TUGAS BESAR
SISTEM CERDAS

# HEPATITIS-C DETECTION USING K-NEAREST NEIGHBOR (KNN)

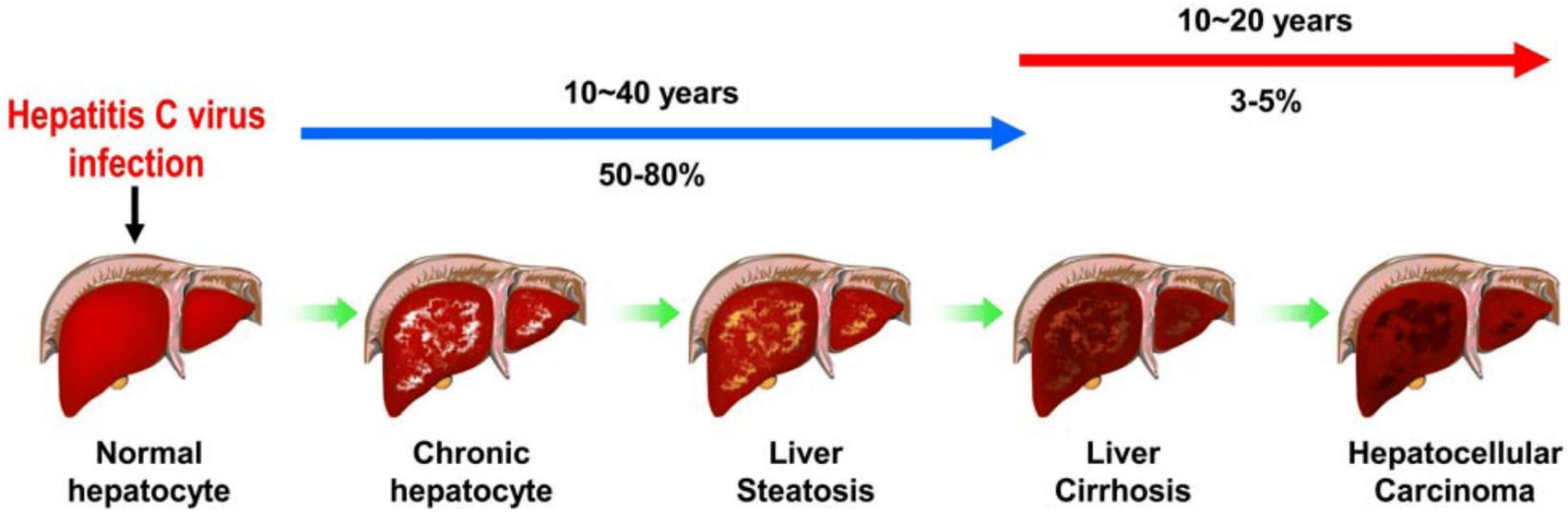

Oleh :

Hanin Nafi'ah (NIM. 2101222083)
Vi Bauty Riska Utami (NIM. 2101222073)

S2 Teknik Elektro

TELKOM UNIVERSITY

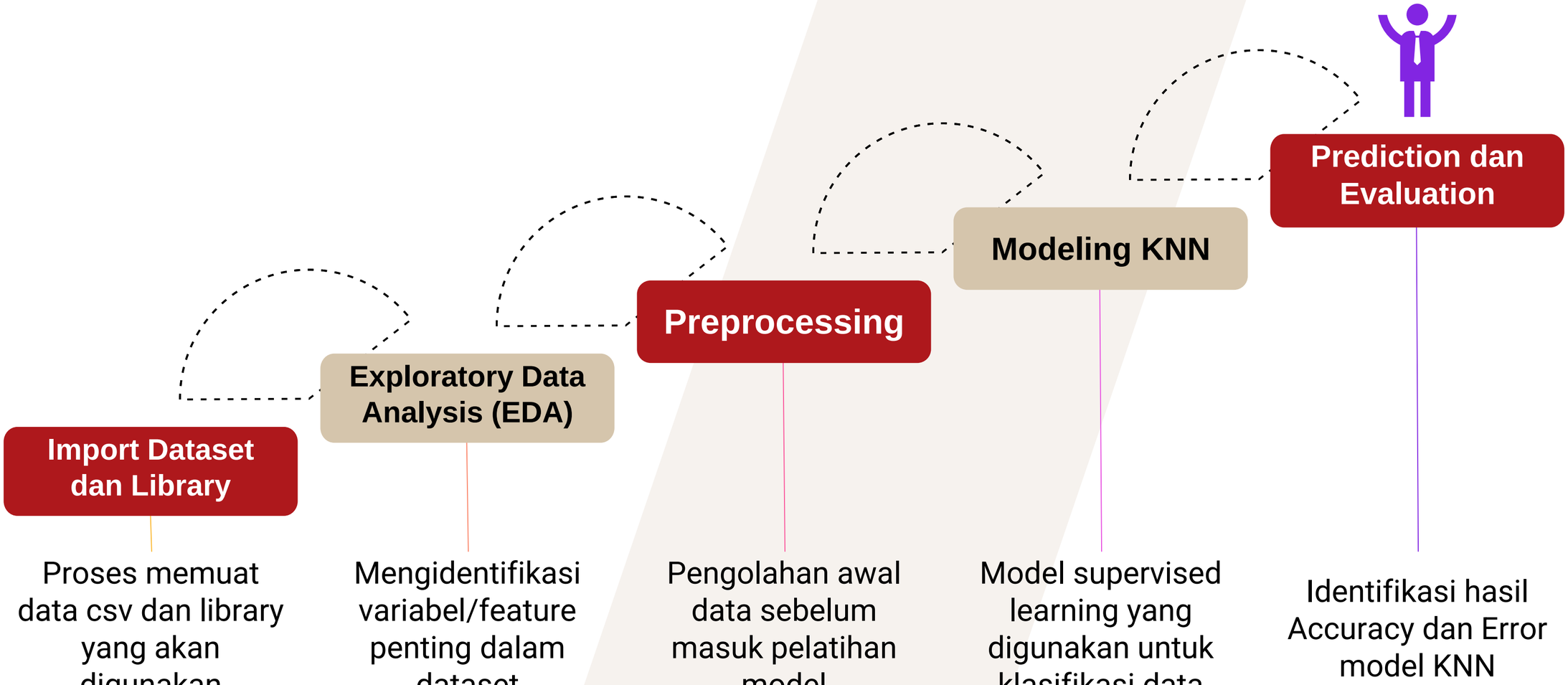
# Background of Problem




Hepatitis merupakan penyakit yang disebabkan oleh beberapa jenis virus yang menyerang dan menyebabkan peradangan serta merusak sel-sel organ hati manusia. Berdasarkan data dari World Health Organization (WHO) pada tahun 2021, ditunjukkan bahwa sebanyak 1% atau 71 juta orang di seluruh dunia terinfeksi virus hepatitis C (HCV) dimana 399 ribu diantaranya meninggal dunia dikarenakan sirosis hati. Melihat data penderita serta dampak dari penyakit hepatitis tersebut, maka perlu dilakukan penanganan untuk menghambat perkembangan penyakit hepatitis C. Salah satu upaya yang dapat dilakukan adalah melakukan screening untuk mendeteksi penyakit hepatitis C.

# Research Methodology





# Import Dataset & Library



```
# Import the csv data
from google.colab import files
files.upload()
```

```
# Import the library
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.impute import SimpleImputer
from sklearn.preprocessing import LabelEncoder
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from imblearn.over_sampling import SMOTE

from sklearn.metrics import classification_report, accuracy_score
from sklearn.metrics import mean_squared_error as mse

import warnings
warnings.filterwarnings('ignore')
```


Data pada penelitian ini diambil dari website UCI Machine Learning Repository yaitu Hepatitis C Prediction Dataset. Data ini terdiri dari 13 fitur dan 615 baris data yang berisi terkait rekam medis dari pasien yang terdeteksi memiliki penyakit hepatitis C maupun pasien yang terdeteksi sehat.

# Exploratory Data Analysis (EDA)



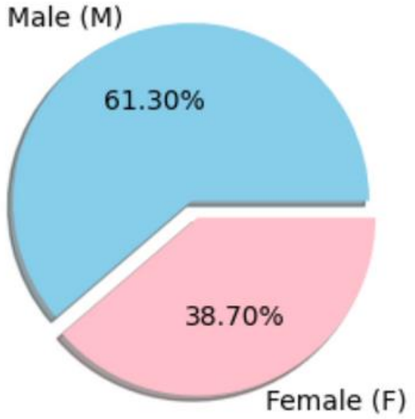
```
# Loading up the dataset
data = pd.read_csv('./HepatitisCdata.csv')
data.head()
```

| index | Unnamed: 0 | Category | Age | Sex | ALB | ALP | ALT | AST | BIL | CHE | CHOL | CREA | GGT | PROT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0=Blood Donor | 32 | m | 38.5 | 52.5 | 7.7 | 22.1 | 7.5 | 6.93 | 3.23 | 106.0 | 12.1 | 69.0 |
| 1 | 2 | 0=Blood Donor | 32 | m | 38.5 | 70.3 | 18.0 | 24.7 | 3.9 | 11.17 | 4.8 | 74.0 | 15.6 | 76.5 |
| 2 | 3 | 0=Blood Donor | 32 | m | 46.9 | 74.7 | 36.2 | 52.6 | 6.1 | 8.84 | 5.2 | 86.0 | 33.2 | 79.3 |
| 3 | 4 | 0=Blood Donor | 32 | m | 43.2 | 52.0 | 30.6 | 22.6 | 18.9 | 7.33 | 4.74 | 80.0 | 33.8 | 75.7 |
| 4 | 5 | 0=Blood Donor | 32 | m | 39.2 | 74.1 | 32.6 | 24.8 | 9.6 | 9.15 | 4.32 | 76.0 | 29.9 | 68.7 |

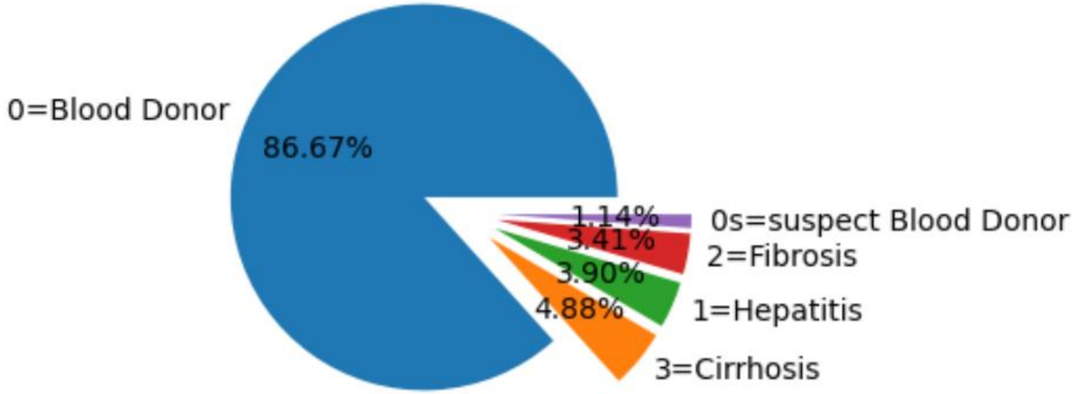
Untuk tiap atribut dari data dapat dilihat pada tabel sebagai berikut :

| Fitur | Penjelasan | Keterangan |
|---|---|---|
| Category | Kategori atau tipe pasien (Healthy Patient, Suspected Patient) | Label |
| Age | Umur pasien | Atribut |
| Sex | Jenis kelamin pasien (Female, Male) | Atribut |
| ALB | Kadar albumin pada darah pasien | Atribut |
| ALP | Kadar alkaline phosphatase pada darah pasien | Atribut |
| ALT | Kadar alanine transaminase pada darah pasien | Atribut |
| AST | Kadar aspartate aminotransferase pada darah pasien | Atribut |
| BIL | Kadar bilirubin pada darah pasien | Atribut |
| CHE | Kadar cholinesterase pada darah pasien | Atribut |
| CHOL | Kadar cholesterol pada darah pasien | Atribut |
| CREA | Kadar creatine pada darah pasien | Atribut |
| GGT | Kadar gamma-glutamyl pada darah pasien | Atribut |
| PROT | Kadar protein pada darah pasien | Atribut |

## Distribusi Gender Dataset

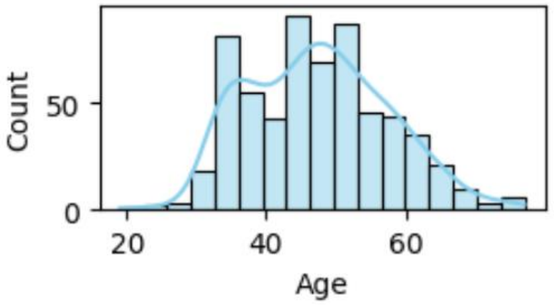


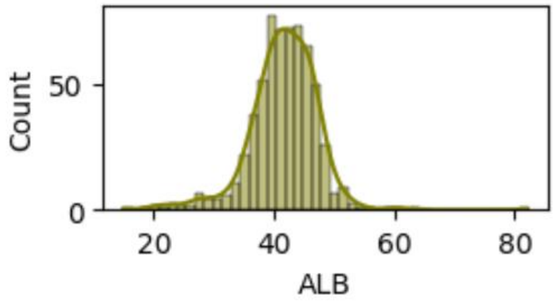

## Distribusi Kategori Rekam Medis Pasien

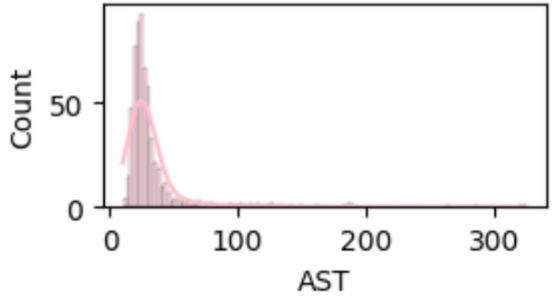


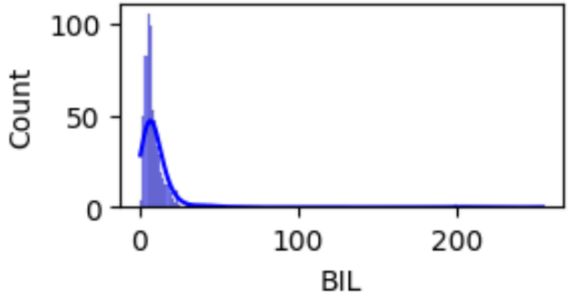
# Exploratory Data Analysis (EDA)

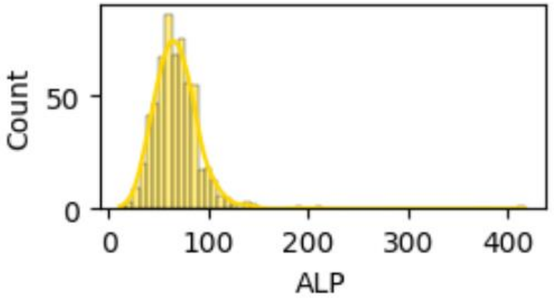


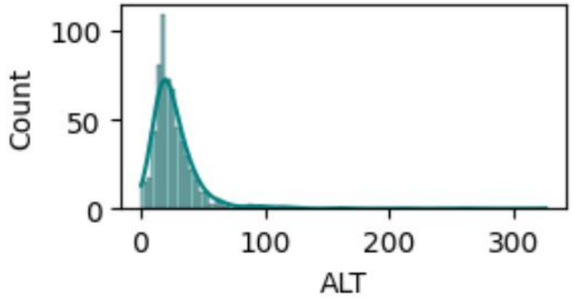


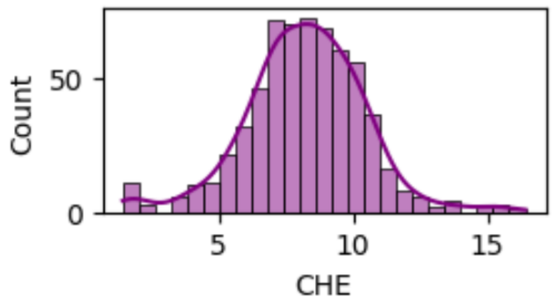



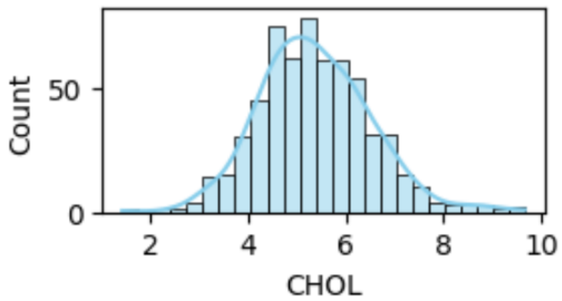



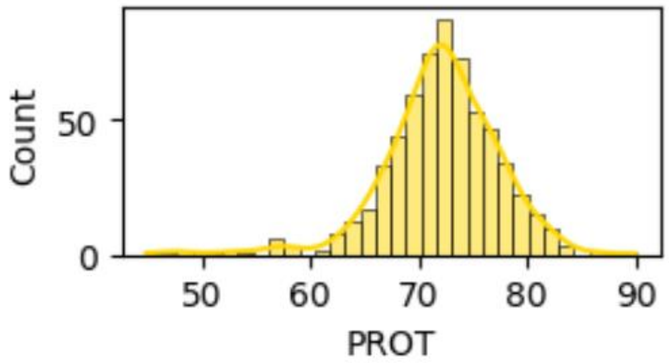



# Preprocessing

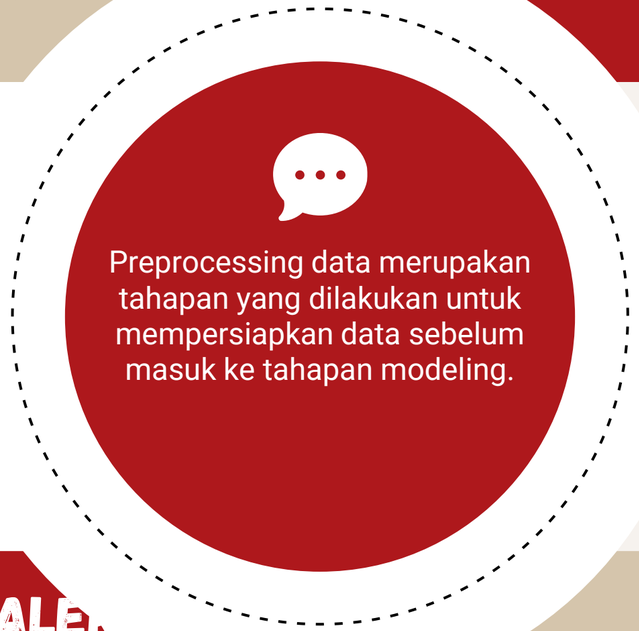


## SIMPLE IMPUTER

Menangani missing value atau nilai kosong (NaN) dengan substitusi nilai diisi dengan mean/rata-rata.

## LABEL ENCODING

Mengkonversi data label menjadi bentuk angka dengan proses transformasi kategorikal menjadi numerik




## STANDAR SCALER

Normalisasi agar data yang digunakan memiliki range yang sama

## SMOTE

Mengatasi ketidakseimbangan data dengan cara sampling ulang kelas minoritas

# Modeling K-Nearest Neighbor

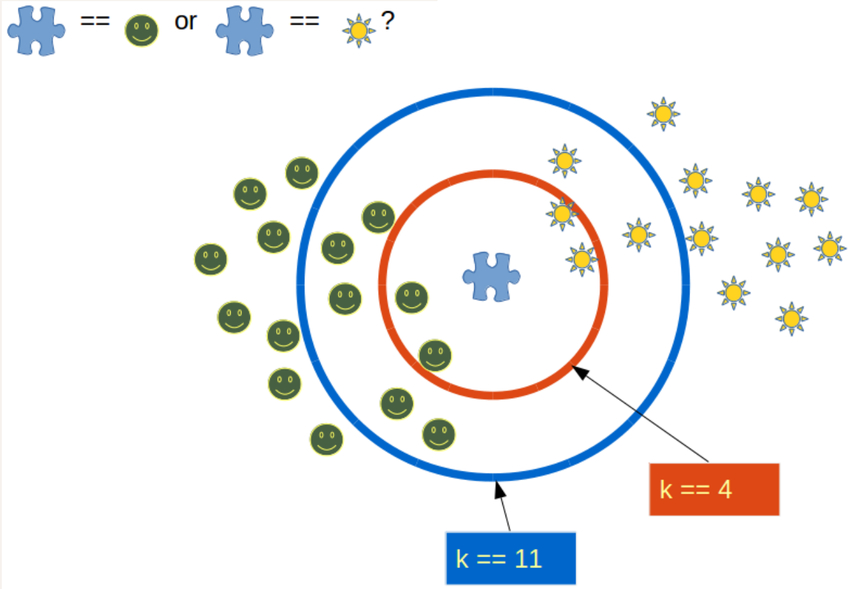


```
KNN = KNeighborsClassifier(n_neighbors=4,metric='euclidean',weights='distance')
KNN.fit(X_train,y_train)
```

```
KNeighborsClassifier
KNeighborsClassifier(metric='euclidean', n_neighbors=4, weights='distance')
```

| | Model | Metric | Weights | Score | MSE |
|---|---|---|---|---|---|
| 0 | K-Nearest Neighbor | Euclidean | Distance | 0.988743 | 0.048780 |
| 1 | K-Nearest Neighbor | Euclidean | Uniform | 0.986867 | 0.045028 |
| 2 | K-Nearest Neighbor | Minkowski | Distance | 0.988743 | 0.048780 |
| 3 | K-Nearest Neighbor | Minkowski | Uniform | 0.986867 | 0.045028 |
| 4 | K-Nearest Neighbor | Manhattan | Distance | 0.986867 | 0.060038 |
| 5 | K-Nearest Neighbor | Manhattan | Uniform | 0.986867 | 0.060038 |

K-Nearest Neighbors (K-NN) adalah salah satu algoritma supervised learning yang digunakan untuk model klasifikasi. K-NN bekerja dengan cara mencari K titik data terdekat dari data yang akan diprediksi. K dalam K-NN merupakan parameter yang menentukan jumlah tetangga terdekat yang akan dipertimbangkan dalam proses prediksi. Nilai K yang lebih besar akan mengambil lebih banyak tetangga terdekat, sementara nilai K yang lebih kecil akan mempertimbangkan tetangga terdekat yang lebih sedikit. Pemilihan nilai K yang tepat sangat penting, karena dapat mempengaruhi akurasi prediksi dan tingkat kompleksitas algoritma.



# Data Prediction



```
# Report Prediction
print(classification_report(y_test,y_pred_type1))
```

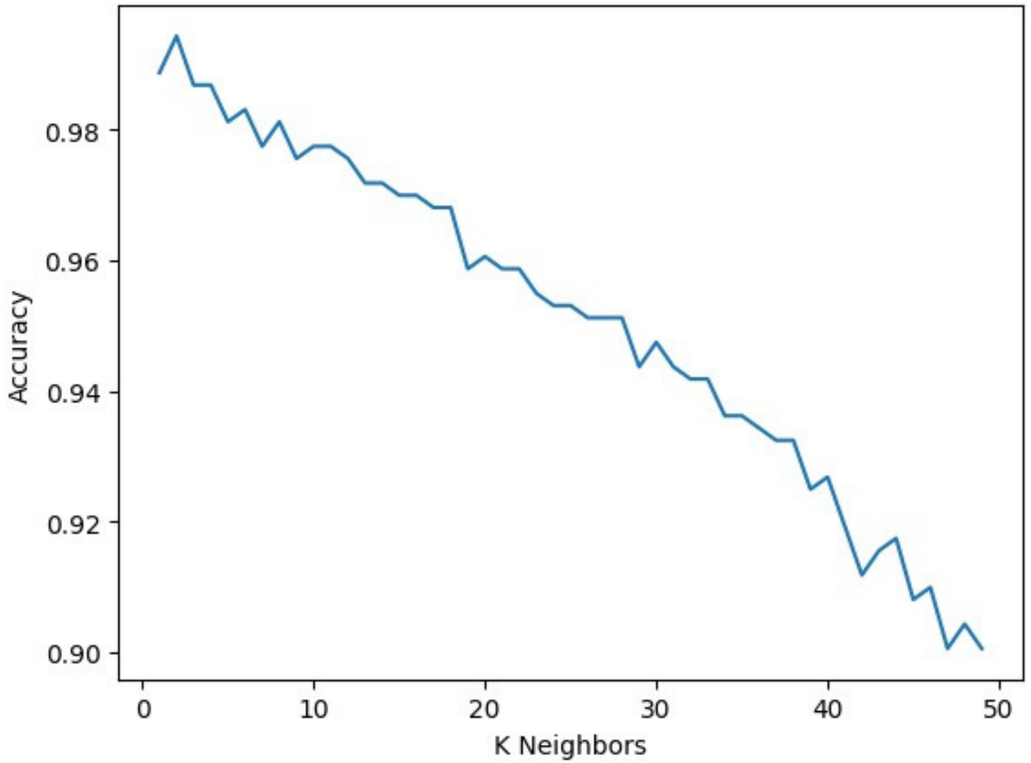
```
              precision    recall  f1-score   support

           0       1.00      0.94      0.97        98
           1       0.99      1.00      1.00       121
           2       0.96      1.00      0.98        98
           3       0.99      1.00      0.99        99
           4       1.00      1.00      1.00       117

    accuracy                           0.99       533
   macro avg       0.99      0.99      0.99       533
weighted avg       0.99      0.99      0.99       533
```

Classification report adalah suatu metode untuk mengevaluasi kinerja model klasifikasi dengan memberikan informasi terperinci tentang precision, recall, F1-score, dan support untuk setiap kelas yang diprediksi. Berikut adalah informasi yang biasanya terdapat dalam classification report:

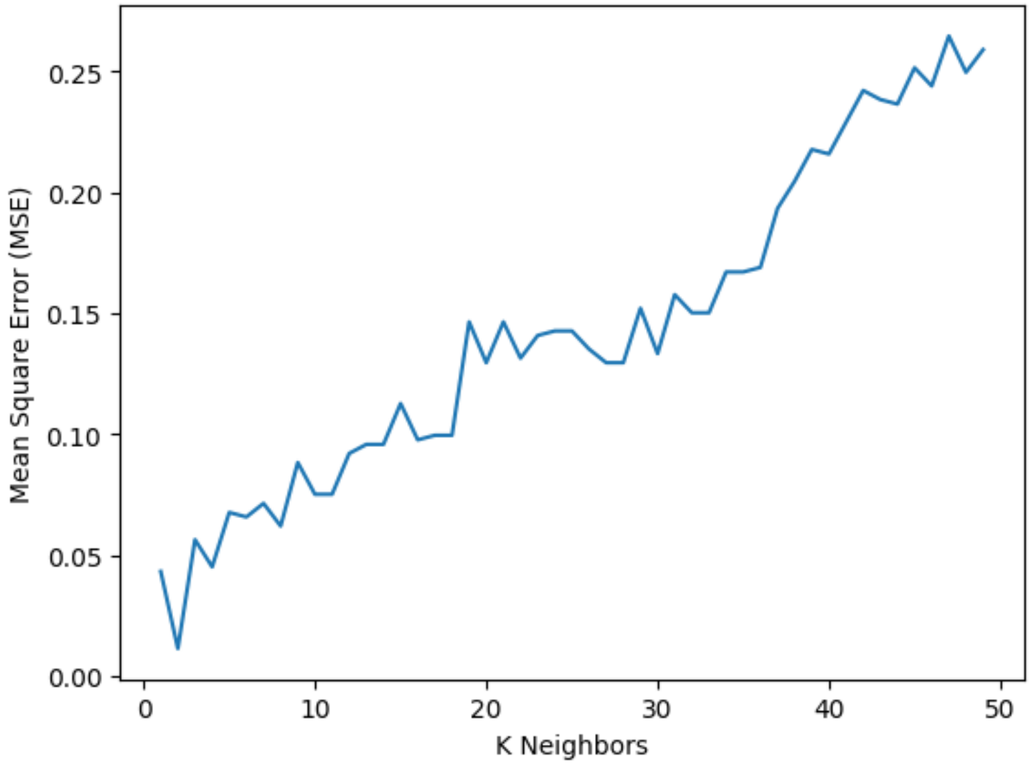
- Precision: Precision mengukur sejauh mana prediksi positif yang dibuat oleh model benar. Precision dihitung dengan rumus: TP / (TP + FP), di mana TP adalah True Positive (jumlah prediksi benar positif) dan FP adalah False Positive (jumlah prediksi salah positif).
- Recall: Recall (sensitivity atau true positive rate) mengukur sejauh mana model dapat mengidentifikasi dengan benar kelas positif. Recall dihitung dengan rumus: TP / (TP + FN), di mana TP adalah True Positive (jumlah prediksi benar positif) dan FN adalah False Negative (jumlah prediksi salah negatif).
- F1-score: F1-score adalah nilai rata-rata harmonik dari precision dan recall. F1-score menggabungkan kedua metrik ini untuk memberikan nilai yang seimbang antara precision dan recall. F1-score dihitung dengan rumus: 2 * ((precision * recall) / (precision + recall)).
- Support: Support adalah jumlah sampel yang termasuk dalam setiap kelas. Support memberikan informasi tentang seberapa banyak sampel yang terdapat dalam setiap kelas.

# Data Evaluation






K-Nearest Neighbors Model dengan range nilai K yaitu 1 sampai 50 diperoleh seperti gambar diatas semakin banyak nilai K maka akurasinya akan semakin turun. Berdasarkan grafik diatas nilai K dengan akurasi yang bagus ada pada range 1 sampai 5.




K-Nearest Neighbors Model dengan range nilai K yaitu 1 sampai 50 diperoleh seperti gambar diatas semakin banyak nilai K maka errornya akan semakin turun. Berdasarkan grafik diatas nilai K dengan error yang rendah ada pada range 1 sampai 5.

# Conclusion

- Model yang digunakan di simulasi ini: Euclidean Distance, Euclidean Uniform, Minkowski Distance, Minkowski Uniform, Manhattan Distance, dan Manhattan Uniform.
- Berdasarkan hasil simulasi yang dilakukan pada deteksi dini penyakit Hepatitis C dengan metode klasifikasi K-Nearest Neighbor dihasilkan akurasi testing tertinggi sebesar 98,8743% dan MSE (Mean Squad Error) 0,048780% menggunakan metode Euclidean Distance dan Minkowski Distance.

# *Catatan Elaborasi Tugas Besar*

*(setelah Revisi)*

- Buatlah model menggunakan beberapa parameter saja dari data tersebut (10 Parameter : ALB, ALP, ALT, AST, BIL, CHE, CHOL, CREA, GGT, PROT). Pilihlah beberapa (3-5 parameter) dengan korelasi atribut yang terbesar dan mudah dijangkau oleh masyarakat. Kemudian simulasikan hasil akurasi dan error sesuai dengan model yang sudah dibuat sebelumnya.

# Features Selection



**Feature selection adalah suatu proses menghapus features yang berlebihan dan tidak relevan dari dataset yang sebenarnya. Sehingga waktu yang digunakan mengeksekusi dari pengklasifikasi yang memproses data berkurang, dan dapat meningkatkan akurasi juga karena features yang tidak relevan dapat memperburuk data mempengaruhi akurasi klasifikasi secara negatif. Dengan feature selection dapat meningkatkan pemahaman dan biaya penanganan data menjadi lebih kecil.**

**Terdapat banyak algoritma features selection, yang akan kami gunakan dalam penelitian ini yaitu algoritma features selection yang bersifat univariate yang disebut select k best.**

| | Input Parameter (Features) | F_Score |
|---|---|---|
| 3 | AST | 444.528329 |
| 4 | BIL | 176.678485 |
| 8 | GGT | 174.913636 |
| 5 | CHE | 74.645270 |
| 6 | CHOL | 60.739136 |

```
Accuracy Score 0.9643527204502814
MSE (Mean Squared Error) 0.1651031894934334
```

## KESIMPULAN

Setelah dilakukan seleksi parameter yang digunakan dalam model prediksi dini sebagai updaya screening awal penyakit Hepatitis C diperoleh hasil 5 parameter tertinggi yang berpengaruh dalam pembuatan model KNN yaitu AST, BIL, GGT, CHE & CHOL.

Namun apabila dilihat dari hasil akurasi dan error setelah dilakukan pengurangan parameter. Diperoleh hasil nilai akurasi menjadi semakin rendah dan nilai error menjadi semakin tinggi. Hal tersebut menunjukan bahwa setiap atribut (10 parameter) mempunyai peran dan saling berpengaruh dalam proses deteksi dini penyakit Hepatitis C.

# Reference and Link:

1. Implementasi Algoritma K-Nearest Neighbor (K-NN) dalam Deteksi Dini Penyakit Hepatitis C | Ni Made Rika Padeswari Kusuma, L. G. Astuti | JNATIA Volume 1, Nomor 1, November 2022

2. https://colab.research.google.com/drive/18knERVkcWOv_kgzC2y-y76m16yl77UGs#scrollTo=lSdpSe2vqczW

# TERIMA KASIH