Machine Learning Approach to Predict the Likelihood of Students Continuing Their Studies

Henry
200710607
Program Studi Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia
200710607 {at} students.uajy.ac.id

Jolly Hans Frankle
200710932
Program Studi Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia
200710681 {at} students.uajy.ac.id

Kenneth Vincentius Theys
200710681
Program Studi Informatika
Universitas Atma Jaya Yogyakarta
Yogyakarta, Indonesia
200710932 {at} students.uajy.ac.id

***Abstract*—Paper ini akan menganalisis dan membandingkan keakuratan dari beberapa algoritma *machine learning* dalam memprediksi kemungkinan para siswa untuk lanjut ke perguruan tinggi. Dalam mengklasifikasi data set siswa, akan digunakan algoritma *Decision Tree, SVM,* dan *Random Forest.* Untuk mencari nilai prediksi akan digunakan beberapa parameter seperti finansial orang tua, nilai akademik, akreditasi sekolah, dan lingkungan. Oleh karena itu, maka akan diteliti lebih lanjut mengenai apakah terdapat hubungan dari semua faktor tersebut terhadap keputusan siswa setelah lulus dari pendidikan formal.**

Keywords—Machine Learning Prediction, Student Admission Classification

Decision Tree, SVM, Random Forest, Robust Scaler, Standard Scaler, MinMax Scaler, Select K-Best, Select from Model, ROC AUC, Pipeline

I. PENDAHULUAN

Pendidikan merupakan faktor yang mempengaruhi pilihan hidup siswa. Dengan adanya keberadaan pendidikan, permasalahan sosial yang biasanya berujung pada diskriminasi setidaknya bisa diminimalisir. Meskipun kekhawatiran permasalahan sosial tidak dapat dihilangkan sepenuhnya, pendidikan menyediakan kesempatan bagi individu untuk mengembangkan diri mereka baik secara akademik maupun non-akademik.

Meskipun setiap siswa dapat dipaparkan dengan pendidikan, berbagai faktor dalam kehidupan mereka dapat menghambat mereka untuk berpartisipasi pada pendidikan yang lebih tinggi. Faktor seperti pendapatan orang tua, kondisi ekonomi, pencapaian nilai, motivasi, keterbatasan akses pada sumber daya serta pengalaman, dan faktor lainnya, membatasi pengetahuan siswa dalam memasuki maupun beradaptasi terhadap sistem pendidikan yang lebih tinggi. [1] [2]

Oleh karena itu, untuk membuktikan teori ini kami mengambil dataset dengan nama *"Go to College"* dari Kaggle yang isinya menyerupai dengan tujuan dari penelitian ini. Penelitian yang kami lakukan diproses menggunakan algoritma *machine learning*, dan menggunakan faktor-faktor internal maupun eksternal seperti yang sudah disebutkan di atas sebagai parameter perbandingannya.

II. TINJAUAN PUSTAKA

Saat ini, perguruan tinggi menjadi pilihan dari banyak siswa untuk memastikan dua hal penting yaitu karier dan finansial [3]. Untuk memprediksi lanjut atau tidak-nya seorang siswa ke perguruan tinggi, maka penulis menggunakan beberapa metodologi *machine learning* seperti Random Forest, Linear Regression, Decision Tree, dll.

Dalam penelitian di [1], banyak faktor yang mempengaruhi diterimanya masuknya siswa di perguruan tinggi seperti hubungan dengan keluarga, atau lingkungan. Tetapi yang paling berpengaruh adalah kemampuan akademik pada saat di pendidikan formal.

Survei yang dilakukan di [2], dari 145 data set terdapat 138 siswa yang ingin melanjutkan Pendidikan ke perguruan tinggi.

Survei yang dilakukan di [4], faktor internal mempengaruhi 90% pada kategori sedang dan 10% pada kategori berat untuk lanjut ke perguruan tinggi. Faktor eksternal mempengaruhi 93.33% pada kategori sedang dan 6.67% pada kategori rendah.

Pada [5], faktor status sosial ekonomi dari orang tua memiliki peran yang cukup besar. Siswa dengan ekonomi berkecukupan memiliki kesempatan yang lebih besar untuk masuk ke perguruan tinggi dalam mengembangkan kemampuannya dibandingkan siswa yang berasal dari keluarga kurang mampu.

Dalam [6], kita perlu melakukan beberapa uji coba pada metodologi untuk mencari nilai yang paling akurat seperti *Random Forest, Linear Regression, Stacked Ensemble Learning, Support Vector Regression, Decision Trees, KNN (K-Nearest Neighbor),* dll.

Pada [7], dari hasil perbandingan semua teknik *Machine Learning*, SVM mendapat nilai akurasi paling tinggi yaitu 84.15% sementara *Decision Tree* mendapat nilai akurasi yang paling rendah yaitu 78.54%.

III. METODE PENELITIAN

Penelitian ini berfokus untuk memprediksi seberapa besar kemungkinan seseorang melanjutkan studinya ke perguruan tinggi menggunakan *dataset*, Teknik pembagian data, dan teknik pembelajaran mesin sebagai berikut:

A. Pengolahan Dataset

Dataset *"Go to College"* merupakan dataset sintetis yang tersedia di bawah lisensi *Creative Commons 0: Public Domain* di situs Kaggle, dengan 1.000 data dan 11 atribut yang berpengaruh cukup signifikan dalam menentukan apakah seorang siswa akan lanjut studi atau tidak. Dalam pemeriksaan data, terdapat 500 data yang memiliki result (*value y*) = True dan 500 data yang memiliki result = false, dengan demikian, data sudah terbagi dengan rata. Agar dapat digunakan dalam penelitian, beberapa atribut dari dataset yang merupakan data kategorikal berisi teks diubah menjadi data integer.

Beberapa atribut yang digunakan dalam dataset ini antara lain:

Tabel 1. Daftar atribut dalam dataset

No	Atribut	Keterangan
1	Tipe Sekolah Asal	Tipe sekolah asal para siswa. 1 jika sekolah Akademik, 2 jika sekolah Vokasi.
2	Akreditasi Sekolah Asal	Akreditasi sekolah asal para siswa. 1 jika A, 2 jika B, 3 jika C.
3	Jenis Kelamin	Jenis kelamin para siswa. 1 jika laki-laki, 2 jika perempuan.
4	Ketertarikan	Skala 1-5, di mana 1 = tidak tertarik hingga 5 = sangat tertarik
5	Tempat Tinggal	Tempat tinggal para siswa. 1 jika di kota (*urban*), 2 jika di luar kota (*rural*).
6	Umur Orang Tua	Umur orang tua siswa
7	Gaji Orang Tua	Gaji orang tua siswa
8	Luas Rumah	Luas lingkungan rumah tempat tinggal para siswa.
9	Nilai Rata-rata	Nilai rata-rata selama menempuh pendidikan menengah.
10	Orang Tua Pernah Melanjutkan Studi	Apakah orang tua siswa pernah melanjutkan studi di pendidikan tinggi.
11	Melanjutkan Studi	*Hasil:* apakah siswa memutuskan melanjutkan studi atau tidak.

B. Pembagian Data dan Pelatihan Model

Dari 1.000 data yang disediakan dataset, 30% (300 data) dipilih menjadi data *training* (uji coba), dan 70% (700 data) dipilih menjadi data *testing* (pengujian). Pembagian data kami lakukan dengan menggunakan fungsi train_test_split yang disediakan Scikit-Learn.

Data dalam penelitian ini ditransformasi menggunakan *preprocessor* antara lain Standard Scaler, Min-Max Scaler, serta Robust Scaler.

Dalam penelitian ini, model dilatih menggunakan metode *pipeline* dengan algoritma *Random Forest*, *Decision Tree*, serta *Support Vector Machine* (*SVM*). Akurasi antara ketiga algoritma tersebut dan *tuning* atau parameter dari ketiga algoritma tersebut akan dijelaskan pada bagian HASIL DAN DISKUSI.

1) Decision Tree (DT): Pohon keputusan adalah sebuah struktur menyerupai diagram alur, di mana memiliki akar, cabang, dan daun. Pohon keputusan mempelajari hierarki pertanyaan if/else, yang mengarah pada keputusan tertentu. Algoritma ini akan mencari semua pertanyaan if/else yang mungkin muncul dari fitur dan menemukan yang paling informatif tentang variabel target.

2) Random Forest Classifier:Random forest merupakan kumpulan dari pohon keputusan (*decision tree*) di mana setiap pohon memiliki struktur yang berbeda dengan pohon lain. Pohon-pohon yang terbentuk belum tentu cocok digunakan untuk semua data, melainkan pada beberapa jenis data saja. Dalam penggunaannya, *random forest* dapat menggunakan ratusan hingga ribuan pohon untuk membentuk *decision boundary* yang lebih halus.

3) Support Vector Machine (SVM): SVM mempelajari tingkat kepentingan setiap data dalam menentukan *decision boundary*. Biasanya hanya sebagian kecil data dianggap penting untuk menentukan *decision boundary*, yaitu data point yang terletak pada batas antar kelas.

4) Standard Scaler: scaler yang mengubah nilai rerata setiap *feature* menjadi 0 dan nilai variasi menjadi 1. Semua *feature* aka nada pada besaran yang sama, namun tidak ada nilai minimum dan maksimum untuk setiap feature.

5) Min-Max Scaler: scaler yang mengubah data pada setiap *feature* agar memiliki nilai minimum 0 dan nilai maksimum 1.

6) Robust Scaler: scaler yang mengubah data pada besaran yang sama dengan menggunakan nilai median dan kuartil, dengan mengabaikan data yang sangat berbeda (*outlier*).

C. Evaluasi Model

Dalam penelitian ini, model dievaluasi menggunakan cross-validation – khususnya 10-fold cross-validation, confusion matrix – guna mendapat akurasi, presisi, *recall*, dan F1-measure, serta *Receiver Operating Characteristics Curve* (*ROC Curve*).

1) Cross-Validatrion: merupakan metode statistik untuk evaluasi performa generalisasi suatu model *machine learning*. Metode *cross-validation* yang paling sering digunakan adalah *K-Fold Cross-Validation*, di mana 'k' adalah jumlah lipatan yang ditentukan oleh pengguna.

2) Confusion matrix: merupakan matriks yang menyajikan visualisasi performa dari suatu algoritma. Matriks konfusi menyajikan dua dimensi, yakni "actual" (hasil sebenarnya), dibandingkan terhadap "predicted" (hasil yang diprediksi model).

3) Receiver Operating Characteristics Curve: merupakan kurva yang menunjukkan *FPR* (*false positive rate*) terhadap *TPR* (*true positive rate*). Kurva yang ideal adalah kurva yang mendekati pojok kiri atas.

IV. HASIL DAN DISKUSI

Dataset yang digunakan diuji menggunakan metode SVM dan RandomForestClassifier, dimana masing-masing metode menggunakan 10-Fold Cross Validation. Fungsi yang digunakan sebagai pengukuran kinerja utama dalam penelitian ini adalah Precision, Recall, F-Measure, dan ROC. Precision atau nilai prediksi positif merupakan nilai yang diprediksi positif yang sebenarnya positif. Recall atau sensitivitas merupakan nilai benar atau positif yang ditangkap oleh prediksi positif. F-Measure merupakan gabungan rata-rata dari hasil Precision dan Recall yang digunakan untuk mengevaluasi classifier, dalam hal ini yaitu SVM dan RandomForestClassifier. ROC memaparkan perbandingan antar FPR (False Positive Rate) terhadap TPR (True Positive

Rate) dalam bentuk kurva, kurva dinilai ideal apabila lebih mendekati bagian pojok kiri atas.

Dataset yang digunakan diuji menggunakan metode SVM dan RandomForestClassifier, dimana masing-masing metode menggunakan 10-Fold Cross Validation. Fungsi yang digunakan sebagai pengukuran kinerja utama dalam penelitian ini adalah Precision, Recall, F-Measure, dan ROC. Precision atau nilai prediksi positif merupakan nilai yang diprediksi positif yang sebenarnya positif. Recall atau sensitivitas merupakan nilai benar atau positif yang ditangkap oleh prediksi positif. F-Measure merupakan gabungan rata-rata dari hasil Precision dan Recall yang digunakan untuk mengevaluasi classifier, dalam hal ini yaitu SVM dan RandomForestClassifier. ROC memaparkan perbandingan antar FPR (False Positive Rate) terhadap TPR (True Positive Rate) dalam bentuk kurva, kurva dinilai ideal apabila lebih mendekati bagian pojok kiri atas.

Tabel 2. Hasil percobaan serta nilai akurasi menggunakan 10-fold cross validation

Classifier	Scaler	Precision	Recall	F-Measure	AUC score	Accuracy (%)
Random Forest Classifier	*Tidak ada*	0.873	0.873	0.873	0.941	89,33
Random Forest Classifier	Standard Scaler	0.874	0.874	0.874	0.941	89,67
Random Forest Classifier	Min-Max Scaler	0.873	0.873	0.873	0.941	89,67
Random Forest Classifier	Robust Scaler	0.873	0.873	0.873	0.941	89,33
Support Vector Machine (SVM)	*Tidak ada*	0.711	0.507	0.359	0.614	52,00
Support Vector Machine (SVM)	Standard Scaler	0.851	0.849	0.848	0.928	89,33
Support Vector Machine (SVM)	Min-Max Scaler	0.864	0.863	0.863	0.934	85,33
Support Vector Machine (SVM)	Robust Scaler	0.886	0.886	0.886	0.955	90,67

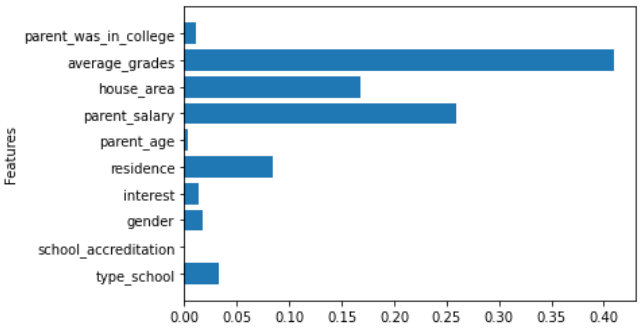
1) Percobaan pertama

Pada percobaan pertama, beberapa algoritma yang dibandingkan antara lain Random Forest Classifier dan Support Vector Machine (SVM), dengan penskalaan data menggunakan Standard Scaler, Min-Max Scaler, dan Robust Scaler. Nilai parameter yang digunakan dalam Random Forest Classifier antara lain n_estimators=100 dan random_state=0. Sedangkan nilai parameter yang digunakan dalam Support Vector Machine (SVM) adalah C=10 dan gamma=0.1.

Mengacu pada tabel dua (2), dataset diaplikasikan kedalam classifier RandomForest dan SVM sembari menggunakan teknik *10-Fold Cross-Validation* untuk memperkirakan performa masing-masing metode, selain itu setiap classifier di-*scale* menggunakan tiga (3) tipe scaler yakni Standard Scaler, Min-Max scaler, dan Robust Scaler. Pada classifier RandomForest, nilai accuracy untuk semua scaler secara rata-rata berkisar di 89,5%, dengan scaler Standard Scaler dan Min-Max Scaler memperoleh nilai accuracy tertinggi yakni 89,67%, dengan rincian: Precision, Recall, dan F-Measure sebesar 0,874 untuk Standard Scaler dan Precision, Recall, dan F-Measure sebesar 0,873 untuk Min-Max Scaler dengan AUC Score sebesar 0,941.

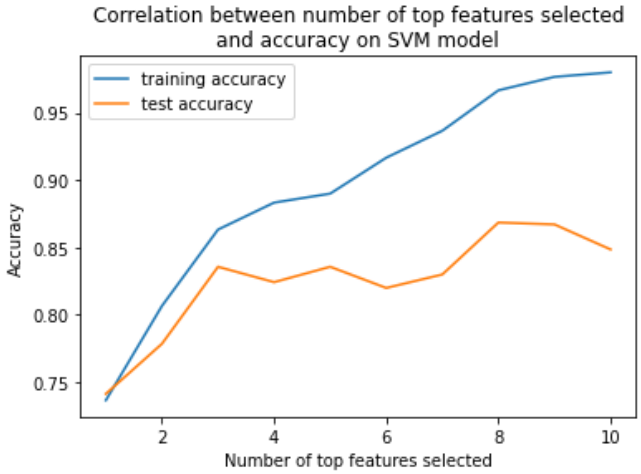
Pada Classifier SVM, nilai accuracy untuk semua scaler secara rata-rata berkisar di 79.33%, dengan scaler Robust Scaler memperoleh nilai accuracy tertinggi yakni 90,67%, dengan rincian: Precision, Recall, dan F-Measure sebesar 0.886 dengan AUC score = 0.955.

2) Percobaan kedua

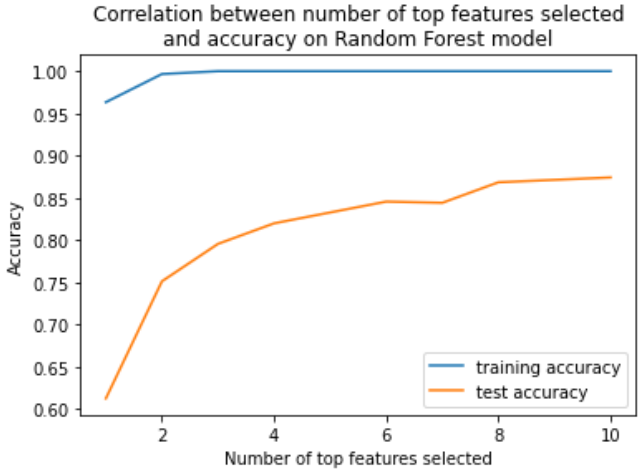


Figur 1. Feature importance (Decision Tree)

Berdasarkan figur satu (1), terdapat 10 fitur yang berpengaruh dalam menentukan hasil akhir. Data diproses menggunakan algoritma Decision Tree dan direpresentasikan dalam bentuk diagram batang. Dapat diamati bahwa terdapat tiga (3) fitur yang memiliki relevansi terbesar terhadap perolehan hasil antara lain nilai rata-rata siswa (average_grades) sebesar 41%, gaji orang tua (parent_salary) sebesar 26%, serta luas lingkungan rumah (house_area) sebesar 17%. *Feature importances diperoleh menggunakan Decision Tree Classifier, dengan parameter max_depth = 8.*

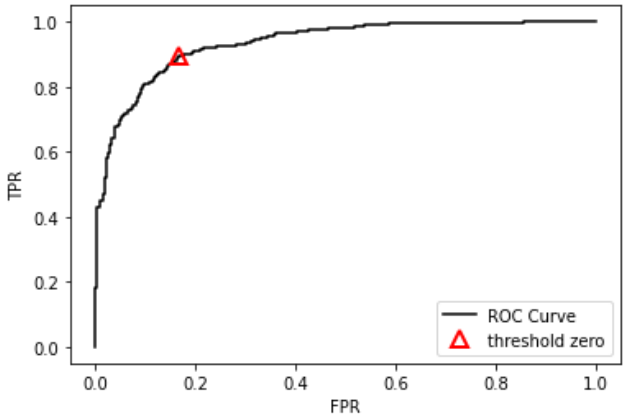


Figur 2. Korelasi antara jumlah fitur terbaik yang dipilih dengan akurasi prediksi pada model SVM

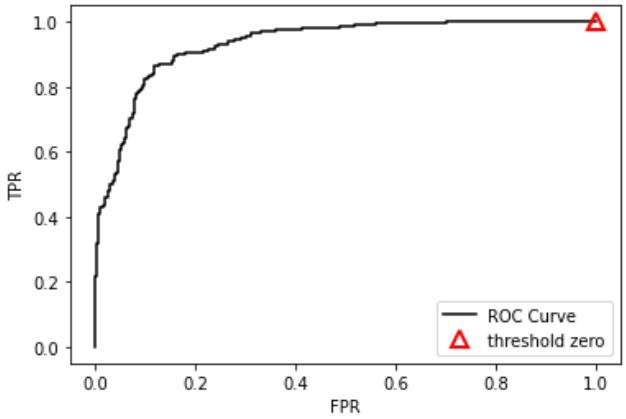


Figur 3. Korelasi antara jumlah fitur terbaik yang dipilih dengan akurasi prediksi pada model Random Forest

Figur dua (2) dan figur tiga (3) merupakan grafik yang merepresentasikan korelasi antara jumlah fitur terbaik yang dipilih oleh algoritma *Select K Best* dan akurasi pada set *training* dan *test* menggunakan dua (2) jenis *classifier* yaitu Support Vector Machine (SVM) dan Random Forest Classifier secara berurut. Dapat diamati bahwa secara umum, akurasi meningkat seiring bertambahnya jumlah fitur yang dipilih. Namun terdapat perbedaan pada SVM dan Random Forest, di mana akurasi set *test* pada SVM tidak meningkat secara konsisten, perilaku yang sebaliknya teramati pada algoritma Random Forest.



Figur 4. ROC curve untuk SVC, C=10, gamma=0.1



Figur 5. ROC curve untuk Random Forest, max_depth=5

Figur empat (4) dan figur lima (5) yang ditampilkan di atas merupakan grafik yang merepresentasikan kurva ROC. Sebuah kurva ROC dikatakan memiliki performa ideal apabila kurva lebih mendekati sudut kiri atas dari diagram XY. Kurva ROC dibangun dengan memplot *True Positive Rate* (TPR) terhadap *False Positive Rate* (FPR), dan kurva ROC digunakan juga untuk membantu dalam menemukan letak ambang (*threshold*) ketika kondisi TPR tinggi dan FPR rendah. Selanjutnya terdapat *threshold* yang direpresentasikan dalam bentuk segitiga berwarna merah pada kedua figur. *Threshold* adalah penanda yang digunakan untuk menemukan titik nilai dari Positive Rate dan False Rate. Dapat diamati bahwa pada figur empat (4), letak *threshold* berada pada posisi di atas 0.8 sumbu TPR, dan di bawah 0.2 sumbu FPR. Namun pada figur lima (5), letak *threshold* berada pada akhir dari garis kurva bagian kanan atas, dimana berbandingan TPR dan FPR sebanding. Letak *threshold* pada figur lima (5) yang menggunakan *classifier* Random Forest menghasilkan hasil yang jauh dari ekspektasi, dibandingkan dengan letak *threshold* pada figur empat (4) yang menggunakan *classifier* SVM. Berdasarkan pengamatan terhadap kedua figur, dapat disimpulkan bahwa figur empat (4) dengan menggunakan *classifier* SVC merupakan *classifier* yang lebih baik dalam hal efisiensi, karena lebih dekat dengan bagian sudut kiri atas dari sumbu XY.

V. KESIMPULAN

Analisis dan perbandingan pada dataset “kemungkinan siswa melanjutkan Pendidikan ke perguruan tinggi” dengan berbagai algoritma *machine learning* seperti Random Forest Classifier dan Support Vector Machine (SVM) ternyata memiliki hasil keakuratan yang berbeda. Setiap jenis scaler yang digunakan seperti Standard Scaler, Min-Max Scaler, dan Robust Scaler juga mempengaruhi nilai akurasi yang dihasilkan. Berdasarkan analisis yang telah dilakukan, Classifier dengan Support Vector Machine (SVM) dengan Robust Scaler merupakan algoritma paling akurat dengan nilai 90,67%.

Korelasi antara jumlah fitur terbaik yang dipilih dengan akurasi prediksi juga bergantung pada jenis algoritma klasifikasi yang digunakan. Pengujian dengan algoritma Support Vector Machine (SVM) mendapat hasil yang lebih baik karena kenaikan antara akurasi *training* dan *test* cukup setara dibandingkan dengan Random Forest.

Decision Tree dapat digunakan untuk memprediksi faktor apa saja yang paling mempengaruhi seorang siswa untuk lanjut ke perguruan tinggi atau tidak. Pada percobaan ini diketahui bahwa faktor yang paling mempengaruhi adalah rata-rata nilai siswa selama menempuh Pendidikan formal.

Berdasarkan analisis yang sudah dilakukan, dapat disimpulkan bahwa ROC Curve dengan menggunakan *classifier Support Vector Machine (SVM)* mendapat hasil yang lebih baik dalam hal efisiensi karena kurva yang terbentuk lebih mendekati titik 0,1 dibandingkan ROC *curve* dengan *Random Forest* yang lebih mendekati *baseline.*

VI. KONTRIBUSI PENULIS

Semua penulis berkontribusi dalam penulisan kode, analisis hasil, serta interpretasi hasil. Semua penulis juga berkontribusi dalam mencari referensi dan menulis isi dari paper ini.

VII. DAFTAR PUSTAKA

[1] L. K. Wayt, "The Impact of Students' Academic and Social Relationships on College Student," *Educational Administration: Theses, Dissertations, and Student Research*, 2012.

[2] S. L. Temple, "Factors that Influence Students' Desires to Attend Higher Education," *Seton Hall University Dissertations and Theses (ETDs)*, 2009.

[3] I. E. Guabassi, M. Rim, Z. Bousalem and A. Qazdar, "A Recommender System for Predicting Students' Admission to a Graduate Program using Machine," *A Recommender System for Predicting Students' Admission to a Graduate*, 2021.

[4] S. Armalita, "Faktor-Faktor Yang Mempengaruhi Minat Untuk Melanjutkan Studi Ke Perguruan Tinggi Siswa Kelas Xll Jurusan Tata Boga, Di Smk Negeri 4 Dan Smk Negeri 6 Yogyakarta," 2016.

[5] S. Khadijah, H. Indrawati and Suarman, "Analisis Minat Peserta Didik untuk Melanjutkan Pendidikan Tinggi," *Jurnal Pendidikan Ilmu Sosial*, vol. 26, no. 2, 2017.

[6] P. Golden, K. Mojesh, L. M. Devarapalli, P. N. S. Reddy, S. Rajesh and A. Chawla, "A Comparative Study on University Admission Predictions Using Machine," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 7, no. 2, pp. 537-548, 2021.

[7] S. Pouriyeh, S. Vahid, G. Sannino, G. D. Pietro, H. Arabnia and J. Gutierrez, "A Comprehensive Investigation and Comparison of Machine Learning Techniques in the Domain of Heart Disease," *Symposium on Computers and Communication*, vol. 22, 2017.