MULTICLASS SVM DENGAN OPTIMIZER PARAMETER GRID UNTUK MEMPREDIKSI PERFORMANCE STUDEN

Alif Apris Setiawan[1]
[1,2,3] Program Studi Pendidikan Informatika, STKIP NU Tegal
kopralvanjono007@gmail.com
Jl. Raya Sel. Banjaran No.21, Procot, Kec. Slawi, Kabupaten Tegal, Jawa Tengah 52412

Keywords:

Data Mining, Classification, Prediction, SVM.

Abstract

Student performance has an important role to measure the quality of students. Prediction is one way that can be used to predict student performance. Prediction can be done with data mining techniques. One of the data mining techniques that can be used is the classification method. This study uses the SVM algorithm method using the Parameter Grid and the dataset used is a public dataset taken from UCI Machine Learning and the tools used are rapidminer. The test is carried out using training data and testing data with a comparison of 70% for training data and 30% for testing data and the accuracy results obtained from testing are 99.64%. Based on these results, the SVM method using the Parameter Grid is more efficient to predict student performance.

Kata Kunci:

Data Mining, Klasifikasi, Prediksi, SVM.

Abstrak

Performance setudent memiliki peranan penting untuk mengukur kualitas siswa prediksi adalah salah satu cara yang dapat digunakan untuk memprediksi performance siswa. Prediksi dapat dilakukan dengan teknik data mining. Salah satu teknik data mining yang dapat digunakan adalah dengan metode klasifikasi. Penelitian ini menggunakan metode algoritma SVM dengan menggunakan Parameter Grid dan dataset yang digunakan adalah dataset publik yang diambil dari UCI Machine Learning dan tools yang digunakan adalah rapidminer. Pengujian dilakukan dengan menggunakan data training dan data testing dengan perbandingan 70% untuk data training dan 30 % untuk data testing dan hasil akurasi yang diperoleh dari pengujian adalah 99.64%. Berdasarkan hasil tersebut maka metode SVM dengan menggunakan Parameter Grid lebih efisien untuk memprediksi performance siswa.

Pendahuluan

Semakin pesatnya perkembangan teknologi dalam pengolahan data melahirkan banyak inovasi didalamnya, data merupakan kumpulan informasi yang dapat diolah dan digunakan untuk menganalisa berbagai macam kebutuhan, data dapat berupa apa saja yang berhubungan dengan kondisi lingkungan, masyarakat, pendidikan dan lain sebagainya. Data baru dapat muncul dari waktu ke waktu. Menganalisa data membutuhkan kemampuan khusus dalam bidang pengolahan data untuk itu dibutuhkan pembelajaran mesin agar dapat menganalisa data lebih efisien dan dapat menyimpan data secara digital dengan format yang baik. Pembelajaran mesin merupakan cabang dari Artifical Intelligence yang dapat memberikan kemampuan untuk pembelajaran mesin secara otomatis.

Dalam dunia pendidikan Educational Data Mining adalah salah satu alternatif untuk pengolahan data dalam bidang pendidikan dengan memanfaatkan data yang berhubungan dengan pendidikan. Organisasi dalam bidang pendidikan menggunakan data untuk memperoleh berbagai macam informasi terutama informasi tentang siswa. data siswa memiliki berbagai macam atribut sehingga dapat diolah untuk membuat prediksi seperti prediksi performance student.

Prediksi dapat dilakukan dengan teknik data mining. Beberapa atribut dalam data siswa memiliki peranan penting untuk mengukur seberapa jauh kualitas performance student. Salah satu teknik data mining yang dapat digunakan untuk prediksi performance student adalah dengan metode klasifikasi. Klasifikasi adalah mengkelompokan data berdasarkan ciri – ciri yang memiliki beberapa persamaan dan perbedaan.

Penelitian ini merupakan implementasi machine learning dalam dunia pendidikan. Hasil dari penelitian ini adalah untuk memprediksi performance student dengan menggunakan algoritma Support Vector Machines (SVM) dengan optimize parameter grid.

Landasan Teori

Beberapa penelitian telah mengajukan berbagai model prediksi untuk memprediksi siswa, bagian ini menguraikan beberapa penelitian terkait prediksi performance student.

Penelitian Cortez terkait dengan memprediksi kinerja siswa sekolah menengah. Prediksi dilakukan dengan menggunakan algoritma Decision Trees, Random Forest, Neural Networks dan Support Vector Machines. Hasilnya akurasi kinerja siswa akan baik jika nilai pertama dan kedua terpenuhi[1].

Emny dkk. comparison of data mining classification algorithms for student performance Hasil penelitian menunjukkan algoritma klasifikasi terbaik untuk kinerja siswa dengan data matematika siswa adalah hutan acak G sebesar 89,78%. Selain itu, tiga algoritma yang memiliki performansi terbaik adalah random forest, adaboost dan K-Nearest Neighboring[2].

Hartatik dkk. Prediksi Kelulusan Mahasiswa Dengan Naive Bayes algoritma Berdasarkan pengolahan hasil observasi observasi dalam penelitian, ditemukan bahwa akurasi tingkat prediksi dengan menggunakan variabel IP siswa menghasilkan akurasi sebesar 75% dan R2 = 68,2%. Model 2 untuk prediksi digunakan 8 variabel yaitu Nim, Jenis Kelamin, Tempat Tinggal, IPS 1, IPS2, IPS3, IPS4, masa studi dan status siswa akurasi hasil prediksi 89% dan dengan tingkat prediksi R2 = 71,4%. Berdasarkan perbandingan pada model 1 dan model 2, model prediksi 1 memberikan hasil yang lebih baik daripada model 1, sehingga dapat menunjukkan bahwa model prediksi dengan menggunakan variabel IP, nilai UN, JK, status kependudukan dapat memberikan prediksi yang lebih baik daripada model yang menggunakan IP. Kemudian pada penelitian selanjutnya dapat digunakan model dengan menambahkan variabel dan algoritma Tambahan[3].

A.Dinesh Kumar dkk. Algoritma Klasifikasi Hibrida untuk Memprediksi Kinerja Siswa Dalam penelitian ini kami telah menggunakan jaringan Radial Basis Function, Multilayer perceptron, dan algoritma klasifikasi J48 dan Random Forest untuk meramalkan kinerja akademik siswa. Keempat algoritma ini diklasifikasikan secara individual dan akurasi klasifikasi dihitung. Kemudian algoritma RBF dan MLP digabungkan bersama dan akurasi dihitung. Kemudian kami membuat algoritma klasifikasi hybrid lain J48 dan Random Forest. Algoritma ini memberikan akurasi yang lebih baik sebesar 76,4583 dibandingkan dengan klasifikasi hybrid RBF dan MLP. Jadi, kami menyimpulkan bahwa algoritma klasifikasi hybrid J48 dan Random Forest bekerja lebih baik daripada algoritma klasifikasi hybrid RBF dan MLP[4].

Chandra Wirawan dkk. Penerapan Data mining untuk Prediksi Ketepatan Waktu Wisuda Mahasiswa Tujuan dari penulisan ini adalah untuk membandingkan metode prediksi terbaik dari 3 metode yang diuji untuk menentukan prediksi yang lolos tepat waktu. Dari 3 metode tersebut, metode pohon keputusan dengan menggunakan algoritma C.4.5 merupakan metode yang memiliki nilai akurasi tertinggi dibandingkan dengan kedua metode lainnya yaitu 89,82%, sedangkan akurasi metode Naïve Bayes sebesar 85,4% dan akurasi metode KNN sebesar 84,07%. Dalam penulisan ini, dataset yang digunakan untuk memprediksi tingkat kelulusan adalah 754 siswa yang terdiri dari 9 atribut prediksi dan 1 atribut hasil, sehingga total atribut yang digunakan adalah 10 atribut parameter. Terdiri dari type_gender, program_studi, types_School, majors_SLTA, region, IPSmt 1-4, dan wisuda[5].

Akarshita Tripathi dkk. Model Klasifikasi Naïve Bayes untuk Siswa Prediksi Kinerja Dalam karya ini, disimpulkan bahwa prediksi kinerja siswa adalah tantangan utama analisis prediksi karena kumpulan data yang kompleks. Pendekatan klasifikasi SVM diterapkan dalam pekerjaan penelitian sebelumnya untuk prediksi kinerja siswa. Untuk meningkatkan akurasi prediksi kinerja siswa, pengklasifikasi nave

bayes diterapkan di pekerjaan penelitian ini. Model yang diusulkan diimplementasikan dalam python dan hasilnya dianalisis dalam hal akurasi, waktu eksekusi. Dianalisis bahwa model yang diusulkan memiliki akurasi yang tinggi dan waktu eksekusi yang rendah dibandingkan dengan model yang ada[6].

Metode

A. Dataset

Dalam penelitian ini kami menggunakan dataset public yang diambil dari: https://archive.ics.uci.edu/ml/datasets/student+performance. Dataset ini merupakan prestasi siswa dari dua sekolah menengah di portugis. Atribut dalam dataset ini meliputi nilai siswa, demografi, sosial dan fitur terkait sekolah. Dataset ini memiliki 32 atribut dari 394 siswa untuk mempermudah proses data mining maka peneliti merubah variabel data menjadi numerik, berikut adalah rincian dari atribut yang ada dalam dataset:

Tabel 1. Deskripsi Atribut Performance Student

Attribute	**Description**
school	student's school (binary: 0 or 1)
Sex	student's sex (binary: 0 or 1)
Age	student's age (numeric: from 15 to 22)
Address	student's home address type (binary: 0 or 1)
Famsize	family size (binary: 0 or 1)
Pstatus	parent's cohabitation status (binary: 0 or 1)
Medu	mother's education (numeric: 1-4)
Fedu	father's education (numeric: 1-4)
Mjob	mother's job (nominal: 1-5)
Fjob	father's job (nominal: 1-5)
Reason	reason to choose this school (nominal: 1-4)
guardian	student's guardian (nominal:1-3)
traveltime	home to school travel time (numeric: 1-4)
studytime	weekly study time (numeric: 1-4)
Failures	number of past class failures (numeric:0-3)
schoolsup	extra educational support (binary: 0 or 1)
Famsup	family educational support (binary: 0 or 1)
Paid	extra paid classes within the course subject (binary: 0 or 1)
activities	extra-curricular activities (binary: 0 or 1)
Nursery	attended nursery school (binary: 0 or 1)
Higher	wants to take higher education (binary: 0 or 1)
Internet	Internet access at home (binary: 0 or 1)
romantic	with a romantic relationship (binary: 0 or 1)
Famrel	quality of family relationships (numeric: 1 - 5)
freetime	free time after school (numeric: 1 - 5)
Goout	going out with friends (numeric: 1 - 5)
Dalc	workday alcohol consumption (numeric: 1 - 5)
Walc	weekend alcohol consumption (numeric: 1 - 5)
Health	current health status (numeric: 1 - 5)
absences	number of school absences (numeric: 0 - 93)
G1	first period grade (numeric: 0 - 20)
G2	second period grade (numeric: 0 - 20)
G3	final grade (numeric: 0 - 20)

B. Data Preprocessing

Setelah menentukan dataset yang sesuai kemudian tahap selanjutnya adalah preprocessing data merupakan tahapan untuk membuat model pembelajaran mesin, data yang awalnya berupa text dan angka semuanya diubah menjadi data numerik untuk memudahkan proses data mining kemudian membuat variabel untuk mengkategorikan grade siswa. Untuk siswa yang memiliki nilai dalam kisaran 15 sampai dengan 20 masuk dalam kategori Baik dengan simbol G, untuk siswa yang memiliki nilai pada kisaran 10 sampai dengan 14 masuk dalam kategori rata-rata dengan simbol A, dan untuk siswa yang memiliki nilai dibawah 10 masuk dalam kategori rendah dengan simbol P.

C. Metode

1. Support Vector Machine (SVM)

Dalam penelitan ini menggunakan metode Support Vector Machine (SVM) untuk mengklasifikasikan data. Support Vector Machine atau SVM adalah algoritma pembelajaran mesin yang diawasi yang dapat digunakan untuk klasifikasi dan regresi. Cara kerja SVM didasarkan pada SRM atau Structural Risk Minimization yang dirancang untuk mengolah data menjadi Hyperplane yang mengklasifikasikan ruang input menjadi dua kelas. Teori SVM diawali dengan pengelompokan kasus-kasus linier yang dapat dipisahkan dengan hyperplane dan dibagi menurut kelasnya.

Konsep SVM diawali dengan masalah klasifikasi dua kelas sehingga membutuhkan set pelatihan positif dan negatif. SVM akan berusaha mendapatkan hyperplane (pemisah) sebaik mungkin untuk memisahkan kedua kelas dan memaksimalkan margin kedua kelas tersebut.

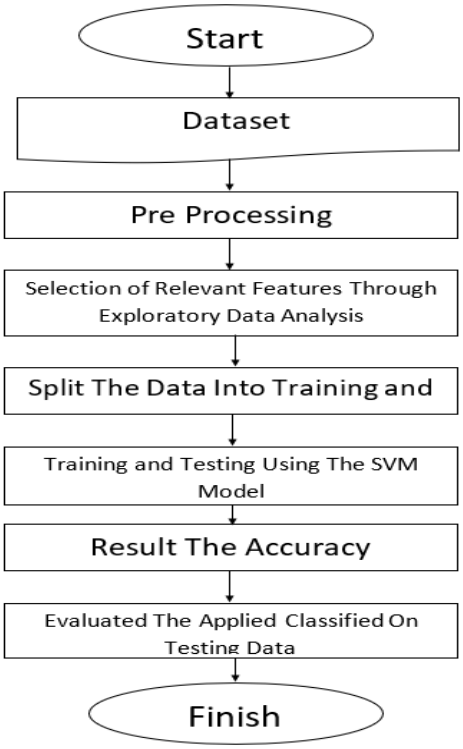
2. Parameter Grid

Pencarian grid merupakan pencarian lengkap berdasarkan subset yang ditentukan dari ruang hiperparameter. Parameter hiper ditentukan dengan menggunakan nilai minimal (batas bawah), nilai maksimal (batas atas) dan jumlah langkah. Ada tiga skala berbeda yang dapat digunakan: skala linier, skala kuadrat dan skala logaritma. Performa setiap kombinasi dievaluasi menggunakan beberapa metrik kinerja. Mengoptimalkan parameter SVM dengan menggunakan teknik Cross Validation (CV) sebagai metrik kinerja. Tujuannya adalah kombinasi untuk mengidentifikasi hyper-parameter yang baik sehingga classifier dapat memprediksi data yang tidak diketahui secara akurat. Teknik Cross Validation dapat mencegah masalah over-fitting. Untuk memilih C dan menggunakan k-fold Cross Validation, pertama-tama kita membagi data yang tersedia menjadi k subset (dalam sebagian besar percobaan kami menetapkan k=10). Kemudian menghitung kesalahan Cross Validation menggunakan kesalahan split untuk mengklasifikasi SVM menggunakan nilai C yang berbeda, dan parameter lainnya. Berbagai kombinasi nilai hyper-parameter dimasukkan dan yang memiliki akurasi cross validation terbaik dipilih dan digunakan untuk melatih SVM di seluruh kumpulan data. Dalam kernel linier hanya ada satu parameter penting untuk dioptimalkan yaitu C, dalam RBF kernel dan kernel sigmoid ada 2 parameter: C dan sedangkan kernel polinomial memiliki 3 parameter: C, dan degree.

Semua kombinasi parameter dihasilkan dan subproses dijalankan untuk setiap kombinasi. Dalam mode tersinkronisasi, tidak ada kombinasi yang dibuat tetapi nilai parameter diperlakukan sebagai daftar kombinasi. Untuk iterasi pada satu parameter tidak ada perbedaan antara kedua mode. Jumlah kemungkinan parameter harus sama untuk semua parameter dalam mode tersinkronisasi.

3. Tahapan Proses Penelitian

Berikuta adalah tahapan penelitian tentang prediksi performance student dengan menggunakan algoritma Support Vector Machines (SVM) dengan Optimize Parameter Grid :



Gambar 1. Tahapan Proses Penelitian

Tahapan penelitian meliputi :

1. Dataset
   Dataset yang digunakan adalah dataset public yang diambil dari dataset UCI Machine Learning. Merupakan data siswa yang memiliki 33 attribut, data berupa text dan numerik.
2. Pre Processing
   Setelah mendapatkan dataset kemudian langkah selanjutnya adalah mengolah data, dimulai dengan mengubah data yang berbentuk text kedalam bentuk numerik dan mengkategorikan grade siswa untuk dijadikan label.
3. Selection of Relevant Features Through Exploratory Data Analysis
   Tahap selanjutnya yaitu menyeleksi data yang berkaitan dengan feature kemudian mengklasifikasikan data dan menganalisa data agar dapat diterapkan kedalam metode SVM dengan menggunakan Parameter Grid.
4. Split The Data Into Training and Testing
   Setelah melakukan proses analisa kemudian bagi data untuk dijadikan data latih dan data uji dalam penelitian ini dilakukan pembagian dengan bobot nilai 0.7 untuk data training dan 0.3 untuk data uji.
5. Training and Testing Using The SVM Model with Parameter Grid
   Setelah melakukan proses pembagian data training dan data uji tahap selanjutnya yaitu validasi data dan melakukan pengujian data kedalam metode algoritma SVM dengan Parameter Grid.
6. Result The Accuracy
   Selanjutnya peneliti melakukan perhitungan berapa nilai akurasi, presisi dan recall yang dihasilkan.
7. Evaluated The Applied Classified On Testing Data
   Setelah melakukan perhitungan akurasi tahap selanjutnya yaitu melakukan evaluasi apakah metode yang digunakan lebih efisien atau perlu dilakukan perbaikan.

Hasil dan Pembahasan

Implementasi dimulai dari pengolahan dataset, data yang berupa text dirubah dalam bentuk numerik kemudian mengkategorikan grade siswa. Sebelum diolah menggunakan aplikasi data RapidMiner, melakukan validasi dengan menghapus data yang tidak lengkap, setelah proses validasi kemudian tahap selanjutnya data dibagi menjadi dua bagian data latih dan data uji dengan perbandingan 70% dan 30% selanjutnya mengimplementasikan algoritma dengan menggunakan model Multiclass SVM dengan Parameter Grid pada data latih, tahapan selanjutnya yaitu menerapkan model pengujian menggunakan operator Apply Model. Dataset kemudian diuji dan dianalisis menggunakan metode SVM dengan Parameter Grid. Tahap terakhir adalah menguji model dengan menggunakan confusion matrix, pada tahap ini menggunakan Operator Performance Classification untuk mengevaluasi model. Berikut adalah hasil prediksi performance student dengan menggunakan metode multiclass SVM dengan Parameter Grid :

D. Akurasi

Hasil akurasi SVM dengan Parameter Grid Data Training 99.64% Data Testing 100%, berikut adalah tabel hasil akurasi metode SVM dan SVM dengan Parameter Grid pada Tabel II dibawah ini.

Tabel II. Hasil Akurasi Data Training dan Data Testing

Metode	**Data Training**	**Data Testing**
SVM	90.22%	92.44%
SVM dengan Parameter Grid	99.64%	100%

E. Precision

Digunakan untuk mengukur proporsi rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positf. Hasil perbandingan Precision disajikan pada Tabel III dibawah ini.

Tabel III. Hasil Presisi Data Training dan Testing

Metode	**Data Training**	**Data Testing**
Pred. P	100%	100%
Pred. A	99.26%	100%
Pred. G	100%	100%

F. Recall

Digunakan untuk menunjukkan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Hasil perbandingan Recall disajikan pada Tabel IV dibawah ini.

Tabel IV. Hasil Recall Data Training dan Testing.

Metode	Data Training	Data Testing
True. P	100%	100%
True. A	100%	100%
True. G	98.04%	100%

Kesimpulan

Dalam penelitian ini dapat disimpulkan bahwa prediksi performance student adalah tantangan utama analisis prediksi karena kumpulan data yang kompleks. Pendekatan klasifikasi SVM diterapkan dalam pekerjaan penelitian sebelumnya untuk prediksi kinerja siswa. Untuk meningkatkan akurasi prediksi kinerja siswa, pengklasifikasi parameter grid diterapkan di pekerjaan penelitian ini. Model yang diusulkan diimplementasikan dalam rapidminer dan hasilnya dianalisis dalam hal akurasi, presisi dan recall. Dianalisis bahwa model yang diusulkan memiliki akurasi yang tinggi dan waktu eksekusi yang cukup lama.

Referensi

[1] P. Cortez, A. M. G. Silva, "Using data mining to predict secondary school student performance", in Proceedings of 5th Future Business Technology Conference, 2008, pp. 5-12.

[2] Yossy, Emny Harna. “Comparasion of Data Mining Classification Algorithms for Student Performance”, IEEE International Conference on Engineering, Technology and Education (TALE), 2019.

[3] Hartatik. “Prediction of Student Graduation with Naive Bayes Algorithm”, Fifth International Conference on Informatics and Computing (ICIC), 2020.

[4] Kumar, A.Dinesh. “Hybrid Classification Algorithms for Predicting Student Performance”, International Conference on Artificial Intelligence and Smart Systems (ICAIS) ISBN: 978-1-7281-9537-7, 2021.

[5] Wirawan, Chandra. “Application of Data mining to Prediction of Timeliness Graduation of Students (A Case Study)”, 7th International Conference on Cyber and IT Service Management (CITSM), 2019.

[6] Tripathi, Akarshita. “Naïve Bayes Classification Model for the Student Performance Prediction”, 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2019.

[7] Huang, Kendrew & Putra, Edi Purnomo. “*Support Vector Machine Algorithm*”, 2022. [Online]https://sis.binus.ac.id/2022/02/14/support-vector-machine-algorithm/. [Accessed: 25 May-2022]

[8] Syarif, Iwan. “SVM Parameter Optimization Using Grid Search and Genetic Algorithm to Improve Classification Performance”, TELKOMNIKA, Vol.14, No.4, December 2016, PP. 1502~1509 ISSN: 1693-6930, 2016.

[9] “UCI Machine Learning Repository:Student Performance Data Set”,2008.[Online].Available:https://archive.ics.uci.edu/ml/datasets/student+performance. [Accessed: 14-Apr-2019].

[10] “Optimize Parameters (Grid)”, 2022. [Online]Available:https://docs.rapidminer.com/latest/studio/operators/modeling/optimization/parameters/optimize_parameters_grid.html. [Accessed: 1-June-2022]