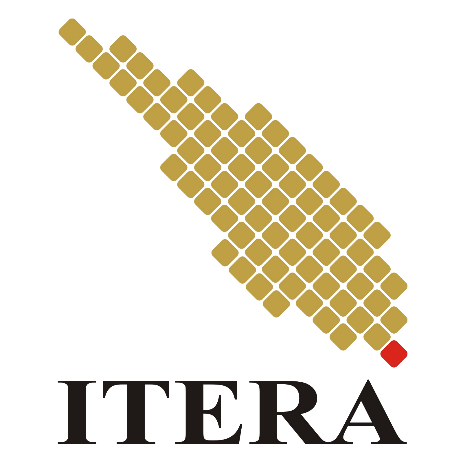
**LAPORAN TUGAS INDIVIDU MODUL 5**

****

DISUSUN OLEH :

MUJADID CHOIRUS SURYA

121450015

RA

**PROGRAM STUDI SAINS DATA**

**JURUSAN SAINS**

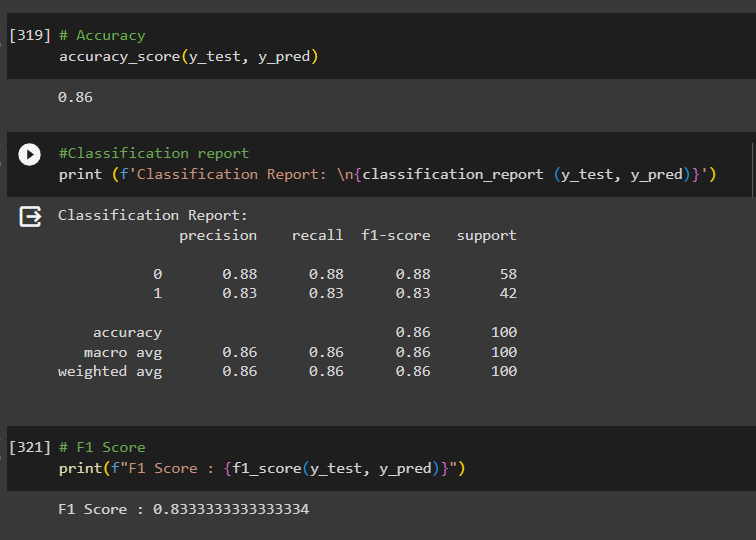
**INSTITUT TEKNOLOGI SUMATERA**

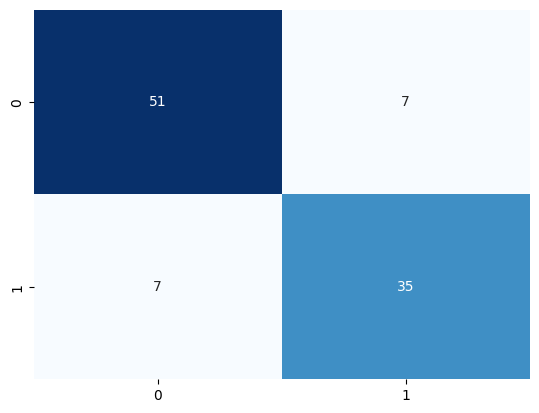
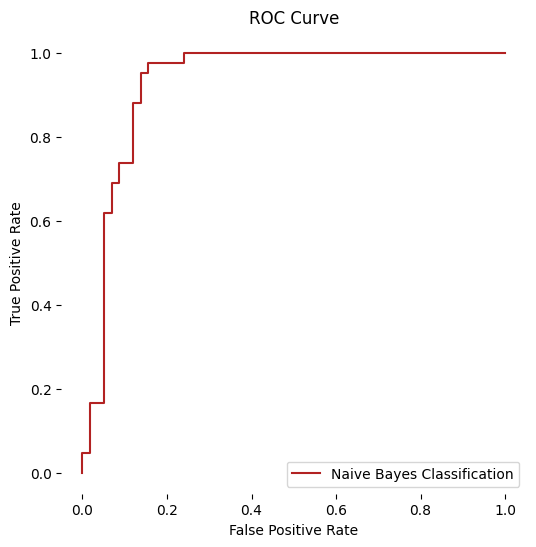
**2023**

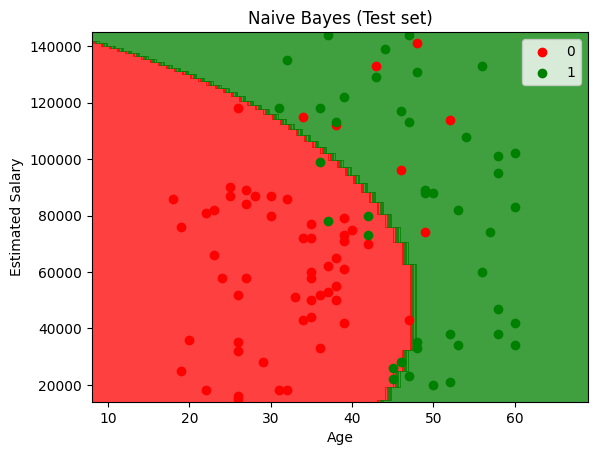
<https://colab.research.google.com/drive/1AJr9ylM7-i7VOYWe7902xGZmEhLc6FrX?usp=sharing>

# Analisis Perbandingan Hasil Evaluasi Performa Penggunaan Model Naïve Bayes dan Support Vector Machine (SVM) dengan dataset Social\_Network\_Ads.csv

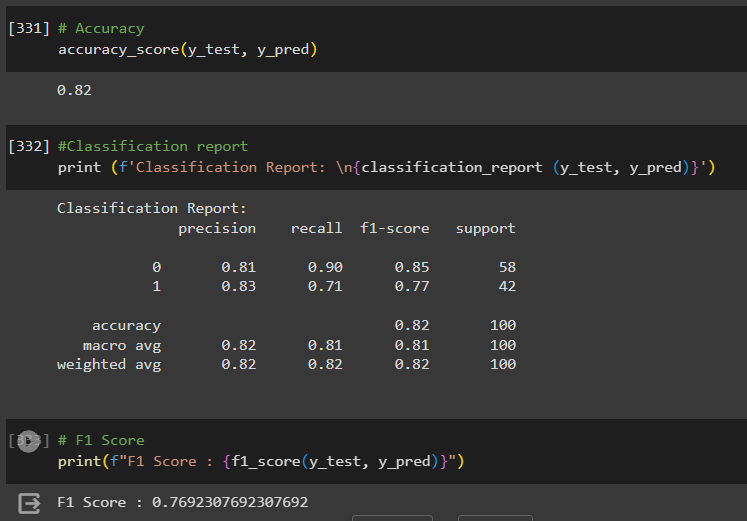
## Naïve Bayes

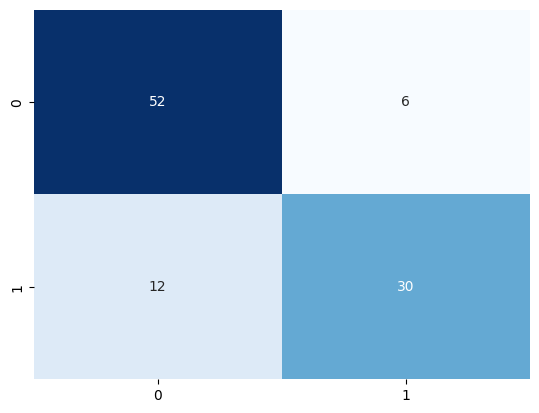
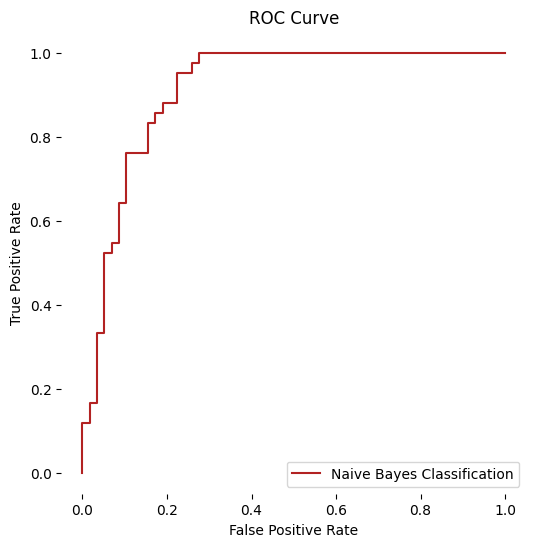


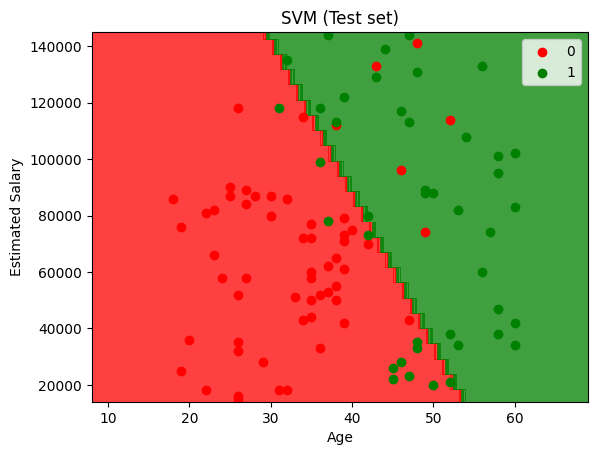
 



## Support Vector Machine (SVM)





## Analisis perbandingan

* Precision:

Naive Bayes memiliki precision yang lebih tinggi untuk kelas 0 (0.88 > 0.81).

Naive Bayes juga memiliki precision yang lebih tinggi untuk kelas 1 (0.83 > 0.83).

* Recall:

SVM memiliki recall yang lebih tinggi untuk kelas 0 (0.90 > 0.88).

Naive Bayes memiliki recall yang lebih tinggi untuk kelas 1 (0.83 > 0.71).

* F1-score:

Naive Bayes memiliki F1-score yang lebih tinggi untuk kelas 0 (0.88 > 0.85).

Naive Bayes juga memiliki F1-score yang lebih tinggi untuk kelas 1 (0.83 > 0.77).

* Accuracy:

Naive Bayes memiliki akurasi yang lebih tinggi (0.86 > 0.82).

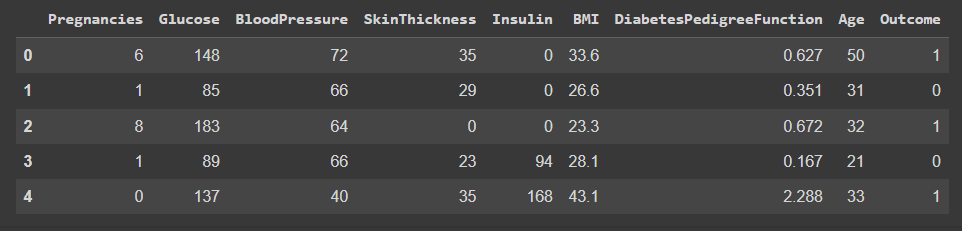
Berdasarkan perbandingan metrik-metrik di atas dan hasil visualisasi dari klasifikasi kelas, dapat disimpulkan bahwa pada dataset ini, model Naive Bayes memberikan performa yang lebih baik dibandingkan dengan SVM. Meskipun SVM memiliki recall yang lebih tinggi untuk kelas 0, Naive Bayes unggul dalam sebagian besar metrik evaluasi, termasuk precision, recall, F1-score, dan akurasi secara keseluruhan.

# B.) Analisis klasifikasi dengan model Naïve Bayes dan Support Vector Machine (SVM) 3 Dataset

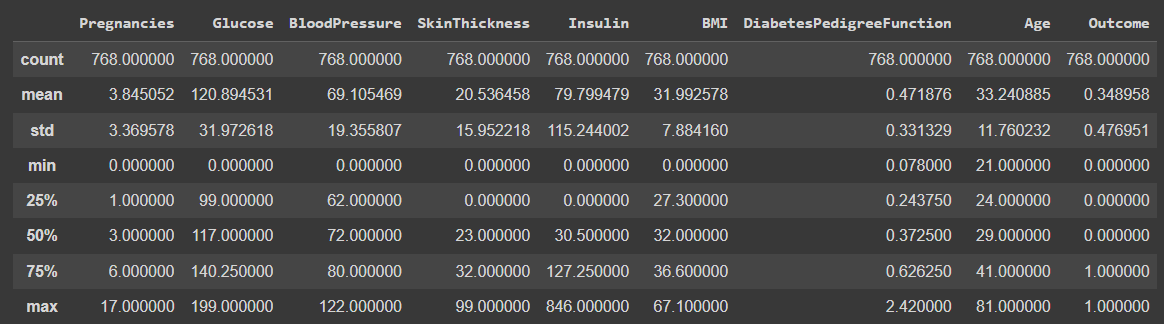
## Deskripsi dan Eksplor Data

### Dataset Diabetes\_RS1

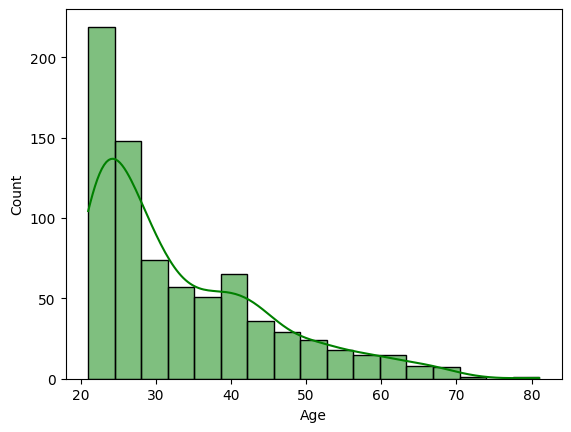
Dataset Diabetes\_RS1 adalah dataset yang berisi data mengenai hubungan antara gen RS1 dengan risiko diabetes mellitus. Dataset ini terdiri dari 2000 orang, dengan 1000 orang memiliki gen RS1 dan 1000 orang tidak memiliki gen RS1. Dataset ini memiliki 8 fitur yaitu:



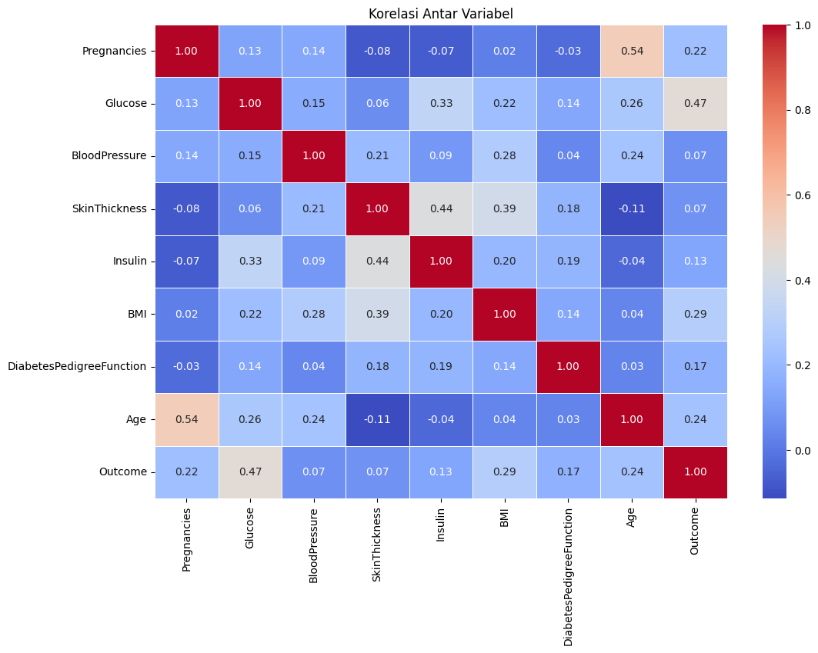
Describe dari dataset Diabetes\_RS1:



Density dari fitur ‘Age’:

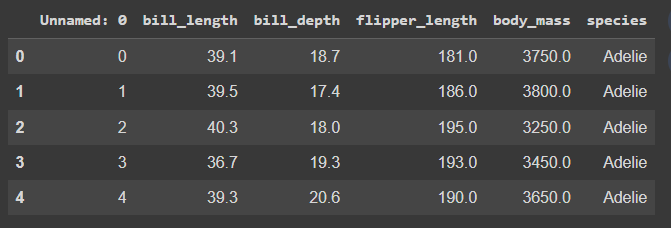


Visualisasi korelasi antar variabel/fitur:

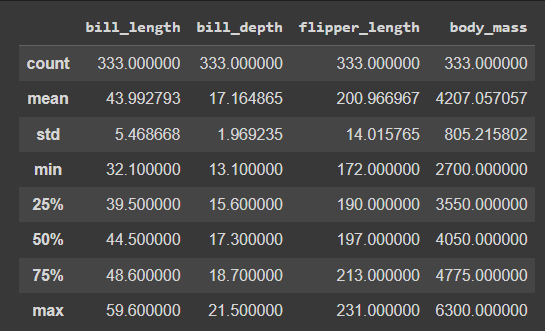


## Dataset Penguins

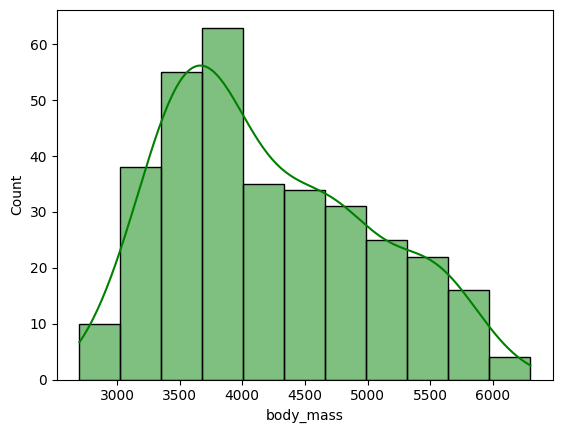
Dataset Penguins adalah dataset yang berisi data mengenai karakteristik fisik dan perilaku dari sepecies penguin. Dataset ini berisi sebanyak 6 variabel yaitu:



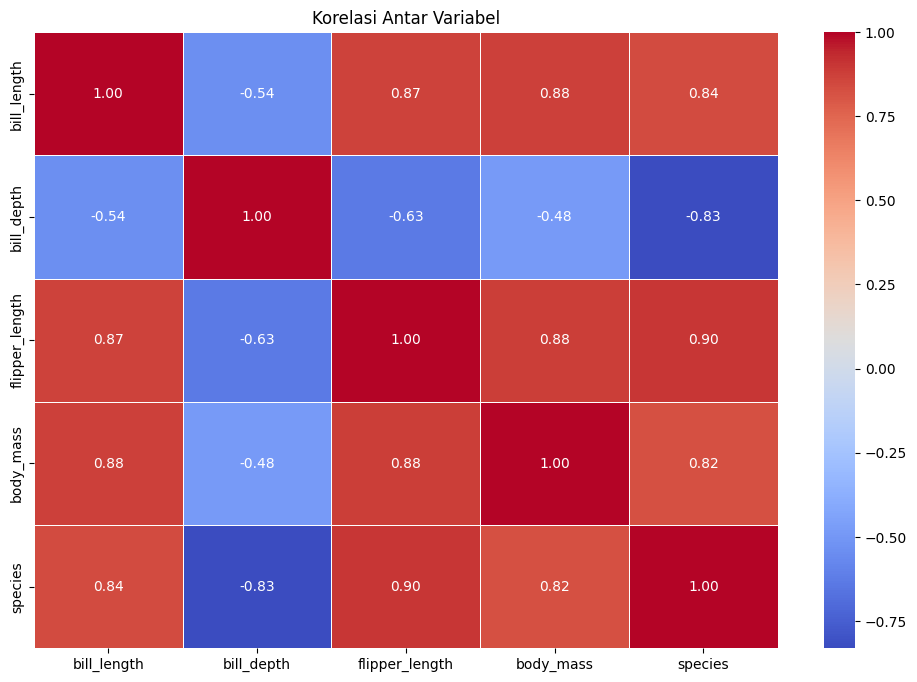
Describe dari dataset Penguins:



Density dari fitur 'body\_mass':

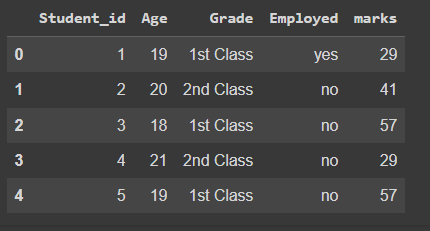


Visualisasi korelasi antar variabel/fitur:



### Dataset Student Bucketing

Dataset Student Bucketing berisi informasi tentang kinerja akademik siswa dan status pekerjaan. Ini terdiri dari lima kolom:

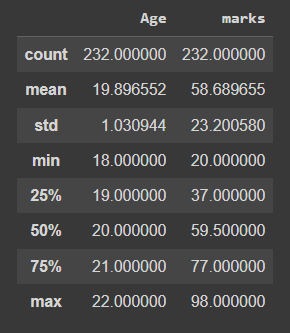


* + Student\_id: Pengidentifikasi unik untuk setiap siswa
  + Usia: Usia siswa dalam tahun
  + Nilai: Tingkat nilai siswa (misalnya, 9, 10, 11)
  + Bekerja: Apakah siswa tersebut bekerja (Ya atau Tidak)
  + nilai: Nilai siswa dalam ujian

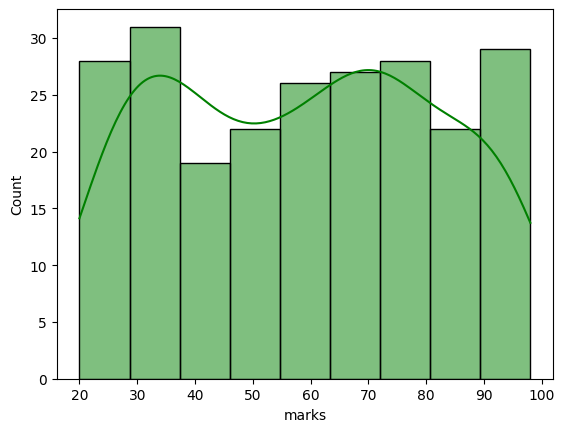
Dataset ini dapat digunakan untuk berbagai keperluan, seperti:

* + Menganalisis hubungan antara usia, tingkat kelas, status pekerjaan, dan nilai
  + Mengidentifikasi siswa yang mungkin memerlukan dukungan tambahan
  + Memprediksi kinerja siswa

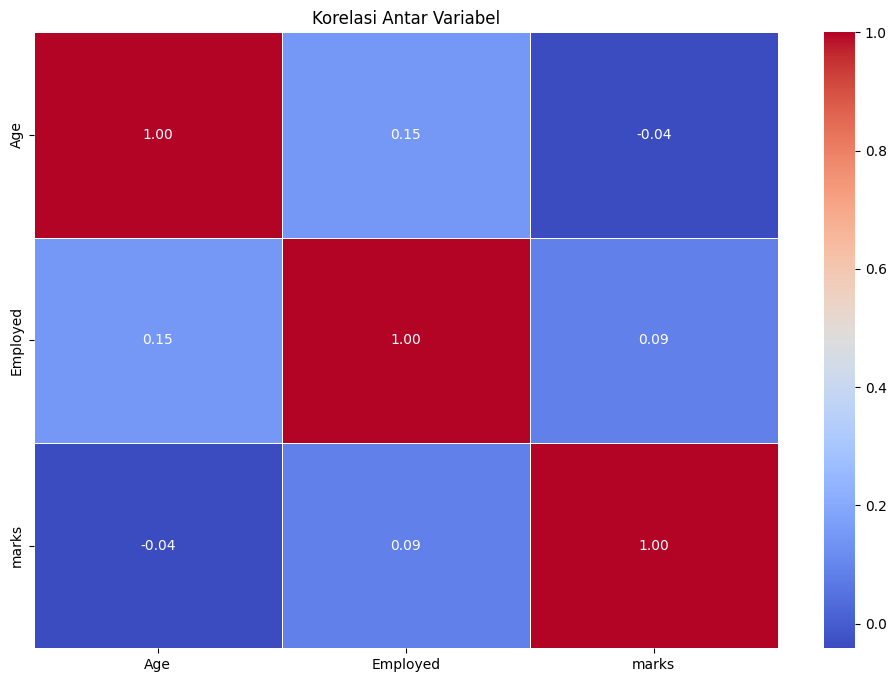
Describe dari dataset Penguins:



Density dari fitur ‘marks’:

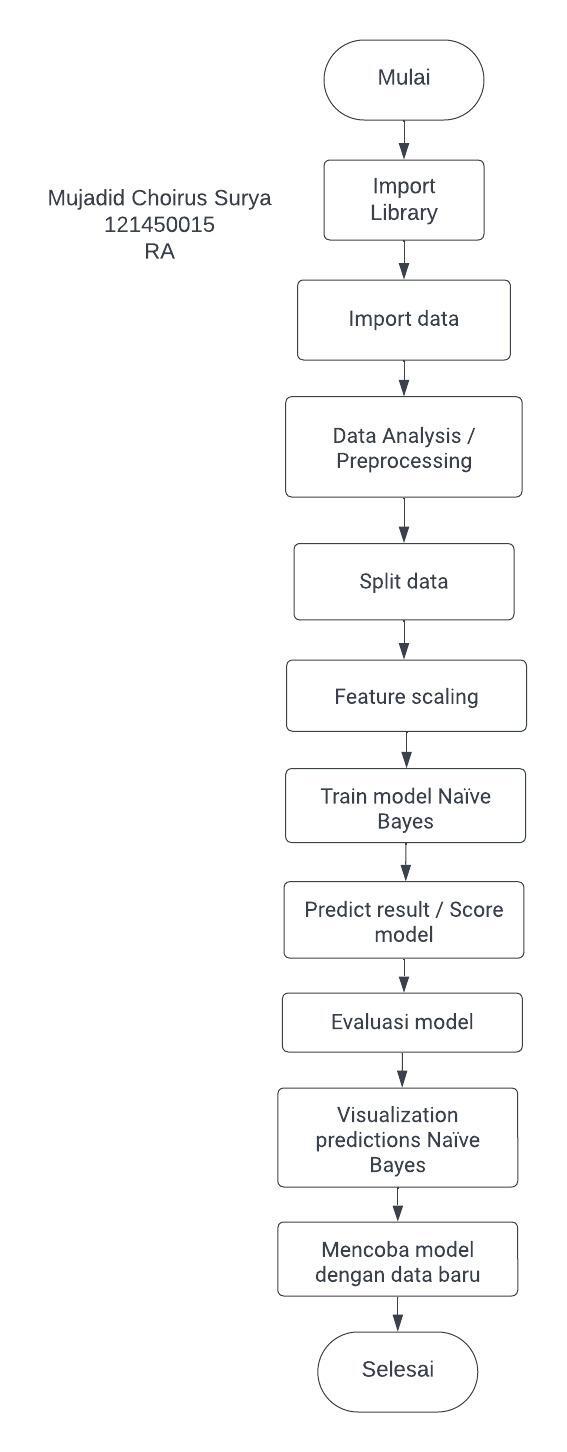


Visualisasi korelasi antar variabel/fitur:

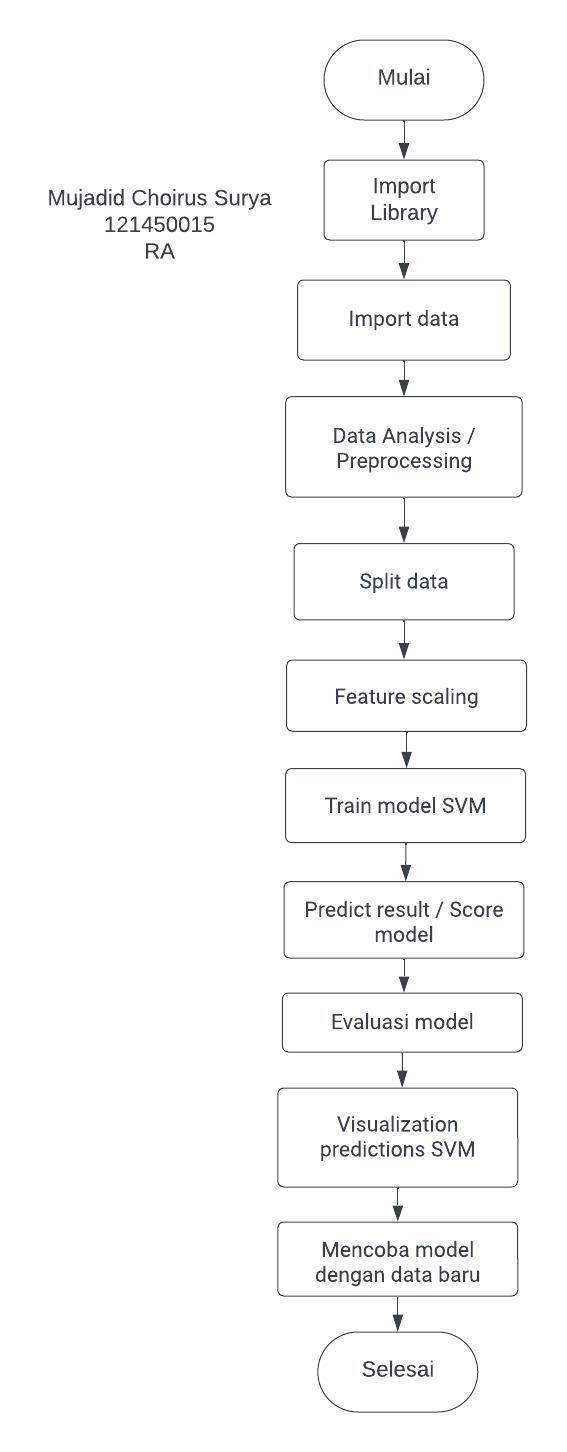


## Flowchart Praktikum

### Naïve Bayes



### Support Vector Machine(SVM)



## Analisis Hasil

### Dataset Diabetes\_RS1

|  |  |  |
| --- | --- | --- |
| Klasifikasi Naïve Bayes |  | Dalam keseluruhan, model cenderung lebih baik untuk mengenali kelas 0, sementara kinerja mengenali kelas 1 perlu perhatian lebih lanjut. Mungkin diperlukan penyesuaian atau peningkatan model, terutama untuk meningkatkan recall kelas 1. |
|  | True Positive (TP): ada 108 jumlah hasil yang benar, di mana model memprediksi kelas positif dan nilai aktual juga kelas positif.  True Negative (TN): ada 34 jumlah hasil yang benar, di mana model memprediksi kelas negatif dan nilai aktual juga kelas negatif.  False Positive (FP): ada 35 jumlah hasil yang salah, di mana model memprediksi kelas positif tetapi nilai aktual kelas negatif.  False Negative (FN): ada 15 jumlah hasil yang salah, di mana model memprediksi kelas negatif tetapi nilai aktual kelas positif. |
|  | Distribusi data yang tidak merata menunjukkan bahwa terdapat hubungan yang signifikan antara tekanan darah dan gula darah pada pasien  Pada kasus ini, algoritma naive bayes dapat memprediksi dengan cukup akurat bahwa pasien dengan kondisi terkena diabetes atau tidak |
| Klasifikasi Support Vector Machine(SVM) |  | Secara keseluruhan, model ini memiliki performa yang lebih seimbang dibandingkan dengan model sebelumnya. Namun, perlu diperhatikan bahwa recall untuk kelas 1 masih relatif rendah, dan hal ini dapat menjadi area potensial untuk peningkatan kinerja model. |
|  | True Positive (TP): ada 112 jumlah hasil yang benar, di mana model memprediksi kelas positif dan nilai aktual juga kelas positif.  True Negative (TN): ada 33 jumlah hasil yang benar, di mana model memprediksi kelas negatif dan nilai aktual juga kelas negatif.  False Positive (FP): ada 36 jumlah hasil yang salah, di mana model memprediksi kelas positif tetapi nilai aktual kelas negatif.  False Negative (FN): ada 11 jumlah hasil yang salah, di mana model memprediksi kelas negatif tetapi nilai aktual kelas positif. |
|  | Pada gambar, hyperplane tersebut terlihat cukup baik dalam memisahkan data kelas 0 dan kelas 1.  Hal ini menunjukkan bahwa algoritma SVM dapat digunakan untuk membedakan antara pasien dengan kondisi terkena diabetes atau tidak dengan akurasi yang cukup. |

### Dataset Penguins

|  |  |  |
| --- | --- | --- |
| Klasifikasi Naïve Bayes |  | Dengan melihat keseluruhan hasil, dapat disimpulkan bahwa model memiliki performa yang sangat baik dalam mengklasifikasikan data, dengan nilai akurasi, precision, recall, dan F1-score mencapai 97%. Model ini dapat dianggap sebagai model yang handal dan dapat dipercaya dalam tugas klasifikasi yang diberikan. |
|  | True Positive (TP): ada 35 jumlah hasil yang benar, di mana model memprediksi kelas positif dan nilai aktual juga kelas positif.  True Negative (TN): ada 30 jumlah hasil yang benar, di mana model memprediksi kelas negatif dan nilai aktual juga kelas negatif.  False Positive (FP): ada 1 jumlah hasil yang salah, di mana model memprediksi kelas positif tetapi nilai aktual kelas negatif.  False Negative (FN): ada 1 jumlah hasil yang salah, di mana model memprediksi kelas negatif tetapi nilai aktual kelas positif. |
|  | Pada kasus ini, algoritma naive bayes dapat memprediksi dengan cukup akurat dengan bisa membedakan kelas 1 dan kelas 0 berdasarkan bill\_length dan body\_mass untuk menentukan species pinguin. |
| Klasifikasi Support Vector Machine(SVM) |  | Sama seperti model sebelumnya, dengan melihat keseluruhan hasil, dapat disimpulkan bahwa model memiliki performa yang sangat baik dalam mengklasifikasikan data, dengan nilai akurasi, precision, recall, dan F1-score mencapai 97%. Model ini dapat dianggap sebagai model yang handal dan dapat dipercaya dalam tugas klasifikasi yang diberikan. |
|  | True Positive (TP): ada 35 jumlah hasil yang benar, di mana model memprediksi kelas positif dan nilai aktual juga kelas positif.  True Negative (TN): ada 30 jumlah hasil yang benar, di mana model memprediksi kelas negatif dan nilai aktual juga kelas negatif.  False Positive (FP): ada 1 jumlah hasil yang salah, di mana model memprediksi kelas positif tetapi nilai aktual kelas negatif.  False Negative (FN): ada 1 jumlah hasil yang salah, di mana model memprediksi kelas negatif tetapi nilai aktual kelas positif. |
|  | Pada kasus ini, algoritma SVM dapat memprediksi dengan cukup akurat dengan bisa membedakan kelas 1 dan kelas 0 berdasarkan bill\_length dan body\_mass untuk menentukan species pinguin. |

### Dataset Student Bucketing

|  |  |  |
| --- | --- | --- |
| Klasifikasi Naïve Bayes |  | Secara keseluruhan, model cenderung lebih baik dalam memprediksi kelas 0 (yang memiliki recall dan F1-score yang lebih tinggi), sementara performa pada kelas 1 relatif lebih rendah, terutama dalam hal recall dan F1-score. Ada potensi untuk meningkatkan performa model terutama pada kelas 1. |
|  | True Positive (TP): ada 29 jumlah hasil yang benar, di mana model memprediksi kelas positif dan nilai aktual juga kelas positif.  True Negative (TN): ada 6 jumlah hasil yang benar, di mana model memprediksi kelas negatif dan nilai aktual juga kelas negatif.  False Positive (FP): ada 18 jumlah hasil yang salah, di mana model memprediksi kelas positif tetapi nilai aktual kelas negatif.  False Negative (FN): ada 5 jumlah hasil yang salah, di mana model memprediksi kelas negatif tetapi nilai aktual kelas positif. |
|  | Pada kasus ini, algoritma naive bayes model cenderung lebih baik dalam memprediksi kelas 0 namun buruk saat memprediksi kelas 1 jika dilihat dari visualisasinya |
| Klasifikasi Support Vector Machine(SVM) |  | Model cenderung memiliki kinerja yang baik dalam memprediksi kelas 0, namun sangat buruk dalam memprediksi kelas 1. Kemampuan model untuk mengenali instance kelas 1 sangat rendah, tercermin dari nilai recall dan F1-score yang nol untuk kelas tersebut. Dalam konteks ini, perlu mempertimbangkan peningkatan atau penyesuaian model untuk meningkatkan kemampuan dalam memprediksi kelas 1. |
|  | True Positive (TP): ada 34 jumlah hasil yang benar, di mana model memprediksi kelas positif dan nilai aktual juga kelas positif.  True Negative (TN): ada 0 jumlah hasil yang benar, di mana model memprediksi kelas negatif dan nilai aktual juga kelas negatif.  False Positive (FP): ada 24 jumlah hasil yang salah, di mana model memprediksi kelas positif tetapi nilai aktual kelas negatif.  False Negative (FN): ada 0 jumlah hasil yang salah, di mana model memprediksi kelas negatif tetapi nilai aktual kelas positif. |
|  | Pada kasus ini, algoritma SVM model cenderung lebih baik dalam memprediksi kelas 0 namun sangat buruk saat memprediksi kelas 1 jika dilihat dari visualisasinya |