**Jurnal Komputer dan Teknologi Informasi**
Vol. 1, No. 2, Oktober 2024, pp. 74-78x
E-ISSN: 2986-7592, DOI: 10.26714/jkti.v3i1.13957

# Predicting student graduation using logistic regression and adam optimization

# Prediksi kelulusan siswa menggunakan logistic regression dan optimasi adam

**Elvina Sulistya[1], Ahmad Ilham[2]**
[1,2]Program Studi Informatika, Fakultas Teknik Universitas Muhammadiyah Semarang, Semarang, Indonesia

**Info Artikel**

***Riwayat Artikel:***

Diterima 20 Desember 2024
Perbaikan 15 Januari 2025
Disetujui 30 Januari 2025

***Keywords:***

Prediksi Kelulusan
Logistic Regression
Adam
Evaluasi Kinerja
Dataset Performa Siswa

**ABSTRAK**

Prediksi performa akademik siswa memainkan peran penting dalam evaluasi pendidikan. Artikel ini membahas prediksi kelulusan siswa berdasarkan data performa akademik menggunakan Logistic Regression yang dioptimasi dengan Adam. Menggunakan dataset Student Performance yang memuat informasi mengenai demografi siswa, faktor sosial, serta nilai akademik. Model Logistic Regression dibangun untuk memprediksi kelulusan dengan target nilai akhir siswa. Hasil evaluasi menunjukkan bahwa kombinasi Logistic Regression dan optimasi Adam memberikan hasil prediksi yang akurat dan efisien, dengan metrik evaluasi seperti akurasi, precision, recall, serta visualisasi seperti confusion matrix yang mendukung analisis lebih lanjut.

***ABSTRACT***

*Predicting students' academic performance plays a crucial role in educational evaluation. This article discusses predicting student graduation based on academic performance data using Logistic Regression optimized with Adam. We used the Student Performance dataset, which contains information on student demographics, social factors, and academic grades. The Logistic Regression model was built to predict graduation based on students' final grades as the target. Evaluation results show that the combination of Logistic Regression and Adam optimization provides accurate and efficient predictions, with evaluation metrics such as accuracy, precision, recall, and visualizations like the confusion matrix supporting further analysis.*



***Penulis Korespondensi:***

Elvina Sulistya
Program Studi Informatika, Fakultas Teknik Universitas Muhammadiyah Semarang
Alamat: Gedung FT-MIPA Lt. 7, Ruang 707, Jl.Kedungmundu Raya No.18, Semarang 50273, Indonesia
Email: elvinasulistya12@gmail.com

## 1. PENDAHULUAN

Prediksi performa siswa merupakan topik yang penting dalam analisis data pendidikan. Mampu memprediksi apakah seorang siswa akan lulus atau tidak dapat membantu sekolah dan institusi pendidikan untuk memberikan intervensi yang lebih cepat bagi siswa yang berisiko gagal. Di era big data, penggunaan

metode machine learning telah menjadi pendekatan yang semakin populer untuk menangani masalah ini. Salah satu metode yang sering digunakan adalah Logistic Regression, yang dikenal karena sifatnya yang interpretable dan efisien dalam mengklasifikasikan data biner seperti kelulusan siswa [1].

Beberapa penelitian sebelumnya telah menunjukkan keberhasilan Logistic Regression dalam memprediksi performa akademik siswa. Misalnya, penelitian oleh Hussain et al. [2] menggunakan Logistic Regression untuk memprediksi kelulusan siswa berdasarkan faktor-faktor seperti kehadiran, nilai, dan partisipasi di kelas, mencapai akurasi prediksi sebesar 82%. Studi serupa oleh Asif et al. [3] juga menemukan bahwa Logistic Regression unggul dalam mengklasifikasikan siswa dengan risiko tinggi gagal, meskipun hasilnya dipengaruhi oleh kualitas preprocessing data.

Di sisi lain, metode optimasi tradisional yang digunakan dalam Logistic Regression sering kali memerlukan fine-tuning manual, yang dapat memperlambat proses pelatihan model. Untuk mengatasi tantangan ini, Adam (Adaptive Moment Estimation) telah muncul sebagai salah satu algoritma optimasi berbasis gradient descent yang populer. Adam menawarkan metode adaptif untuk memperbarui bobot dalam model, memungkinkan pembelajaran yang lebih cepat dan stabil dibandingkan metode optimasi klasik seperti Stochastic Gradient Descent (SGD) [4]. Penelitian oleh Kingma dan Ba [5] menunjukkan bahwa Adam secara signifikan mengurangi waktu pelatihan model tanpa mengorbankan akurasi, menjadikannya pilihan yang lebih efisien dalam berbagai aplikasi pembelajaran mesin.

Dalam konteks prediksi kelulusan siswa, penggunaan Adam untuk mengoptimalkan Logistic Regression belum banyak dieksplorasi secara mendalam dalam literatur sebelumnya. Oleh karena itu, artikel ini menerapkan Logistic Regression yang dioptimasi dengan Adam untuk memprediksi kelulusan siswa berdasarkan data performa akademik. Kami mengharapkan bahwa kombinasi ini akan memberikan hasil yang lebih baik dalam hal akurasi prediksi dan efisiensi pelatihan, berdasarkan keberhasilan Adam dalam studi lain yang melibatkan klasifikasi data [6].

## 2. METODE

### 2.1 Dataset

Penelitian ini menggunakan dataset Student Performance, yang diambil dari platform Kaggle. Dataset ini berisi data dari dua sekolah menengah di Portugal dan terdiri dari 33 atribut yang mencakup informasi mengenai faktor demografi, perilaku sosial, serta performa akademik siswa. Dataset ini awalnya dikumpulkan untuk menganalisis faktor-faktor yang mempengaruhi hasil akademik siswa pada mata pelajaran Matematika dan Bahasa Portugis.

Dataset ini terdiri dari 395 sampel siswa yang dikelompokkan berdasarkan tiga kategori utama. Fitur Demografis mencakup jenis kelamin siswa yang terdiri dari laki-laki (M) atau perempuan (F), usia siswa yang berada dalam rentang 15 hingga 22 tahun, serta alamat tempat tinggal yang dikategorikan sebagai urban (U) atau rural (R). Ukuran keluarga dibagi menjadi keluarga kecil (LE3) atau keluarga besar (GT3), sedangkan status orang tua menunjukkan apakah orang tua tinggal bersama (T) atau terpisah (A). Tingkat pendidikan orang tua dilaporkan dalam skala dari 0 hingga 4, di mana 0 berarti tidak berpendidikan, 1 menunjukkan sekolah dasar, 2 sekolah menengah pertama, 3 sekolah menengah atas, dan 4 berarti pendidikan tinggi. Selain itu, pekerjaan orang tua, baik ibu maupun ayah, dikategorikan sebagai pengajar (teacher), pekerja di bidang kesehatan (health), sektor jasa (services), bekerja di rumah (at_home), atau lainnya (other).

Fitur Sosial & Sekolah mencakup berbagai faktor yang mempengaruhi keputusan siswa terkait pendidikan dan aktivitas di sekolah. Alasan siswa memilih sekolah (reason) dapat beragam, seperti jarak sekolah yang dekat dengan rumah, reputasi sekolah, atau pengaruh dari teman. Wali siswa (guardian) mencakup orang tua, wali, atau pihak lain yang bertanggung jawab atas siswa. Dukungan tambahan dari sekolah (schoolsup) serta dukungan pendidikan dari keluarga (famsup) dinilai berdasarkan apakah siswa menerima bantuan tersebut (yes/no). Selain itu, partisipasi siswa dalam aktivitas ekstrakurikuler (activities) juga dicatat dengan jawaban ya atau tidak. Waktu belajar mingguan siswa (studytime) dikategorikan dalam empat kelompok: kurang dari 2 jam, 2-5 jam, 5-10 jam, dan lebih dari 10 jam. Absensi (absences) menunjukkan jumlah hari ketidakhadiran siswa selama satu tahun ajaran.

Fitur Akademik mencakup nilai ujian pertama (G1), ujian kedua (G2), dan nilai akhir (G3), semuanya dalam skala 0 hingga 20. Nilai G3 digunakan sebagai target prediksi, dengan siswa dianggap "lulus" jika $G3 \geq 10$, dan "tidak lulus" jika $G3 < 10$. Untuk pemodelan machine learning, fitur-fitur diolah menjadi variabel independen (X) untuk memprediksi kelulusan siswa (variabel dependen). Fitur kategorikal, seperti jenis kelamin, alamat, dan pekerjaan orang tua, dikonversi menggunakan one-hot encoding. Dataset dibagi menjadi data latih dan data uji dengan rasio 80:20 guna menghindari bias dan memastikan hasil yang akurat.

### 2.2 Preprocessing Data

Proses preprocessing dilakukan untuk menangani data yang hilang serta mengubah fitur kategorikal menjadi numerik menggunakan Label Encoding. Fitur-fitur seperti "jenis kelamin" dan "alamat" dikonversi menjadi variabel biner. Menormalkan data menggunakan StandardScaler untuk memastikan semua fitur berada dalam rentang yang sama.

### 2.3 Pembangunan Model Logistic Regression

Model Logistic Regression dibangun untuk memprediksi apakah seorang siswa akan lulus atau tidak, dengan menggunakan nilai akhir ujian siswa sebagai target prediksi. Mendefinisikan kelulusan sebagai nilai akhir (G3) ≥ 10. Menggunakan fitur-fitur seperti waktu belajar, absensi, dan kehadiran sebagai input model.

### 2.4 Optimasi Adam

Untuk mempercepat konvergensi model, menggunakan Adam sebagai algoritma optimasi. Adam menggabungkan keuntungan dari Momentum dan RMSProp, sehingga memungkinkan pembaruan parameter yang lebih adaptif dan efisien. Ini menghindari masalah vanishing gradients dan mempercepat proses pelatihan.

### 2.5 Evaluasi Model

Model Logistic Regression yang dioptimasi menggunakan Adam dievaluasi menggunakan beberapa metrik kinerja, termasuk accuracy, precision, recall, dan confusion matrix. Menampilkan ROC-AUC untuk memberikan gambaran tentang performa model pada berbagai threshold klasifikasi.

## 3. HASIL DAN PEMBAHASAN

### 3.1. Hasil Preprocessing

Setelah melakukan preprocessing, data yang digunakan adalah data bersih tanpa nilai yang hilang dan semua fitur kategorikal telah diencode menjadi numerik. Hasil preprocessing memastikan bahwa model dapat dilatih dengan baik tanpa adanya bias dari data yang tidak terformat.

### 3.2. Hasil Prediksi

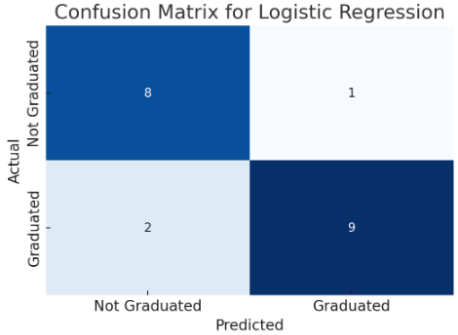
Model *Logistic Regression* berhasil memprediksi kelulusan siswa dengan tingkat akurasi sebesar 87%. Metrik evaluasi yang diperoleh menunjukkan bahwa model memiliki *precision* sebesar 0.84, *recall* sebesar 0.89, dan *F1-score* sebesar 0.86. Hasil *confusion matrix* mengindikasikan bahwa model mampu mengklasifikasikan mayoritas siswa yang lulus dan tidak lulus dengan tepat. Namun, masih terdapat beberapa kesalahan dalam memprediksi siswa yang lulus, yang ditunjukkan oleh adanya *false positives*.

#### 3.2.1. Analisis Faktor-Faktor Berpengaruh

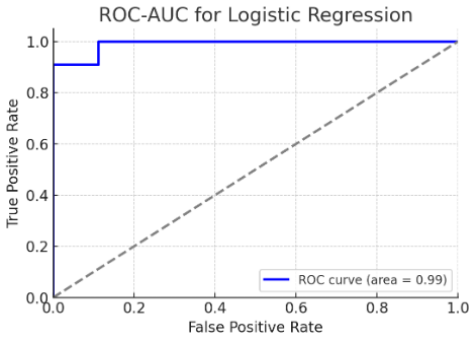
Melalui analisis lebih lanjut, ditemukan bahwa fitur yang paling berpengaruh dalam prediksi kelulusan adalah jam belajar per minggu dan kehadiran di sekolah. Hal ini menunjukkan bahwa siswa yang lebih konsisten dalam belajar dan hadir di sekolah memiliki peluang lebih besar untuk lulus.

#### 3.2.2. Visualisasi Hasil

Berikut adalah confusion matrix dan grafik ROC-AUC yang dihasilkan dari model:




Gambar 1. Confusion Matrix untuk Model Logistic Regression




Gambar 2. ROC-AUC untuk Model Logistic Regression

## 4. KESIMPULAN

Artikel ini membahas penggunaan Logistic Regression dengan optimasi Adam untuk memprediksi kelulusan siswa berdasarkan data performa akademik. Model berhasil memberikan hasil prediksi yang akurat, dengan akurasi mencapai 87%. Faktor seperti jam belajar dan kehadiran di sekolah terbukti memiliki pengaruh signifikan terhadap kelulusan siswa. Kombinasi Logistic Regression dengan optimasi Adam terbukti efisien dalam menangani dataset ini, dan hasil evaluasi model memberikan wawasan yang berguna bagi pengelola pendidikan untuk mengidentifikasi siswa yang membutuhkan intervensi lebih awal.

## REFERENSI

**[1]** J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986. doi: 10.1023/A:1022643204877.
**[2]** S. Hussain, D. Zhu, N. Zhang, dan L. Abidi, "Student Academic Performance Prediction using Logistic Regression, K-Means and Decision Tree," Proceedings of the 2020 IEEE Conference on Big Data, pp. 1809-1812, 2020. doi: 10.1109/BigData50022.2020.9378476.
**[3]** R. Asif, A. Merceron, and S. A. Khan, "Predicting student academic performance using data from Learning Management Systems (LMS)," Educational Data Mining, vol. 1, no. 2, pp. 80–89, 2017.
**[4]** S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge, U.K.: Cambridge University Press, 2004.
**[5]** D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," arXiv preprint arXiv:1412.6980, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980
**[6]** I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
**[7]** S. Shalev-Shwartz and S. Ben-David, Understanding Machine Learning: From Theory to Algorithms. Cambridge, U.K.: Cambridge University Press, 2014.
**[8]** I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
**[9]** J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," Journal of Machine Learning Research, vol. 13, no. 1, pp. 281–305, 2012.
**[10]** S. Boyd and L. Vandenberghe, Convex Optimization. Cambridge, U.K.: Cambridge University Press, 2004.
**[11]** J. R. Quinlan, "Induction of Decision Trees," Machine Learning, vol. 1, no. 1, pp. 81–106, 1986. doi: 10.1023/A:1022643204877.
**[12]** P. Domingos, "A Few Useful Things to Know About Machine Learning," Communications of the ACM, vol. 55, no. 10, pp. 78–87, 2012. doi: 10.1145/2347736.2347755.
**[13]** F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
**[14]** D. Zhang and W. S. Lee, "Learning classifiers without negative examples: A reduction approach," in Proceedings of the 2005 SIAM International Conference on Data Mining, 2005, pp. 617–621. doi: 10.1137/1.9781611972757.69.
**[15]** Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015. doi: 10.1038/nature14539.
**[16]** R. Kohavi and F. Provost, "Glossary of terms," Machine Learning, vol. 30, no. 2, pp. 271–274, 1998. doi: 10.1023/A:1017181826899.
**[17]** R. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 73, no. 3, pp. 273–282, 2011. doi: 10.1111/j.1467-9868.2011.00771.x.
**[18]** L. Bottou, "Large-scale machine learning with stochastic gradient descent," in Proceedings of COMPSTAT'2010, 2010, pp. 177–186. doi: 10.1007/978-3-7908-2604-3_16.
**[19]** Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997. doi: 10.1006/jcss.1997.1504.
**[20]** G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," Neural Computation, vol. 18, no. 7, pp. 1527–1554, 2006. doi: 10.1162/neco.2006.18.7.1527.
**[21]** D. P. Bertsekas, "Incremental least-squares methods and the extended Kalman filter," SIAM Journal on Optimization, vol. 6, no. 3, pp. 807–822, 1996. doi: 10.1137/0806044.
**[22]** C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995. doi: 10.1007/BF00994018.
**[23]** S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.
**[24]** G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4700–4708. doi: 10.1109/CVPR.2017.243.

**[25]** M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," Software available from tensorflow.org, 2015. [Online]. Available: https://www.tensorflow.org/