

Pembangunan Model Pendeteksi Risiko Preeklamsia pada Ibu Hamil dengan Menggunakan Metode *Data Mining*

Muhammad Ilham¹, Ni Luh Saddhwi Saraswati Adnyani¹, dan Kadarsah Suryadi¹

¹Program Studi Teknik Industri, Fakultas Teknologi Industri, Institut Teknologi Bandung, Bandung, Indonesia

mohammad.ilham362@gmail.com, saddhwi@itb.ac.id, kadarsah@itb.ac.id

Abstrak

Preeklamsia adalah komplikasi kehamilan yang diindikasikan dengan peningkatan tekanan darah yang terjadi setelah usia kehamilan 20 minggu serta keberadaan protein di dalam urine. Apabila tidak ditangani dengan cepat, preeklamsia dapat berujung pada kematian ibu dan janin. Oleh karena itu, diperlukan suatu metode yang dapat membantu para tenaga kesehatan untuk dapat memberikan deteksi awal penyakit preeklamsia yang diderita oleh ibu hamil. Salah satu metode yang dapat digunakan adalah metode *data mining*. Penelitian ini dilakukan dengan tujuan untuk mengembangkan model berdasarkan metode *data mining* yang dapat digunakan sebagai alat untuk mengidentifikasi pasien dengan preeklamsia dan juga untuk mengidentifikasi faktor risiko terkait. Penelitian dilakukan dengan menggunakan enam algoritma klasifikasi *data mining* pada 109 data pasien klinik kandungan di Rumah Sakit Islam Jakarta Pondok Kopi (RSIJPK). Fitur *input* yang digunakan sebagai atribut deteksi Preeklamsia diperoleh berdasarkan hasil studi literatur dan hasil konsultasi dengan dokter spesialis kandungan. Berdasarkan hasil evaluasi model, diperoleh *logistic regression* sebagai algoritma yang memiliki performansi terbaik dalam mendeteksi penyakit preeklamsia pada pasien kehamilan RSIJPK dengan nilai akurasi sebesar 98% dan tingkat presisi sebesar 100%. Selain itu, pada penelitian ini juga dirancang suatu prototipe aplikasi yang dapat digunakan oleh tenaga kesehatan untuk dapat mendeteksi risiko preeklamsia pada ibu hamil secara cepat.

Kata kunci: *data mining*, *logistic regression*, preeklamsia

Abstract

Preeclampsia is a pregnancy complication indicated by an increase in blood pressure that occurs after 20 weeks of gestation and the presence of protein in the urine. If not treated quickly, preeclampsia can lead to maternal and fetal death. Therefore, a method that can help health workers to provide early detection of preeclampsia is needed. One method that can be used is data mining. This study was conducted with the aim of developing a model based on data mining methods that can be used as a tool to identify patients with preeclampsia and also to identify associated risk factors. This study was conducted using six data mining classification algorithms on 109 obstetric clinic patient data at the Jakarta Pondok Kopi Islamic Hospital (RSIJPK). The input features used as preeclampsia detection attributes were obtained based on the results of a literature study and consultations with obstetricians. Based on the results of the model evaluation, logistic regression has the best performance in detecting preeclampsia with accuracy value of 98% and precision level of 100%. In addition, this study also designed an application prototype that can be used by health workers to quickly detect the risk of preeclampsia in pregnant women.

Keywords: data mining, logistic regression, preeclampsia

1. Pendahuluan

Pada tahun 2020, angka kematian ibu akibat kehamilan cukup tinggi yaitu sekitar 287.000 ibu meninggal di seluruh dunia pada saat atau sesudah masa kehamilan (World Health Organization, 2023) Angka kematian ibu di Indonesia merupakan salah satu yang paling tinggi di Asia Tenggara (Baharuddin dkk., 2019). Terdapat beberapa komplikasi utama yang mengakibatkan kematian ibu akibat kehamilan, diantaranya yaitu pendarahan setelah melahirkan, infeksi setelah melahirkan, serta preeklamsia dan eklamsia (World Health Organization, 2023). Berdasarkan data kematian ibu pada bulan September sampai Oktober pada tahun 2014 di rumah sakit yang tersebar di provinsi Sumatera Utara, Banten, Jawa Barat, Jawa Tengah, Jawa Timur, dan Sulawesi Selatan, ditemukan bahwa preeklamsia adalah penyebab kematian ibu dengan angka tertinggi yaitu sebanyak 38 kasus dari total 90 kasus atau sebesar 42% dari sampel (Baharuddin dkk., 2019).

Preeklamsia adalah komplikasi kehamilan yang diindikasikan dengan peningkatan tekanan darah yang terjadi setelah usia kehamilan 20 minggu serta keberadaan protein di dalam urin (Roberts & Gammill, 2005). Kelainan ini memberikan dampak kepada ibu dan juga anak yang berada di dalam janin. Kelanjutan dari preeklamsia adalah eklamsia yang ditandai dengan kejang-kejang yang diderita oleh pasien. Apabila tidak ditangani dengan cepat, preeklamsia dapat berujung pada kematian ibu dan janin. Terdapat dugaan bahwa preeklamsia disebabkan oleh kelainan dari fungsi dan perkembangan plasenta, yaitu organ yang memiliki fungsi untuk menyalurkan nutrisi dan darah untuk janin.

Info Makalah:

Dikirim : 07-13-23;
Revisi 1 : 02-16-24;
Diterima : 02-20-24.

Penulis Korespondensi:

Telp : +62-817110695
e-mail : saddhwi@itb.ac.id

Kematian ibu yang disebabkan oleh preeklamsia dapat dicegah sejak dini (Kustiyahningsih dkk., 2020). Deteksi dini preeklamsia sangat penting dan disarankan untuk dilakukan selama trimester pertama kehamilan agar tindakan pencegahan dapat dimulai lebih awal (Chaemsaitong dkk., 2022). Oleh karena itu, diperlukan suatu metode yang dapat membantu para tenaga kesehatan untuk dapat memberikan deteksi awal terkait penyakit preeklamsia yang diderita oleh ibu hamil. Salah satu metode yang dapat digunakan untuk mendeteksi Preeklamsia adalah metode *data mining*. *Data mining* merupakan suatu tahapan yang dilakukan pada pencarian pengetahuan dari *database* untuk memproduksi pola atau model yang berguna dari data (Silwattananusarn & KulthidaTuamsuk, 2012). *Data mining* terletak di antara ilmu statistik, teknologi basis data, pengenalan pola, pembelajaran mesin, visualisasi data, dan sistem pakar (Obenshain, 2004). Berdasarkan data-data *health record* yang telah dikumpulkan oleh pihak rumah sakit, dapat dibangun model deteksi dengan berbagai metode *data mining* untuk memilih model terbaik berdasarkan beberapa parameter evaluasi model. Penggunaan metode *data mining* untuk mendeteksi Preeklamsia pada pasien diharapkan dapat membantu tenaga kesehatan untuk mempercepat pendeteksian preeklamsia pada pasien untuk menekan angka kematian ibu.

Terdapat beberapa penelitian terdahulu yang menggunakan penerapan metode *data mining* untuk mendeteksi penyakit preeklamsia pada pasien kehamilan. Poon dan Nicolaides (2014) mengembangkan metode skrining yang efektif untuk memprediksi Preeklamsia pada trimester pertama kehamilan dengan menggunakan *Bayes Theorem*. Zhang dkk. (2019) melakukan penelitian untuk mengestimasi kinerja prediktif dari model skrining untuk preeklamsia pada pasien hamil trimester pertama di Cina berdasarkan karakteristik ibu, MAP, serum PLGF, dan PAPP-A dengan menerapkan algoritma PREDICTOR. Sufriyana dkk. (2020) mengembangkan dan memvalidasi model prediksi Preeklamsia dengan menggunakan algoritma *Random Forest* yang diterapkan pada dataset asuransi kesehatan nasional di Indonesia. Marić dkk. (2020) mengembangkan model prediksi Preeklamsia dengan menggunakan algoritma *Elastic Net*, *Gradient Boosting*, dan *Logistic Regression*.

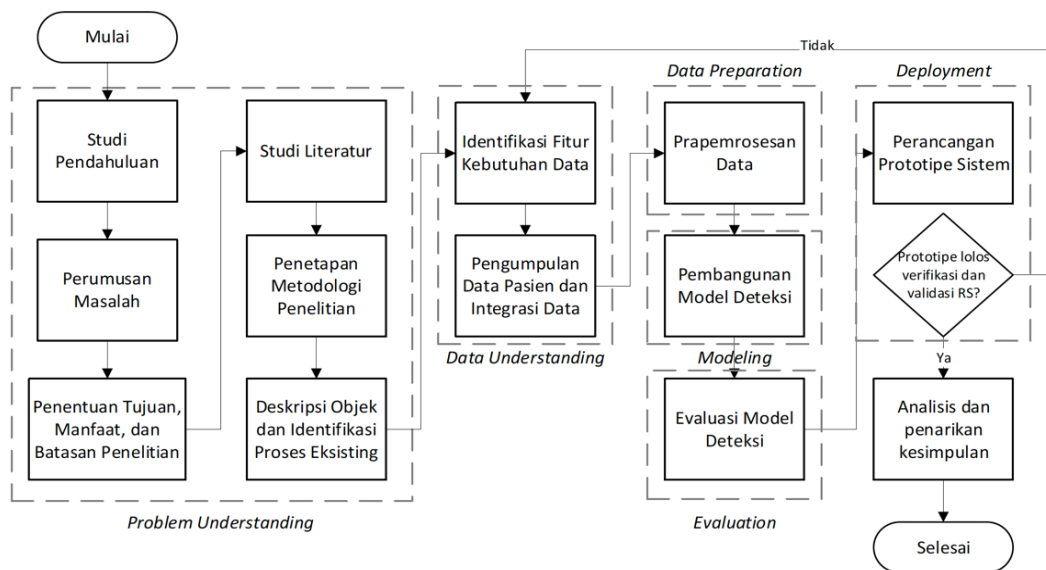
Penelitian ini dilakukan dengan tujuan untuk mengembangkan model berdasarkan metode *data mining* yang dapat digunakan sebagai alat pendeteksi untuk mengidentifikasi pasien dengan preeklamsia dan juga untuk mengidentifikasi faktor risiko yang memiliki kontribusi terbesar terhadap penyakit preeklamsia pada ibu hamil. Pada penelitian ini, akan dipilih algoritma terbaik untuk membangun model prediksi risiko preeklamsia tersebut. Algoritma yang akan dibandingkan dalam penelitian ini yaitu *decision tree*, *support vector machine*, *naïve bayes*, *logistic regression*, *k-nearest neighbors*, dan *random forest*. Penggunaan algoritma-algoritma *machine learning* ini disebabkan oleh jenis *task* yang akan diselesaikan pada penelitian ini. Fitur target dari penelitian ini adalah fitur risiko Preeklamsia yang direpresentasikan oleh dua nilai yaitu positif dan negatif. *Task* dari *data mining* yang digunakan untuk melakukan deteksi tersebut adalah *task* klasifikasi. Oleh karena itu, dipilih algoritma-algoritma yang dapat digunakan untuk melakukan *task* klasifikasi. Selain itu, pada penelitian ini juga akan dirancang suatu prototipe aplikasi yang dapat digunakan oleh tenaga kesehatan untuk dapat mendeteksi risiko preeklamsia pada ibu hamil secara cepat.

2. Metode

Penelitian ini dilakukan dengan langkah-langkah seperti yang diilustrasikan pada Gambar 1. Langkah-langkah tersebut mengadopsi metodologi *Cross Industry Standard Process for Data Mining* (CRISP-DM) untuk domain kedokteran yang disebut CRISP-MED-DM (Niakšu, 2015). Terdapat enam tahap pada metodologi CRISP-MED-DM, yaitu tahap *problem understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment* (Niakšu, 2015).

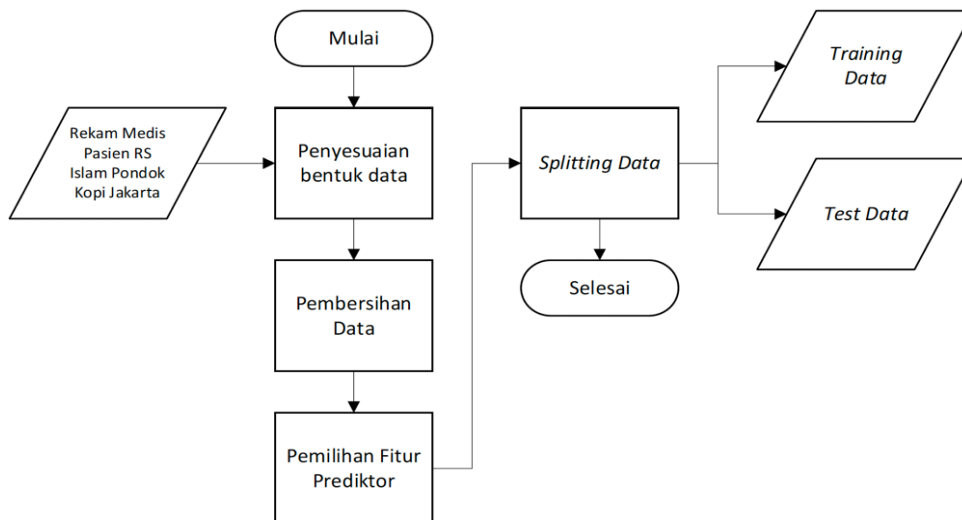
Tahap *problem understanding* merupakan tahapan awal yang dilakukan pada penelitian ini. Tahap ini terdiri dari studi pendahuluan, perumusan masalah, penentuan tujuan, manfaat, dan batasan penelitian, studi literatur, serta penetapan metodologi penelitian, serta deskripsi objek penelitian dan identifikasi proses eksisting.

Selanjutnya, *data understanding* merupakan tahapan yang terdiri dari identifikasi kebutuhan fitur data serta pengumpulan data pasien dan integrasi data. Data yang dibutuhkan untuk kebutuhan penelitian diidentifikasi berdasarkan studi literatur yang telah dilakukan dan hasil wawancara dengan dokter spesialis kehamilan rumah sakit Rumah Sakit Islam Jakarta Pondok Kopi (RSIJPK). Identifikasi fitur kebutuhan data dilakukan untuk melakukan seleksi terhadap seluruh data yang didapatkan dari pihak rumah sakit menjadi beberapa bagian data yang relevan dengan kebutuhan penelitian. Seleksi data tersebut dilakukan berdasarkan studi literatur dan hasil wawancara dengan dokter kandungan pihak rumah sakit. Pengurangan jumlah atribut dilakukan untuk mengurangi kompleksitas dari proses *data mining* serta untuk meningkatkan performa dari model yang akan dibangun. Pengumpulan data pasien dilakukan setelah identifikasi fitur kebutuhan data. Data dikumpulkan dari data pasien yang memiliki penyakit preeklamsia dan data pasien yang tidak memiliki penyakit preeklamsia. Data pasien diambil dalam horizon waktu tertentu dari bagian rekam medis dan laboratorium untuk pasien ibu hamil. Pengumpulan data yang dilakukan bersesuaian dengan standar prosedur yang akan diberikan oleh pihak rumah sakit. Data yang dikumpulkan dari bentuk berkas fisik milik rumah sakit dipindahkan ke dalam format *excel* untuk memudahkan kegiatan komputasi.



Gambar 1. Diagram Alir Penelitian.

Tahap *data preparation* dilakukan dengan melakukan pra pemrosesan data. Pra Pemrosesan data dilakukan setelah dilakukan pengumpulan data pasien. Tahapan ini dilakukan agar data yang telah dikumpulkan siap untuk diolah dan memiliki format yang sesuai dengan kebutuhan pembangunan model deteksi. Tahapan pra pemrosesan data secara lebih detail ditampilkan pada Gambar 2. Tahapan dimulai dengan penyesuaian bentuk data, identifikasi kebutuhan atribut, pembersihan data, sampai *splitting data* menjadi 2 bagian yaitu data *training* dan data *test*.



Gambar 2. Tahap Pra Pemrosesan Data.

Kemudian, pada tahap *modeling* dilakukan pembangunan model deteksi preeklamsia. Data yang sudah siap akan diproses dengan menggunakan algoritma-algoritma yang telah ditentukan agar model yang dibangun dapat digunakan untuk mendeteksi risiko penyakit preeklamsia pada ibu hamil. Proses *data mining* dilakukan dengan menggunakan bahasa pemrograman Python 3. Algoritma *machine learning* yang akan digunakan yaitu *decision tree*, *support vector machine*, *naïve bayes*, *logistic regression*, *k-nearest neighbors*, dan *random forest*. Setelah model dibangun dengan berbagai algoritma yang berbeda, akan dilakukan evaluasi sebagaimana dijelaskan pada tahapan selanjutnya. Kemudian, algoritma yang memiliki performansi terbaik pada tahapan evaluasi model akan ditingkatkan performannya dengan cara mencari kombinasi dari parameter-parameter yang digunakan pada model, atau yang biasa disebut dengan *hyperparameter tuning*. Percobaan kombinasi dari parameter-parameter tersebut sehingga didapatkan nilai yang optimal dapat dilakukan dengan menggunakan metode *grid search* maupun *random search* dan melakukan estimasi dari nilai *error* pada tahapan *training* model dengan menggunakan metode *cross validation*. Jumlah *fold* yang paling umum digunakan pada metode *cross validation* adalah 5 dan 10. Sufriyana dkk. (2020) menggunakan *10-fold cross validation*, sedangkan Marić dkk. (2020) menggunakan *5-fold cross validation*. Pada penelitian yang dilakukan, akan

digunakan *10-fold cross validation*. Namun, perlu diingat bahwa tidak seluruh kegiatan *hyperparameter tuning* dapat memberikan performa yang lebih baik apabila dibandingkan dengan proses konstruksi model yang tidak menggunakan *hyperparameter tuning* (Nair, 2021). Apabila hal tersebut terjadi, pembangunan model yang tidak menggunakan *hyperparameter tuning* dapat terpilih karena memiliki performa yang lebih baik.

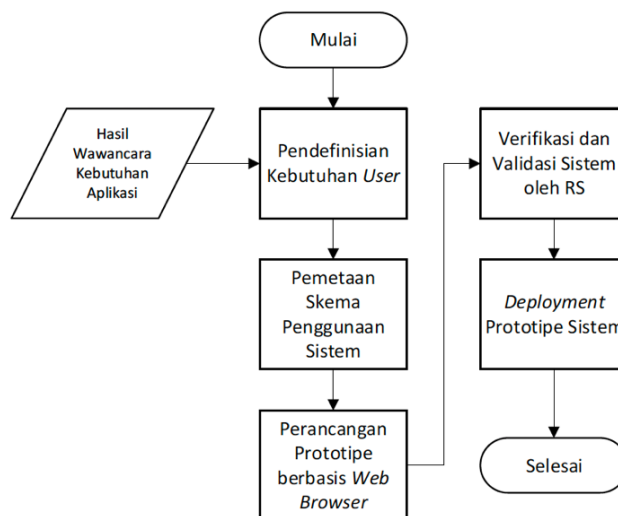
Tahap evaluasi model merupakan tahapan yang dilakukan untuk menguji hasil kinerja model klasifikasi penelitian. Model deteksi yang telah dibangun pada tahapan pembangunan model akan dilakukan evaluasi dengan menggunakan *confusion matrix* yang ditunjukkan pada Gambar 3 (Draelos, 2019). Perhitungan-perhitungan yang dilakukan pada TP, FP, FN, dan TN dapat menghasilkan beberapa parameter-parameter yang dapat digunakan untuk mengukur performansi dari model klasifikasi yang dibangun. *Metric* utama yang dijadikan acuan untuk mengevaluasi model adalah akurasi dan *precision*. Alasan digunakan *precision* adalah karena model tidak ingin salah mendeteksi pasien yang sebenarnya memiliki risiko tinggi untuk mengidap preeklamsia namun dideteksi tidak memiliki risiko tinggi untuk mengidap preeklamsia oleh model. Hal tersebut dapat berakibat fatal dan bahkan dapat memakan korban manusia. Setelah algoritma model deteksi dengan tingkat akurasi dan presisi tertinggi terpilih, model tersebut ditingkatkan performanya dengan menggunakan teknik *hyperparameter tuning* sebagaimana dijelaskan pada tahapan pembangunan model.

	<i>Actually Positive</i>	<i>Actually Negative</i>
<i>Predicted Positive</i>	<i>True Positives (TP)</i>	<i>False Positives (FP)</i>
<i>Predicted Negative</i>	<i>False Negatives (FN)</i>	<i>True Negatives (TN)</i>

Gambar 3. *Confusion Matrix*.

Tahap *deployment* merupakan tahapan pembangunan sistem informasi yang berguna bagi pemangku kepentingan dari model yang telah dibangun dan dievaluasi. Perancangan sistem dilakukan setelah model deteksi terbaik telah dipilih pada tahapan evaluasi model. Pada tahap ini, akan diidentifikasi spesifikasi dari sistem yang dibutuhkan oleh pihak rumah sakit. Spesifikasi dari sistem didapatkan dengan membagi kebutuhan menjadi dua bagian yaitu kebutuhan fungsional dan kebutuhan non fungsional. *Output* dari sistem yang akan dihasilkan pada penelitian berupa purwarupa atau *prototype* yang dapat digunakan oleh pihak rumah sakit. Perancangan dari aplikasi dan *user interface* akan dilakukan dengan menggunakan *package Python* bernama *Streamlit*. *Streamlit* merupakan *framework* yang dapat ditulis dengan menggunakan bahasa pemrograman *Python* untuk membangun aplikasi *website*. Purwarupa yang telah dihasilkan akan diverifikasi dan divalidasi oleh pihak rumah sakit. Apabila terdapat bagian dari sistem yang perlu untuk dilakukan perbaikan, maka purwarupa dari sistem akan dimodifikasi untuk memenuhi revisi perbaikan yang diberikan oleh pihak rumah sakit. Gambar 4 menunjukkan diagram alir dari proses perancangan prototipe sistem secara lengkap.

Terakhir, pada penelitian ini akan dilakukan analisis serta penarikan kesimpulan dan saran. Analisis dan kesimpulan dilakukan terhadap pengolahan data dan juga perancangan sistem yang telah dibuat.



Gambar 4. Diagram Alir Perancangan Prototipe Sistem.

3. Hasil dan Pembahasan

3.1. Identifikasi Kebutuhan Data

Penentuan fitur dan transformasi data dilakukan untuk memenuhi tahapan *data understanding* dan *data preparation* pada tahapan CRISP-DM. Fitur-fitur deteksi dipilih berdasarkan studi literatur dan beberapa fitur yang digunakan pada penelitian serupa mengenai deteksi penyakit preeklamsia pada ibu hamil. Fitur-fitur tersebut dapat dilihat pada Tabel 1.

Tabel 1. Fitur yang Diambil dari Penelitian Sebelumnya.

No	Data	Tipe	Penjelasan	Sumber Identifikasi
1	Umur	Numerik	Usia dari pasien kehamilan	(Marić dkk., 2020), (Sufriyana dkk., 2020), (Zhang dkk., 2019)
2	Umur Kehamilan	Numerik	Usia kehamilan dari pasien	(Roberts & Gammill, 2005)
3	Tinggi Badan	Numerik	Tinggi badan dari pasien kehamilan	(Marić dkk., 2020), (Zhang dkk., 2019)
4	Berat Badan	Numerik	Berat badan dari pasien kehamilan	(Zhang dkk., 2019)
5	Tekanan Darah	Numerik	Tekanan darah dari pasien kehamilan. Dibagi menjadi 2 bagian yaitu tekanan darah diastolik dan tekanan darah sistolik	(Marić dkk., 2020), (Poon & Nicolaides, 2014), (Zhang dkk., 2019)
6	Kehamilan ke-	Numerik	Urutan kehamilan dari pasien kehamilan	(Pittara, 2022)
7	Sejarah Penyakit	Kategori	Sejarah penyakit yang dimiliki oleh pasien kehamilan. Dibagi menjadi beberapa kategori riwayat penyakit	(Marić dkk., 2020), (Sufriyana dkk., 2020), (Zhang dkk., 2019), (Poon & Nicolaides, 2014)

Selain dari itu, fitur-fitur yang akan digunakan juga didapatkan dari hasil wawancara dengan dokter spesialis kandungan dari pihak RSIJPK. Fitur-fitur tersebut dapat dilihat pada Tabel 2.

Tabel 2. Fitur yang Diambil dari Hasil Wawancara dengan Dokter Spesialis.

No	Data	Tipe	Penjelasan	Sumber Identifikasi
1	Keluhan	Kategori	Keluhan fisik yang dirasakan oleh pasien. Contohnya adalah keluhan pusing, kejang, mual, dan sebagainya	Dokter Spesialis Kandungan
2	Albumin Urea	Numerik	Hasil uji laboratorium urin yang telah dilakukan pasien. Mengindikasikan keberadaan protein pada urin	Dokter Spesialis Kandungan
3	Trombosit	Numerik	Hasil uji laboratorium tes darah yang telah dilakukan pasien. Merepresentasikan jumlah sel darah putih	Dokter Spesialis Kandungan
4	Hb	Numerik	Hasil uji laboratorium tes darah yang telah dilakukan pasien. Merepresentasikan jumlah sel darah merah	Dokter Spesialis Kandungan
5	<i>Serum Glutamic-Oxaloacetic Transaminase (SGOT)</i>	Numerik	Hasil uji laboratorium tes fungsi hati yang telah dilakukan oleh pasien. Merepresentasikan jumlah enzim SGOT di dalam tubuh	Dokter Spesialis Kandungan
6	<i>Serum Glutamic Pyruvic Transaminase (SGPT)</i>	Numerik	Hasil uji laboratorium tes fungsi hati yang telah dilakukan oleh pasien. Merepresentasikan jumlah enzim SGPT di dalam tubuh	Dokter Spesialis Kandungan
7	Bilirubin	Kategori	Hasil uji laboratorium yang merepresentasikan jumlah bilirubin di dalam tubuh	Dokter Spesialis Kandungan

3.2. Pengumpulan Data Pasien dan Integrasi Data

Pemenuhan data yang dibutuhkan untuk keperluan penelitian menggunakan data yang didapatkan dari berkas rekam medis setiap pasien. Pada akhir dari tahapan pengumpulan data, dihasilkan sebanyak 109 data yang terdiri dari 48 data pasien positif preeklamsia dan 61 data pasien negatif preeklamsia. Selain dari itu, bagian dari berkas rekam medis yang akan diserap meliputi layanan rawat jalan, layanan rawat inap, fasilitas laboratorium. Data dari seluruh berkas tersebut disatukan ke dalam satu *dataset* untuk dilakukan analisis secara lebih lanjut. Tabel 3 menunjukkan rangkuman dari fitur-fitur yang telah berhasil dikumpulkan dari berkas rekam medis pasien milik RSIJPK.

Tabel 3. Fitur yang Berhasil Diambil.

No	Fitur	Peran	Pecahan dari Data
1	Risiko Preeklamsia	Target	-
2	Pusing	Prediktor	Keluhan
3	Pandangan Kabur	Prediktor	Keluhan
4	Kejang	Prediktor	Keluhan
5	Bagian Tubuh Bengkak	Prediktor	Keluhan
6	Sakit Perut/Mual	Prediktor	Keluhan
7	Umur	Prediktor	Umur
8	Umur Kehamilan	Prediktor	Umur Kehamilan
9	Tekanan Darah Diastolik	Prediktor	Tekanan Darah
10	Tekanan Darah Sistolik	Prediktor	Tekanan Darah
11	Kehamilan ke-	Prediktor	Kehamilan ke-
12	Riwayat Preeklamsia	Prediktor	Sejarah Penyakit
13	Riwayat Hipertensi Keluarga	Prediktor	Sejarah Penyakit
14	Riwayat Hipertensi	Prediktor	Sejarah Penyakit
15	Protein (Albumin Urea)	Prediktor	Albumin Urea
16	Trombosit	Prediktor	Trombosit
17	Hb	Prediktor	Hb
18	SGOT	Prediktor	SGOT
19	SGPT	Prediktor	SGPT
20	Bilirubin	Prediktor	Bilirubin

Tidak seluruh fitur yang disebutkan pada Tabel 1 dan Tabel 2 berhasil diserap dari data rekam medis pasien. Fitur tinggi badan dan berat badan tidak digunakan karena data tersebut tidak selalu tersedia untuk berkas rekam medis pasien ibu hamil sehingga dapat menyebabkan nilai *missing value* yang besar.

3.3. Prapemrosesan Data

Langkah pertama yang dilakukan yaitu penyesuaian bentuk data. Untuk melakukan pembersihan data, eksplorasi data, dan pembangunan model, tipe-tipe data yang digunakan perlu untuk dilakukan transformasi terlebih dahulu agar tipe data dapat sesuai dan seluruh tahapan dari analisis data dapat berjalan dengan mulus. Untuk bagian transformasi data, akan dilakukan beberapa penyesuaian sebagai berikut:

1. Fitur tekanan darah dibagi menjadi dua bagian yaitu tekanan darah sistolik dan tekanan darah diastolik.
2. Fitur keluhan dibagi menjadi lima bagian yaitu pusing, pandangan kabur, kejang, bengkak, dan sakit perut/mual.
3. Fitur sejarah penyakit dibagi menjadi tiga bagian yaitu riwayat preeklamsia, riwayat hipertensi keluarga, dan riwayat hipertensi pribadi.
4. Fitur bilirubin dibagi menjadi dua bagian yaitu positif dan negatif yang akan ditandai dengan *label* 1 atau 0.
5. Fitur protein dibagi menjadi dua bagian yaitu positif dan negatif yang akan ditandai dengan *label* 1 atau 0.

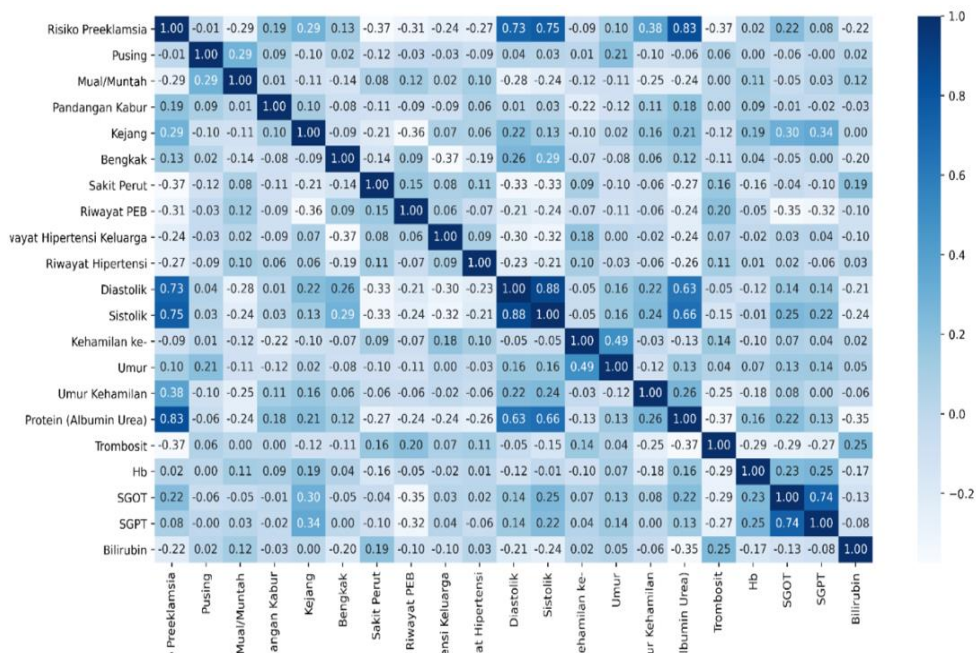
Dengan begitu, total fitur deteksi yang akan digunakan berjumlah sebanyak 19 buah. Setelah itu, akan ditambahkan satu fitur untuk dijadikan sebagai fitur target yaitu fitur Risiko preeklamsia. Risiko preeklamsia akan diberikan *label* positif apabila memiliki probabilitas yang tinggi untuk mengidap penyakit preeklamsia, sedangkan untuk *label* negatif, akan diberikan kepada pasien yang memiliki probabilitas yang rendah untuk mengidap penyakit preeklamsia.

Setelah fitur-fitur berhasil diidentifikasi secara lengkap, dilakukan transformasi data dengan menggunakan metode *label encoding* untuk fitur yang memiliki tipe data nominal agar menjadi tipe numerik sehingga dapat dimasukkan ke dalam model yang akan dibangun. Untuk fitur Pusing, Pandangan Kabur, Kejang, Bagian Tubuh Bengkak, dan Sakit Perut/Mual, perubahan tipe data nominal menjadi tipe numerik dilakukan melalui aplikasi *Microsoft Excel*, sedangkan untuk fitur-fitur lainnya, *Python* digunakan untuk mentransformasikan data. Untuk fitur protein (albumin urea) dan fitur bilirubin, nilai dari fitur tersebut diubah menjadi berlabel 1 atau 0 yang mengindikasikan keberadaan dari fitur-fitur tersebut di dalam tubuh pasien atau tidak. Jumlah yang semakin banyak tidak memiliki pengaruh yang signifikan terhadap tujuan utama dari sistem yang dibangun yaitu untuk mendeteksi Preeklamsia.

Setelah fitur-fitur pada *dataset* telah diidentifikasi secara lengkap serta dilakukan transformasi data agar dapat dimasukkan ke dalam model yang dibangun, langkah yang harus dilakukan berikutnya adalah tahapan pembersihan data. Pada tahapan ini, dilakukan penghapusan fitur yang memiliki peran kurang atau tidak signifikan terhadap fitur target. Selain dari itu, fitur yang memiliki nilai *missing value* lebih dari 50% juga dihapus dari peran prediktor. Fitur yang memiliki nilai *missing value* lebih dari 50% dianggap kurang efektif untuk dijadikan prediktor dalam model karena tidak dapat dilakukan imputasi nilai secara efektif (Hartini, 2016).

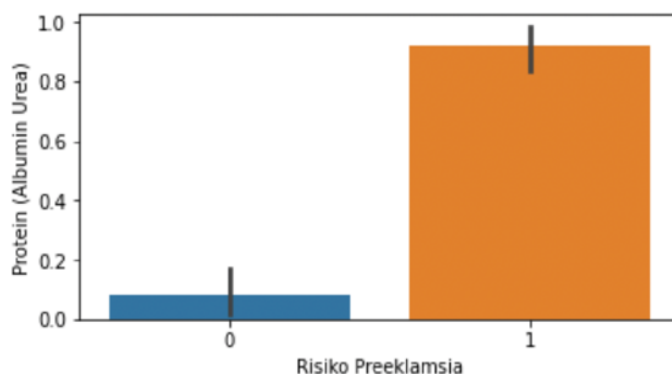
Terdapat 3 fitur yang memiliki jumlah *missing value* lebih besar atau sama dengan 50% yaitu fitur SGOT, SGPT, dan Bilirubin. Ketiga atribut tersebut dihapus dari model. Kemudian, dilakukan imputasi terhadap *missing values* yang terdapat pada *dataset*. Metode imputasi yang dilakukan ada dengan metode *similar case imputation* yaitu dengan cara menggantikan nilai kosong pada suatu kolom dengan nilai rata-rata untuk fitur numerik dan nilai modus untuk fitur nominal.

Setelah dilakukan pembersihan data, selanjutnya dilakukan pemilihan fitur prediktor. Sebelum tahapan konstruksi model, perlu dipilih fitur-fitur prediktor yang memiliki signifikansi yang tinggi pada fitur target. Tujuan dari dilakukannya proses pemilihan fitur prediktor adalah agar dapat mengoptimalkan performa model, mengurangi ongkos komputasi dan pengumpulan data, serta mengurangi kebisingan yang dapat mengakibatkan bias pada model. Oleh karena itu, perlu dilakukan *exploratory data analysis* untuk memvisualisasikan dan memutuskan jenis-jenis fitur yang akan dimasukkan ke dalam model deteksi preeklamsia. Pada Gambar 4, dapat dilihat bahwa terdapat beberapa fitur yang memiliki nilai korelasi tinggi, korelasi sedang, dan korelasi rendah terhadap target.

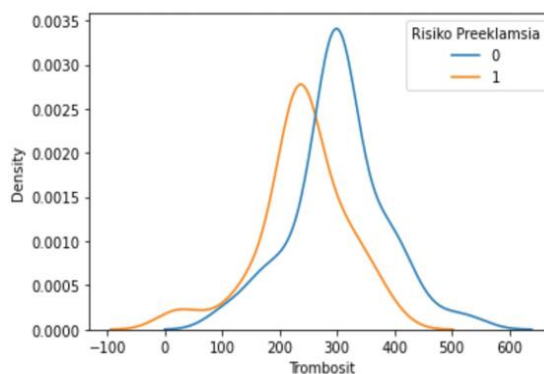


Gambar 5. Heatmap Fitur-Fitur Prediktor.

Besar angka korelasi akan berkisar antara nilai 0 sampai dengan 1. Selanjutnya, akan dipilih fitur-fitur yang memiliki angka korelasi lebih besar atau sama dengan 0,20. Prediktor yang memiliki nilai korelasi tertinggi adalah variabel Protein (Albumin Urea) dengan nilai korelasi sebesar 0,83, sedangkan prediktor yang memiliki nilai korelasi rendah namun masih dapat dipertimbangkan untuk dimasukkan ke dalam model deteksi Preeklamsia adalah fitur pandangan kabur dengan nilai korelasi sebesar 0,19. Fitur pandangan kabur memiliki koefisien korelasi yang sangat rendah. Namun, karena nilainya sangat mendekati tingkat hubungan rendah dan keluhan fisik pandangan kabur divalidasi oleh dokter sebagai salah satu fitur yang dapat digunakan untuk mendeteksi Preeklamsia, maka fitur pandangan kabur tetap dimasukkan ke dalam model. Selanjutnya, variabel prediktor yang memiliki tipe data nominal akan dibuatkan visualisasi *bar chart* agar dapat dilihat perbandingan probabilitas suatu kategori terhadap fitur target, sedangkan variabel prediktor yang memiliki tipe data numerik dibuatkan visualisasi *kdeplot* agar dapat ditemukan apabila terdapat nilai atau titik-titik tertentu yang cenderung untuk memiliki pengaruh probabilitas lebih besar terhadap fitur target. Contoh dari *bar chart* yang dibuatkan untuk fitur prediktor bertipe nominal dapat dilihat pada Gambar 6. Dari gambar tersebut, dapat dilihat bahwa pasien-pasien yang memiliki risiko positif untuk mengidap penyakit preeklamsia cenderung untuk memiliki keberadaan protein pada darah, diindikasikan dengan ukuran dari *bar* yang lebih tinggi terhadap nilai rata-rata dari fitur protein (albumin urea). Contoh dari *kdeplot* yang dibuatkan untuk fitur prediktor terhadap fitur target dapat dilihat pada Gambar 7. Dari gambar tersebut, dapat dilihat bahwa pasien-pasien yang memiliki risiko positif untuk mengidap preeklamsia cenderung untuk memiliki jumlah sel darah putih atau trombosit yang lebih rendah. Dapat disimpulkan bahwa fitur prediktor yang akan dimasukkan ke dalam model deteksi yaitu Mual/Muntah, Pandangan Kabur, Kejang, Bengkak, Sakit Perut, Riwayat PEB, Riwayat Hipertensi Keluarga, Riwayat Hipertensi, Umur Kehamilan, Diastolik, Sistolik, Protein (Albumin Urea), dan Trombosit.



Gambar 6. Bar Chart Fitur Protein vs Fitur Target.



Gambar 7. Kdeplot Persebaran Angka Trombosit terhadap Risiko Preeklamsia.

Setelah data dibersihkan dan siap untuk dimasukkan ke dalam model deteksi, langkah yang harus dilakukan berikutnya adalah membagi *dataset* menjadi 2 bagian, yaitu *train set* dan *test set*. *Train set* merupakan bagian yang akan digunakan untuk melakukan pembangunan dengan melatih model, sedangkan *test set* merupakan bagian yang akan digunakan untuk melakukan pengujian terhadap performa model. Proporsi yang akan digunakan adalah 70% untuk *train set* dan 30% untuk *test set*. Penggunaan proporsi 70% untuk *train set* dan 30% untuk *test set* dalam pembagian *dataset* adalah praktek yang umum dilakukan (Prabhakaran, 2023). Selain itu, studi empiris menunjukkan bahwa hasil terbaik dapat diperoleh jika 20-30% data digunakan untuk pengujian, dan sisanya 70-80% data digunakan untuk pelatihan (Gholamy dkk., 2018). Namun, proporsi pembagian dataset ini tidak bersifat kaku dan bisa disesuaikan berdasarkan kebutuhan penelitian (Prabhakaran, 2023). Pada penelitian ini, pembagian dataset menghasilkan data train set sebanyak 76 dan data test set sebanyak 33.

3.4. Pembangunan Model Deteksi

Setelah fitur-fitur prediktor yang akan dimasukkan ke dalam model deteksi telah ditentukan serta *dataset* telah dibagi menjadi 2 bagian, yaitu *train set* dan *test set*, tahapan berikutnya adalah dengan menggunakan berbagai algoritma berbeda untuk membandingkan hasil dan performa dari model yang berhasil dibangun. Algoritma klasifikasi yang akan digunakan pada tahapan konstruksi model meliputi algoritma *decision tree*, *support vector machine*, *naïve bayes*, *logistic regression*, *k-nearest neighbors*, dan *random forest*.

Tahapan utama yang dilakukan pada konstruksi model adalah menetapkan variabel *random state* agar menghindari kemungkinan untuk mendapatkan hasil performa yang berbeda-beda setiap kali model dibangun dengan menggunakan algoritma yang sama. Selain dari itu, penggunaan nilai *random state* yang sama membuat perbandingan dari performa model menjadi semakin objektif karena menggunakan persebaran data yang sama. Pada Tabel 4 dapat dilihat bahwa algoritma *logistic regression* memiliki performa yang terbaik dan tidak mengalami *overfitting* diindikasikan oleh nilai *test score* dan *train score* yang sama. Oleh karena itu, 10-fold *cross validation* dilakukan pada algoritma *logistic regression* dengan kombinasi dari parameter-parameter yang digunakan pada model, atau yang biasa disebut dengan *hyperparameter tuning*. Percobaan kombinasi dari parameter-parameter tersebut sehingga didapatkan nilai yang optimal dapat dilakukan dengan menggunakan metode *grid search*. *Hyperparameter* yang akan dicari kombinasi terbaiknya terdiri dari Nilai C, solver, dan max iteration dengan nilai-nilai sebagai berikut.

1. Nilai C [100, 10, 1.0, 0.1, 0.01]
2. Solver ['lbfgs', 'newton-cg', 'liblinear', 'sag', 'saga']
3. Max Iteration [100, 1000, 2500, 5000]

3.5. Evaluasi Model Deteksi

Setelah seluruh model berhasil dibangun, perlu dilakukan evaluasi model untuk membandingkan performansi dari seluruh model deteksi yang telah dibangun dengan berbagai algoritma yang berbeda. Pada Tabel 4 dapat dilihat rangkuman dari performa seluruh model yang dibangun dalam penelitian ini.

Tabel 4. Performa Model.

Algoritma	Akurasi	Presisi	Recall	F1-Score	AUC	Train Score	Test Score
<i>Decision Tree</i>	100%	100%	100%	100%	100%	100%	100%
<i>Support Vector Machine</i>	100%	100%	100%	100%	100%	99%	100%
<i>Naïve Bayes</i>	73%	91%	56%	69%	74%	84%	73%
<i>Logistic Regression</i>	97%	95%	100%	97%	97%	97%	97%
<i>K-Nearest Neighbors</i>	94%	100%	89%	94%	94%	95%	94%
<i>Random Forest</i>	97%	100%	94%	97%	97%	100%	97%

Dalam melakukan evaluasi model, metrik akurasi dan presisi dipilih sebagai metrik utama dalam keputusan pemilihan model. Alasan digunakan *precision* adalah karena model tidak ingin salah mendeteksi pasien yang sebenarnya memiliki risiko tinggi untuk mengidap preeklamsia namun dideteksi tidak memiliki risiko tinggi untuk mengidap preeklamsia oleh model. Hal tersebut dapat berakibat fatal dan bahkan dapat memakan korban manusia.

Pada proses pemilihan model, tingkat performansi model yang mencapai 100% untuk seluruh metrik bukan merupakan performa yang baik karena ada kemungkinan bahwa model yang dibuat memiliki kendala *overfitting* (Ying, 2019). Hal tersebut dapat memiliki arti bahwa model yang dibangun dapat memprediksi data yang tersedia dengan baik, namun ada kemungkinan bahwa tidak dapat memprediksi data-data yang belum pernah dilihat sebelumnya dengan akurat. Menurut (Ying, 2019), terdapat beberapa hal yang dapat dilakukan untuk mengurangi *overfitting* pada *dataset*, diantaranya yaitu *early stopping*, *regularization*, dan memilih model dengan kompleksitas rendah. Dalam pembangunan model deteksi Preeklamsia pada penelitian ini, telah dilakukan beberapa langkah tersebut dalam upaya untuk mengurangi potensi *overfitting* sebagai berikut.

1. *Early stopping*

Pada penelitian ini telah dilakukan percobaan terhadap beberapa proporsi antara data pelatihan dengan data pengujian yang berbeda-beda. Tahapan tersebut dilakukan untuk menemukan titik dimana penghentian pelatihan data perlu untuk dilakukan sehingga ditemukan titik yang memiliki keseimbangan antara performa pelatihan dengan performa pengujian.

2. *Regularization*

Pada penelitian ini dilakukan pemilihan fitur prediktor untuk mengurangi dimensi dari *dataset* sebelum dimasukkan ke dalam model deteksi. Pada pembelajaran mesin, semakin banyak fitur yang digunakan untuk dijadikan sebagai data pelatihan, maka tingkat kerumitan dari model juga akan meningkat. Model-model yang cenderung untuk mengalami *overfitting* adalah model-model yang memasukkan seluruh fiturnya ke dalam model yang dibangun. Terlebih lagi apabila *dataset* yang digunakan berjumlah relatif kecil (Ying, 2019). Oleh karena itu, pada pembangunan model deteksi Preeklamsia pada pasien kehamilan, hanya digunakan fitur-fitur yang memiliki signifikansi tinggi terhadap diagnosis berdasarkan hasil perhitungan statistik dan juga hasil validasi oleh dokter spesialis kandungan.

3. Memilih Model dengan Kompleksitas Rendah

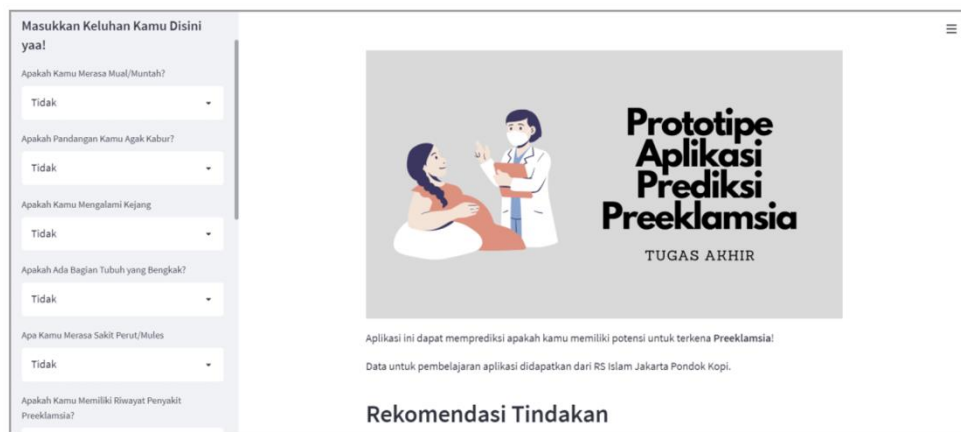
Perlu diingat bahwa jumlah data yang digunakan untuk pelatihan dan pengujian model berjumlah relatif kecil. Oleh karena itu, kompleksitas model juga memiliki peranan penting untuk mengurangi potensi *overfitting*. Pelatihan model secara sederhana adalah kegiatan yang dilakukan untuk membuat parameter-parameter dari model agar memiliki nilai yang optimal sehingga dapat menghasilkan performa yang memuaskan. Semakin kompleks model, semakin banyak parameter yang perlu untuk ditemukan nilai optimalnya, mengakibatkan jumlah data pelatihan yang diperlukan juga meningkat. Oleh karena itu, pilihan yang dapat dilakukan adalah antara menambah jumlah data pelatihan atau menurunkan tingkat kompleksitas dari model (Ying, 2019). Maka, model *logistic regression* dipilih karena merupakan salah satu model yang memiliki tingkat kompleksitas rendah apabila dilihat dari jumlah *hyperparameter*. Penambahan jumlah pelatihan tidak dapat dilakukan karena terdapat beberapa kebijakan dari rumah sakit yang membatasi jumlah data yang dapat diambil.

Berdasarkan matrik evaluasi model yang dipilih yaitu tingkat akurasi dan tingkat presisi, *logistic regression* dipilih sebagai algoritma yang dipilih karena memiliki tingkat akurasi dan presisi yang tinggi yaitu sebesar 97% dan 95%. Untuk melihat apakah model memiliki kecenderungan untuk *overfit*, dapat dilihat dari nilai *train score* dan *test score*. Model yang dibangun memiliki nilai *train score* dan *test score* sebesar 97% dan 97%. Dengan demikian, dapat disimpulkan bahwa model yang dibangun tidak mengalami *overfit* karena nilai akurasi dari data pelatihan dan nilai akurasi dari data-data yang belum pernah dilihat sebelumnya yaitu *data testing* memiliki nilai yang hampir sama (Subramanian & Simon, 2013).

Setelah itu, dilakukan perubahan proporsi *data splitting* dari 70:30 menjadi 55:45 yang meningkatkan performansi dari model menjadi memiliki tingkat akurasi sebesar 98% dan tingkat presisi sebesar 100% tanpa mengalami *overfit* karena memiliki nilai tes dan nilai latih yang sama yaitu sebesar 98%. Terakhir, dilakukan proses *hyperparameter tuning* dalam upaya untuk meningkatkan performansi model. Namun, kegiatan tersebut justru menurunkan nilai akurasi dari model menjadi 91%. Oleh karena itu, digunakan pembangunan model yang tidak menggunakan *hyperparameter tuning* dengan proporsi *splitting data* 55:45.

3.6. Perancangan Prototipe Sistem

Setelah model dibangun, dievaluasi, dan siap untuk digunakan oleh pengguna, langkah yang harus dilakukan berikutnya adalah merancang prototipe yang bersesuaian dengan kebutuhan pengguna. Tahapan ini merupakan fase *deployment* pada siklus CRISP-DM. Model deteksi yang telah dibangun akan ditampilkan dan digunakan melalui prototipe aplikasi sederhana. Halaman utama dari prototipe aplikasi yang dirancang pada penelitian ini dapat dilihat pada Gambar 8.



Gambar 8. Halaman Utama Prototipe Aplikasi.

Kesimpulan

Algoritma yang digunakan untuk membangun model deteksi Preeklamsia pada pasien kehamilan di RSIJPK adalah algoritma *logistic regression* karena memiliki performa dan generalisasi terbaik apabila dibandingkan dengan algoritma lainnya. Algoritma *logistic regression* dengan proporsi *splitting data* sebesar 55:45 menghasilkan nilai akurasi sebesar 98% dan nilai presisi sebesar 100%. Algoritma *decision tree* yang juga digunakan dalam penelitian ini menghasilkan nilai akurasi sebesar 100% dan nilai presisi sebesar 100%, hal ini mengindikasikan bahwa algoritma *decision tree* mengalami *overfitting*. Berdasarkan hasil diskusi dengan dokter spesialis kandungan, studi literatur yang dilakukan sebelum penelitian, dan proses eksplorasi serta pra pemrosesan data, diperoleh sebanyak 12 fitur prediktor Preeklamsia yang terdiri dari protein (albumin urea), diastolik, sistolik, umur kehamilan, trombosit, sakit perut, riwayat PEB, kejang, mual/muntah, riwayat hipertensi, riwayat hipertensi keluarga, dan pandangan kabur.

Pada penelitian ini juga dirancang suatu sistem deteksi yang dapat digunakan oleh *user* untuk melakukan deteksi terhadap penyakit preeklamsia pada pasien kehamilan. *User* yang dimaksud terbagi menjadi 3 pihak yaitu dokter spesialis kandungan, tenaga medis RSIJPK, dan pasien kehamilan. Dalam melakukan perancangan sistem deteksi preeklamsia, dilakukan identifikasi kebutuhan fungsional dan nonfungsional dari sistem berdasarkan kebutuhan dari pihak RSIJPK. Rancangan sistem dibangun dalam bentuk prototipe yang dibuat dengan menggunakan *packages streamlit python*. Setelah prototipe berhasil dibuat, dokter spesialis kandungan akan secara langsung melakukan verifikasi dan validasi atas tampilan, fitur, dan performa dari sistem yang dibangun. Sistem deteksi Preeklamsia yang dibangun diharapkan dapat membantu pihak RSIJPK agar dapat mengidentifikasi pasien-pasien yang memiliki risiko tinggi untuk mengidap penyakit preeklamsia.

Berdasarkan hasil penelitian yang dilakukan, diajukan beberapa saran untuk penelitian berikutnya yakni sebagai berikut:

1. Menambahkan jumlah data pelatihan agar dapat meningkatkan tingkat keyakinan dari korelasi antara fitur prediktor dengan fitur target. Selain dari itu, menurunkan kemungkinan untuk terjadi *overfitting*.
2. Menggunakan fitur-fitur seperti SGOT, SGPT, dan bilirubin karena memiliki pengaruh yang signifikan terhadap hasil diagnosis risiko preeklamsia pada pasien. Untuk itu, perlu dilakukan pengumpulan data-data pasien yang telah melakukan uji laboratorium terkait dalam jumlah yang lebih besar.
3. Menambahkan fitur-fitur yang belum tersedia pada data rekam medis milik RSIJPK misalkan atribut detak jantung janin pada kandungan karena memiliki signifikansi yang cukup tinggi untuk mendiagnosis risiko preeklamsia pada pasien kehamilan berdasarkan studi literatur yang dilakukan

Daftar Pustaka

- Baharuddin, M., Amelia, D., Suhowatsky, S., Kusuma, A., Suhargono, M. H., & Eng, B. (2019). Maternal death reviews: A retrospective case series of 90 hospital-based maternal deaths in 11 hospitals in Indonesia. *International Journal of Gynecology and Obstetrics*, 144, 59–64.
- Chaemsaihong, P., Sahota, D. S., & Poon, L. C. (2022). First trimester preeclampsia screening and prediction. *American Journal of Obstetrics and Gynecology*, 226(2), 1071-1097.
- Draelos, R. (2019). *Measuring Performance: The Confusion Matrix*. <https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/>
- Gholamy, A., Kreinovich, V., & Kosheleva, O. (2018). *Why 70/30 or 80/20 relation between training and testing sets: A pedagogical explanation*. https://scholarworks.utep.edu/cs_techrep/1209/
- Hartini, E. (2016). Efficiency comparison of method of handling missing value in data evaluation system or component. *Prosiding Seminar Nasional Teknologi Energi Nuklir 2016*. <https://karya.brin.go.id/id/eprint/1561/>
- Kustiyahningsih, Y., Mula'ab, & Hasanah, N. (2020). Metode Fuzzy ID3 untuk klasifikasi status preeklamsi ibu hamil. *Teknika*, 9(1), 74–80.
- Marić, I., Tsur, A., Aghaeepour, N., Montanari, A., Stevenson, D. K., Shaw, G. M., & Winn, V. D. (2020). Early prediction of preeclampsia via machine learning. *American Journal of Obstetrics & Gynecology MFM*, 2(2), 100100.
- Nair, A. (2021). *Improve Your Hyperparameter Tuning Experience With The Random Search*. <https://towardsdatascience.com/improve-your-hyperparameter-tuning-experience-with-the-random-search-2c05d789175f>
- Niakšu, O. (2015). CRISP data mining methodology extension for medical domain. *Baltic J. Modern Computing*, 3(2), 92–109.
- Obenshain, M. K. (2004). Application of data mining techniques to healthcare data. *Infection control and hospital epidemiology*, 25(8), 690–695.
- Pittara. (2022). *Preeklamsia*. <https://www.alodokter.com/preeklamsia>
- Poon, L. C., & Nicolaides, K. H. (2014). Early prediction of preeclampsia. *Obstetrics and Gynecology International*, 2014, 1–11.
- Prabhakaran, S. (2023). *Train test split - How to split data into train and test for validating machine learning models?* <https://www.machinelearningplus.com/machine-learning/train-test-split/>
- Roberts, J. M., & Gammill, H. S. (2005). Preeclampsia. *Hypertension*, 46(6), 1243–1249.
- Silwattananusarn, T., & KulthidaTuamsuk, A. (2012). Data mining and its applications for knowledge management : A literature review from 2007 to 2012. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, 2(5). <https://doi.org/10.5121/ijdkp.2012.2502>
- Subramanian, J., & Simon, R. (2013). Overfitting in prediction models – is it a problem only in high dimensions? *Contemporary Clinical Trials*, 36(2), 636–641.
- Sufriyana, H., Wu, Y.-W., & Su, E. C.-Y. (2020). Artificial Intelligence-assisted prediction of preeclampsia: Development and external validation of a nationwide health insurance dataset of the BPJS kesehatan in Indonesia. *EBioMedicine*, 54, 102710.
- World Health Organization. (2023). *Maternal mortality*. <https://www.who.int/news-room/fact-sheets/detail/maternal-mortality>
- Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, 1168, 022022.
- Zhang, J., Han, L., Li, W., Chen, Q., Lei, J., Long, M., Yang, W., Li, W., Zeng, L., & Zeng, S. (2019). Early prediction of preeclampsia and small-for-gestational-age via multi-marker model in Chinese pregnancies: A prospective screening study. *BMC Pregnancy and Childbirth*, 19(1).