MODELING STUDENT GRADUATION IN PTN ENTRANCE SELECTION THROUGH REPORT CARD SCORES USING STOCHASTIC GRADIENT DESCENT CLASSIFICATION

Firman Matiinu Sigit[1], Rachmawati Findiana [2]

Universitas Tulungagung

Keywords:
Public University, Report Score, Stochastic Gradient Descent

***Correspondence Address:**
firman.matiinu@gmail.com

Abstract: There are many ways to enter state universities in Indonesia, namely by taking an entrance test selection or by selecting a report card. The simplest way is to take part in the section based on report card scores because students are not tired of preparing to take part in the selection. Still, in the Mojokerto area, guidance and counseling teachers do not know whether students' report card scores meet the requirements or do not meet the criteria to take part in the selection, and student selection is still being carried out. Manually after that, the teacher provides a letter of recommendation to take part in the section of students who are deemed to meet the criteria. In this experiment, we tried to build a prediction machine based on existing data, so it is hoped that this machine can provide an overview for consideration for guidance and counseling teachers to give recommendations to students who meet the criteria. From this experiment, a training set with 2764 data was studied using the Stochastic Gradient Descent Algorithm and gave a recall score = 0.178, precision = 0.394, and F1 score = 0.245.

INTRODUCTION

The number of upper secondary and equivalent students (SMA and MA) who continue to higher levels of education (tertiary education) is one indicator of the success of learning at the school. The more school graduates who continue to higher education, the better the quality of education at the school. Currently, the level of rigor for entrance selection at State Universities (PTN) is relatively high, based on data collected from the official SNMPTN website in 2019, from 478,608 participants who registered, 92,331 participants were accepted, so the ratio of accepted participants is 5:1. However, the level of rigor depends on the study program in the respective department and PTN being targeted.

Based on data collected through several official websites of institutions administering new student admissions to PTNs in Indonesia, this is done through:

1. National Selection to Enter PTN, which is carried out based on the results of achievement tracking academic/report marks and portfolio of prospective students without going through a written exam. So this route is known as the National

Higher Education Entrance Selection through the report card route.

2. Joint Selection for PTN Entrance, which is carried out based on the results of the Written Based Examination Computer (UTBK), as well as other criteria agreed upon by PTN
3. Independent Selection for PTN Entrance is a selection that depends on the material and examination procedures University/institution that holds.

In SMA and MA in the sub-district area of Mojokerto district, students who wish to take part in the selection to enter State Universities through the report card route are still recorded manually. The students come to the Guidance Counseling office, and then the guidance and counseling teacher provides a letter of recommendation. In the sub-districts in the Mojokerto district in 2018, it was recorded that 100 students out of 348 students were accepted at PTN through the report card route, while in 2019, as many as 71 students out of 350 students were accepted at PTN, so in the sub-district of Mojokerto district the number of students accepted at PTN is still relatively low, namely around 20% – 30% through this route [1].

Because of these problems, in this research, existing data, report card score data along with information 'accepted' or 'not accepted' were used to carry out predictive modeling, so it is hoped that the predictive modeling will be able to provide 'insight' to guidance and counseling teachers as material for consideration and recommendations, from the BK teacher to students who meet the selection criteria for entering PTN through the report card route..

RESEARCH METHODS

The stages in this research can be described through the following block diagram,

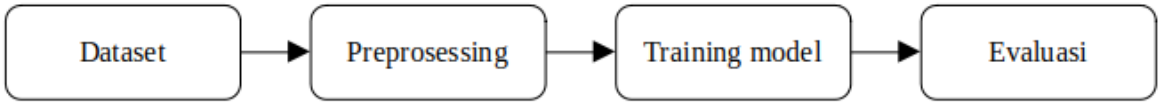

Figure 1. Research block diagram

A. Datasets

The Dataset used in this research is a dataset containing a number of report card scores for each student from State High Schools and MANs in sub-district areas in the Mojokerto district. The data taken are the grades of the last five semesters, namely data on the number of report card grades for each semester along with the average grade data, along with information about whether they were accepted at PTN 'PASSES' or not

accepted at PTN 'NOT PASSED.' Data received at PTN 'LULUS' is received via the SNMPTN, SNMPN, SPANPTKIN, and PMDKPN channels.

The Dataset used consists of 2764 rows with 13 columns consisting of (average 1 to 5, sum 1 to 5, and standard deviation as well as the classification target, namely 'ENTRY PATH'), while the 'PTN' column was removed in the preprocessing process This is to simplify the model with a classification target of only one category, namely 'ENTRY PATH.'

An overview of the Dataset (taken only five rows) can be seen in the table below:

	JUMLAH1	RATA21	JUMLAH2	RATA22	JUMLAH3	RATA23	JUMLAH4	RATA24	JUMLAH5	RATA25	JALUR MASUK	PTN	STD
2750	1284.0	80.250000	1311.0	81.937500	1267.0	84.466667	1291.0	86.066667	1312	87.466667	TIDAK LULUS	TIDAK LULUS	637.308283
1849	1229.0	81.933333	1232.0	82.133333	1165.0	83.214286	1186.0	84.714286	1204	86.000000	TIDAK LULUS	TIDAK LULUS	590.388630
651	1609.0	84.700000	1604.0	84.500000	1621.0	85.400000	1678.0	88.400000	1676	88.300000	TIDAK LULUS	TIDAK LULUS	817.991015
261	1156.0	82.600000	1182.0	84.466667	1221.0	86.866667	1227.0	87.333333	1257	89.666667	TIDAK LULUS	TIDAK LULUS	592.158775
1253	1261.0	84.066667	1256.0	83.733333	1223.0	87.357143	1217.0	86.928571	1243	88.785714	SNMPTN	UB	608.260510

Figure 2. Dataframe from Dataset (5 data taken)

What needs attention here is the classification target, namely 'ENTRY PATH' where the 'NOT PASS' category consists of 2300 data, then SNMPTN consists of 279 data, SNMPN consists of 93 data, SPANPTKIN consists of 81 data, and PMDKPN consists of 11 data, as can be seen in the table below:

Table 1. Number of each category in the 'ENTRY PATH' column

'ENTRY PATH' (category)	Amount
NOT PASS	2300
SNMPTN	279
SNMPN	93
SPAN PTKIN	81
PMDKPN	11

It can be seen in Table 1 above that some categories have a more significant amount of data than the other categories, namely the 'NOT PASS' category with 2300 data. A dataset like this is called an imbalanced dataset.

B. Preprocessing

The Dataset consisting of 2764 rows and 13 columns was split into 80 percent training set and 20 percent testing set (with appropriate proportions referring to the classification target, namely in the 'ENTRY PATH' column) so that we obtained 2211 rows of data for the training set and 553 rows of data for testing sets. The proportion of datasets in the training set refers to the ratio of data in the classification target in the 'ENTRY PATH' column in the primary Dataset, which can be seen in the table below:

'ENTRY PATH' (category)	Amount
NOT PASS	1836
SNMPTN	219
SNMPN	78
SPAN PTKIN	69
PMDKPN	9

Table 2. Number of each category in the 'ENTRY PATH' column in the training set

If you add up all the data categories in Table 2 above, it produces 2764 data, which corresponds to the data in the training set.

After obtaining the training set and testing set, the next step is to determine the classification target where the classification target is in the 'ENTRY PATH' category column so that here we separate the 'ENTRY PATH' column into classification targets (Y vector) for both the training set and testing set. After determining the classification target (Y vector), we focus on the data that will be trained on machine learning/symbolized as X input. The X input data can be seen in the image below:

	JUMLAH1	RATA21	JUMLAH2	RATA22	JUMLAH3	RATA23	JUMLAH4	RATA24	JUMLAH5	RATA25	STD
2750	1284.0	80.250000	1311.000000	81.937500	1267.0	84.466667	1291.0	86.066667	1312	87.466667	637.308283
1849	1229.0	81.933333	1232.000000	82.133333	1165.0	83.214286	1186.0	84.714286	1204	86.000000	590.388630
651	1609.0	84.700000	1604.000000	84.500000	1621.0	85.400000	1678.0	88.400000	1676	88.300000	817.991015

Figure 3. Data frame input matrix (X input)

From the data frame in Figure 3 above, it can be seen that there are 11 columns where initially there were 13 columns (Figure 1); this can be understood where the 'ENTRY PATH' column was separated and designated as the classification target (Y), and the 'PTN' column was deliberately omitted/not used with the intention of simplifying machine learning modeling. The next stage is that the data frame X input is subjected to

a standard scaler so that the symbolized as X is prepared. Next, this crafted X matrix data will be input as learning material for the Machine Learning Algorithm used.

As previously known, referring to Table 2 in the training set, the data categories contained in the 'ENTRY PATH' classification target column have different data compositions in each category, where the 'NOT PASS' data category has the most significant number, namely 1836 / 83 percent of the total data, this causes unbalanced data. Apart from that, the data contained in the training set has a total of 2764 data (relatively small), so to make it easier to evaluate the performance of machine learning modeling, classification will be carried out using binary type (with only two classification targets) rather than multiclass classification (with more than two classification targets). Multiclass sort is not used because, in the target class, there is a data category, namely PMDKPN (Table 2), which only has 9 data, which is relatively small. In this study, the binary classification used has two classification targets, namely 'NOT PASS' and 'PASS,' where in the classification target 'PASS' are all categories other than 'NOT PASS,' namely SNMPTN, SNMPN, SPAN PTKIN, and PMDKPN being the four categories (SNMPTN, SNMPN, SPAN PTKIN, and PMDKPN) are compressed into one classification target, 'PASS', then 'NOT PASS' is given the 'FALSE' label and 'PASS' is given the 'TRUE' title.

C. Training model

In this experiment, the machine learning classification algorithm used is Stochastic Gradient Descent (SGD) with two classification targets, namely 'NOT PASS' and 'PASS' with each target label 'FALSE' for 'NOT PASS' and 'TRUE' for ' PASSED'. The following is a block diagram of the training model used in this experiment,

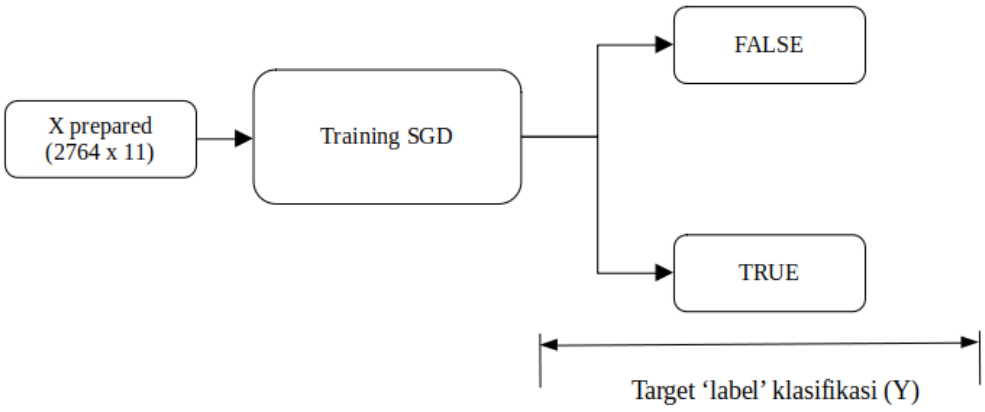

Figure 4. SGD model training

D. Evaluation

After training/learning the model using a machine learning algorithm with two target classification 'labels' 'FALSE' and 'TRUE', the next step is to evaluate the performance of the machine learning modeling. Evaluations carried out to analyze the performance of this modeling include:

1. Determine accuracy with K Fold Cross Validation,
2. Using the confusion matrix, determine the Recall and Precision as well as the F1 value,
3. Using the Recall vs Precision curve,
4. Using the ROC (Receiver Operating Characteristic) curve.

RESULTS AND DISCUSSION

After training the model using Stochastic Gradient Descent (SGD) classification, the next stage is to evaluate the performance of this classification modeling, whether it has good performance (the model can recognize the Dataset and provide good classification performance or not).

A. Determine accuracy with K Fold cross-validation

The first stage is to measure the performance of this machine learning modeling using K Fold cross-validation. K Fold Cross-validation is widely used to measure the performance of machine learning modeling, namely by dividing the training set by many Folds [4]; in this experiment using K=3 foldings, so for each Fold division, there is one-Fold as a test set, it can be explained as in Table 3 below,

Table 3. K Folds Cross Validation, K=3

Train sets	Train sets	Test set
Train sets	Test set	Train sets
Test set	Train sets	Train sets

So, Table 3 above provides information that the training set is divided into three parts (Folds), namely the train set and the test set. The SGD modeling was trained using 2 Folds train sets; then, accuracy tests were carried out using the Folds test set, repeated using different Folds test sets three times (K=3). This SGD modeling gives accuracy numbers of [0.7734057, 0.83310719, 0.8358209], respectively (in the form of a matrix array with three columns because the k-fold used uses k=3). The accuracy figures above

look relatively good (because training and testing use the same training set as that used for previous SGD training)

B. Confusion Matrix

After getting the previous accuracy value using Cross-validation on the training set and getting a relatively good accuracy value, the next stage is to determine the recall, precision, and F1 score values using the Confusion Matrix.

The confusion matrix that was formed initially used cross-validation with the same treatment as in the previous cross-validation experiment. SGD was trained using an existing training set but did not produce an accuracy score like the previous experiment; instead, it had a prediction model. This Prediction Model is tested for classification performance using the original training set data and provides a matrix with row=2, column=2, or confusion matrix with size (2x2); size (2x2) is formed because the modeling used only has two classification targets (2 classification label). The confusion matrix formed in this modeling is as in Table 4 below:

Table 4. Confusion Matrix

	DO NOT PASS (Prediction)	PASS (Prediction)
NOT PASS (Actual)	1733	103
PASS (Actual)	308	67

Information:

: True Positives (TP)

: False Positives (FP)

: True Negatives (TN)

: False Negatives (FN)

From the confusion matrix above, it can be concluded that the 'PASS' category was correctly classified as 'PASS' (TRUE) 67 times (TP). In comparison, it was classified as 'NOT PASS' (FALSE) 308 times (FN), while the 'NOT PASS' category was correctly classified as 'NOT PASS' 1733 times (TN) while classified as 'PASS' 103 times (FP). From this information, the Recall, Precision, and F1 score values can be seen.

Recall = TP/(TP+FN) = 67/(67+308) = 0.17866

Precision = TP/(TP+FP)= 67/(67+103) = 0.39411

F1 score = 0.2458

From the information above, it can be concluded that the performance of our modeling providing Recall, Precision, and F1 score values is not as good as the accuracy values obtained previously (with cross-validation and testing using the training set), so it can be said that our modeling is very overfitting [2].

C. Evaluation using the Precision, Recall versus Decision Threshold curve (Precision Recall Tradeoff)

From the Recall, Precision values that have been obtained, the Precision, Recall versus Decision Threshold curve (Precision Recall Tradeoff) is obtained as shown in the image below,

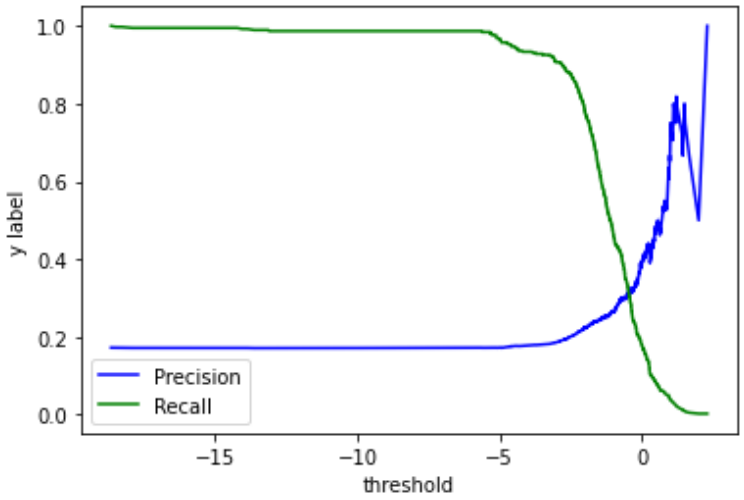

Figure 5. Curve*Precision, Recall versus Decision Threshold (Precision Recall Tradeoff)*

From the description of the curve above, information is obtained that the recall and precision values will give the same value as the previous calculation if the threshold = 0. By changing the point, you will get a change in the recall and precision values (Precision Recall Tradeoff).

D. Evaluation using the ROC (Receiver Operating Characteristic) Curve

Like the treatment to produce the confusion matrix above, the ROC curve is also generated first using cross-validation to create predictive modeling (here, using the decision function to make the y score). After obtaining the prediction model, the actual training set is tested using the existing prediction model and produces a ROC curve like the following,

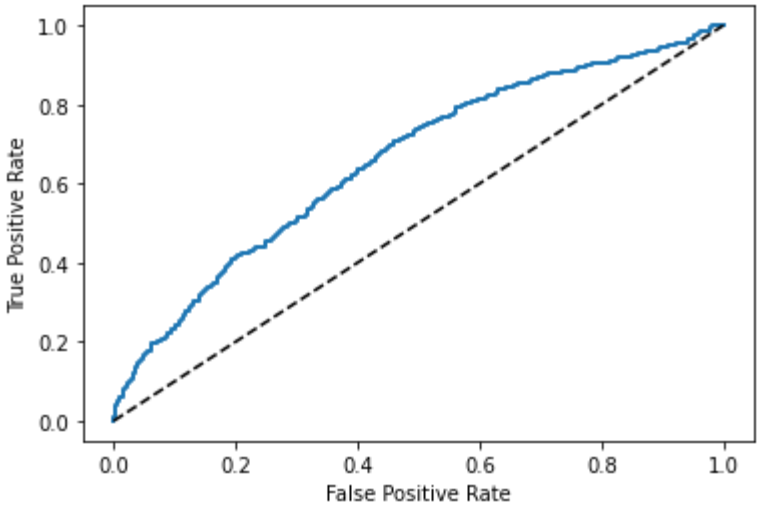

Figure 5. CurveROC (Receiver Operating Characteristics)

The red dashed diagonal line provides information without classification modeling, while the blue line provides classification modeling information in this experiment. The blue line where the False Positive Rate (FPR) is getting closer to 0 and the True Positive Rate (TPR) is getting closer to 1, or the blue line tends to approach the top left corner, then the classification performance used will also be better. From this graph, it can be obtained that the ROC_AUC score in this modeling is equal to 0.6593.

CONCLUSIONS AND RECOMMENDATIONS

From the existing Dataset using the Stochastic Gradient Descent (SGD) classification, the modeling obtained (binary type) still experiences overfitting, as proven by the accuracy values respectively [0.7734057, 0.83310719, 0.8358209] while the recall value = 0.17866, precision value = 0.39411, and F1 = 0.248, this indicates that the modeling has a small error/high accuracy in training. In contrast, validation using cross-validation prediction/testing has a significant error/small precision [3]. One way to overcome this overfitting is to increase the data in the Dataset [2].

REFERENCES

Rachmawati Findiana. (2021). Classification of Student Graduation in State University Entrance Selection through the Report Card Score Route using the Support Vector Machine Method. Sepuluh Nopember Institute of Technology.

Aurelien Geron. (2019). Hands-on Machine Learning with Scikit-Learn and TensorFlow. O'Reilly.

Will Koehrsen, 2018. Overfitting vs. Overfitting Underfitting: A Complete Example. Retrieved August 20, 2023, fromhttps://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765.

Cross Validation (2019). Cross Validation — Why & How. Retrieved July 25, 2023, fromhttps://towardsdatascience.com/cross-validation-430d9a5fee22.