BAB IV
HASIL DAN PEMBAHASAN

Pada bab ini penulis akan membahas mengenai perhitungan data yang digunakan. Data tersebut akan dihitung menggunakan algoritma K-Nearest Neighbor (KNN) kemudian di optimasi menggunakan Teknik Bagging.

4.1. Persiapan Data

Data pada penelitian penyakit Stroke ini di dapat langsung dari situs resmi www.kaggle.com(https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset) yang di akses di akses pada tanggal 17 Juli 2024, Jumlah record data sebanyak 5.110 data yang merupakan banyaknya pasien yang mengalami gejala dan jumlah atribut dataset sebanyak 12 atribut yang merupakan gejala-gejala yang dialami pasien Stroke. Tabel berikut merupakan data asli yang akan di proses mulai dari tahap preprocesing, Pembagian data uji dan data latih, implementasi metode KNN serta optimasi menggunakan Teknik Bagging. Atribut-atribut tersebut dijelaskan pada Gambar 4.1.

id	stroke	gender	age	hypertension	heart_disea...	ever_married	work_type	Residence_t...	avg_glucos...	bmi	smoking_st...
29540	0	Male	67	0	0	Yes	Private	Rural	97.040	26.9	smokes
53525	0	Female	72	0	0	Yes	Private	Urban	83.890	33.1	formerly smo...
65411	0	Female	51	0	0	Yes	Private	Urban	152.560	21.8	Unknown
26214	0	Female	63	0	0	Yes	Self-employed	Rural	75.930	34.7	formerly smo...
22190	0	Female	64	1	0	Yes	Self-employed	Urban	76.890	30.2	Unknown
56714	0	Female	1	0	0	No	children	Rural	62.130	16.8	Unknown
4211	0	Male	26	0	0	No	Govt_job	Rural	100.850	21	smokes
6369	0	Male	59	1	0	Yes	Private	Rural	95.050	30.9	never smoked
56799	0	Male	76	0	0	Yes	Govt_job	Urban	82.350	38.9	never smoked
32235	0	Female	45	1	0	Yes	Govt_job	Rural	95.020	N/A	smokes
28048	0	Male	13	0	0	No	children	Urban	82.380	24.3	Unknown
68598	0	Male	1	0	0	No	children	Rural	79.150	17.4	Unknown
41512	0	Male	57	0	0	Yes	Govt_job	Rural	76.620	28.2	never smoked
64520	0	Male	68	0	0	Yes	Self-employed	Urban	91.680	40.8	Unknown
579	0	Male	9	0	0	No	children	Urban	71.880	17.5	Unknown
7293	0	Male	40	0	0	Yes	Private	Rural	83.940	N/A	smokes
68398	0	Male	82	1	0	Yes	Self-employed	Rural	71.970	28.3	never smoked
36901	0	Female	45	0	0	Yes	Private	Urban	97.950	24.5	Unknown
45010	0	Female	57	0	0	Yes	Private	Rural	77.930	21.7	never smoked
22127	0	Female	18	0	0	No	Private	Urban	82.850	46.9	Unknown
14180	0	Female	13	0	0	No	children	Rural	103.080	18.6	Unknown

Gambar 4.1 Potongan Dataset

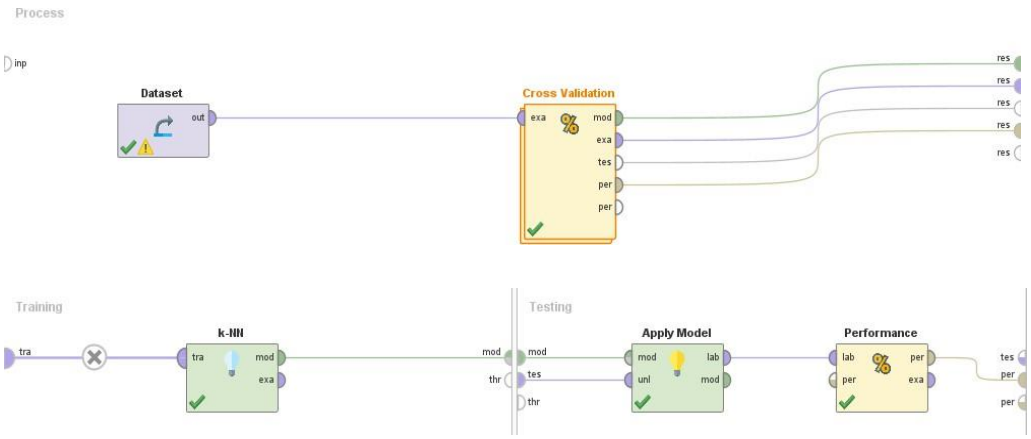
Gambar 4.1. menunjukan karakter data pada penelitian ini, data tersebut memiliki label pada atribut Stroke yang berarti data ini merupakan data supervised karena memiliki label, label pada data ini adalah 1 jika pasien mengalami stroke atau 0 jika tidak. Alat yang digunakan pada penilitian ini adalah aplikasi RapidMiner(Altair AI Studio).

4.2. Pemodelan

Pemilihan dan penerapan teknik pemodelan yang sesuai dilakukan pada tahapan ini. Pemodelan pada penelitian ini menggunakan teknik data mining prediksi.

4.2.1. Penelitian Menggunakan Algoritma K-Nearest Neighbor

Penerapan data pada Rapidminer untuk Prediksi Penyakit Stroke menggunakan algoritma K-Nearest Neighbor ditunjukan pada gambar 4.2.



Gambar 4.2 Pemodelan Algoritma KNN pada Rapidminer

Pada gambar 4.2 dataset yang sudah disiapkan diterapkan pada aplikasi rapidminer dengan melakukan eksperimen menggunakan cross validation yang secara langsung dapat membagi data menjadi data training dan data testing karena data yang digunakan adalah supervised dan algoritma yang digunakan adalah KNN. Hasil eksperimen dapat kita lihat pada gambar 4.3.

accuracy: 94.50% +/- 0.28% (micro average: 94.50%)

	true 1	true 0	class precision
pred. 1	3	35	7.89%
pred. 0	246	4826	95.15%
class recall	1.20%	99.28%	

Gambar 4.3 Confusion Matrix Algoritma KNN

Gambar 4.3 adalah confusion matrix yang menunjukan hasil eksperimen, didalam confusion matrix kita dapat melihat hasil akurasi, class presisi, dan class recall. Akurasi yang dihasilkan adalah 94,50 %. Hasil akurasi pada penelitian sebelumnya dengan eksperimen penelitian ini dengan menggunakan algoritma KNN untuk prediksi penyakit Stroke ditunjukan pada tabel 4.1.

Tabel 4.1 Pebandingan Hasil

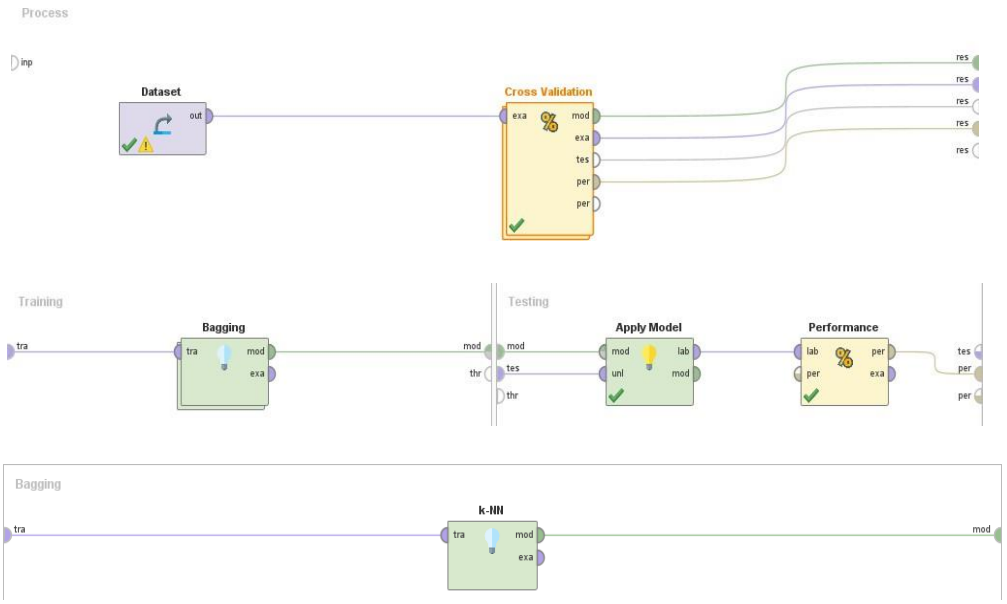
Penelitian	**Tingkat Akurasi**
Penelitian sebelumnya	**93,54%**
Penelitian saat Ini	**94,50 %**

Eksperimen penelitian ini dengan pengujian K-Fold 9 pada cross validation menghasilkan tingkat akurasi paling tinggi dan tingkat akurasi lebih tinggi dari penelitian sebelumnya yang dilakukan oleh Zuriati dkk, akan tetapi masih dapat ditingkatkan tingkat akurasinya.

4.2.2. Penelitian Menggunakan Algoritma KNN dan Bagging

Metode yang dapat meningkatkan tingkat akurasi adalah penggunaan teknik optimasi. Eksperimen lanjutan yang dilakukan pada penelitian ini adalah penggunaan teknik optimasi menggunakan teknik Bagging atau Bootstrap

aggregating. Penerapan pada rapidminer prediksi penyakit stroke menggunakan Algoritma KNN dan Bagging dapat kita lihat pada gambar 4.4.



Gambar 4.4 Penerapan Rapidminer Algoritma KNN dan Bagging

Ekperimen pada gambar 4.4 yaitu teknik validasi untuk membagi data training dan data testing menggunakan teknik validasi cross validation. Dari ekperimen tersebut menndapatkan hasil yang dapat kita lihat pada gambar 4.5.

accuracy: 95.03% +/- 0.11% (micro average: 95.03%)

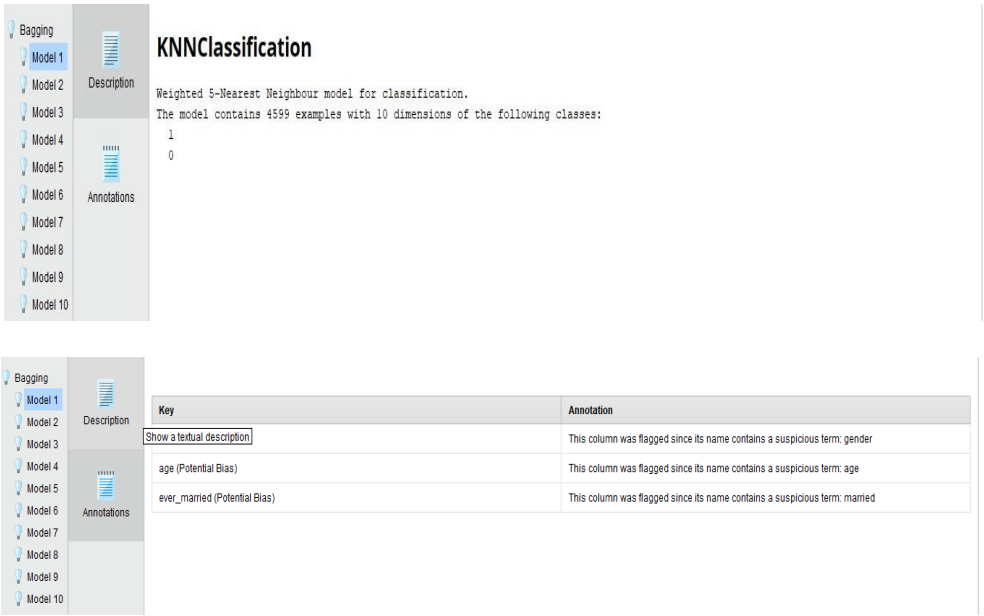
	true 1	true 0	class precision
pred. 1	0	5	0.00%
pred. 0	249	4856	95.12%
class recall	0.00%	99.90%	

Gambar 4.5 Confusion Matrix Algoritma KNN dan Bagging

Hasil dari eksperimen dapat diketahui number of fold yang menghasilkan baik akurasi,adalah penggunaan number of fold 2. dimana hasil akurasi 95,03%.

Hasil yang didapatkan pada ekperimen ini tidak lebih baik dari ekperimen yang hanya menggunakan algoritma KNN.

Dalam RapidMiner, model bagging akan mengambil beberapa subset dari dataset (bootstrap sampling) dan melatih beberapa model KNN pada subset tersebut. Setelah model dilatih, prediksi dari masing-masing model KNN akan digabungkan, umumnya dengan voting mayoritas untuk prediksi. Model yang dilatih dapat dilihat pada Gambar 4.6.



Gambar 4.6 Model Bagging

Model diuji dengan 4599 contoh yang terdiri dari 10 dimensi. Kelas yang diidentifikasi adalah 1 dan 0, di mana 1 mungkin menunjukkan adanya stroke dan 0 menunjukkan tidak ada. Dalam pengujian ini, model menggunakan metode weighted, yang berarti bahwa tetangga yang lebih dekat memiliki pengaruh lebih

besar dalam keputusan klasifikasi dibandingkan tetangga yang lebih jauh. Hasil dari pengujian ini dapat diukur menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Kemudian Hasil pengujian di RapidMiner menunjukkan bahwa beberapa kolom dalam dataset Anda telah ditandai sebagai "Potential Bias" karena nama kolom tersebut mengandung istilah yang dianggap sensitif atau mencurigakan. Berikut penjelasannya:

1. Age (Potensi Bias)

   Penjelasan: Kolom ini berisi informasi tentang usia individu. Istilah "age" dapat menjadi sensitif karena berkaitan dengan diskriminasi usia dalam analisis.

   Dampak: Penggunaan informasi usia dalam model dapat menyebabkan bias dalam prediksi, terutama jika usia berpengaruh besar terhadap hasil (misalnya, risiko stroke).

2. *Ever Married* (Potensi Bias)

   Penjelasan: Kolom ini menunjukkan status pernikahan individu (apakah pernah menikah atau tidak). Istilah "married" dapat memunculkan bias gender atau sosial.

   Dampak: Penggunaan status pernikahan dapat memperkuat stereotip atau menimbulkan diskriminasi berdasarkan status sosial.

3. Gender (Potensi Bias)

   Penjelasan: Kolom ini berisi informasi tentang jenis kelamin individu. Istilah "gender" sering dianggap sensitif dalam konteks analisis data.

   Dampak: Keterlibatan informasi gender dapat menyebabkan bias dalam hasil prediksi, mengingat risiko kesehatan dan perilaku bisa berbeda antara gender.

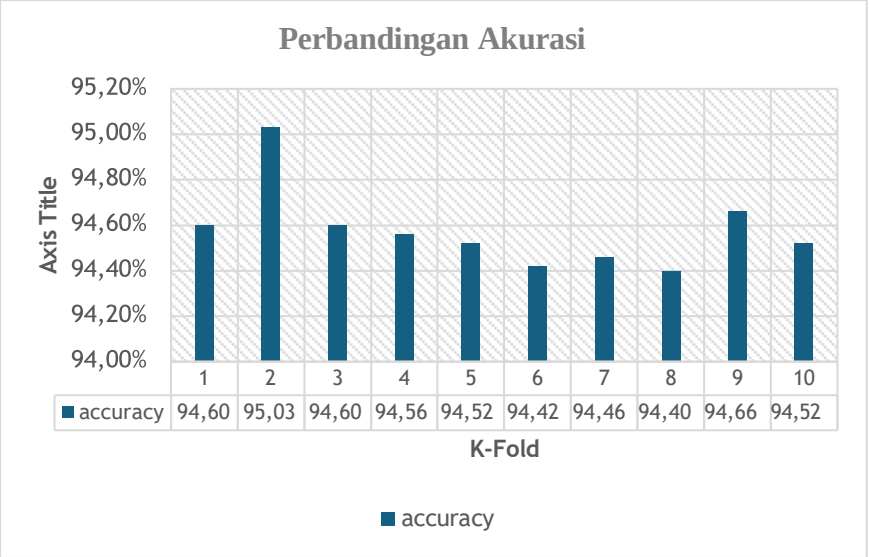
4.2.3. Perbandingan Penelitian Menggunakan Bagging dan Tidak

Perbandingan hasil eksperimen yang tidak menggunakan metode Bagging dan eksperimen yang menggunakan Bagging ditunjukan pada tabel 4.2.

Tabel 4.2 Perbandingan Hasil Akurasi Teknik Bagging

Penelitian	Tingkat Akurasi
Eksperimen Tidak Menggunakan MetodeBagging	94,50%
Eksperimen Yang Menggunakan MetodeBagging	95,03 %

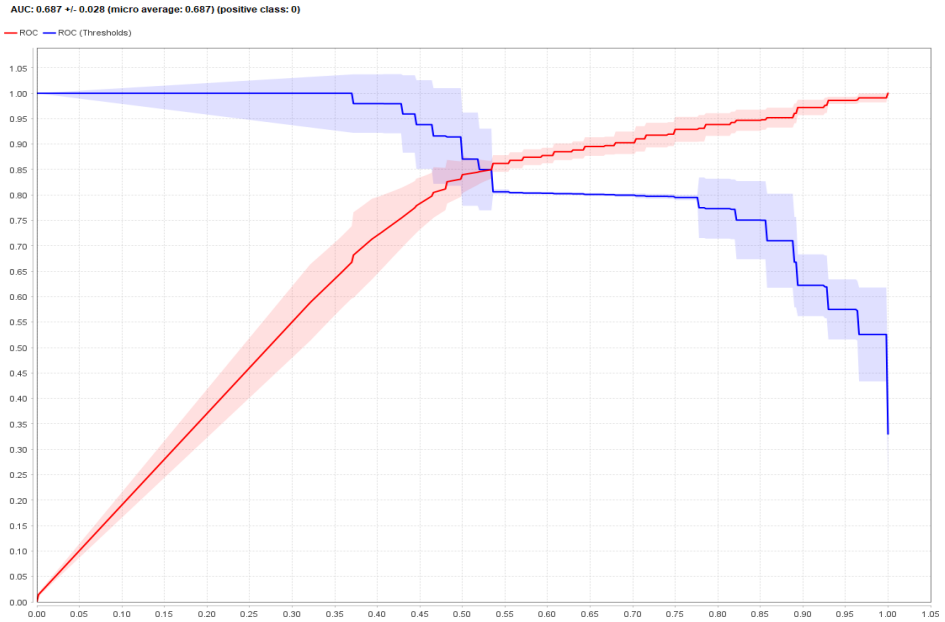
Hasil dari eksperimen ini untuk menghasilkan tingkat akurasi yang tinggi sebesar 95,03% menggunakan K-Fold 2, ini merupakan hasil akurasi paling tinggi dalam pengujian. Perbandingan akurasi, dan recall dapat kita lihat pada table dan gambar 4.7.



Gambar 4.7 Perbandingan Akurasi

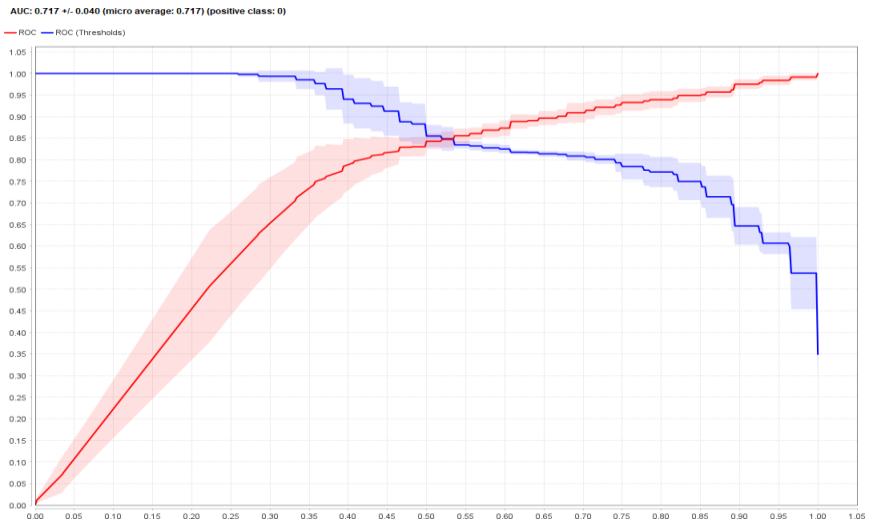
Berdasarkan tabel dan gambar diagram diatas, menunjukan bahwa perubahan jumlah K-Fold pada Cross Validation dalam melakukan klasifikasi akan menghasilkan akurasi yang berbeda sehingga kita dapat menghasilkan akurasi yang terbaik. Dapat kita lihat pada K-Fold 2 hasil akurasi dan presisi merupakan yang paling tinggi berbeda dengan hasil recall yang tidak lebih tinggi dari penggunaan K-Fold lainnya.

Selain Confusion matrix untuk mengetahui kinerja dari eksperimen ini kita dapat mengandalkan kurva ROC-AUC yang dihasilkan. Perbandingan hasil Kurva ROC-AUC pada penelitian yang tidak menggunaakan Bagging dan yang menggunakan Bagging dapat kita lihat pada gambar 4.8 dan gambar 4.9 secara berurutan.



Gambar 4.8 Kurva ROC-AUC Hasil Eksperimen Tidak Menggunakan Bagging

Gambar 4.8 menunjukan kinerja algoritma KNN dalam memprediksi penyakit stroke, kinerja algoritma pada eksperimen ini menghasilkan Area Under Curva (AUC) 68.7% termasuk dalam kategori Poor Classification.



Gambar 4.9 Kurva ROC-AUC Hasil Eksperimen Menggunakan Bagging

Gambar 4.9 menunjukan kinerja algoritma KNN dalam memprediksi penyakit stroke, kinerja algoritma pada eksperimen ini menghasilkan Area Under Curva (AUC) 71.7% termasuk dalam kategori Fair Classification.

4.3. Evaluasi

Pemodelan yang telah dilakukan dengan melakukan ekperimen menggunakan Bootstrap aggregating menghasilkan akurasi yang lebih baik. Penggunaan Algoritma KNN dan Bagging memperoleh akurasi yang tinggi yaitu 95,03 %. Merujuk dari model proses yang telah dilakukan, maka Mengatasi potensi bias

dalam data sangat penting untuk membangun model prediksi yang adil dan tidak diskriminatif. Analisis lebih lanjut menunjukkan bahwa saat menggunakan Bagging, atribut yang paling memberikan kontribusi signifikan terhadap peningkatan akurasi meliputi Age, Heart Disease dan Ever Married. Usia pasien memiliki pengaruh signifikan terhadap prediksi stroke. Atribut ini terbukti menjadi indikator utama, kontribusinya lebih stabil berkat teknik Bagging yang mengurangi varians model. Heart Disease Atribut ini menunjukkan kontribusi yang meningkat saat menggunakan Bagging, di mana model dapat mengidentifikasi pola lebih baik dalam subset data yang berbeda. Kemudian Ever Married Atribut ini menunjukkan peningkatan kontribusi dalam model Bagging, berpotensi memberikan wawasan lebih tentang risiko stroke berdasarkan faktor sosial. Sebaliknya, tanpa menggunakan Bagging, kontribusi atribut seperti Age dan Hypertension menjadi lebih dominan, namun menghasilkan model yang lebih rentan terhadap overfitting. Peningkatan usia berkorelasi dengan meningkatnya risiko stroke serta Memiliki riwayat hipertensi juga memberikan kontribusi penting dalam klasifikasi, di mana pasien dengan hipertensi menunjukkan risiko lebih tinggi. Selain itu kinerja algoritma pada penelitian ini juga menunjukkan performa yang sangat baik termasuk dalam kategori Fair Clasification berdasarkan hasil AUC yang memiliki nilai 71.7%.