

A Novel Approach in Determining the Reasons of Student Attrition based on Enhanced Genetic Algorithm with Cross-Average Crossover Operator

Markdy Y. Orong, Melphen G. Dumoloy, Ardie G. Besiohan, Warnner D. Amin,
Herwina A. Lonzon & Gretel T. Ricalde

Misamis University, Ozamiz City, Philippines.

DOI: <http://doi.org/10.46382/MJBAS.2021.5106>

Copyright: ©2021 Markdy Y. Orong et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Article Received: 24 September 2020

Article Accepted: 28 January 2021

Article Published: 15 February 2021

ABSTRACT

Prediction helps organizations in their decision-making activities for the improvement of the different functions, services, and even their income. The C4.5 algorithm is the most commonly used prediction algorithm having difficulty in choosing attributes that lead to low prediction accuracy. Hence, integration of *k*-means segmentation for feature selection has been done showing an improvement of prediction accuracy. However, the accuracy result is still not good enough for a prediction model. Thus, there is a need to develop a new strategy to increase the accuracy to an acceptable level. In the present study, Genetic Algorithm (GA) is enhanced and integrated into the prediction model. A new crossover mechanism called Cross-Average Crossover (CAX) is introduced to address the problem of the original GA that can lead to not so accurate prediction results. The student's reasons for attrition in an educational institution was used as the dataset to determine the accuracy of the newly developed model having GA with CAX as against to a model having unenhanced GA. Results showed that an enhanced GA outperformed the unenhanced GA in terms of variable optimization. Further, a prediction model having an improved GA yields a more accurate prediction results as against to a model having unenhanced GA.

Keywords: Accuracy, Cross average crossover, C4.5, Genetic algorithm, Prediction.

1. Introduction

Prediction in research fulfills the desires of humanity to detect the future and know what fate holds. Technologically, prediction is a data mining approach that finds extensive use in the field of business (Al Sonosy, Rady, Badr, & Hashem, 2017; Megahed, Yin, & Nezhad, 2016), health (Alshurafa et al., 2017; Basu & Roy, 2017), education (Amornsinlaphachai, 2016; Kumar, Chowdary, Venkatramaphanikumar, & Kishore, 2016), and many other fields. For organizations, prediction helped management in analyzing data needed for decision-making and ultimately improves performance and gaining competitive advantage over others (Chiu, C., & Shu, 2017; Sugiarto, V. C., Sarno, R., & Sunaryono, 2016). Organizations like department stores, small business enterprises, multi-corporations, churches, hospitals, banks and even schools are making use of the capabilities of prediction mechanism in improving their performance and incomes (Padilla, W. R., Jesús, G. H., & Molina, 2018; Pai, P. F., & Liu, 2018). As the educational systems of the country is improving, schools are embracing the advantages of technology to improve their services to the stakeholders and helped them in decision-making activities (Hasbun, T., Araya, A., & Villalon, 2016). Further, data mining techniques economically offer more customized education, improved system efficiency, and reduce the education process expenses for universities (Devasia, T., Vinushree, T. P., & Hegde, 2016). Thus, predictive analysis is essential for the school administrators in order to help them in their day to day undertakings. Several algorithms are available for organizations in their decision-making undertakings. The predictive algorithms like Naïve Bayes Classification (Aneja & Lal, 2015; Mori, 2016; Walia, Kalra, & Mehrotra, 2016), Support Vector Machine (SVM) (Patel, 2017; H. Zhang, Zhao, Yong, Zhang, & Ji,

2017), Apriori (Jena & Kamila, 2014), Linear Regression (Rahim, M. S., Chowdhury, A. E., Islam, M. A., & Islam, 2017), Logistic Regression Algorithm (Tostes, Silva, Assuncao, Duarte-Figueiredo, & Loureiro, 2016; Usman & Subject, 2016), Artificial Neural Networks (Jiao, Chen, Tian, Hospital, & University, 2016; Kayri & Gencoglu, 2017), Random Forest (Khongchai & Songmuang, 2017; Mago, Srivastava, D., Lewis, & M, 2016; Xu, Zhang, Zhang, & Wei, 2017), C4.5 for Decision Tree (Bjij, Hdoufane, & Jarid, 2016) and Nearest Neighbors (Al-Shehri et al., 2017) has been used by many organization for data analysis and prediction. Among the algorithms, C4.5 is the most commonly used for prediction. However, difficulty to choose attributes is an identified problem of the algorithm that leads to low prediction accuracy. Hence, Moedjiono, Isak, & Kusdaryono, (2016) integrated k-means to C4.5 resulting to a better accuracy from 69% (without K-means) to 79% (with K-means). However, the accuracy result is still not good enough for a prediction model (Zhang et al., 2017). Thus, there is a need to develop a new strategy to increase the accuracy to an acceptable level.

One best strategy in increasing the accuracy of a prediction model is to integrate feature selection or variable optimization (Chandrasekar, Qian, Shahriar, & Bhattacharya, 2017; Maldonado, Flores, Verbraken, Baesens, & Weber, 2015) since it is evident in the study of Almayan & Al Mayyan, (2016), Chou, Hsieh, & Qiu, (2017) that there is an increase in prediction accuracy having a limited number of features. An effective strategy to optimize the number of variables is through Genetic Algorithm (GA) (Paul et al., 2016; Wu et al., 2017). The algorithm is a search heuristic that generates useful solutions to optimization and search problems (Hosseini & Hosseini, 2016). Further, it is one of the powerful algorithms based on evolutionary ideas of natural genetics which generate a population of chromosomes as the solution of the problem, and survival of the fittest function of the GA will be the output optimized subset of chromosomes. Moreover, GA has been applied to solve different kinds of difficult optimization problems with the simulation of natural selection and genetic evolution (Zheng, Liu, Liu, & Du, 2009) and is used to optimize parameters to speed up the prediction process (Ding et al., 2014). However, GA suffers coupling problem in the crossover function (Umbarkar & Sheth, 2015) which can lead to not so accurate results. Hence, there is a need to develop new crossover mechanism for the genetic algorithm to be used for variable optimization that could contribute to a promising accuracy of a prediction model since any decision-making undertakings of an organization may be affected with the accuracy issues of prediction models.

2. Literature Review

Prediction

Prediction is a mechanism that is used to determine what the future may bring to a particular organization such as in business, health, and education. It is a data mining approach that utilized historical data within an organization that is trained to come up with a prediction model. Further, it is a widely used mechanism that helps organizations in many decision-making activities for the improvement of its processes and even its income (Rahim, M. S., Chowdhury, A. E., Islam, M. A., & Islam, 2017; Siregar, B., Nababan, E. B., Yap, A., & Andayani, 2017). Thus, the prediction has an essential role in improving the performance of the business, quality of health care services and in education. There are several existing prediction algorithms that are widely used by organizations (Dong, Z., Zhao, Y., & Chen, 2018; Somwanshi, H., & Ganjewar, 2018) in helping them with their decision-making

undertakings. On the other hand, there are several studies in the field of information technology that combined different algorithms to come up with a prediction model.

The model of Aye, Balcilar, Gupta, & Majumdar, (2015) used 26 forecasting models to forecast South Africa's aggregate seasonal retail sales. They presented some theoretical as well as practical results out of their study. On the other hand, (Ramos, Santos, & Rebelo, 2015) compared the performance of state space models and ARIMA models for predicting sales by applying both to a case study of women footwear retail sales. Further, (Arunraj & Ahrens, 2015) developed an enhanced seasonal ARIMA model for daily food sales forecasting. They showed that their method provides better prediction and profound insights for the effect of demand influencing factors. Moreover, (Žliobaite, Bakker, & Pechenizkiy, 2012) presented a two-level switch model and studied a case of food wholesaler. This sales prediction approach divides the sales time series into predictable and random, then uses intelligent predictor for predictable and moving average for random.

Hybrid Prediction Model

Prediction becomes more advantageous in some organization. Hence, some authors developed a hybrid prediction model with the aim of improving its performance in terms of its accuracy. The study of (Liao, Chen, Liu, & Chiu, 2016) proposed a concept for predicting churn prediction in virtual worlds. In this work, the authors employed a hybrid classification model based on meta-heuristic and machine learning algorithms. It considers and combines the monetary cost, user behavior and social neighbor features in the determination of customer behavior. Thus the proposed model detects the churn with high accuracy within less prediction time. Though the hybrid model with combined features enhances churn prediction, the multi-objective problem occurs when considering many features. Further, the study of (Dalvi, Khandge, Deomore, Bankar, & Kanade, 2016) proposed to build a model for churn prediction for telecommunication companies using data mining and machine learning techniques namely logistic regression and decision trees. The model showed high accuracy with less time of churn prediction. However, the model is only suitable for a few classes. Recently, the study of (Moedjiono et al., 2016) developed a prediction model that combines k-means segmentation and C4.5 algorithms in predicting customer loyalty. The model showed an increase prediction accuracy of 79.33%. However, the accuracy result can still be improved to a much higher percentage. Moreover, the author suggested to include variable optimization to improve its accuracy level.

Accuracy of Prediction Model

The study of (Chandrasekar et al., 2017) improves the prediction accuracy of decision tree mining with data preprocessing utilizing the supervised filter discretization. The study showed that there was an increased number of performances of J48 by approximately 2.63% for training dataset and 10.53% for test dataset which proves that the optimal level of discretization improves both the model and construction time and prediction accuracy of the J48 classifier.

The study of (Leijoto et al., 2014) proposed a physical-chemical feature selection methodology calculated utilizing the structures that compose the proteins through a simple genetic algorithm. The results obtained with the proposed

method were superior to those found in the literature outlined in the study, reaching a precision of 71% and a sensitivity of 68%.

On the other hand, the study of (Lei, Cai, & Zhao, 2017) summarizes on his paper a hybrid model combining EMD and extreme learning machine (ELM), where high-frequency signals removed and processed time series is then modeled and predicted. The prediction performance of the hybrid model compared with that of the ELM-only method created from the original time series. The results show that the proposed hybrid model outperforms the simple ELM method for both short-term and long-term prediction of pole coordinates. The improvement of prediction accuracy up to 360 days in the future is found to be 24.91% and 26.79% on average in terms of mean absolute error (MAE) for the x_p and y_p components of pole coordinates, respectively.

Genetic Algorithm

The genetic algorithm is a search heuristic that generates useful solutions to optimization and search problems (Hossein & Hosseinvand, 2016). It is one of the powerful algorithms based on evolutionary ideas of natural genetics which generate a population of chromosomes as the solution of the problem, and survival of the fitness function of the GA will be the output optimized subset of chromosomes. The essential component of GA is the fitness function which evaluates the chromosomes. Moreover, GA has been applied to solve different kinds of difficult optimization problems with the simulation of natural selection and genetic evolution (Zheng et al., 2009). Further, the algorithm uses the operations of selection, crossover, and mutation to generate the next generation, and they search for an optimal solution from a set of points (Mester, 2014). Moreover, GA is used to optimize parameters to speed up the prediction process (Ding et al., 2014).

Some studies integrate GA into a model with the aim of improving its accuracy. The study of (Liu, M., Zhang, M., Zhao, W., Song, C., Wang, D., Li, Q., & Wang, 2017) utilized genetic algorithm based approaches to predict the plate production process defects, and increase the rate of finished products, and improved enterprise profits, on the base of large-scale industrial data accumulated in medium-thick plate production process. According to the experimental results, 86.7% of the plates in the training set correctly classified. The accuracy rate of good plate prediction was 86.5%, and the bad board prediction accuracy was 100%. 89.6% of the plates correctly classified in the test set, with a good plate hit rate of 90.5%, the wrong board hit rate is relatively low to 55.6%, which means that we can use pre-determine more than 89.6% quality of the plate.

Further, the study of (Wu et al., 2017) proposed a GA method to calibrate the parameter of an improved nonlinear hydrological nitration prediction model. Based on the data from RMSE, PBIAS, and NSE, the Genetic Algorithm method is better than the LOADSET. GA predictions are significantly different from the LOADTEST predictions. The authors also used some other favorite machine learning techniques (generalized linear regression, gradient boosted tree regression, random forest regression, and decision tree regression) to predict nitrate content with the same dataset. The results show GA has the best PBIAS value than other methods. This means GA does least overestimates and underestimate compared with the other six introduced methods. Furthermore, (Yang et al., 2016) proposed a genetic algorithm optimized back propagation (GA-BP) training to solve the problem that the

neural network based spectrum prediction model always trapped in optimal local solution. The performance of GA-BP training compared to the original BP training. The result showed that the prediction accuracy of GA-BP training improved by 50% compared to BP training. The study of (Paul et al., 2016) proposed a new feature selection strategy called GARF (genetic algorithm based on random forest) extracted from positron emission tomography (PET) images and clinical data for outcome prediction in oesophageal cancer. Results showed that the proposed feature selection method improves the outcome prediction compared to other tested methods by at least 8% for predictive study and 11% for the prognostic one. These excellent results confirmed by other evaluation criteria (AUC, sensitivity, and specificity). Machine learning techniques, and particularly RF, provide useful expertise in the selection of subsets of multimodal features. Moreover, (Yang et al., 2016) proposed a prediction model based on clustering search strategy improved genetic algorithm (IGA) and wavelet neural network (WNN) (IGA-WNN) to improve the prediction accuracy of short-term traffic flow. The experimental results show that IGA-WNN model has higher prediction accuracy and a better nonlinear fitting ability compared with the traditional WNN and GA-WNN prediction models. On the other hand, (Ding et al., 2014) designed a hybrid intelligent algorithm to predict river water quality utilizing Combined Principal Component Analysis (PCA), Genetic Algorithm (GA) and Back Propagation Neural Network (BPNN). The use of GA is to optimize the parameters of BPNN. The average prediction rates of non-polluted and polluted water quality are 88.9% and 93.1% respectively, and the global prediction rate is approximately 91%. The water quality prediction system based on the combination of Neural Networks and Genetic Algorithms can accurately predict water quality and provide useful support for real-time early warning systems. The study of (Kanumuri, Pushpalatha, Naidu, & Kumar Singh, 2017) implemented a genetic algorithm to optimize the artificial neural networks used, to predict the mechanical properties of Austenitic Stainless Steel 304 (ASS-304) at elevated temperatures. The results show that the proposed hybrid, neural network - the genetic model is a more accurate and effective method for predicting the mechanical properties of ASS-304 at elevated temperatures.

C4.5 Algorithm

One related algorithm in data mining concept is C4.5, which is an algorithm of classification problem in machine learning and data mining. C4.5 created by J. Ross Quinlan, named like that because C4.5 is a descent from ID3 approaching that popular in decision tree (Moedjiono et al., 2016). Further, C4.5 is a known algorithm widely used to design decision trees (Mantas & Abellán, 2014). In the tree diagram, the tree's root illustrated as the first question, and every branch will be called tree's branch which consisted of tested value in attributes testing. Existing branch will branch until the last branch that is called leaf. Leaf is a type of data label which has been testing, can be named as the result of classification or the result of data prediction (Moedjiono et al., 2016).

3. Research Framework

The enhanced model is composed of two significant stages, the variable optimization stage and the clustering and prediction stage as presented in Figure 2-1. In variable optimization stage, the genetic algorithm is enhanced by integrating a new crossover mechanism called Cross-average crossover to obtain the optimized variables out of the given dataset. The end product of the variable optimization stage is the removal of insignificant and retaining the

significant variables that are used as input for the clustering and prediction stage. The clustering and prediction stage utilized the k-means and C4.5 algorithm.

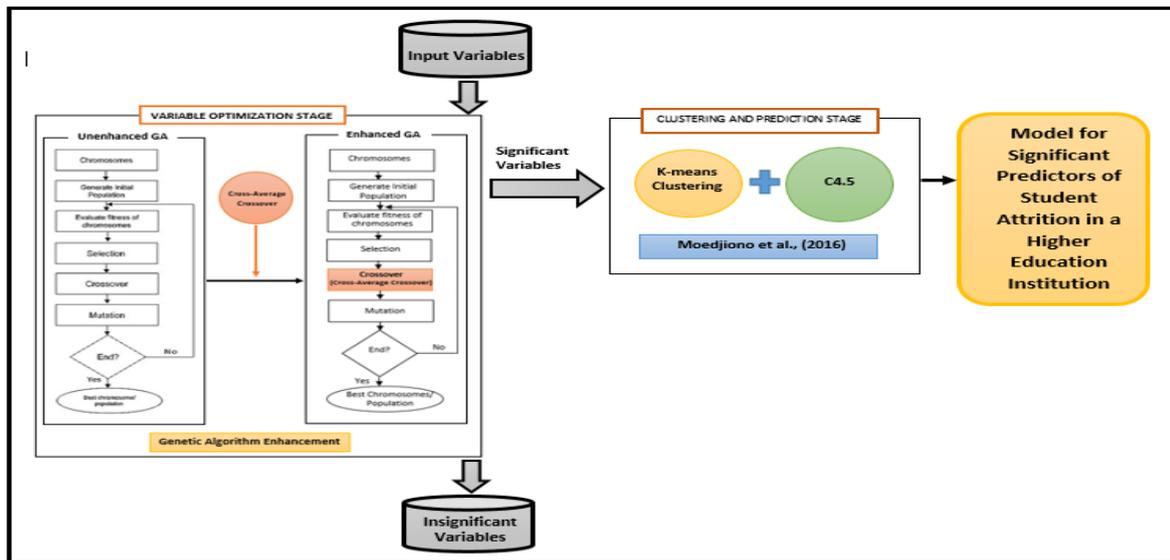


Fig.1. Conceptual Framework of the Study

4. Methods

GA Crossover Operator Enhancement

The new GA operator called cross-average crossover choose the first gene of the first chromosome and the last gene of the second chromosome. The offspring of the mated genes is produced by calculating the average. The process is repeated until the last gene of the first chromosome and first gene of the second chromosome is mated.

Variable Optimization

The number of variables in the datasets used is optimized utilizing the enhanced GA having the Cross-Average Crossover mechanism.

Data Clustering

The optimized variables in the dataset are used for data clustering. The study utilized K-means as the clustering algorithm. It is one of the simplest unsupervised learning algorithms that solve the known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. The clustering procedure is simulated using the WEKA software.

Decision Tree Generation

To complete the prediction process of the model, the study determined the decision tree based on the clustered datasets using C4.5 algorithm. Simulation of the C4.5 is done using the WEKA software.

Accuracy Testing

A 10-fold cross-validation is used to determine the accuracy of the model having GA with cross-average crossover integrated to k-means and C4.4 and a model with original GA to k-means and C4.5.

Model Validation

To validate the accuracy of the derived prediction model, the test data which is the 30% of the total dataset was used. The validation is done by the WEKA software. The correctly classified instances is the basis of the research to determine the accuracy. Accuracy having 85% and above is highly acceptable accuracy (Ding et al., 2014; Liu, M., Zhang, M., Zhao, W., Song, C., Wang, D., Li, Q., & Wang, 2017).

5. Results and Discussions

Variable Optimization Results

The Figure 2 presented the comparison of the enhanced GA and the original GA in terms of the number of optimized number of variables.

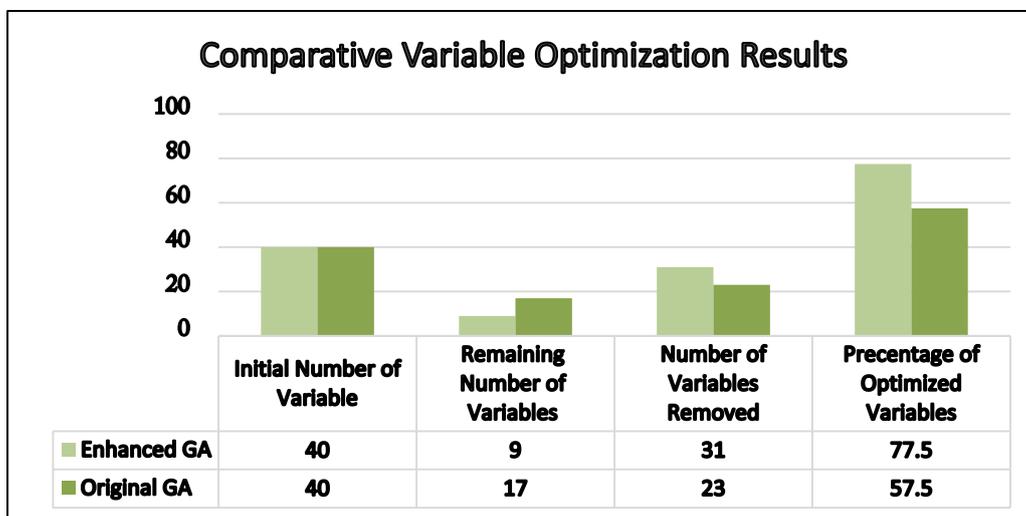


Fig.2. Comparative Variable Optimization Result

From the Figure 2, it can be seen that the number of optimized variables is 77.5% using the enhanced GA is comparable to that of the number of optimized variables using original GA showing only 57.5% of variables that are removed from the dataset.

Comparative Accuracy Results

In order to test the accuracy of the model having an enhanced GA against the model having original GA, a 10-fold cross validation is used. Utilizing WEKA software, the accuracy of the two model is determined. Table 1 presented the comparative results of the accuracy of the model having enhanced GA against the model having original GA.

Table 1. Comparative Accuracy results of a Model with enhanced GA and original GA

Prediction Model	Correctly Classified Instances	Incorrectly Classified Instances	Root Mean Squared Error	ROC Area
Model with Enhanced GA	93.17 %	6.82%	0.25	0.5
Model with Original GA	87.20%	12.80%	0.33	0.5

As presented in Table 1, it is evident that the prediction model having an enhanced GA yields a high prediction rate based on the correctly classified data of 93.317% that outperformed the accuracy of a model having original GA that obtained an accuracy of 87.20%. The graphical results of the accuracy of the models are presented in Figure 3.

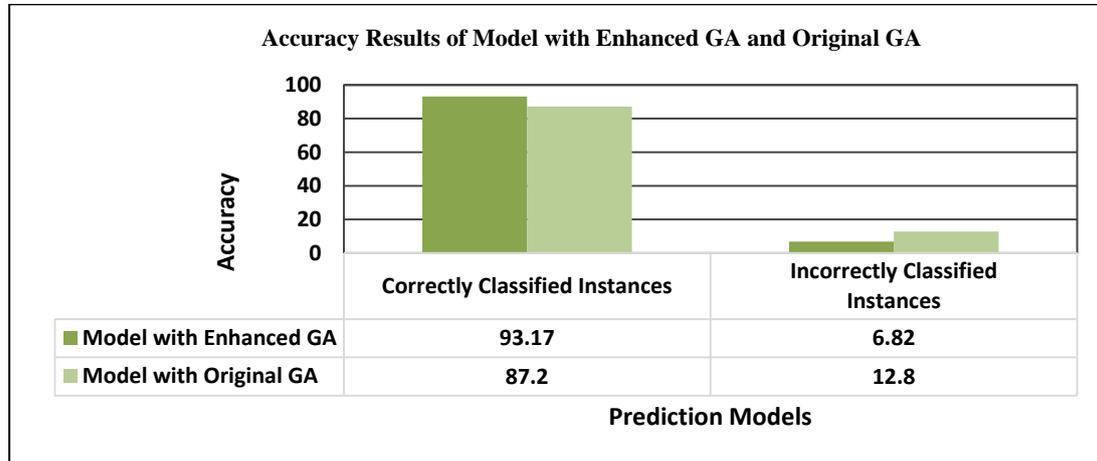


Fig.3. Accuracy Results of Model with Enhanced GA and Original GA

Model Validation Results

To validate the accuracy of the enhanced prediction model having an enhanced GA, a 10-fold cross validation was used with the test data. Table 2 presented the validation results of the prediction models.

Table 2. Model Validation Results

Prediction Model	Correctly Classified Instances	Incorrectly Classified Instances	Root Mean Squared Error	ROC Area
Model with Enhanced GA	91.24%	8.76%	0.28	0.526
Model with Original GA	80.51%	19.49%	0.40	0.49

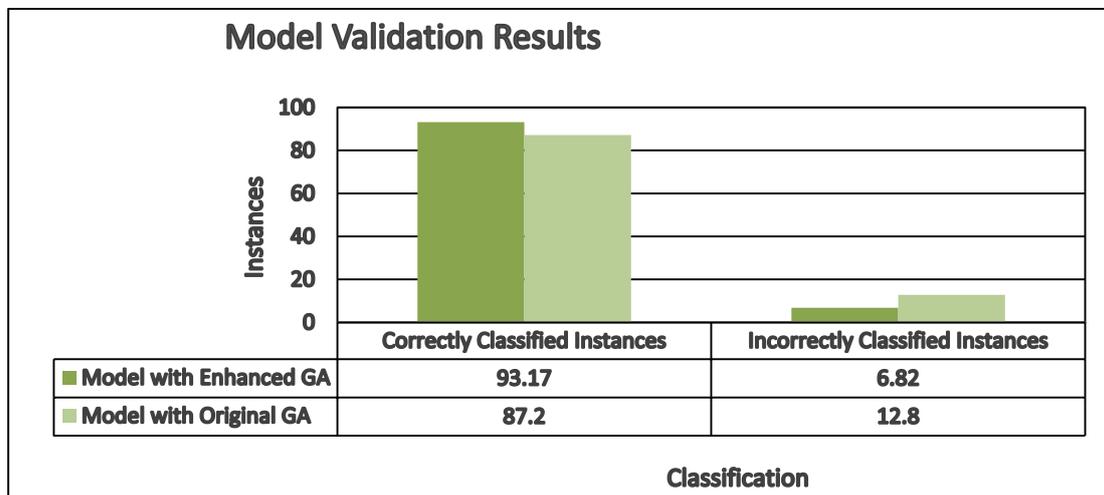


Fig.4. Model Validation Results

As presented in Table 2, the validation results showed that the model having an enhanced GA yields 91.24% accuracy which is good enough to outperform the model having original GA having 80.51% accuracy. The graphical model validation results are presented in Figure 4.

Predictors of Student Attrition

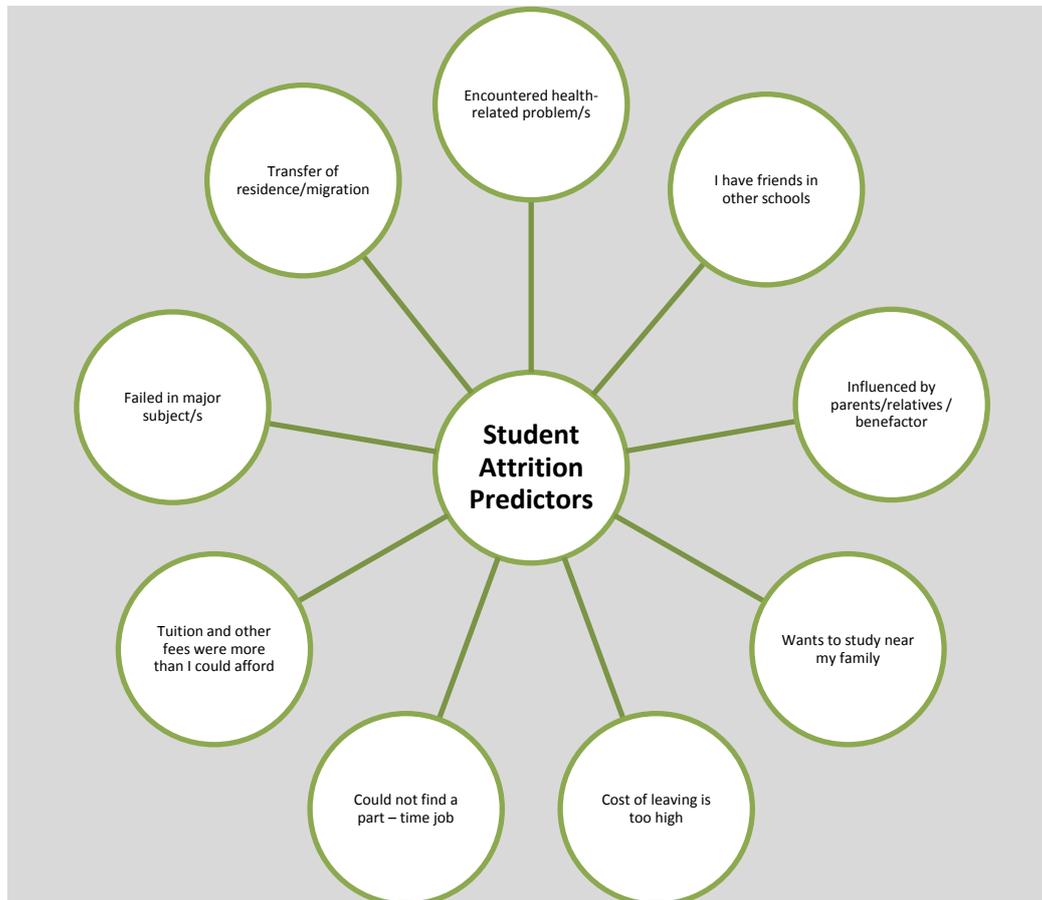


Fig.5. Predictors of Student Attrition

Figure 5 presented the different predictors of student's attrition in a university. The list was derived from the optimized number of variables in the datasets used utilizing the enhanced genetic algorithm. There are nine (9) identified predictors as presented in the figure.

6. Conclusion and Recommendations

An enhanced GA with CAX produces a more optimized number of variables which is 77.5 % as against to the original GA having 57.5% of the variables removed. Integration of enhanced GA with CAX to K-means segmentation and C4.5 algorithm produces a high accuracy hybrid prediction model which is 93.17% as against to a model having an original GA having 87.20%. The model having enhanced GA yields 91.24% as against the model having an original GA having 80.51% based on the model validation results. The list of student's attrition predictors produced in the study helped educational institutions in designing preventive measures in mitigating the attrition rate in an educational institution. The conclusions drawn affirm that the objectives of the study were realized. Utilization of other feature selection algorithm to improve the accuracy of the model is recommended and

present a comparative analysis with the results. Further, there is a need to conduct comparative study with the variable optimization utilizing the enhanced GA with CAX as against to GA having different existing crossover operators. On the other hand, the author suggested to explore other possible operators used in crossover function in GA and conduct comparative analysis on its desired output.

References

- Al-Shehri, H., Al-Qarni, A., Al-Saati, L., Batoaq, A., Badukhen, H., Alrashed, S., Olatunji, S. O. (2017). Student performance prediction using Support Vector Machine and K-Nearest Neighbor. Canadian Conference on Electrical and Computer Engineering, 17–20. <https://doi.org/10.1109/CCECE.2017.7946847>
- Al Sonosy, O., Rady, S., Badr, N. L., & Hashem, M. (2017). A study of spatial machine learning for business behavior prediction in location based social networks. Proceedings of 2016 11th International Conference on Computer Engineering and Systems, ICCES 2016, 266–272. <https://doi.org/10.1109/ICCES.2016.7822012>
- Almayan, H., & Al Mayyan, W. (2016). Improving accuracy of students' final grade prediction model using PSO. Proceedings of the 6th International Conference on Information Communication and Management, ICICM 2016, 35–39. <https://doi.org/10.1109/INFOCOMAN.2016.7784211>
- Alshurafa, N., Sideris, C., Pourhomayoun, M., Kalantarian, H., Sarrafzadeh, M., & Eastwood, J. A. (2017). Remote Health Monitoring Outcome Success Prediction Using Baseline and First Month Intervention Data. IEEE Journal of Biomedical and Health Informatics, 21(2), 507–514. <https://doi.org/10.1109/JBHI.2016.2518673>
- Amornsinsinlaphachai, P. (2016). Efficiency of data mining models to predict academic performance and a cooperative learning model. 2016 8th International Conference on Knowledge and Smart Tech. (KST), 66–71.
- Aneja, S., & Lal, S. (2015). Effective asthma disease prediction using naive Bayes - Neural network fusion technique. Proceedings of 2014 3rd International Conference on Parallel, Distributed and Grid Computing, PDGC 2014, 137–140. <https://doi.org/10.1109/PDGC.2014.7030730>
- Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. International Journal of Production Economics, 170, 321–335. <https://doi.org/10.1016/j.ijpe.2015.09.039>
- Aye, G. C., Balcilar, M., Gupta, R., & Majumdar, A. (2015). Forecasting aggregate retail sales: The case of South Africa. International Journal of Production Economics, 160, 66–79. <https://doi.org/10.1016/j.ijpe.2014.09.033>
- Basu, S., & Roy, S. (2017). Convolutional Regression Framework for Health Behavior Prediction, 214–221.
- Bjij, I., Hdoufane, I., & Jarid, A. (2016). Molecular modeling : Application of Support Vector Machines and Decision Trees for anti-HIV activity prediction of organic compounds, 1–4.
- Chandrasekar, P., Qian, K., Shahriar, H., & Bhattacharya, P. (2017). Improving the Prediction Accuracy of Decision Tree Mining with Data Preprocessing. 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 481–484. <https://doi.org/10.1109/COMPSAC.2017.146>

- Chiu, C., & Shu, C. H. (2017). Monthly car sales prediction using Internet Word-of-Mouth (eWOM). In *INnovations in Intelligent SysTems and Applications (INISTA)*. 2017 IEEE Int. Conf. (pp. 345-348).
- Chou, C.-H., Hsieh, S.-C., & Qiu, C.-J. (2017). Hybrid genetic algorithm and fuzzy clustering for bankruptcy prediction. *Applied Soft Computing*, 56, 298–316. <https://doi.org/https://doi.org/10.1016/j.asoc.2017.03.014>
- Dalvi, P. K., Khandge, S. K., Deomore, A., Bankar, A., & Kanade, V. A. (2016). Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression.
- Devasia, T., Vinushree, T. P., & Hegde, V. (2016). Prediction of students performance using Educational Data Mining. In *2016 International Conference on Data Mining and Advanced Computing*. (pp. 91-95). IEEE.
- Ding, Y. R., Cai, Y. J., Sun, P. D., & Chen, B. (2014). The Use of Combined Neural Networks and Genetic Algorithms for Prediction of River Water Quality. *Journal of Applied Research and Technology*, 12(3), 493–499. [https://doi.org/10.1016/S1665-6423\(14\)71629-3](https://doi.org/10.1016/S1665-6423(14)71629-3)
- Dong, Z., Zhao, Y., & Chen, Z. (2018). Support vector machine for channel prediction in high-speed railway communication systems. In *2018 IEEE MTT-S International Wireless Symposium (IWS)* (pp. 1-3). IEEE.
- Hasbun, T., Araya, A., & Villalon, J. (2016). Extracurricular activities as dropout prediction factors in higher education using decision trees. In *Advanced Learning Technologies (ICALT), 2016 IEEE 16th Int. Conference*.
- Hossein, G., & Hosseinvand, H. (2016). A New Feature Selection IDS based on Genetic Algorithm and SVM. *IEEE 2016 8th International Symposium on Telecommunications (IST)*, 1–6.
- Jena, L., & Kamila, N. K. (2014). A Model for Prediction of Human Depression Using Apriori Algorithm. *2014 International Conference on Information Technology*, 240–244. <https://doi.org/10.1109/ICIT.2014.65>
- Jiao, J., Chen, C., Tian, X., Hospital, C. U., & Universiy, J. (2016). propagation artificial neural network model (BP-ANN) model in cirrhosis patients. <https://doi.org/10.1109/ITME.2016.47>
- Kanumuri, L., Pushpalatha, D. V., Naidu, A. S. K., & Kumar Singh, S. (2017). A Hybrid Neural Network - Genetic Algorithm for Prediction of Mechanical Properties of ASS-304 at Elevated Temperatures. *Materials Today: Proceedings*, 4(2), 746–751. <https://doi.org/10.1016/j.matpr.2017.01.081>
- Kayri, I., & Gencoglu, M. T. (2017). Prediction of Power Production from a Single Axis Photovoltaic System by Artificial Neural Networks, 212–215.
- Khongchai, P., & Songmuang, P. (2017). Random Forest for Salary Prediction System to Improve Students' Motivation. *Proceedings - 12th International Conference on Signal Image Technology and Internet-Based Systems, SITIS 2016*, 637–642. <https://doi.org/10.1109/SITIS.2016.106>
- Kumar, S. C., Chowdary, E. D., Venkatramaphanikumar, S., & Kishore, K. V. K. (2016). M5P model tree in predicting student performance: A case study. *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, 1103–1107. <https://doi.org/10.1109/RTEICT.2016.7808002>

- Lei, Y., Cai, H., & Zhao, D. (2017). Improvement of the prediction accuracy of polar motion using empirical mode decomposition. *Geodesy and Geodynamics*, 8(2), 141–146. <https://doi.org/10.1016/j.geog.2016.09.007>
- Leijoto, L. F., Rodrigues, T. A. D. O., Zaratey, L. E., & Nobre, C. N. (2014). A Genetic Algorithm for the Selection of Features Used in the Prediction of Protein Function. 2014 IEEE International Conference on Bioinformatics and Bioengineering, 168–174. <https://doi.org/10.1109/BIBE.2014.42>
- Liao, H. Y., Chen, K. Y., Liu, D. R., & Chiu, Y. L. (2016). Customer Churn Prediction in Virtual Worlds. *Proceedings - 2015 IIAI 4th International Congress on Advanced Applied Informatics, IIAI-AAI 2015*, 115–120. <https://doi.org/10.1109/IIAI-AAI.2015.265>
- Liu, M., Zhang, M., Zhao, W., Song, C., Wang, D., Li, Q., & Wang, Z. (2017). Prediction of congestion degree for optical networks based on bp artificial neural network. In *Optical Communications and Networks (ICOON), 2017 16th International Conference on* (pp. 1-3). IEEE.
- Mago, N., Srivastava, S., D., R., Lewis, L. E. S., & M, S. (2016). Prediction of Apnea of Prematurity in Neonates using Support Vector Machines and Random Forests, 693–697.
- Maldonado, S., Flores, Á., Verbraken, T., Baesens, B., & Weber, R. (2015). Profit-based feature selection using support vector machines - General framework and an application for customer retention. *Applied Soft Computing Journal*, 35, 240–248. <https://doi.org/10.1016/j.asoc.2015.05.058>
- Mantas, C. J., & Abellón, J. (2014). Credal-C4.5: Decision tree based on imprecise probabilities to classify noisy data. *Expert Systems with Applications*, 41(10), 4625–4637. <https://doi.org/10.1016/j.eswa.2014.01.017>
- Megahed, A., Yin, P., & Nezhad, H. R. M. (2016). An optimization approach to services sales forecasting in a multi-staged sales pipeline. *Proceedings - 2016 IEEE International Conference on Services Computing, SCC 2016*, 713–719. <https://doi.org/10.1109/SCC.2016.98>
- Mester, G. (2014). Design of the Fuzzy Control Systems Based on Genetic Algorithm for Intelligent Robots. *Interdisciplinary Description of Complex Systems*, 12(3), 245–254. <https://doi.org/10.7906/indecs.12.3.4>
- Moedjiono, S., Isak, Y. R., & Kusdaryono, A. (2016). Customer Loyalty Prediction In Multimedia Service Provider Company With K-Means Segmentation And C4 . 5 Algorithm, (Icic).
- Mori, T. (2016). Superposed naive bayes for accurate and interpretable prediction. *Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015*, 1228–1233. <https://doi.org/10.1109/ICMLA.2015.147>
- Padilla, W. R., Jesús, G. H., & Molina, J. M. (2018). Model learning and spatial data fusion for predicting sales in local agricultural markets. In *2018 21st International Conf. on Information Fusion (FUSION)* (pp. 1-5). IEEE.
- Pai, P. F., & Liu, C. H. (2018). Predicting Vehicle Sales by Sentiment Analysis of Twitter Data and Stock Market Values. *IEEE Access*, 6, 57655-57662.

Patel, A. K. (2017). Development of Online Machine Vision System using Support Vector Regression (SVR) Algorithm for Grade Prediction of Iron Ores, 149–152.

Paul, D., Su, R., Romain, M., Sébastien, V., Pierre, V., & Isabelle, G. (2016). Feature selection for outcome prediction in oesophageal cancer using genetic algorithm and random forest classifier. *Computerized Medical Imaging and Graphics*. <https://doi.org/10.1016/j.compmedimag.2016.12.002>

Rahim, M. S., Chowdhury, A. E., Islam, M. A., & Islam, M. R. (2017). Mining trailers data from youtube for predicting gross income of movies. In *Humanitarian Technology Conference (R10-HTC), 2017 IEEE Region 10* (pp. 551-554). IEEE.

Ramos, P., Santos, N., & Rebelo, R. (2015). Performance of state space and ARIMA models for consumer retail sales forecasting. *Robotics and Computer-Integrated Manufacturing*, 34, 151–163. <https://doi.org/10.1016/j.rcim.2014.12.015>

Siregar, B., Nababan, E. B., Yap, A., & Andayani, U. (2017). Forecasting of raw material needed for plastic products based in income data using ARIMA method. In *Electrical, Electronics and Information Engineering (ICEEIE), 2017 5th International Conference on* (pp. 135-139). IEEE.

Somwanshi, H., & Ganjewar, P. (2018). Real-Time Dengue Prediction Using Naive Bayes Predictor in the IoT. In *2018 International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 725-728). IEEE.

Sugiarto, V. C., Sarno, R., & Sunaryono, D. (2016). Sales forecasting using Holt-Winters in Enterprise Resource Planning at sales and distribution module. In *Information & Communication Technology and Systems (ICTS), 2016 International Conference on* (pp. 8-13). IEEE.

Tostes, A. I. J., Silva, T. H., Assuncao, R., Duarte-Figueiredo, F. L. P., & Loureiro, A. A. F. (2016). STRIP: A Short-Term Traffic Jam Prediction Based on Logistic Regression. *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, 1–5. <https://doi.org/10.1109/VTCFall.2016.7881239>

Umbarkar, A. J., & Sheth, P. D. (2015). Crossover Operators in Genetic Algorithms : A Review, 6956(October), 1083–1092. <https://doi.org/10.21917/ijsc.2015.0150>

Usman, S., & Subject, A. (2016). Risk Prediction Of Having Increased Arterial Stiffness Among Diabetic Patients Using Logistic Regression, 699–701.

Walia, N. K., Kalra, P., & Mehrotra, D. (2016). Prediction of carbon stock available in forest using naive bayes approach. *Proceedings - 2016 2nd International Conference on Computational Intelligence and Communication Technology, CICT 2016*, 275–279. <https://doi.org/10.1109/CICT.2016.61>

Wu, L., & Li, M. (2018). Applying the CG-logistic Regression Method to Predict the Customer Churn Problem. In *2018 5th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS)* (pp. 1-5). IEEE.

Wu, R., Painumkal, J. T., Volk, J. M., Liu, S., Louis, S. J., Tyler, S., ... Harris, F. C. (2017). Parameter estimation of nonlinear nitrate prediction model using genetic algorithm. *Evolutionary Computation (CEC), 2017 IEEE Congress On*, (3), 1893–1899.

Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2017). Risk prediction of type II diabetes based on random forest model. *3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics*, 382–386.

Yang, H., & Hu Man Zhou Li, X. (2016). Wavelet neural network with improved genetic algorithm for traffic flow time series prediction. *Optik*, 127, 8103–8110. <https://doi.org/10.1016/j.ijleo.2016.06.017>

Yang, J., Zhao, H., & Chen, X. (2016). Genetic Algorithm Optimized Training for Neural Network Spectrum Prediction, 2949–2954.

Zhang, Z., & Zhu, W. (2017). Location and Motion Prediction of Consumers in a Large Shopping Mall. In *Advanced Cloud and Big Data (CBD), 2017 Fifth International Conference on* (pp. 250-255). IEEE.

Zhang, H., Zhao, J., Yong, X., Zhang, C., & Ji, Y. (2017). Genetic Algorithm Based Approaches for Medium-thick Plate Stress Analysis Feature Extraction and Product Defect Prediction, *ᵀ(2)*, 6366–6369.

Zheng, Z. R., Liu, Z. C., Liu, X. Y., & Du, P. (2009). Genetic algorithm-based image preprocessing for volume rendering optimization. *ITME2009 - Proceedings 2009 IEEE International Symposium on IT in Medicine and Education*, 389–393. <https://doi.org/10.1109/ITIME.2009.5236394>

Žliobaite, I., Bakker, J., & Pechenizkiy, M. (2012). Beating the baseline prediction in food sales: How intelligent an intelligent predictor is? *Expert Systems with Applications*, 39(1), 806–815.

<https://doi.org/10.1016/j.eswa.2011.07.078>