**SUPERVISED LEARNING**

**(CLASSIFICATION)**

Mata Kuliah Big Data Analitik

Dosen Pengampu : Fathoni Maharadika, S.Kom, M. T.



Kelompok 3

Ilham Budi Kristiawan 220660121053

Nayaka Sahasika 220660121121

Herman Sanjaya 220660121119

Muhammad Farid Abdillah 220660121082

Muhammad Reza Fadlillah 230660121146

**PRODI INFORMATIKA**

**FAKULTAS TEKNOLOGI INFORMASI**

**UNIVERSITAS SEBELAS APRIL SUMEDANG**

# **DAFTAR ISI**

[BAB 1 3](#_Toc168864762)

[PENDAHULUAN 3](#_Toc168864763)

[**1.1** **Latar Belakang** 3](#_Toc168864764)

[**1.2** **Tujuan** 3](#_Toc168864765)

[**1.3 Ruang Lingkup** 3](#_Toc168864766)

[**1.4 Manfaat** 4](#_Toc168864767)

[BAB 2 4](#_Toc168864768)

[PEMBAHASAN 4](#_Toc168864769)

[**Teori Supervised Learning Classification** 4](#_Toc168864770)

[**2.1 Definisi** 4](#_Toc168864771)

[**2.2 Algoritma Umum** 4](#_Toc168864772)

[**2.3 Metrik Evaluasi** 5](#_Toc168864773)

[**Metode Supervised Learning Classification** 5](#_Toc168864774)

[**3.1 Persiapan Data** 5](#_Toc168864775)

[**3.2 Pemilihan Algoritma** 5](#_Toc168864776)

[**3.3 Pelatihan Model** 5](#_Toc168864777)

[**3.4 Evaluasi Model** 5](#_Toc168864778)

[**Hasil Supervised Learning Classification** 5](#_Toc168864779)

[**4.1 Interpretasi Model** 5](#_Toc168864780)

[**4.2 Prediksi pada Data Baru** 6](#_Toc168864781)

[**4.3 Evaluasi Kinerja** 6](#_Toc168864782)

[BAB 3 6](#_Toc168864783)

[PENUTUP 6](#_Toc168864784)

[**Kesimpulan** 6](#_Toc168864785)

[DAFTAR PUSTAKA 7](#_Toc168864786)

# **KATA PENGANTAR**

Puji syukur kehadirat Allah SWT dengan segala kerahmatan nya kami masih diberikan kesehatan serta kesempatan dalam menyusun tugas makalah dengan materi machine learning. Dalam hal ini kami sangat puas dengan hasil dan bahan yang sudah kami siapkan sebelumnya. Makalah ini membahas tentang materi Supervised Learning (classification) guna memenuhi tugas mata kuliah Big Data Analitik.

Kami ucapkan terima kasih Bapak Fathoni Mahardika, S. Kom., M.T, selaku dosen Matakuliah Big Data Analitik yang telah memberikan tema yang kami dapatkan. Kami menyadari bahwa makalah kami masih jauh dari sempurna, oleh karena itu kritik dan saran dari semua pihak yang bersifat membangun selalu kami harapkan demi kesempurnaan makalah ini. Akhir kata, kami sampaikan terima kasih atas semua rekan yang telah berperan serta dalam penyusunan makalah ini dari awal sampai akhir. Semoga Allah SWT selalu meridhoi segala usaha kita. Amin.

**Sumedang, 4 Juli 2024**

**Penulis**

# **BAB 1**

# **PENDAHULUAN**

## **Latar Belakang**

Penyebaran informasi dalam bentuk dokumen digital telah berkembang dengan pesat dan setiap waktu terus mengalami pertumbuhan dan jumlahnya semakin besar. Media massa versi elektronik dan situs web di internet merupakan dua contoh media yang menggunakan dan menyebarkan informasi berbentuk dokumen digital. Mengelola informasi dari kumpulan dokumen teks yang jumlahnya sangat besar tentunya bukan pekerjaan yang mudah. Oleh karena itu diperlukan sebuah metode yang dapat mengorganisir dan mengklasifikasi dokumen secara otomatis, sehingga dapat mempermudah dalam pencarian informasi yang relevan dengan kebutuhan (Samodra, Sumpeno, & Hariadi, 2009).

Bidang yang mempelajari teknik-teknik untuk pengorganisasian dokumen teks secara umum dibagi menjadi dua kelompok, yaitu classification dan clustering. Clustering teks berhubungan dengan menemukan sebuah struktur kelompok yang belum kelihatan (tak terpandu atau unsupervised) dari sekumpulan dokumen. Sedangkan pengklasifikasian teks dapat dianggap sebagai proses untuk membentuk golongan-golongan (kelas-kelas) dari dokumen berdasarkan pada kelas kelompok yang sudah diketahui sebelumnya (terpandu atau supervised) (Darujati et al., 2012).

Dalam penelitiannya, Trisedya (2009) menjelaskan bahwa teknik klasifikasi dapat dilakukan dengan dua cara yaitu dengan pendekatan supervised learning dan pendekatan unsupervised learning. Teknik yang banyak digunakan dalam unsupervised learning adalah teknik clustering. Clustering merupakan teknik mengelompokkan dokumen-dokumen, sehingga dokumen yang memiliki kemiripan dikumpulkan dalam sebuah cluster tertentu. Pendekatan kedua adalah supervised learning. Pendekatan ini dilakukan dengan membangun sebuah classifier dari proses pembelajaran mengenai ciri dari tiap-tiap kategori yang ada. Pendekatan supervised learning dapat dibagi menjadi fully supervised learning dan semi supervised learning. Fully supervised learning adalah teknik klasifikasi dimana semua dokumen training telah diketahui kategorinya. Naïve Bayes adalah contoh dari teknik fully supervised learning, sedangkan semi supervised learning adalah teknik klasifikasi dimana pembelajaran dilakukan dari dokumen training yang telah diketahui kategorinya dan dokumen training yang belum diketahui kategorinya.

Berdasarkan penelitian di atas, untuk mempermudah pencarian informasi yang sesuai dengan yang kita inginkan dan sesuai dengan kategorinya, maka pengklasifikasifikasian dokumen akan membantu bagaimana mendapatkan informasi, sehingga mempermudahan pengolahan dan penggunaannya sesuai kebutuhan dan tujuan yang ingin dicapai. Selain itu, hal yang harus diperhatikan adalah bagaimana cara melakukan klasifikasi dokumen medis saat data yang ada terdiri dari dua jenis dokumen yang berbeda yaitu dokumen berlabel dan dokumen tidak berlabel. Selain itu, labeled documents (dokumen berlabel) hanya tersedia dalam jumlah yang kecil. Permasalahan dokumen pembelajaran untuk melakukan klasifikasi dokumen ini dapat diatasi dengan pendekatan baru yang dapat mempelajari labeled data maupun unlabeled data walaupun labeled data hanya tersedia dalam jumlah yang kecil. Pendekatan ini dikenal dengan nama pendekatan semi supervised learning.

Teknik semi-supervised learning adalah metode yang efisien untuk menambah data training secara otomatis dari data yang tidak berlabel (unlabeled data). Selain itu, perkembangan dari banyak aplikasi pengolahan bahasa (natural language app) menganggap masalah ini adalah sebuah tantangan dimana data yang tidak berlabel (unlabeled data) relative dalam jumlah yang berlimpah sedangkan data berlabel (labeled data) jumlahnya agak terbatas (Qiu, Cho, Ma, & Campbell, 2019). Berbeda dengan pendekatan supervised learning, Teknik semi-supervised learning dapat meningkatkan kinerjanya dengan meningkatkan informasi dalam data yang tidak berlabel. Beberapa hasil terbaru dari Laine & Aila (2017); Miyato et al (2019); Tarvainen & Valpola (2017) menunjukkan bahwa teknik semi-supervised learning dapat mencapai kinerja dari teknik supervised learning dalam skenario tertentu.

Pada penelitiannya, (Andini, 2013) menjelaskan bahwa saat ini sulit untuk mengetahui dokumen berdasarkan kebutuhan. Oleh karena itu, untuk mengetahui dokumen berdasarkan kebutuhan perlu dibantu oleh klasifikasi dokumen teks, yaitu suatu proses pengelompokan dokumen ke kategori yang dapat digunakan untuk melakukan analisis.

Di era big data saat ini, supervised learning classification menjadi alat penting untuk mengolah dan menganalisis data dalam berbagai bidang. Teknik ini memungkinkan kita untuk membangun model prediktif yang dapat mengklasifikasikan data dengan akurasi tinggi, membuka peluang baru untuk pengambilan keputusan yang lebih cerdas dan efisien.

## **Rumusan Masalah**

Berdasarkan latar belakang yang telah diuraikan sebelumnya, maka diperoleh sebuah rumusan masalah, yaitu :

1. Apakah Pengertian Supervised Learning?
2. Apa saja jenis – jenis algoritma pada Supervised Learning?
3. Bagaimana metode Supervised Learning dilakukan?

## **Tujuan**

Makalah ini bertujuan untuk membahas dasar-dasar supervised learning classification, termasuk algoritma umum, aplikasi, dan manfaatnya.

## **Ruang Lingkup**

Makalah ini akan fokus pada supervised learning classification untuk tugas klasifikasi biner, di mana data diklasifikasikan ke dalam dua kategori.

## **Manfaat**

Memahami supervised learning classification dapat membantu Anda:

1. Menerapkan teknik ini untuk menyelesaikan masalah klasifikasi dalam berbagai bidang.
2. Memilih algoritma classification yang tepat untuk aplikasi Anda.
3. Menginterpretasikan hasil model classification.
4. Mengevaluasi kinerja model classification.

# **BAB 2**

# **PEMBAHASAN**

## **Landasan Teori Supervised Learning (Classification)**

## **2.1 Definisi**

Supervised learning classification adalah teknik machine learning yang digunakan untuk melatih model untuk mengklasifikasikan data baru ke dalam kategori yang telah ditentukan. Model ini dilatih menggunakan dataset yang berisi data berlabel, di mana setiap data point memiliki label yang menunjukkan kategorinya. Pada algoritma Supervised Learning, system diberikan training data set berupa informasi masukan dan keluaran yang diinginkan, sehingga sistem akan mempelajari berdasarkan data yang telah ada. Sistem akan mencari pola dari data set, kemudian pola itu akan dijadikan sebagai acuan untuk kumpulan data berikutnya. Machine leaning menjalankan algoritma Supervised Learning dengan melakukan pendekatan dari satu data ke data yang lain. Pada Bahasa pemrograman Python, pedekatan dilakukan dengan terlebih dahulu memecah data menjadi dua, yaitu 1) data yang akan dipelajari, dan 2) data yang akan dites terhadap data yang telah dipelajari. Supervised learning bersifat prediktif, sehingga dari banyak hasil pendekatan yang telah dilakukan akan didapatkan sebuah prediksi. Algoritma ini dapat menyelesaikan permasalahan data secara linear, multilinear, maupun polynomial.

Supervised Learning merupakan metode yang paling populer dalam implementasi algoritma untuk Machine Learning. Namun kelemahan dari metode iniadalah kita harus menyediakan atau menetapkan sebuah output yang benar. Kita mengajari mesin untuk melakukan sesuatu yang benar, tentu kita harus memiliki contoh apa sesuatu yang benar tersebut. Sebagai contoh, jika kita ingin mengajari sebuah program untuk menentukan sebuah opini yang negatif, positif, atau netral tentu kita harus menyediakan sampel atau contoh opini yang negatif, positif, atau netral itu seperti apa.

## **2.2 Algoritma Umum**

Beberapa algoritma supervised learning classification yang umum digunakan antara lain:

1. **Logistic Regression**: Algoritma ini menggunakan regresi linear untuk memprediksi probabilitas data milik suatu kelas.
2. **Support Vector Machines (SVM)**: Algoritma ini memisahkan data ke dalam kelas yang berbeda dengan mencari hyperplane yang memaksimalkan margin antara kelas.
3. **Decision Trees:** Algoritma ini membangun pohon keputusan berdasarkan aturan yang dapat diinterpretasikan untuk mengklasifikasikan data. Decision tree merepresentasikan sebuah tree dimana internal nodenya mengetes sebuah atribut, masing-masing cabangnya berhubungan dengan nilai dari atribut dan masing-masing lead nodenya berisi sebuah klasifikasi. Algoritma
4. **K-Nearest Neighbors (KNN):** Algoritma ini mengklasifikasikan data baru berdasarkan data terdekat di ruang fitur.

## **2.3 Metrik Evaluasi**

Kinerja model supervised learning classification dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score.

## **Metode Supervised Learning Classification**

## **3.1 Persiapan Data**

Sebelum melatih model classification, data perlu dipersiapkan dengan cara:

1. Pembersihan Data: Mengatasi missing values dan outlier.
2. Normalisasi Data: Menormalisasi data ke skala yang sama.
3. Feature Engineering: Menciptakan fitur baru yang dapat meningkatkan kinerja model.

## **3.2 Pemilihan Algoritma**

Algoritma classification yang tepat untuk aplikasi Anda tergantung pada beberapa faktor, seperti:

1. Jenis data
2. Ukuran dataset
3. Kinerja komputasi yang tersedia
4. Kebutuhan interpretasi

## **3.3 Pelatihan Model**

Model classification dilatih menggunakan dataset berlabel. Algoritma akan mempelajari pola dalam data dan membangun model yang dapat memprediksi label untuk data baru.

## **3.4 Evaluasi Model**

Kinerja model classification dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Model dengan kinerja terbaik dipilih untuk digunakan pada data baru.

## **Hasil Supervised Learning Classification**

## **4.1 Interpretasi Model**

Beberapa algoritma classification, seperti decision trees, dapat diinterpretasikan untuk memahami bagaimana model membuat keputusan.

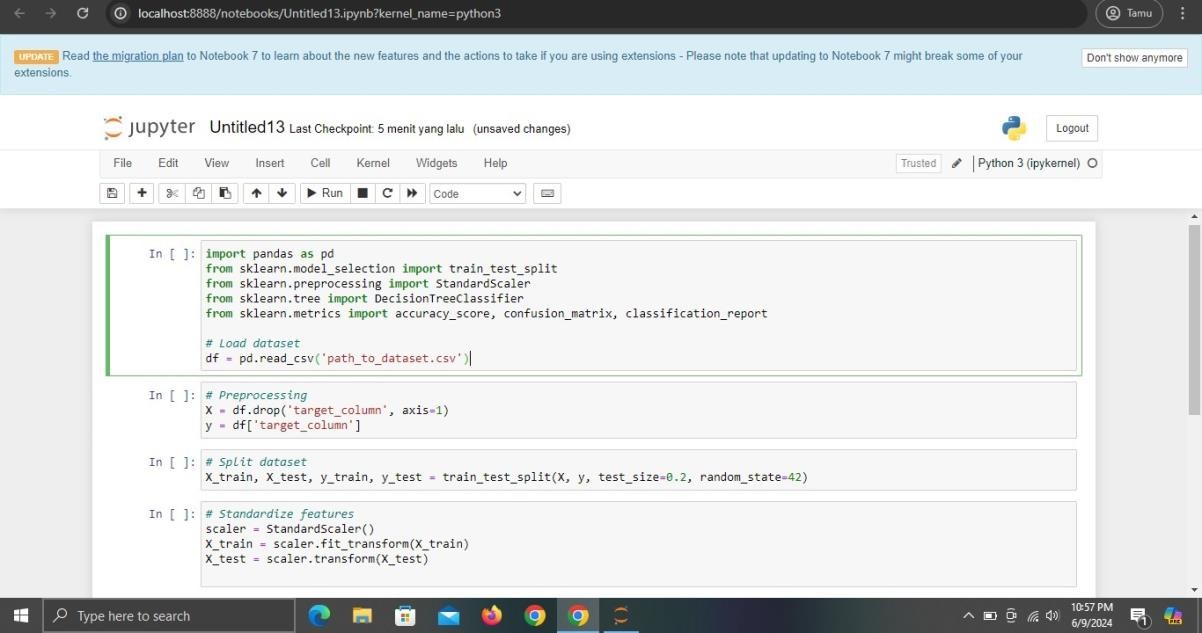
## **4.2 Prediksi pada Data Baru**

Model classification yang telah dilatih dapat digunakan untuk memprediksi label untuk data baru.

## **4.3 Evaluasi Kinerja**

Kinerja model classification pada data baru dievaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score.

## **Lampiran**



1. **Blok 1: Impor Modul dan Membaca Dataset**

* Baris pertama mengimpor modul **pandas** sebagai **pd**, serta berbagai fungsi dari **sklearn** (**scikit-learn**) untuk memisahkan dataset, skala fitur, membuat model klasifikasi, dan menghitung metrik evaluasi.
* **df = pd.read\_csv('path\_to\_dataset.csv')** membaca dataset dari file CSV yang lokasinya ditentukan oleh **'path\_to\_dataset.csv'.**

1. **Blok 2: Pra-pemrosesan Data**

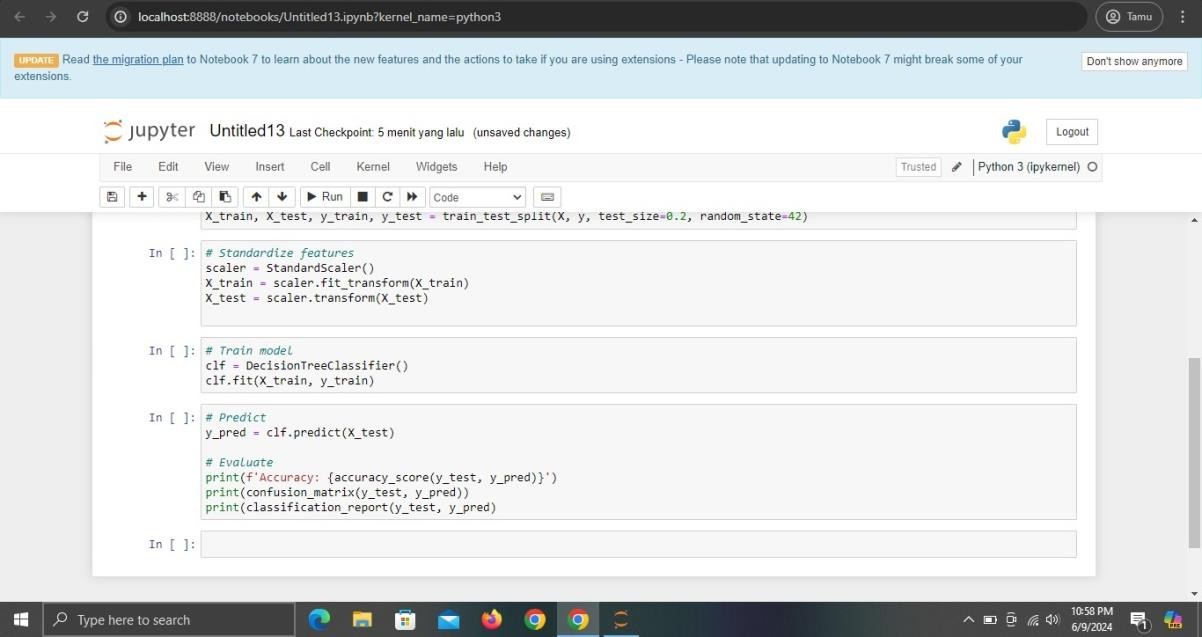
* **X** menyimpan fitur (semua kolom kecuali kolom target) dengan menghapus kolom **target\_column** dari **df**.
* **y** menyimpan kolom target dari **df**.

1. **Blok 3: Membagi Dataset**

* Dataset dibagi menjadi data pelatihan (**X\_train, y\_train**) dan data pengujian (**X\_test, y\_test**) dengan proporsi 80% pelatihan dan 20% pengujian. **random\_state=42** digunakan untuk memastikan hasil pembagian yang konsisten setiap kali kode dijalankan.

1. **Blok 4: Standarisasi Fitur**

* Data pelatihan dan pengujian di-standardisasi menggunakan **StandardScaler**, yang mengubah data sehingga memiliki rata-rata 0 dan standar deviasi 1. **scaler.fit\_transform(X\_train)** menghitung statistik yang diperlukan dari data pelatihan dan menerapkannya pada **X\_train**, sedangkan **scaler.transform(X\_test)** hanya menerapkan transformasi yang telah dihitung pada **X\_test.**



1. **Blok 4 (lanjutan): Standarisasi Fitur**

* Blok ini sama dengan yang sudah dijelaskan sebelumnya, di mana fitur data pelatihan dan pengujian di-standardisasi.

1. **Blok 5: Melatih Model**

* **clf = DecisionTreeClassifier()** membuat instance dari model Decision Tree Classifier.
* **clf.fit(X\_train, y\_train)** melatih model menggunakan data pelatihan **X\_train** dan label **y\_train.**

1. **Blok 6: Prediksi**

* **y\_pred = clf.predict(X\_test)** menggunakan model yang sudah dilatih untuk memprediksi label dari data pengujian **X\_test.**

1. **Blok 7: Evaluasi**

* **print(f'Accuracy: {accuracy\_score(y\_test, y\_pred)}')** mencetak akurasi dari model dengan membandingkan prediksi **y\_pred** dengan label sebenarnya **y\_test.**
* **print(confusion\_matrix(y\_test, y\_pred))** mencetak confusion matrix yang memberikan gambaran lebih rinci tentang performa klasifikasi.
* **print(classification\_report(y\_test, y\_pred))** mencetak laporan klasifikasi yang mencakup metrik seperti precision, recall, f1-score untuk setiap kelas.

# **BAB 3**

# **PENUTUP**

## **Kesimpulan**

Supervised learning classification merupakan teknik machine learning yang kuat dan serbaguna yang dapat digunakan untuk mengklasifikasikan data dengan akurasi tinggi. Algoritma classification memiliki berbagai aplikasi dalam berbagai bidang dan menawarkan banyak manfaat, seperti interpretasi model, skalabilitas, dan aplikasi luas. Dengan terus berkembangnya teknologi machine learning, supervised learning classification akan terus memainkan peran penting dalam pengolahan dan analisis data di masa depan.

# **DAFTAR PUSTAKA**

* ⁠James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning. Springer.
* ⁠Murphy, K. P. (2012). Machine learning: A probabilistic perspective. The MIT Press.
* Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. Springer.

Catatan

Makalah ini hanya memberikan gambaran umum tentang supervised learning classification. Untuk mempelajari lebih lanjut tentang topik ini, Anda disarankan untuk membaca buku, artikel, dan tutorial yang lebih mendalam. Anda dapat menambahkan sub-bab dan contoh spesifik untuk memperkaya mak