion, 2017, pp. 204-207, doi: 10.1109/ISCC.2017.8024530.

[4] G. Ayyappan ,K.Sivakumar, Heart Disease Data Set Classifications: Comparisons Of Correlation Co Efficient By Applying Various Parameters In Gaussian Processes, Indian Journal of Computer Science and Engineering (IJCSE) , Vol. 9 No. 5 Oct-Nov 2018, Page Number130-134, e-ISSN : 0976-5166, p-ISSN : 2231-3850.

[5] https://www.kaggle.com/mruanova/predict-heart-disease-using-random-forests#Random-Forest-Classifier

[6] https://www.kaggle.com/nyjoey/heart-disease

[7] https://towardsdatascience.com/exploratory-data-analysis-on-heart-disease-uci-data-set-ae129e47b323

[8] Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J., Sandhu, S., Guppy, K., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64,304--310.

[9] David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database.

[10]S. Mohan, C. Thirumalai, G. Srivastava, 2019. Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. J. IEEE Access, vol. 7, pp. 81542-81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[11]Chandna, Deepali, 2014. Diagnosis of Heart Disease Using Data Mining Algorithm. International Journal of Computer Science and Information Technologies, Vol. 5 (2) , 2014, 1678-1680.

[1]Rani, K., 2011. Analysis Of Heart Diseases Dataset Using Neural Network Approach. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.5, September 2011

[13]Karthiga, A. Sankari, M. Safish Mary, M. Yogasins, 2017. Early Prediction of Heart Disease Using Decision Tree Algorithm. International Journal of Advanced Research in Basic Engineering Sciences and Technology 3.3 (2017).

[14]C. Sowmiya, P. Sumitra, 2017. Analytical study of heart disease diagnosis using classification techniques. IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Srivilliputhur, 2017, pp. 1-5, doi: 10.1109/ITCOSP.2017.8303115.

[15]Bahadur, Shamsher, 2013. Predict the Diagnosis of Heart Disease Patients Using Classification Mining Techniques. IOSR Journal of Agriculture and Veterinary Science. 4. 60-64. 10.9790/2380-0426164.

[16]C. Sowmiya, P. Sumitra, 2017. Analytical study of heart disease diagnosis using classification techniques. IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Srivilliputhur, 2017, pp. 1-5, doi: 10.1109/ITCOSP.2017.8303115.

[17]Parthiban, G., Srivatsa, Shesh, 2012. Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients. International Journal of Applied Information Systems. 3. 25-30. 10.5120/ijais12-450593.

[18] G. Ayyappan ,K.Sivakumar, Heart Disease Data Set Classifications: Comparisons Of Correlation Co Efficient By Applying Various Parameters In Gaussian Processes, Indian Journal of Computer Science and Engineering (IJCSE) , Vol. 9 No. 5 Oct-Nov 2018, Page Number135-140, e-ISSN : 0976-5166, p-ISSN : 2231-3850.

[19]Gennari, J.H., Langley, P, & Fisher, D. (1989). Models of incremental concept formation. Artificial Intelligence, 40, 11--61.