**PROPOSAL PROYEK**

**Classification University using Naïve Bayes Algorithm**



**Disusun oleh:**

1. 12S17002 – Megawati L.D Sianturi
2. 12S17014 – David Simamora
3. 12S17035 – Desriyani Silaen

**PROGRAM STUDI SARJANA SISTEM INFORMASI**

**FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO INSTITUT TEKNOLOGI DEL**

**November 2020**

# DAFTAR ISI

[DAFTAR ISI 1](#_Toc57972985)

[DAFTAR GAMBAR 2](#_Toc57972986)

[Bab 1. *Business Understanding* 3](#_Toc57972987)

[1.1 Determine Business Objectives 3](#_Toc57972988)

[1.2 Access the Situation 4](#_Toc57972989)

[1.3 Determine Data Mining Goals 5](#_Toc57972990)

[1.4 Produce Project Plan 5](#_Toc57972991)

[Bab 2. *Data Understanding* 5](#_Toc57972992)

[2.1 Collect Initial Data 6](#_Toc57972993)

[2.2 Describe Data 6](#_Toc57972994)

[2.3 Explore Data 6](#_Toc57972995)

[**2.3.1 Membaca dataset** 6](#_Toc57972996)

[**2.3.2 Melihat kolom yang tersedia pada dataset** 7](#_Toc57972997)

[**2.3.4 Melihat dimensi dataset** 7](#_Toc57972998)

[**2.3.5 Detail Statistik** 8](#_Toc57972999)

[**2.3.6 Menemukan relasi antar variabel** 8](#_Toc57973000)

[2.4 Data Preparation 8](#_Toc57973001)

[**2.4.1 Data Cleaning** 8](#_Toc57973002)

[Daftar Pustaka 9](#_Toc57973003)

# DAFTAR GAMBAR

[**Gambar 1. Fase CRISM-DM** 4](#_Toc57973245)

[**Gambar 2. Tampilan Dataset University** 7](#_Toc57973246)

[**Gambar 3. Tampilan Kolom Dataset** 7](#_Toc57973247)

[**Gambar 4. Tampilan Dimensi Dataset** 7](#_Toc57973248)

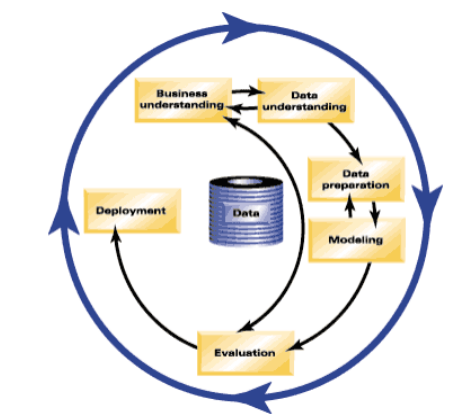
[**Gambar 5. Tampilan Detail Statistik** 8](#_Toc57973249)

# Bab 1. *Business Understanding*

Tahap pertama dari proses CRISP-DM adalah memahami apa yang ingin dicapai dari perspektif bisnis atau penelitian secara keseluruhan. Dalam tahap ini tujuan dan batasan bisnis ataupun penelitian, rencana proyek untuk mencapai data mining dan tujuan proyek, dan bagaimana strategi awal yang diperlukan serta bagamaimana rancangan yang akan dibangun untuk mencapai tujuan.

## 1.1 Determine Business Objectives

Tahap ini akan menentukan tujuan bisnis dan menentukan faktor-faktor penting yang terlibat dalam penelitian yang direncanakan. Business Objektif dari proyek ini adalah mengimplementasikan algoritma Naïve Bayes untuk klasifikasi penerimaan mahasiswa pada universitas sesuai dengan *performance* atau kemampuan mahasiswa. Klasifikasi prestasi mahasiswa merupakan hal yang sangat penting untuk meningkatkan proses penerimaan. Begitu juga dengan tingkat kinerja dari mahasiswa dapat dipengaruhi oleh banyak faktor yang terkait nilai ujian toefl, profil mahasiswa, program yang diambil, dan departemen mereka. Klasifikasi awal kinerja mahasiswa dapat membantu meningkatkan proses penerimaan mahasiswa dalam sebuah universitas. Penampilan klasifikasi dapat diperoleh dengan menggunakan teknik data mining teknik pada dataset univerity. Klasifikasi datanya adalah teknik terpenting dalam penelitian data mining. Itu tergantung pada kategorisasi (pemberian kelas) data berdasarkan nilai dari atribut klasikasi. Pengklasifikasi adalah diwakili oleh berbagai jenis model. Dan variasi algoritma sangat bagus dalam beberapa waktu untuk mendorong pengklasifikasi dari data [1]. Klasifikasi merupakan sebuah teknik pengelompokan data kedalam beberapa kategori yang ditentukan. Dalam klasifikasi, data yang diperoleh terlebih dahulu dilakukan pengolahan dengan menggunakan variabel yang ada untuk menentukan data tersebut termasuk kategori yang mana. Metode klasifikasi Naïve Bayes adalah metode pembelajaran Bayesian yang praktis; hal tersebut didasarkan pada teorema Bayesian. Dalam proses menghitung klasifikasi naive Bayes, terdapat 3 alur yaitu membaca data training, menghitung nilai probabilitas setiap atribut pada setiap kelasnya dengan memperhatikan apakah data tersebut merupakan data kualitatif atau kuantitatif dan menentukan probabilitas akhir. Probabilitas akhir yang didapat digunakan untuk menentukan termasuk dalam kelas ya atau tidak untuk data yang ingin diuji [2]. Ihsan dan Ashraf dalam penelitiannya melakukan prediksi kinerja siswa menggunakan KNN dan Naïve Bayes. Penelitian ini membuat klasifikasi siswa untuk membantu kementrian pendidikan untuk memprediksi kinerja siswa tersebut. Hasil percobaan menunjukkan bahwa Naïve Bayes lebih baik daripada KNN dengan nilai akurasi 93,6% [1]. Azwa dkk dalam penelitiannya menggunakan metode Naïve Bayes classifier dalam memprediksi kinerja mahasiswa tahun pertama mata kuliah Ilmu Komputer. Digunakan 6 parameter dalam memprediksi yaitu ras, jenis kelamin, pendapatan keluarga, mode masuk universitas, dan grade point average sebagai nilai kategorik:buruk, sedang, dan baik. Penelitian membantu dosen dan manajemen fakultas dalam mengidentifikasi mahasiswa yang kinerjanya lemah sehingga mampu mengambil kesimpulan untuk meningkatkan prestasi akademik mahasiswa [3].Terlepas Naïve Bayes merupakan salah satu teknik klasifikasi yang canggih, Sehingga pada kasus ini diteliti menggunakan metode Naïve Bayes untuk pengklasifikasian penerimaan mahasiswa pada sebuah universitas.



**Gambar 1. Fase CRISM-DM**

(Sumber : <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.735.7124&rep=rep1&type=pdf>)

## 1.2 Access the Situation

Sumber daya yang digunakan untuk mengerjakan proyek ini adalah :

* *Data sources* yang digunakan pada penelitian ini adalah dataset yang diambil dari kaggle melalui url <https://www.kaggle.com/nitishabharathi/university-recommendation>.
* Pada penelitian ini terdiri dari 3 orang mahasiswa yang berperan pada pengerjaan proyek yang berjudul *classification university using naïve bayes algorithm* mulai dari tahap *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation,* dan *deployment.*

## 1.3 Determine Data Mining Goals

Adapun tujuan data mining dalam penelitian *classification university using naïve bayes algorithm* adalah untuk memprediksi mahasiswa dapat diterima disebuah universitas atau tidak berdasarkan kemampuan mahasiswa.

## 1.4 Produce Project Plan

Tahap ini akan menjelaskan rencana yang ditujukan untuk mencapai tujuan data mining dan dengan demikian dapat mencapai tujuan bisnis, kemudian menentukan teknik dan tools yang akan dipergunakan.

Tahapan secara umum dalam perencanaan proyek ini adalah :

1. Pengumpulan data
2. Melakukan data preprocessing
3. Experimen dan pengujian model dengan metode naïve bayes
4. Evaluasi dan Validasi Hasil

Adapun *tools* yang digunakan pada proyek penelitian ini adalah python.

# Bab 2. *Data Understanding*

Tahap selanjutnya dari proses CRISP-DM adalah memperoleh data yang akan digunakan dalam proyek untuk melakukan analisis data. Pada tahap ini juga dilakukan pemeriksaan terhadap kualitas data yang akan digunakan, apakah datanya mencakup semua kasus yang diperlukan, apakah terdapat *error* pada data, dan apakah ada *missing values* di dalam data.

## 2.1 Collect Initial Data

Pada tahap ini dilakukan pengumpulan data. Dataset yang digunakan pada penelitian ini adalah dataset yang diambil dari kaggle melalui url <https://www.kaggle.com/nitishabharathi/university-recommendation>.

## 2.2 Describe Data

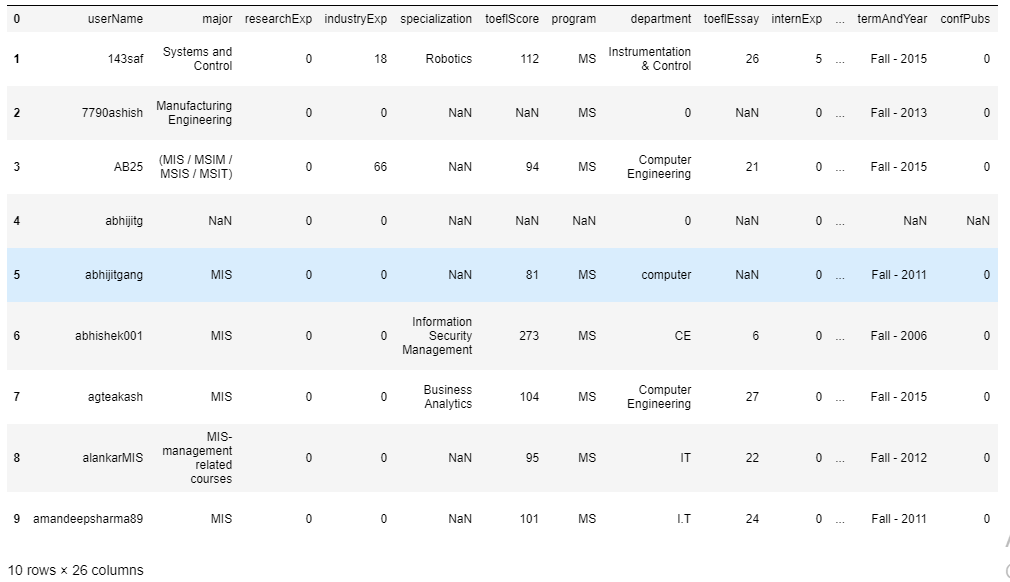
Data yang akan digunakan dalam proyek ini adalah university dataset yang diperoleh kaggle melalui url <https://www.kaggle.com/nitishabharathi/university-recommendation>. Dataset terdiri dari 53645 username mahasiswa dan 26 atribut dengan tipe atribut *categorical, integer*, dan *real*. Data ini akan digunakan untuk menerapkan teknik data mining untuk mendapatkan hasil klasifikasi.

## 2.3 Explore Data

Berikut merupakan hasil data explorasi yang kami lakukan menggunakan *Exploratory Data* *Analysis*. *Exploratory Data Analysis* (EDA) merupakan sebuah proses untuk memahami kumpulan data dengan meringkas karakteristik utamanya yang sering dilakukan dengan memplotnya secara visual. Langkah ini sangat penting terutama ketika kita sampai pada pemodelan data untuk menerapkan *Machine Learning*. EDA dicirikan oleh fleksibilitas ekstrim yang diperlukan untuk mengidentifikasi dan menyelidiki berbagai fenomena statistik dan substantif yang muncul selama penelitian empiris [4].

### **2.3.1 Membaca dataset**

Kami menggunakan library *pandas* untuk membaca data dan melihat 10 data pertama dari dataset untuk melihat gambaran data secara umum.



**Gambar 2. Tampilan Dataset University**

### **2.3.2 Melihat kolom yang tersedia pada dataset**

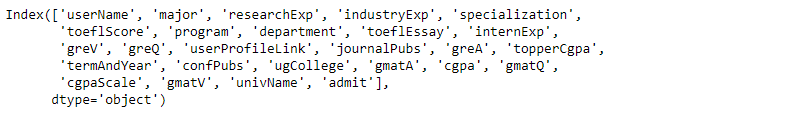
Terdapat 26 kolom pada dataset yaitu yang terdiri dari :

'userName', 'major', 'researchExp', 'industryExp', 'specialization',

'toeflScore', 'program', 'department', 'toeflEssay', 'internExp'

'greV', 'greQ', 'userProfileLink', 'journalPubs', 'greA', 'topperCgpa' 'termAndYear', 'confPubs', 'ugCollege', 'gmatA', 'cgpa', 'gmatQ’

'cgpaScale', 'gmatV', 'univName', 'admit'



**Gambar 3. Tampilan Kolom Dataset**

### **2.3.4 Melihat dimensi dataset**

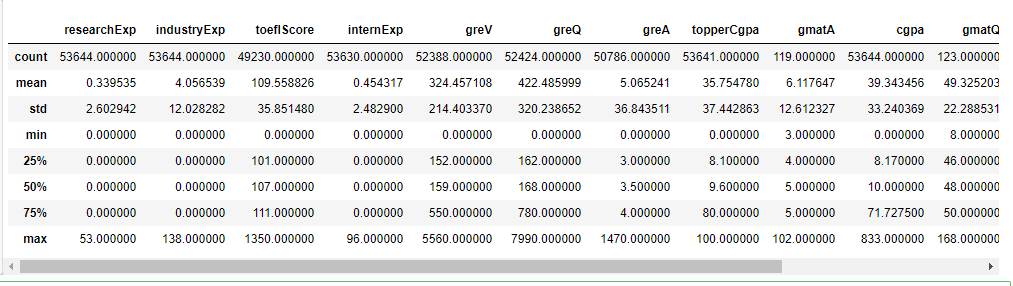
Melihat dimensi dataset untuk melihat berapa jumlah kolom dan baris yang terdapat pada dataset yang akan digunakan. Dimensi dari dataset yang digunakan adalah 53645 baris dan 26 kolom



**Gambar 4. Tampilan Dimensi Dataset**

### **2.3.5 Detail Statistik**

Salah satu fungsi yang terdapat pada library pandas adalah describe(). describe() digunakan untuk melihat detail statistik seperti persentil, rata-rata, standart deviasi dan lain-lain.



**Gambar 5. Tampilan Detail Statistik**

### **2.3.6 Menemukan relasi antar variabel**

Heatmaps dalam python digunakan untuk menentukan relasi antar variabel. Relasi antar variabel sering juga disebut sebagai korelasi.

## 2.4 Data Preparation

Tahap persiapan data mencakup semua aktivitas untuk menyusun dataset akhir dari data mentah awal.

### **2.4.1 Data Cleaning**

Cleaning yang dilakukan adalah dengan mengecek adanya nilai null, adanya data yang duplikat, dan mendeteksi adanya outlier didalam dataset. Data Cleaning juga masih termasuk kedalam bagian dari Exploratory Data Analysis.

**2.4.1.1 Feature Selection**

Feature selection dilakukan untuk membuang sebagian kolom atau atribut yang dianggap tidak terlalu dibutuhkan, dalam proses penambangan data, dengan menggunakan fungsi drop untuk men-drop beberapa atribut.



Sehingga tersisa 9 atribut yang akan digunakan.

#### 2.4.1.2 Menangani Missing Values

Missing value dapat dilakukan dengan menghapus adanya nilain NaN atau null pada dataset yang kita miliki



Dengan menggunakan fungsi dropna maka tabel yang mengandung nilai NaN atau null akan d drop dari dataset.

#### 2.4.1.2. Dropping non admitted data

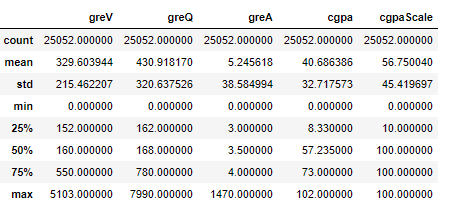
Pada dataset yang dimiliki terdapat atribut admit yant menyatakan apakah data tersebut diterima atau tidak. Dalam konteks ini kita akan mengurangi data yang dimiliki secara massive dengan cara tidak mengolah data yang tidak admit atau nilai dari atribut admin = 0.



Setelah dipilih nilai dari admit yang sesuai maka , atribut admit juga akan d drop karena tidak diperlukan lagi untuk keperluan klasifikasi.

#### 2.4.1.3 Melakukan analisis Statistik

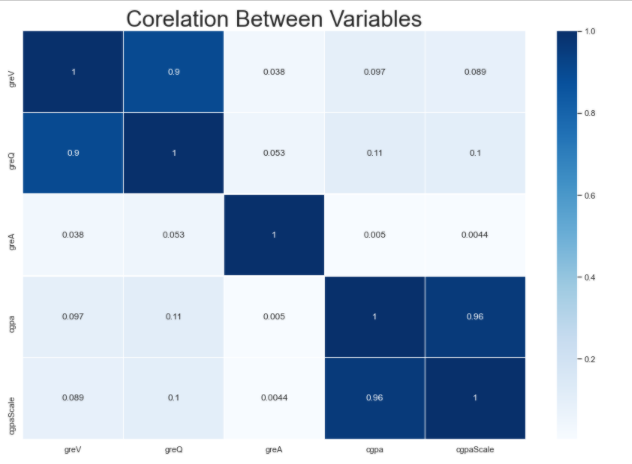
1. Data Statistik



Terdapat 5 atribut numerik dan kita melihat penjelasan mengenai data statistic yang ada pada dataset tersebut.

2. Korelasi

Korelasi diperukan untuk mengetahui hubungan antar atribut yang diniliki oleh dataset tersebut. Dengan adanya korelasi kita bisa menilai apakah atribut-atribut tersebut mengalami korelasi positif atau negatif.



Dari gambar diatas kita bisa melihat bahwa semua data yang kita miliki memiliki korelasi positif. Dimana korelasi positif artinya bahwa perubahan nilai pada atribut yang kita miliki menyatakan bahwa atribut yang kita miliki akan bergerak dan akan berbanding lurus terhadap perubahan data yang kita miliki.

# Daftar Pustaka

[1] I. A. A. Amra and A. Y. A. Maghari, “Students Performance Prediction Using KNN and Naïve Bayesian,” 2017, pp. 909–913.

[2] M. Sabransyah, Y. N. Nasution, and F. D. T. Amijaya, “Aplikasi Metode Naive Bayes dalam Prediksi Risiko Penyakit Jantung,” *J. EKSPONENSIAL*, vol. 8, no. 2, pp. 111–118, 2017.

[3] A. A. Aziz, N. H. Ismail, F. Ahmad, and H. Hassan, “A Framework for Students’ Academic Performance Analysis using Naïve Bayes Classifier,” *J. Teknol.*, vol. 75, no. 3, pp. 13–19, 2015, doi: 10.11113/jt.v75.5037.

[4] A. T. Jebb, S. Parrigon, and S. E. Woo, “Exploratory data analysis as a foundation of inductive research,” *Hum. Resour. Manag. Rev.*, vol. 27, no. 2, pp. 265–276, Jun. 2017, doi: 10.1016/j.hrmr.2016.08.003.