BAB III
METODOLOGI PENELITIAN

3.1 Objek Penelitian

Objek penelitian pada penelitian ini adalah data dari mahasiswa Strata 1 pada Universitas Multimedia Nusantara. Secara spesifik, data yang digunakan dimulai dari data mahasiswa angkatan 2017 sampai 2023. Variabel yang tersedia pada data ini antara lain adalah; program studi, angkatan, jenis kelamin, status beasiswa, dan status mahasiswa (status *churn*).

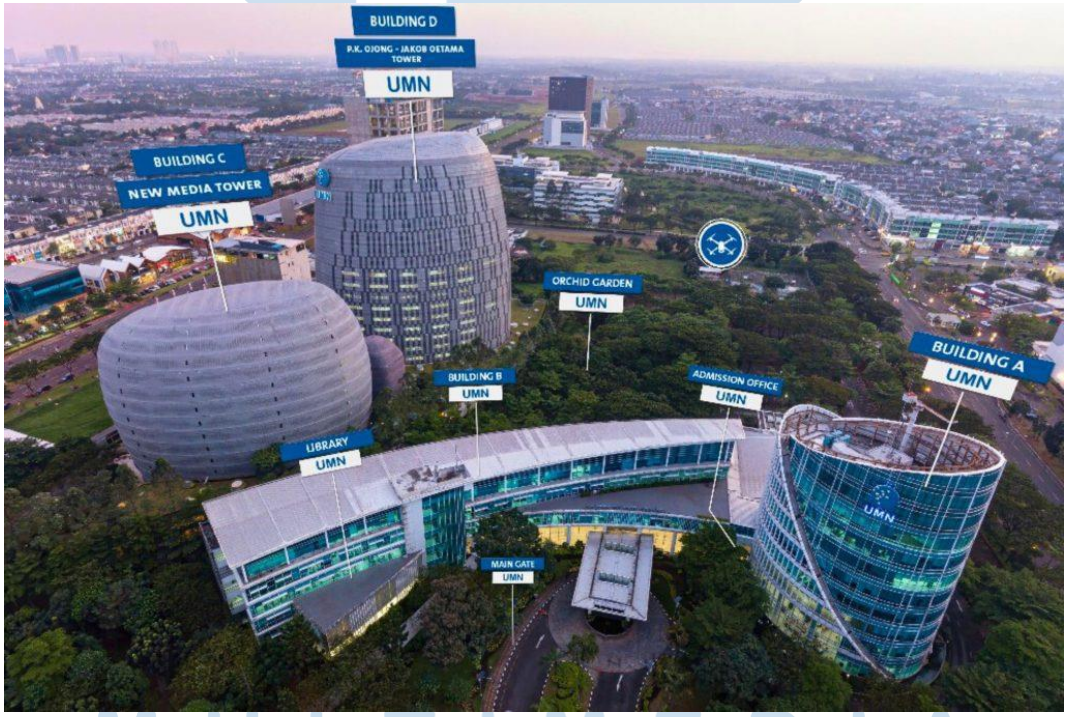
3.1.1. Profil Perusahaan

Universitas Multimedia Nusantara merupakan sebuah universitas swasta yang didirikan oleh Dr. Jakob Oetama, pendiri dari Kompas Gramedia. UMN hadir di Indonesia pada 25 November 2005, namun baru diumumkan secara resmi oleh Sekretaris Jenderal Kementerian Pendidikan Nasional pada 20 November 2006 di Hotel Santika. Lokasi dari UMN terletak pada Gading Serpong, Tangerang, Banten, Indonesia. Saat ini, UMN telah memiliki 4 gedung dengan gedung D sebagai gedung terbaru. Gedung D atau biasa disebut sebagai PK Ojong – Jakob Oetama Tower meraih penghargaan sebagai 1st *Runner Up Energy Efficient Building* pada ASEAN *Energy Award* di Bangkok, Thailand pada tahun 2019. Saat ini, UMN memiliki 4 fakultas dan 16 jurusan, serta telah terakreditasi A. Berikut merupakan fakultas dan jurusan yang ada di UMN:

1. Fakultas Teknik & Informatika
    a. Informatika (S1)
    b. Teknik Komputer (S1)
    c. Teknik Elektro
    d. Teknik Fisika
    e. Sistem Informasi

f. Gelar Bersama – Program Informatika

2. Fakultas Bisnis
    a. Perhotelan
    b. Akuntansi
    c. Manajemen
    d. Magister Manajemen Teknologi
3. Fakultas Ilmu Komunikasi
    a. Komunikasi Strategis
    b. Jurnalistik
    c. Magister Ilmu Komunikasi
4. Fakultas Seni & Desain
    a. Desain Komunikasi Visual
    b. Arsitektur
    c. Film & Animasi



Gambar 3. 1 Universitas Multimedia Nusantara

Sumber: [33]

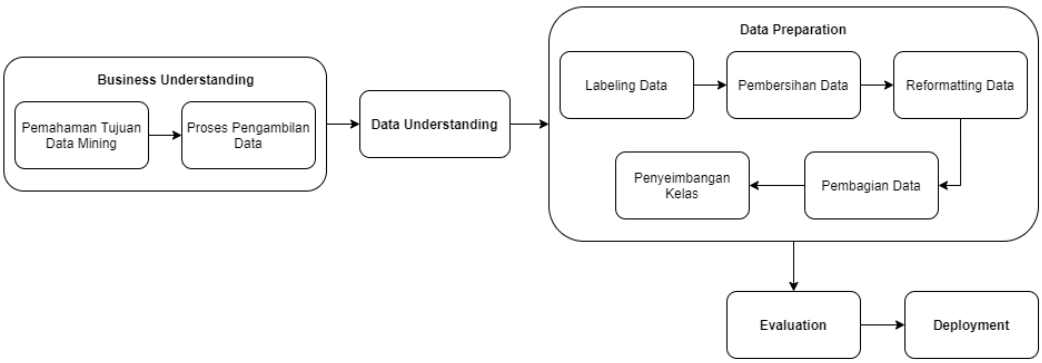
Gambar 3.1 di atas merupakan tampilan komplek Universitas Multimedia Nusantara yang berlokasi di Gading Serpong, Tangerang. Dapat dilihat bahwa terdapat 4 gedung dalam komplek tersebut. Universitas Multimedia Nusantara tidak hanya menyediakan pendidikan berkualitas tinggi, namun juga keindahan dari arsitektur yang unik seperti yang terlihat pada Gedung New Media Tower dan Gedung P.K. Ojong-Jakoeb Oetama Tower [33] yang terlihat pada bagian belakang kiri dan belakang kanan pada gambar 3.1.

3.2 Metode Penelitian

Pada penelitian ini, mahasiswa akan diklasifikasikan ke dalam 2 label berbeda, yaitu *churn* dan *not churn*. Oleh karena itu, pembuatan model klasifikasi merupakan salah satu hal yang perlu dilakukan. Model untuk klasifikasi *churn* dan *not churn* pada mahasiswa akan dilakukan dengan menggunakan 3 algoritma berbeda untuk menghasilkan 3 hasil yang berbeda. Algoritma tersebut antara lain adalah Decision Tree, Random Forest, dan XGBoost. Adapun *framework* yang akan digunakan pada penelitian ini adalah salah satu *framework* data *mining*, yaitu CRISP-DM.

3.2.1. CRISP-DM

CRISP-DM atau *Cross-Industry Standard Process for Data Mining* merupakan sebuah model proses independen industri yang digunakan untuk melaksanakan proyek data *mining*. Saat ini, CRISP-DM juga merupakan standar dilakukannya data *mining* yang paling populer. Visualisasi untuk alur kerja CRISP-DM dapat dilihat pada gambar 2.1. CRISP-DM memiliki 6 tahapan di dalamnya, yaitu *business understanding*, data *understanding*, data *preparation*, *modeling*, *evaluation*, dan *deployment* [19].



Gambar 3. 2 Alur Penelitian

Gambar 3.2 di atas merupakan alur dari penelitian yang dilakukan. Penelitian akan dimulai dari tahap *business understanding* yang berisi pemahaman tujuan data *mining* dan proses pengambilan data, lalu dilanjutkan dengan data *understanding*, yaitu pemahaman dari data yang telah diperoleh. Selanjutnya, proses data *preparation* akan dilewati. Data *preparation* dimulai dari proses *labeling* data, pembersihan data, *reformatting* data, pembagian data, dan diakhiri dengan proses penyeimbangan kelas. Proses akan dilanjutkan ke tahap *evaluation* dan pada akhirnya tahap *deployment*.

3.2.1.1. Business Understanding

a. Pemahaman Tujuan Data Mining

Tujuan dilakukannya data *mining* terhadap suatu bisnis penting untuk diketahui sebelum masuk ke dalam tahapan berikutnya. Dengan mengetahui tujuan dilakukannya data *mining*, hasil yang diperoleh akan lebih maksimal dan sesuai karena setiap proses dipastikan mengacu pada tujuan tersebut. Data *mining* yang dilakukan terhadap mahasiswa Strata 1 Universitas Multimedia Nusantara memiliki tujuan untuk mengetahui *customer churn* yang ada pada Universitas Multimedia Nusantara, faktor apa yang mempengaruhi terjadinya *customer churn*, dan pada

akhirnya berupaya untuk menyediakan informasi yang dapat digunakan dalam pengurangan terhadap angka *churn* yang ada pada Universitas Multimedia Nusantara. Tiga algoritma *machine learning* akan dimanfaatkan dalam membuat model klasifikasi, yaitu Decision Tree, Random Forest, dan XGBoost. Keseluruhan proses hingga tahap *deployment* akan dilakukan dengan menggunakan bahasa pemrograman Python pada *platform* Visual Studio Code.

b. Proses Pengambilan Data

Dalam upaya melakukan data *mining*, tentunya diperlukan data. Pada penelitian ini data yang diperlukan adalah data mahasiswa Strata 1 Universitas Multimedia Nusantara. Data diperoleh dari Biro Informasi Akademik (BIA) Universitas Multimedia Nusantara (UMN) melalui surat daring dan format *file* Excel.

3.2.1.2. Data Understanding

Tahap selanjutnya adalah memahami data yang akan digunakan dalam proses data *mining*. Dalam penelitian ini, data yang akan digunakan merupakan data mahasiswa Universitas Multimedia Nusantara Strata 1 mulai dari angkatan 2017 sampai dengan 2023. Pada data ini, terdapat beberapa *field* seperti semester masuk, fakultas, program studi, jenis kelamin, jalur masuk, IPS per semester, IPK per semester, SKS per semester, total SKS, status mahasiswa, dan semester lulus.

3.2.1.3. Data Preparation

a. *Labeling* Data

Proses preparasi data dimulai dengan membagi data ke dalam 2 label/kelas yang berbeda. Pada penelitian ini, label

akan dibagi menjadi *churn* dan *not churn*. Label diberikan agar algoritma *machine learning* dapat mempelajari pola yang ada untuk tiap label.

b. Pembersihan Data

Proses selanjutnya adalah pembersihan data. Data harus dibersihkan terlebih dahulu dari *null value* sehingga kualitas dan kestabilan data menjadi lebih baik. Data yang sudah bersih selanjutnya akan melalui proses *reformatting*.

c. Reformatting Data

Reformatting data dilakukan dalam penelitian ini karena format awal data yang kurang baik dan sulit untuk dibaca serta dimengerti. Format awal data untuk kolom performa akademik seperti IPS dan IPK yang tadinya ditampilkan per periode diubah menjadi per semester. Hal ini membuat data menjadi lebih mudah untuk dibaca dan dimengerti. Tidak hanya itu, *reformatting* data juga dapat menghasilkan wawasan yang lebih baik.

d. Pembagian Data

Pembagian data atau disebut juga sebagai *splitting* perlu untuk dilakukan dalam penggunaan algoritma *machine learning*. Pada penelitian ini, data dibagi ke dalam 3 *set*, *training*, *validation*, dan *testing*. Hal ini dilakukan karena adanya keterbatasan data untuk tahap *deployment*. Rasio pembagian data adalah 70% *training* dan 30% *testing* [34]. Dari 70% data *training*, akan dibagi lagi menjadi 70% dari 70% untuk data *training* dan 30% dari 70% untuk data *validation*. Data *training* berfungsi untuk melatih algoritma dalam mengklasifikasikan data ke dalam label. Data *validation* berfungsi untuk mengetahui performa tiap

algoritma dalam mengklasifikasikan data. Data testing berperan sebagai *unseen* data yang akan digunakan pada tahap *deployment*.

e. Penyeimbangan Kelas

Dalam permasalahan klasifikasi, seringkali terjadi *class imbalance* dimana anggota suatu kelas jauh lebih banyak dibandingkan kelas lainnya. Hal ini dapat diatasi dengan menggunakan teknik *rebalancing*, yaitu *oversampling*. Metode *oversampling* yang digunakan adalah SMOTE. *Oversampling* akan diterapkan terhadap *training* data, sehingga dalam proses melatih algoritma, data yang diterima oleh algoritma imbang untuk tiap kelasnya. Teknik *rebalancing* dilakukan agar performa model tidak bias terhadap kelas mayoritas saja.

3.2.1.4. Modeling

a. Feature Importance

Feature importance merupakan proses pencarian fitur-fitur yang dianggap penting dalam mengklasifikasikan data ke dalam suatu kelas oleh tiap algoritma. Fitur yang dianggap penting oleh tiap algoritma mungkin saja berbeda. Setelah menemukan fitur-fitur yang dianggap penting, model akan dilatih ulang hanya dengan menggunakan fitur-fitur tersebut. Batasan yang digunakan dalam pemilihan fitur adalah memiliki skor *feature importance* yang lebih besar dari nol. Jumlah fitur yang dipilih dapat mempengaruhi performa model [35].

b. Retraining

Berdasarkan hasil *feature importance*, algoritma akan dilatih ulang. Model yang dibuat berdasarkan hasil dari *feature importance* tentunya memiliki tingkat relevansi yang lebih baik karena sudah dipastikan bahwa tiap fitur dengan skor lebih besar dari nol berkontribusi terhadap proses klasifikasi. Dengan demikian, diharapkan bahwa model dapat memiliki performa yang maksimal.

3.2.1.5. Evaluation

Setelah melakukan pemodelan dengan menggunakan algoritma *machine learning*, hasil klasifikasi mahasiswa ke dalam label *churn* dan *not churn* akan dievaluasi dengan menggunakan data validation. Evaluasi klasifikasi dapat dilakukan dengan melihat nilai akurasi, presisi, *f-measure/f1-score* dan *recall* dari 3 model yang telah dibuat dengan 3 algoritma berbeda. Akurasi digunakan karena dapat secara sederhana menunjukkan performa dari model dalam mengklasifikasikan data. Presisi dan *recall* adalah 2 metrik lain yang perlu digunakan selain akurasi untuk melihat performa model secara lebih detail, sebab terkadang nilai akurasi yang tinggi bisa menyesatkan. Presisi menunjukkan jumlah dari prediksi positif yang memang sesungguhnya adalah positif dan *recall* menunjukkan jumlah dari data positif yang diprediksi positif. *F1-measure* merupakan nilai rata-rata harmonik antara presisi dan *recall*. Dengan melakukan evaluasi, algoritma yang terbaik dalam mengklasifikasikan mahasiswa ke dalam label *churn* dan *not churn* dapat diketahui.

3.2.1.6. Deployment

Deployment merupakan tahapan dilakukannya penerapan model (*deploy*) yang telah dibuat pada tahap sebelumnya ke dalam data baru/*unseen data*. Data baru tersebut nantinya akan di *labeling*

ke dalam kelas *churn* atau *not churn*. Pada tahap *deployment*, Streamlit akan digunakan untuk membuat *web application* yang mampu memprediksi *customer churn* pada Universitas Multimedia Nusantara. Untuk melakukan prediksi pada *web application* tersebut, *user* harus melakukan *input* data berupa *file* Excel. *Output* yang dihasilkan akan berupa *file* Excel yang dapat diunduh dan telah ditambahkan kolom hasil prediksi.

3.3 Teknik Pengumpulan Data

Data dalam penilitian ini merupakan data sekunder yang diperoleh dari pihak Biro Informasi Akademik (BIA) Universitas Multimedia Nusantara. Data yang diperoleh merupakan data mahasiswa Strata 1 dari angkatan 2017 – 2023. Total banyaknya data adalah sebanyak 16,178 *records*.

3.4 Teknik Pengambilan Sampel

Populasi dari data adalah keseluruhan mahasiswa angkatan 2017 – 2023, yaitu sebanyak 16,178. Untuk sampel yang digunakan adalah keseluruhan dari populasi. Dari seluruh data ini, akan dilakukan pembagian dengan perbandingan sebesar 70:30, yaitu data *training* sebesar 70% dan data *testing* sebesar 30%.

3.5 Teknik Analisis Data

Analisis data dimulai dengan mengambil data mahasiswa Strata 1 Universitas Multimedia Nusantara melalui Biro Informasi Akademik (BIA). Data yang diperoleh merupakan data mahasiswa mulai dari angkatan 2017 sampai dengan 2023. Total *records* dalam data adalah sebanyak 16,178. *Field* yang ada di dalam data antara lain adalah semester masuk, fakultas, program studi, jenis kelamin, jalur masuk, IPS per semester, IPK per semester, SKS per semester, total SKS, status mahasiswa, dan semester lulus. Data yang ada kemudian akan masuk ke dalam tahap data *preprocessing*, yaitu tahap dilakukannya pembersihan data untuk menjaga kestabilan dan kualitas data. Pembagian data ke dalam data *training* dan data *testing* kemudian akan dilakukan. Tahap pemodelan data dengan

menggunakan algoritma *machine learning* khusus untuk teknik klasifikasi akan menjadi tahap berikutnya. Algoritma yang digunakan ada 3, yaitu Decision Tree, Random Forest, dan XGBoost. Tahap berikutnya yang akan dilakukan adalah mengklasifikasikan mahasiswa ke dalam label *churn* atau *not churn* dengan menggunakan model yang telah dibuat dengan 3 algoritma berbeda. Evaluasi kemudian akan dilakukan terhadap model yang telah dibuat dengan melihat tingkat akurasinya dalam mengklasifikasikan mahasiswa ke label *churn* atau *not churn*.

Tools yang digunakan dalam penelitian antara lain adalah Microsoft Excel sebagai tempat penyimpanan data, Visual Studio Code sebagai *code editor platform*, dan Python sebagai bahasa pemrograman yang digunakan dalam melakukan data *mining*.

Tabel 3. 1 Kelebihan dan Kekurangan *Tools*

Tools	**Kelebihan**	**Kekurangan**
Microsoft Excel	Merupakan *software* penyimpan data paling populer. Tidak hanya menyimpan data, Excel juga dapat melakukan komputasi statistikal serta membuat chart dan grafik [30].	Terletak pada segi keamanan data yang kurang baik serta ketidakmampuannya dalam menangani *dataset* besar [30].
Visual Studio Code	*Code editor* yang sangat simple namun memiliki banyak *tools* serta mendukung berbagai bahasa pemrograman. Salah satu fitur yaitu IntelliSense membantu *developer* dalam *code completion* dan *debugging*. VS Code juga menyediakan Extension Marketplace bagi para *developer* untuk menambahkan bahasa pemrograman dan *tools* [36].	Dengan adanya *extension* yang sangat banyak pada VS Code, developer kemungkinan mengalami kesulitan dalam memilih *extension* yang paling cocok untuk digunakan [37].
Python	Fleksibel dalam *read* data, *machine learning package*, dan *running code*. Python juga merupakan *open source programming* [38].	Kekurangan Python terletak pada tingkat kecepatannya saat *launch* aplikasi [38].

Microsoft Excel digunakan dalam penelitian sebagai tempat penyimpanan dataset. Data yang ada kemudian akan diproses dengan menggunakan bahasa

pemrograman Python. Bahasa pemrograman Python akan digunakan pada VS Code sebagai *code editor platform*. Pada tahap ini, 3 algoritma *machine learning*, yaitu Decision Tree, Random Forest, dan XGBoost akan digunakan untuk membuat model.