**PROPOSAL PROYEK**

**12S4054 – DATA MINING**

**Klasifikasi Penyakit Hepatitis Menggunakan Algoritma Naïve** **Bayes**



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** 2](#_Toc57037701)

[**1.** **Business Understanding** 4](#_Toc57037702)

[**1.1 Objektif** 4](#_Toc57037703)

[**1.3 Perencanaan proyek** 5](#_Toc57037704)

[**1.4 Business Success Criteria** 5](#_Toc57037705)

[**2.Data Understanding** 5](#_Toc57037706)

[**2.1 Deskripsi data** 5](#_Toc57037707)

[**2.2 Data exploration** 5](#_Toc57037708)

[**2.2.1 Membaca dataset** 6](#_Toc57037709)

[**2.2.2 Melihat kolom yang tersedia pada dataset** 6](#_Toc57037710)

[**2.2.3 Melihat dimensi dataset** 6](#_Toc57037711)

[**2.2.4 Detail Statistik** 7](#_Toc57037712)

[**3.** **Data Preparation** 9](#_Toc57037713)

[**3.1.1 Dataset** 9](#_Toc57037714)

[**3.1.2 Cleaning Data** 9](#_Toc57037715)

[**Referensi** 13](#_Toc57037716)

**DAFTAR GAMBAR**

[Gambar 1. Tampilan Dataset Penyakit Hepatitis 6](file:///C:\Users\David\Downloads\Dami.docx#_Toc56704361)

[Gambar 2. Tampilan Kolom Dataset Penyakit Hepatitis 6](#_Toc56704362)

[Gambar 3. Tampilan Dimensi Dataset Penyakit Hepatitis 7](#_Toc56704363)

[Gambar 4. Tampilan Detail Statistik Dataset Penyakit Hepatitis 7](file:///C:\Users\David\Downloads\Dami.docx#_Toc56704364)

[Gambar 5 Korelasi 8](#_Toc56704365)

[Gambar 6 Korelasi dengan visualisasi heatmap 8](#_Toc56704366)

[Gambar 7. Pengecekan data null 9](file:///C:\Users\David\Downloads\Dami.docx#_Toc56704367)

[Gambar 8 tampilan boxplot pada atribut ‘bilirubin’ 10](file:///C:\Users\David\Downloads\Dami.docx#_Toc56704368)

[Gambar 9 Tampilan boxplot pada atribut protime 10](file:///C:\Users\David\Downloads\Dami.docx#_Toc56704369)

[Gambar 10 Tampilan boxplot pada atribut 'age' 11](#_Toc56704370)

[Gambar 11 Tampilan boxplot pada atribut 'alt\_phospate' 11](file:///C:\Users\David\Downloads\Dami.docx#_Toc56704371)

[Gambar 12 Tampilan boxplot pada albumin 11](file:///C:\Users\David\Downloads\Dami.docx#_Toc56704372)

[Gambar 13 Menghitung IQR 12](#_Toc56704373)

[Gambar 14 Menerapkan IQR Score 12](file:///C:\Users\David\Downloads\Dami.docx#_Toc56704374)

# **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 Objektif**

Objektif dari proyek ini adalah mengimplementasikan algoritma Naïve Bayes untuk diagnosis penyakit hepatitis. Hepatitis merupakan penyakit yang diderita oleh banyak orang, bahkan bisa menyebabkan kematian. Prediksi awal dapat mencegah kematian tersebut yaitu dengan mengumpulkan data pasien hepatitis yang dilihat dari faktor – faktornya. Untuk mengolah data tersebut, dibutuhkan Data Mining. Salah satu metode data mining yang digunakan pada penelitian ini adalah klasifikasi. 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. Terlepas Naïve Bayes merupakan salah satu teknik klasifikasi yang canggih yang juga digunakan untuk diagnosis penyakit. Sehingga pada kasus ini diteliti menggunakan metode Naïve Bayes untuk pengklasifikasian penyakit hepatitis.

**1.2 Tujuan**

Tujuan dari penelitian ini adalah :

1. Menerapakan teknik data mining untuk melakukan klasifikasi menggunakan teknik naïve bayes terhadap dataset hepatitis dan memprediksi seseorang apakah berpotensi terserang hepatitis atau tidak.

## **1.3 Perencanaan proyek**

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

## **1.4 Business Success Criteria**

*Business Succes Criteria* dari penelitan ini adalah ketika penelitian dapat membuat klasifikasi pada dataset dan menentukan pasien yang berpotensi hepatitis dan tidak, terlepas dari akurasi dan presisi yang dihasilkan.

# **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 Deskripsi data**

Data yang akan digunakan dalam proyek ini adalah Hepatitis dataset yang diperoleh dari UCI *Machine Learning* (https://archive.ics.uci.edu/ml/datasets/hepatitis) dan tersedia di *kaggle (*https://kaggle.com*)*. Dataset terdiri dari 20 atribut dengan tipe atribut *categorical, integer*, dan *real*. Data ini akan digunakan untuk menerapkan teknik data mining untuk mendapatkan hasil klasifikasi.

## **2.2 Data exploration**

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. [2]

### **2.2.1 Membaca dataset**

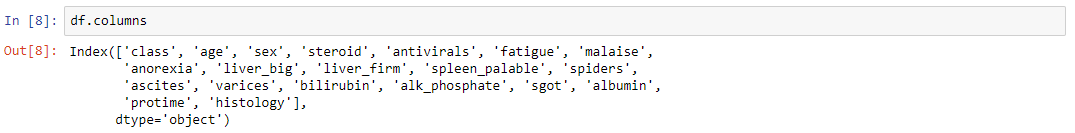
Kami menggunakan library *pandas* untuk membaca data dan melihat 10 data pertama dari dataset untuk melihat gambaran data secara umum.

Gambar 1. Tampilan Dataset Penyakit Hepatitis

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

Terdapat 12 kolom pada dataset yaitu yang terdiri dari :

‘class’,’age’,’sex’,’steroid’,’antivirals’,’fatigue’,’malaise’,’anorexia’,’liver\_big’,’liver\_firm’,’spleen\_palable’,’spiders’,’ascites’,’varices’,’bilirubin’,’alk\_phoshate’,’sgot’,’albumin,’protime’,’histology’.



Gambar 2. Tampilan Kolom Dataset Penyakit Hepatitis

### **2.2.3 Melihat dimensi dataset**

Melihat dimensi dataset untuk melihat berapa jumlah kolom dan baris yang terdapat pada dataset yang kami gunakan

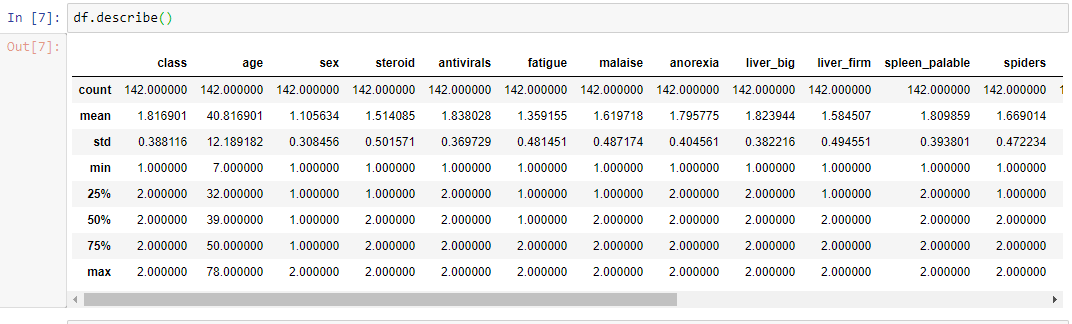


Gambar 3. Tampilan Dimensi Dataset Penyakit Hepatitis

Dimensi dari dataset yang kami gunakan ialah 142 baris dan 20 kolom

### **2.2.4 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.

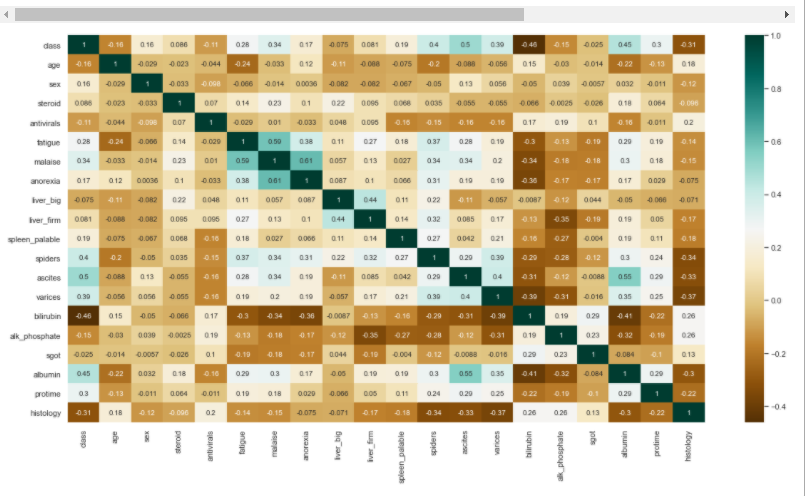


Gambar 4. Tampilan Detail Statistik Dataset Penyakit Hepatitis

**2.2.5 Menemukan relasi antar variabel**

Menggunakan heatmaps untuk menentukan relasi antar variabel atau kita sering kita sebut sebagai korelasi



Gambar 5 Korelasi

Gambar 6 Korelasi dengan visualisasi heatmap

# **3. Data Preparation**

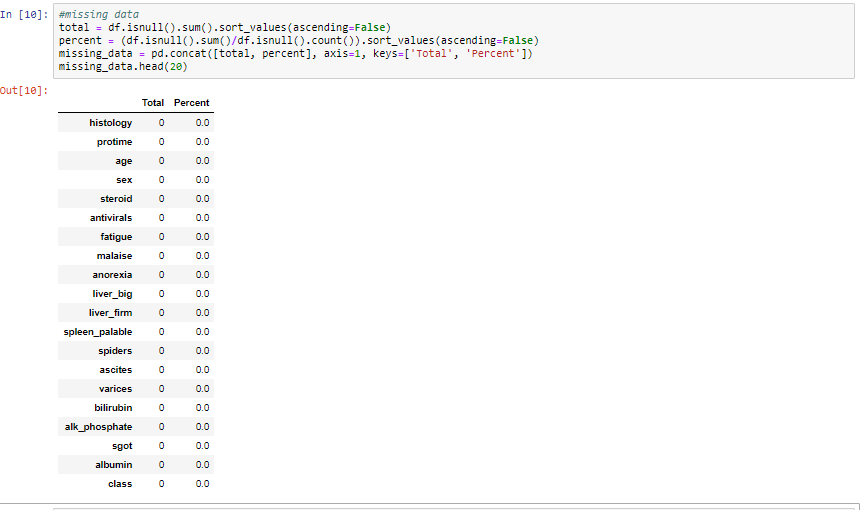
Tahap persiapan data mencakup semua aktivitas untuk menyusun dataset akhir dari data mentah awal.

## **3.1.1 Dataset**

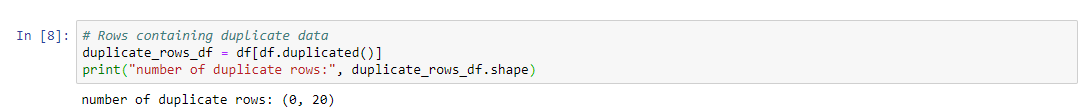
Data yang akan digunakan dalam proyek ini adalah Hepatitis dataset yang diperoleh dari UCI *Machine Learning* (https://archive.ics.uci.edu/ml/datasets/hepatitis) dan tersedia di *kaggle (*https://kaggle.com*)*. Dataset terdiri dari 20 atribut dengan tipe atribut *categorical, integer*, dan *real*.

## **3.1.2 Cleaning Data**

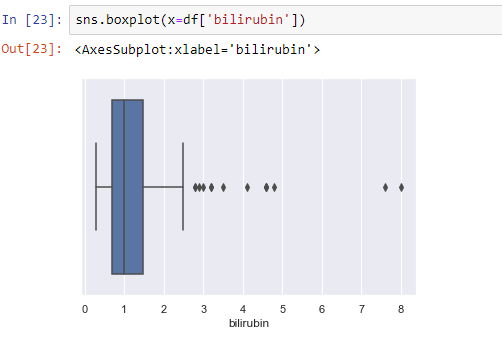
Cleaning yang dilakukan adalah dengan mengecek adanya nilai null, adanya data yang duplikat, dan mendeteksi adanya outlier didalam dataset. Data Cleaning juga masih termasuk kedalam bagian dari Exploratory Data Analysis.

3.1.2.1 Menangani missing data

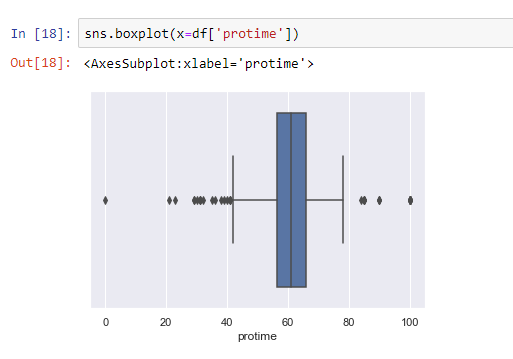
Gambar 7. Pengecekan data null

3.1.2.2 Mendeteksi adanya data yang duplikat

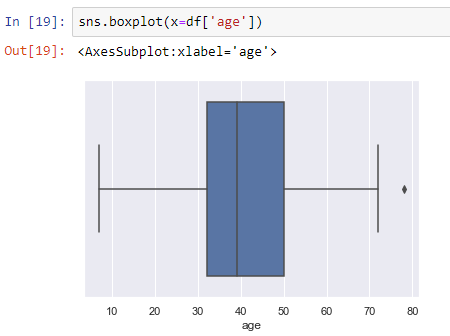
3.1.2.3 Mendeteksi outlier dengan seaborn pada kelas data numerical



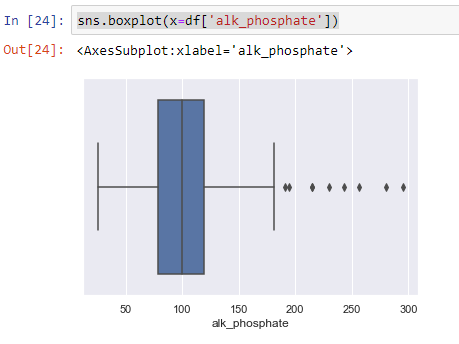
Gambar 8 tampilan boxplot pada atribut ‘bilirubin’



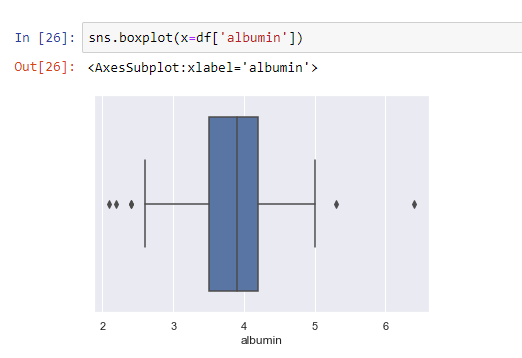
Gambar 9 Tampilan boxplot pada atribut protime



Gambar 10 Tampilan boxplot pada atribut 'age'

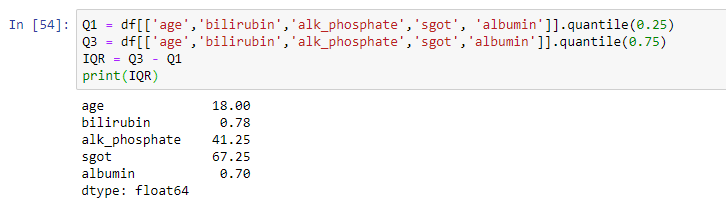


Gambar 11 Tampilan boxplot pada atribut 'alt\_phospate'



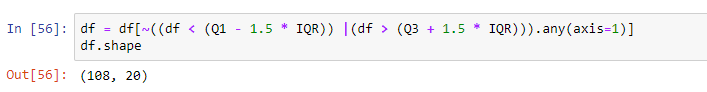
Gambar 12 Tampilan boxplot pada albumin

Setelah menggunakan boxplot dari library seaborn kita melihat adanya beberapa outlier pada dataset yang kita miliki , kemudian kita akan menghilangkan outlier dengan menggunakan teknik IQR Score. Kita menghitung kuartil atas dan kuartil bawah dari masing-masing atribut, kemudian mengurangkannya, dimana hasilnya kita sebut sebagai IQR.



Gambar 13 Menghitung IQR

Kemudian kita menghitung IQR Scorenya lalu menampilkan dimensi dataset yang kita miliki dengan menggunakan .shape()



Gambar 14 Menerapkan IQR Score

Kita melihat bahwa dimensi dari data yang kita miliki telah berubah dari (142, 20) menjadi (108, 20), dimana beberapa outlier berhasil dihilangkan dengan menggunakan IQR Score

# **Referensi**

|  |  |
| --- | --- |
| [1] | ,. K. H. M. T. A. Sulastri, "ANALISIS PERBANDINGAN KLASIFIKASI PREDIKSI PENYAKIT HEPATITIS DENGAN MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR, NAÏVE BAYES DAN NEURAL NETWORK," *Jurnal DINAMIK,* vol. 24, no. 2, pp. 82-91, 2019. |
| [2] | "Exploratory data analysis as a foundation of inductive research," *Human Resource Management ,* 2016. |
| [3] | M. B. Agus Byna, "Penerapan Metode Adaboost Untuk Mengoptimasi," *Jurnal SISFOKOM (Sistem Informasi dan Komputer),* vol. 09, no. 3, pp. 407-411, 2020. |