Machine Learning Technique Analysis and Applications for Predicting Student Performance

Evaristus Didik Madyatmadja, Albert Jonathan Susanto

Abstract*: *Machine Learning is an emerging research field concerned with developing methods to answer uncommon problems. There are many problems that can be answered with Machine Learning method, one of them is on educational scope. Many Educators right now cannot identify whether a certain student is on the brink of failing or not. As a result, many college students failed because the educators cannot help them. In this paper, we present our user-friendly decision support tool made from Machine Learning algorithm and to answer the problem we focus, which is to prevent college student from failing by providing educational agents necessary information and predictions. Our objective is to know which machine learning algorithm that can be used to predict the student's performance and to create a decision support tool that can be used by educational agents so that educational agents can prevent student from failing the course.

Keywords* : *Classifier, Dataset, Data Training, Model

I. INTRODUCTION

Lately, Business Intelligence (BI) topics has risen in IT Executives league and the Business Intelligence software product market continued to grow rapidly, even though the macro-economic poses a great challenge to the market itself [1]. There are many BI-related topics that contributed to the BI-Software market growth, such as "Big Data".

Despite the increased demands for Business Intelligence product is [2], the wider academic research community has just known the topic, and until today Business Intelligence is still fragmented and ambiguous. There are several differences between contemporary Business Intelligence and the early form of Decision Support System: First, they typically involve systematic integration, aggregation, structured and unstructured data management in pseudo 'real time' data warehouses, which enable a new form of fact-based Decision Support System (DSS) [4]. Second, BI solutions on this era deals with very big and

increasing amount of data ('Big Data') and can rely on more-modern processing capabilities and technology, which can create a new opportunities of knowledge discovery (e.g. data mining). Third, Business Intelligence grows from newly discovered ways of data interrogation and information delivery.During the last decade, data mining technique application become a very popular topics, enabling the development of efficient and accurate models that can predict student's academic performance. The researcher has spent so much time for developing an efficient and accurate prediction model for predicting the students' future academic performance based on a classifier. There are so much dedication from the researchers, but the development of such prediction model is a very challenging task, given that the technology to do such thing is very limited [5,6,7,10]. The reason is that the dataset from this domain class distribution are usually located in one class on most cases [8].

Many of the business decision in this era is based on the information and predictions generated by system. This system, known as Decision Support System (DSS), is becoming a critical system in the business and educations because of its abilities to predict the outcome of a decision and provide accurate information based on provided data. The ability to predict the outcome based on the available data came from Machine Learning Classifier algorithm. Machine Learning is a branch of Artificial Intelligence (AI) that enables machines to perform and learn from their jobs by using intelligent software and structured algorithm [9]. Statistics plays a very big role in Machine Learning algorithm because it defines every Machine Learning Classifier algorithm. Because Machine Learning algorithm requires data, therefore each Machine Learning algorithm produces different results on each case.In this research, we present the analysis, design, and implementation of the Decision Support tool for predicting student's performance in Algorithm and Programming course. The purpose of this research is to know which machine learning algorithm that can be used to predict the student's performance and to create a decision support tool that can be used by educational agents so that educational agents can prevent student from failing the course.

II. LITERATURE REVIEW

2.1. Decision Support System

Decision Support System (DSS) is a computer software system that helps decision maker to solve existing abnormal problems by using previous data provided by them.

A Decision Support System is a computer-based system that helps decision makers decide semi-structured and unstructured problems through direct interaction.

Revised Manuscript Received on October 15, 2019.

Dr. Evaristus Didik Madyatmadja, Associate Professor**,** Information Systems Department, School of Information Systems, Bina Nusantara University, Jakarta. Email: emadyatmadja@binus.edu

Albert Jonathan Susanto, Research Assistant**,** Information Systems Department, School of Information Systems, Bina Nusantara University.

[12] There are several difference between Decision Support System and Management Information system, since Decision Support System is created to emphasis these issues: First, to provide useful information to higher-level management to support relatively unstructured decision making activities. Second, to fuse models into the information system software. Lastly, to provide the system's user with powerful but simple languages for problem solving [13].

2.2. Supervised Machine Learning Techniques

Supervised machine learning is a special case of data mining that concerns the process of predicting unknown attribute values from a given set of known attributes values [14]. For this purpose, many techniques and algorithms have been developed based on artificial intelligence and statistics. In the rest of this section, we present the popular classes of classification algorithms, which include Linear Regression, K-Nearest Neighbor (K-NN), Gaussian Naïve Bayes, and Support Vector Machines.

A Bayesian network is a combination of direct acyclic graphs of vertices and links and a set of conditional probability tables. [15]. Each node in the graph is associated with a feature whereby the links between nodes represent the relationships between them and the strength of the links is determined by conditional probability tables. More analytically, each node in the network has an associated probability table that describes the conditional probability distribution of that node given its parents nodes. If a node has one or more parents the probability distribution is a conditional distribution, where the probability of each attribute depends on the values of the parents while in case a node has no parents the probability distribution is unconditional. Using a suitable training method, one can induce the structure of the Bayesian network from a given training set [15]. The classifier based on this network and on the given set of attributes X1, X2,…,Xn returns the label c that maximizes the posterior probability p(c|X1, X2,...,Xn).

The K-Nearest Neighbors (KNN) are non-parametric, lazy learning supervised learning method used for classification and regression. The basic theory behind K-Nearest Neighbors is that in the calibration dataset, it finds a group of k- amount of samples that are nearest to unknown samples (e.g., based on distance functions) [16]. From these k-amount of samples, the label (class) of unknown samples are determined by calculating the average of the response variables (i.e., the class attributes of the k-nearest neighbors) [17].

The Support Vector Machines (SVM) are a group of supervised learning methods established as part of the most precise discriminatory methods used in classification. They represent an extension to nonlinear models of the generalized portrait algorithm of Vapnik [18] which is based on structural risk minimization, an inductive principle of use in machine learning. However, in most real-world problems there exists no such hyperplane that successfully separates the instances in the training set since they involve non-separable data.

III. RESULT

2.3. Methodology

The aim of this study is to find the appropriate classifier and develop Decision Support System for predicting the students" performance at the final examinations. For this purpose, we have adopted the Agile Methodology that consists of several stages as illustrated in Figure 1:

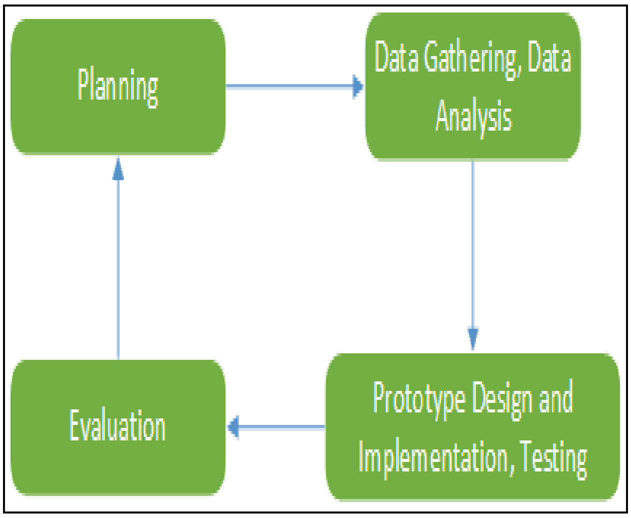

Planning
Data Gathering, Data Analysis
Prototype Design and Implementation, Testing
Evaluation

Fig. 1. Agile Methodology diagram

There are 4 stages in this development cycle:

1. Planning Stage
   In this stage, we plan how to collect the data, and how to create the prototype.
2. Data Gathering and Data Analysis Stage
   In this stage, we gather the necessary data, which is grade data from 1st semester Information Technology student who took Algorithm and Programming course. After gather the data, we analyze the data to choose the appropriate classifier and then test the selected classifier to select the classifier that has the highest accuracy score.
3. Prototype Design and Implementation Stage
   After we decided which classifier we used, we generate the model by using the chosen classifier and then develop the Decision Support tool.
4. Evaluation Stage
   In this stage, we evaluate the prototype that we designed to find feature that may be vital to the prototype.

2.4. Datasets

The data used in our study concerns about the student's performance in Algorithm and Programming of the first year of Information Technology department that is students of ages 17-19 years. The data have been collected by the university during 2018-2019 and consists of 279 patterns. The attributes concern information about the students' Online Quiz performance and has value between 0 and 100, in accordance with standard Indonesian education system.

Furthermore, the student score data were classified into two classification: Passed or Not Passed that has been visualized by Figure 2.

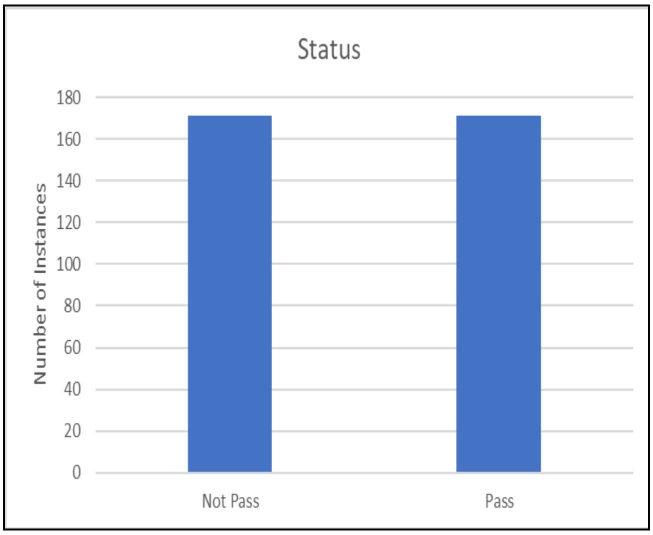


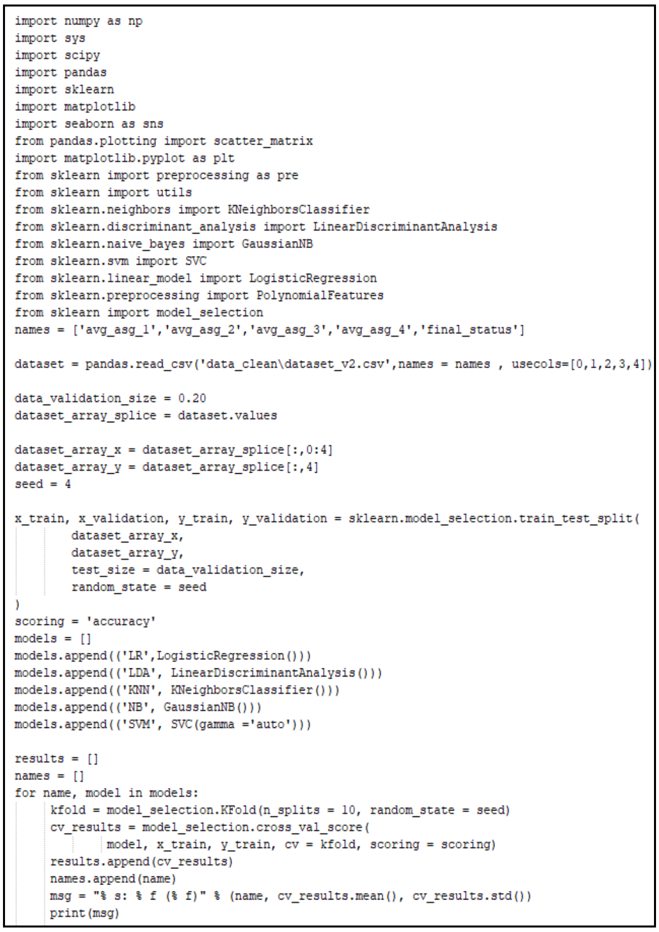
Fig. 2. Class Distribution

2.5. Evaluation/Experimental result

Next, we conduct a series of tests in order to establish which learning algorithm predicts the class "Pass" or "Not Pass" in which a student belongs based on its grades on Algorithm and Programming course. Thus, we have selected the popular and frequently used classifier for each described machine learning technique.

The first classifier that we used came from Naïve Bayes algorithm family, the Gaussian Naïve Bayes. Gaussian Naïve Bayes is a simple learning classifier that captures the assumption that every attribute is independent from the rest of the attributes, given the state of the class attribute. Since we made assumption that each class is distributed to a normal distribution (Gaussian distribution), we decided to use Gaussian Naïve Bayes as our selected classifier. Logistic Regression is one of regression analysis and used to model a binary dependent variable, in this case are "Pass" or "Not Pass". The Linear Discriminant Analysis is quite similar to Logistic Regression algorithm but enables to model many classes. The K-Nearest Neighbours is a pattern recognition classifier. Support-Vector Machine is an associated learning classifier that can analyze data for classification and regression. In order to minimize the effect of any expert bias by not attempting to tune any of the algorithms to the specific datasets we have utilized the default values of all learning parameters.

To test each selected classifier, we used Scikit-learn from Python Library as shown in Figure 3. First, we split the dataset into 2: the training dataset and the testing dataset. The Training dataset is used to create a learning model, and then the Testing dataset is used to measure the accuracy rate of each algorithm based on Training dataset. To split the dataset, we used K-Fold Cross Validation Technique. K-Fold Cross-Validation technique is a statistical method to evaluate learning algorithms or classifiers by splitting the data into k-size segments or folds. These segments will be used to train and test several learning algorithms or classifiers with each classifier have their own segments of data [19]. After split the dataset, we input the selected classifier into an array variable, and then we begin test each selected classifier. We use several classifiers that has been stored into the array ('models' variable) and then began train the data according the classifier so that we don't waste much time train the data one-by-one with several classifier. And then we print the result of our classifier testing.



```
import numpy as np
import sys
import scipy
import pandas
import sklearn
import matplotlib
import seaborn as sns
from pandas.plotting import scatter_matrix
import matplotlib.pyplot as plt
from sklearn import preprocessing as pre
from sklearn import utils
from sklearn.neighbors import KNeighborsClassifier
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.naive_bayes import GaussianNB
from sklearn.svm import SVC
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn import model_selection
names = ['avg_asg_1','avg_asg_2','avg_asg_3','avg_asg_4','final_status']

dataset = pandas.read_csv('data_clean\dataset_v2.csv',names = names , usecols=[0,1,2,3,4])

data_validation_size = 0.20
dataset_array_splice = dataset.values

dataset_array_x = dataset_array_splice[:,0:4]
dataset_array_y = dataset_array_splice[:,4]
seed = 4

x_train, x_validation, y_train, y_validation = sklearn.model_selection.train_test_split(
        dataset_array_x,
        dataset_array_y,
        test_size = data_validation_size,
        random_state = seed
)
scoring = 'accuracy'
models = []
models.append(('LR',LogisticRegression()))
models.append(('LDA', LinearDiscriminantAnalysis()))
models.append(('KNN', KNeighborsClassifier()))
models.append(('NB', GaussianNB()))
models.append(('SVM', SVC(gamma ='auto')))

results = []
names = []
for name, model in models:
     kfold = model_selection.KFold(n_splits = 10, random_state = seed)
     cv_results = model_selection.cross_val_score(
             model, x_train, y_train, cv = kfold, scoring = scoring)
     results.append(cv_results)
     names.append(name)
     msg = "% s: % f (% f)" % (name, cv_results.mean(), cv_results.std())
     print(msg)
```

Fig. 3. Code to Test Each Selected Algorithm

Table 1 summarizes the performance of each classifier, measured by the percentage of patterns that were classified correctly in the presented datasets.

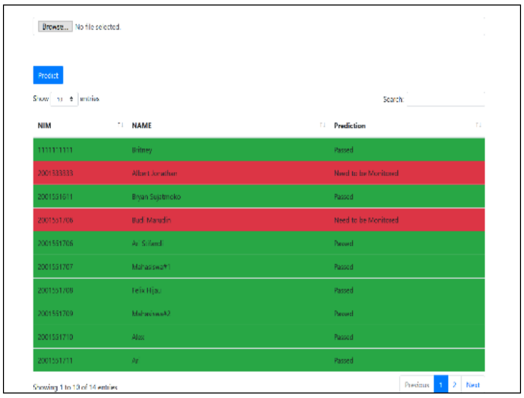
Table-I: Classifier's accuracy

Algorithms	Accuracy
Logistic Regression	92.9%
Linear Discriminant Analysis	94.1%
K-Nearest Neighbour	96%
Gaussian Naïve Bayes	93.7%
Support-Vector Machine	85.7%

Based on accuracy score from each classifier, K-Nearest Neighbour has the highest accuracy, scoring 96% accuracy from provided dataset.

2.6. Decision Support Tool

In this section, we present a prototype version of our software support tool for predicting the students' performance at Algorithm and Programming course (Figure 3). The tool has been developed using HTML, Flask library and Scikit-learn Python library. This tool uses our generated model to help the application predict the future result from input. To generate the model, we train the data with our chosen classifier, which is K-Nearest Neighbor and then save the outcome of data training into the model file. Flask Python Library is used to host the website and become a bridge between model data and the application.



NIM	NAME	Prediction
1111111111	Britney	Passed
2001533333	Albert Jonathan	Need to be Monitored
2001551611	Bryan Sujatmoko	Passed
2001551706	Budi Marudin	Need to be Monitored
2001551706	Ari Sifandi	Passed
2001551707	Mahasiswa*1	Passed
2001551708	Felix Hijau	Passed
2001551709	Mahasiswa#2	Passed
2001551710	Alex	Passed
2001551711	Ar	Passed

Showing 1 to 10 of 14 entries

Fig. 3. The Prediction Tool

The main feature of our software tool is:

- Import student data: this module is used to import several student data such as Student ID, Student Name, and their score by using excel file (.xlsx).
- Prediction: this module is used to predict the course graduation on each listed student based on generated model.

IV. CONCLUSION

2.7. Conclusion

Prediction using Machine Learning and Data Mining techniques is a significant tools and act as a first step to help educators identify students who are likely to have poor score and performance in specific subjects. In this research we developed user-friendly decision support tool for predicting the students' performance, together with a case study concerning the Algorithm and Programming graduation rate of the first semester of Information Technology. Our proposed tool is based on K-Nearest Neighbor Machine Learning classifier. We choose K-Nearest Neighbor classifier because based on our test against Test dataset, K-Nearest Neighbor produces the highest accuracy value compare to 4 other chosen classifiers. To test it, we first split the data into several segments by using K-Fold Cross Validation and after that we train and test the data by using several data segment that has been distributed into our chosen classifier. After we test the classifiers, we then create a model and use it on proposed tool to predict the outcome of several student based on their previous quiz score. Furthermore, significant advantages of the presented tool are that it has an easy-to-use user interface and it can be deployed in online server. We have illustrated the main features of our software tool and we have also presented a case study to illustrate its functionalities and the experiment set up processes. Our preliminary results revealed that we can gain insights early about student progress and provide several possible actions such as further study or additional learning activities, resources and learning tasks. Furthermore, it is worth mentioning that the used attributes in our Decision Support Tool are not a conclusive list and can be changed according to necessity.

2.8. Research Suggestion

Currently, our prediction tool is still under development and given that this is a pilot study, our evaluators and testers (Lecturers and Educators) are rather small. Hence, in our plans is to do a systematic and extensive evaluation of the tool by several groups of potential users (in this case, Lecturers and Educators) to evaluate its usability. Moreover, since the data we used to test each classifier came from 1st semester Information Technology student at 2018/2019, another direction for future research would be to collect data from every subject on Information Technology course and apply our methodology to predicting the students' graduation rate on Information Technology course.

REFERENCES

1. Sommer, D, Sood, B. Market Share Analysis: Business Intelligence and Analytics Software,, Gartner Research Report, 2014.
2. Negash, S. Business Intelligence. Communications of AIS; 13. 177-195, 2003.
3. Kowalczyk, M, Buxmann, P, Besier, J. Investigating Business Intelligence and Analytics from a Decision Process Perspective: A Structured Literature Review. 21st European Conference on Information Systems, Utrecht, The Netherlands, 2013.
4. Baars, H, Kemper, HG. Management Support with Structured and Unstructured Data - an Integrated Business Intelligence Framework. Information Systems Management, 25:2. 132-148, 2008.
5. Baker, R. &Yacef, K.. The state of educational data mining : A review future visions. Journal of Educational Data Mining, 1(1), 3–17, 2009.
6. Romero, C. & Ventura, S.. Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, 33, 135–146, 2007.
7. Romero, C. & Ventura, S.. Educational data mining: A review of the state of the art. IEEE on Transactions on Systems, Man, and Cybernetics - Part C: Applications and Reviews, 40(6), 601–618, 2010.
8. Kotsiantis, S.. Use of machine learning techniques for educational proposes: a decision support system for forecasting students‟ grades. Artificial Intelligence Review, 37, 331–344, 2012.
9. Mohammed, Mohssen & Badruddin Khan, Muhammad & Bashier, Eihab.. Machine Learning: Algorithms and Applications, 2016.
10. Romero, C., Ventura, S., Pechenizkiy, S., & Baker, M. Handbook of Educational Data Mining. London: Chapman & Hall, 2010.
11. Kotsiantis, S. Use of machine learning techniques for educational proposes: a decision support system for forecasting students‟ grades. Artificial Intelligence Review, 37, 331–344, 2012.
12. Sprague, Ralph H., and Eric Carlson, Building Effective Decision Support Systems, Prentice Hall, Upper Saddle River, NJ, 1982.
13. Robert H. Bonczek, Clyde W. Holsapple, Andrew B. Whinston, Foundations of Decision Support Systems, Academic Press, 1981.
14. Mitchell, T. Machine Learning. USA: McGraw Hill, 1997.
15. Jensen, F. An Introduction to Bayesian Networks. Heidelberg: Springer-Verlag., 1996.
16. Thanh Noi, P.; Kappas, M. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. Sensors., 18, 18, 2018.
17. Akbulut, Y.; Sengur, A.; Guo, Y.; Smarandache, F NS-k-NN: Neutrosophic Set-Basedk-Nearest Neighborsclassifier.Symmetry, 9, 179, 2017.
18. Vapnik, V. The Nature of Statistical Learning Theory. Heidelberg: Springer-Verlag, 1995.
19. Refaeilzadeh P., Tang L., Liu H. Cross-Validation. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA, 2009.

AUTHORS PROFILE

Evaristus Didik Madyatmadja received the master degree Computer Science from Gadjah Mada University (UGM), Yogyakarta, Indonesia, in 2005. He received the Doctor of Computer Science from Bina Nusantara University, Jakarta, Indonesia, in 2019. Currently, he is a Associate Professor at Scchool of Information Systems, Bina Nusantara University, Jakarta, Indonesia. His interests are in e-government, decision support system, data mining and business intelligence.

Albert Jonathan Susanto is an undergraduate information system student at Bina Nusantara University. He currently also a junior researcher at Bina Nusantara University, School of Information Systems. His interest are in Database, Machine Learning, and Data Mining.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication