Software Effort Estimation using Bio-inspired Rao algorithm

Submitted in partial fulfilment of the requirements of the degree of

Bachelor of Technology (B.Tech) by

K Sai Vivek (197139)
S Sai Vamsi (197171)
K Vasantha Eswari Devi (197140)

Supervisors:

Dr. Manjubala Bisi

Department of Computer Science and Engineering

NIT Warangal

India

Acknowledgement

I would like to express my heartily gratitude towards Dr.Manjubala Bisi Madam for her valuable guidance, supervision, suggestions, encouragement and the help throughout the semester, for the completion of our project work. She kept me going when we were down and gave me the courage to keep moving forward.

I also want to thank evaluation committee for their valuable suggestions and conducting smooth presentation of the project.

(Signature)	(Signature)	(Signature)
K Sai Vivek	S Sai Vamsi	K Vasantha Eswari Devi
197139	197171	197140

Declaration

I declare that this written submission represents our ideas, my supervisor's ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

(Signature)	(Signature)	(Signature)
K Sai Vivek	S Sai Vamsi	K Vasantha Eswari Devi
197139	197171	197140

APPROVAL SHEET

This Dissertation Work entitled **"Software Effort Estimation using Bio-inspired Rao algorithm"** by **K Sai Vivek (197139),S Sai Vamsi (197171),K Vasantha Eswari Devi (197140)** is approved for

the degree of Bachelor of Technology (B.Tech).

Examiners

____

Supervisor

____

Dr. Manjubala Bisi

____

Dr. S. Ravi Chandra (CSE-HOD)

NIT Warangal

Certificate

This is to certify that the Dissertation work entitled **"Software Effort Estimation using Bio-inspired Rao algorithm"** is a bonafide record of work carried out by **"K Sai Vivek (197139),S Sai Vamsi (197171),K Vasantha Eswari Devi (197140)"**, submitted to Dr. Manjubala Bisi of "Department of Computer Science and Engineering", in partial fulfilment of the requirements for the award of the degree of B.Tech at "National Institute of Technology, Warangal" during the 2022-2023.

(Signature)
Dr. Manjubala Bisi
Assistant Professor
"Department of Computer Science and Engineering"

Abstract

Context and Objective: Software effort estimation is an essential feature of software engineering for effective planning, controlling and delivering successful software projects. The failure in Effort estimation accuracy leads to customer disappointment and poor software development process. The objective is to analyze different feature selection algorithms and asses the role they play to increase the accuracy of software development effort predictions.
Dataset: COCOMO (Constructive Cost Model) Dataset, CHINA Dataset. Albrecth Dataset, Kemerer Dataset are used.
Methods: Effort Estimation techniques like Linear Regression, Support Vector Regression (SVR), Random Forest (for Decision Tree) are used and applied to above dataset. Used Non-Bio Inspired Algorithms like Information Gain,Correlation Based feature selection and Bio Inspired Rao Algorithm.
Result: Mean Absolute error (MAE) and Root Mean Squared error (RMSE) are calculated for each Estimate Techniques and compared to predict which model is most suitable for Software Effort Estimation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Software Effort Estimation:

Effort estimation would enable managers to estimate, forecast and accurately quote the requirement for schedule, budget and manpower to successfully complete software projects. Learning about this planning process can help managers to apply it to new development or project initiatives. Software Effort Estimation reduces the failure of a software project indicates recognizable scope, effort and quality failure.

1.2 Feature Selection:

Feature selection is used to improve the performance of predictive model and reduce the computational cost of modelling. It reduces Overfitting, improves the accuracy of the model and reduces Training time. Feature Selection methods include and remove attributes available in data without modifying the attributes.

1.3 Objectives

Objective is to analyze various feature selection algorithms (Bio and Non Bio Inspired Algorithms) Using Rao Algorithm for selecting best subset of features.

Chapter 2

Background

Effort Estimation techniques like Linear Regression, Support Vector Regression (SVR), Random Forest (for Decision Tree) are used and applied to COCOMO,CHINA,ALBRECHT,KEMERER dataset.

Linear Regression: This model attempts to draw a linear relationship between independent variables and a single dependent variable.

Support Vector Regression: SVR is a supervised learning algorithm that is used to predict values and to find the best fit line in hyperplane.

Random Forest: This model creates different decision trees and they are combined by majority vote.

Non-Bio Inspired Feature Selection:

Information Gain Feature Selection: This feature selection technique evaluates the gain of each variable in the context of the target variable.

Correlation Based Feature Selection: Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable.

Bio Inspired Feature Selection [2]:

Binary Rao Algorithm: [8]

It uses both best and worst solutions in each iteration and random interactions among candidate solutions to quickly find an optimum solution.

Chapter 3

Related Work

Software Effort Estimation

According to the chaos report (2015) of The Standish Group International, 60% of IT projects were not on their scheduled time and 56% were not on the budget [4]. From the Literature we observe that Software Effort Estimation is an important task without which it leads to inaccurate budget and scheduling problems. Bio-inspired feature selection algorithms can further improve the accuracy of existing estimation techniques. Past Projects Information can be used to predict software effort estimation of target project. Knowing software development effort is very important before working on a software, we need to create a method which can predict the effort. Using machine learning, past project's information can be used to predict effort for current/target project.

While the bio-inspired algorithms are employed in SDEE for parameter optimization to configure the hyperparameters of the estimation techniques,very few studies have investigated them for feature selection.Genetic Algorithm managed to improve the prediction accuracy of the estimation techniques such as SVM and regression models. Authors employed Genetic Algorithm with a variety of datasets, such as COCOMO, Albrecht, China, Kemerer to extract the best features and produceda best prediction accuracy when compared to the one of the baselines, i.e., Neural Networks (NN), SVM and Bayesian models. Similarly, Genetic Algorithm is used as feature selection algorithm to improve the performance of ensemble methods. The authors have employed and compared the performance of three bio-inspired algorithms (Ant Colony Optimisation, Genetic Algorithm, and Particle Swarm Optimization) which provided different results when used with different datasets. PSO is used to improve the performance of the baseline ANN and compared the performance of Differential Evolution (DE) with PSO and found that DE provided better accuracy than PSO when used in combination with the Analogy based estimation approach.

In [6], Genetic Algorithm is used both for parameter tuning and feature selection to improve the prediction accuracy (when using Cococmo, Nasa, Albrecht, Desharnais datasets). Apart from the bio-inspired algorithms, some traditional feature selection algorithms are also investigated in the context of SDEE.

Two traditional feature selection algorithms, namely, Correlation-based Feature Selection (CFS) and RRelief, have been used in combination with the estimation techniques Support Vector Regression (SVR), Multilayer Perceptron, and Decision Tree, employing a variety of datasets. Similarly,Principal Component Analysis and Correlation-based Feature Selection and the results show that the effort predictions were better than those obtained with Artificial Neural Network. Sarro et al. analyzed the use of the Genetic Algorithm (GA) to configure Support Vector Machine for inter-release fault prediction [7].

Chapter 4

Proposed Solution

In this chapter, we are going to discuss the approach.

4.1 System overview

This is the sample way of including the image in latex

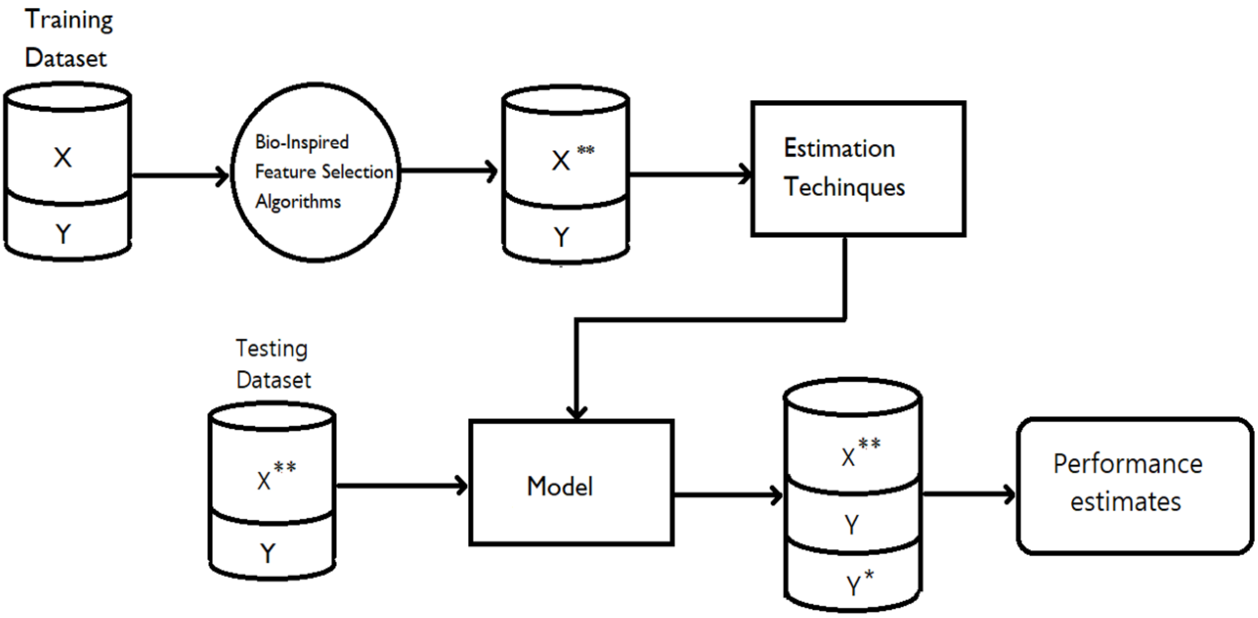


Figure 4.1: Frame Work for proposed Software Estimation Model

X is representing the original dataset features. X** is representing the reduced features. Y is the dependent variable. Y* is the predicted output. Training data is given to the bio-inspired feature selection algorithms which selects subset of features. This dataset is given as input to estimation techniques, and a model is created. Testing data is given as input to this model to predict Y* for each observation. As a conclusion Performance Estimates (MAE, RMSE) are calculated.

4.2 Algorithms

Single Objective Binary Rao Algorithm : It uses both best and worst solutions in each iteration and random interactions among candidate solutions to quickly find an optimum solution.
Information Gain Feature selection Algorithm : This feature selection technique evaluates the gain of each variable in the context of the target variable.
Correlation Based Feature selection Algorithm : Features with high correlation are more linearly dependent and hence have almost the same effect on the dependent variable.
Multi Objective Feature selection Algorithm :[5] It uses both best and worst solutions in each iteration and random interactions among candidate solutions to quickly find an optimum solution.It also considers multi objective function like maximization of R-squared error and minimization of number of features.

Chapter 5

Experiments

COCOMO dataset is used which contains 17 features and 63 instances. Out of these 17 attributes 15 are effort multipliers. These effort multipliers are divided into two types. Some attributes (acap, pcap, tool etc.) are increased to decrease the actual development effort and some attributes (data, time, turn etc.) are decreased to decrease the actual development effort. Similarly with COCOMO,we also used Kemerer, China, Albrecht for effort estimation techniques like: Linear Regression , Support Vector Regression (SVR) , Random Forest (RF).

For each model Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) are calculated to predict best model suitable for effort estimation. Non-Bio Inspired algorithms like : Information Gain , Correlation Based [3] .

Bio Inspired Feature Selection Binary Rao Algorithm : Randomly initialize the Matrix (size: Number of Samples X Number of features) with binary values. Calculate the fitness values [1] of each sample using any estimate technique and find best and worst samples. Update the worst sample using the below objective function if fitness of new sample is more than the present.
Objective Function :
$\mathrm{S}_{new} = S_{worst} + r * (S_{best} - S_{worst})$

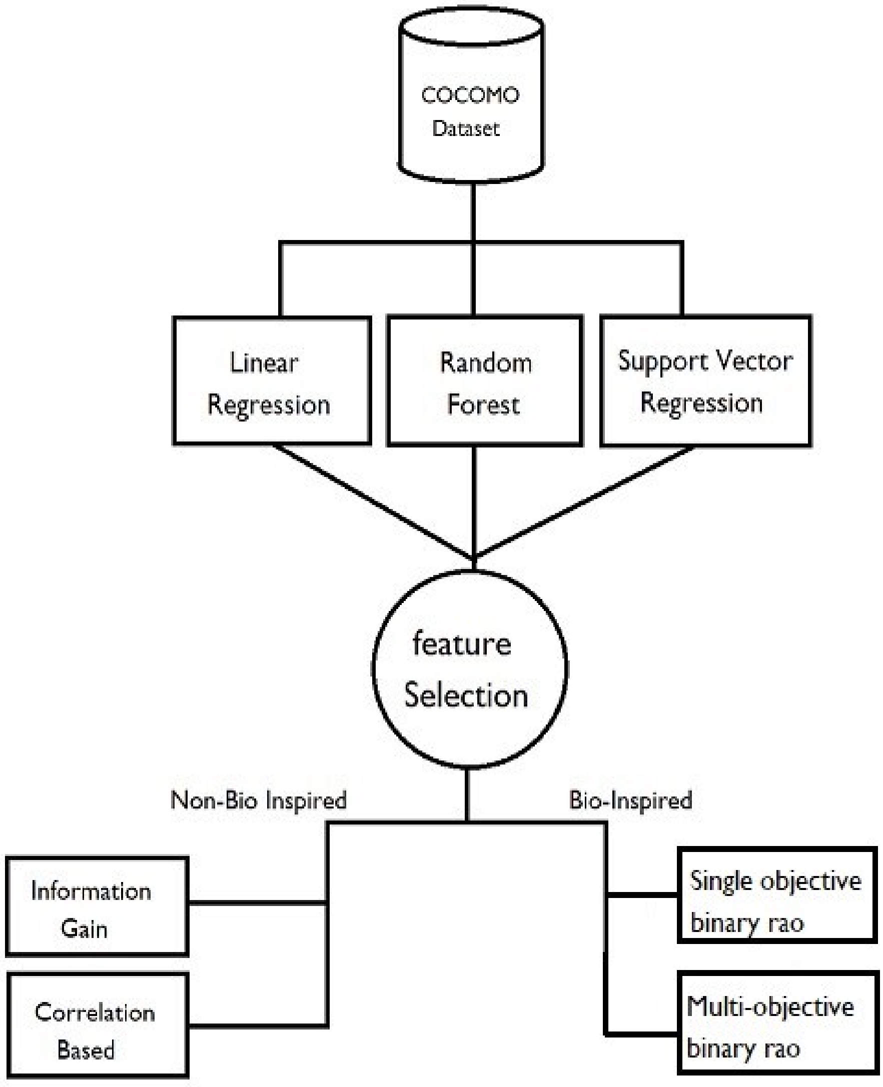
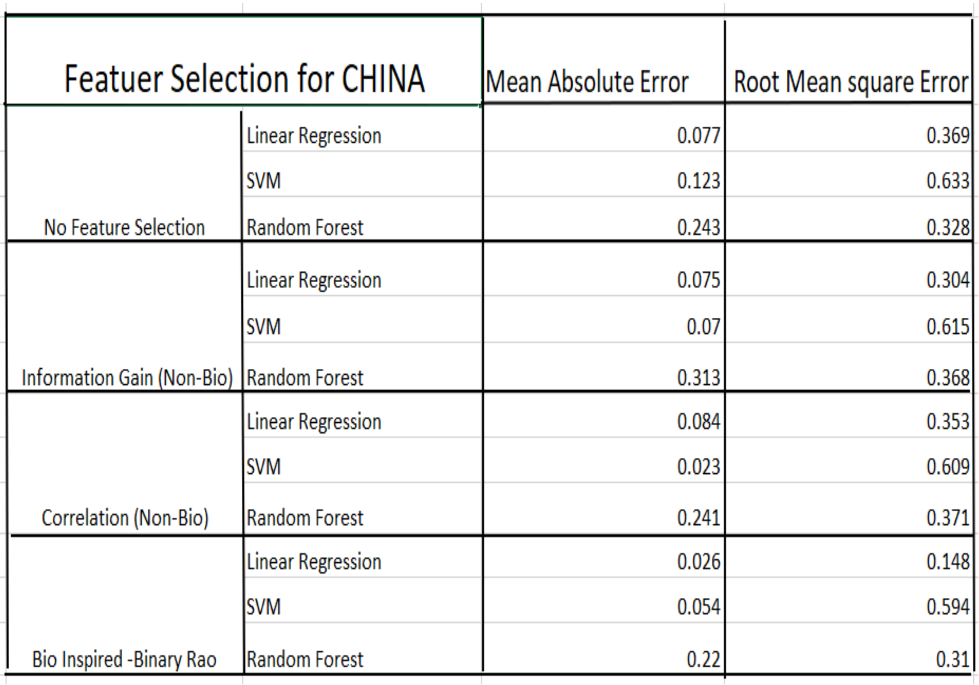
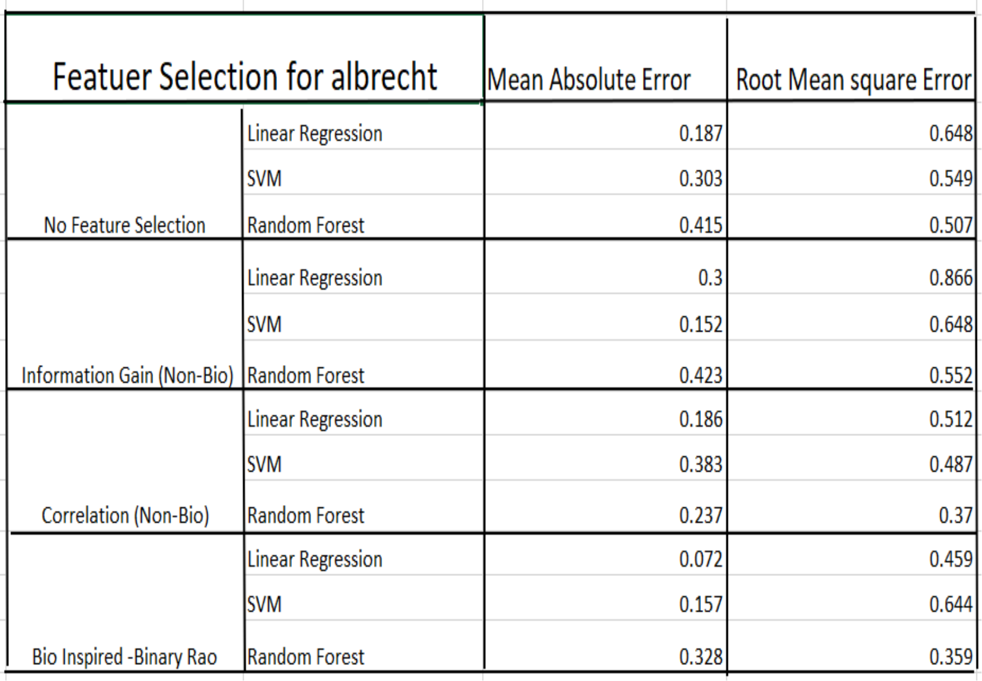


Figure 5.1: feature selection models considered for Software Effort Estimation



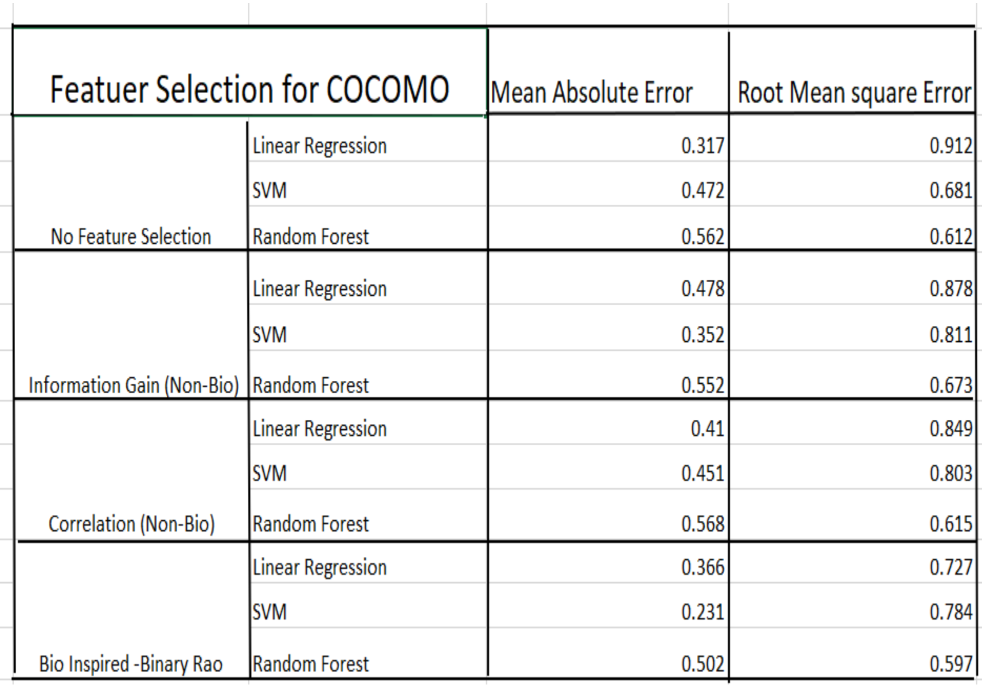
Featuer Selection for CHINA		Mean Absolute Error	Root Mean square Error
No Feature Selection	Linear Regression	0.077	0.369
	SVM	0.123	0.633
	Random Forest	0.243	0.328
Information Gain (Non-Bio)	Linear Regression	0.075	0.304
	SVM	0.07	0.615
	Random Forest	0.313	0.368
Correlation (Non-Bio)	Linear Regression	0.084	0.353
	SVM	0.023	0.609
	Random Forest	0.241	0.371
Bio Inspired -Binary Rao	Linear Regression	0.026	0.148
	SVM	0.054	0.594
	Random Forest	0.22	0.31

Table 5.1: Errors for CHINA dataset



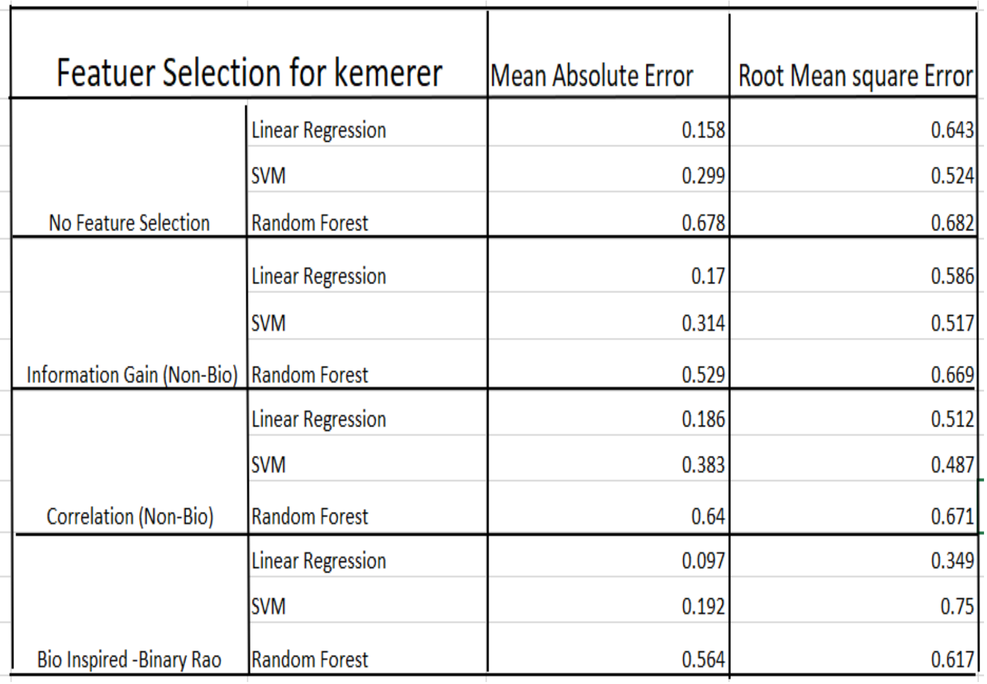
Featuer Selection for albrecht		Mean Absolute Error	Root Mean square Error
No Feature Selection	Linear Regression	0.187	0.648
	SVM	0.303	0.549
	Random Forest	0.415	0.507
Information Gain (Non-Bio)	Linear Regression	0.3	0.866
	SVM	0.152	0.648
	Random Forest	0.423	0.552
Correlation (Non-Bio)	Linear Regression	0.186	0.512
	SVM	0.383	0.487
	Random Forest	0.237	0.37
Bio Inspired -Binary Rao	Linear Regression	0.072	0.459
	SVM	0.157	0.644
	Random Forest	0.328	0.359

Table 5.2: Errors for ALBRECHT dataset



Featuer Selection for COCOMO		Mean Absolute Error	Root Mean square Error
No Feature Selection	Linear Regression	0.317	0.912
	SVM	0.472	0.681
	Random Forest	0.562	0.612
Information Gain (Non-Bio)	Linear Regression	0.478	0.878
	SVM	0.352	0.811
	Random Forest	0.552	0.673
Correlation (Non-Bio)	Linear Regression	0.41	0.849
	SVM	0.451	0.803
	Random Forest	0.568	0.615
Bio Inspired -Binary Rao	Linear Regression	0.366	0.727
	SVM	0.231	0.784
	Random Forest	0.502	0.597

Table 5.3: Errors for COCOMO dataset



Featuer Selection for kemerer		Mean Absolute Error	Root Mean square Error
No Feature Selection	Linear Regression	0.158	0.643
	SVM	0.299	0.524
	Random Forest	0.678	0.682
Information Gain (Non-Bio)	Linear Regression	0.17	0.586
	SVM	0.314	0.517
	Random Forest	0.529	0.669
Correlation (Non-Bio)	Linear Regression	0.186	0.512
	SVM	0.383	0.487
	Random Forest	0.64	0.671
Bio Inspired -Binary Rao	Linear Regression	0.097	0.349
	SVM	0.192	0.75
	Random Forest	0.564	0.617

Table 5.4: Errors for Kemerer dataset

Mean Absolute Error,Root Mean Squared Error are calculated for CHINA dataset (Table 5.1), ALBRECHT dataset (Table 5.2), COCOMO dataset (Table 5.3), KEMERER Dataset(Table 5.4) using the above models with and without feature selection.

From the above tables, we observed that the errors are less, when feature selection is done using single objective binary rao algorithm than feature selection done using non-bio inspired algorithms.

Chapter 6

Conclusion and Future Work

Conclusion:

Estimate Techniques (like Linear Regression,SVM,Random Forest) are applied on COCOMO, China, Albrecht and kemerer datasets without doing any feature selection.Non-Bio Inspired(Information Gain, Correlation Based) and Bio Inspired feature selection(Single Objective Binary Rao Algorithm) are used for feature Selection and same estimate techinques are used to calculate Mean Absolute Error(MAE) and Root Mean Squared Error(RMSE).

From the above tables,we can conclude that Bio Inspired rao algorithm is optimal algorithm for software development effort estimation when compared with Non Bio Inspired algorithms like Information Gain and Correlation Based feature selection models.

Future Work:

Single Objective Binary Rao Algorithm can be further improved by considering Multi Objective Functions.In Single Objective Binary Rao Algorithm, it focuses mainly on Minimization of error, which can be further improved by considering Multi Objective Binary Rao Algorithm where it also focuses on Minimization of number of features and Maximization of Adjusted R Squared.

Bibliography

[1] Farah B Ahmad and Laheeb M Ibrahim. Software development effort estimation techniques using long short term memory. In *2022 International Conference on Computer Science and Software Engineering (CSASE)*, pages 182–187. IEEE, 2022. 5

[2] Asad Ali and Carmine Gravino. Improving software effort estimation using bio-inspired algorithms to select relevant features: An empirical study. *Science of Computer Programming*, 205:102621, 2021. 2

[3] Halcyon Davys Pereira De Carvalho, Roberta Fagundes, and Wylliams Santos. Extreme learning machine applied to software development effort estimation. *IEEE Access*, 9:92676–92687, 2021. 5

[4] Yasir Mahmood, Nazri Kama, Azri Azmi, and Mazlan Ali. Improving estimation accuracy prediction of software development effort: A proposed ensemble model. In *2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, pages 1–6. IEEE, 2020. 3

[5] Chao Ni, Xiang Chen, Fangfang Wu, Yuxiang Shen, and Qing Gu. An empirical study on pareto based multi-objective feature selection for software defect prediction. *Journal of Systems and Software*, 152:215–238, 2019. 4.2

[6] Adriano LI Oliveira, Petronio L Braga, Ricardo MF Lima, and Márcio L Cornélio. Ga-based method for feature selection and parameters optimization for machine learning regression applied to software effort

estimation. *information and Software Technology*, 52(11):1155–1166, 2010. 3

[7] Federica Sarro, Sergio Di Martino, Filomena Ferrucci, and Carmine Gravino. A further analysis on the use of genetic algorithm to configure support vector machines for inter-release fault prediction. In *Proceedings of the 27th annual ACM symposium on applied computing*, pages 1215–1220, 2012. 3

[8] Karpagalingam Thirumoorthy et al. A feature selection model for software defect prediction using binary rao optimization algorithm. *Applied Soft Computing*, 131:109737, 2022. 2