# BAB II
# LANDASAN TEORI

## 2.1. Tinjauan Pustaka

Pada tahapan ini Peneliti akan melakukan tinjauan pustka yang telah dilakukan dalam penerapan *data mining* menggunakan algoritma XGboost. Sehingga, dalam penelitian ini diperlukan tinjauan pustaka sebagai alat dalam penerapan algoritma ini, agar dapat menghindari pembuatan ulang, mengindentifikasi kesenjangan, mengetahui algoritma yang sudah diterapkan, mengetahui penelitian yang sama dibidang ini, serta melanjutkan untuk penelitian sebelumnya. Perbedaan dalam penelitian ini dengan penelitian terdahulu adalah peneliti menggunakan data siswa sekolah menengah pertama untuk memprediksi calon penerima KIP .

**Tabel 2. 1** Daftar Literatur

| No | Nama Peneliti | Tahun | Judul |
|---|---|---|---|
| 1 | (Karo, 2020) | 2020 | Implementasi Algoritma XGboost dan Feature Importance untuk Klasifikasi pada Kebakaran Hutan dan Lahan |
| 2 | (Ngakan Nyoman Pandika Pinata, 2020) | 2020 | Prediksi Kecelakaan Lalu Lintas di Bali dengan XGboost pada Python |

**Tabel 2. 2** Daftar Literatur (Lanjutan)

| No | Nama Peneliti | Tahun | Judul |
|---|---|---|---|
| 3 | (Rachmi, 2020) | 2020 | Implementasi Algoritma Random Forest Dan XGboost pada Klasifikasi *Customer Churn* |
| 4 | (Rahman, 2020) | 2020 | Implementasi Algoritma Svm, MlLP dan XGboost pada Data Ekspresi Gen |
| 5 | (Mardiansyah, 2019) | 2019 | Penanganan Masalah Data Kredit untuk Kelas Tidak Seimbang Menggunakan Smote XGboost |

### 2.1.1. Literatur 1

Kebakaran hutan dan lahan di Indonesia telah menjadi masalah krisis lingkungan tahunan. Sebaran kebakaran hutan terbesar terjadi dipulau Sumatera. Salah satu upaya tindakan dalam pencegahan dan meminimalisasikan resiko kebakaran hutan dengan cara mengklasifikasikan jenis titik panas di lahan, sehingga di dapat skala prioritas dalam pemadaman titik api. Penelitian ini bertujuan mengklasifikasikan type titik panas dengan algoritma XGboost dan feature importance yang terdapat di pulau Sumatera. Data titik panas diperoleh dari Globalforestwatch.com. Proses mengurangi variabel dari data yang diperoleh menghasil dampak yang sangat signifikan pada model klasifikasi. Terapat enam dan atau tujuh variabel yang sangat berpengaruh dalam menentukan titik panas,

variabel tersebut jugalah yang menghasilkan model klasfikasi terbaik. XGboost dan feature importance menghasilkan akurasi sebesar 89.52%. Sensitivity (SE), Specificity (SP), dan Matthews Correlation Coefficient (MCC) secara berturut turut 91.32 %, 93.16 % dan 92.75 %. Algoritma ini juga lebih baik dibandingkan dengan hasil penelitian sebelumnya.

### 2.1.2. Literatur 2

Tingginya pertumbuhan penduduk di Indonesia menyebabkan kepemilikan kendaraan bermotor pribadi semakin tinggi yang mempengaruhi meningkatnya kemacetan dan juga kecelakaan lalu lintas. Peramalan angka kecelelakan lalu lintas dilakukan pada penelitian ini sebagai salah satu upaya yang dapat dimanfaatkan sebagai dasar tindakan antisipasi terkait peningkatan angka kecelelakaan lalu lintas. Penelitian ini bertujuan untuk meramalkan kecelakaan lalu lintas menurut akibatnya menggunakan Xtreme Gradient Boosting (XGboost) dengan bahasa pemrograman Python. Data yang digunakan dalam penelitian ini yaitu data dari Badan Pusat Statistik Provinsi Bali dengan periode dari Tahun 1996 sampai dengan Tahun 2019 dalam selang waktu tahunan. Hasil peramalan diukur menggunakan RMSE (Root Mean Square Error). Penerapan XGboost untuk meramalkan data kecelakaan lalu lintas menurut akibatnya, menunjukkan model XGboost memiliki performa yang sangat baik pada dua kategori yaitu kategori jumlah orang meninggal akibat kecelakaan dengan nilai RMSE 4,92 dan jumlah orang yang mengalami luka berat dengan nilai RMSE 4,11. Nilai RMSE model XGboost untuk kategori jumlah kejadian kecelakaan lalu lintas yaitu sebesar 21,69 dan kategori orang yang mengalami luka ringan akibat kecelakaan yaitu sebesar 77,24.

### 2.1.3. Literatur 3

Teknologi komunikasi yang terus berkembang pesat mengakibatkan masyarakat konsumtif akan informasi dan komunikasi. Hal ini dimanfaatkan oleh penyedia jasa layanan telekomunikasi yang terus berinovasi untuk mempertahankan customer. Karena terbukanya persaingan antara penyedia jasa layanan telekomunikasi dapat mengakibatkan customer churn. Prediksi churn dapat dilakukan untuk mengidentifikasi customer churn sejak awal dan membantu sektor CRM (Customer Relationship Management) agar dapat mempertahankan customer, sehingga mengurangi potensi kerugian yang dialami perusahaan. Penelitian ini bertujuan untuk memprediksi customer churn dengan data data telecommunications churn pada District of Columbia menggunakan algoritma klasifikasi Random Forest dan Extreme Gradient Boosting, kedua algoritma ini merupakan bagian dari algoritma ensemble. Hasil analisis klasifikasi bahwa algoritma Extreme Gradient Boosting (XGboost) lebih unggul dibandingkan algoritma Random Forest dilihat dari tingkat akurasi dan nilai AUC. Algoritma Extreme Gradient Boosting (XGboost) mendapatkan nilai akurasi dan AUC sebesar 95.6% dan 0.876, sedangkan algoritma Random Forest mampu menghasilkan nilai akurasi dan AUC sebesar 93.5% dan 0.799.

### 2.1.4. Literatur 4

Diabetes Melitus (DM) adalah penyakit yang berlangsung lama atau kronis yang ditandai dengan naiknya kadar gula (glucose) pada darah yang tinggi. Diabetes melitus diklasifikasikan menjadi diabetes melitus tipe 1 (DM tipe 1), diabetes melitus tipe 2 (DMT2), dan diabetes melitus yang terjadi pada saat

kehamilan. Menurut organisasi kesehatan dunia World Health Organization (WHO), pada tahun 2000 terdapat 171 juta orang menderita diabetes melitus dan diprediksi akan mengalami peningkatan 2 kali lipat menjadi 366 juta jiwa pada tahun 2030. Berdasarkan data dari IDF 95% dari total penderita diabetes melitus merupakan penderita diabetes melitus tipe 2 (DMT2). Diabetes melitus tipe 2 merupakan penyakit metabolik dengan karakteristik hiperglikemia yang terjadi karena kelainan sekresi insulin, kerja insulin atau kedua-duanya. Berbeda dengan diabetes melitus tipe 1 yang disebabkan karena kerusakan pankreas sehingga tidak dapat cukup memproduksi insulin, penderita diabetes melitus tipe 2 biasanya disebabkan karena pola makan yang tidak sehat dan jarang berolahraga, sedangkan diabetes melitus tipe 1 biasanya menyerang anak-anak dikarenakan kelainan genetik sejak lahir. Seiring majunya perkembangan tekhnologi, kini telah berkembang suatu bidang ilmu baru yaitu bioinformatika. Salah satu implementasi dari bioinformatika adalah digunakannya algoritma-algoritma komputasi, matematika, dan statistika dalam membantu menyelesaikan permasalahan-permasalahan biologi melalu analisis data ekspresi gen.

Pada penelitian ini dilakukan analisis klasifikasi pada data microarray hasil dari ekspresi gen pada sampel diabetes melitus tipe 2, Impaired Glucose Tolerance (IGT), dan sampel dengan kadar gula darah normal (NGT) dengan kode series GSE18732. Pada penelitian ini digunakan algoritma klasifikasi Support Vector Machine (SVM), Arsitektur Multilayer Perceptron (MLP), dan Xtreme Gradient Boosting (XGboost) untuk menganalisis data tersebut. Hasil dari analisis yang dilakukan, didapatkan bahwa algoritma klasifikasi Support Vector Machine dengan

menggunakan kernel Linear mampu mendapatkan nilai akurasi terbesar yaitu sebesar 91,30%. Sedangkan arsitektur Multilayer Perceptron dengan satu hidden layer dan 100 hidden node mampu mendapatkan nilai akurasi sebesar 78,26% dan algoritma klasifikasi terakhir Xtreme Gradient Boosting mampu mendapatkan nilai akurasi sebesar 71,39%.

### 2.1.5. Literatur 5

Beberapa peneliti banyak menemukan data dengan kondisi kelas tidak seimbang, dimana terdapat data dengan jumlah minoritas dan mayoritas. SMOTE merupakan satu pendekatan data untuk kelas tidak seimbang dan XGboost merupakan salah satu algoritma untuk permasalahan data tidak seimbang. Penelitian ini menggunakan SMOTE dan XGboost atau disingkat dengan SMOTEXGboost untuk penanganan data dengan kelas tidak seimbang. Hasil penelitian memperlihatkan nilai AUCSMOTEXGboost lebih baik dari model Logistic Regression, Random Forest, Support Vector Machine, dan XGboost.

### 2.1. Contoh Data Training

Berikut ini adalah contoh data yang akan memprediksi siswa akan menerima bantuan KIP atau tidak yang ditujukan pada tabel 2.3.

**Tabel 2. 3** Data Siswa Prediksi Bantuan KIP

| UA | PA | HA | UB | PB | HB | JTK | Nilai | LABEL |
|---|---|---|---|---|---|---|---|---|
| Dewasa | 0,2 | 0,5 | Dewasa | 1 | 0 | Cukup | Baik | Tidak Dapat |
| Lansia | 0,5 | 0,5 | Lansia | 1 | 0 | Banyak | Sangat Baik | Dapat |
| Lansia | 0,5 | 0,8 | Dewasa | 1 | 0 | Cukup | Cukup | Tidak Dapat |
| Lansia | 0,5 | 0,5 | Dewasa | 1 | 0 | Sedikit | Cukup | Tidak Dapat |
| Lansia | 0,5 | 0,8 | Dewasa | 1 | 0 | Sedikit | Baik | Tidak Dapat |
| Lansia | 0,5 | 0,5 | Lansia | 1 | 0 | Banyak | Cukup | Dapat |

Keterangan:

UA, UB : Usia Ayah, Usia Ibu

PA, PB : Pendidikan Ayah, Pendidikan Ibu

HA, HB : Penghasilan Ayah, Penghasilan Ibu

JTK : Jumlah Tanggungan Keluarga

Nilai Siswa : Nilai Rata-rata Siswa

## 2.2. *Data Mining*

Menurut Prasetyo, Eko (2014:1) mengartikan data mining sebagai berikut: "Data mining adalah campuran dari statistik, kecerdasan buatan, dan riset basis data yang masih berkembang".

Proses data mining secara sistematis, ada tiga langkah utama yaitu:

1. Eksplorasi/pemrosesan awal data

   Eksplorasi/pemrosesan awal data terdiri dari 'pembersihan' data, normalisasi data, transformasi data, penanganan data yang salah, reduksi dimensi, pemilihan subset fitur, dan sebagainya.

2. Membangun model dan melakukan validasi terhadapnya

   Membangun model dan melakukan validari terhadapnya berarti melakukan analisis berbagai model dengan kinerja prediksi yang 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 perkiraan/prediksi masalah yang diinvestigasi (Prasetyo, 2014).

### 2.3.1. Tahapan-Tahapan Data Mining

Sebagai suatu rangkaian proses, *Data Mining* dapat dibagi menjadi beberapa tahapan proses. Tahapan-tahapan tersebut bersifat interaktif, pemakai terlibat langsung atau dengan perantara *knowlage base.* Tahapan-tahapan *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* 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.

### 2.2.1. 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 algoritma-algoritma 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. Klasifikasi**

Klasifikasi merupakan proses untuk menemukan fungsi dan model yang dapat membedakan 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:

1. 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.

2. 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).

**2.5. XGboost**

XGboost merupakan salah satu algoritma boosting yaitu kumpulan decission tree yang pembangungan pohon berikutnya akan bergantung pada pohon sebelumnya. Pohon pertama dalam XGboost akan lemah dalam melakukan klasifikasi dengan inisialisasi probability yang ditentukan oleh peneliti dan kemudian akan dilakukan update bobot pada setiap pohon yang dibangun sehingga menghasilkan kumpulan pohon klasifikasi yang kuat. Prediksi dilakukan dengan menjumlahkan seluruh bobot yang ada di setiap pohon dan kemudian memasukkan nilai tesebut ke fungsi logistik. XGboost akan meminimumkan fungsi objektif sebagai berikut:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_{i,}\hat{y_i}^{t-1} + f_t(x_i)\right) + \Omega(f_t) \quad \text{.....(2.1)}$$

Dimana $\sum_{i=1}^{n} l\left(y_{i,}\hat{y}_{i}^{t-1}\right)$ adalah *differentiable loss function* untuk mengukur perbedaan antara prediksi $\hat{y}_i$ dan nilai aktual $y_i$, sedangkan $f_t(xi)$ adalah model baru yang dibangun dan $\Omega(f_t)$ merupakan kompleksitas model pada fungsi regresi contoh ke-$i$ pada iterasi ke-$i$. *function* pada klasifikasi kelas respon biner bisa menggunakan log loss. Persamaan Omega merupakan parameter regularisasi yang akan membuat model berusaha menghindari overfitting. Nilai gain bisa ditentukan untuk penentuan splitting node. Berikut ini adalah rumus untuk mencari nilai gain sebagai ukuran sebuah atribut yang berpengaruh terhadap hasil pada algoritma XGboost: $(x + a)^n =$

$$\mathcal{L}split = \frac{1}{2}\left[\frac{(\sum_{i\in I_L} g_i)^2}{\sum_{i\in I_L} hi + \lambda} + \frac{(\sum_{i\in I_R} g_i)^2}{\sum_{i\in I_R} hi + \lambda} + \frac{(\sum_{i\in I_I} g_i)^2}{\sum_{i\in I_I} hi + \lambda}\right] - \gamma \quad \text{.....(2.2)}$$

Untuk menekan pertumbuhan pohon dan mencegah model overfitting, ditambahkan ambang pemisah $\gamma$ yang merupakan koefisien bernilai 0 dan λ bernilai 1. Nilai $gi$ dan $hi$ merupakan turunan pertama dan kedua *loss function* pada XGboost (Chen, 2016).

### 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 algoritma yang dapat menghitung akurasi, presisi, dan recall. Akurasi adalah hasil dari membagi semua nilai yang diprediksi dengan benar dengan jumlah total data. Nilai akurasi tertinggi adalah 100 dan nilai akurasi terendah adalah 0%. (Sigit, 2018).

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FN+FP} x100\% \quad (2.3)$$

*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).

$$Precision = \frac{TP}{TP+FP} x100\% \quad (2.4)$$

Sedangkan *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).

$$Recall = \frac{TP}{TP+FN} x100\% \quad (2.5)$$

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

$$\frac{\mathbf{1}}{\boldsymbol{f1}} = \frac{1}{2}\left(\frac{1}{precision} + \frac{1}{recall}\right) \quad (2.6)$$

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).