

Perbandingan Pengklasifikasian Penyakit Parkinson Menggunakan Algoritma Naive Bayes, Random Forest, dan Regresi Logistik

Michael Emmanuel Purba¹, Angga Zefanya Situmorang², Muhammad Wahyu Pratama Lubis³

^{1,2,3}Universitas Mikroskil, Jl. Thamrin No. 112, 124, 140, Telp. (061) 4573767, Fax. (061) 4567789

^{1,2,3}Fakultas Informatika, Teknik Informatika, Universitas Mikroskil, Medan

e-mail: ¹211111304@students.mikroskil.ac.id, ²211111287@students.mikroskil.ac.id,
³211110731@students.mikroskil.ac.id

Dikirim: 19-01-2025 | Diterima: 08-04-2025 | Diterbitkan: 30-04-2025

Abstrak

Penyakit Parkinson (PP) merupakan kondisi neurodegeneratif kronis dan progresif yang menjadi penyebab kedua tertinggi setelah Alzheimer. Penyakit ini disebabkan oleh kerusakan sel saraf pada substantia nigra dan akumulasi protein abnormal berupa Lewy bodies, dengan prevalensi yang meningkat pada kelompok usia lanjut. Penelitian ini bertujuan untuk membandingkan performa algoritma Naive Bayes, Random Forest, dan Regresi Logistik dalam mengklasifikasikan penyakit Parkinson menggunakan dataset dari Kaggle yang terdiri dari 2.105 data dengan 35 atribut. Tahapan preprocessing mencakup penghapusan atribut yang tidak relevan, pengecekan distribusi kelas, serta pembagian data latih dan uji secara stratifikasi. Hasil evaluasi menunjukkan bahwa algoritma Random Forest memiliki performa terbaik dengan skor ROC mencapai 0,97, menjadikannya model paling efektif untuk mendukung diagnosis dini penyakit Parkinson secara cepat, akurat, dan efisien.

Kata kunci: Penyakit Parkinson, algoritma Naive Bayes, Random Forest, Regresi Logistik, klasifikasi

Abstract

Parkinson's Disease (PD) is a chronic and progressive neurodegenerative condition, ranking as the second most common after Alzheimer's disease. It is caused by damage to nerve cells in the substantia nigra and the accumulation of abnormal protein aggregates known as Lewy bodies, with its prevalence increasing among the elderly population. This study aims to compare the performance of the Naive Bayes, Random Forest, and Linear Logistic algorithms in classifying Parkinson's disease using a dataset from Kaggle, consisting of 2,105 records with 35 attributes. The preprocessing steps included removing irrelevant attributes, checking class distribution, and stratified splitting of training and testing data. The evaluation results revealed that the Random Forest algorithm outperformed others with an impressive ROC score of 0.97, making it the most effective model for supporting early, accurate, and efficient diagnosis of Parkinson's disease.

Keywords: Parkinson's disease, Naive Bayes algorithm, Random Forest, Linear Regression, classification

1. PENDAHULUAN

1.1. Latar Belakang

Penyakit Parkinson (PP) merupakan salah satu penyakit neurodegeneratif yang paling sering terjadi, menempati peringkat kedua setelah Alzheimer [1]. Penyakit ini bersifat kronis, progresif, dan tidak dapat disembuhkan, sehingga memberikan dampak sosial yang signifikan. Pengobatan yang tersedia saat ini belum mampu menghentikan perkembangan penyakit ini dan sering kali disertai efek samping yang merugikan. Oleh karena itu, diperlukan alternatif terapi tambahan dengan risiko efek samping yang lebih rendah, seperti penggunaan vitamin [2]. Gejala Parkinson yang dapat dikenali meliputi rasa lemas atau kekakuan pada tubuh serta tremor atau getaran halus pada salah satu tangan. Selain itu, penyakit ini memengaruhi substantia nigra, area kecil di otak tengah yang berfungsi mengirimkan sinyal ke saraf tulang belakang untuk mengontrol otot tubuh [3].

Prevalensi penyakit Parkinson di negara-negara industri diperkirakan mencapai 0,3% dari total populasi, dengan sekitar 1% pada individu yang berusia di atas 60 tahun. Angka ini meningkat seiring bertambahnya usia, baik pada pria maupun wanita. Di Eropa, prevalensi Parkinson pada usia 85-89 tahun dilaporkan sebesar 3,5% [4]. Insiden penyakit ini berkisar antara 10 hingga 50 per 100.000 orang per tahun, dengan prevalensi mencapai 100 hingga 300 per 100.000 populasi. Frekuensi Parkinson meningkat tajam sesuai dengan bertambahnya usia [5]. Puncak insiden terjadi pada dekade keenam kehidupan, meskipun kasus dengan onset lebih awal dapat ditemukan pada dekade keempat [6]. Parkinson jarang terjadi sebelum usia 50 tahun, tetapi insidensi dan prevalensinya meningkat secara signifikan setelah usia 60 tahun. Berdasarkan studi meta-analisis, prevalensi Parkinson meningkat dari 107 per 100.000 pada usia 50-59 tahun menjadi 1.087 per 100.000 pada usia 70-79 tahun [5]. Karena kaitannya dengan usia, jumlah kasus Parkinson diproyeksikan meningkat sebesar 25-30% dalam 25 tahun mendatang. Prevalensi tertinggi penyakit ini tercatat pada ras Kaukasia di Amerika Utara dan Eropa (0,98% hingga 1,84%), sedangkan prevalensi lebih rendah ditemukan pada ras Asia (0,018%) dan yang terendah pada ras kulit hitam di Afrika (0,01%) [7].

Penelitian ini bertujuan untuk mengatasi tantangan dalam mendiagnosis penyakit Parkinson dengan mengevaluasi dan membandingkan kinerja tiga algoritma pembelajaran mesin, yaitu Naive Bayes, Random Forest, dan Regresi Logistik. Ketiga algoritma tersebut dipilih berdasarkan pendekatan unik yang ditawarkan, seperti analisis probabilistik, metode berbasis pohon keputusan, dan model prediksi linier. Penelitian ini diharapkan dapat memberikan panduan dalam menentukan algoritma paling efektif untuk mendukung diagnosis dini sekaligus meningkatkan akurasi deteksi penyakit Parkinson. Dalam penelitian sebelumnya, pendekatan Naive Bayes dan Random Forest dibandingkan. Hasil menunjukkan bahwa metode Naive Bayes memiliki akurasi 49,06%, Random Forest 74,28%, C4.5 57,53%, Bayesian Network 48,07%, dan Decision Stump 49,95%, dengan Random Forest memberikan performa lebih baik dibandingkan lainnya [3], [8]. Menurut penelitian yang dilakukan oleh Chandel, Kunwar, Sabitha, Choudhury, dan Mukherjee pada 2017 tentang perbandingan deteksi penyakit tiroid menggunakan klasifikasi K-Nearest Neighbor (KNN) dan Naive Bayes, algoritma KNN menunjukkan akurasi 93,44%, sedangkan Naive Bayes hanya 22,56% [9].

Penelitian ini juga bertujuan untuk mengidentifikasi algoritma yang paling efektif dalam meningkatkan akurasi diagnosis dini guna mendukung pengembangan metode diagnosis yang lebih cepat, efisien, dan akurat. Penelitian ini diharapkan memberikan kontribusi signifikan terhadap pengembangan teknologi cerdas dalam diagnosis penyakit neurodegeneratif, khususnya Parkinson, baik untuk penelitian maupun aplikasi klinis. Data yang digunakan dalam penelitian ini berasal dari situs Kaggle, dengan total 2.105 data yang digunakan untuk melatih dan mengevaluasi model. Dataset ini dipilih karena relevan dengan tujuan penelitian, yaitu meningkatkan akurasi diagnosis dini penyakit Parkinson. Data tersebut mencakup berbagai fitur yang mendukung analisis pola dan pengambilan keputusan yang lebih baik dalam identifikasi penyakit.

1.2. Rumusan Masalah

Penyakit Parkinson adalah salah satu penyakit neurodegeneratif yang bersifat progresif dan tidak dapat disembuhkan, dengan dampak sosial yang signifikan karena keterbatasan pengobatan yang ada saat ini. Diagnosis dini sangat penting untuk memperlambat perkembangan penyakit dan meningkatkan kualitas hidup pasien, namun metode konvensional sering kali memiliki keterbatasan dalam hal akurasi. Dalam konteks ini, penerapan pembelajaran mesin berpotensi menjadi solusi yang efektif dalam mendukung diagnosis dini penyakit Parkinson. Berdasarkan latar belakang tersebut, rumusan masalah yang diangkat dalam penelitian ini adalah:

1. Bagaimana performa algoritma pembelajaran mesin, yaitu Naive Bayes, Random Forest, dan Regresi Logistik, dalam mendeteksi penyakit Parkinson secara akurat?
2. Algoritma mana yang paling efektif dalam meningkatkan akurasi diagnosis dini penyakit Parkinson berdasarkan data yang tersedia?
3. Sejauh mana penerapan pembelajaran mesin dapat berkontribusi dalam pengembangan metode diagnosis yang lebih cepat dan akurat?

1.3. Tujuan

Penelitian ini bertujuan untuk:

1. Menganalisis dan membandingkan performa tiga algoritma pembelajaran mesin, yaitu Naive Bayes, Random Forest, dan Regresi Logistik, dalam mendeteksi penyakit Parkinson.
2. Mengidentifikasi algoritma yang memberikan akurasi terbaik untuk deteksi dini penyakit Parkinson.
3. Mendukung pengembangan metode diagnosis berbasis teknologi cerdas yang lebih cepat, efisien, dan akurat untuk mendukung penelitian maupun aplikasi klinis.

1.4. Ruang Lingkup

Adapun batasan masalah yang ditetapkan dari penelitian ini adalah sebagai berikut:

1. Penelitian menggunakan dataset berjumlah 2.105 data yang diunduh dari situs Kaggle, dengan fitur-fitur relevan yang mendukung diagnosis penyakit Parkinson.
2. Fokus penelitian terletak pada penerapan dan evaluasi tiga algoritma pembelajaran mesin, yaitu Naive Bayes, Random Forest, dan Regresi Logistik.
3. Pengukuran kinerja algoritma didasarkan pada metrik akurasi dalam mendeteksi penyakit Parkinson, serta dilakukan perbandingan dengan hasil penelitian sebelumnya.
4. Penelitian ini tidak mencakup pengembangan algoritma baru, melainkan memanfaatkan algoritma yang sudah ada untuk mengevaluasi efektivitasnya dalam diagnosis penyakit Parkinson.

2. TINJAUAN PUSTAKA

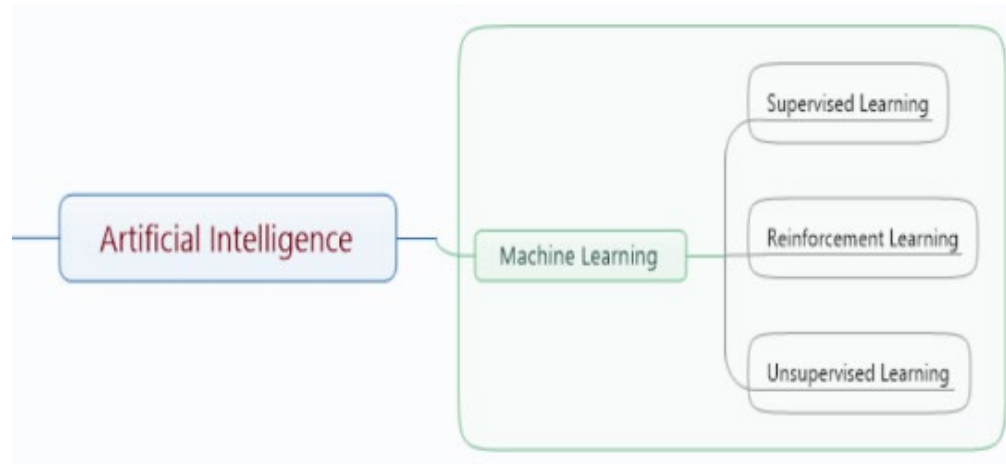
2.1. Penyakit Parkinson

Penyakit Parkinson adalah gangguan neurodegeneratif yang menduduki peringkat kedua setelah Alzheimer dalam prevalensinya. Penyakit ini disebabkan oleh penurunan kadar dopamin di otak, yang berperan dalam mengontrol gerakan tubuh. Penurunan dopamin tersebut terjadi akibat kerusakan sel saraf di Substantia Nigra Pars Compacta (SNc) pada batang otak, serta penumpukan protein abnormal berupa Lewy bodies yang mengandung α -synuclein. Meskipun sudah banyak penelitian yang dilakukan, penyebab pasti dari penyakit Parkinson masih belum ditemukan. Penyakit ini termasuk salah satu gangguan sistem saraf dengan insidensi tertinggi setelah Alzheimer [10].

2.2. Pembelajaran Mesin

Teknologi machine learning (ML) adalah sistem yang dirancang untuk belajar secara otomatis tanpa memerlukan instruksi langsung dari pengguna. Machine learning dibangun dengan menggabungkan berbagai disiplin ilmu, seperti statistika, matematika, dan data mining, yang memungkinkan mesin untuk menganalisis data secara otomatis tanpa perlu pemrograman ulang atau perintah tambahan [11].

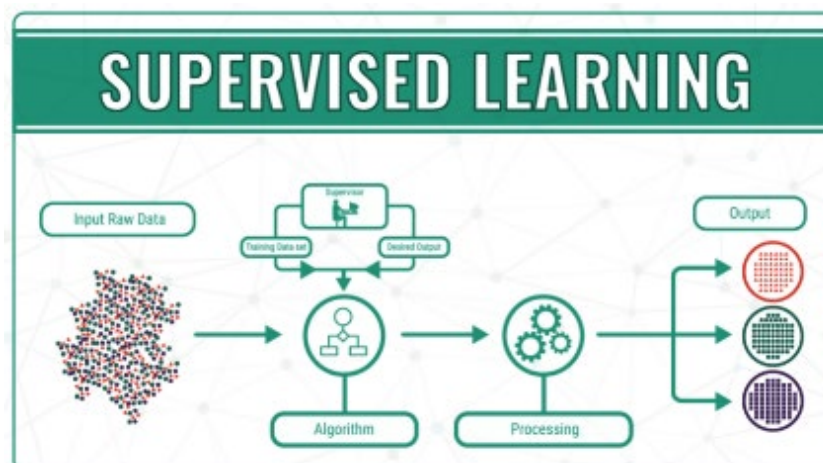
Penelitian terkini mengidentifikasi tiga kategori utama dalam machine learning, yaitu Supervised Learning, Unsupervised Learning, dan Reinforcement Learning [12]. Skema keterkaitan antara artificial intelligence dan machine learning dapat digambarkan dalam Gambar 1.



Gambar 1. Skema keterkaitan artificial intelligence dan machine learning

2.3. Pembelajaran Supervised

Pada algoritma Supervised Learning, sistem diberikan dataset yang mencakup informasi masukan dan keluaran yang diinginkan. Dengan demikian, sistem dapat mempelajari pola dari data yang sudah ada. Setelah mempelajari pola tersebut, sistem akan menggunakannya sebagai referensi untuk menganalisis kumpulan data berikutnya. Gambar 2 menunjukkan diagram blok yang menggambarkan cara kerja dari Supervised Learning [13].



Gambar 2. Skema cara kerja supervised learning

2.4. Naïve Bayes

Algoritma Naïve Bayes adalah teknik klasifikasi yang menggunakan pendekatan probabilitas dan statistik, yang pertama kali diajukan oleh ilmuwan Inggris, Thomas Bayes. Prinsip dasar dari algoritma ini adalah memperkirakan kemungkinan di masa depan berdasarkan pengalaman masa lalu, yang dikenal dengan Teorema Bayes [14]. Dalam algoritma ini, terdapat asumsi bahwa setiap atribut saling independen, yang dikenal dengan istilah "Naive." Artinya, dalam penerapannya, algoritma Naïve Bayes menganggap bahwa tidak ada hubungan antara satu atribut dengan atribut lainnya, meskipun sebenarnya atribut tersebut mungkin memiliki keterkaitan.

Pada tahap pelatihan, algoritma klasifikasi akan membangun model dengan menganalisis data pelatihan. Proses ini dapat dianggap sebagai pembentukan fungsi atau pemetaan dari $Y=F(X)$, di mana Y adalah kelas hasil prediksi dan X adalah tuple yang ingin diprediksi kelasnya [2]:

1. Menghitung nilai peluang kondisi baru (X_k) dari setiap hipotesa terhadap kelas (C_i) yang ada.
2. Menghitung nilai akumulasi peluang dari setiap kelas (C_i)
3. Menghitung Nilai $P(X|C_i) \times P(C_i)$
4. Menentukan kelas dari kasus baru tersebut.

Rumus Naive Bayes yang digunakan dalam proses analisis klasifikasi adalah sebagai berikut:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)} \quad (1)$$

1. $P(C|X)$: Probabilitas posterior yang menunjukkan kemungkinan bahwa kelas C adalah benar, diberikan informasi fitur X .
2. $P(X|C)$: Probabilitas likelihood, yaitu kemungkinan data X muncul jika kelas C sudah diketahui.
3. $P(C)$: Probabilitas prior, yaitu peluang awal dari kelas C berdasarkan distribusi kelas dalam dataset.
4. $P(X)$: Probabilitas dari data X , sering dianggap konstan dalam perhitungan karena nilainya sama untuk semua kelas.

2.5. Random Forest

Random Forest adalah pengembangan dari metode Decision Tree yang menggabungkan beberapa Decision Tree, di mana setiap pohon keputusan dilatih menggunakan sampel individu, dan setiap atribut dipecah pada pohon yang dipilih secara acak dari subset atribut. Random Forest memiliki beberapa keunggulan, seperti peningkatan akurasi ketika ada data yang hilang, ketahanan terhadap outlier, dan efisiensi dalam penyimpanan data. Selain itu, Random Forest juga memiliki proses seleksi fitur yang memungkinkan untuk memilih fitur terbaik, sehingga meningkatkan performa model klasifikasi. Dengan adanya seleksi fitur, Random Forest dapat bekerja secara efektif pada big data dengan parameter yang kompleks [15].

Metode ini terdiri dari root node, internal node, dan leaf node. Root node adalah simpul yang terletak paling atas, sering disebut sebagai akar dari pohon keputusan. Internal node adalah simpul percabangan yang memiliki minimal dua output dan hanya satu input. Sedangkan leaf node, atau terminal node, adalah simpul terakhir yang hanya memiliki satu input dan tidak memiliki output. Pohon keputusan dimulai dengan menghitung nilai entropy sebagai indikator ketidakmurnian atribut, dan nilai information gain digunakan untuk menentukan pembagian terbaik. Untuk menghitung nilai entropy digunakan rumus seperti pada persamaan 2, sedangkan nilai information gain dihitung menggunakan persamaan 3 [16].

Berikut ini merupakan rumus dari Random Forest [17] :

$$Entropy(Y) = - \sum_i p(c|Y) \log^2 p(c|Y), \quad (2)$$

Keterangan :

Y : Himpunan kasus

$P(c|Y)$: Proporsi nilai Y terhadap kelas c .

$$Information\ Gain(Y, a) = Entropy(Y) - \sum_{v \in Values(a)} \frac{|Y_v|}{|Y_a|} Entropy(Y_v) \quad (3)$$

Keterangan :

$Values(a)$ = Nilai yang mungkin dalam himpunan a .

Y_v = Subkelas dari Y dengan kelas v yang berhubungan dengan kelas a .

Y_a = Semua nilai yang sesuai dengan a .

2.6. Regresi Liner

Dengan menggunakan regresi logistik, kita dapat menentukan bagaimana variabel respons dikotomis—yaitu variabel dengan dua kategori dalam skala nominal atau ordinal—terhubung dengan satu atau lebih prediktor yang berskala kategori atau kontinu [18].

Metode ini menghubungkan output biner dengan variabel independen berdasarkan probabilitas, untuk memprediksi nilai dari variabel dependen, yang kemudian akan mengklasifikasikan data ke dalam dua kategori yang berbeda [19]. Regresi logistik adalah salah satu metode klasifikasi yang umum digunakan. Regresi logistik biner digunakan ketika variabel dependen berupa variabel dikotomis. Menurut Hosmer, D.W., dan Lemeshow, S. (2000) dalam Purwa (2019) [20], secara umum, model regresi logistik adalah [21]:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n}} \quad (4)$$

Dimana $\pi(x)$ adalah nilai probabilitas yang berada dalam rentang $0 \leq \pi(x) \leq 1$, yang berarti regresi logistik menggambarkan suatu probabilitas. Dengan mentransformasikan $\pi(x)$ pada persamaan di atas menggunakan transformasi logit $g(x)$, di mana:

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) \quad (5)$$

Maka diperoleh bentuk logit :

$$l = \log p \left(\frac{p_y}{1 - p_y} \right) \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n \quad (6)$$

Keterangan :

l = log-odds, p adalah basis dari algoritma

β_n = adalah paramater model

p_y = adalah probabilitas dari kejadian tersebut

2.7. Evaluasi Model

Parameter recall, f1-measure, presisi, dan akurasi adalah beberapa metrik untuk mengukur kinerja algoritma, yang dapat dihitung berdasarkan confusion matrix yang ditunjukkan pada Tabel 1 [22].

Tabel 1. Tabel Confusion Matrix

	Label atau kelas	
	Positif	Negatif
Positif	True Positive	False Positive
Negatif	False Negative	True Negative

Pada tahap ini, pengujian dilakukan dengan menghitung nilai Akurasi, Sensitivitas, dan Presisi [23].

$$akurasi = \frac{TN + TP}{TN + TP + FN + FP} \quad (7)$$

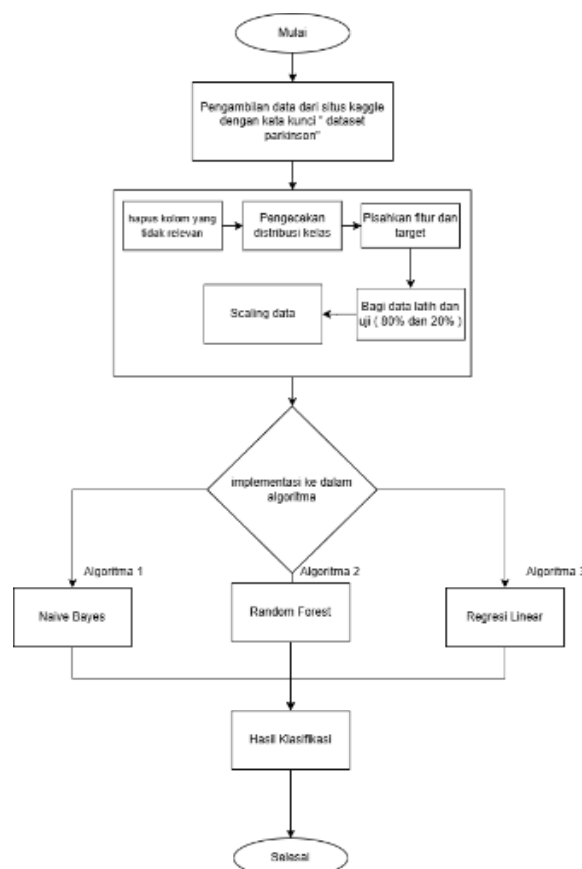
Selain akurasi, performa klasifikasi juga dapat dinilai berdasarkan nilai sensitivitas dan spesifisitas. Sensitivitas mengukur akurasi pada kelas positif, sementara spesifisitas mengukur akurasi pada kelas negatif. Rumus untuk menghitung sensitivitas dan spesifisitas adalah sebagai berikut.

$$Sensitivity = \frac{TP}{(TP + FP)} \times 100\% \quad (8)$$

$$akurasi = \frac{TN + TP}{TN + TP + FN + FP} \quad (9)$$

3. METODE PENELITIAN

Pemecahan masalah dalam penelitian ini dimulai dengan pengumpulan data yang relevan, diikuti oleh tahap preprocessing untuk memastikan data siap digunakan oleh model. Setelah itu, data yang telah diproses diuji dengan tiga algoritma pembelajaran mesin, yaitu Naive Bayes, Random Forest, dan Regresi Logistik. Hasil dari setiap model kemudian dianalisis untuk menentukan algoritma yang paling efektif dalam mengklasifikasikan penyakit Parkinson, berdasarkan akurasi dan kinerja keseluruhannya.



Gambar 3. Alur Penelitian

3.1. Pengambilan Data

Dataset yang digunakan dalam penelitian ini diperoleh dari situs Kaggle, dengan total sebanyak 2.106 data yang terdiri dari 35 atribut. Dataset ini dipilih karena kelengkapan dan relevansinya dalam mendukung analisis serta pengklasifikasian penyakit Parkinson.

id	sex	age	lwt	lwt2	lwt3	lwt4	lwt5	lwt6	lwt7	lwt8	lwt9	lwt10	lwt11	lwt12	lwt13	lwt14	lwt15	lwt16	lwt17	lwt18	lwt19	lwt20	lwt21	lwt22	lwt23	lwt24	lwt25	lwt26	lwt27	lwt28	lwt29	lwt30	lwt31	lwt32	lwt33	lwt34	lwt35	lwt36	lwt37	lwt38	lwt39	lwt40	lwt41	lwt42	lwt43	lwt44	lwt45	lwt46	lwt47	lwt48	lwt49	lwt50	lwt51	lwt52	lwt53	lwt54	lwt55	lwt56	lwt57	lwt58	lwt59	lwt60	lwt61	lwt62	lwt63	lwt64	lwt65	lwt66	lwt67	lwt68	lwt69	lwt70	lwt71	lwt72	lwt73	lwt74	lwt75	lwt76	lwt77	lwt78	lwt79	lwt80	lwt81	lwt82	lwt83	lwt84	lwt85	lwt86	lwt87	lwt88	lwt89	lwt90	lwt91	lwt92	lwt93	lwt94	lwt95	lwt96	lwt97	lwt98	lwt99	lwt100	lwt101	lwt102	lwt103	lwt104	lwt105	lwt106	lwt107	lwt108	lwt109	lwt110	lwt111	lwt112	lwt113	lwt114	lwt115	lwt116	lwt117	lwt118	lwt119	lwt120	lwt121	lwt122	lwt123	lwt124	lwt125	lwt126	lwt127	lwt128	lwt129	lwt130	lwt131	lwt132	lwt133	lwt134	lwt135	lwt136	lwt137	lwt138	lwt139	lwt140	lwt141	lwt142	lwt143	lwt144	lwt145	lwt146	lwt147	lwt148	lwt149	lwt150	lwt151	lwt152	lwt153	lwt154	lwt155	lwt156	lwt157	lwt158	lwt159	lwt160	lwt161	lwt162	lwt163	lwt164	lwt165	lwt166	lwt167	lwt168	lwt169	lwt170	lwt171	lwt172	lwt173	lwt174	lwt175	lwt176	lwt177	lwt178	lwt179	lwt180	lwt181	lwt182	lwt183	lwt184	lwt185	lwt186	lwt187	lwt188	lwt189	lwt190	lwt191	lwt192	lwt193	lwt194	lwt195	lwt196	lwt197	lwt198	lwt199	lwt200	lwt201	lwt202	lwt203	lwt204	lwt205	lwt206	lwt207	lwt208	lwt209	lwt210	lwt211	lwt212	lwt213	lwt214	lwt215	lwt216	lwt217	lwt218	lwt219	lwt220	lwt221	lwt222	lwt223	lwt224	lwt225	lwt226	lwt227	lwt228	lwt229	lwt230	lwt231	lwt232	lwt233	lwt234	lwt235	lwt236	lwt237	lwt238	lwt239	lwt240	lwt241	lwt242	lwt243	lwt244	lwt245	lwt246	lwt247	lwt248	lwt249	lwt250	lwt251	lwt252	lwt253	lwt254	lwt255	lwt256	lwt257	lwt258	lwt259	lwt260	lwt261	lwt262	lwt263	lwt264	lwt265	lwt266	lwt267	lwt268	lwt269	lwt270	lwt271	lwt272	lwt273	lwt274	lwt275	lwt276	lwt277	lwt278	lwt279	lwt280	lwt281	lwt282	lwt283	lwt284	lwt285	lwt286	lwt287	lwt288	lwt289	lwt290	lwt291	lwt292	lwt293	lwt294	lwt295	lwt296	lwt297	lwt298	lwt299	lwt300	lwt301	lwt302	lwt303	lwt304	lwt305	lwt306	lwt307	lwt308	lwt309	lwt310	lwt311	lwt312	lwt313	lwt314	lwt315	lwt316	lwt317	lwt318	lwt319	lwt320	lwt321	lwt322	lwt323	lwt324	lwt325	lwt326	lwt327	lwt328	lwt329	lwt330	lwt331	lwt332	lwt333	lwt334	lwt335	lwt336	lwt337	lwt338	lwt339	lwt340	lwt341	lwt342	lwt343	lwt344	lwt345	lwt346	lwt347	lwt348	lwt349	lwt350	lwt351	lwt352	lwt353	lwt354	lwt355	lwt356	lwt357	lwt358	lwt359	lwt360	lwt361	lwt362	lwt363	lwt364	lwt365	lwt366	lwt367	lwt368	lwt369	lwt370	lwt371	lwt372	lwt373	lwt374	lwt375	lwt376	lwt377	lwt378	lwt379	lwt380	lwt381	lwt382	lwt383	lwt384	lwt385	lwt386	lwt387	lwt388	lwt389	lwt390	lwt391	lwt392	lwt393	lwt394	lwt395	lwt396	lwt397	lwt398	lwt399	lwt400	lwt401	lwt402	lwt403	lwt404	lwt405	lwt406	lwt407	lwt408	lwt409	lwt410	lwt411	lwt412	lwt413	lwt414	lwt415	lwt416	lwt417	lwt418	lwt419	lwt420	lwt421	lwt422	lwt423	lwt424	lwt425	lwt426	lwt427	lwt428	lwt429	lwt430	lwt431	lwt432	lwt433	lwt434	lwt435	lwt436	lwt437	lwt438	lwt439	lwt440	lwt441	lwt442	lwt443	lwt444	lwt445	lwt446	lwt447	lwt448	lwt449	lwt450	lwt451	lwt452	lwt453	lwt454	lwt455	lwt456	lwt457	lwt458	lwt459	lwt460	lwt461	lwt462	lwt463	lwt464	lwt465	lwt466	lwt467	lwt468	lwt469	lwt470	lwt471	lwt472	lwt473	lwt474	lwt475	lwt476	lwt477	lwt478	lwt479	lwt480	lwt481	lwt482	lwt483	lwt484	lwt485	lwt486	lwt487	lwt488	lwt489	lwt490	lwt491	lwt492	lwt493	lwt494	lwt495	lwt496	lwt497	lwt498	lwt499	lwt500	lwt501	lwt502	lwt503	lwt504	lwt505	lwt506	lwt507	lwt508	lwt509	lwt510	lwt511	lwt512	lwt513	lwt514	lwt515	lwt516	lwt517	lwt518	lwt519	lwt520	lwt521	lwt522	lwt523	lwt524	lwt525	lwt526	lwt527	lwt528	lwt529	lwt530	lwt531	lwt532	lwt533	lwt534	lwt535	lwt536	lwt537	lwt538	lwt539	lwt540	lwt541	lwt542	lwt543	lwt544	lwt545	lwt546	lwt547	lwt548	lwt549	lwt550	lwt551	lwt552	lwt553	lwt554	lwt555	lwt556	lwt557	lwt558	lwt559	lwt560	lwt561	lwt562	lwt563	lwt564	lwt565	lwt566	lwt567	lwt568	lwt569	lwt570	lwt571	lwt572	lwt573	lwt574	lwt575	lwt576	lwt577	lwt578	lwt579	lwt580	lwt581	lwt582	lwt583	lwt584	lwt585	lwt586	lwt587	lwt588	lwt589	lwt590	lwt591	lwt592	lwt593	lwt594	lwt595	lwt596	lwt597	lwt598	lwt599	lwt600	lwt601	lwt602	lwt603	lwt604	lwt605	lwt606	lwt607	lwt608	lwt609	lwt610	lwt611	lwt612	lwt613	lwt614	lwt615	lwt616	lwt617	lwt618	lwt619	lwt620	lwt621	lwt622	lwt623	lwt624	lwt625	lwt626	lwt627	lwt628	lwt629	lwt630	lwt631	lwt632	lwt633	lwt634	lwt635	lwt636	lwt637	lwt638	lwt639	lwt640	lwt641	lwt642	lwt643	lwt644	lwt645	lwt646	lwt647	lwt648	lwt649	lwt650	lwt651	lwt652	lwt653	lwt654	lwt655	lwt656	lwt657	lwt658	lwt659	lwt660	lwt661	lwt662	lwt663	lwt664	lwt665	lwt666	lwt667	lwt668	lwt669	lwt670	lwt671	lwt672	lwt673	lwt674	lwt675	lwt676	lwt677	lwt678	lwt679	lwt680	lwt681	lwt682	lwt683	lwt684	lwt685	lwt686	lwt687	lwt688	lwt689	lwt690	lwt691	lwt692	lwt693	lwt694	lwt695	lwt696	lwt697	lwt698	lwt699	lwt700	lwt701	lwt702	lwt703	lwt704	lwt705	lwt706	lwt707	lwt708	lwt709	lwt710	lwt711	lwt712	lwt713	lwt714	lwt715	lwt716	lwt717	lwt718	lwt719	lwt720	lwt721	lwt722	lwt723	lwt724	lwt725	lwt726	lwt727	lwt728	lwt729	lwt730	lwt731	lwt732	lwt733	lwt734	lwt735	lwt736	lwt737	lwt738	lwt739	lwt740	lwt741	lwt742	lwt743	lwt744	lwt745	lwt746	lwt747	lwt748	lwt749	lwt750	lwt751	lwt752	lwt753	lwt754	lwt755	lwt756	lwt757	lwt758	lwt759	lwt760	lwt761	lwt762	lwt763	lwt764	lwt765	lwt766	lwt767	lwt768	lwt769	lwt770	lwt771	lwt772	lwt773	lwt774	lwt775	lwt776	lwt777	lwt778	lwt779	lwt780	lwt781	lwt782	lwt783	lwt784	lwt785	lwt786	lwt787	lwt788	lwt789	lwt790	lwt791	lwt792	lwt793	lwt794	lwt795	lwt796	lwt797	lwt798	lwt799	lwt800	lwt801	lwt802	lwt803	lwt804	lwt805	lwt806	lwt807	lwt808	lwt809	lwt810	lwt811	lwt812	lwt813	lwt814	lwt815	lwt816	lwt817	lwt818	lwt819	lwt820	lwt821	lwt822	lwt823	lwt824	lwt825	lwt826	lwt827	lwt828	lwt829	lwt830	lwt831	lwt832	lwt833	lwt834	lwt835	lwt836	lwt837	lwt838	lwt839	lwt840	lwt841	lwt842	lwt843	lwt844	lwt845	lwt846	lwt847	lwt848	lwt849	lwt850	lwt851	lwt852	lwt853	lwt854	lwt855	lwt856	lwt857	lwt858	lwt859	lwt860	lwt861	lwt862	lwt863	lwt864	lwt865	lwt866	lwt867	lwt868	lwt869	lwt870	lwt871	lwt872	lwt873	lwt874	lwt875	lwt876	lwt877	lwt878	lwt879	lwt880	lwt881	lwt882	lwt883	lwt884	lwt885	lwt886	lwt887	lwt888	lwt889	lwt890	lwt891	lwt892	lwt893	lwt894	lwt895	lwt896	lwt897	lwt898	lwt899	lwt900	lwt901	lwt902	lwt903	lwt904	lwt905	lwt906	lwt907	lwt908	lwt909	lwt910	lwt911	lwt912	lwt913	lwt914	lwt915	lwt916	lwt917	lwt918	lwt919	lwt920	lwt921	lwt922	lwt923	lwt924	lwt925	lwt926	lwt927	lwt928	lwt929	lwt930	lwt931	lwt932	lwt933	lwt934	lwt935	lwt936	lwt937	lwt938	lwt939	lwt940	lwt941	lwt942	lwt943	lwt944	lwt945	lwt946	lwt947	lwt948	lwt949	lwt950	lwt951	lwt952	lwt953	lwt954	lwt955	lwt956	lwt957	lwt958	lwt959	lwt960	lwt961	lwt962	lwt963	lwt964	lwt965	lwt966	lwt967	lwt968	lwt969	lwt970	lwt971	lwt972	lwt973	lwt974	lwt975	lwt976	lwt977	lwt978	lwt979	lwt980	lwt981	lwt982	lwt983	lwt984	lwt985	lwt986	lwt987	lwt988	lwt989	lwt990	lwt991	lwt992	lwt993	lwt994	lwt995	lwt996	lwt997	lwt998	lwt999	lwt1000
----	-----	-----	-----	------	------	------	------	------	------	------	------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	-------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	--------	---------

Gambar 4. Data yang didapatkan dari situs kaggle

3.2. Preprocessing

Pada tahap ini, data yang telah dikumpulkan akan melalui proses preprocessing untuk memastikan kualitas dan kelayakannya dalam melatih model secara optimal. Proses preprocessing yang dilakukan meliputi:

1. Hapus kolom yang tidak relevan
2. Pengecekan distribusi kelas
3. Pisahkan fitur dan target
4. Split data (80% data training dan 20% data testing)
5. Scaling data

3.3. Pelatihan Model

Pelatihan model merupakan bagian penting dalam metode penelitian yang bertujuan untuk menguji seluruh data menggunakan tiga algoritma yang telah ditentukan. Proses ini dilakukan untuk mengevaluasi kinerja setiap algoritma dan menentukan nilai atau persentase akurasi terbaik dalam klasifikasi penyakit Parkinson. Model yang akan diuji adalah :

1. Naïve bayes
2. Random forest
3. Regresi Logistik

3.4. Hasil Klasifikasi

Hasil klasifikasi merupakan output akhir dari proses pelatihan model terhadap data yang telah diproses. Pada tahap ini, akan terlihat model mana yang memiliki performa terbaik dibandingkan model lainnya dalam mengklasifikasikan penyakit Parkinson. Evaluasi ini dilakukan berdasarkan metrik tertentu, seperti akurasi, presisi, *recall*, dan F1-score, sehingga dapat memberikan gambaran yang jelas mengenai efektivitas masing-masing model dalam mendukung diagnosis penyakit Parkinson.

4. HASIL DAN PEMBAHASAN

4.1. Preprocessing

4.1.1. Hapus kolom yang tidak relevan

Pada tahap awal, dilakukan penghapusan kolom yang tidak relevan terhadap analisis, seperti *PatientID* dan *DoctorInCharge*. Kolom-kolom ini dianggap tidak berkontribusi pada proses klasifikasi karena *PatientID* hanya berfungsi sebagai identitas unik pasien, sedangkan *DoctorInCharge* merupakan atribut administratif yang tidak mempengaruhi diagnosis. Penghapusan kolom-kolom ini bertujuan

untuk mengurangi kompleksitas data dan memastikan bahwa model hanya menggunakan fitur yang relevan.

4.1.2. Pengecekan distribusi kelas

Setelah data dipersiapkan, langkah berikutnya adalah memeriksa distribusi kelas pada target variabel (*Diagnosis*). Pengecekan ini penting untuk memastikan bahwa data terdistribusi dengan seimbang antara kelas-kelas yang ada. Distribusi kelas yang seimbang membantu menghindari terjadinya bias pada model, di mana model lebih cenderung memprediksi kelas yang memiliki jumlah data lebih banyak. Jika distribusi kelas tidak seimbang, teknik seperti oversampling atau undersampling dapat dipertimbangkan.

4.1.3. Pisahkan fitur dan target

Proses berikutnya adalah pemisahan antara fitur (X) dan target (y). Fitur (X) adalah variabel independen yang digunakan untuk memprediksi target, sementara target (y) adalah variabel dependen yang ingin diprediksi oleh model, yaitu *Diagnosis*. Pemisahan ini penting untuk memastikan bahwa data siap untuk pelatihan model, sehingga fitur dapat diproses secara terpisah dari label yang ingin diprediksi.

4.1.4. Pembagian Data untuk Pelatihan dan Pengujian

Data kemudian dibagi menjadi dua bagian: 80% digunakan untuk pelatihan (*training*) dan 20% untuk pengujian (*testing*). Pembagian ini dilakukan dengan menggunakan stratifikasi untuk menjaga proporsi kelas yang sama antara data latih dan data uji. Stratifikasi memastikan bahwa setiap subset data (latih dan uji) memiliki distribusi kelas yang serupa dengan dataset asli, yang penting agar evaluasi model mencerminkan performa yang sebenarnya. Pembagian ini memungkinkan model untuk dilatih dengan data yang cukup dan diuji dengan data yang belum pernah dilihat sebelumnya.

4.2. Pelatihan model

4.2.1. Naïve Bayes

Algoritma Naïve Bayes adalah algoritma yang memperkirakan kemungkinan suatu data uji termasuk ke dalam kategori tertentu berdasarkan probabilitas dari setiap fitur data uji tersebut [24].

```
Confusion Matrix:
[[121  39]
 [ 40 221]]
True Positive (TP): 221
True Negative (TN): 121
False Positive (FP): 39
False Negative (FN): 40
Classification Report:

```

	precision	recall	f1-score	support
0	0.75	0.76	0.75	160
1	0.85	0.85	0.85	261
accuracy			0.81	421
macro avg	0.80	0.80	0.80	421
weighted avg	0.81	0.81	0.81	421

```

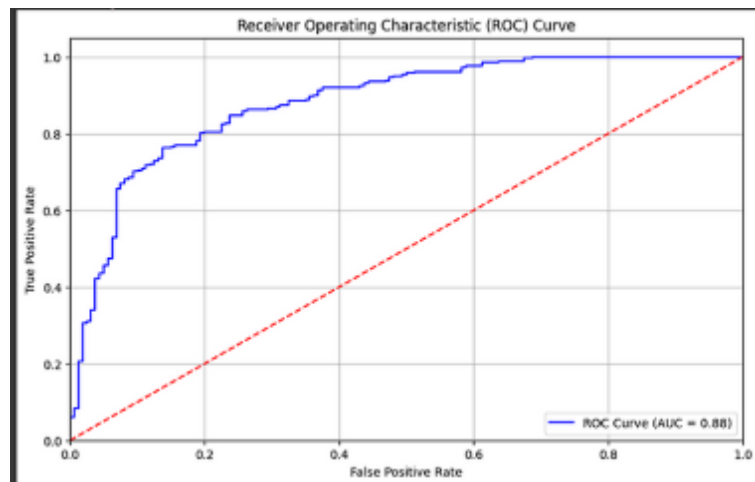
Akurasi Data Latih: 81.12%
Akurasi Data Uji: 81.24%

```

Gambar 5. Hasil uji Naïve Bayes

Setelah melakukan pelatihan pada data yang sudah di preprocessing, bisa disimpulkan bahwa sanya model yang digunakan untuk pengklasifikasian dapat dikatakan sangat baik, dengan akurasi pada data latih mencapai 97,57% dan akurasi pada data uji sebesar 93,35%. Namun, terdapat risiko overfitting

karena model terlalu unggul dalam memproses data latih. Berdasarkan nilai F1-Score, dapat disimpulkan bahwa model bekerja lebih baik dalam mengklasifikasikan kelas 1 dibandingkan kelas 0. Untuk hasil evaluasi menggunakan ROC dapat dilihat pada gambar .



Gambar 6. Hasil evaluasi ROC pada model Naïve Bayes

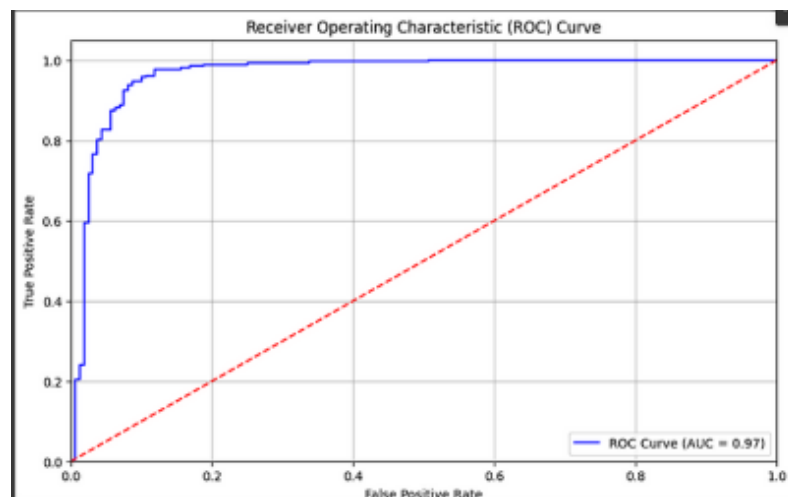
4.2.2. Random forest

Random Forest adalah gabungan dari sejumlah pohon keputusan (*trees*) yang dibangun sedemikian rupa, di mana setiap pohon bergantung pada sampel acak (*random*) yang diambil secara independen dan memiliki distribusi yang sama untuk semua pohon dalam hutan (*forests*) [25].

Classification Report:				
	precision	recall	f1-score	support
0	0.92	0.90	0.91	160
1	0.94	0.95	0.95	261
accuracy			0.93	421
macro avg	0.93	0.93	0.93	421
weighted avg	0.93	0.93	0.93	421
Akurasi Data Latih: 97.57%				
Akurasi Data Uji: 93.35%				

Gambar 7. Hasil uji Random Forest

Model yang digunakan untuk pengklasifikasian dapat dinilai sangat baik, dengan akurasi pada data latih sebesar 97,57% dan akurasi pada data uji sebesar 93,35%. Namun, model ini masih berisiko mengalami overfitting karena performanya yang terlalu optimal pada data latih. Berdasarkan analisis F1-Score, model menunjukkan kinerja yang lebih baik dalam mengklasifikasikan kelas 1 dibandingkan kelas 0. Untuk hasil evaluasi menggunakan ROC dapat dilihat pada gambar .



Gambar 8. Hasil evaluasi ROC pada model Random Forest

4.2.3. Regresi Logistik

Regresi logistik (*logistic regression*) adalah bagian dari analisis regresi yang digunakan ketika variabel dependen (*respon*) berupa variabel dikotomi, yaitu yang terdiri dari dua nilai yang mewakili terjadinya atau tidaknya suatu kejadian, yang biasanya diberi angka 0 atau 1. Berbeda dengan regresi linier biasa, regresi logistik tidak mengasumsikan hubungan linier antara variabel independen dan dependen [26].

```
Confusion Matrix:
[[116  44]
 [ 32 229]]
True Positive (TP): 229
True Negative (TN): 116
False Positive (FP): 44
False Negative (FN): 32
Classification Report:
              precision    recall  f1-score   support

     0       0.78         0.72         0.75         160
     1       0.84         0.88         0.86         261

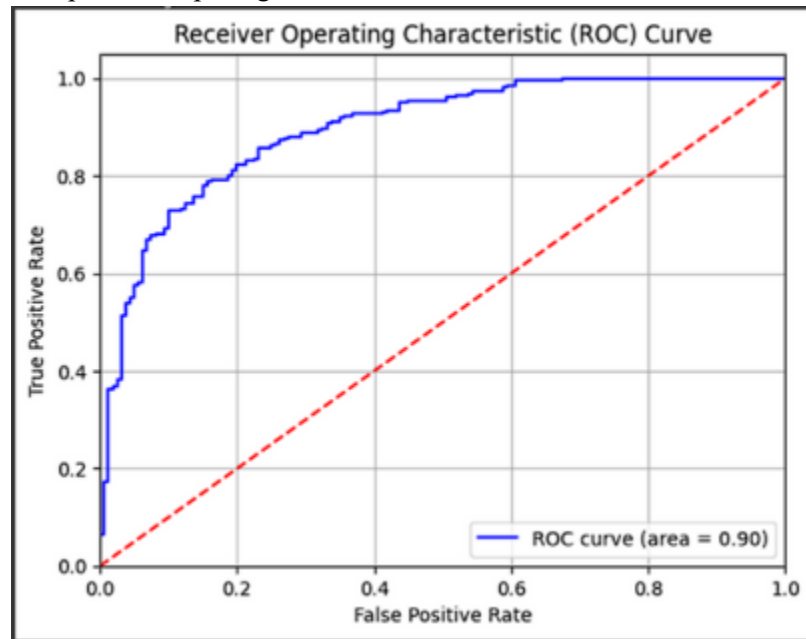
 accuracy          0.82         0.82         0.82         421
 macro avg         0.81         0.80         0.81         421
 weighted avg      0.82         0.82         0.82         421

Akurasi Data Latih: 82.13%
Akurasi Data Uji: 81.95%
```

Gambar 9. Hasil uji Regresi Logistik

Setelah dilakukan pelatihan, dapat disimpulkan bahwa model Logistic Regression memiliki performa yang baik dalam pengklasifikasian. Hal ini ditunjukkan oleh akurasi data latih sebesar 82,13% dan akurasi data uji sebesar 81,95%, yang menunjukkan perbedaan yang kecil sehingga model tidak mengalami underfitting maupun overfitting. Berdasarkan data F1-Score, model juga menunjukkan

kinerja yang lebih optimal dalam mengklasifikasikan kelas 1 dibandingkan kelas 0. Untuk hasil evaluasi menggunakan ROC dapat dilihat pada gambar 10.



Gambar 10. Hasil evaluasi ROC pada model Regresi Logistik

4.3. Hasil Klasifikasi

Model Confusion Matrix akan digunakan untuk menguji akurasi prediksi yang dibuat berdasarkan hasil klasifikasi menggunakan algoritma perbandingan Naïve Bayes, Random Forest, dan Regresi Logistik dalam klasifikasi diagnosis penyakit Parkinson [3].

Tabel 2. Hasil perbandingan klasifikasi antar model

Jenis data	Akurasi		
	Naïve Bayes	Random Forest	Regresi Logistik
Data latih	81.12%	97.57%	82.13%
Data uji	81.24%	93.35%	81.95%

Dalam perbandingan performa ketiga model menggunakan skor ROC dan Matriks Konfusi, model Random Forest menunjukkan hasil yang sangat unggul dengan skor ROC mencapai 0,97. Hal ini mengindikasikan kemampuannya yang luar biasa dalam membedakan antar kelas secara efektif. Secara keseluruhan, model Random Forest muncul sebagai pilihan terbaik untuk tugas klasifikasi ini, berkat akurasinya yang tinggi dalam memprediksi hasil, menjadikannya model yang paling andal di antara model-model yang diuji.

5. KESIMPULAN

Kesimpulan dari penelitian ini menunjukkan bahwa model Random Forest unggul dalam mengklasifikasikan penyakit Parkinson dibandingkan dengan model lain yang diuji, seperti Logistic Regression. Dengan skor ROC yang mengesankan sebesar 0,97, atau 97%, Random Forest mampu membedakan antara kelas dengan sangat baik, menunjukkan kemampuannya dalam menangani

perbedaan yang kompleks antara kelas yang ada. Model ini juga menunjukkan akurasi yang tinggi baik pada data latih maupun data uji, menjadikannya pilihan yang paling andal dalam tugas klasifikasi ini.

Di sisi lain, meskipun model Logistic Regression memiliki akurasi yang baik dengan nilai yang hampir setara antara data latih dan data uji, serta tidak menunjukkan gejala underfitting atau overfitting, Random Forest tetap menunjukkan performa superior dalam hal akurasi dan kemampuan prediksi yang lebih stabil.

Secara keseluruhan, model Random Forest terbukti menjadi model yang paling efektif dalam mengklasifikasikan penyakit Parkinson, dengan keunggulan dalam membedakan antara kelas serta ketahanan terhadap masalah overfitting. Hal ini membuatnya lebih dapat diandalkan dan cocok untuk digunakan dalam diagnosis dini penyakit Parkinson menggunakan data yang tersedia.

Namun, perlu diperhatikan bahwa dataset yang digunakan dalam penelitian ini berasal dari Kaggle. Meskipun dataset tersebut menyediakan fitur yang relevan untuk analisis, penting untuk mempertimbangkan keterbatasannya dalam hal representasi data di dunia nyata. Ukuran sampel, distribusi data, serta kemungkinan adanya bias dalam pengambilan data dapat mempengaruhi generalisasi model ketika diterapkan pada populasi yang lebih luas. Oleh karena itu, penelitian lanjutan dengan dataset yang lebih beragam dan realistis diperlukan untuk memastikan keandalan model dalam skenario medis yang sebenarnya.

6. SARAN

Berdasarkan hasil penelitian ini, beberapa saran untuk penelitian selanjutnya adalah sebagai berikut:

1. Peningkatan Proses Preprocessing: Meskipun langkah preprocessing yang dilakukan sudah cukup efektif, kemungkinan ada fitur tambahan yang relevan namun belum dimanfaatkan sepenuhnya. Penelitian selanjutnya bisa lebih mengeksplorasi teknik seleksi fitur atau penerapan metode ekstraksi fitur lanjutan untuk meningkatkan akurasi model.
2. Eksplorasi Algoritma Lain: Penelitian ini hanya menggunakan tiga algoritma, yaitu Naive Bayes, Random Forest, dan Logistic Regression. Untuk penelitian lebih lanjut, bisa mencoba algoritma lain seperti Support Vector Machines (SVM), Gradient Boosting, atau Neural Networks, yang mungkin memiliki kemampuan lebih baik dalam mengklasifikasikan penyakit Parkinson.
3. Peningkatan Ukuran Dataset: Dalam penelitian ini, dataset yang digunakan berjumlah 2.105 data. Dengan jumlah data yang terbatas, penelitian berikutnya sebaiknya melibatkan dataset yang lebih besar atau menggabungkan data dari berbagai sumber, untuk meningkatkan akurasi dan generalisasi model. Penggunaan data yang lebih beragam dapat membantu model untuk memahami lebih banyak variasi terkait penyakit Parkinson.
4. Pengujian dengan Data Dunia Nyata: Untuk meningkatkan validitas temuan, uji model menggunakan data dunia nyata atau data klinis yang lebih luas dapat dilakukan. Ini dapat memberikan gambaran lebih jelas mengenai bagaimana model dapat diterapkan dalam kondisi nyata untuk mendukung diagnosis dini penyakit Parkinson.

DAFTAR PUSTAKA

- [1] S. Alia *et al.*, "Penyakit Parkinson: Tinjauan Tentang Salah Satu Penyakit Neurodegeneratif yang Paling Umum," *Aksona*, vol. 1, no. 2, pp. 95–99, 2022, doi: 10.20473/aksona.v1i2.145.
- [2] A. R. Onibala, C. D. Mambo, and A. S. R. Masengi, "Peran Vitamin dalam Penanganan Penyakit Parkinson," *Jurnal Biomedik (Jbm)*, vol. 13, no. 3, p. 322, 2021, doi: 10.35790/jbm.13.3.2021.31956.

- [3] W. Z. Aprilita, R. Akbar, R. Cahyadi Prayogi, T. Informatika, and S. Amik Riau, "SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat Comparison of K-Nearest Neighbor (KNN) and Naive Bayes Algorithms in the Classification of Parkinson's Disease Komparasi Algoritma K-Nearest Neighbor (KNN) dan Naive Bayes dalam Klasifikasi P," pp. 188–193, 2023, [Online]. Available: <https://journal.irpi.or.id/index.php/sentimas>
- [4] N. Fajar Susanti, S. Ns Ida Djafar, Mk. Ns Robiul Fitri Masithoh, Mk. dr Frisca Angreni, Mb. Deniyati, and Ms. dr Atika Indah Sari Ns Tria Prasetya Hadi, *Penyakit Muskuloskeletal*. 2024. [Online]. Available: <http://repository.uki.ac.id/15073/>
- [5] O. B. Tysnes and A. Storstein, "Epidemiology of Parkinson's disease," *J Neural Transm*, vol. 124, no. 8, pp. 901–905, 2017, doi: 10.1007/s00702-017-1686-y.
- [6] W. Muangpaisan, H. Hori, and C. Brayne, "Systematic review of the prevalence and incidence of Parkinson's disease in Asia," *J Epidemiol*, vol. 19, no. 6, pp. 281–293, 2009, doi: 10.2188/jea.JE20081034.
- [7] S. A. Tanazza and L. M. Erawati, "Analisis Intervensi Fisioterapi Pada Penyakit Parkinson Menggunakan Vosviewer," *Physio Journal*, vol. 2, no. 2, pp. 49–60, 2022, doi: 10.30787/phyjou.v2i2.877.
- [8] F. M. Januarsyah, E. Zuhairi, and R. M. Firsandaya, "Perbandingan Algoritma Random Forest, Decision Stump, Naïve Bayes, Bayesian Network dan Algoritma C4.5 Untuk Prediksi Pola Kartu Poker," *Prosiding Annual Research Seminar*, vol. 5, no. 1, pp. 122–126, 2019, [Online]. Available: <https://archive.ics.uci.edu/ml/datasets/Poker+Hand>.
- [9] K. Chandel, V. Kunwar, S. Sabitha, T. Choudhury, and S. Mukherjee, "A comparative study on thyroid disease detection using K-nearest neighbor and Naive Bayes classification techniques," *CSI Transactions on ICT*, vol. 4, no. 2–4, pp. 313–319, 2016, doi: 10.1007/s40012-016-0100-5.
- [10] Linlin Lindayani, Dewi Marfuah, Diwa Agus Sudrajat, and Eva Supriatin, "Literature Review Efektivitas Latihan Aerobik Dalam Meningkatkan Fungsi Motorik Pada Lansia Dengan Penyakit Parkinson," *Risenologi*, vol. 6, no. 1a, pp. 100–108, 2021, doi: 10.47028/j.risenologi.2021.61a.220.
- [11] D. Intern, "Apa itu Machine Learning? Beserta Pengertian dan Cara Kerjanya," dicoding. [Online]. Available: <https://www.dicoding.com/blog/machine-learning-adalah/>
- [12] A. Roihan, P. A. Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *IJCIT (Indonesian Journal on Computer and Information Technology)*, vol. 5, no. 1, pp. 75–82, 2020, doi: 10.31294/ijcit.v5i1.7951.
- [13] H. Abijono, P. Santoso, and N. L. Anggreini, "Algoritma Supervised Learning Dan Unsupervised Learning Dalam Pengolahan Data," *Jurnal Teknologi Terapan: G-Tech*, vol. 4, no. 2, pp. 315–318, 2021, doi: 10.33379/gtech.v4i2.635.
- [14] A. Z. Macfud, A. P. Kusuma, and W. D. Puspitasari, "Analisis Algoritma Naive Bayes Classifier (Nbc)," vol. 7, no. 1, pp. 87–94, 2023.
- [15] R. Supriyadi, W. Gata, N. Maulidah, and A. Fauzi, "Penerapan Algoritma Random Forest Untuk Menentukan Kualitas Anggur Merah," *E-Bisnis : Jurnal Ilmiah Ekonomi dan Bisnis*, vol. 13, no. 2, pp. 67–75, 2020, doi: 10.51903/e-bisnis.v13i2.247.
- [16] Y. S. Nugroho and N. Emiliyawati, "Sistem Klasifikasi Variabel Tingkat Penerimaan Konsumen Terhadap Mobil Menggunakan Metode Random Forest," *Jurnal Teknik Elektro*, vol. 9, no. 1, pp. 24–29, 2017.
- [17] G. A. Sandag, "Prediksi Rating Aplikasi App Store Menggunakan Algoritma Random Forest," *CogITo Smart Journal*, vol. 6, no. 2, pp. 167–178, 2020, doi: 10.31154/cogito.v6i2.270.167-178.
- [18] N. Ranti, M. 1*, K. H. Hanif, C. Nisa, S. Informasi, and B. Kaltara, "Perbandingan Algoritma Regresi Logistik, Support Vector Machine, dan Gradient Boosting Pada Analisis Sentimen Data Komentar Siswa," vol. 4, no. 2, pp. 27–32, 2023, [Online]. Available: <http://creativecommons.org/licences/by/4.0/%0Ahttp://ejournal.uhb.ac.id/index.php/IKOMTI>

- [19] A. Regresi, "Penerapan CRISP-DM untuk Deteksi Eksoplanet menggunakan," vol. 4, pp. 160–169, 2024.
- [20] T. Purwa, "Perbandingan Metode Regresi Logistik dan Random Forest untuk Klasifikasi Data Imbalanced (Studi Kasus: Klasifikasi Rumah Tangga Miskin di Kabupaten Karangasem, Bali Tahun 2017)," *Jurnal Matematika, Statistika dan Komputasi*, vol. 16, no. 1, p. 58, 2019, doi: 10.20956/jmsk.v16i1.6494.
- [21] A. Salim and M. R. Alfian, "Optimalisasi Regresi Logistik Menggunakan Algoritma Genetika Pada Data Klasifikasi," *Jurnal Teknologi Informasi dan Terapan*, vol. 6, no. 2, pp. 50–55, 2019, doi: 10.25047/jtit.v6i2.109.
- [22] W. Li, P. Liu, Q. Zhang, and W. Liu, "An improved approach for text sentiment classification based on a deep neural network via a sentiment attention mechanism," *Future Internet*, vol. 11, no. 4, 2019, doi: 10.3390/FI11040096.
- [23] E. Fauziningrum, M. Pd and M. P. Encis Indah Suryaningsih, S.T., "Evaluasi Dan Prediksi Penguasaan Bahasa Inggris Maritim Menggunakan Metode Decision Tree Dan Confusion Matrix (Studi Kasus Di Universitas Maritim Amni)," *Angewandte Chemie International Edition*, 6(11), 951–952., pp. 5–24, 2021.
- [24] Widia, Z. Y. Aqsalia, S. Sari, N. U. Khoirunisa, and F. Kurniawan, "Optimasi Algoritma Naive Bayes Untuk Menganalisis Sentimen Pada Konten Pemindahan Ibu Kota di Youtube," *Journal of Computer and Information Systems Ampera*, vol. 5, no. 2, pp. 68–83, 2024.
- [25] C. Yulianto, "Model Penilaian Tanah Massal Berbasis Parcel-Based Mass Valuation Using Random Forest in Surakarta City," pp. 26–39, 2024.
- [26] A. M. Widodo, Y. S. Anggraeni, N. Anwar, A. Ichwani, and B. A. Sekti, "Performansi K-NN, J48, Naive Bayes dan Regresi Logistik sebagai Algoritma Pengklasifikasi Diabetes," *Prosiding SISFOTEK*, vol. 5, no. 1, pp. 27–33, 2021, [Online]. Available: https://scholar.google.com/citations?view_op=view_citation&hl=en&user=FOwZ8hUAAAJ&pagesize=100&citation_for_view=FOwZ8hUAAAJ:a3BOISfXSfwC

