BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Sebelumnya

Penelitian sebelumnya dikumpulkan sebanyak 10 jurnal diantaranya 7 (empat) jurnal nasional dan 3 (tiga) jurnal internasional dengan meninjau kembali jurnal-jurnal tersebut menggunakan analisa 3C2S. Jurnal yang dikumpulkan tidak hanya sama persis dengan pembahasan penelitian yang akan dilakukan, tetapi jurnal yang terdapat kesamaan bidang, metode, ataupun obyek penelitian satu sama lain juga akan direview dengan 3C2S. Detail review penelitian sebelumnya bisa dilihat pada Tabel 2.1:

Tabel 2. 1 Analisa 3C2S

No	Judul	Comparing	Contrasting	Criticize	Synthesize	Summarize
1	Prediksi Kelulusan Mahasiswa Stikom Tunas Bangsa Prodi Sistem Informasi Menggunakan AlgoritmaC4.5[10]	Mahasiswa Prodi SI STIKOM Tunas Bangsa akan selesai studinya setelah kelulusan.	Faktor mempengaruhi kelulusan mahasiswa dengan pengaruh Prodi SI STIKOM Tunas Bangsa.	Hanya dilakukan studi kasus Prodi SI STIKOM Tunas Bangsa dengan 2 atribut yaitu IPK dan absensi.	Penelitian ini menggunakan metode klasifikasi Algoritme C4.5 selaras dengan prediksi kelulusan mahasiswa.	Atribut yang berpengaruh yaitu IPK dan diperoleh tingkat akurasi sebesar 90.00%, *precision* sebesar 91,38% dan *recall* sebesar 98,15%.
2	Prediksi Ketepatan Waktu Kelulusan Mahasiswa dengan Metode Algoritma C5.0[11]	Studi untuk memperkirakan kapan mahasiswa akan lulus menggunakan teknik prediksi.Algoritme C5.0 serta pembuatan sistem berbasis web.	Membahas atribut yang mempengaruhi waktu kelulusan mahasiswa angkatan 2015.	Penelitian ini tidak mencantumkan sumber dataset yang digunakan sebagai atribut dalam perhitungan Algoritme C5.0, hanya mencantumkan tahun dataset mahasiswa angkatan 2015.	Penelitian ini menggunakan metode klasifikasi Algoritme C5.0 agar sesuai dengan tujuan penelitian yaitu prediksi waktu kelulusan mahasiswa.	Penelitian ini mengatakan bahwa dari 5 atribut yang digunakan, Beberapa atribut tidak dimasukkan sebagai bagian dari pohon keputusan dalam proses pembuatannya. Serta hasil akurasi dengan Algoritme C5.0 yaitu 83,78% dan nilai *error* sebesar 16,21%.

No	Judul	Comparing	Contrasting	Criticize	Synthesize	Summarize
3	Analisa Data Pertanian Tanaman Pangan Untuk Memprediksi Hasil Panen Dengan Data Mining Algoritma C.45 (Studi Kasus: Dinas Tanaman Pangan Dan Holtikutura Provinsi Sumut)[12]	Melakukan pnelitian untuk analisa data pangan dalam prediksi hasil panen menggunakan metode Algoritme C.4.5, sama seperti penelitian yang dilakukan.	Studi untuk mengidentifikasi variabel yang memiliki dampak paling besar terhadap hasil panen, sedangkan penelitian yang dilaukan adalah atribut yang mempengaruhi kelulusan mahasiswa.	Hanya dilakukan sampai dengan penerapan Algoritme C4.5 tidak sampai membuat suatu sistem yang dapat prediksi hasil panen padi seperti apa.	Penelitian ini menggunakan algoritma C4.5 dengan mencari tahu variabel apa saja yang mempengaruhi hasil panen, agar selaras dengan prediksi suatu hal yang belum diketahui.	Hasi uji Tanagra dengan penerapan Algoritme C4.5 menghasilkan pola kombinasi *itemsets* dan *rules* serta Algoritme C4.5 akan bermanfaat sekali untuk mendapatkan hasil panen yang maksimal.

No	Judul	Comparing	Contrasting	Criticize	Synthesize	Summarize
	Mahasiswa Tepat Waktu Menggunakan Metode Decision Tree Dan Artificial Neural Network[16]	meningkatkan kelulusan secara tepat waktu supaya memberikan saran pembuatan kebijakan di masa mendatang untuk perguruan tinggi.	mahsiswa secara tepat waktu membandingkan dua meotde dengan 3 atribut yaitu angkatan, jenis kelamin, dan keterangan lulus tepat waktu atau tidak, sedangkan penelitian yang akan dilakukan menggunakan 3 atribut berbeda.	kurang meyakinkan untuk prediksi kelulusan mahasiswa secara tepat waktu atua tidak, yaitu hanya angkatan, jenis kelamin, serta jumlah keterangan lulus tepat waktu.	berupa data label dan dilakukan pengujian dalam perbandingan 2 metode yaitu *Decision Tree* dan *Artificial Neural Network* dengan Rapid Miner.	hasil akurasi pengujian dengan metode *Artificial Neural Network* lebih tinggi dari hasil pengujian metode *decision tree* karena *neural network* memiliki *backward* yang mengembalikan jika terdapat error dengan akurasi *decision tree* sebesar 74,51% dan *artificial neural network* sebesar 79,74%
8	Analysis K-Nearest Neighbor Algorithm for Improving Prediction Student Graduation Time[17]	Melakukan penelitian untuk meningkatkan kelulusan secara optimal supaya bisa memperoleh kualitas akademik yang baik.	Membuat pengujian dengan aplikasi K-NN berdasarkan NPM, sedangkan studi yang akan dilakukan melakukan klasifikasi berdasarkan 3 atribut dalam prediksi waktu	Atribut yang digunakan tidak terlalu dipaparkan secara jelas dalam prediksi kelulusan mahasiswa dengan metode K-NN.	Metode k-Nearest Neighbor digunakan pada studi ini dengan melakukan pengujian pada aplikasi K-NN berbasis web.	Penelitian ini menunjukkan bahwa prediksi keluulusan mahasiswa bisa meningkat akurasinya dengan mengimplementasikan metode k-Nearest Neighbor.

No	Judul	Comparing	Contrasting	Criticize	Synthesize	Summarize
			kelulusan mahasiswa.			
9	Analysis and Design of Decision Support System Dashboard for Predicting Student Graduation Time[18]	Melakukan penelitian untuk meningkatkan prediksi kelulusan mahasiswa tepat waktu dengan adanya prototype aplikasi program studi Sisem Informasi.	Membuat pengujian dengan Algoritme C4.5 kemudian implementasi *protoype* aplikasi prediksi, sedangkan penelitian yang akan dilakukan melakukan klasifikasi menggunakan Algoritme C4.5 dengan 3 atribut.	Prototype dashboard yang dibangun masih kurang analisis dalam pengambilan keputusan sehingga baru menunjukkan informasi dan prediksi sementara.	Penelitian ini menggunakan metode Algoritme C4.5 dalam pembuatan dashboard.	Penelitian ini menunjukkan bahwa pembuatan prototype dashboard menggunakan Algoritme C4.5 bisa dapat diimplementasikan untuk meningkatkan prediksi kelulusan mahasiswa tepat waktu.
10	University Student Satisfaction Analysis on Academic Services by Using Decision Tree C4.5 Algorithm (Case Study:	Melakukan penelitian untuk mengukur kepuasan mahasiswa Universitas Putra Indonesia menggunakan Algoritme C4.5 supaya bisa	Membuat pohon keputusan dalam mengukur kepuasan mahasiswa terhadap akademik kampus.	Aturan yang didapat setelah perhitungan Algoritme C4.5 dalam mengukur kepuasan mahasiswa, tidak dipaparkan secara detail.	Penelitian ini menggunakan metode Algoritme C4.5 dalam menganalisis kepuasan mahasiswa terhadap pelayanan akademik.	Penelitian menunjukkan bahwa Algoritme C4.5 merupakan metode yang sangat baik dalam ketepatan dalam menganalisis kepuasan mahasiswa terhadap pelayanan akademik dengan akurasi 95%

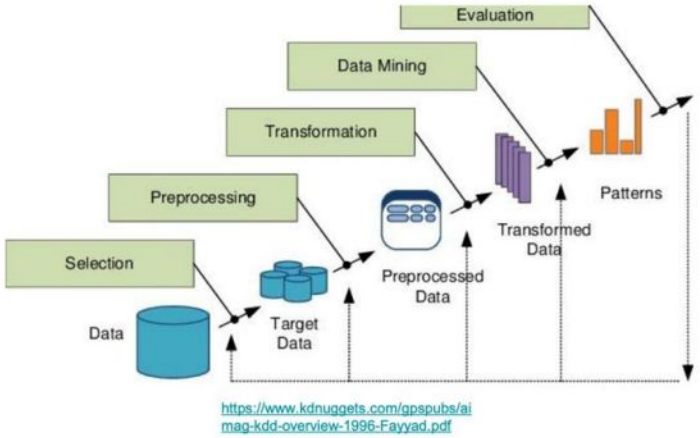
No	Judul	Comparing	Contrasting	Criticize	Synthesize	Summarize
	Universitas Putra Indonesia "YPTK" Padang)[19]	meningkatkan kebijakan Akademik				dan *precision* sebesar 97,53% d serta *recall* sebesar 96,34%.

Penelitian yang akan dilakukan berjudul "**Prediksi Waktu Kelulusan Mahasiswa menggunakan Algoritme C4.5"** yang memiliki perbedaan dengan penelitian sebelumnya. Kelebihan Algoritme C4.5 diantaranya yaitu mampu menangani atribut dengan tipe data *diskret* dan kontinu, mampu menangani atribut yang kosong, serta mampu memangkas pohon keputusan (menghilangkan cabang pada pohon keputusan). Langkah menggunakan Algoritme C4.5 yaitu dataset penelitian, menghitung nilai *entropy*, menghitung nilai *gain*, memilih atribut sebagai akar, membuat cabang untuk masing masing nilai, mengulangi proses sehingga tiap atribut memiliki kelas yang sama. Pembuatan pohon keputusan menghasilkan suatu klasifikasi dari atribut yang digunakan dan diperoleh pada pengujian data menggunakan sistem library *Scikit Learn* dengan *Jupyter notebook*.

2.2 Dasar Teori

2.2.1 *Knowledge Discovery in Database* (KDD)

Keseluruhan proses identifikasi pola, pengetahuan, dan informasi dari sekumpulan *big data* merupkan definisi dari *Knowledge Discovery in Database* (KDD) [16]. Istilah KDD atau mencari pengetahuan dalam sebuah basis data, menekankan pada penerapan metode penambangan data tertentu yang meliputi proses pencarian pengetahuan dalam data yang luas. *Knowledge Discovery in Database* (KDD) merupakan metode untuk memperoleh pengetahuan dari database yang berisi tabel – tabel saling berhubungan [20]. Hasil *Knowledge Discovery in Database* (KDD) bisa digunakan untuk pengambilan keputusan. Gambar proses KDD bisa dilihat pada Gambar 2.1 sebagai berikut:



Gambar 2. 1 Proses *Knowledge Discovery in Database* (KDD)[8]

Proses *Knowledge Discovery in Database* (KDD) terdiri dari enam proses yaitu[21]:

a. Pemilihan data (*Data Selection*)

Data dipilih dari dataset keseluruhan kemudian menjadi data latih yang digunakan untuk melewati proses *data mining* Algoritme C4.5.

b. Pembersihan data (*Preprocessing/Cleansing*)

Pembersihan data dilakukan dengan pembuangan data yang *double, inkonsisten* data diperiksa, kesalahan diperbaiki.

c. *Transformation*

Jenis atau pola informasi yang akan dicari dalam basis data menjadi pengaruh dalam proses tranformasi pada data yang dipilih. Transformasi diperlukan karena adanya kebutuhan inputan data yang tidak sama dengan data yang kita dapatkan. Data yang sudah melewati proses KDD sebelumnya, diubah terlebih dahulu sesuai algoritme atau teknik dalam penambangan data.

d. Penambangan data (*data mining*)

Tahapan utama dalam KDD yaitu *data mining* yang data yang terpilih melewati proses pencarian pola atau informasi dengan Algoritme atau teknik *data mining* tertentu.

e. Evaluasi (*Evaluation*)

Evaluasi dilakukan dengan memeriksa hasil prediksi akan sesuai atau bertentangan dengan hasil actual yang ada.

f. Presentasi/visualisasi

Grafik, pohon keputusan, merupakan bentuk visualisasi dalam penyajian informasi mengenai *rule* yang terbentuk setelah melewati proses diatas.

2.2.2 *Data Mining*

Data Mining merupakan sebuah teknik mengekrasi data yang besar untuk mencari sebuah informasi yang bermanfaat bagi pengguna. *Data mining* berguna dalam banyak bidang ilmu untuk mencari sebuah pengetahuan dan informasi didalam sebuah data yang ukurannya besar[22]. Fungsi penambangan data berdasarkan kategori yaitu prediksi (*predictive*) dan deksripsi (*descriptive*). Fungsi prediksi yaitu penambangan data dapat memprediksi sesuatu yang sudah terjadi dan sesuatu yang akan terjadi. Model dari fungsi *predictive* yaitu *Classification*, *Forecasting*, dan

Regression. Fungsi Deskripsi yaitu penambangan data dapat menjelaskan dan menggambarkan sesuatu yang sedang terjadi saat ini. Selain fungsi, penambangan data juga memiliki peran utama, yaitu sebanyak 5 peran utama *Data mining*, sebagai berikut [21].

a. Estimasi (*Estimation)*

Estimasi yaitu menerka sebauh nilai yang belum diketaui dan *label* atau *class* yang digunakan bertipe numerik. Seperti menerka penghasilan disaat informasi mengenai orang tersebut belum diketahui. Algoritme estimasi mirip dengan algoritme klasifikasi, tetapi *target variable* berupa bilangan numerik (kontinyu) dan bukan kategorikal (nominal dan diskrit). Algoritme estimasi yang digunakan yaitu *Linear Regression, Neural Network, Support Vector Machine*.

b. Prediksi (*Prediction)*

Prediksi digunakan untuk menentukan faktor yang akan mempengaruhi suatu kondisi mendatang dan memprediksi dampak dari perubahan saat ini. Seperti contoh seorang investor menggunakan prediksi untuk menentukan adakah faktor yang akan mempengaruhi dari naik turunnya saham perusahaan. Contoh prediksi lainnya yaitu ahli statistik memprediksi dampak signifikan dari perubahan operasi bisnis. Prediksi memiliki *label* atau *class* bertipe numerik dengan data yang digunakan merupakan data *time series* (data rentet waktu). Istilah prediksi kadang digunakan juga untuk klasifikasi, tidak hanya untuk prediksi *time series*, karena sifatnya yang bisa menghasilkan *class* berdasarkan berbagai atribut yang disediakan.

c. Klasifikasi (*Classification*)

Klasifikasi merupakan fungsi atau model untuk yang dapat membedakan kelas data, agar dapat diperkirakan kelas dari suatu objek ketika labelnya tidak diketahui. Data dengan

label atau *class* berupa nilai kategorikal (nominal). Algoritme klasifkasi diantaranya yaitu ID3, C4.5, *Artificial Neural Network* (ANM), *Naïve Bayes*, *Genetic Algorithm*, Fuzzy, *Case-Based Reasoning* (CBR), *k-Nearest Neighbor*(k-NN). Contoh klasifikasi, apabila *label* atau *class* adalah pendapatan, maka bisa digunakan nilai nominal (kategorikal) yaitu pendapatan besar, menengah, dan kecil.

d. Pengklusteran (*Clustering*)

Clustering merupakan fungsi atau model untuk memperoleh pola yang labelnya tidak diketahui (*unsupervised learning*). Pengelompokkan data, hasil observasi dan kasus kedalam *label* atau *class* yang mirip. Mengidentifikasi kelompok dari barang atau produk yang mempunyai karakteristik khusus, berbeda dengan klasifikasi, dimana *clustering* tidak terdapat definisi-definisi karakteristik yang diberikan pada waktu *classification.* Algoritme yang digunakan untuk memperoleh pola dan trend *clustering* diantaranya yaitu *K-Means*, *K-Medoid*, *Fuzzy C-Means*, *Subtractive*, *Mountain*, *Hierarki*.

e. Asosiasi (*Association*)

Mengidentifikasi hubungan antara kejadian-kejadian yang terjadi pada suatu waktu, atau analisis keranjang pasar untuk mengidentifikasi item-item produk yang kemungkinan dibeli konsumen bersamaan dengan produk lain. Setiap item dipertimbangkan sebagai informasi dengan tujuan agar dapat mencari produk yang biasanya terjual bersamaan dan juga agar dapat mencari tahu apa saja aturan yang menyebabkan kesamaan itu. Algoritme asosiasi diantaranya yaitu *Apriori algorithm, FP-Growth Algorithm, GRI Algorithm.*

Manfaat *data mining* dilihat dari dua sudut pandang, yaitu sudut pandang komersial dan sudut pandang keilmuan. Manfaat

data mining dari sudut pandang komersial untuk penanganan melonjaknya volume data sehingga penggunakan *data mining* dapat memberikan informasi yang diperlukan. Contoh dari sudut pandang komersial yaitu cara identifikasi produk yang terjual bersama dengan produk lain, cara prediksi tingkat penjualan, cara prediksi perilaku bisnis di masa depan. Manfaat *data mining* dari sudut pandang keilmuan untuk menganalisa serta menyimpan data yang bersifat *real time* dan sangat besar. Contohnya pemindaian langit dengan *telescope,* penempatan pada suatu satelit untuk *remote sensor*.

Transformasi data merupakan sebuah skala data dalam bentuk lain sehingga data memiliki distribusi yang diharapkan. Jenis transformasi data memiliki 7 jenis, yaitu Kuadrat, kubik, akar, logaritma, invers, arcsin, dan invers skor. Secara umum untuk mentransformasi data tergantung jenis kasus dan algoritma yang dipakai.

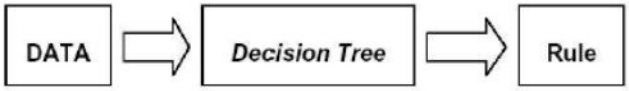
Teknik pembelajaran *data mining* terdiri dari dua macam yaitu *Supervised Learning* dan *Unsupervised Learning*. Teknik *supervised learning* yaitu data yang memiliki label sehingga sudah diketahui dari awal hasilnya akan dipetakan kemana. Klasifikasi adalah salah satu contoh dari *supervised learning* karena hasil akhir pada klasifikasi yaitu tertuju pada satu kelas, seperti apakah mahasiswa itu lulus tepat waktu atau tidak. Proses klasifikasi terbagi menjadi dua, yaitu *data training* dan *data testing*. Algoritme Klasifikasi yaitu pengolahan data latih untuk menciptakan aturan (*rules)*. *Rules* merupakan hasil dari model klasifikasi yang akan digunakan untuk prediksi pada data baru sebagai data uji [2].

Teknik *unsupervised learning* merupakan kebalikan dari teknik *supervised learning* dimana teknik ini dianggap datanya tidak memiliki label. *Unsupervised learning* tergantung algoritme

yang mendeteksi semua pola seperti *association* dan *sequence* yang muncul dari kriteria penting yang spesifik dalam data yang dimasukkan. Klasterisasi adalah salah satu contoh dari *unsupervised learning* karena kategori ini mengelompokkan data yang memiliki kemiripan yang identik untuk menjadi suatu kelas yang sama.

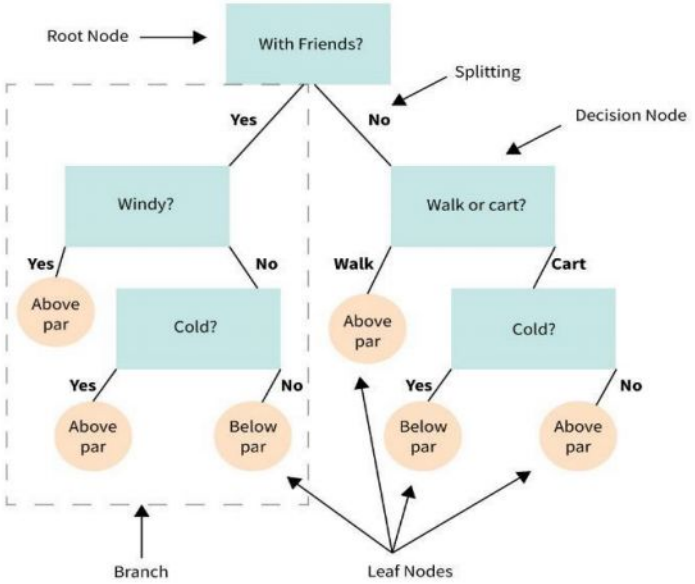
2.2.3 *Decision Tree*

Pohon keputusan (*Decision Tree*) adalah mengubah bentuk data yang awalnya dalam bentuk tabel yaitu atribut dan *record* menjadi bentuk pohon (*tree*) sehingga pohon tersebut bisa merepresentasikan aturan (*rule*). Pohon (*tree*) terdiri dari simpul (*node*) dan rusuk (*edge*). Misalkan untuk mdafenentukan kelulusan mahasiswa secara tepat waktu, kriteria yang diperhatikan seperti SKS, IPS, dimana atribut DECISION sebagai atribut target. Konsep pohon keputusan dalam dilihat pada Gambar 2.2:



Gambar 2. 2 Konsep Pohon Keputusan[8]

Simpul adalah struktur yang berisikan data sedangkan rusuk adalah penghubung antar simpul. Algoritme yang digunakan untuk *decision tree* diantaranya ID3, CART, C4.5. Alur pada *decision tree* ditelusuri dari simpul akar (*root*) ke simpul daun yang memegang prediksi kelas. Parameter yang dinyatakan kemudian dibuat sebagai kriteria dalam pembentukan *tree.* Atribut memiliki nilai yang menjadi target atribut yang disebut *instance*. Contoh *instance* yaitu seperti atribut predikat kelulusan memiliki *instance* berupa sangat memuaskan dan pujian. Contoh pohon keputusan bisa dilihat pada Gambar 2.3 berikut:



Gambar 2. 2 Contoh Decision Tree[8]

Keunggulan *decision tree* yaitu mudah diinterpretasikan, tingkat akurasi yang dapat diterima, penanganan atribut bertipe *diskret* secara efisien, serta bisa menangani atribut bertipe *diskret* dan numerik [17].

2.2.4 Algoritme C4.5

Algoritme yang populer digunakan untuk membangun *DecisionTree* yang mudah mnengerti, Algoritme C4.5 merupakan algoritme yang populer digunakan untuk membangun sebuah pohon keputusan yang mudah dimengerti, fleksibel, dan menarik karena dapat divisualisasikan dalam bentuk gambar [23]. Algoritme C4.5 yang digunakan untuk proses klasifikasi merupakan pengembangan dari ID3.

Langkah-langkah dalam membangun pohon keputusan menggunakan Algoritme C4.5 adalah sebagai berikut [18]:

1. Siapkan dataset pelatihan. Dataset pelatihan biasanya diperoleh dari riwayat data yang sudah ada sebelumnya dan telah dikelompokkan ke dalam kelas-kelas tertentu.

2. Menghitung nilai *entropy*

   Entropy merupakan nilai informasi yang menyatakan ukuran ketidakpastian. Nilai *entropy* setiap atribut yaitu menggunakan persamaan sebagai berikut:

   $$Entropy(S) = -\sum_{i=1}^{n} pi * log_2\ pi \qquad (2.1)$$

   Keterangan:

   S = jumlah atribut

   S_i = jumlah keseluruhan

   N = jumlah partisi

   P_i = proporsi dari S_i terhadap S

3. Menghitung nilai *gain*

   Gain merupakan ukuran efektifitas suatu variabel dalam mengklasifikasikan data. Nilai *Gain* dari suatu variabel merupakan selisih antara nilai *entropy* total dengan *entropy* dari variabel tersebut. Atibut yang diperoleh dengan rumus 2.1 akan menghasil *gain* tertinggi untuk dijadikan akar pohon keptuusan.

   $$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} \times Entropy(S_i) \qquad (2.2)$$

   Keterangan:

   S = Himpunan Kasus

   A = Atribut

   n = jumlah partisi atribut a

   $|S_i|$ = jumlah kasus pada partisi ke- i

   |S| = jumlah kasus dalam S

4. Menentukan atribut sebagai akar (*root*) dari pohon keputusan yang didasarkan pada nilai *gain* tertinggi dari atribut yang ada.

5. Ulangi proses kedua untuk setiap cabang sampai semua kasus di cabang memiliki kelas yang sama.

2.2.5 *Scikit Learn*

Jupyter Notebook merupakan pengembangan ilmu data yang berbeda dari proyek Python lainnya yaitu adanya penggunaan IPPython Jupyternotebooks. *Jupyter Notebook* menyediakan sarana untuk membuat dan berbagi dokumen interaktif dengan potongan kode langsung yang dapat dieksekusi dan plot serta rendering persamaan matematika melalui sistem penyusunan huruf Latext. Keuntungan *Jupyter Notebook* salah satunya yaitu dapat menjalankan perintah secara langsung di dalam *prompt Anaconda* dengan menyertakan awalan tanda seru (!). Fitur tebaik *Jupyter Notebook* adalah kemampuan untuk membuat laporan langsung yang berisi kode sehingga dapat dieksekusi.

Paket dan modul Python sebagai berikut[24]:

a. *NumPy* adalah salah satu komponen inti komputasi ilmiah dengan Python.
b. *SciPy* adalah paket komputasi ilmiah inti
c. *pandas* adalah perpustakaan berkinerja tinggi untuk memuat, membersihkan, menganalisis, dan memanipulasi struktur data.
d. *matplotlib* adalah pustaka Python dasar untuk membuat grafik dan plot kumpulan data dan juga merupakan paket dasar dari pustaka plotting Python lainnya.
e. *Scikit-learn* adalah perpustakaan pembelajaran mesin Python yang menyediakan sejumlah teknik penambangan, pemodelan, dan analisis data dalam API sederhana.

Scikit-learn atau dikenal *sklearn* adalah *machine learning library software* gratis untuk bahasa pemrograman Python. Fitur berbagai algoritma klasifikasi, refresi, dan pengelompokkan termasuk mesin vector dukungan, hutan acak, peningkatan gradien, k-means dan DBSCAN dan dirancang untuk beroperasi dengan perpustakaan numerik dan ilmiah Python NumPy dan SciPy [25]. Dasar dari *machine learning Scikit Learn* diantaranya yaitu *Load*

Data, Splitting Data (train dan test), *Modelling(*Decision Tree Algoritem C4.5), Prediksi dan Akurasi, Visualiasasi.

2.2.6 *K-Fold Validation*

K-Fold Validation adalah metode tambahan dari teknik data mining yang bertujuan unutk memperoleh hasil akurasi yang maksimal dengan mengevaluasi kinerja Algoritme. Data dipisahkan mejadi dua subset yaitu data *training* dan data *testing*. Algoritme dilatih oleh data *training* dan divalidasi oleh data *testing*. Pemilihan jenis CV dapat didasarkan pada ukuran dataset.

2.2.7 *Confusion Matrix*

Confusion Matrix adalah sebuah metode yang kegunaannya untuk melakukan perhitungan akurasi pada konsep dalam penambangan data. Evaluasi dengan menggunakan metode *confusion matrix* menghasilkan nilai akurasi, presisi, dan *recall*. Pengukuran akurasi dilakukan dengan metode pengujian *confusion matrix* dapat dilihat pada Tabel 2.2:

Tabel 2. 2 *Confusion Matrix*

Actual	Prediksi	
	Negatif (LTTW)	Positif (LTW)
Negatif (LTTW)	False Negative (FN)	False Positive (FP)
Positif (LTW)	True Positive (TP)	True Negative (TN)

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \times 100\% \quad (2.1)$$

$$Precision = \frac{TP}{TP+FP} \times 100\% \quad (2.2)$$

$$Recall = \frac{TP}{TP+FN} \times 100\% \quad (2.3)$$

Akurasi (*Accuracy*)dalam klasifikasi penambangan data merupakan persentase ketepatan pada *record data* yang telah diklasifikasikan secara benar dan dilakukan pengujian pada hasil klasifikasi [2]. Presisi (*precision*)merupakan proporsi pada kasus yang diprediksi positif, dimana data yang sebenarnya juga positif.

Recall atau *sensitivity* adalah proporsi kasus positif yang diprediksi dengan benar [26].