Prediksi Penyakit Jantung dengan Menggunakan Algoritma XgBoost dan Randomized Search Optimizer

Proposal

Disusun oleh:
Reo Sahobby
123170067

PROGRAM STUDI INFORMATIKA
JURUSAN TEKNIK INFORMATIKA
FAKULTAS TEKNIK INDUSTRI
UNIVERSITAS PEMBANGUNAN NASIONAL "VETERAN"
YOGYAKARTA

2021

Daftar Isi

Daftar Isi ii
Daftar Gambar iv
Daftar Tabel vi
BAB I PENDAHULUAN 1
1.1. Latar Belakang 1
1.2. Rumusan Masalah 3
1.3. Batasan Masalah 3
1.4. Tujuan Penelitian 3
1.5. Manfaat Penelitian 3
1.6. Tahapan Penelitian 3
1.7. Sistematika Penulisan 4
BAB II TINJAUAN LITERATUR 6
2.1. Tinjauan Studi 6
2.2. Tinjauan Pustaka 19
2.2.1. Jantung 19
2.2.2. Deteksi Penyakit Jantung 19
2.2.3. Machine Learning 20
2.2.4. Klasifikasi 21
2.2.5. Algoritma XgBoost 23
2.2.6. Parameter Algoritma XgBoost 25
2.2.7. Randomized Search Optimizer 25
2.2.8. Cross Validation 26
2.2.9. Confusion Matrix 27
BAB III METODOLOGI PENELITIAN 28
3.1. Metode Penelitian 28
3.1. Pengumpulan Data 29
3.2. Pengolahan Data Awal 29
3.3. Analisis kebutuhan 31
3.3.1. Kebutuhan Fungsional 31
3.3.2. Kebutuhan Non Fungsional 31
3.4. Pembuatan Model Klasifikasi 32
3.4.1. Preprocessing 32

3.4.2. Training 34
3.4.3. Evaluasi 58
3.5. Proses Desain 59
3.5.1. Perancangan Sistem 60
3.5.2. Perancangan Proses 61
3.5.3. Perancangan Antarmuka 63
DAFTAR PUSTAKA 64
LAMPIRAN 68

Daftar Gambar

Gambar 2. 1 Grid Search dan Random Search 26
Gambar 2. 2 K-Fold Cross Validation 26
Gambar 3. 1 Tahapan Penelitian 28
Gambar 3. 2 Flowchart Preprocessing 30
Gambar 3. 3 Flowchart Umum XgBoost 35
Gambar 3. 4 Tree Untuk Node Cp=3 37
Gambar 3. 5 Tree Untuk Node Cp=0 38
Gambar 3. 6 Tree Untuk Node Cp=1 39
Gambar 3. 7 Tree Untuk Node Thalac<123 39
Gambar 3. 8 Tree Untuk Node Thalac<163 40
Gambar 3. 9 Tree Untuk Node Thalac<182 41
Gambar 3. 10 Tree Untuk Node Thalac<196 41
Gambar 3. 11 Tree Sementara Yang Terbentuk 42
Gambar 3. 12 Tree Untuk Node Cp=0 42
Gambar 3. 13 Tree Untuk Node Cp=1 43
Gambar 3. 14 Tree Untuk Node Thalac<135 43
Gambar 3. 15 Tree Untuk Node Thalac<188 44
Gambar 3. 16 Tree Sementara Yang Sudah Dibuat 45
Gambar 3. 17 Tree Untuk Node Cp=0 45
Gambar 3. 18 Tree Untuk Node Cp=1 46
Gambar 3. 19 Tree Dengan Node Thalac<135 46
Gambar 3. 20 Hasil Tree Pertama Yang Dibuat 47
Gambar 3. 21 Tree Dan $OValue$ 48
Gambar 3. 22 Tree Kedua Untuk Node Cp=3 50
Gambar 3. 23 Tree Kedua Untuk Node Cp=1 50
Gambar 3. 24 Tree Kedua Untuk Node Cp=0 51
Gambar 3. 25 Tree Kedua Untuk Node Thalac<123 51
Gambar 3. 26 Tree Kedua Untuk Node Thalac<163 52
Gambar 3. 27 Tree Kedua Untuk Node Thalac<182 52
Gambar 3. 28 Tree Kedua Untuk Node Thalac<196 53
Gambar 3. 29 Hasil Tree Kedua Sementara Yang Terbentuk 53
Gambar 3. 30 Tree Kedua Untuk Node Cp=0 54
Gambar 3. 31 Tree Kedua Untuk Node Cp=1 54
Gambar 3. 32 Tree Kedua Untuk Node Thalac<135 55
Gambar 3. 33 Tree Kedua Untuk Node Thalac<188 55
Gambar 3. 34 Hasil Tree Kedua 56
Gambar 3. 35 Tree Kedua Dan $OValue$ 56
Gambar 3. 36 Flowchart Randomized Search 58
Gambar 3. 37 Evaluasi Dengan Confusion Matrix 59
Gambar 3. 38 Model Pengembangan Waterfall 60
Gambar 3. 39 Arsitektur Perangkat Lunak 61
Gambar 3. 40 Flowchart Sistem 61
Gambar 3. 41 Dfd Level 0 62

Gambar 3. 42 Dfd Level 1 62
Gambar 3. 43 Dfd Level 2 Proses Identifikasi 63
Gambar 3. 44 User Interface Sistem 63

Daftar Tabel

Tabel 2. 1 Tabel State Of The Art..... 16
Tabel 2. 2 Parameter Data..... 19
Tabel 2. 3 Parameter Xgboost..... 25
Tabel 2. 4 Tabel Confusion Matrix..... 27
Tabel 3. 1 Kolom Pada Dataset 29
Tabel 3. 2 Spesifikasi Perangkat Keras..... 32
Tabel 3. 3 Spesifikasi Kebutuhan Perangkat Lunak 32
Tabel 3. 4 Contoh Menghapus Outlier..... 32
Tabel 3. 5 Perhitungan Normalisasi Minmax 33
Tabel 3. 6 Contoh Data Onehot Encoding..... 34
Tabel 3. 7 Data Hasil One Hot Encoding 34
Tabel 3. 8 Contoh Dataset Untuk Training..... 36
Tabel 3. 9 Tabel Dengan Nilai Residual..... 37
Tabel 3. 10 Data Terbaru Dan Residu 49
Tabel 3. 11 Data Dan Residual Tree Kedua 57
Tabel 3. 12 Hasil Prediksi..... 59

BAB I
PENDAHULUAN

1.1.Latar Belakang

Jantung merupakan organ dalam manusia yang fungsinya sangatlah penting yaitu untuk mengedarkan darah yang berisi oksigen dan nutrisi ke seluruh tubuh dan untuk mengangkut sisa hasil metobolisme tubuh, sehingga tubuh dapat bekerja dengan optimal. Akan sangat fatal apabila di dalam organ jantung terdapat gangguan seperti penyumbatan pembuluh darah, dan lain-lain. Sehingga menyebabkan jantung tidak dapat bekerja dan dapat menyebabkan kematian. Berdasarkan data dari WHO terdapat sebanyak 7,3 juta penduduk di seluruh dunia meninggal karena penyakit jantung. Penyakit jantung adalah penyakit yang menyerang pada organ jantung yang berkaitan dengan pembuluh darah, contohnya adalah pembuluh darah di organ jantung yang tersumbat. Penyakit ini menyerang pada pembuluh darah arteri karena tejadi proses *arterosklerosis* pada dinding arteri yang menyebabkan penyempitan (Marleni & Alhabib, 2017). Penyakit jantung juga bisa disebut dengan istilah *suddent death* (Widiastuti et al., 2014), karena penyakit jantung tersebut sering kali tidak menimbulkan gejala namun tiba-tiba pembuluh darah di jantung yang tersumbat tidak dapat memompa darah dan menyalurkannya ke seluruh tubuh, sehingga dapat menyebabkan kematian.

Proses pendeteksian apakah seseorang tersebut terkena penyakit jantung dapat dilakukan dengan melakukan konsultasi kepada dokter spesialis jantung yang nantinya akan dilakukan pemeriksaan laboratorium dan dikonsultasikan oleh dokter spesialis jantung (Wibisono & Fahrurozi, 2019). Namun cara tersebut tidaklah efisien, selain memakan waktu yang lama karena proses pemeriksaan, menunggu hasil pemeriksaan, dan konsultasi tentunya memakan waktu yang lama, juga karena memakan biaya yang cukup tinggi. Oleh karena itu perlu dilakukan pendeteksian penyakit jantung secara digital supaya dapat meningkatkan efektifitas kerja. Penelitian yang sudah dilakukan untuk menciptakan pendeteksian penyakit jantung secara digital seperti penelitian yang dilakukan oleh Retnasari (Retnasari & Rahmawati, 2017), penelitian yang dilakukan oleh Wibisono (Wibisono & Fahrurozi, 2019), dan penelitian yang dilakukan oleh Prasetyo (Prasetyo & Prasetiyo, 2020). Penelitian tersebut dilakukan menggunakan data-data hasil rekam jantung yang ada, yang nantinya dipelajari pola-pola datanya dan akan menghasilkan prediksi, berdasarkan data tersebut apakah seseorang ini berpotensi menderita penyakit jantung atau tidak. Salah satu teknik identifikasi penyakit jantung adalah menggunakan metode klasifikasi. Klasifikasi adalah jenis analisis data yang digunakan untuk memprediksi label kelas dari data tersebut (Annisa, 2019).

Dalam kasus prediksi penyakit jantung ini, penelitian-penelitian sebelumnya telah banyak dilakukan dengan menggunakan berbagai algoritma klasifikasi yang ada. Diantaranya adalah penelitian yang dilakukan oleh Retnasari dan Rahmawati, yang melakukan penelitian dengan menggunakan algoritma *Naïve Bayes* dan algoritma C4,5. Penelitian tersebut dilakukan dengan menggunakan 270 data yang bersumber dari UCI *Machine Learning Repository* dengan jumlah *features* yaitu 13, penelitian tersebut dilakukan dengan menggunakan *rapid mider* dan *confusion matrix* untuk menghitung akurasi masing-masing algoritma. Hasil dari penelitian yang dilakukan tersebut menunjukkan bahwa

algoritma *Naïve Bayes* lebih baik dengan mendapatkan nilai akurasi sebesar 86,67% dan algoritma C4.5 mendapat akurasi sebesar 83,70% (Retnasari & Rahmawati, 2017). Penelitian selanjutnya yang dilakukan oleh Ardea dan Achmad, penelitian tersebut dilakukan untuk mencari algoritma terbaik dengan cara membandingkan masing-masing hasil dari algoritma tersebut. Algoritma yang dibandingkan di dalam penelitian tersebut adalah algoritma *Naïve Bayes*, algoritma *Random Forest*, algoritma *Decision Tree*, dan algoritma *K-Nearest Neighbohr*. Hasil dari penelitian tersebut untuk masing-masing algoritma dihitung dengan menggunakan *confusion matrix* dan didapat hasil akurasi untuk masing-masing algoritma sebagai berikut. Algoritma *Random Forest* memiliki nilai akurasi tertinggi dengan 85,67%, kemudian algoritma *Naïve Bayes* dan algoritma *Decision Tree* memiliki nilai akurasi yang sama dengan nilai akurasi 80,33%, dan algoritma *K-Nearest Neighbor* memiliki nilai akurasi paling rendah yaitu 69,67%. Dengan hasil tersebut, algoritma yang terbaik adalah algoritma Random Forest (Wibisono & Fahrurozi, 2019). Selanjutnya penelitian yang dilakukan oleh Erwin Prasetyo dan Budi Prasetiyo, penelitian tersebut dilakukan dengan menerapkan teknik *bagging* pada algoritma C4.5 untuk melihat apakah teknik *bagging* dapat meningkatkan akurasi dari model klasifikasi yang dibuat. Data yang digunakan dalam penelitian tersebut adalah data *Heart Disease* yang diambil dari UCI *Machine Learning* sejumlah 300 data. Hasil dari penelitian tersebut membuktikan bahwa penerapan teknik *bagging* pada algoritma C4.5 dapat meningkatkan akurasi model yang dibuat dengan kenaikan yaitu 8,86% dengan hasil akurasi algoritma C4.5 sebesar 72,98% dan akurasi algoritma C4.5 yang dikombinasikan dengan teknik *bagging* adalah 81,84% (Prasetyo & Prasetiyo, 2020)

Dari berbagai macam algoritma yang sudah digunakan dalam penelitian sebelumnya, tentunya masing-masing algoritma memiliki kelebihan dan kelemahan. Contohnya adalah terjadinya *overfitting*, ciri *overfitting* adalah memiliki hasil *training* yang sangat bagus, namun pada saat dilakukan pengujian terhadap *data testing* diperoleh performa yang buruk (Septadaya et al., 2019). Metode yang memiliki overfitting contohnya adalah C4,5 (Rahayu et al., 2015), yang dalam penelitian tersebut dapat diselesaikan dengan menggunakan *threshold pruning*. Kemudian penelitian dengan algoritma serupa (Afianto et al., 2017), dapat mengatasi *overfitting* menggunakan algoritma Random Forest. Penelitian yang dilakukan untuk memprediksi ketahanan hidup pasien jantung koroner (Kusuma & Srinandi, 2013), menggunakan metode *Partial Least Square* (PLS) untuk mengatasi *overfitting*. Sedangkan penelitian yang dilakukan untuk mendiagnosis penyakit jantung berdasarkan suara (Lubis & Gondawijaya, 2019), menerapkan *cross validation* dalam pembuatan modelnya untuk mengatasi *overfitting*.

Pada penelitian ini algoritma yang dipilih untuk mengatasi permasalahan *overfitting* adalah menggunakan algoritma XgBoost. Algoritma XgBoost adalah algoritma *gradient boosting* yang dibuat dengan *tree-based* yang dapat membuat *boosted tree* secara efisien dan dapat dikerjakan secara paralel (Karo, 2020). Algoritma XgBoost juga memiliki regularization yang dapat berfungsi untuk menghindari overfitting yang terjadi (Zhang et al., 2018).

Berdasarkan latar belakang dan analisis permasalahan yang telah dilakukan. Tujuan dari penelitian ini adalah untuk mengetahui hasil identifikasi penyakit jantung dengan

menggunakan algoritma XgBoost. Selanjutnya, akan dilakukan analisis performa dari model yang dibuat, seperti hasil akurasi. Penelitian ini akan dilakukan dengan menggunakan data rekam jantung berupa data *tabular* yang nantinya berdasarkan data tersebut akan diklasifikasikan ke dalam terkena penyakit jantung, atau tidak terkena penyakit jantung.

1.2. Rumusan Masalah

Sesuai dengan uraian latar belakang yang sudah dijelaskan di atas, rumusan masalah dalam penelitian ini adalah sebagai berikut:

a. Proses identifikasi penyakit jantung yang tidak efisien jika dilakukan dengan cara konvensional.
b. *Overfitting* yang masih terjadi pada algoritma lain dalam mengidentifikasi penyakit jantung.

1.3. Batasan Masalah

Batasan masalah yang ada di dalam penelitian ini adalah sebagai berikut:

a. Data yang digunakan dalam penelitian ini adalah data *Heart Disease* yang diambil dari UCI *Machine Learning*.
b. Algoritma klasifikasi yang digunakan adalah algoritma XgBoost.
c. Hasil identifikasi penyakit jantung adalah terkena penyakit jantung dengan label 1, atau tidak terkena penyakit jantung dengan label 0.

1.4. Tujuan Penelitian

Tujuan yang ingin dicapai dari penelitian ini adalah sebagai berikut:

a. Mengidentifikasi penyakit jantung berdasarkan data *tabular* data rekam jantung.
b. Menerapkan algoritma klasifikasi XgBoost untuk mengidentifikasi penyakit jantung dan mengatasi *overfitting*.

1.5. Manfaat Penelitian

a. Membantu dalam proses deteksi dini penyakit jantung. Sehingga membuat kita semakin sadar akan kesehatan jantung kita.
b. Mengetahui performa model yang dibuat menggunakan algoritma XgBoost dalam menyelesaikan permasalahan identifikasi penyakit jantung

1.6. Tahapan Penelitian

Pada penelitian yang akan dilakukan ini, terdapat beberapa tahapan yang akan dilakukan yaitu sebagai berikut:

a. Studi Literatur

   Tahap pertama yang dilakukan dalam penelitian ini adalah melakukan studi literatur untuk mencari referensi, penelitian sebelumnya, data yang akan digunakan, dan lain-lain. Study literature dapat dicari dari jurnal-jurnal yang membahas penelitian serupa.
b. Pengumpulan Data

   Tahap selanjutnya adalah melakukan pengumpulan data, data yang akan digunakan dalam penelitian ini adalah data sekunder, yaitu data *Heart Disease* yang bersumber dari UCI *Machine Learning Repository*.
c. Pembuatan Model *Machine Learning*

   Tahap selanjutnya adalah pembuatan model *machine learning*, yaitu pada tahap ini dilakukan pembuatan model menggunakan algoritma dan teknik yang sudah dipilih.

d. Pengujian dan Evaluasi Model

   Setelah model *machine learning* dibuat, tahap selanjutnya adalah memastikan model yang dibuat memiliki performa yang baik dalam menangani data. Apabila model dirasa belum maksimal, dapat dilakukan pembuatan model ulang dengan *hyper parameter* yang berbeda dan dilakukan pengujian lagi, diharapkan mendapat peningkatan performa.

e. Implementasi Perangkat Lunak dan Pengujian

   Selanjutnya, setelah model yang dibuat memiliki performa yang bagus, model tersebut diimplementasikan dalam bentuk perangkat lunak yang bisa digunakan oleh pengguna. Dalam pembuatan perangkat lunak ini, menggunakan metodologi *waterfall*. Setelah perangkat lunak selesai dibuat, dilakukan pengujian perangkat lunak untuk memastikan perangkat lunak yang dibuat berjalan normal tanpa ada kendala.

f. Kesimpulan dan Saran

   Setelah semua tahap dilakukan, didapatkan kesimpulan dari penelitian yang sudah dilakukan tentang bagaimana performa algoritma XgBoost dalam menangani kasus permasalahan yang dipilih.

1.7.Sistematika Penulisan

Penelitian ini disusun berdasarkan sistematika penulisan yang terdiri dari 5 bab yang terdiri dari:

BAB 1 PENDAHULUAN

Pada BAB I ini, membahas latar belakang penelitian ini dilakukan, rumusan masalah yang ada di dalam penelitian ini, batasan masalah, tujuan, dan manfaat penelitian ini dilakukan, serta sistematika penulisan laporan mengenai penelitian yang dilakukan.

BAB II TINJAUAN LITERATUR

Dalam BAB II ini, berisi landasan teori mengenai obyek penelitian dan metode yang akan dilakukan di dalam penelitian ini, kemudian juga membahas penelitian-penelitian serupa yang sudah dilakukan sehingga menjadi referensi penulis dalam mengadakan melakukan penelitian ini.

BAB III METODE PENELITIAN

Pada BAB III ini berisi penjelasan tentang metode yang akan digunakan oleh penulis di dalam melakukan penelitian ini. Metode-metode yang dipilih nantinya akan digunakan untuk menyelesaikan permasalahan pada kasus yang sedang diteliti, yaitu identifikasi penyakit jantung.

BAB IV HASIL DAN PEMBAHASAN

Pada BAB IV ini, berisi pemaparan dan penjelasan hasil dari tahapan demi tahapan penelitian yang sudah dilakukan oleh penulis dengan menggunakan metode yang sudah dijelaskan pada bab sebelumnya. Penjelasan hasil penelitian akan berisi evaluasi performa model yang sudah dibuat dengan menggunakan algoritma yang dipilih.

BAB V PENUTUP

Bab ini akan berisi kesimpulan hasil dari penelitian yang sudah dilakukan oleh penulis. Kemudian penulis juga menambahkan kekurangan dari penelitian yang sudah dilakukan ditambahkan dengan saran yang bisa dilakukan pada penelitian yang akan

datang, dapat berupa saran perbaikan data ataupun saran mengenai perbaikan metode supaya penelitian yang akan datang dapat menghasilkan hasil yang lebih maksimal.