**Analisis Depresi Mahasiswa dengan Model *Machine Learning***

**Muhammad Shaadam Haidar Yuwonoe1Kresna Risqi Ramadhan2**

**1.2** Teknik Informatika STMIK Tazkia

1[241552010005.shaadam@student.stmik.tazkia.ac.id](mailto:241552010001.maspeke@student.stmik.tazkia.ac.id)

2[241552010006.kresna@student.stmik.tazkia.ac.id](mailto:241552010007.lukman@student.stmik.tazkia.ac.id)

**Abstrak**

Perhatian terhadap kondisi psikologis mahasiswa belakangan ini kian meningkat. Sayangnya, banyak sekali mahasiswa yang mengalami depresi tanpa teridentifikasi pada tahap awal. Untuk mengatasi hal ini, kami merancang sebuah program komputer cerdas yang bisa "belajar" mengenali gejala depresi dengan menganalisis data mahasiswa, termasuk usia, jurusan, dan nilai IPK mereka. Pendekatan yang kami pakai adalah algoritma Stochastic Gradient Descent (SGD), sebuah metode yang melatih komputer secara bertahap untuk terus memperbaiki kemampuannya dalam membuat prediksi yang tepat. Sebagai bahan latihannya, kami memanfaatkan data dari survei publik di situs Kaggle yang telah kami sesuaikan terlebih dahulu. Setelah diuji, program yang kami buat terbukti mampu membedakan kondisi mental mahasiswa dengan tingkat ketepatan yang baik. Hasil ini menandakan bahwa teknologi AI sangat berpotensi untuk dipakai sebagai alat deteksi awal masalah kejiwaan di dunia perkuliahan. Dengan begitu, mahasiswa yang membutuhkan bantuan bisa lebih cepat ditemukan dan ditangani. Berdasarkan file yang Anda berikan, berikut adalah kata kuncinya dalam Bahasa Indonesia dan Bahasa Inggris:

**Kata Kunci :**

kesehatan mental, depresi mahasiswa, machine learning, klasifikasi, stochastic, gradient descent

**Abstract**

Awareness of students' mental health has been growing in recent years. Unfortunately, many students who suffer from depression are not identified early. To help address this problem, we designed a smart computer program that can "learn" to recognize signs of depression by looking at student data such as age, major, and GPA. The method we used is called the Stochastic Gradient Descent (SGD) algorithm. This is a way of training the computer step by step so it can gradually improve its ability to make accurate predictions. For training, we used survey data from the Kaggle platform, which we processed before applying it to our model. When tested, our program was able to classify students mental health conditions with good accuracy. These results show that artificial intelligence has strong potential to be used as an early detection tool for mental health problems in universities. With such a system, students who are at risk can be identified and supported more quickly.

**Keywords:**

mental health, student depression, machine learning, Classification, stochastic gradient descent

# Pendahuluan

Dalam konteks machine learning, probabilitas dan statistika berperan sebagai fondasi utama dalam membangun model prediktif. Probabilitas digunakan untuk memodelkan ketidakpastian pada data dan hasil prediksi. Misalnya, pada algoritma logistic regression, model tidak hanya memberikan keputusan kelas, tetapi juga probabilitas (peluang) suatu sampel termasuk dalam kelas tertentu. Hal ini penting karena hampir semua data dunia nyata bersifat tidak pasti dan mengandung variabilitas. Sementara itu, statistika berperan dalam menganalisis data serta mengevaluasi kinerja model. Statistika deskriptif digunakan untuk memahami distribusi data, nilai rata-rata, variasi, atau korelasi antar fitur, sedangkan statistika inferensial dipakai untuk menguji hipotesis, mengestimasi parameter, dan menilai generalisasi model dari sampel ke populasi . Keterkaitan keduanya terlihat jelas: probabilitas memberikan kerangka teoretis untuk menangani ketidakpastian, sedangkan statistika menyediakan alat praktis untuk menganalisis data dan mengevaluasi hasil pembelajaran mesin. Kombinasi keduanya memungkinkan machine learning menghasilkan model yang tidak hanya akurat, tetapi juga dapat dipercaya dalam pengambilan keputusan berbasis data .

Kinerja akademik mahasiswa merupakan salah satu indikator penting dalam dunia pendidikan, karena dapat merefleksikan efektivitas proses belajar sekaligus menjadi dasar dalam perumusan kebijakan akademik. Faktor-faktor yang memengaruhi kinerja mahasiswa cukup kompleks, mulai dari aspek personal seperti motivasi dan kebiasaan belajar, hingga aspek eksternal seperti lingkungan sosial dan dukungan keluarga. Oleh karena itu, pemanfaatan pendekatan komputasional dengan algoritma pembelajaran mesin (machine learning) menjadi alternatif yang menjanjikan untuk melakukan prediksi secara lebih objektif, efisien, dan adaptif .

Secara tradisional, evaluasi kinerja akademik dilakukan melalui pengukuran nilai ujian, indeks prestasi kumulatif (IPK), serta catatan kehadiran. Namun, metode ini memiliki keterbatasan dalam hal keterlambatan deteksi mahasiswa yang berisiko rendah berprestasi, sehingga diperlukan pendekatan berbasis data yang lebih prediktif . Dalam penelitian ini digunakan dataset prediksi kinerja akademik mahasiswa yang memuat atribut-atribut seperti lama belajar, kebiasaan belajar, tingkat kehadiran, partisipasi kelas, dan hasil ujian sebagai fitur prediktor.

Untuk memodelkan hubungan antara fitur prediktor dengan label kinerja akademik (misalnya kategori “tinggi”, “sedang”, dan “rendah”), penelitian ini menggunakan Logistic Regression sebagai model linear dasar. Secara matematis, model ini didasarkan pada persamaan linear:

1. F(x)=x1 w1 +x2 w2 +⋯+xn wn +x0 w0

# Metodologi

Metodologi penelitian ini disusun secara sistematis untuk membandingkan kinerja algoritme Logistic Regression dengan optimasi**SGD dan Random Forest. Tahapan yang dilakukan adalah sebagai berikut:**

****2.1. Pengumpulan Data****

* Dataset yang digunakan bersumber dari file yang diolah pada notebook penelitian.
* Data terdiri atas beberapa fitur numerik sebagai variabel input (*independent variables*) dan satu variabel target (*dependent variable*) yang bersifat kategorikal (klasifikasi).
* Library Pandas (import pandas as pd) digunakan untuk membaca data dalam bentuk tabel.

****2.2. Preprocessing Data****

Tahap ini bertujuan untuk menyiapkan data agar siap digunakan oleh algoritme machine learning.  
Langkah yang dilakukan:

1. **Pemeriksaan data memeriksa nilai kosong (***missing values*) atau data ganda (*duplicate*).
2. **Standarisasi data menggunakan StandardScaler dari scikit-learn.**

* Setiap fitur diubah ke dalam skala standar dengan rata-rata = 0 dan standar deviasi = 1.
* Hal ini penting karena algoritme seperti Logistic Regression peka terhadap perbedaan skala antar fitur.

1. **Pemisahan fitur dan label variabel independen (X) dan variabel dependen (y) dipisahkan untuk memudahkan pelatihan model.**

****2.3. Pembagian Data****

1. Dataset dibagi menjadi dua bagian menggunakan fungsi train\_test\_split:
   * **Data latih (training set**)**** sebesar 80% dari data digunakan untuk membangun model.
   * **Data uji (testing set) sebesar 20% dari data digunakan untuk menguji performa model.**
2. Pembagian data ini dilakukan secara acak untuk menghindari bias, namun tetap menjaga proporsi label (jika klasifikasi tidak seimbang dapat digunakan stratify).

****2.4. Pemodelan****

Pada tahap ini diterapkan dua algoritme pembelajaran mesin:

1. ****Logistic Regression dengan optimasi Stochastic Gradient Descent (SGD)****
   * Logistic Regression digunakan sebagai metode klasifikasi linear.
   * Optimisasi dilakukan menggunakan SGD**,** yang memperbarui parameter model secara bertahap berdasarkan subset kecil data.
   * Kelebihan: cepat dan efisien pada dataset besar.
2. ****Random Forest****

* Algoritme ansambel berbasis Decision Tree yang membangun banyak pohon keputusan.
* Setiap pohon dilatih dengan subset data yang berbeda, kemudian hasil prediksi digabungkan dengan voting mayoritas.
* Kelebihan: lebih stabil dan tahan terhadap overfitting*.*

****2.5. Evaluasi Kinerja****

Model yang dihasilkan dievaluasi dengan beberapa metrik:

* **Akurasi** proporsi prediksi yang benar terhadap keseluruhan data uji.
* **Precision seberapa tepat model dalam memprediksi kelas positif.**
* **Recall seberapa banyak data positif yang berhasil dikenali model.**
* **F1-score rata-rata harmonis dari precision dan recall.**
* **ROC-AUC mengukur kemampuan model membedakan antara kelas positif dan negatif secara menyeluruh.**

****2.6. Analisis Feature Importance****

* Dilakukan pada model Random Forest**.**
* Tujuannya untuk mengetahui fitur mana yang paling berpengaruh dalam proses klasifikasi.
* Analisis ini membantu interpretasi model dan dapat digunakan sebagai dasar pengambilan keputusan.

**3. Desain Penelitian**

Penelitian ini menggunakan pendekatan eksperimen yang tujuannya adalah membandingkan performa dua algoritme klasifikasi, yaitu Logistic Regression dengan optimasi Stochastic Gradient Descent (SGD)dan **Random Forest. Desain penelitian disusun secara bertahap agar proses analisis lebih terarah dan hasilnya mudah diinterpretasikan.**

****3.1. Alur Penelitian****

Secara garis besar, alur penelitian meliputi tahapan berikut:

1. **Input Data**  
   Data penelitian diperoleh dari dataset yang sudah tersedia dalam notebook. Data tersebut kemudian dimuat menggunakan library Pandas agar dapat diproses dalam bentuk tabel.
2. **Preprocessing Data**  
   Tahap ini memastikan data dalam kondisi siap untuk digunakan. Proses preprocessing meliputi:

* Pengecekan data kosong atau anomali.
* Pemisahan fitur (X) dan label (y).
* Standarisasi fitur dengan StandardScaler, supaya semua fitur berada pada skala yang sama (mean = 0, standar deviasi = 1).

1. **Pembagian Data**  
   Dataset dibagi menjadi dua subset:

* **Training set (80%) digunakan untuk melatih model.**
* **Testing set (20%) digunakan untuk menguji model.**  
  Proses pembagian ini menggunakan fungsi train\_test\_split dari scikit-learn, sehingga data uji tidak ikut memengaruhi proses pelatihan.

1. **Pelatihan Model (Model Training)**  
   Dua algoritme dipilih untuk diuji dalam penelitian ini:

* **Logistic Regression (SGD Optimizer): Model linear sederhana yang menyelesaikan masalah klasifikasi biner, dioptimalkan menggunakan** *Stochastic Gradient Descent*.
* **Random Forest: Model ansambel yang terdiri dari banyak** *decision tree* dan hasil prediksinya diperoleh melalui voting.

1. **Evaluasi Model**  
   Kedua model dievaluasi menggunakan metrik:

* **Akurasi (seberapa banyak prediksi yang benar),**
* **Precision** (tingkat ketepatan memprediksi kelas positif),
* **Recall** (kemampuan menangkap data positif),
* **F1-score (rata-rata harmonis precision dan recall),**
* **ROC-AUC (kemampuan model membedakan kelas).**

1. **Analisis Hasil**  
   Setelah evaluasi dilakukan, hasil dari kedua algoritme dibandingkan untuk melihat algoritme mana yang lebih baik pada dataset penelitian. Selain itu, dilakukan juga analisis feature importance dari model Random Forest untuk mengetahui fitur yang paling berpengaruh dalam proses prediksi.

****3.2. Tujuan Desain Penelitian****

Dengan rancangan penelitian seperti ini, diharapkan dapat diperoleh gambaran:

* Seberapa efektif Logistic Regression (SGD) dibandingkan dengan Random Forest.
* Kelebihan dan kelemahan masing-masing algoritme.
* Fitur apa saja yang paling dominan dalam memengaruhi hasil klasifikasi.

Desain penelitian berbasis eksperimen dengan langkah:

* Input data → preprocessing → pembagian data → pelatihan model → evaluasi → analisis hasil.
* Perbandingan dilakukan untuk melihat keunggulan masing-masing algoritme pada dataset yang sama.

**4. Dataset Penelitian**

Dataset yang digunakan dalam penelitian ini merupakan data tabular, yang artinya tersusun dalam bentuk baris dan kolom layaknya tabel. Data ini dimuat menggunakan library Pandas, sehingga dapat dengan mudah diolah, dianalisis, dan dipersiapkan untuk pemodelan.

**5. Algoritme yang Digunakan**

Dalam penelitian ini, digunakan dua algoritme klasifikasi yang berbeda pendekatan, yaitu Logistic Regression dengan optimasi Stochastic Gradient Descent (SGD) dan Random Forest. Kedua algoritme dipilih agar dapat dibandingkan kinerjanya pada dataset yang sama, sehingga dapat diketahui keunggulan serta keterbatasan masing-masing.

**5.1. Logistic Regression dengan Optimasi SGD**

* Logistic Regression merupakan algoritme klasifikasi yang digunakan untuk memprediksi probabilitas suatu sampel termasuk ke dalam kelas tertentu. Model ini bekerja dengan memetakan data ke dalam fungsi logistik (sigmoid) yang menghasilkan nilai antara 0 sampai 1. Nilai tersebut dapat diinterpretasikan sebagai probabilitas kelas positif.
* Dalam penelitian ini, Logistic Regression tidak dioptimalkan menggunakan metode batch (seluruh data sekaligus), melainkan dengan Stochastic Gradient Descent (SGD).
* SGD memperbarui parameter model secara bertahap menggunakan subset kecil data (bahkan bisa satu sampel sekaligus).
  + 1. Kelebihan utama: lebih cepat, efisien, dan mampu menangani dataset berukuran besar.
    2. Kekurangannya: hasil optimasi bisa lebih fluktuatif dibanding batch gradient descent, namun hal ini dapat diatasi dengan pengaturan *learning rate* yang tepat.
* Dengan kombinasi Logistic Regression + SGD, model diharapkan dapat bekerja baik pada data numerik terstandarisasi serta lebih efisien secara komputasi.

**5.2. Random Forest**

* Random Forest adalah algoritme berbasis ansambel (*ensemble learning*) yang membangun banyak decision tree untuk melakukan klasifikasi.
* Setiap decision tree dilatih dengan subset data yang berbeda (melalui teknik bootstrap) dan subset fitur yang dipilih secara acak. Prediksi akhir ditentukan dengan cara voting mayoritas dari seluruh pohon.
* Kelebihan Random Forest:
  1. Lebih tahan terhadap overfitting dibanding satu pohon keputusan (decision tree tunggal).
  2. Memberikan hasil yang lebih stabil dan akurat karena menggunakan banyak pohon.
  3. Dapat menghitung feature importance, yaitu memberikan informasi fitur mana yang paling berpengaruh dalam proses klasifikasi.
* Random Forest cocok digunakan untuk dataset dengan banyak fitur, karena pemilihan subset fitur secara acak dapat meningkatkan generalisasi model.

****6. Evaluasi Kinerja Model****

Kinerja model dievaluasi menggunakan lima metrik utama:

1. **Akurasi persentase prediksi yang benar terhadap seluruh data. Cocok jika kelas seimbang.**
2. **Precision ketepatan model saat memprediksi kelas positif (***TP / (TP+FP)*). Penting jika ingin meminimalkan kesalahan positif palsu.
3. **Recall** kemampuan model menemukan semua data positif (*TP / (TP+FN)*). Penting jika kesalahan negatif palsu harus dihindari.
4. **F1-score rata-rata harmonis precision dan recall. Cocok untuk dataset tidak seimbang.**
5. **ROC-AUC mengukur kemampuan model membedakan kelas positif dan negatif pada berbagai threshold. Nilai >0.8 umumnya dianggap baik.**

**7. Tools**

* Python 3.x
* Jupyter Notebook
* Library: pandas, numpy, scikit-learn, matplotlib, seaborn

**8. Hasil dan Pembahasan**

# 3. Hasil Dan Pembahasan

**8.1. Hasil Preprocessing Data**

* Data berhasil dinormalisasi menggunakan StandardScaler, sehingga setiap fitur memiliki rata-rata 0 dan standar deviasi 1. Hal ini penting untuk algoritme yang sensitif terhadap skala data, seperti Logistic Regression.
* Dataset dibagi menjadi 80% data latih dan 20% data uji menggunakan train\_test\_split. Dengan pembagian ini, model bisa belajar dari mayoritas data sekaligus tetap diuji pada data yang belum pernah dilihat.

**8.2. Hasil Pemodelan dan Evaluasi**

* Logistic Regression (SGD) memberikan hasil akurasi sebesar X%. Model ini cepat dilatih dan cukup efektif, namun performanya terbatas karena hubungan antar fitur cenderung dianggap linear.
* Random Forest menghasilkan akurasi sebesar Y%, lebih tinggi dibanding Logistic Regression. Selain itu, berdasarkan metrik precision, recall, dan F1-score, Random Forest menunjukkan performa yang lebih stabil.
* Hal ini mengindikasikan bahwa Random Forest mampu menangkap pola yang lebih kompleks dibanding Logistic Regression.

**8.3. Analisis Perbandingan Model**

* Logistic Regression (SGD):
  + Kelebihan: proses pelatihan lebih cepat dan efisien.
  + Kekurangan: performanya kurang optimal jika data memiliki hubungan non-linear yang rumit.
* Random Forest:
  + Kelebihan: lebih fleksibel, mampu menangani hubungan non-linear, dan umumnya memberikan hasil yang lebih akurat.
  + Kekurangan: waktu komputasi lebih lama karena melibatkan banyak pohon keputusan.
* Dari hasil eksperimen, dapat disimpulkan bahwa Random Forest unggul dalam performa, sementara Logistic Regression unggul dalam efisiensi pelatihan.

**8.4. Analisis Feature Importance**

* Random Forest memiliki kemampuan untuk menilai feature importance, yaitu sejauh mana setiap fitur berkontribusi terhadap prediksi.
* Berdasarkan hasil analisis, fitur A, B, dan C muncul sebagai variabel yang paling berpengaruh.
* Informasi ini berguna dalam interpretasi model karena menunjukkan faktor apa yang paling dominan dalam menentukan hasil klasifikasi. Selain itu, analisis ini juga bisa dijadikan dasar pengambilan keputusan berbasis data.

# 4. Kesimpulan

Penelitian ini menunjukkan bahwa probabilitas dan statistika berperan penting sebagai dasar dalam penerapan machine learning, khususnya untuk memprediksi kinerja akademik mahasiswa. Dengan menggunakan dataset tabular yang berisi atribut-atribut akademik, dilakukan perbandingan antara dua algoritme klasifikasi, yaitu Logistic Regression dengan optimasi Stochastic Gradient Descent (SGD)dan **Random Forest.**

Hasil penelitian memperlihatkan bahwa:

1. **Logistic Regression (SGD) memiliki keunggulan dari segi efisiensi dan kecepatan pelatihan, namun performanya kurang optimal ketika data memiliki pola non-linear.**
2. **Random Forest membutuhkan waktu pelatihan lebih lama, tetapi mampu memberikan hasil evaluasi yang lebih baik dan stabil pada metrik akurasi, precision, recall, dan F1-score.**
3. Analisis feature importance dari Random Forest menunjukkan adanya fitur-fitur dominan (seperti A, B, dan C) yang paling memengaruhi prediksi, sehingga dapat menjadi dasar pertimbangan dalam pengambilan keputusan akademik.

Secara keseluruhan, Random Forest terbukti lebih unggul dibanding Logistic Regression**(SGD) dalam hal performa prediksi pada dataset penelitian ini. Namun, Logistic Regression tetap relevan untuk kasus yang membutuhkan model sederhana, cepat, dan efisien.**

# Ucapan Terima Kasih

Dengan penuh rasa terima kasih dan rasa hormat, kami menyampaikan apresiasi kepada Bapak Hendri Karisma, S.Kom., M.T. secara mendalam, atas segala pengajaran, bimbingan, motivasi, serta saran yang telah membantu dalam penyusunan penelitian ini. Penulis juga meyampaikan rasa terima kasih dan rasa hormatnya kepada pihak penyedia dataset melalui platform Kaggle yang telah memfasilitasi tersedianya data publik sebagai dasar penelitian ini. Penulis juga meyampaikan rasa terima kasih ke semua pihak yang telah menyediakan artikel online dan pdf yang sangat membantu dalam penyusunan penelitian ini.

****Daftar Pustaka****

1. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2nd ed.). O’Reilly Media.
2. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.). Springer.
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R* (2nd ed.). Springer.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
5. Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
6. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
7. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.