DRAFT PROPOSAL SKRIPSI

PREDIKSI PELUANG LOLOS KUALIFIKASI PIALA DUNIA TIMNAS INDONESIA MENGGUNAKAN METODE DECISION TREE DAN REGRESI LOGISTIK

ERICA APRILIA SUTRISNI
NPM 21081010285

DOSEN PEMBIMBING
Dr. Basuki Rahmat, S.Si. MT.

KEMENTERIAN PENDIDIKAN, KEBUDAYAAN, RISET, DAN TEKNOLOGI
UNIVERSITAS PEMBANGUNAN NASIONAL VETERAN JAWA TIMUR
FAKULTAS ILMU KOMPUTER
PROGRAM STUDI INFORMATIKA
SURABAYA
2024

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

BAB I
PENDAHULUAN

1.1. Latar Belakang

Sepak bola merupakan salah satu olahraga paling populer di dunia, termasuk di Indonesia[1]. Dukungan masif dari masyarakat menjadikan sepak bola tidak hanya sekadar olahraga, tetapi juga bagian dari identitas nasional. Sejak pertama kali berpartisipasi dalam kualifikasi Piala Dunia FIFA, harapan agar Tim Nasional (Timnas) Indonesia dapat berlaga di ajang sepak bola tertinggi tersebut selalu menjadi impian besar bagi seluruh lapisan masyarakat. Namun, realisasi dari impian ini masih menjadi tantangan yang perlu dihadapi, mengingat berbagai faktor yang mempengaruhi performa tim dan kompetisi yang semakin ketat di tingkat Asia dan dunia. Meskipun Timnas Indonesia telah mengalami perkembangan yang positif dalam beberapa tahun terakhir, tantangan besar tetap ada, termasuk kualitas lawan yang semakin kuat dan konsistensi permainan yang menjadi kunci keberhasilan.

Dalam beberapa tahun terakhir, Timnas Indonesia menunjukkan perkembangan positif yang signifikan, baik dari segi teknik, taktik, maupun fisik. Peningkatan kualitas pemain, ditunjang dengan strategi pelatihan yang lebih baik, serta dukungan dari berbagai pihak, termasuk federasi sepak bola, pelatih berkualitas, dan pemain muda berbakat, memberikan optimisme bagi masyarakat Indonesia akan peluang timnas untuk lolos ke Piala Dunia. Namun demikian, meskipun banyak kemajuan yang telah dicapai, sejumlah tantangan tetap memengaruhi peluang keberhasilan Indonesia, termasuk kualitas tim lawan yang semakin meningkat, serta faktor non-teknis yang sering kali menjadi hambatan dalam menjaga konsistensi performa tim di kompetisi besar.

Dengan adanya kemajuan teknologi informasi dan komputasi, analisis prediktif berbasis machine learning dan statistik menawarkan pendekatan baru yang dapat digunakan untuk menilai peluang Timnas Indonesia dalam kualifikasi Piala Dunia[2]. Prediksi berbasis data ini dapat memberikan gambaran objektif mengenai potensi Timnas Indonesia berdasarkan berbagai variabel, seperti hasil pertandingan sebelumnya, kualitas tim lawan, jadwal pertandingan, serta faktor-faktor internal tim[3]. Teknologi ini diharapkan dapat membantu pelatih, manajemen, dan penggemar sepak bola Indonesia dalam merencanakan strategi

yang lebih baik dan menyusun langkah-langkah strategis untuk meningkatkan peluang lolos ke Piala Dunia. Dengan pendekatan yang berbasis data, analisis prediktif memberikan wawasan yang lebih dalam dan bisa menjadi referensi penting dalam membuat keputusan yang lebih rasional dan terukur.

Penelitian ini berfokus pada pengembangan sistem prediksi untuk menilai peluang lolos Timnas Indonesia dalam kualifikasi Piala Dunia menggunakan dua algoritma utama, yaitu decision tree dan regresi logistik. Decision tree dipilih karena kemampuannya untuk menghasilkan model prediksi yang jelas dan mudah dipahami. Dalam decision tree, data dibagi berdasarkan atribut yang relevan untuk menghasilkan keputusan, yang membuatnya sangat transparan dan mudah diinterpretasikan[4]. Ini memungkinkan kita untuk mengidentifikasi dengan jelas faktor-faktor utama yang mempengaruhi peluang lolos, serta hubungan antar variabel yang berkontribusi terhadap hasil akhir. Selain itu, decision tree tidak memerlukan asumsi yang kuat tentang distribusi data, sehingga cocok digunakan dalam kasus dengan data yang tidak terstruktur atau kompleks.

Sementara itu, regresi logistik dipilih karena kemampuannya untuk memprediksi probabilitas dari suatu kejadian. Dalam konteks ini, regresi logistik digunakan untuk menghitung peluang lolos Timnas Indonesia ke Piala Dunia berdasarkan variabel yang ada[5]. Hasil dari regresi logistik berupa nilai probabilitas yang berada dalam rentang 0 hingga 1, yang dapat dengan mudah digunakan untuk menilai kemungkinan suatu kejadian terjadi. Keunggulan utama dari regresi logistik adalah kemampuannya untuk memberikan prediksi yang lebih halus dan berbasis probabilitas, yang penting dalam konteks perencanaan jangka panjang dan pengambilan keputusan.

Proses penelitian ini dimulai dengan pengumpulan data yang relevan, seperti hasil pertandingan, gol yang dicetak, jumlah kekalahan, serta posisi tim lawan dalam kualifikasi. Data ini kemudian akan diproses melalui tahap preprocessing, di mana data yang hilang akan diatasi dan data kategorikal akan diubah menjadi numerik. Setelah itu, data akan dibagi menjadi dua bagian: data pelatihan dan data pengujian. Data pelatihan digunakan untuk melatih model menggunakan algoritma decision tree dan regresi logistik, sementara data pengujian digunakan untuk mengukur kinerja model yang telah dibangun. Selanjutnya, model yang dihasilkan akan dievaluasi menggunakan metrik seperti

akurasi, precision, recall, dan F-Score untuk menilai sejauh mana model dapat memprediksi hasil dengan tepat.

Berdasarkan paparan di atas, maka peneliti akan mengangkat sebuah judul penelitian "Prediksi Peluang Lolos Kualifikasi Piala Dunia 2026 Zona Asia Timnas Indonesia menggunakan metode Decision Tree dan Regresi Logistik".

1.2. Rumusan Masalah

Berdasarkan latar belakang di atas, maka dapat dibentuk beberapa poin rumusan masalah yaitu:

1. Bagaimana cara memprediksi peluang lolos Timnas Indonesia dalam kualifikasi Piala Dunia menggunakan algoritma decision tree dan regresi logistik?
2. Faktor-faktor apa saja yang paling berpengaruh terhadap peluang lolos Timnas Indonesia dalam kualifikasi Piala Dunia berdasarkan hasil prediksi yang menggunakan algoritma decision tree?
3. Sejauh mana akurasi prediksi peluang lolos Timnas Indonesia menggunakan algoritma regresi logistik dibandingkan dengan algoritma decision tree dalam kualifikasi Piala Dunia?

1.3. Batasan Masalah

Batasan masalah berisi hal-hal yang membatasi lingkup penelitian

1. Penelitian ini berfokus pada Kualifikasi Piala Dunia 2026 Zona Asia
2. Penelitian ini hanya menggunakan data ranking FIFA, klasemen, jumlah gol, tembakan, tembakan ke arah gawang, penguasaan bola, operan, akurasi operan, kartu kuning, kartu merah, offside, dan tendangan sudut.
3. Metode yang digunakan Decision Tree untuk menentukan Faktor yang mempengaruhi Peluang Lolos, dan metode Regresi Logistik untuk prediksi
4. Evaluasi model dilakukan menggunakan metrik akurasi dengan metode confusion matriks dan evaluasi model.

1.4.Tujuan Penelitian

Berdasarkan rumusan masalah yang ada, maka tujuan dari penelitian ini adalah:

1. Membangun sistem prediksi untuk menilai peluang lolos Timnas Indonesia dalam kualifikasi Piala Dunia menggunakan algoritma decision tree dan regresi logistik.
2. Mengidentifikasi faktor-faktor utama yang mempengaruhi peluang lolos Timnas Indonesia dalam kualifikasi Piala Dunia berdasarkan hasil prediksi yang menggunakan algoritma decision tree.
3. Mengevaluasi dan membandingkan akurasi prediksi peluang lolos Timnas Indonesia menggunakan algoritma regresi logistik dan decision tree untuk menentukan algoritma yang paling efektif.

1.5.Manfaat Penelitian

Penelitian ini memiliki beberapa manfaat, diantaranya:

1. Memberikan wawasan bagi pelatih dan manajemen Timnas Indonesia dalam merencanakan strategi kualifikasi Piala Dunia.
2. Menjadi referensi bagi penggemar sepak bola dan pihak terkait lainnya dalam memahami dan mengikuti perkembangan Timnas Indonesia dalam kualifikasi Piala Dunia.
3. Pengembangan metode analisis prediktif berbasis data dalam bidang olahraga, khususnya sepak bola.

BAB II
TINJAUN PUSTAKA

2.1 Penelitian Terdahulu

Penelitian-penelitian yang sudah dilakukan peneliti sebelumnya yang terkait dengan penelitian yang akan dilakukan adalah sebagai berikut.

1. Penelitian mengenai prediksi hasil pertandingan FIFA World Cup Qatar 2022 menggunakan teknologi Machine Learning dengan Python, yang merupakan kontribusi signifikan dalam bidang analisis data olahraga. Penelitian ini mengaplikasikan berbagai algoritma seperti Logistic Regression, K-Nearest Neighbors, Naïve-Bayes, Support Vector Machine (SVM), Neural Network, dan Random Forest untuk mengevaluasi akurasi prediksi berdasarkan data pertandingan FIFA World Cup dari tahun 2006 hingga 2018. Hasil penelitian menunjukkan bahwa algoritma SVM dan Neural Network memiliki akurasi tertinggi pada edisi tertentu, dengan SVM menunjukkan keunggulan pada tahun 2010 dan 2014, sedangkan Neural Network unggul pada tahun 2006 dan 2018. Selain itu, penelitian ini juga memberikan prediksi bahwa Jerman memiliki peluang besar untuk memenangkan FIFA World Cup 2022, yang dapat menjadi referensi penting bagi penelitian selanjutnya dalam memanfaatkan Machine Learning untuk analisis hasil pertandingan olahraga.
2. Jurnal ini mengkaji penerapan metode klasifikasi Naïve Bayes untuk memprediksi potensi juara La Liga oleh FC Barcelona, menggunakan data historis klasemen dari tahun 1929 hingga 2024. Penelitian ini menunjukkan bahwa algoritma Naïve Bayes efektif dalam menganalisis dan memprediksi peluang keberhasilan klub sepak bola berdasarkan atribut kunci seperti jumlah gol yang dicetak, jumlah kebobolan, dan total poin yang diperoleh. Dengan menggunakan aplikasi RapidMiner, penelitian ini tidak hanya berhasil mengklasifikasikan data, tetapi juga memberikan wawasan yang lebih dalam mengenai dinamika kompetisi La Liga. Hasil penelitian ini dapat dijadikan referensi penting bagi studi-studi selanjutnya yang ingin mengeksplorasi metode klasifikasi dalam konteks olahraga, serta memperluas pemahaman tentang faktor-faktor yang mempengaruhi keberhasilan tim dalam kompetisi sepak bola.

3. Jurnal ini menyajikan penelitian yang berfokus pada penerapan algoritme Naïve Bayes dalam memprediksi juara Liga Primer Inggris musim 2022/2023, dengan menggunakan data statistik historis dari 10 klub teratas selama 20 tahun terakhir. Penelitian ini menunjukkan bahwa metode Naïve Bayes dapat memberikan akurasi yang tinggi, mencapai 91,7%, serta presisi dan recall yang signifikan, yaitu 84,2% dan 88,9% masing-masing. Hasil analisis menunjukkan bahwa hanya dua klub, Arsenal dan Manchester City, memiliki peluang untuk menjadi juara, dengan Manchester City memiliki probabilitas yang jauh lebih tinggi. Temuan ini memberikan kontribusi penting dalam bidang data mining dan pengolahan data, serta menunjukkan potensi algoritme Naïve Bayes sebagai alat yang efektif untuk analisis prediktif dalam konteks olahraga. Penelitian ini dapat dijadikan referensi bagi studi-studi selanjutnya yang ingin mengeksplorasi penggunaan teknik serupa dalam memprediksi hasil kompetisi olahraga lainnya.
4. Jurnal ini membahas penerapan algoritma C.45 dalam memprediksi pemenang klub sepak bola pada ajang Liga Champions UEFA, dengan fokus pada pengolahan data historis yang mencakup berbagai faktor seperti kualitas pemain, umur, kualitas pelatih, dan aspek finansial klub. Penelitian ini menunjukkan bahwa algoritma C.45, yang dikenal dengan kemampuannya dalam membangun pohon keputusan, dapat secara efektif mengklasifikasikan dan memprediksi hasil pertandingan berdasarkan data yang relevan. Hasil penelitian ini memberikan wawasan yang berharga bagi penggemar sepak bola dan pelatih dalam merumuskan strategi, serta menyoroti pentingnya analisis data dalam olahraga modern. Dengan demikian, penelitian ini dapat dijadikan referensi penting bagi studi-studi selanjutnya yang ingin mengeksplorasi penggunaan teknik data mining dan algoritma prediktif dalam konteks kompetisi olahraga lainnya.

2.2 Landasan Teori

Landasan teori merupakan bagian penting dalam penelitian yang memuat teori-teori serta hasil penelitian sebelumnya yang relevan dengan topik yang dibahas. Bagian ini menjadi dasar yang esensial dan wajib ada dalam setiap proses penelitian.

2.2.1 Kualifikasi Piala Dunia

Kualifikasi Piala Dunia FIFA adalah proses penting bagi tim sepak bola

nasional untuk mencapai Putaran Final Piala Dunia, yang diadakan di enam zona kontinental FIFA, termasuk Afrika, Asia, Amerika Utara dan Tengah, Amerika Selatan, Oceania, dan Eropa.

Untuk Piala Dunia FIFA 2026 yang akan berlangsung di Amerika Serikat, Kanada, dan Meksiko, zona Asia akan mengadakan kualifikasi dengan total 8 ⅓ tempat yang tersedia, terdiri dari 8 tempat lolos langsung dan 1 tempat perebutan antar konfederasi. Proses kualifikasi ini melibatkan lima babak, di mana dua babak pertama juga berfungsi sebagai kualifikasi untuk Piala Asia AFC 2027, dengan format yang telah diperbarui pada 1 Agustus 2022 untuk mencerminkan peningkatan alokasi tempat di putaran final. Babak pertama melibatkan 20 tim yang berada di peringkat 27 hingga 46, di mana mereka akan bertanding dalam dua leg untuk menentukan sepuluh pemenang yang melaju ke babak kedua. Pada babak kedua, 36 tim dibagi menjadi sembilan grup, dan juara serta runner-up grup akan melanjutkan ke babak ketiga, di mana 18 tim akan bersaing dalam tiga grup. Dua tim teratas dari setiap grup di babak ketiga akan lolos langsung ke Piala Dunia, sementara tim peringkat ketiga dan keempat akan melanjutkan ke babak keempat. Di babak keempat, enam tim yang tersisa akan dibagi menjadi dua grup, dan pemenang grup akan lolos ke Piala Dunia, sedangkan runner-up akan bertanding di babak kelima untuk menentukan perwakilan Asia dalam play-off antar konfederasi.

2.2.2 Data Mining

Data Mining adalah suatu proses yang bertujuan untuk menggali dan mengekstraksi informasi atau pengetahuan yang berharga dari kumpulan data yang besar, seperti basis data atau repositori database lainnya. Proses ini dilakukan dengan menggunakan teknik dan metode tertentu untuk mengidentifikasi pola, hubungan, atau informasi yang sebelumnya tidak terlihat atau tersembunyi di dalam data tersebut. Tujuan utama dari data mining adalah untuk menemukan wawasan baru yang dapat memberikan nilai tambah atau mendukung pengambilan keputusan berdasarkan informasi yang tersembunyi dalam database tersebut. Pekerjaan dalam data mining dapat diklasifikasikan ke dalam empat kategori utama, yaitu model prediksi (prediction modeling), analisis kluster (cluster analysis), analisis asosiasi (association analysis), dan deteksi anomali (anomaly detection)[6]

2.2.3 Decision Tree

Decision tree adalah suatu metode klasifikasi menggunakan struktur pohon,

dimana setiap node menerangkan atribut dan cabangnya menerangkan nilai dari atribut, sedangkan daunnya digunakan untuk menerangkan kelas. Node paling atas dari decision tree ini dinamakan root. Decision tree adalah salah satu metode yang dapat diterapkan untuk mengklasifikasikan tim objek atau data untuk menghasilkan sebuah keputusan .

Pendekatan ini terbagi dari serangkaian node pilihan, dihubungkan melalui cabang, bergerak menurun ke bawah dari simpul akar sampai berakhir di simpul daun. Pengembangan decision tree dimulai dari simpul akar, terutama berdasarkan pada konvensi yang diposisikan di bagian atas diagram pohon keputusan, semua atribut dievaluasi pada simpul seleksi, dengan setiap outcome yang mungkin menghasilkan cabang. Tiap cabang dapat masuk baik ke decision node yang lain ataupun ke leaf node.

Dalam pohon keputusan dikenal tiga jenis node, yaitu:

1. Akar sebagai node paling atas, node ini tidak memiliki input dan dapat tidak memiliki output atau dapat memiliki output lebih dari satu.
2. Internal node sebagai node percabangan, node ini hanya memiliki satu input dan output minimal dua.
3. Daun sebagai node akhir atau terminal node, node ini hanya memiliki satu input dan tidak memiliki output (simpul terminal).

Konsep dasar dari pohon keputusan adalah mengubah data menjadi model pohon keputusan, kemudian mengubah model pohon menjadi aturan (rule) dan menyederhanakannya. Manfaat utama dari menggunakan pohon keputusan adalah kemampuan untuk menyederhanakan proses pengambilan keputusan yang awalnya kompleks sehingga pembuat keputusan dapat menafsirkan solusi untuk masalah. Proses pembuatan decision tree menggunakan Google Colaboratory diawali dengan membuat perpustakaan yang diperlukan terlebih dahulu, kemudian dilanjutkan dengan membuat Dataset 1 yang sudah siap untuk ditambang yang berasal dari file Excel.

Selanjutnya membagi kolom yang terdapat pada Dataset 1 menjadi dua jenis variabel dependen (variabel target) dan variabel independen (variabel fitur). Variabel dependen atau target adalah data historis pertandingan. Sedangkan variabel independen atau fitur yang digunakan berasal dari kolom “Peringkat FIFA”, “Gol”, “Tembakan”, “Tembakan ke arah gawang”, “Penguasaan bola”, “operan”, dan “akurasi operan”, Berikutnya dataset tersebut dibagi kembali menjadi training set dan test set. Besar test set yang digunakan adalah sebesar 10% dari data yang ada.

Langkah terakhir adalah membuat Model Pohon Keputusan menggunakan Scikit-learn. Pohon keputusan tersebut pun kemudian dievaluasi dengan mengetahui seberapa akurat pengklasifikasian dari model tersebut.

2.2.4 Algoritma C4.5

Algoritma C45 yaitu sebuah algoritma yang digunakann untuk membangun decision tree (pengambilan keputusan). Algoritma C4.5 adalah salah satu algoritma induksi pohon keputusan, yaitu ID3 (Iterative Dichotomiser 3). ID3 dikembangkan oleh J. Ross Quinlan. Dalam prosedur algoritma ID3, input berupa sampel training, label training, dan atribut. Algoritma pengembangan dari ID3.[7]

Secara singkat logika algoritma C4.5 yang digunakan adalah sebagai berikut:

1. Pilih atribut sebagai akar
2. Buat cabang untuk masing-masing nilai
3. Bagi kasus dalam cabang
4. Ulangi proses untuk masing - masing cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Ada beberapa tahap dalam membuat sebuah pohon keputusan dengan algoritma C.45, yaitu :

1. Menyiapkan data traning. Data traning biasanya diambil dari data histori yang pernah terjadi sebelumnya dan sudah dikelompokan ke dalam kelas – kelas tertentu .
2. Menentukan akar dari pohon. Akar akan diambil dari atribut yang terpilih, dengan cara menghitung nilai gain dari masing – masing atribut, nilai gain yang paling tinggi yang akan menjadi akar pertama. Sebelum menghitung nilai gain dari atribut, hitung dahulu nilai entropy. Untuk menghitung nilai entropy digunakan rumus.

   $Entropy(\mathrm{S}) - \sum_{i=1}^{n} - pi.log2\ pi$

   Keterangan :
   S = himpunan kasus
   n = jumlah partisi S
   Pi = proporsi Si terhadap S
3. Menghitung nilai gain menggunakan rumus:

Gain(A) = Entropy(S) – $\sum_{i=n}^{n} \frac{|Si|}{|S|}$ Entropy(Si)

Keterangan :

S = himpunan kasus

A = atribut

n = jumlah partisi atribut A

|Si| = jumlah kasus pada partisi ke i

|S| = jumlah kasus dalam S

4. Ulangi langkah ke-2 hingga semua record terpartisi.
5. Proses partisi pohon keputusan akan berhenti saat :
    5. Semua record dalam simpul N mendapat kelas yang sama.
    6. Tidak ada atribut di dalam record yang dipartisi lagi.
    7. Tidak ada record di dalam cabang yang kosong.

2.2.5 Regresi Logistik

Regresi logistik dalam statistika merupakan suatu teknik analisis data yang bertujuan untuk mengetahui hubungan antara beberapa variabel dimana variabel responnya bersifat kategorik, baik nominal maupun ordinal dengan variabel penjelasnya bersifat kategorik atau kontinu[8].

Secara khusus regresi logistik yang digunakan pada penelitian ini adalah regresi logistik biner. Regresi logistik biner merupakan salah satu pendekatan model matematis yang digunakan untuk menganalisis hubungan beberapa faktor dengan sebuah variabel yang bersifat biner, yaitu variabel respon yang terdiri dari dua kategori misalnya y = 1 menyatakan hasil yang diperoleh ”lolos” dan y = 0 menyatakan hasil yang diperoleh “tidak lolos”.

Dalam kondisi demikian, variabel y mengikuti distribusi Bernoulli untuk setiap observasi tunggal. Fungsi Probabilitas untuk setiap observasi adalah diberikan sebagai berikut

$$f(y) = n^{y}(1-\pi)^{1-y};\ y = 0,\ 1$$

Suatu fungsi $P(X|Y=1)$ dicari dengan menggunakan transformasi logit yaitu $g(x)$ dinyatakan sebagai berikut:

$$P(X|Y=1)\frac{e^{(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}}{1+e^{(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)}}$$

Dimana:

P($X|Y$=1) : probabilitas kejadian berhasil

P : banyaknya variabel prediktor

X_1… : Variabel independen kuantitatif atau kualitatif

β_0 : Konstanta

$\beta_1 \ldots \beta p$: Koefisien variabel regresi

Berdasarkan persamaan, p adalah banyaknya variabel prediktor. Untuk mempermudah pendugaan parameter regresi, P ($X|Y$ = 1) merupakan peluang kejadian sukses dengan nilai probabilitas $0 \leq P(X|Y = 1) \leq 1$ dan merupakan nilai parameter untuk j = 1,2,…,p. $P(X|Y = 1)$ merupakan fungsi yang non linear, sehingga perlu ditransformasi ke dalam bentuk logit untuk memperoleh fungsi yang linear agar hubungan antara variabel independen dan variabel dependen dapat terlihat. Model transformasi logit dari $\pi(x)$ dari persamaan (4) dapat dituliskan sebagai berikut:

$$g(x) = In\left(\frac{P(X|Y=1)}{1-P(X|Y=1)}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 \ldots + \beta_pX_p$$

$g(x)$: transformasi logit dari $P(X|Y = 1)$

Model tersebut merupakan fungsi linier dari parameter-parameternya. Dalam model regresi linier, diasumsikan bahwa amatan dari variabel respon diekspresikan sebagai $y = E(Y|x) + \varepsilon$ terdapat pada persamaan

$$E(Y|x) = \beta_0 + \beta_1X_1 + \cdots + \beta_pX_p$$

merupakan rataan dari populasi dan ε merupakan komponen acak yang menunjukkan penyimpangan amatan dari rataannya dan ε diasumsikan mengikuti sebaran normal dengan rataan nol dan varians konstan.

Untuk mengembalikan nilai probabilitas yang dihasilkan kepada dua kategori variabel responnya, yaitu y = 0 atau y = 1 digunakan konversi sebagai berikut:

$$\hat{y}\ (x) = \{0, P(X|Y = 1) < 0.5$$
$$\hat{y}\ (x) = \{0, P(X|Y = 1)\ \geq 0.5$$

Proses pembuatan model regresi logistik juga menggunakan Google Colaboratory sebagai coding environment dengan format notebook serupa dengan Jupyter.

2.2.6 Confusion Matrix

Confusion matrix adalah suatu metode yang digunakan untuk melakukan perhitungan dan meggambarkan kinerja model klasifikasi (classifier) pada satu set data uji yang nilai sebenarnya diketahui. Confusion matrix dapat membantu dalam visualisasi kinerja suatu algoritma[5]. Seperti yang pada tabel 2.1 yang menampilkan hasil prediksi pada masalah klasifikasi, jumlah prediksi benar dan salah dirangkum dengan nilai-nilai dan dipecah kepada masing-masing label. Hal ini dapat membantu untuk mengetahui kesalahan yang dibuat oleh classifier.

Tabel 2. 1 Confusion Matrix

		Predict Label	
		Positive (P)	Negative (N)
Actual Label	P	True Positive (TP)	False Negative (FN)
	N	False Positive (FP)	True Negative (TN)

Dimana :

True Positive : merupakan data positif dan diprediksi benar

True Negative : merupakan data negatif dan diprediksi benar

False Positive : merupakan data negatif tetapi diprediksi sebagai data positif (error tipe 1)

False Negative : meupakan data positif tetapi diprediksi sebagai data negatif (error tipe 2)

2.2.7 Akurasi dan error

Akurasi merupakan jumlah prediksi yang benar dibagi dengan keseluruhan data prediksi. Akurasi dapat diperoleh meggunakan persamaan

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN}$$

Error merupakan jumlah seluruh data prediksi yang salah dan dibagi dengan keseluruhan data prediksi. Dinyatakan dalam pesamaan

$$\text{Error} = \frac{FP+FN}{TP+TN+FP+FN}$$

Atau persamaannya dapat dituliskan menjadi persamaan

$$\text{Error} = 1 - \text{Akurasi}$$

2.2.8 Presisi

Presisi adalah rasio jumlah prediksi yang benar-benar positif dari semua kelas positif yang diprediksi benar. Nilai presisi dapat dihitung dengan persamaan (2.12).

$$\text{Presisi} = \frac{TP}{TP+FP}$$

2.2.9 Recall

Recall atau Sensitivity menggambarkan seberapa banyak prediksi yang benar dari seluruh kelas positif. Nilai Recall dapat dihitung dengan menggunakan persamaan

$$\text{Recall} = \frac{TP}{TP+FN}$$

2.2.10 F-Score

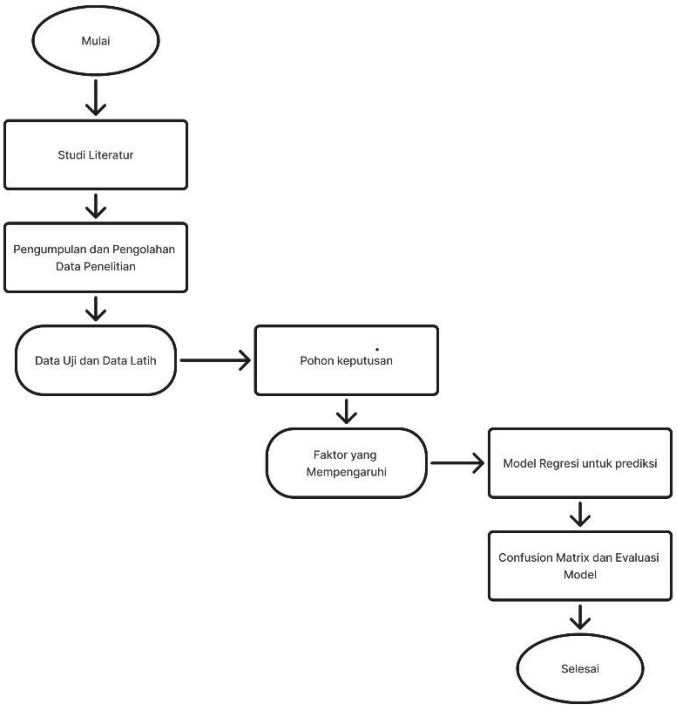
Pada pengukuran kinerja model, presisi dan recall merupakan pengukuran yang sama pentingnya. Presisi dan recall juga saling bertolak belakang. Recall bisa ditingkatkan semaksimal mungkin dengan cara memperbanyak prediksi sampel pada kelas positif yang mengakibatkan FP juga bertambah. Hal ini akan membuat presisi mejadi semakin berkurang karena tujuan utama dari evaluasi model menggunakan presisi adalah mengurangi jumlah FP. Berlaku juga sebaliknya untuk recall yang dipengaruhi oleh FN. Oleh sebab itu, dibutuhkan suatu cara untuk mendapatkan model yang seimbang. F-Score merupakan perbandingan rata-rata presisi dan recall yang dibobotkan [9]. F-Score dapat dihitung dengan menggunakan persamaan

$$F - Score = 2\ x \frac{Presisi\ x\ \ Recall}{Presisi\ x\ Recall}$$

BAB III
DESAIN DAN IMPLEMENTASI SISTEM

3.1 Alur Penelitian

Penelitian ini menggunakan metode Decision Tree dan Regresi Logistik untuk memprediksi peluang lolos. Tahapan penelitian dapat dilihat pada gambar berikut.



Gambar 3. 1 Alur Penelitian

3.1.1 Studi Literatur

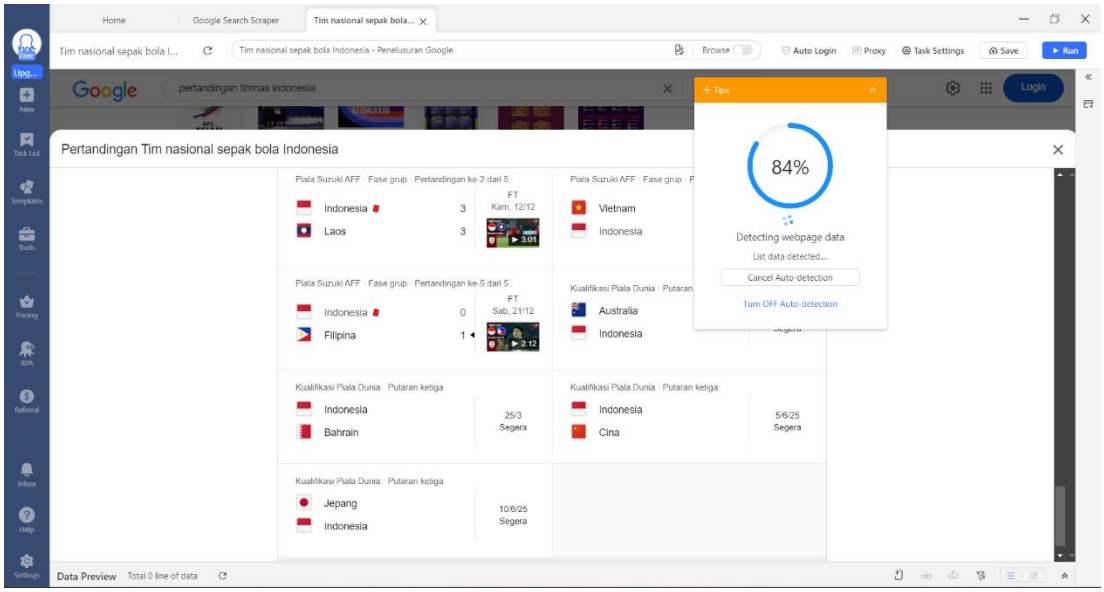
Studi literatur merupakan metode pengumpulan data dan penelitian yang dilakukan melalui kegiatan membaca, mencatat, serta menganalisis berbagai sumber literatur yang memiliki relevansi dengan topik penelitian. Metode ini juga sering disebut sebagai penelitian pustaka atau penelitian perpustakaan. Sumber literatur berupa Jurnal, buku, website yang berkaitan dengan topik pembahasan Timnas Indonesia, Kualifikasi Piala Dunia, Data Mining, Decision Tree, Algoritma C4.5, dan Regresi Logistik serta metode klasifikasi lainnya.

3.1.2 Pengumpulan dan Pengolahan Data Penelitian

Pada tahap ini, dilakukan proses pengumpulan dan pengolahan data. Tahapan ini cukup penting karena metode Data Mining memerlukan data untuk melakukan

analisis dan pembelajaran agar dapat menentukan suatu keputusan atau prediksi. Jenis pembelajaran yang dilakukan pada penelitian ini adalah prediksi (prediction), di mana variabel output yang coba diprediksi berupa kategori "Lolos" atau "Tidak Lolos". Pada dataset ini, terdapat fitur atau atribut seperti Peringkat FIFA, Jumlah Gol, Penguasaan Bola, dan lainnya yang dibutuhkan untuk proses pembelajaran.

Octoparse digunakan sebagai tools web scraping untuk memproses pengambilan data dari media berupa halaman FIFA, dan google search. Dalam penelitian ini data yang diambil berupa Ranking FIFA, klasemen, jumlah gol, tembakan, tembakan ke arah gawang, penguasaan bola, operan, akurasi operan, kartu kuning, kartu merah, offside, dan tendangan sudut. Langkah dalam pengambilan data dapat dilihat pada Gambar 2



Gambar 3. 2 Proses Pengumpulan Data pada Web Scraping

Setelah web scraping selesai mengumpulkan data dari sumber, langkah berikutnya adalah menyimpannya dalam format Excel agar lebih mudah dianalisis.

pertandingan indonesia

Tanggal	Lawan	Jenis Pertandingan	Hasil	Peringkat FIFA	Klasemen	Gol	tembakan	tembakan ke arah gawang	penguasaan bola	operan	akurasi operan	pelanggaran	kartu kuning	kartu merah	offside	tendangan sudut
12/10/2023	Brunei	round 1	Indonesia	127/184	1/2	6/0										
17/10/2023	Brunei	round 1	Indonesia	127/184	1/2	6/0										
16/11/2023	Irak	round 2	Irak	127/56	2/1	1/5										
21/11/2023	Filipina	round 2	draw	127/150	2/4	1/1										
21/03/2024	Vietnam	round 2	Indonesia	127/114	2/3	1/1										
26/03/2024	Vietnam	round 2	Indonesia	127/114	2/3	3/0										
06/06/2024	Irak	round 2	Irak	127/56	2/1	0/2	9/11	1/2	55%/45%	362/209	77%/78%	13/9	1/1	1/0	1/2	8/3
11/06/2024	Filipina	round 2	Indonesia	127/150	2/4	2/0	20/7	7/1	54%/46%	411/354	79%/76%	9/11	1/3	0/0	3/1	7/4
06/09/2024	Arab Saudi	round 3	Draw	127/59	3/4	1/1	8/18	2/4	34%/66%	313/593	73%/85%	15/6	3/2	0/0	2/4	1/6
10/09/2024	Australia	round 3	Draw	127/26	3/2	0/0	5/19	2/5	36%/64%	277/457	62%/79%	12/7	1/3	0/0	0/3	3/15
10/10/2024	Bahrain	round 3	draw	127/81	3/5	2/2	5/24	5/5	43%/57%	322/396	71%/78%	27/10	3/2	0/0	2/0	1/7
15/10/2024	Cina	round 3	Cina	127/90	3/6	1/2	14/5	6/3	76%/24%	604/163	83%/50%	8/11	2/1	0/0	1/0	6/1

Gambar 3. 3 Dataset

Setelah data diperoleh melalui web scraping dan dikonversi ke dalam format Excel, informasi tersebut disusun dalam tabel yang terstruktur dengan baik, seperti yang

ditunjukkan pada Gambar 5. Format Excel ini memudahkan peneliti untuk memfilter, mengurutkan, menganalisis data, dan mengaplikasikan metode data mining secara lebih efisien.

3.1.3 Pembuatan program

Program dibuat dengan memanfaatkan perangkat lunak *Spyder*, yang dipilih karena antarmukanya mudah digunakan, terutama bagi pemula, dan kompatibel dengan *virtual environment* Anaconda. Anaconda Environment digunakan untuk menginstal library yang diperlukan dalam penelitian ini, seperti *pandas* dan *scikit-learn*. *Pandas* adalah library Python yang dirancang untuk memanipulasi dan menganalisis data, sedangkan *scikit-learn* merupakan library yang di dalamnya terdapat berbagai macam algoritma.

3.1.4 Pembagian data latih dan data uji

Dataset yang telah diolah kemudian dipecah menjadi data latih dan data uji. Data latih adalah data yang akan digunakan untuk membangun model, sedangkan data uji digunakan untuk menguji seberapa baik sistem bekerja. Pada penelitian ini pembagian data latih dan data uji akan dilakukan dengan 2 macam skenario pengujian seperti pada table

Tabel 3. 1 Skenario Pengujian

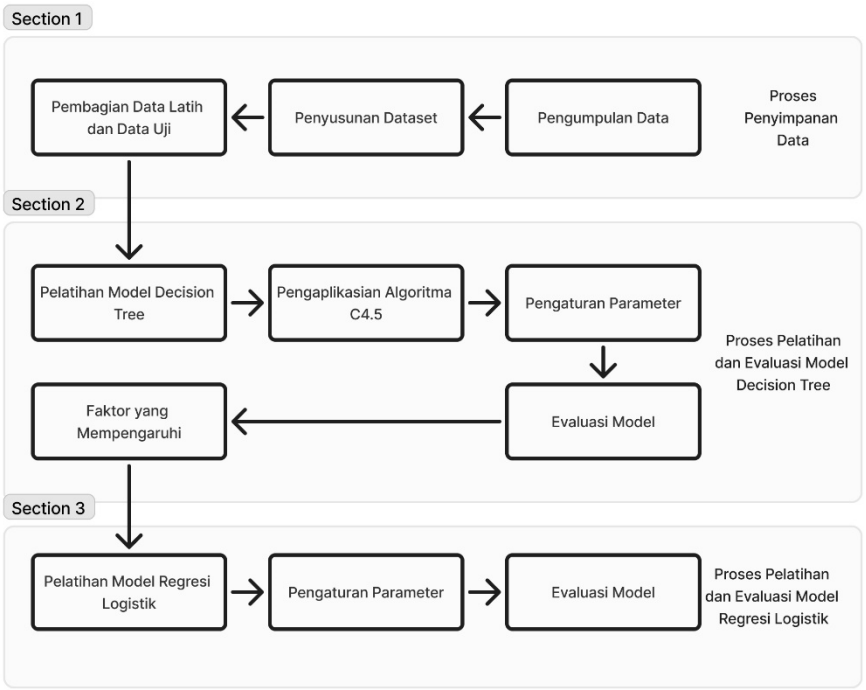
Pengujian	Data Latih	Data Uji
Skenario 1	75%	25%
Skenario 2	80%	20%

3.1.5 Analisis Hasil Kerja

Analisis kinerja dari simulasi yang dilakukan pada penelitian ini dibagi menjadi 2 bagian. Pertama menentukan faktor yang mempengaruhi lolos dengan menggunakan Decision Tree kemudian dianalisis perfoma. Kedua memprediksi peluang lolos dari hasil Decision Tree dengan menggunakan Regresi Logostik kemudian dianalisis performa. Parameter yang digunakan untuk menganalisisa performa kedua bagian tersebut adalah akurasi, presisi, recall, dan f-score.

3.2 Perancangan Sistem

Perancangan sistem dibutuhkan agar sistem berjalan dengan sistematis dan tidak terjadi eror pada saat dilakukan pengujian.



Gambar 3. 4 Diagram Sistem

3.2.1 Proses Penyimpanan Data

Tahap pertama pada sistem berisi proses penyiapan dataset sebelum kemudian masuk ke dalam algoritma untuk dilatih. Dimulai dengan menggumpulkan setiap atribut data dari sumber data yang telah dilampirkan sebelumnya untuk membuat suatu dataset yang terdiri dari 15 input dan 1 output. Output atau kelas dibagi menjadi 2 label kategori untuk mempermudah proses klasifikasi, kemudian masing-masing label tersebut diubah menjadi bentuk numerik untuk mempercepat proses komputasi dengan bantuan perangkat lunak Microsoft Excel. Pada proses terakhir. di fase ini, dataset dibagi menjadi 2 bagian, yaitu data latih dan data uji berdasarkan beberapa skenario pengujian.

3.2.2 Proses Pelatihan Model

Data latih yang telah dipisahkan akan dimanfaatkan untuk membangun model prediksi. Penelitian ini menggunakan algoritma *Decision Tree* dan *Regresi Logistik* sebagai metode Prediksi.

DAFTAR PUSTAKA

[1] S. Zein and G. Gunawan, "Prediksi Hasil FIFA World Cup Qatar 2022 Menggunakan Machine Learning dengan Python," *Jurnal Riset Matematika*, pp. 153–162, Dec. 2022, doi: 10.29313/jrm.v2i2.1382.

[2] I. Metode *et al.*, "Implementation of Naïve Bayes Classification Method To Predict La Liga Champion," *Jurnal Teknologi Sistem Informasi E-ISSN*, vol. 5, no. 2, pp. 128–139, 2024, doi: 10.35957/jtsi.v5i2.8028.

[3] D. Walangare, "Sistem Prediksi Pertandingan Sepak Bola Dengan Metode Analytical Hierarchy Process (AHP)."

[4] E. D. Anggara, A. Widjaja, and B. R. Suteja, "Prediksi Kinerja Pegawai sebagai Rekomendasi Kenaikan Golongan dengan Metode Decision Tree dan Regresi Logistik," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 1, Apr. 2022, doi: 10.28932/jutisi.v8i1.4479.

[5] R. Syahranita and S. Zaman, "Regresi Logistik Multinomial untuk Prediksi Kategori Kelulusan Mahasiswa," MEI, 2023.

[6] Y. Laia, C. Tandian, and A. Saputra, "PENERAPAN DATA MINING DALAM MEMPREDIKSI PEMENANG KLUB SEPAK BOLA PADA AJANG LIGA CHAMPION DENGAN ALGORITMA C.45," *Jurnal Sistem Informasi Ilmu Komputer Prima (JUSIKOM PRIMA*, vol. 2, no. 2, 2019.

[7] D. K. Utomo, A. A. Supianto, and W. Purnomo, "Sistem Prediksi Penerimaan SNMPTN menggunakan Algoritme Decision Tree C4.5," 2019. [Online]. Available: http://j-ptiik.ub.ac.id

[8] G. E. Putra and T. Fatimah, "3 rd Seminar Nasional Mahasiswa Fakultas Teknologi Informasi (SENAFTI) 30 Agustus 2023-Jakarta," 2023.

[9] (FIFA, 2024)

[10] (Wikipedia, 2023)

[11] (wikipedia, 2024)