AdaBoost integration with Genetic Algorithm for Psychological Aptitude Result Interpretation Model

1st Bayu Hanif Pratama
Department of Information Technology
Politeknik Caltex Riau
Pekanbaru, Indonesia
bayu20s2tk@mahasiswa.pcr.ac.id

2nd Dadang Syarif Sihabudin Sahid
Department of Information Technology
Politeknik Caltex Riau
Pekanbaru, Indonesia
dadang@pcr.ac.id

***Abstract*—SMA Darma Yudha Pekanbaru has a special program to facilitate students in knowing their interests and talents. The process is carried out by conducting special tests that can only be carried out by a certified psychologist. The psychologist is an educational worker under the division of guidance and counseling. The results of the test will be interpreted by psychologists with an output in the form of interests that can be used to determine the selection of majors in further study in college. Moreover, the results of the interpretation also contain recommendations of interest and talent for a career after the learner graduated from college. However, the recommendations of interest and talent are the result of unilateral analysis by psychologists without a supporting device to confirm the truth and accuracy. In this case the author will conduct research in the form of analysis of the results of interpretation issued by the psychologist of course with a variety of assessment instruments and not only that, to further ensure the results of interpretation, the author conducts analysis using 2 (two) machine learning methods and will then be done in order to get the best results. In this study, the authors used machine learning by comparing the results of analysis from 2 (two) methods namely Naïve Bayes and Decision Trees Classifier which then the classification results will be improved with AdaBoost.**

Keywords—Machine Learning, Supervised Learning, Classification, AdaBoost, Genetic Algorithm.

I. INTRODUCTION

Aptitude assessment and intervention plays sustainable and distinguished roles for every type and phase of evaluation for kids and teenagers concerning to their learning and behaviour issues. Aptitude, intelligence, and achievement as psychological constructs or types of tests are not easily distinguished It is simply distinguished that an achievement test describes people's present status, and aptitude test predict their future behaviour and ability tests asses their innate potential. Both aptitude and intelligence are enduring traits of individual, not easily modified by experiences and special training. In some cases both aptitude and intelligence-tests results were regarded as indications of innate capacity [1].

A psychological testing is a series of for an individual's different abilities, such as their aptitude in a particular field, cognitive functions like memory and spatial recognition, or even traits. These tests are based on scientifically tested psychological theories. The test is administered by the school and the result is interpreted by a licensed and nationally certified school psychologist, with a specialization in multicultural school psychology.

This research aims to develop a model that has been built on previous research namely Ada Boost – Genetic Algorithm [2]. Development focuses on optimizing the process of split datasets into data trains and test data. Optimization uses two algorithms that have the best performance in the classification, namely naïve bayes and decision trees. The accuracy of the predictions of the two algorithms will be compared to determine the quality of the model being worked on.

II. RELATED WORK

Ahmed Sharaf Elden et al [2], implement and test The Ada/ GA algorithm on ASSISTments dataset. The results showed that this algorithm has improved the detection accuracy as well as it reduces the complexity of computation. The accuracy value of the proposed algorithm prediction is 82.07% which is not much different from the application of a single AdaBoost algorithm, it is 81.85%.

Achmad bisri dan Rinna Rachmatika [3], use a sampling technique, SMOTE (synthetic Minority Over-Sampling Technique) and bagging technique as an ensemble in the Gradient Boosted Trees (GBT) classification method for handling the class imbalance problem. The proposed method is able to provide significant results with an accuracy of 80.57% and an AUC of 0.858, in the category of good classification.

Saifudin [4], A research on the selection of prospective new students by proposing several data mining methods and the one with the best value is the Support Vector Machine (SVM) method with an accuracy of 65%. Model testing with a 10-fold cross validation technique is implemented in this work. With this validation technique, the split process between training data and testing data should be a much better way. However, the accuracy of the predictions of each algorithm is not more than 65%.

Al-Radaideh et al [5], proposed a decision tree model with three different classification methods (ID3, C4.5, and nave Bayes) which allows students to predict the final grade in a course under study. From the test results they found that the decision tree model is the best of several models ever.

Dekker et al [6], presented a case study to predict students' drop out after the first semester of their studies or even before they enter the study program as well as identifying success-factors specific to the EE program by demonstrating the effectiveness of several classification techniques and cost-sensitive datasets. The experimental results found that using simple classification (J48, CART) gave satisfactory results compared to other algorithms such as Bayes Net or JRip.

Kalles and Perrakeas [7], studied the performance of different machine learning techniques (decision tree, neural network, nave Bayes, instance-based learning, logistic regression and support vector machine). By comparing those techniques with genetic algorithm based on decision tree induction, they can analyze the students' academic performance with students' homework as the paramete. They also got short rules to explain and predict success/failure on students' exams.

Based on the literacy review conducted, we found 2 (two) algorithms that will be developed with previous research, namely ada-ga model by focusing on optimizing the training process and testing datasets. These algorithms are naïve bayes and decision trees classifiers.

III. RESEARCH METHODOLOGY

This research methodology systematically defined the experiment models with research stages from data collection method, data processing, proposed models, model experiment, and evaluation and result validation.

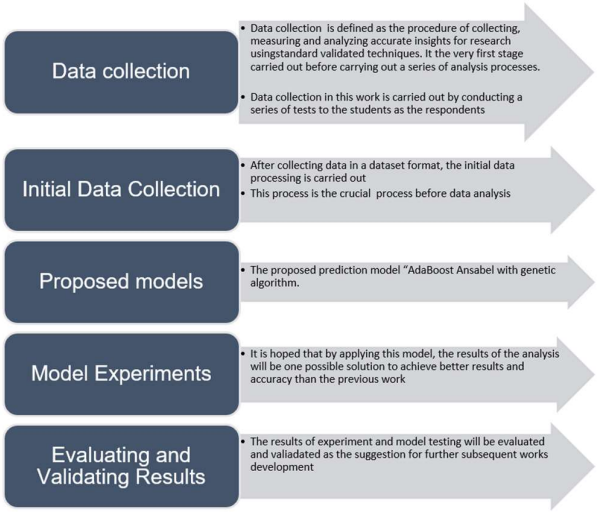

Fig. 1. Flowchart of Methodology Research

Figure 1 describes the stages of work. Starting from the stage of data collection, initial data processing, applying the proposed model, testing and finally is the process of evaluation and validation.

IV. THE PROPOSED MODEL

The objectives of this paper are framed to analyse and predict of the psychological test results by constructing a predictive model using AdaBoost algorithm and genetic algorithm, then validating developed model with original datasets and reliable sources. The next subsections will describe the AdaBoost algorithm, the genetic algorithm and a brief description of the proposed algorithm.

The researcher in this field decided to use the classification techniques, decision tree (DT) dan naïve bayes (NB) due to some advantages they may have over traditional statistical models. Mostly, DT has advantages over traditional statistics on two issues: Primarily, they can handle a large number of predictor variables, far more than the traditional statistics. Moreover, the DM techniques are non-parametric and capture nonlinear relationships and complex interactions between predictors and dependent variable [8].

NB is a classification method with a simple probabilistic-based prediction technique which refers to Bayes' theory by using strong assumptions (naives) and based on probability functions for each instance in mapping attribute classification on a stable efficiency and low complexity system [9][10]

A. *Boosting and AdaBoost*

Boosting is a common machine learning algorithm that increases accuracy of Learning algorithm. It is a widely used and powerful prediction technique due to sequentially builds an ensemble of weak classifiers. In boosting, a weak classifier is a model for binary classification that performs slightly better than random guessing. Formally, a weak classifier achieves slightly better than 50 percent accuracy on the training data. Weak classifier sets are built repeatedly from training data more than thousands of iterations. At each iteration, the training data are re-weighed on how good they are classified (greater weight is given to the classification error sample). Weights are calculated for weak classifiers based on their classification accuracy. The weighted predictions of the weak classifiers are combined by voting to calculate the final outcome prediction [11].

AdaBoost is the most common optimization algorithm for binary classification proposed by Freund and Schapire [12]. It takes as input a training set "S" of "m" sample ($S = \{(x_1,y_1),\ldots,(x_m,y_m)\}$), where each instance of x_i is a vector of attribute values that belongs to a domain or instance space X, and each label y_i is the class label associated with xi that belongs to a finite label space $Y=\{-1, +1\}$ for binary classification problems.

The following figure is a general illustration of the AdaBoost algorithm for binary classification problem.

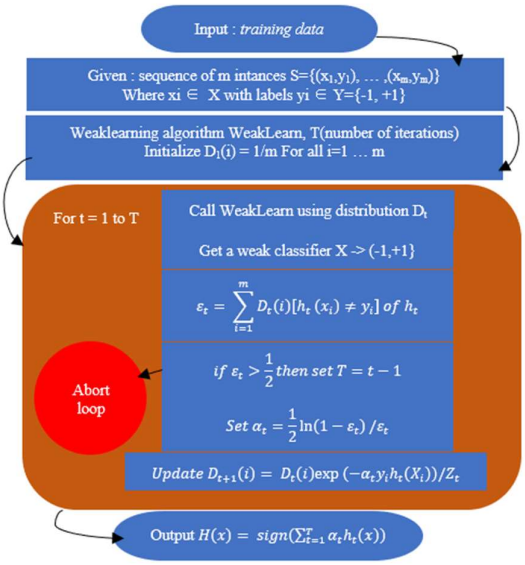

Fig. 2. A generalized version of the AdaBoost Algorithm

Ada Boost weights the training sample with a probability distribution of $D_t(x)$ in each iteration. WeakLearn algorithm is applied to generate ht classifier with error rate et on training sample. The effect of changing the weights is placing more error classifiers in the final stage. This process continues during the T round until the of the final classifier H, is constructed by weighting the weak classifier $h_1, h_2, \ldots, h_T$. Each classifier is weighted according to its accuracy of the distribution Dt that it was trained on [12]. The weak classifier is the core of an AdaBoost algorithm. In this work, classification and regression tree (CART) algorithm, proposed by Breiman et al. [13], was used as WeakLearn to AdaBoost algorithm.

B. Genetic Algorithm

Genetic algorithm (GA)is an evolutionary based stochastic optimization algorithm with a global search potential proposed by Holland (1973) [14]. GA is among the most successful class of algorithms under EAs(Evolutionary Algorithms) which are inspired by the evolutionary ideas of natural selection. Because of its outstanding performance with optimization, GA has been regarded as a function optimizer.

The algorithm starts by initializing the solution population (chromosomes) and comprises representation of the problem usually in the form of a bit vector. Chromosomes evolve through successive iterations called generations. During each generation, the chromosomes are evaluated, using some measures of fitness (using an appropriate fitness function suitable for the problem). To create the next generation, new chromosomes, called offspring, are formed by combining two chromosomes from the current generation using the crossover operator or modifying the chromosomes using the mutation operator. A new generation is formed by voting; fitter chromosomes have a higher probability of being selected. After several generations, the algorithm encounters the best chromosome, which hopefully represents an optimal or suboptimal solution to the problem. The three main genetic operators in GA involve selection, crossover, and mutation[15].

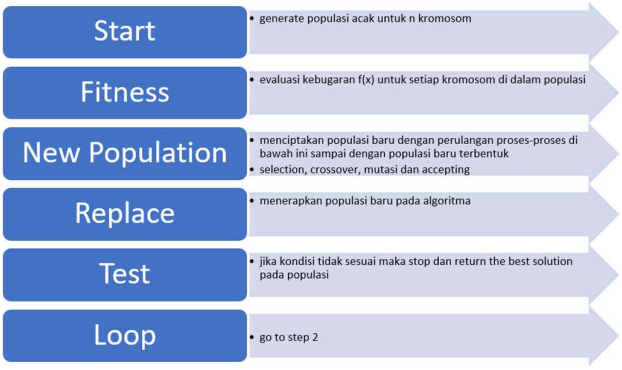

Fig. 3. The outline of the GA algorithm

C. Overview of the Proposed Model: AdaBoost-GA

Freund dan Schapire [12] concluded that the AdaBoost algorithm is less prone to overfitting problems compared to most learning algorithms, because it increases sensitivity to data and noise outliers. Thus, mislabelled cases or outliers may cause the overfitting problems, for the new classifier to focus more on those observations that have been misclassified, resulting in a large number of weak classifiers to achieve better performance [11].

In this study, a new boosting algorithm called "Ada-GA" which strengthen the classification of the decision tree and nave Bayes algorithms. In the previous work, applied another classification algorithm after the output was corroborated by AdaBoost – GA [2].

Input:a set S of m instances: $S=\{(x_1,y_1), \ldots, (x_m,y_m)\}$ where $x_i \in X$ with labels $y_i \in Y=\{-1, +1\}$, P (population size), G (maximum number of generations), T(initial number of weak learners).

Initialize: a randomly generated population of m solutions (consists of a set of Tweak learners withtheir weightsproduced by AdaBoost)

Evolve: fork = 1, 2, …, G.
1-Generate a population of bit strings b
2- Evaluate the fitness of the solutions:$f(b)=w_1 * (1-L/T)+w_2 * (1/E_b^s)$
where b isthe evolved best individual, w_1, w_2 are fitness weights ,$E_b^s$is the validation error and $L=\sum_{i=1}^{T} b_i$
3-Use vector b to update the weak classifiers (T)and their weights.
4-Produce new generation of solutions using the genetic operations (selection, mutation and crossover).
5- If end condition, stop, and return the best solution; else loop

Output:final hypothesis (with optimized classifiers and their weights).

Fig. 4. The Proposed Procedure of Ada-GA

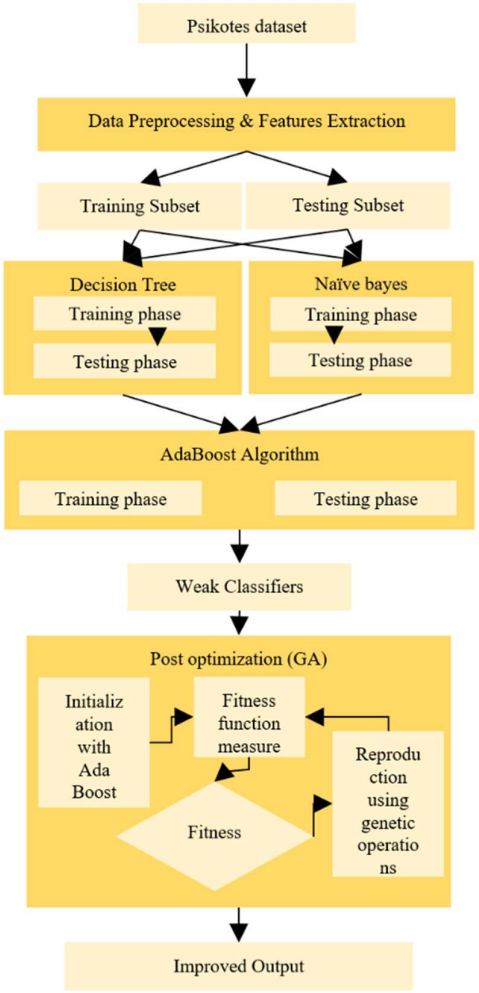

Fig. 5. The Structure of the Proposed Model Ada-GA

The structure of "Ada-GA" is detailed in Figure 4 which consists of three following phases:

1. Pre-processing and feature extraction phase: This phase is specifically to split the dataset into training and testing randomly.
2. Pre-processing and feature extraction phase: This phase is specifically to split the dataset into training and testing randomly.
3. The difference between this work and the previous work is that there is a combination of higher values two algorithms, Decision Tree and Naïve Bayes, and a comparative analysis of both models.
4. Post optimization procedure phase: this phase is composed of three parts:
   a. Initialization with AdaBoost
   b. Fitness function, and
   c. GA

V. CONCLUSIONS

Modeling for optimization in the training and testing process carried out in this study can produce a higher score than in the previous work of 82.07%. We are very optimistic because based on the reference classification algorithm that has always been at the top in terms of performance are decision trees and naïve bayes.

REFERENCES

[1] G. Goldstein and M. Hersen, *Handbook of Psychological Assessment 3rd Edition*, 3rd ed. Pergamon, 2000.

[2] A. ElDen, M. A. Moustafa, H. M. Harb, and A. H.Emara, "Adaboost Ensemble with Simple Genetic Algorithm for Student Prediction Model," *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 2, pp. 73–85, 2013, doi: 10.5121/ijcsit.2013.5207.

[3] A. Bisri and R. Rachmatika, "PREDIKSI KELULUSAN MAHASISWA MENGGUNAKAN TEKNIK MACHINE LEARNING PADA LEVEL DATA UNTUK MENANGANI KETIDAKSEIMBANGAN KELAS," *Direktorat Ris. dan Pengabdi. Masy. Direktorat Jenderal Penguatan Ris. dan Pengemb. Kementrian Riset, Teknol. dan Pendidik. Tinggi*, 2019.

[4] A. Saifudin, "Metode Data Mining Untuk Seleksi Calon Mahasiswa," vol. 10, no. 1, pp. 25–36, 2018.

[5] Q. A. Al-Radaideh and E. Al-Shawakfa, "Mining Student Data Using Decision Trees," *Expert Syst. Appl.*, vol. 40, no. 2, pp. 1–18, 2015, [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2008.02.021%0Ahttp://mechanicaldesign.asmedigitalcollection.asme.org/article.aspx?doi=10.1115/1.4026094%0Awww.wiley.com%0Ahttp://www.sciencedirect.com/science/article/pii/S22125671150.

[6] G. W. Dekker, M. Pechenizkiy, and J. M. Vleeshouwers, "Predicting students drop out: A case study," *EDM'09 - Educ. Data Min. 2009 2nd Int. Conf. Educ. Data Min.*, no. January, pp. 41–50, 2009.

[7] D. Kalles and C. Pierrakeas, *Analyzing student performance in distance learning with genetic algorithms and decision trees*, vol. 20, no. 8. 2006.

[8] Z. J. Kovacic, "Early Prediction of Student Success: Mining Students Enrolment Data," *Proc. 2010 InSITE Conf.*, pp. 647–665, 2010, doi: 10.28945/1281.

[9] T. D. Salma and Y. S. Nugroho, "Sistem Rekomendasi Pemilihan Sekolah Menengah Tingkat Atas Menggunakan Metode Naive Bayes," *Khazanah Inform. J. Ilmu Komput. dan Inform.*, vol. 2, no. 2, p. 85, 2016, doi: 10.23917/khif.v2i2.2306.

[10] N. Ailmi, Z. Saharuna, E. Tungadi, and A. M. Learning, "Metode Klasifikasi Pada Aplikasi Pendukung Keputusan Untuk Pemilihan Unit Kegiatan Mahasiswa," *Pros. Semin. Nas. Tek. Elektro dan Inform. 2020*, pp. 142–147, 2020.

[11] DONG-YOP OH, "GA-BOOST: A GENETIC ALGORITHM FOR ROBUST BOOSTING," 2012.

[12] Y. Freund, R. E. Schapire, and M. Hill, "Experiments with a New Boosting Algorithm Rooms f 2B-428 , 2A-424 g," 1996.

[13] L. Breiman, J. H. Friedman, R. A. Olshen, Charles, and J. Stone., *Classification and Regression Trees*, vol. 10, no. 4. Brooks/Cole Publishing, 1984.

[14] J. H. Holland, "The Optimal Allocation of Trials," *Adapt. Nat. Artif. Syst.*, vol. 2, no. 2, pp. 88–105, 2019, doi: 10.7551/mitpress/1090.003.0008.

[15] B. S and S. S. Sathy, "A Survey of Bio inspired Optimization Algorithms," *Int. J. Soft Comput. Eng.*, vol. 2, no. 7, pp. 2323–2339, 2012, doi: 10.1007/s11269-015-0943-9.