# Pendekatan *Supervised Learning* untuk Diagnosa Kehamilan

**FAHIRA1, ZIAN ASTI DWIYANTI, RONI HABIBI**

Universitas Logistik dan Bisnis Internasional, Indonesia

Email : 1204044@std.poltekpos.ac.id

## Abstrak

Dalam studi ini dilakukan evaluasi performa dari beberapa algoritma *machine learning* dalam mendiagnosis kehamilan. Tujuan dari studi ini adalah untuk menemukan algoritma yang paling efektif antara *decision tree* dengan *random forest* dalam mendiagnosis kehamilan. Studi ini juga bertujuan untuk memberikan pengetahuan yang lebih baik mengenai bagaimana teknik pembelajaran mesin dapat digunakan dalam proses diagnosa kehamilan dan memberikan informasi yang berguna bagi para dokter dan peneliti dalam membuat keputusan yang tepat dalam mendiagnosis kehamilan. Dari hasil evaluasi dapat disimpulkan bahwa model *Random Forest* dengan menggunakan dataset yang seimbang dan metode Gini memiliki akurasi terbaik sebesar 81%. Hal ini menunjukkan bahwa menggunakan dataset yang seimbang dapat meningkatkan performa dalam mendiagnosis kehamilan. Algoritma Random Forest merupakan metode yang sering digunakan dalam proses pengklasifikasian karena kinerjanya yang baik. Algoritma ini bekerja dengan membuat pohon keputusan yang digunakan untuk membuat prediksi. Pada studi ini, kami menggunakan algoritma Random Forest dengan metode Gini untuk melakukan prediksi kehamilan.

**Kata kunci**: kehamilan, diagnosa, *random forest*, *decision tree*, *machine learning*, gini, entropy

## *Abstract*

*This study evaluates the performance of several machine learning algorithms in diagnosing pregnancy. The purpose of this study is to find the most effective algorithm between decision trees and random forests in diagnosing pregnancy. This study also aims to provide better knowledge about how machine learning techniques can be used in the process of diagnosing pregnancy and provide useful information for doctors and researchers in making the right decisions in diagnosing pregnancy. From the evaluation results it can be concluded that the Random Forest model using a balanced dataset and the Gini method has the best accuracy of 81%. This shows that using a balanced dataset can improve performance in diagnosing pregnancy. Random Forest Algorithm is a method that is often used in the classification process because of its good performance. This algorithm works by creating a decision tree that is used to make predictions. In this study, we used the Random Forest algorithm with the Gini method to predict pregnancy.*

***Keywords****: pregnancy, diagnosis, random forest, decision tree, machine learning, gini, entropy*

## 1. PENDAHULUAN

Data merupakan sekumpulan fakta atau deskripsi tentang sesuatu yang diperoleh melalui observasi atau mencari sumber tertentu **(Helmi, 2021)**. Informasi yang dikumpulkan dapat berupa fakta maupun asumsi. Data merupakan kumpulan informasi yang diperoleh dari sumber-sumber tertentu yang belum diolah lebih lanjut. Namun setelah melalui proses analisis melalui penelitian atau eksperimen, data tersebut dapat menjadi lebih kompleks dan bermanfaat sebagai *database*, informasi, atau solusi untuk menyelesaikan masalah tertentu**(Timotius, 2017)**.

Menurut data yang didapatkan, Indonesia merupakan negara dengan jumlah penduduk yang besar dan berkembang pesat. Menurut laporan Kompas.com pada tanggal 3 Februari 2022, jumlah penduduk di Indonesia terus bertambah setiap tahun **(Kompas.com, 2022)**. Kondisi pertumbuhan penduduk di Indonesia tentu saja membutuhkan perhatian yang serius untuk mengurangi jumlah kematian. Salah satu solusi yang dapat diterapkan untuk memudahkan ibu hamil dalam mengetahui kondisi kehamilannya adalah dengan memberikan akses yang mudah. Kehamilan adalah di mana janin berkembang di dalam rahim atau uterus wanita**(Indrawati, dkk, 2021)**.

Profesi baru yang dikenal sebagai *data science* telah muncul di era ini. *Data science* atau ilmu data adalah seperangkat prinsip fundamental yang mendukung dan memandu ekstraksi informasi dan pengetahuan dari data**(Muttaqin, 2022)**. Untuk mengelola data, kita dapat menggunakan teknik *machine learning* yang menggunakan data untuk membuat model dan Membuat pilihan dengan menggunakan model yang sudah dibentuk sebelumnya. *Machine learning* dapat diinstruksikan sesuai dengan cara belajarnya, sehingga dapat melakukan tugasnya. Perbedaan cara belajar dapat dibagi menjadi tiga jenis, yaitu *supervised learning, unsupervised learning dan reinfo*rcement learning**(Id, 2021)**.

Pada penelitian ini digunakan *supervised learning* untuk melakukan pengklasifikasian data. Klasifikasi (*Classification*) dalam *Supervised Learning* adalah saat kita dapat menetapkan label khusus untuk pengamatan**(Abijono, dkk, 2021)**. Contoh klasifikasi adalah memprediksi apakah gejala seseorang akan menyebabkan kanker payudara atau tidak, berdasarkan data sebelumnya yang dikumpulkan dalam gejala dan hasil yang diharapkan (Ya/Tidak). Pada dasarnya, ketika proses pembelajaran dilakukan, *system* atau hasil yang diharapkan yaitu peluang untuk memecahkan masalah *system* tersebut **(Siswanto, 2013)**. Contoh algoritma yang termasuk dalam kategori *supervised learning* adalah Regresi Linear, Regresi Logistik, Analisis Diskriminan Linear, k-Nearest Neighbors, Support Verctor Machines (SVMs), Random Forest, Decision Tree, dan Naive Bayes.

Seperti yang telah dijelaskan, ada ratusan jenis metode atau algoritma yang dapat digunakan dalam *supervised learning*, salah satu algoritma yang digunakan untuk penelitian ini yaitu algoritma *decision tree* dan *random forest*.

*Decision Tree* merupakan salah satu algoritma yang paling bagus, karena berbagai fitur yang menarik, sederhana, lengkap, tanpa parameter, dan mampu menangani tipe data campuran. *Decision tree* merupakan teknik *supervised learning* yang dapat digunakan untuk menyelesaikan masalah klasifikasi maupun regresi, namun lebih sering digunakan untuk masalah klasifikasi**(Muzakir & Wulandari, 2016)**. *Decision tree* merupakan metode pembelajaran terawasi yang menggunakan struktur pohon yang mewakili kumpulan fitur data pada setiap simpul internalnya, cabang mewakili aturan tentang keputusan yang diambil dan setiap simpul daun akan menunjukkan hasil dari aturan tersebut**(Ginantra, dkk, 2021)**.

Proses pembuatan keputusan dalam *decision tree* dimulai dari simpul akar, dengan mengevaluasi nilai atribut yang ada pada data dan membandingkannya dengan atribut dari dataset yang akan diuji. Kemudian, algoritma akan mengikuti cabang yang sesuai dan melompat ke node berikutnya, sehingga akhirnya mencapai simpul daun yang menentukan kelas dari dataset tersebut.

Sedangkan, *Random Forest* adalah algoritma *machine learning* yang termasuk dalam teknik *supervised learning*  dan merupakan algoritma yang populer **(Riaddy, dkk, 2016)**. Dalam *machine learning* algoritma ini dapat menyelesaikan masalah dalam regresi dan klasifikasi. Konsep ini didasarkan pada pembelajaran *ensemble*, yang merupakan proses menggabungkan beberapa klasifikasi untuk meningkatkan kinerja model dan menyelesaikan masalah yang kompleks **(Rachmi, 2020)**. *Random Forest* ini merupakan algoritma klasifikasi yang terdiri dari banyak pohon keputusan yang dikelompokkan dalam satu kumpulan data yang berbeda-beda dan diambil rata-ratanya untuk meningkatkan akurasi prediksi pada kumpulan data tersebut.

Bahasa pemrograman yang digunakan adalah *Python*. Jadi, *python* adalah sebuah bahasa tingkat tinggi yang difokuskan pada kesederhanaannya yang mudah dipelajari. Struktur data yang disediakan dalam *Python* sangat efisien dan sederhana, tetapi juga menawarkan pendekatan yang efektif untuk pemrograman berorientasi objek. Sintaksis yang elegan dan pengetikan dinamis, bersama dengan sifat yang ditafsirkan, membuat *Python* sangat cocok untuk *scripting* dan pengembangan aplikasi cepat dalam berbagai bidang dan platform. *Python* juga memiliki perpustakaan yang dikembangkan oleh komunitas *Python* yang dapat dengan mudah diunduh dan digunakan sesuai dengan kebutuhan pengembangan.

## 2. METODE

Dalam penelitian ini, metode yang digunakan adalah analisis data secara kuantitatif. Metode kuantitatif dalam penelitian adalah metode yang digunakan untuk mengumpulkan data numerik dan menganalisis data tersebut secara statistik(Rudini, 2017). Metode ini digunakan untuk menguji hipotesis atau untuk menemukan hubungan antara variabel. Oleh karena itu, proses pengumpulan data dan analisis data tidak dapat dipisahkan. Sehingga kami menyelesaikan masalah dengan mengacu pada kejadian yang terjadi secara langsung dan proses bisnis di sebuah klinik.

Data dikumpulkan dengan cara melakukan pengamatan langsung terhadap klinik dan menganalisis proses bisnis yang terjadi. Penelitian yang digunakan adalah penelitian prediktif analitik. Penelitian prediktif analitik adalah penelitian yang digunakan untuk memprediksi kejadian masa depan berdasarkan data masa lalu. Dalam penelitian prediktif analitik, data dikumpulkan dan dianalisis untuk menemukan pola dan hubungan antara variabel. Kemudian, model matematis atau algoritma dikembangkan untuk memprediksi kejadian masa depan. Model ini dapat digunakan untuk memprediksi kehamilan.

Sebagai referensi, penulis mengikuti tahapan model *Cross-Industry Standard Process for Data Mining* (CRISP-DM), terdiri dari *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, dan *Deployment* **(Singgalen, 2022)**. CRISP-DM digunakan untuk memandu dalam proses *machine learning* dari awal sampai akhir. ini memungkinkan dalam pengelolaan proyek menjadi lebih efektif dan meningkatkan hasil akhir dari proyek/penelitian yang dilakukan.



**Gambar 1. CRISP-DM**

### 2.1 *Business Understanding*

*Business understanding* ditujukan untuk memahami tentang tujuan bisnis, mengevaluasi tujuan bisnis dan menguraikan tujuan bisnis ke dalam tujuan analisis data **(Septian, 2009)**. Dalam kajian ini, diperlukan kompetensi untuk mengumpulkan data dari berbagai sumber untuk diagnosa kehamilan dan melakukan analisis untuk membuat model prediksi yang dibutuhkan.

### 2.2 *Data Understanding*

Pada proses ini, kami melakukan proses pencarian data, analisis data, dan evaluasi kualitas data yang digunakan dalam proyek ini. Sumber data yang akan digunakan berasal dari berbagai sumber seperti penelitian sebelumnya dan wawancara dengan beberapa ahli. Sumber literatur yang digunakan banyak diperoleh dari buku, *paper* atau jurnal, karya ilmiah, dan situs web yang relevan.

### 2.3 *Data Preparation*

*Data Preparation* atau adalah proses yang dilakukan untuk mengubah data mentah menjadi data yang berkualitas dan cocok sebagai input untuk alat-alat analisis data.

### 2.4 *Modelling*

Pada tahap ini, beberapa metode pemodelan yang pemilihan dan implementasi pada dataset yang telah dipersiapkan untuk menangani keperluan bisnis yang harus sesuai. Kedua metode utama yang digunakan pada proyek ini adalah *Decision Tree* dan *Random Forest* sebagai metode klasifikasi.

### 2.5 *Evaluation*

Pada proses ini, model yang telah dibuat akan diuji dan dinilai keakurasiannya. Tahap ini digunakan untuk mengukur seberapa baik model yang dipilih dan memenuhi tujuan. Selain itu, untuk menentukan apakah dibutuhkan model lain untuk dibuat dan diukur.

### 2.6 *Deployment*

*Deployment* adalah membuat sistem dari model yang telah dibuat dan dikembangkan ke sistem yang telah ada sebelumnya. Pastinya akan melihat evaluasi terlebih dahulu, jika evaluasi sudah baik, maka *deployment* dilakukan.

## 3. HASIL DAN PEMBAHASAN

Hasil dan pembahasan pada penelitian ini akan menjelaskan mengenai model prediksi dan hasil *deployment*.

1. **Model Prediksi**

Data yang digunakan dalam penelitian ini berjumlah 1014 dengan 7 fitur Seperti yang terlihat pada Gambar 2 di bawah ini :



**Gambar 2. Data yang Digunakan**

Fitur yang digunakan sebagai input yaitu *age, SystolcBP, DiastolicBP, BS(Blood Sugar), BodyTemp, dan HeartRate*. Sedangkan *RiskLevel* digunakan sebagai *output*/label dengan memiliki 3 kategori yaitu : *low risk, mid risk dan high risk*. Setelah dilakukan *data understanding* maka tahap selanjutnya dilakukan *data preparation/* *data preprocessing* dilakukan untuk mengubah data mentah menjadi data yang berkualitas dan siap digunakan.



**Gambar 3. Data Preparation**

Pada gambar 3 diketahui bahwa tidak ada data yang *null*. Namun, terdapat data duplikat sebanyak 562 data. Agar menjadi data yang berkualitas maka data duplikat tersebut dihapus dan menyisakan data sejumlah 452.



**Gambar 4. Proses Encodding**

Pada kolom label dilakukan proses *encoding* pengubahan *text* menjadi angka agar dipahami *python*. Seperti pada gambar diatas, setiap data dengan *label low risk* diubah menjadi 1, *mid risk* diubah menjadi 2, dan *high risk* diubah menjadi 3. Berikut ini merupakan jumlah hasil dari setiap label ditunjukkan pada diagram batang di bawah ini :



**Gambar 5. Jumlah Data**

Untuk melihat *correlasi* dari setiap fitur bisa dilihat dari heatmap berikut :



**Gambar 6. Correlasi Setiap Fitur**

Setelah data di *preprocessing* maka Langkah selanjutnya yaitu melakukan *modelling* karena data tersebut sudah siap digunakan. Pada tahap *modelling* ini dilakukan beberapa pengujian model *machine learning* pada dataset dengan menggunakan algoritma *Random forest* dan *decision tree*. Untuk mengetahui akurasi tersebut dilakukan pembagian dataset baik itu data yang seimbang maupun yang tidak seimbang menjadi dua bagian yaitu : *train* dan *test*. Pada kasus ini data untuk *train* sebesar 80% dan untuk i sebesar 20%.

Hasil dari evaluasi *decision tree* dan *random forest* dengan data yang tidak seimbang menggunakan algoritma gini dan entropy ditunjukkan pada tabel :

**Tabel 1. Hasil Evaluasi dengan Data Tidak Seimbang**

|  |  |  |
| --- | --- | --- |
|  | **Decision Tree** | **Random Forest** |
| **Gini** | **Entropy** | **Gini** | **Entropy** |
| Accuracy | 59 % | 58 % | 66 % | 63 % |
| Precision | 59 % | 57 % | 64 % | 61 % |
| Recall | 60 % | 58 % | 66 % | 63 % |
| F1 score | 59 % | 58 % | 64 % | 61 % |

Dapat dilihat pada Tabel 1. bahwa akurasi terbesar dari model evaluasi dengan data yang tidak seimbang yaitu random forest menggunakan algoritma gini dengan akurasi sebesar 66%. Selain itu, pada tabel 2. terdapat hasil evaluasi decision tree dan random forest dengan data yang seimbang menggunakan algoritma gini dan entropy. Agar data menjadi seimbang, maka digunakan metode SMOTE.

SMOTE (*Synthetic Minority Over-sampling Technique*) adalah metode untuk mengatasi masalah ketidakseimbangan data dalam *machine learning*. SMOTE menggunakan teknik *oversampling* untuk meningkatkan jumlah data minoritas dalam dataset dengan menciptakan sampel sintetis yang berdekatan dengan data minoritas yang ada.

SMOTE bekerja dengan mengambil sampel acak dari data minoritas dan menemukan sampel yang paling dekat dengannya di dalam dataset. Kemudian, ia menghitung jarak antara sampel yang diambil dan sampel yang ditemukan, dan menciptakan sampel sintetis baru dengan mengambil titik yang berada di antara kedua sampel tersebut.

**Tabel 2. Hasil Evaluasi dengan Data Seimbang**

|  |  |  |
| --- | --- | --- |
|  | **Decision Tree** | **Random Forest** |
| **Gini** | **Entropy** | **Gini** | **Entropy** |
| Accuracy | 80 % | 79 % | 81 % | 78 % |
| Precision | 80 % | 79 % | 81 % | 78 % |
| Recall | 80 % | 79 % | 81 % | 78 % |
| F1 score | 80 % | 79 % | 81 % | 78 % |

Berdasarkan tabel akurasi terbesar pada model evaluasi dengan data yang seimbang menggunakan algoritma gini dan entropy yaitu *random forest* dengan algoritma gini yang memiliki akurasi yaitu 81 %.

1. **Implementasi Model**

Implementasi model *Machine Learning* adalah proses mengintegrasikan model yang telah dibangun ke dalam sistem atau aplikasi yang dapat digunakan oleh pengguna. Website ini menggunakan *framework Flask*, berikut ini tampilannya :

1. **Halaman Register**

Untuk dapat masuk ke dalam aplikasi, maka pengguna harus melakukan register terlebih dahulu. Dalam halaman register ini terdapat form inputan yang harus diisi oleh pengguna, yaitu : username, email dan password.



**Gambar 7. Halaman Register**

1. **Halaman Login**

Setelah melakukan register, maka pengguna dapat login dengan *email* dan password yang sudah diinputkan sebelumnya. Jika inputannya sesuai maka pengguna akan diarahkan ke halaman dashboard aplikasi.



**Gambar 8. Halaman Login**

1. **Halaman Dashboard**

Halaman dashboard ini merupakan tampilan awal ketika pengguna sudah melakukan login. Terdapat beberapa fitur yang bisa diakses oleh pengguna. Bisa dilihat pada sidebar aplikasi. Terdapat menu logout di pojok kanan atas jika pengguna ingin keluar dari aplikasi yang nantinya akan diarahkan ke halaman login.



**Gambar 9. Halaman Dashboard**

1. **Halaman Prediksi**

Pada halaman prediksi ini merupakan implementasi model yang sudah dibuat. Tujuan dibuatnya halaman prediksi ini yaitu untuk mempermudah dalam penggunaannya dan mudah dimengerti. Terdapat form inputan yang harus diisi oleh pengguna yaitu, NIK Pasien , Umur, Sistole, Diastole, Gula Darah, Suhu Tubuh, dan Detak Jantung. Jika pengguna sudah mengisi form tersebut, penguna dapat mengklik tombol Prediksi.



**Gambar 10. Halaman Prediksi**

1. **Halaman Hasil Prediksi**

Jika form inputan dalam halaman prediksi sudah diisi, maka jika di klik tombol Prediksi akan menampilkan halaman hasil dari prediksi berdasarkan data yang telah diinputkan. Berikut ini tampilan dari halaman hasil prediksi :



**Gambar 11. Halaman Hasil Prediksi**

## 4. KESIMPULAN

Dari hasil evaluasi yang ditunjukkan dalam tabel 1 dan 2, dapat disimpulkan bahwa model *Random Forest* yang menggunakan data yang seimbang dan metode Gini memiliki akurasi terbaik sebesar 81%. Hal ini menunjukkan bahwa menggunakan data yang seimbang dapat meningkatkan performa dalam mendiagnosis kehamilan. Algoritma *Random Forest* merupakan salah satu algoritma yang banyak digunakan dalam klasifikasi karena kinerjanya yang baik, jika digabungkan dengan data yang seimbang dan metode Gini, dapat menghasilkan performa yang lebih baik dalam mengdiagnosis kehamilan. Oleh karena itu, dalam praktiknya, algoritma *Random Forest* dengan data yang seimbang dan metode Gini dapat digunakan untuk memprediksi kehamilan.

## UCAPAN TERIMA KASIH

Terima kasih kepada Bapak Roni Habibi selaku pembimbing pada penelitian ini dan kepada teman-teman D4 Teknik Informatika B Angkatan 2020 yang telah memberikan *support*. Tak lupa untuk mengucapkan terima kasih kepada kedua orang tua yang telah mendukung baik itu materil maupun non-materil.

## DAFTAR rujukan

Abijono, H., Santoso, P., & Anggreini, N. L. (2021). Algoritma Supervised Learning Dan Unsupervised Learning Dalam Pengolahan Data. *Jurnal Teknologi Terapan: G-Tech*, *4*(2), 315–318. https://doi.org/10.33379/gtech.v4i2.635

Ginantra, Ni Luh Wiwik Sri Rahayu, dkk. (2021). *Data Mining dan Penerapan Algortima* (J. Watrianthos, Ronal & Simarmata, Ed.). https://books.google.co.id/books?id=v0gtEAAAQBAJ&printsec=frontcover&hl=id#v=onepage&q&f=false

Helmi, S. (2021). *Analisis data* (Issue July).

Id, I. D. (2021). *Machine Learning: Teori, Studi Kasus dan Implementasi Menggunakan Python* (1st ed.). UR PRESS.

Indrawati, S. K., Ns, M., Elfira, E., Ns, M. K., Yufdel, S. K., & Ns, M. K. (2021). *Terapi Komplementer pada Kehamilan. Media Sains Indonesia.* (R. R. Rerung, Ed.). Penerbit Media Sains Indonesia.

 Kompas.com. (2022). *Jumlah Penduduk Indonesia 2022*. Kompas. https://nasional.kompas.com/read/2022/04/27/03000051/jumlah-penduduk-indonesia-2022.

Muttaqin, D. (2022). *Big Data : Informasi Dalam Dunia Digital* (R. Watrianthos, Ed.). Yayasan Kita Menulis.

 Muzakir, A., & Wulandari, R. A. (2016). Model Data Mining sebagai Prediksi Penyakit Hipertensi Kehamilan dengan Teknik Decision Tree. *Scientific Journal of Informatics*, *3*(1), 19–26. https://doi.org/10.15294/sji.v3i1.4610

Rachmi, A. N. (2020). *Implementasi Metode Random Forest Dan Xgboost Pada Klasifikasi Customer Churn*.

Riaddy, A. I., Sibaroni, Y., & Aditsania, A. (2016). Ekstraksi Informasi pada Makalah Ilmiah dengan Pendekatan Supervised Learning. *E-Proceeding of Engineering*, *3*(1), 1184–1190.

Rudini, R. (2017). Peranan Statistika Dalam Penelitian Sosial Kuantitatif. *Jurnal SAINTEKOM*, *6*(2), 53. https://doi.org/10.33020/saintekom.v6i2.13

Septian, N. Y. (2009). Data Mining Menggunakan Algoritma Naïve Bayes Untuk Klasifikasi Kelulusan Mahasiswa Universitas Dian Nuswantoro. *Jurnal Semantik 2013*, 1–11.

Singgalen, Y. A. (2022). *Analisis Performa Algoritma NBC , DT , SVM dalam Klasifikasi Data Ulasan Pengunjung Candi Borobudur Berbasis CRISP-DM*. *4*(3), 1634–1646. https://doi.org/10.47065/bits.v4i3.2766

Siswanto, M. M. (2013). Metode Logika Fuzzy Tsukamoto Dalam Sistem Pengambilan Keputusan Penerimaan Beasiswa. *Jurnal Media Infotama*, *9*(1), 140–165.

Timotius, K. H. (2017). *Pendekatan Manajemen Pengetahuan untuk Perkembangan Pengetahuan* (P. Christian, Ed.; 1st ed.). Penerbit ANDI. https://books.google.co.id/books?id=yVJLDwAAQBAJ&printsec=frontcover&hl=id#v=onepage&q&f=false