JURNAL TIMES
TECHNOLOGY INFORMATICS & COMPUTER SYSTEM

Volume XIII No 2, Desember 2024
*p*ISSN : 2337 – 3601
*e*ISSN : 2549 – 015X

Tersedia *online* di http://ejournal.stmik-time.ac.id

PREDIKSI EPIDEMIOLOGI PENYAKIT TIDAK MENULAR MENGGUNAKAN ALGORITMA *RANDOM FOREST* PADA PUSKESMAS

Ferry Saptawan[1)], David[2)], Tony Wijaya[3)], Sandy Kosasi[4)], Susanti Margaretha kuway[5)]
Program Studi Teknik Informatika
STMIK Pontianak
Jl. Merdeka Barat No 372
email: ferry.saptawan@gmail.com[1)], david@stmikpontianak.ac.id[2)], tony_wijaya@stmikpontianak.ac.id[3)], sandykosasi@stmikpontianak.ac.id[4)], shantykuway@stmikpontianak.ac.id[5)]

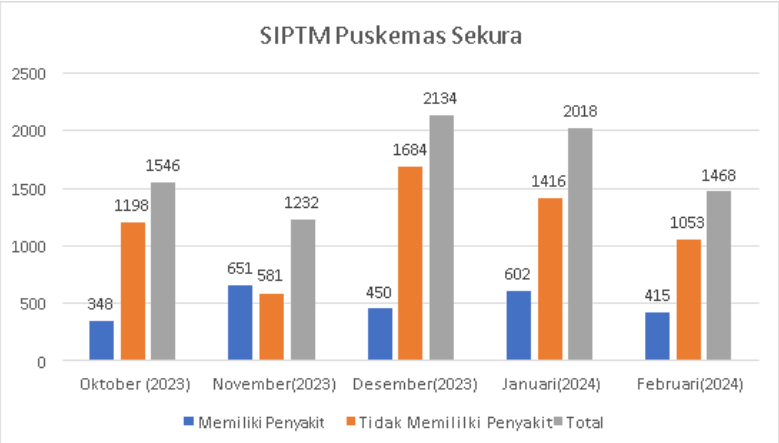
Abstrak

Puskesmas Sekura yang merupakan puskesmas rawat inap dan terletak di kabupaten sambas kecamatan sekura menghadapi tantangan dalam mengurangi angka penderita penyakit tidak menular. Penelitian ini bertujuan mengatasi masalah tersebut dengan menerapkan machine learning untuk proses klasifikasi penyakit berdasarkan data rekam medis. Algoritma yang digunakan pada penelitian kali ialah *Random Forest*. Metode yang digunakan adalah *Design Science Research* (DSR) dengan enam tahap: identifikasi masalah, tujuan solusi, perancangan, pengembangan, demonstrasi, evaluasi, dan komunikasi. Penelitian ini menghasilkan perangkat lunak berbasis web menggunakan framework CodeIgniter dan PHP untuk memudahkan proses prediksi. Dalam penelitian kali ini algoritma *Random Forest* dipastikan dapat melakukan prediksi penyakit tidak menular, namun untuk tingkat akurasinya masih sangat rendah. Puskesmas sekura harus menambahkan lebih banyak data pada sistem prediksi ini agar tingkat akurasi menjadi bertambah. Hasil akurasi 27.64% tergolong cukup rendah untuk dapat digunakan dalam prediksi, sistem masih harus dilatih dengan menambahkan dataset yang dibuat dari data rekam medis pasien. Puskesmas sekura masih belum dapat menggunakan sistem prediksi ini secara langsung kelapangan.

Kata Kunci: *Epidemiologi, Random Forest,* Penyakit Tidak Menular, *Machine Learning.*

1. Pendahuluan

Kesehatan adalah hal penting dalam hidup, dan menjadi mahal saat penyakit menyerang. Pencegahan terbaik memerlukan analisis faktor penyebab penyakit, seperti yang dipelajari dalam epidemiologi. Epidemiologi mempelajari distribusi penyakit dan kondisi kesehatan serta faktor risiko yang terkait [1]. Sosialisasi untuk meningkatkan kesadarakan masyarakaat perlu dilakukan, seperti yang dilakukan oleh Puskesmas Sekura. Berdasarkan karakteristik wilayah, Puskesmas Sekura merupakan Puskesmas kawasan pedesaan, sedangkan berdasarkan kemampuan penyelenggaraan termasuk dalam kategori Puskesmas Rawat Inap. Wilayah kerja sebanyak 10 desa di wilayah Kecamatan Teluk Keramat. UPT Puskesmas Sekura didukung jejaring di bawahnya sebanyak 4 Pustu, 10 Poskesdes dan 27 Posyandu Balita serta 20 Posbindu. Jumlah sekolah 16 Paud, 4 TK, 22 SD, 7 SLTP, 4 SLTA. Puskesmas Sekura memiliki program-program kesehatan yang dapat meningkatkan kesadaran masyarakat, salah satunya penyuluhan kesadaran akan bahaya penyakit tidak menular. Penyakit tidak menular merupakan penyakit atau kondisi medis yang tidak dapat ditularkan dari satu individu ke individu lainnya [2]. Dengan adanya program tersebut diharapkan dapat menurunkan angka masyarakat penyandang penyakit tidak menular. Namun, kenyataannya jumlah masyarakat yang mengidap penyakit tidak menular tidak menurun sama sekali dibuktikan dengan grafik berikut yang akan ditampilkan pada gambar 1.



Gambar 1. Grafik SIPTM Puskesmas Sekura

Solusi untuk masalah yang dihadapi puskesmas sekura adalah diperlukan penyusunan perencanaan yang baik agar dapat menurunkan penderita penyakit tidak menular, dalam mewujudkannya puskesmas sekura harus melakukan kegiatan screening lebih sering dan lebih tersebar. Sistem klasifikasi sangat diperlukan pada permasalahan kali ini, karena dengan adanya sistem klasifikasi puskesmas sekura dapat melakukan kegiatan screening lebih sering dan lebih tersebat areanya karena sistem klasifikasi dapat mengoptimalkan waktu dan tenaga pekerja yang dibutuhkan. Klasifikasi merupakan teknik dalam data mining untuk mengelompokkan data berdasarkan keterikatan data terhadap data sampe [3]. Algoritma *Random Forest* dapat menggunakan untuk mengklasifikasi big data. Prunning atau pemangkasan variable seperti descission tree tidak terdapat dalam algoritma *Random Forest* namun keunggulan dari *Random Forest* dapat menggabungkan banyak pohon dan untuk single tree yang terdiri atas satu pohon dalam melakukan klasifikasi dan prediksi kelas [4]. *Random Forest* juga merupakan bagian dari machine learning. Penerapan machine learning cukup lazim dan sering dilakukan pada jaman modern. Machine learning dapat mempermudah pekerjaan dalam melakukan analisis berdasarkan data yang sudah dimiliki dengan mengkomputasi data input untuk mencapai tugas yang diinginkan untuk menghasilkan hasil tertentu [5]. Dataset yang akan digunakan diolah dari data rekam medis pasien puskesmas sekura.

Penelitian sebelumnya oleh [6], penerapan algoritma Decision Tree, Naïve Bayes, k-Nearest Neighbour, *Random Forest*, dan Decison Stump dalam prediksi penyakit jantung, menunjukkan *Random Forest* dan decision stump memiliki akurasi paling tinggi dibanding algoritma lain. Penelitian oleh [7], uji *Random Forest* untuk prediksi penyakit liver menunjukkan bahwa *Random Forest* mampu melakukan klasifikasi dengan akurasi 0,713326 dengan f1 score 81%. Penelitian terdahulu oleh [8] penerapan *Random Forest* untuk prediksi penyakit stroke. beberapa tahapan dilakukan dalam penelitian ini diantaranya adalah tahapan preprocessing, processing dan evaluasi. Hasil dari penelitian ini yaitu akurasi sebesar 99%.

Pada tiga jurnal terdahulu diatas hanya terdapat 1 penyakit yang di prediksi. Pada penelitian kali ini terdapat 10 penyakit yaitu hipertensi, obesitas, diabetes melitus tipe 2, paru paru basah, ppok, asma bronchiale, diabetes melitus tipe 3, glaukoma, dan presbicusis. Penelitian ini menggunakan perangkat lunak berbasis web-based, berbeda dengan peneltian sebelumnya yang berbentuk program dekstop. Penelitian kali ini juga menggunakan dataset yang dibuat dari data rekam medis pasien. Penelitian ini bisa mengembangkan pemahaman baru tentang bagaimana klasifikasi menggunakan *Random Forest* dapat mempermudah perkejaan tenaga medis. Penerapan klasifikasi menggunakan algoritma *Random Forest* dapat mengoptimalkan waktu dan tenaga medis yang bertugas dalam proses screening, sehingga screening dapat dilakukan lebih sering dan lebih tersebar areanya.

2. Landasan Teori

Proses *Random Forest* melibatkan pembuatan beberapa pohon keputusan dari subset dataset yang berbeda, di mana prediksi akhir diperoleh melalui voting untuk klasifikasi atau rata-rata untuk regresi. Langkah-langkah membangun model *Random Forest* meliputi pengambilan sampel acak dengan penggantian (bootstrapping), pembangunan pohon keputusan dari sampel tersebut, dan penggabungan prediksi dari semua pohon untuk mendapatkan hasil yang lebih akurat dan tahan terhadap overfitting.

Machine Learning

Machine learning (ML) adalah sebuah algoritma yang dapat menemukan korelasi dari karakteristik data yang diberikan. ML memerlukan data latih (training data) yang cukup untuk membangun model algoritma yang cerdas dan perlu diverifikasi akurasinya menggunakan data pengujian (testing data). Setelah dilatih, model algoritma ML ini secara mandiri mampu memberikan prediksi berupa hasil regresi, klasifikasi, atau klastering yang dapat digunakan untuk proses diagnosis dan prognosis[9].

Random Forest

Proses klasifikasi akan menggunakan algoritma *Random Forest*. Metode *Random Forest* merupakan pengembangan dari metode CART (Classification and Regression Trees), yaitu menerapkan metode bootstrap aggregating (bagging) dan random feature selection [10].

Entropi digunakan untuk mengukur tingkat ketidakpastian atau impurity dalam dataset. Semakin tinggi entropi, semakin tidak murni dataset tersebut. Berikut ini adalah gambar 8 yang merupakan rumus dari entropi

$$Entropy\ (S) = -\sum\_(i = 1)\text{^}c\ p\_i\ *\ log_2(p\_i) \quad (2.1)$$

Keterangan:

S: dataset yang akan di analisis

C: jumlah kelas yang berbeda dalam kelas S

Ꝑi: probabilitas dari kelas i, yang merupakan jumlah sampel dari kelas i dibagi dengan jumlah total sampel dalam kelas S

$$\text{Split Information (S,A)} = -\sum\_(i = 1)\text{^}c\ (|S\_i| \ / \ |S|)\ *\ log_2(|S\_i| \ / \ |S|) \quad (2.2)$$

Keterangan:
S: dataset yang akan di analisis
C: jumlah kelas yang berbeda dalam kelas S

Confusion Matrix

Confusion Matrix berfungsi untuk melihat tingkat akurasi dari hasil yang akan diklasifikasikan secara manual dengan menggunakan tabel matrix. Apabila data set terdapat dua kelas, maka yang satu dianggap positif dan lainya dianggap negative.

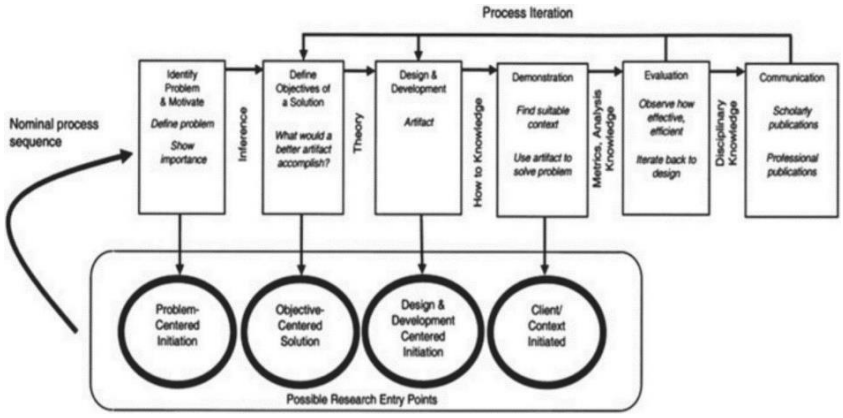
$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (2.3)$$

Keterangan:
TP: Jumlah data positif yang diprediksi benar sebagai positif oleh model
TN: Jumlah data negatif yang diprediksi benar sebagai negatif oleh model.
FP: Jumlah data negatif yang diprediksi salah sebagai positif oleh model.
FN: Jumlah data positif yang diprediksi salah sebagai negatif oleh model

3. Metode Penelitian

Bentuk penelitian yang digunakan pada penelitian ini yaitu Studi Kasus untuk memahami dan mendalami fenomena tertentu, pada penelitian ini yaitu untuk mendalami kasus atau kendala yang ada pada Puskesmas Sekura yaitu analisis epidemiologi.

Metode penelitian yang digunakan yaitu *Design Science Research* (DSR) adalah sebuah metodologi yang berorentasikan desain informasi sistem. Tujuan dari proyek penelitian DSR adalah untuk memperluas batas kemampuan manusia dan organisasi dengan merancang artefak baru dan inovatif yang diwakili oleh konstruksi, model, metode, dan instantiasi [11]. DSR juga merupakan kerangka prosedur yang digunakan untuk mempermudah penelitian di bidang teknologi informasi yang digunakan sebagai proses pemahaman serta mengulas untuk mengenali dan mengevaluasi hasil penelitian [12].



Gambar 2. Alur Proses *Design Science Research* (DSR)

Metode pengembangan yang digunakan pada penelitian ini yaitu Metode Extreme Programming (XP) yaitu pendekatan berorientasi objek yang bertujuan untuk mempersingkat waktu dalam menghasilkan suatu system dengan tahapan-tahapan yang ada[13].

Metode Pengujian yang digunakan yaitu white box testing yaitu pengujian untuk menilai di setiap bagian yang terdapat di dalam perangkat lunak dan hasilnya telah sesuai dengan rencana yang telah ditetapkan dan diinginkan[14]. Sistem pemodelan perangkat lunak yang digunakan dalam penelitian ini yaitu Unified Modelling Language. Unified Modeling Language (UML) adalah standar industri yang digunakan untuk menggambarkan, merancang, dan mendokumentasikan struktur dan perilaku sistem perangkat lunak, UML menyediakan bahasa visual yang kaya untuk menggambarkan berbagai aspek dari system [15]. Dalam merancang sebuah sistem prediksi epidemiologi penyakit tidak menular pada Puskesmas Sekura, peneliti menggunakan UML (Unified Modelling Language) sebagai alat bantu dalam perancangan pemodelan. Dalam penelitian ini menggunakan empat diagram UML yaitu usecase diagram, activity diagram, sequence diagram dan class diagram.

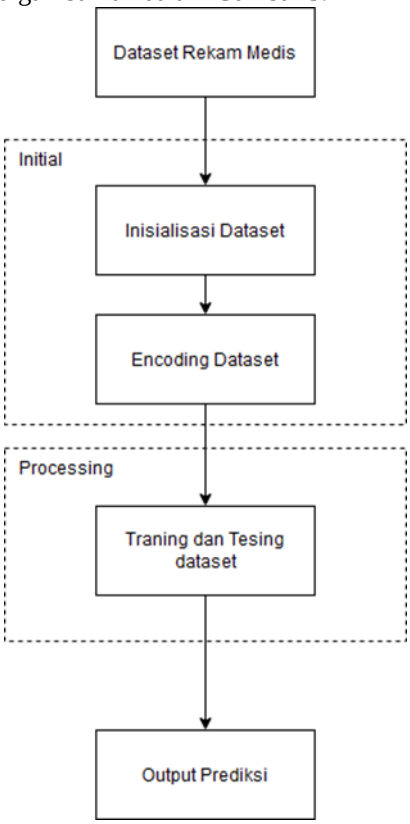
Teknik pengumpulan data yang digunakan dalam penelitian ini terdiri dari wawancara dan studi dokumentasi. Wawancara dilakukan kepada seorang pemegang program kegiatan screening di puskesmas sekura yaitu feni pratiwi berupa data rekam medis pasien yang diambil selama proses screening yang dilakukan oleh puskesmas sekura. Dokumentasi penelitian ini dilakukan dengan mempelajari data berupa arsip dokumentasi. Data sekunder adalah sumber data yang tidak memberikan data secara langsung kepada pengumpul data[16]

Dataset dibuat dari data rekam medis pasien yang diambil oleh puskesmas sekura. Data rekam medis pasien memiliki beberapa kolom yaitu, tanggal pemeriksaan, NIK, Nama Pasien, Tanggal Lahir, Jenis Kelamin, Provinsi Asal, Kab Asal, Alamat, No telp, Status Pendidikan, Pekerjaan, Status Perkawinan, Golongan Darah, Riwayat penyakit tidak menular pada keluarga, Riwayat penyakit tidak menular pada diri sendiri, faktor resiko (merokok), Faktor Resiko (Kurang Aktifitas Fisik), Pola Makan (gula berlebihan), Pola Makan (Garam Berlebihan), Lemak Berlebihan, Pola

Makan (Konsumsi Alkohol), Tekanan Darah (sistol dan diastol), IMT (Tinggi Badan dan Berat Badan), Lingkar Perut, Kadar Gula, Diagnosis. Dataset akan mengambil beberapa kolom data rekam medis

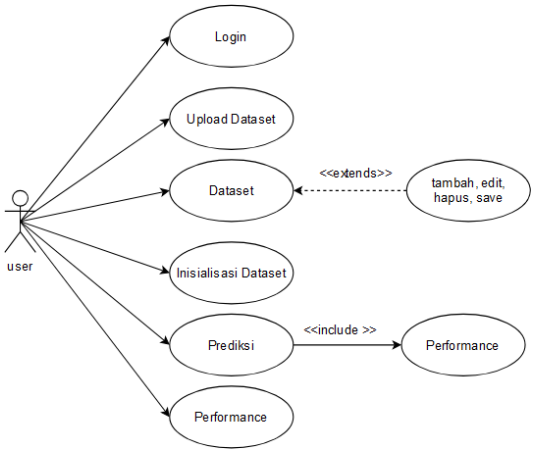
4. Hasil Penelitian

Arsitektur kerangka menunjukkan alur pemrosesan dataset rekam medis hingga menghasilkan output prediksi. Proses ini terbagi menjadi tiga tahapan utama, yaitu input, proses, dan output. Tahap pertama adalah input, di mana dataset rekam medis diinput sebagai data mentah yang akan digunakan dalam tahap selanjutnya. Pada tahap kedua, yaitu proses, terdapat dua sub-tahap: Initial dan Processing. Dalam sub- tahap Initial, dataset melalui inisialisasi dan encoding untuk menyiapkan data agar siap digunakan oleh model pembelajaran mesin. Setelah itu, pada sub-tahap Processing, dataset yang telah diproses dilatih dan diuji melalui langkah-langkah training dan testing untuk membangun model prediksi. Hasil akhir dari keseluruhan proses ini adalah output prediksi, yaitu hasil yang dihasilkan oleh model yang telah dilatih dan diuji. Output ini dapat digunakan untuk analisis lebih lanjut atau sebagai dasar pengambilan keputusan medis. Secara keseluruhan, kerangka kerja ini menjelaskan alur pengolahan data medis hingga menjadi output yang bermanfaat, dengan dataset yang disimpan di LocalhostDB sebagai basis data. Arsitektur kerangka kerja yang digambarkan dalam Gambar 3.



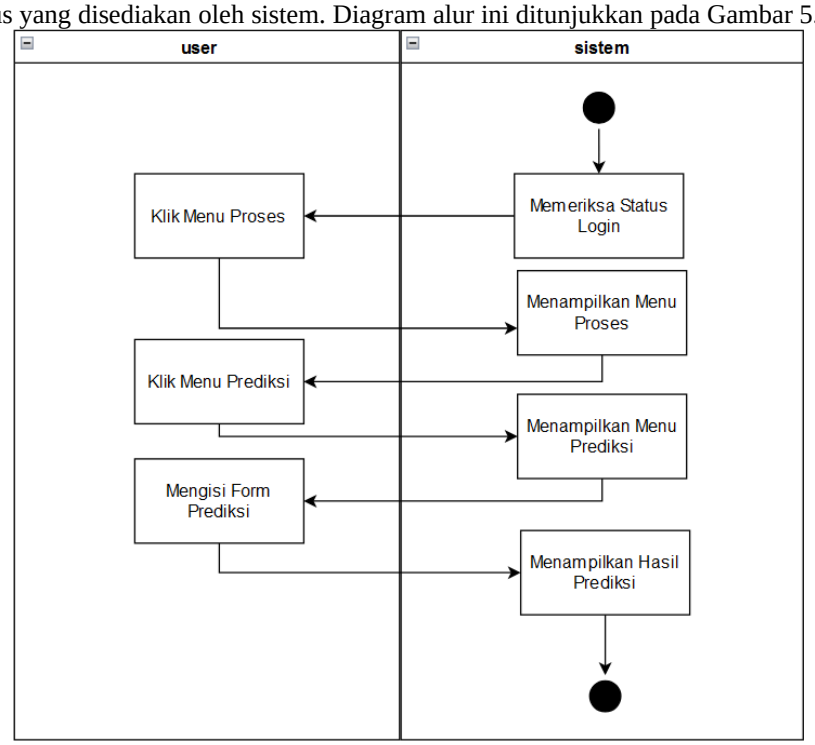
Gambar 3. Arstitektur Tahap Kerja

Use case diagram menjelaskan aktivitas-aktivitas dalam sistem. Diagram ini dimulai dengan satu aktor, yaitu user. Ketika user membuka aplikasi dan menuju halaman login, mereka harus menginput username dan password. Jika input tersebut benar, user dapat mengakses halaman utama; jika salah, mereka diminta memasukkan kembali informasi yang benar. Untuk menggunakan sistem, user harus mengunggah file dataset melalui menu "Upload Dataset" dan dapat menyunting dataset melalui menu "Dataset." User juga harus melakukan inisialisasi danencoding dataset di menu "Inisialisasi Dataset." Setelah itu, user dapat memproses dataset dengan algoritma *Random Forest* dan melakukan prediksi di menu "Prediksi." Hasil performa algoritma dapat dilihat di menu "Performance," yang menunjukkan kinerja algoritma terhadap dataset yang diinputkan. Use Case diagram akan ditampilkan pada gambar 4.



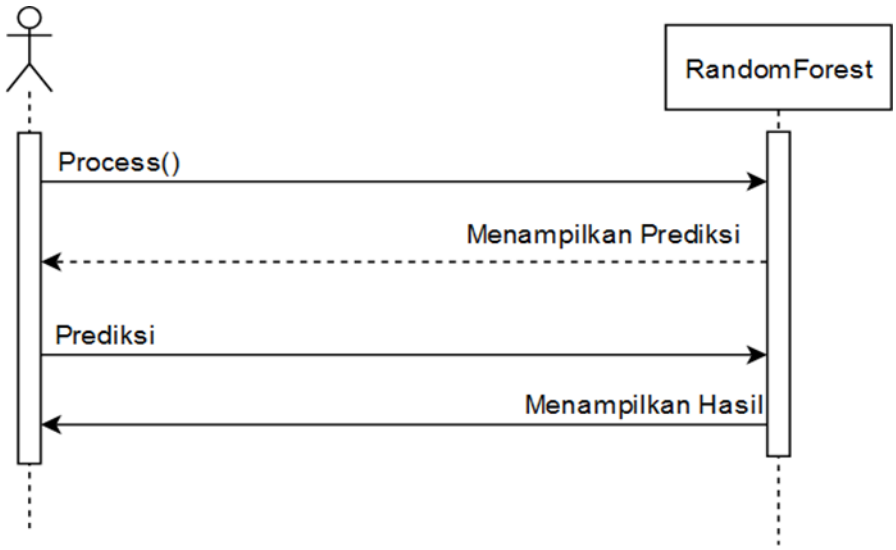
Gambar 4. Use Case Diagram

Diagram activity Prediksi menjelaskan alur proses ketika pengguna melakukan prediksi menggunakan sistem. Proses dimulai saat pengguna mengklik menu Proses untuk mengakses fitur prediksi. Sistem kemudian memeriksa status login pengguna. Jika pengguna sudah login, menu proses akan tampil, termasuk pilihan menu Prediksi. Setelah pengguna memilih menu prediksi, sistem menampilkan form prediksi yang harus diisi dengan data atau parameter yang diperlukan. Setelah form diisi, sistem memproses data tersebut dan menampilkan hasil prediksi pada halaman khusus yang disediakan oleh sistem. Diagram alur ini ditunjukkan pada Gambar 5.



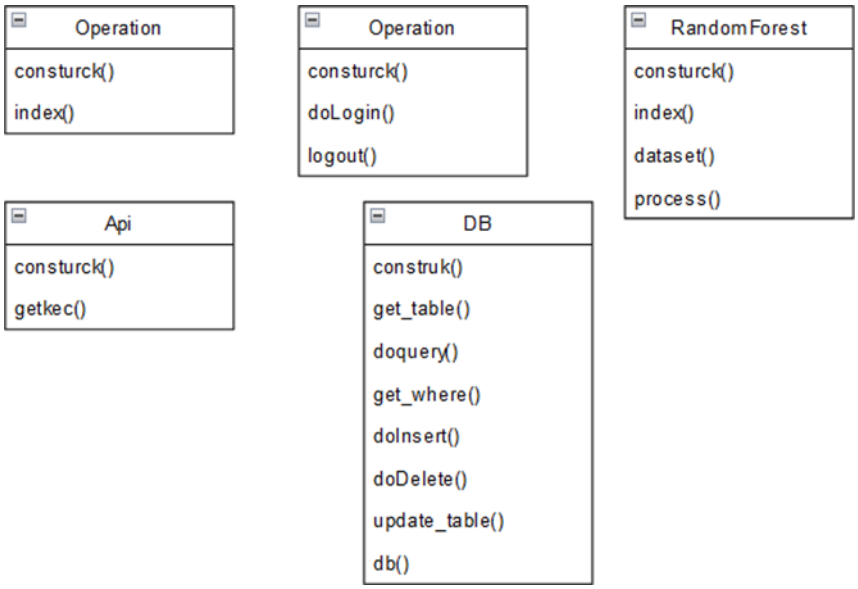
Gambar 5. Diagram Activity Prediksi

Diagram sequence Prediksi menjelaskan alur proses ketika pengguna melakukan prediksi melalui sistem. Proses dimulai saat pengguna berhasil login dan mengakses halaman utama. Sistem memverifikasi status login untuk memastikan bahwa pengguna memiliki hak akses. Setelah itu, pengguna mengklik menu Proses untuk mengakses fitur-fitur yang tersedia, termasuk opsi Prediksi. Sistem menampilkan halaman prediksi, di mana pengguna memasukkan data atau parameter yang diperlukan dalam form prediksi. Setelah form diisi, pengguna menekan tombol Prediksi untuk memulai proses. Sistem kemudian memproses data berdasarkan model atau algoritma yang ada dan menampilkan hasil prediksi kepada pengguna, yang dapat berupa angka, grafik, atau informasi lain yang relevan. Diagram ini ditunjukkan pada Gambar 6.



Gambar 6. Diagram Sequance Prediksi

Class diagram dari sistem analisis epidemiologi yang menggambarkan struktur beberapa kelas utama dalam sistem, masing-masing dengan tanggung jawab spesifik. Kelas Operation bertugas mengelola operasi dasar seperti inisialisasi komponen serta menangani login dan logout pengguna. Kelas RandomForest menangani pemuatan halaman utama, pengelolaan dataset, serta pelaksanaan analisis dengan algoritma *Random Forest*. Kelas Api bertugas membangun API dan mengambil data dari sumber eksternal. Sementara itu, kelas DB berperan penting dalam mengelola akses dan manipulasi data, termasuk operasi seperti pengambilan data, eksekusi query, serta insert, update, dan delete pada tabel database. Setiap kelas mendukung analisis epidemiologi, mulai dari autentikasi pengguna hingga pengelolaan dan analisis data yang efisien. Diagram class ini ditunjukkan pada Gambar 7.



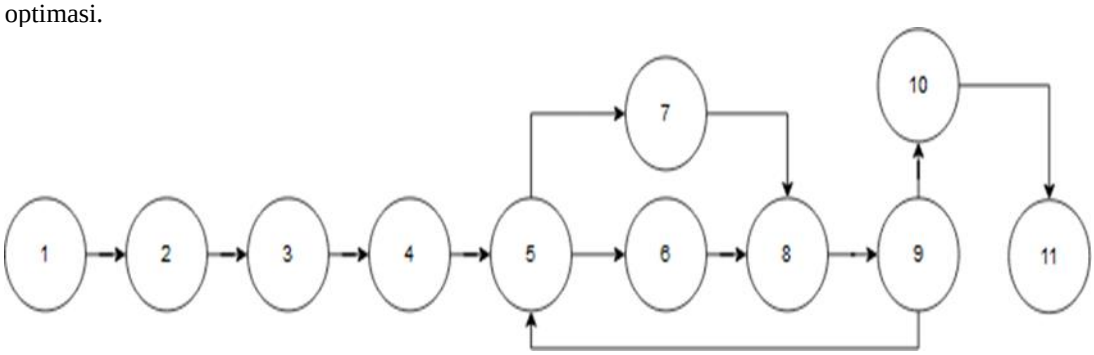
Gambar 7. Diagram Class

Berikut Gambar 8 merupakan coding yang akan diuji menggunakan metode white-box dari Prediksi

```
}else if($page=="predict"){
  ?><div class="row"><div class="col-4"><?php
  echo form_open("");
  ?>
  <div style="display:<?=$x==0||$x==(sizeof($column)-1)?'none':'block'?>" class="form-group">
    <label>Banyak Tree</label>
    <select class="form-control" name="ntree">
      <option value="3" <?=$request->getPost("ntree")==3?'selected':''?>>3 Tree</option>
      <option value="4" <?=$request->getPost("ntree")==4?'selected':''?>>4 Tree</option>
      <option value="5" <?=$request->getPost("ntree")==5?'selected':''?>>5 Tree</option>
      <option value="6" <?=$request->getPost("ntree")==6?'selected':''?>>6 Tree</option>
      <option value="7" <?=$request->getPost("ntree")==7?'selected':''?>>7 Tree</option>
      <option value="8" <?=$request->getPost("ntree")==8?'selected':''?>>8 Tree</option>
      <option value="9" <?=$request->getPost("ntree")==9?'selected':''?>>9 Tree</option>
      <option value="10" <?=$request->getPost("ntree")==10?'selected':''?>>10 Tree</option>
    </select>
  </div>
  <?php
  $x=0;
  $predictval = $request->getPost("predict");
  foreach ($column as $key) {
    ?>
    <div style="display:<?=$x==0||$x==(sizeof($column)-1)?'none':'block'?>" class="form-group">
      <label><?=$key?></label>
      <input type="number" class="form-control" min="0" step="0.01" name="predict[]" value="<?=$x==0?'1':@$predictval[$x]?>" />
    </div>
    <?php
    $x++;
  }
  ?>
  <div class="form-group mt-2">
    <button class="btn btn-primary" type="submit" name="prediksi" value="1">Prediksi</button>
  </div>
  <?php
  form_close();
  ?></div><?php
  ?><div class="col-6">
    <?php
      if($request->getPost("prediksi")==1){
        shuffle($data);
        $trees = random_forest_train($data, $request->getPost("ntree"), 2);
        $predictval = $request->getPost("predict");
        $result = random_forest_classify($trees,$predictval);
        ?>
        <div class="card text-white bg-success">
            <div class="card-body">
                <blockquote class="card-bodyquote mb-0">
                    <p><b>Hasil Prediksi : <?=$result?></b></p>
                </blockquote>
            </div>
        </div>
        <?php
      }
  ?></div>
  </div>
<?php
  if($request->getPost("prediksi")==1){
    ?><div class="row mt-2">
      <?php
      for ($x=0;$x<$request->getPost("ntree");$x++) {
        ?><div class="col-12">
            <div class="card text-white bg-warning p-1" style="overflow-y:auto;">
              <div class="card-body">
              Tree <?=$x+1?> <br />
              <ul class="tree">
                <li><b><?=$column[$trees[$x]->getFeatureId()]?> <= <?=$trees[$x]->getThreshold()?></b>
                  <ul>
                    <?php echo method_exists($trees[$x],'getLeft')?pohon($data,$column,$trees[$x]->getLeft()):$trees[$x]->species($data); ?>
                    <?php echo method_exists($trees[$x],'getRight')?pohon($data,$column,$trees[$x]->getRight()):$trees[$x]->species($data); ?>
                  </ul>
                </li>
              </ul>
            </div>
          </div>
        </div><?php
      }
    ?>
  </div><?php
}
?>
```

Gambar 8. Code Prediksi

Berdasarkan Gambar 9 dapat dibuat flowgraph Prediksi. Adapun Gambar 9 akan memperlihatkan flowgraph optimasi.



Gambar 9. Flowgraph Prediksi

Gambar 10 merupakan flowgraph fungsi optimasi. Adapun jalur pengujian sebagai berikut.
V(G) = E – N + 2
E = jumlah dari edge flowgraph N = jumlah dari node flowgraph
Sehingga kompleksitas dari flowgraph tampil produk adalah: V(G) = 12 – 11 + 2 = 3
Path 1 : 1-2-3-5-6-8-9-10-11
Path 2 : 1-2-4-5-7-8-9-5
Path 3 : 1-2-4-5-7-8-9-5-6-8-9-10-11

Tabel 1. Variabel Data Input dan Data Output

Variabel	Nama Variabel	Kriteria
V1	Jenis Kelamin	1 = Pria 2 = Perempuan
V2	Merokok	1 = Tidak 2 = Iya
V3	Garam_Berlebih	1 = Tidak

		2 = Iya
V4	Sistol	90.....240
V5	Diastol	60…....160
V6	Tinggi Badan	130…....200
V7	Berat Badan	27….....120
V8	Lingkar Perut	50.....104
V9	Kadar_Gula	83….....520
V10	Diagnosa	Diabetes Melitus Tipe 2 = 1 Diabetes Melitus Tipe 3 = 2 PPOK = 3 Hipertensi = 4 Obesitas = 5 Glaukoma = 6 Paru Paru Basah = 7 Presbicusis = 8 Asma Bronchiale = 9

Preprocessing data dilakukan pada tahap awal untuk mempersiapkan data mentah sebelum dilakukanya proses pemodelan. Preprocessing dilakukan dengan cara mengubah data menjadi bentuk yang lebih mudah atau mengeliminasi data yang tidak sesuai. Preprocessing data dilakukan dengan tujuan mendapatkan hasil yang lebih akurat, pengurangan waktu perhitungan dan membuat nilai data menjadi lebih kecil.

Tabel 2. Data Set Hasil Diinisialisasi

no	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	1	1	1	130	90	160	52	64	153	1
2	2	2	1	150	90	155	65	96	200	2
3	2	1	1	110	80	148	42	72	169	1
4	2	1	1	130	80	155	53	88	138	3
…	…	…	…	…	…	…	…	…	…	…
250	1	2	1	160	100	152	49	77	177	1

Dari dataset yang sudah diinisialisasi, bagi data menjadi 2 yaitu untuk data traning dan data testing, untuk contoh pada kali ini data dibagi menjadi 70:30, 7 data pertama untuk data traning dan 3 data terakhir untuk data testing. data training mendapatkan 4 data untuk diagnosis diabetes melitus tipe 2 yang diberi label 0, 1 data untuk diagnosis diabetes melitus tipe 3 yang diberi label 1, 1 data untuk diagnosis PPOK yang diberi label 2, dan 1 data untuk diagnosis obesitas yang diberi label 3, jadi total data tersebut ada 7 sesuai dengan pembagian diawal. Hitung nilai entropi dari seluruh data training untuk menentukan seberapa tidak teraturnya data tersebut.

H(S) = -(4/7log2¬¬(4/7)+1/7log2(1/7)+1/7log2(1/7)+1/7log2(1/7))

Hitung per komponen

4/7 ≈ 0.571, log2(0.571) ≈ -0,485, jadi – 0,571 x – 0,485 ≈ 0,277

1/7 ≈ 0,143, log2(0,143) ≈ - 2,807, jadi – 0,143 x -2,807 ≈ 0,401

Total entropi awal

H(S)=0.277+0.401+0.401+0.401≈1.48

Setelah itu menghitung informasi gain, untuk kali ini kita coba dengan variabel sistol, bagi data berdasarkan variabel sistol, misalnya kita bagi pada nilai variabel sistol > 130 maka akan mendapatkan data 2,5,6 untuk kategori 1 dengan sistol > 130, dan data 1,3,4,7 untuk kategori 2 dengan sistol <= 130. Hitung entropi pada setiap kategori, kategori 1 data 2,5,6.

H(S1) = - (2/3log2(2/3)+1/3log2(1/3))

H(S1) = - (0,667 x - 0,585 + 0,333 x -1,585) ≈ 0,918

Kategori 2 data 1,3,4,7.

H(S2) = - (2/4 log2(2/4) + 1/4log2(1/4) + ¬1/4 log2(1/4))

H(S2) = - (0,5 x -1 + 0,25 + -2 + 0,25 x -2) ≈ 1,5

Informasi gain

IG(S, Sistol) = H (S) – (S1/S x H(S1) + S2/S x H(S2))

IG(S, Sistol) = 1,48 – (3/7 x 0,918 + 4/7 x 1,5))

IG(S, Sistol) = 1,48 – (0,393 + 0,857) ≈ 1,48 – 1,25 = 0,23

Membentuk pohon keputusan dengan variabel sistol
Jika sistol > 130, maka diagnosis adalah “0” atau “1”.
Jika sistol <= 130, maka diagnosis adalah “0”, “2”, atau “3”
Membentuk *Random Forest*, Langkah dalam membentuk *Random Forest* ialah dengan membuat pohon keputusan menggunakan variabel yang lain dengan proses yang serupa. Setelah semua proses selesai dapat dilakukan testing dengan data testing, sebagai contoh data testing ke 1 (jenis_kelamin = 0, merokok = 0, garam_berlebih = 0, sistol = 100, diastol = 80, tinggi¬_badan = 144, berat_badan = 41, lingkar_perut = 66, kadar_gula = 110). Dapat dihasilkan prediksi sebagai berikut, pada pohon keputusan 1 dengan menggunakan variabel sistol mendapat hasil diagnosis penyakit PPOK, pada pohon keputsan 2 dengan menggunakan variabel kadar_gula mendapat hasil diagnosis diabetes melitus tipe 2, pada pohon keputusan 3 dengan menggunakan variabel berat_badan mendapat hasil diagnosis PPOK, maka dari hasil voting hasil diagnosis dari data testing 1 adalah PPOK. Akurasi dari hasil dapat dihitung dengan rumus sebagai berikut Akurasi=jumlah prediksi benar / jumlah total data uji x 100%.

Tabel 3. Confusion Matrix *Random Forest*

Kelas Aktual	DMT2	OS	HT	PPOK	GK	PPB	PBC	DMT3	AB
DMT2	13	5	18	0	0	0	0	0	0
OS	17	4	14	1	0	0	0	0	0
HT	9	5	17	1	0	0	0	1	0
PPOK	4	0	5	0	0	0	0	0	0
GK	1	0	0	0	0	0	0	0	0
PPB	1	0	1	0	0	0	0	0	0
PBC	0	0	1	0	0	0	0	0	0
DMT3	1	0	0	0	0	0	0	0	0
AB	0	0	0	0	0	0	0	0	0

Hasil pengujian yang dilakukan otomatis oleh sistem mendapatkan nilai sebagai berikut yang akan ditampilkan pada tabel 4

Tabel 4. Akurasi label

Label	Accuracy	Precision	Recall
DIABETES MILITUS TIPE 2	0.39837398373984	0.36111111111111	0.27659574468085
OBESITAS	0.33333333333333	0.10810810810811	0.26666666666667
HIPERTENSI	0.40650406504065	0.51515151515152	0.29824561403509
PPOK	0.073170731707317	0	0
GLAUKOMA	0.008130081300813	0	0
PARU PARU BASAH	0.016260162601626	0	0
PRESBICUSIS	0.008130081300813	0	0
DIABETES MILITUS TIPE 3	0.008130081300813	0	0
ASMA BRONCHIALE	0	0	0

Total akurasi yang didapat pada klasifikasi penyakit tidak menular pada penelitian kali ini adalah hanya 27.64%. ini dikarenakan dataset yang dimiliki masih kurang banyak dan kurang bagus kualitasnya.

5. Kesimpulan

Dalam penelitian kali ini algoritma *random forest* dipastikan dapat melakukan prediksi penyakit tidak menular, namun untuk tingkat akurasinya masih sangat rendah. Puskesmas sekura harus menambahkan lebih banyak data pada sistem prediksi ini agar tingkat akurasi menjadi bertambah. Hasil akurasi 27.64% tergolong cukup rendah untuk dapat digunakan dalam prediksi. Rendahnya hasil total akurasi dari sistem ini dipengaruhi oleh 2 faktor yaitu dataset yang masih sedikit dan memliki kuatlitas yang tidak terlalu baik, dan fakor kedua label pada penelitian kali ini lebih dari 2.

Terdapat beberapa saran untuk pengembangan lebih lanjut dalam sistem analisis epidemiologi menggunakan algoritma *random forest*, mengingat adanya beberapa kelemahan yang perlu diperbaiki. Pertama, disarankan untuk menambahkan form upload file pada halaman prediksi agar pengguna dapat melakukan prediksi secara massal tanpa harus melakukannya satu per satu. Kedua, disarankan untuk menambahkan halaman khusus untuk inisialisasi

dan encoding data, sehingga pengguna dapat dengan mudah melakukan penyesuaian apabila terdapat perubahan pada variabel prediktor dalam dataset.

6. Daftar Pustaka

[1] Zebua, R., Gulo, V. E., Purba, I., dan Gulo, M. J. K., 2023, Perubahan Epidemiologi Demam Berdarah Dengue (DBD) di Indonesia Tahun 2017–2021, *Jurnal Ilmiah Kesehatan Masyarakat*, vol. 2, no. 1, pp. 129-136.

[2] Rahayu, D., Irawan, H., Santoso, P., Susilowati, E., Atmojo, D. S., dan Kristanto, H., 2021, Deteksi Dini Penyakit Tidak Menular Pada Lansia, *Jurnal Peduli Masyarakat*, vol. 3, no. 1, pp. 91-96.

[3] Oktanisa, I., & Supianto, A. A. (2018). Perbandingan Teknik Klasifikasi Dalam Data Mining Untuk Bank Direct Marketing. *Jurnal Teknologi Informasi dan Ilmu Komputer*, *5*(5), 567-576.

[4] Suswanto, D. (2016). Analisis Perbandingan Metode Machine Learning pada Prediksi Khasiat Jamu.

[5] El Naqa, I., dan Murphy, M. J., 2015, What is Machine Learning?, *Springer*, New York.

[6] Annisa, R. (2019). Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung. *JTIK (Jurnal Teknik Informatika Kaputama)*, *3*(1), 22-28.

[7] Kesuma, M. (2023). Prediksi Penyakit Liver Menggunakan Algoritma *Random Forest*. *Jurnal Informasi dan Komputer*, *11*(02), 184-189.

[8] Aji, P. W. S., Suprianto, S., & Dijaya, R. (2023). Prediksi Penyakit Stroke Menggunakan Metode *Random Forest*. *Kesatria: Jurnal Penerapan Sistem Informasi (Komputer dan Manajemen)*, *4*(4), 916-924.

[9] Han, J., Kamber, M., dan Tong, H., 2023, Data Mining: Concepts and Techniques (Edisi keempat), Katey Birtcher, Cambridge.

[10] Priantama, Y., dan Siswa, T. A. Y., 2022, Optimasi Correlation-Based Feature Selection Untuk Perbaikan Akurasi *Random Forest* Classifier Dalam Prediksi Performa Akademik Mahasiswa, *Jurnal Informatika dan Komputer*, vol. 6, no. 2, pp. 35-50.

[11] Brocke, J., et al., 2020, Accumulation and Evolution of Design Knowledge in *Design Science Research* - A Journey Through Time and Space, *Journal of the Association for Information Systems*, vol. 21, no. 3, pp. 520- 544.

[12] Peffers, K., Tuunanen, T., Rothenberger, M. A., dan Chatterjee, S., 2007, A *Design Science Research* Methodology for Information Systems Research, *Journal of Management Information Systems*, vol. 24, no. 3, pp. 45-77.

[13] Anggraini, A, R., Oliver, J, 2019, *Journal of Chemical Information and Modeling*, Vol 53, No 9, pp 1689–1699.

[14] Pressman, Maxim. 2020. Software Engineering. New York: McGraw-Hill.

[15] Sundaramoorthy. 2022. UML Diagramming A Case Study Approach. New York: Auerbach Publications.

[16] Sugiyono, 2018, Metode Penelitian Kuantitatif, Alfabeta, Bandung