BAB II
LANDASAN TEORI

2.1 Studi Literatur

Studi literatur digunakan untuk menghindari adanya pembuatan ulang, mengindentifikasi kesenjangan, megetahui penelitian terdahulu dengan medote yang serupa, melanjutkan penelitian yang telah ada serta menjadi referensi dalam penelitian yang sedang berlangsung. Hal tersebut dilakukan karena ada kemungkinan banyak penelitian yang telah menggunakan metode yang sama yakni *Decision Tree* dengan algoritma *C5.0* dalam penerapan *data mining*. Perbedaan penelitian ini dengan penelitian terdahulu adalah Peneliti menggunakan data prediksi pada data balita yang ada di Desa Kebun Dalam untuk mengetahui apakah balita tersebut mengalami gizi buruk atau tidak. Berikut daftar studi literatur yang digunakan dalam penelitian ini:

Tabel 2. 1 Daftar Literatur

Peneliti	Judul	Atribut/Fitur	Metode	Perbedaan Penelitian
(Rindya Ella Sari, 2021)	Penerapan Algoritma *C5.0* dalam Memprediksi Persediaan Buah pada UD.Bunda Syafira Buah	Nama buah, kualitas, harga	*C5.0*	Penelitian ini mengolah data dari data persediaan untuk melakukan pengendalian stok buah dengan tujuan agar buah yang di stok dapat terjaga kualitasnya dan tidak mengalami kerugian.

Tabel 2. 2 Daftar Literatur (Lanjutan)

Peniliti	Judul	Atribut/Fitur	Metode	Perbedaan Penelitian
(Apriani Candra Wijaya, 2018)	Implementasi Algoritma *C5.0* dalam Klasifikasi Pendapatan Masyarakat (Studi Kasus: Kelurahan Mesjid Kecamatan Medan Kota)	Umur, pendidikan, pekerjaan, pemilikan rumah, jumlah anggota keluarga, pendapatan	*C5.0*	Pada penelitian ini dilakukan pengambilan keputusan untuk penerimaan bantuan langsung tunai (BLT) dengan penggunaan implementasi dari sistem weka.
(Afnia Sartika Hutasoit, 2018)	Implementasi *Data Mining* Klasifikasi Status Gizi Balita Pada Posyandu Medan Timur Dengan Menggunakan Metode *C4.5*	Nama balita, jenis kelamin, nama orangtua, berat badan, umur	*C4.5*	Pada penelitian ini data yang digunakan adalah data balita yang akan diukur tingkat gizinya baik, buruk, atau berlebihan dengan menggunakan algoritma *C4.5*
(Natanael Benekditus, 2020)	Algoritma Klasifikasi *Decision Tree C5.0* untuk Memprediksi Performa Akademik Siswa	siswa mengangkat tangan, berpartisipasi dalam diskusi, insisiatif siswa untuk belajar di luar sekolah,	*C5.0*	Penelitian ini menggunakan aparameter yang diuji dari setiap atribut yang ada,

		dan absensi siswa		sehingga penilaian performa siswa dipantau dari berbagai atribut yang saling dikomparasi.
(Kastawan, 2018)	Implementasi Algoritma C5.0 pada Penilaian Kinerja Pegawai Negeri Sipil	Sasaran kerja pegawai, orientasi pelayanan, integritas, komitmen, disiplin, kerjasama, kepemimpinan	*C5.0*	Penelitian ini membahas tentang ilmu kemenajemenan dengan memanfaatkan kriteria penunjang keberhasilan dari profesionalitas.

2.2 *Data Mining*

Data Mining merupakan suatu teknik yang membahas mengenai penggalian atau pengumpulan data informasi yang dikumpulkan biasanya berupa pola-pola yang tersembunyi pada data, serta hubungan setiap elemen-elemen data, atapun model untuk keperluan data (Sigit & Yuita, 2018). Pada tahun 1990-an istilah *Data mining* mulai populer dikomunitas pengguna basisdata, akan tetapi sebenarnya perkembangan awal *data mining* berawal pada tahun 1763 ketika Thomas Bayes mempublikasikan Teorema Bayes.

Ada juga yang berpendapat bahwa, *data mining* merupakan suatu disiplin ilmu yang mempelajari metode untuk mengekstrak pengetahuan atau menemukan pola dari suatu data. Data mining sering juga disebut *Knowledge Discovery in Database* (KDD), yaitu kegiatan pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan di masa depan (Prasetyo, 2014).

2.2.1 Teknik-Teknik Data Mining

Data mining dapat dibagi menjadi beberapa tahapan proses. Teknik – teknik tersebut bersifat interaktif, pemakai terlibat langsung atau dengan perantara *knowlage base.* Teknik – teknik *data mining* adalah sebagai berikut:

1. *Predictive modelling* yang merupakan pengolahan *data mining* dengan melakukan prediksi/peramalan. Tujuan metode ini untuk membangun model prediksi suatu nilai yang mempunyai ciri-ciri tertentu.
2. *Association* (Asosiasi) merupakan teknik dalam *data mining* yang mempelajari hubungan antar data.
3. *Clustering* (Klastering) atau pengelompokkan merupakan teknik untuk mengelompokkan data ke dalam suatu kelompok tertentu.
4. *Classification* (Klasifikasi) merupakan teknik mengklasifikasikan data. Perbedaannya dengan metode *clustering* terletak pada data, dimana pada *clustering* variabel dependen tidak ada, sedangkan pada classification diharuskan ada variabel dependen. (Prasetyo, 2014)

2.2.2 Proses Data Mining

Secara sistematis ada tiga langkah utama dalam data mining:

1. Eksplorasi/pemrosesan awal data

   Eksplorasi/pemrosesan awal data terdiri dari *cleansing* data untuk menghindari redudansi data dan menghilangkan data terdapat *record* yang kosong, menormalkan data guna memberikan jangkauan antar data, transformasi data, penanganan data yang salah, reduksi dimensi, pemilihan subset fitur, dan sebagainya.

2. Membangun model dan melakukan validasi terhadapnya

   Memodelkan dan memvalidasi terhadap kinerja dan tingkat akurasi dari algoritma prediksi yang digunakan untuk mendapatkan hasil terbaik. Dalam langkah ini digunakan metode-metode seperti, klasifikasi, regresi, analisis *cluster*, deteksi anomali juga masuk dalam langkah eksplorasi.

3. Penerapan

   Penerapan berarti menerapkan model pada data yang baru untuk menghasilkan prediksi masalah yang akan terjadi pada kejadian dimasa mendatang (Prasetyo, 2014).

2.3 Prediksi

Prediksi atau peramalan merupakan suatu metode pengolahan dalam data mining untuk melakukan prediksi atau peramalan. Tujuan dari teknik ini berkaitan dengan pembuatan sebuah model yang dapat melakukan pemetaan dari setiap himpunan variabel ke setiap targetnya untuk memodelkan prediksi suatu nilai yang mempunyai ciri-ciri tertentu. Kemudian menggunkan model tersebut untuk memberikan nilai target pada himpunan baru yang didapat. Ada 2 jenis model prediksi, yaitu klasifikasi dan regresi. Klasifikasi digunakan untuk variabel target disket, sedangkan regresi digunakan untuk variabel target kontinu. Target yang

didapatkan dan tidak ada nilai seri waktu (*time series*) yang harus didapatkan untuk mendapat target nilai akhir. Sementara melakukan prediksi jumlah penjualan yang didapatkan 3 bulan kedepan harus didapatkan terlebih dahulu nilai penjualan bulan kedua dan pertama dan hal ini masuk kategori regresi. Dalam hal ini ada nilai seri waktu yang harus dihitung untuk sampai pada target akhir yang diinginkan dan ada nilai kontinu yang harus dihitung untuk mendapatkan nilai akhir yang diinginkan (Prasetyo, 2014).

2.4 Decision Tree

Decision tree atau pohon keputusan adalah pohon yang digunakan sebagai prosedur penalaran untuk mendapatkan jawaban dari masalah yang dimasukan. Pohon yang dibentuk tidak selalu berupa pohon biner. Jika semua fitur dalam set menggunakan 2 macam nilai kategorikal maka bentuk pohin yang didapatkan berupa pohon biner. Jika dalan fitur berisi lebih dari 2 macam nilai kategorikal atau menggunakan tipe numerik maka bentuk pohon yang didapatkan biasanya tidak berupa pohon biner.

Kefleksibelan membuat metode ini atraktif, khususnya karena memberikan keuntungan berups visualisasi saran (saran bentuk *decision tree*) yang membuat procedure prediksinya dapat diamati (Gorunescu, 2011). *Decision tree* banyak digunakan untuk menyelesaikan kasus penentuan keputusan seperti di bidang kedokteran (diagnosis penyakit pasien), ilmu komputer (struktur data), psikologi (teori pengambilan keputusan), dan sebagainya.

Data pada tabel 2.2 berisi contoh untuk melakukan prediksi “apakah harus bermain baseball?” dengan menjawab ya atau tidak dengan menggunakan empat

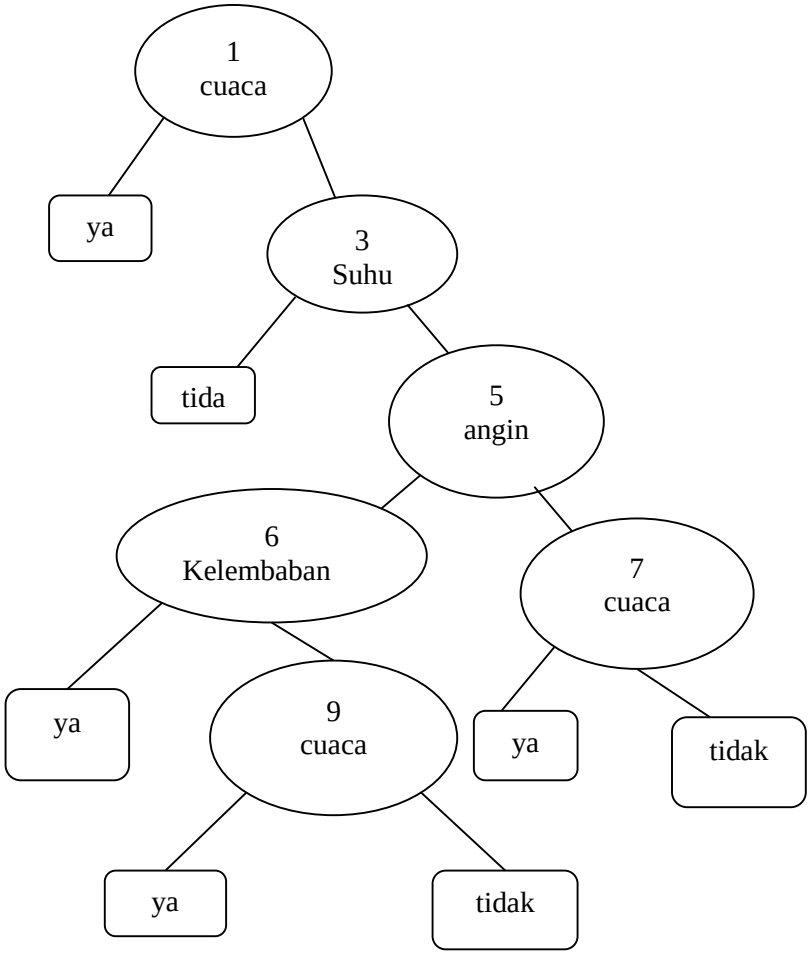
fitur diantaranya cuaca, suhu, kelembaban dan angin. Dengan rincian sebagai berikut (Prasetyo, 2014):

Tabel 2.3 Data Klasifikasi Bermain *Baseball*

Cuaca	Suhu	Kelembaban	Angin	Bermain
Cerah	Panas	Tinggi	Pelan	tidak
Cerah	Panas	Tinggi	Kencang	Tidak
Mendung	Panas	Tinggi	Pelan	Ya
Hujan	Lembut	Tinggi	Pelan	Ya
Hujan	Dingin	Normal	Pelan	Ya
Hujan	Dingin	Normal	Pelan	Tidak
Mendung	Dingin	Normal	Kencang	Ya
Cerah	Lembut	Tinggi	Pelan	Tidak
Hujan	Lembut	Normal	Pelan	Ya

Tabel 2.4 Data Klasifikasi Bermain *Baseball*

Cuaca	Suhu	Kelembaban	Angin	Bermain
Mendung	Lembut	Tinggi	Kencang	Ya
Mendung	Panas	Normal	Pelan	Ya
Hujan	Lembut	Tinggi	Kencang	Tidak



Gambar 2.1 Contoh *Decision Tree*
Sumber : Prasetyo (2014)

2.5 Algoritma C5.0

Algoritma C5.0 adalah pengembangan dari algoritma C4.5, yang mana memiliki kelebihan khususnya pada kumpulan data yang besar. Algoritma C5.0 lebih baik dari algoritma C4.5 pada efisiensi dan memori. Secara umum, alur proses pembuatan tree pada algoritma C5.0 dan algoritma C4.5 memiliki kemiripan, dimana kedua algoritma tersebut melakukan perhitungan entropy dan gain. Algoritma C4.5 akan berhenti hanya pada perhitungan gain, sedangkan algoritma C5.0 akan melanjutkannya dengan menghitung gain ratio berdasarkan nilai gain dan entropy. Ukuran gain ratio digunakan untuk memilih atribut uji pada setiap node di dalam tree. Atribut dengan nilai gain ratio tertinggi akan terpilih sebagai

parent bagi node selanjutnya. Untuk menghitung nilai entropy, digunakan Persamaan 2.1.

$$Entropy(S) = -\sum_{i=1}^{m} p_i log_2^{p_i} \quad \text{..... (2.1)}$$

dimana S set data yang terdiri atas n data sampel, p_i adalah proporsi yang bisa dihitung dengan $p_i = \frac{n_i}{|s|}$, n_i adalah jumlah data yang termasuk kelas ke-i dan $|s|$ adalah banyaknya data pada set S. Untuk menghitung *entropy* bersyarat atribut A digunakan Persamaan 2.2.

$$E(S|A) = -\sum_{j=1}^{v} p_i' \sum_{i=1}^{m} p_{ij} log_2(p_{ij}) \quad \text{..... (2.2)}$$

dimana p_j' adalah proporsi yang dapat dihitung dengan $p_j' = \frac{|s_j|}{s} = \frac{\sum_i^m n_{ij}}{n}$, p_{ij} adalah peluang bersyarat yang dapat dihitung dengan $p_{ij} = \frac{n_{ij}}{|s_j|}$ dan $|s_j|$ adalah jumlah data dengan atribut A. Maka, nilai *gain* dari atribut A dapat dihitung dengan Persamaan 2.3.

$$Gain\ (A) = E(A) - E(S|A) \quad \text{..... (2.3)}$$

Nilai *gain ratio* dari atribut A dihitung dengan Persamaan 4.

$$Gain\ Ratio\ (A) = \frac{Gain\ (A)}{Split\ (A)} \quad \text{..... (2.4)}$$

dimana,

$$Split(A) = -\sum_{j=1}^{v} p_j'\, log_2(p_j') \quad \text{..... (2.5)}$$

Algoritma C5.0 memecah data latih berdasarkan atribut yang memiliki nilai informasi *gain* terbesar. Prosedur *split* terus dilakukan hingga tidak ada lagi *subset* data yang dapat di-*split*. Untuk memperoleh hasil terbaik, dilakukan evaluasi terhadap model dengan menghitung akurasi yang akan menunjukkan tingkat

kebenaran pengklasifikasian data terhadap kelas yang sebenarnya. Semakin tinggi nilai akurasi berarti semakin rendah kesalahan prediksi terhadap data uji, sehingga merefleksikan model memiliki performa yang baik. Metode evaluasi yang digunakan dalam penelitian ini adalah K-fold cross validation yang membagi himpunan contoh secara acak menjadi *k* himpunan bagian (subset). Pada metode ini dilakukan pengulangan sebanyak *k* kali untuk data pelatihan dan pengujian. Pada setiap pengulangan, satu subset digunakan untuk pengujian sedangkan subset sisanya digunakan untuk pelatihan. Akurasi diperoleh berdasarkan data pengujian terhadap model klasifikasi menggunakan Persamaan 6 (Andi Nurkholis, 2020).

$$Accuracy(\%) = \frac{\sum Test\ data\ is\ correctly\ classified}{\sum Test\ data} \times 100 \quad \text{..... (2.6)}$$

2.6 *Python*

Python merupakan sebuah bahasa pemrograman yang cukup terkenal yang memiliki banyak manfaat untuk mendukung pemrograman yang berorientasi objek dan dapat berjalan diberbagai macam platform sistem operasi, seperti PCs, Macintosh, UNIX. Beberapa kelebihan dari bahasa pemrograman *python* diantara lain:

1. Pengembangan program dilakukan dengan cepat dan coding yang lebih sedikit
2. Mendukung multiplatform
3. Memiliki sistem pengelolaan memori yang otomatis
4. *Python* bersifat Object Oriented Programming (OOP)

2.7 *Confusion Matrix*

Confusion matrix adalah metode yang dapat melakukan perhitungan akurasi, presisi, dan recall. Akurasi merupakan hasil perhitungan dari semua nilai prediksi

yang benar dibagi dengan jumlah keseluruhan data. Nilai akurasi terbaik jika nilai akurasi tersebut sama dengan 100% dan yang terburuk 0% (Sigit, 2018).

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FN+FP} x100\% \quad (2.7)$$

Precision didapat dengan menghitung jumlah keseluruhan nilai prediksi positif yang benar dibagi dengan jumlah keseluruhan prediksi kelas yang benar. Nilai terbaik *precision* adalah 100% sementara yang terburuk 0% (Sigit, 2018).

$$Precision = \frac{TP}{TP+FP} x100\% \quad (2.8)$$

True positive rate atau biasa disebut dengan *recall* adalah jumlah prediksi benar dibagi dengan keseluruhan jumlah kelas yang salah. Untuk nilai terbaik *recall* adalah 100% sementara yang terburuk adalah 0% (Sigit, 2018).

$$Recall = \frac{TP}{TP+FN} x100\% \quad (2.9)$$

Secara definisi, *F1-Score* adalah *harmonic mean* dari *precision* dan *recall*, yang secara matematik dapat ditulis begini:

$$\mathbf{\frac{1}{f1}} = \frac{1}{2}\left(\frac{1}{precision} + \frac{1}{recall}\right) \quad (2.10)$$

Nilai terbaik *F1-Score* adalah 1.0 dan nilai terburuknya adalah 0. Secara representasi, jika *F1-Score* punya skor yang baik mengindikasikan bahwa model klasifikasi kita punya precision dan recall yang baik (Sigit, 2018).

2.8 Klasifikasi

1. Klasifikasi merupakan proses untuk menemukan fungsi dan model yang dapatmembedakan atau menjelaskan konsep atau kelas data dengan tujuan memperkirakan kelas yang tidak diketahui dari suatu objek. Dalam proses pengklasifikasian biasa terdapat dua proses yang harus dilakukan, yaitu:

2. Proses *Training*

   Pada proses ini akan digunakan data training set atau data sampel yang telah diketahui label – label atau atribut dari data sampel tersebut untuk membangun model dengan menggunkan 80% dari keseluruhan data set.

3. Proses *Testing*

Pada proses testing ini dilakukan untuk mengetahui keakuratan model yang telah dibuat pada proses *training* maka dibangun data yang disebut dengan data *testing* untuk mengklasifikasi label – labelnya. Klasifikasi merupakan penempatan objek – objek kesalah satu dari beberapa kategori yang telah ditetapkan sebelumnya. Klasifikasi sekarang ini telah banyak digunakan dalam berbagai aplikasi, sebagai contoh pendeteksian pesan email, spam berdasarkan *header* dan isi atau mengklasifikasikan galaksi berdasarkan bentuk – bentuknya. Pada proses klasifikasi data yang di-input-kan adalah data *record* atau data sampel. Pada setiap *record* dikenal sebagai *instance* atau atau contoh yang ditentukan oleh sebuah *tuple* (x,y). Dimana x adalah himpunan atribut dan y adalah atribut tertentu yang menyatakan sebagai label *class* (Nugroho & Subanar,2013).