

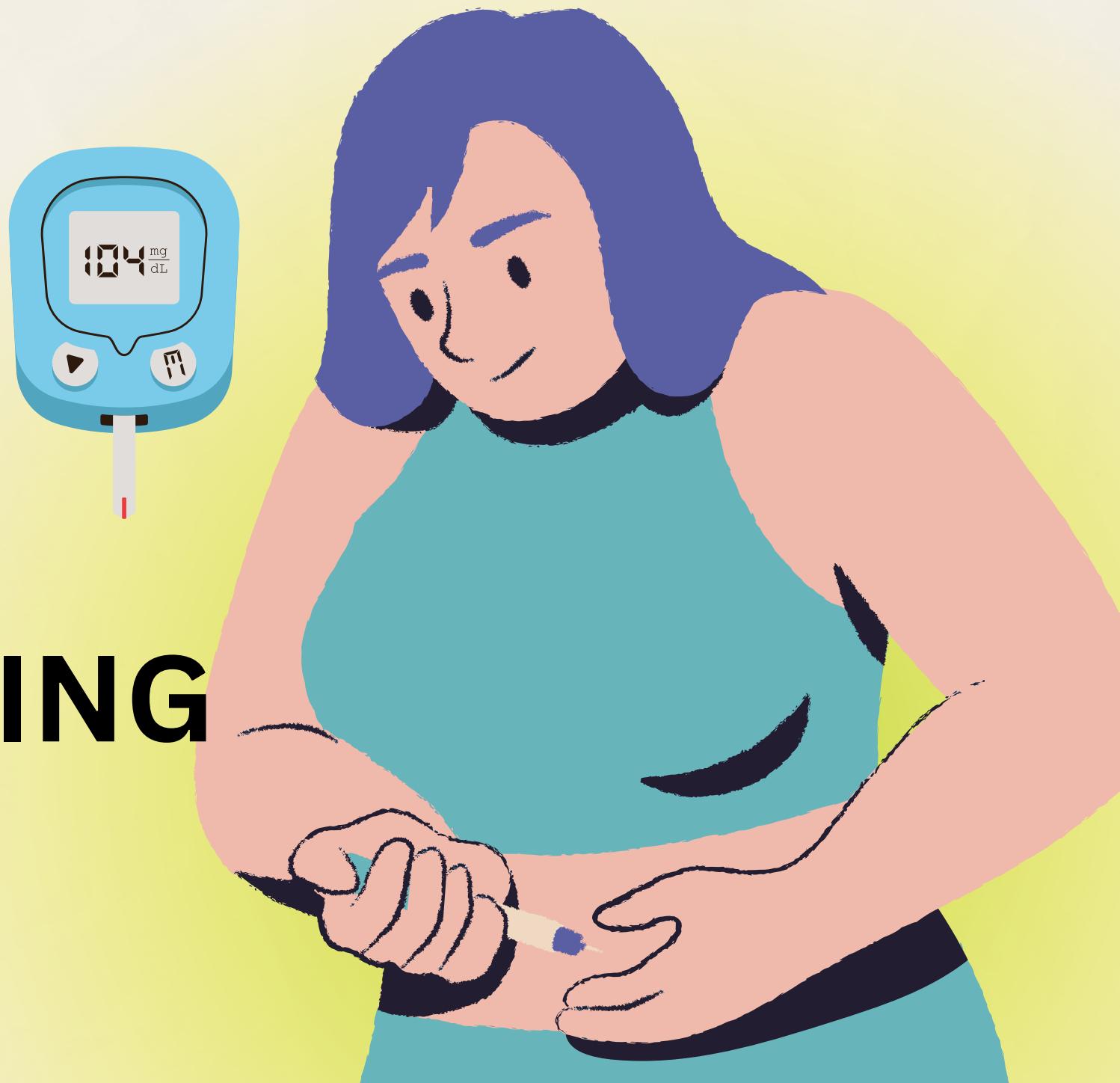
TU 155

DIABETES DATA SET

APPLY BY

NAÏVE BAYES

PROJECT MACHINE LEARNING

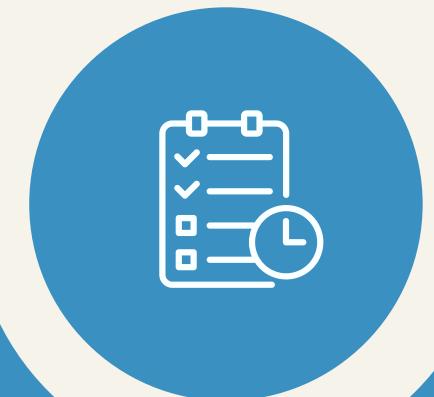


OVERVIEW

Choose
data



Clean data



Train model



Analyze
&
Conclusion



ทำไมถึงต้องเป็น

DIABETES DATA SET



CLEAN DATA



หา MISSING VALUE

ตัดค่าที่เป็นไปไม่ได้ออก

ตัด OUTLIER



BEFORE

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | Diabetes | PedigreeFunction | Age | Outcome |
|-------|-------------|------------|---------------|---------------|------------|------------|----------|------------------|------------|------------|
| count | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | 768.000000 | | 768.000000 | 768.000000 | 768.000000 |
| mean | 3.845052 | 120.894531 | 69.105469 | 20.536458 | 79.799479 | 31.992578 | | 0.471876 | 33.240885 | 0.348958 |
| std | 3.369578 | 31.972618 | 19.355807 | 15.952218 | 115.244002 | 7.884160 | | 0.331329 | 11.760232 | 0.476951 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | | 0.078000 | 21.000000 | 0.000000 |
| 25% | 1.000000 | 99.000000 | 62.000000 | 0.000000 | 0.000000 | 27.300000 | | 0.243750 | 24.000000 | 0.000000 |
| 50% | 3.000000 | 117.000000 | 72.000000 | 23.000000 | 30.500000 | 32.000000 | | 0.372500 | 29.000000 | 0.000000 |
| 75% | 6.000000 | 140.250000 | 80.000000 | 32.000000 | 127.250000 | 36.600000 | | 0.626250 | 41.000000 | 1.000000 |
| max | 17.000000 | 199.000000 | 122.000000 | 99.000000 | 846.000000 | 67.100000 | | 2.420000 | 81.000000 | 1.000000 |

AFTER

| | Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | Diabetes | PedigreeFunction | Age | Outcome |
|-------|-------------|------------|---------------|---------------|------------|------------|----------|------------------|------------|------------|
| count | 392.000000 | 392.000000 | 392.000000 | 392.000000 | 392.000000 | 392.000000 | | 392.000000 | 392.000000 | 392.000000 |
| mean | 3.301020 | 122.627551 | 70.663265 | 29.145408 | 156.056122 | 33.086224 | | 0.523046 | 30.864796 | 0.331633 |
| std | 3.211424 | 30.860781 | 12.496092 | 10.516424 | 118.841690 | 7.027659 | | 0.345488 | 10.200777 | 0.471401 |
| min | 0.000000 | 56.000000 | 24.000000 | 7.000000 | 14.000000 | 18.200000 | | 0.085000 | 21.000000 | 0.000000 |
| 25% | 1.000000 | 99.000000 | 62.000000 | 21.000000 | 76.750000 | 28.400000 | | 0.269750 | 23.000000 | 0.000000 |
| 50% | 2.000000 | 119.000000 | 70.000000 | 29.000000 | 125.500000 | 33.200000 | | 0.449500 | 27.000000 | 0.000000 |
| 75% | 5.000000 | 143.000000 | 78.000000 | 37.000000 | 190.000000 | 37.100000 | | 0.687000 | 36.000000 | 1.000000 |
| max | 17.000000 | 198.000000 | 110.000000 | 63.000000 | 846.000000 | 67.100000 | | 2.420000 | 81.000000 | 1.000000 |

ATTRIBUTES

BloodPressure

Pregnancies

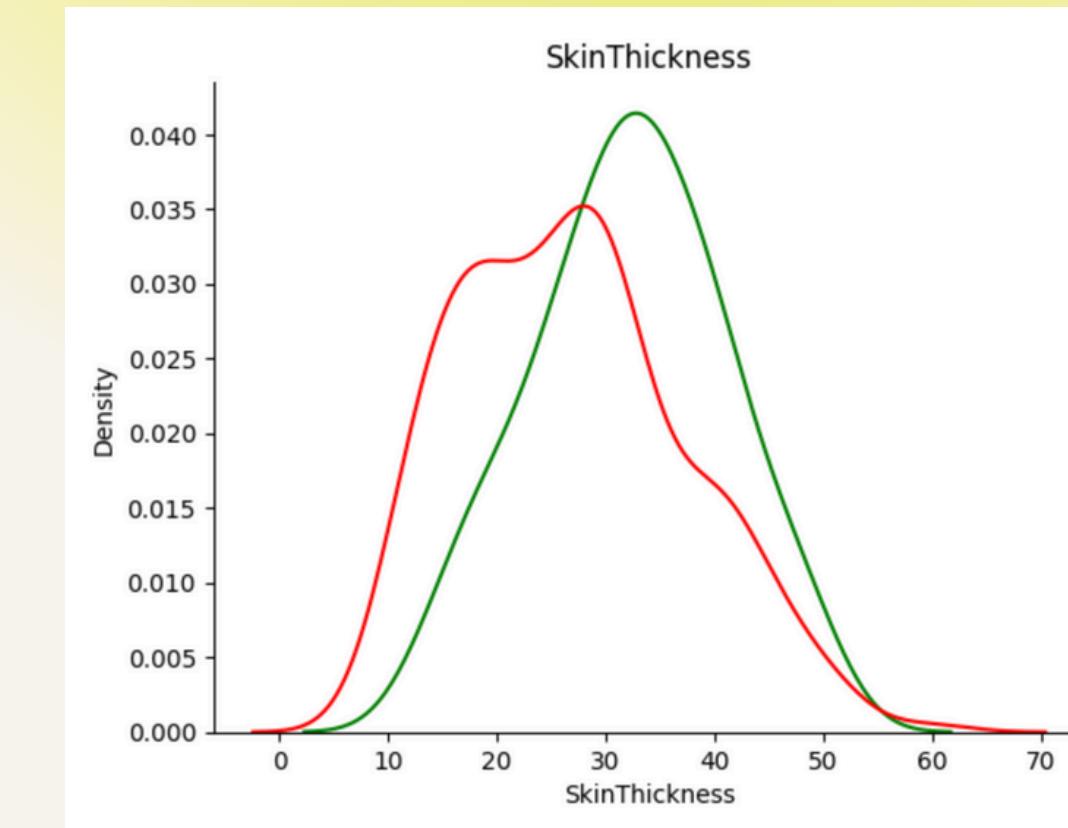
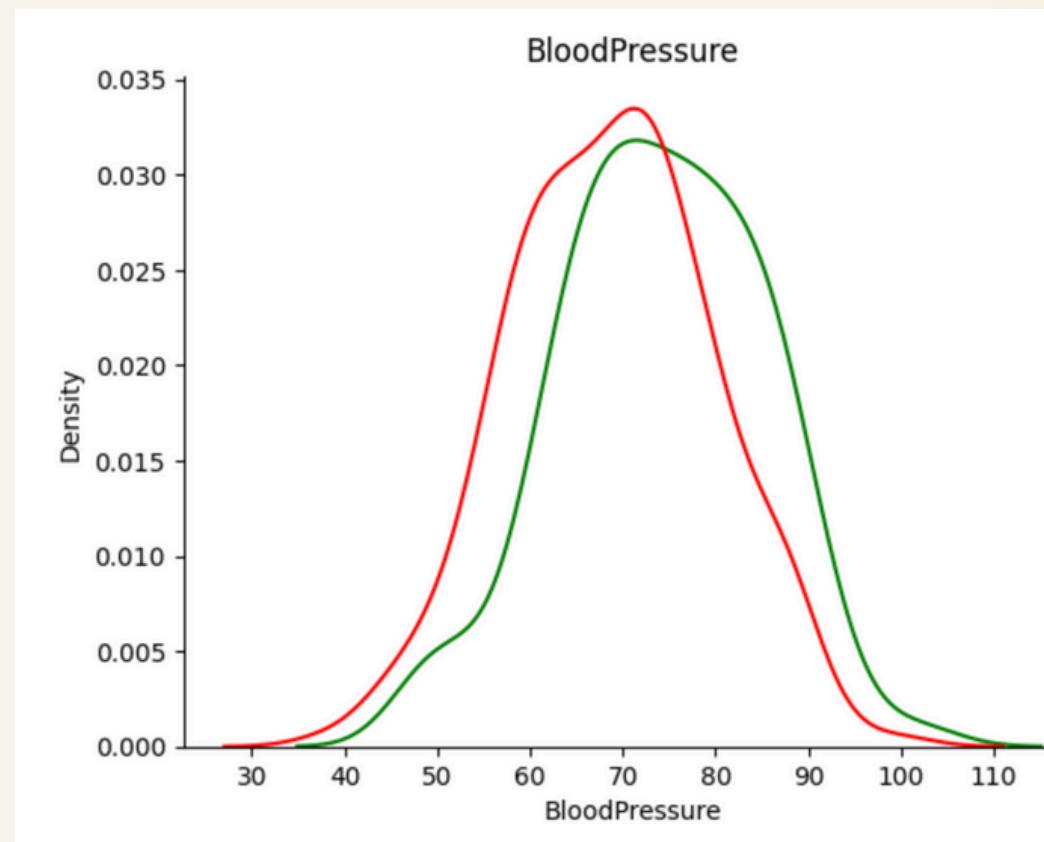
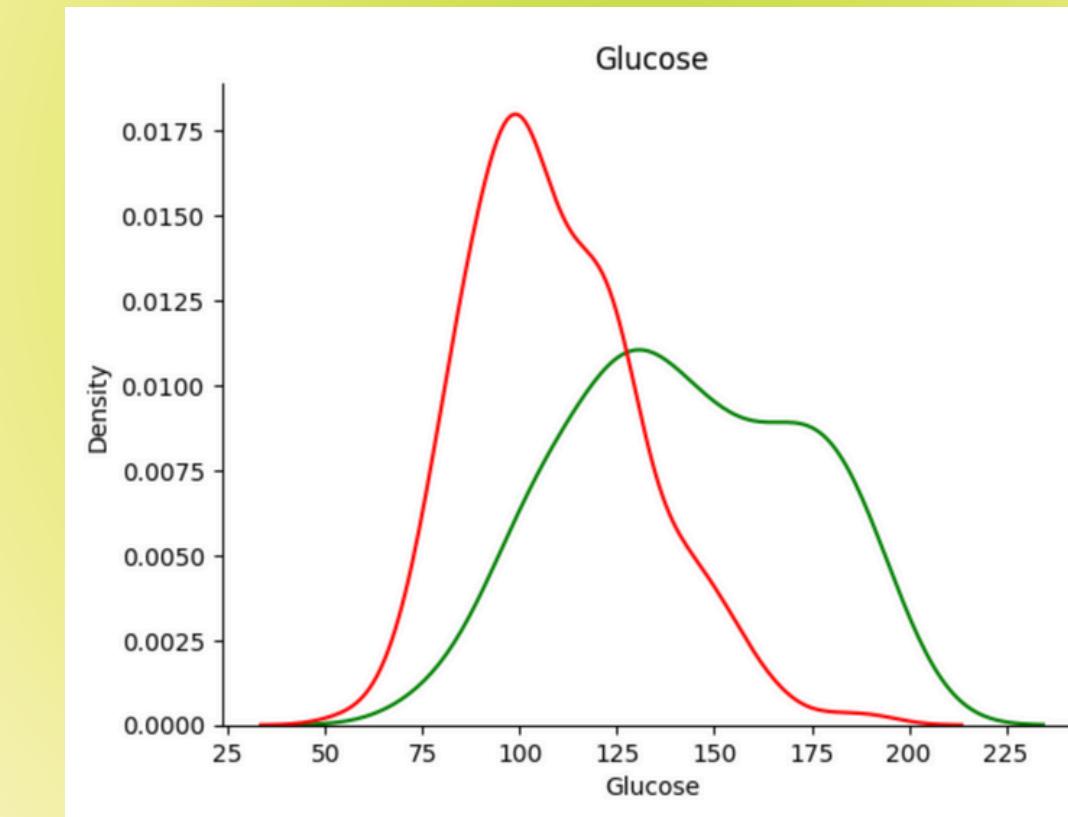
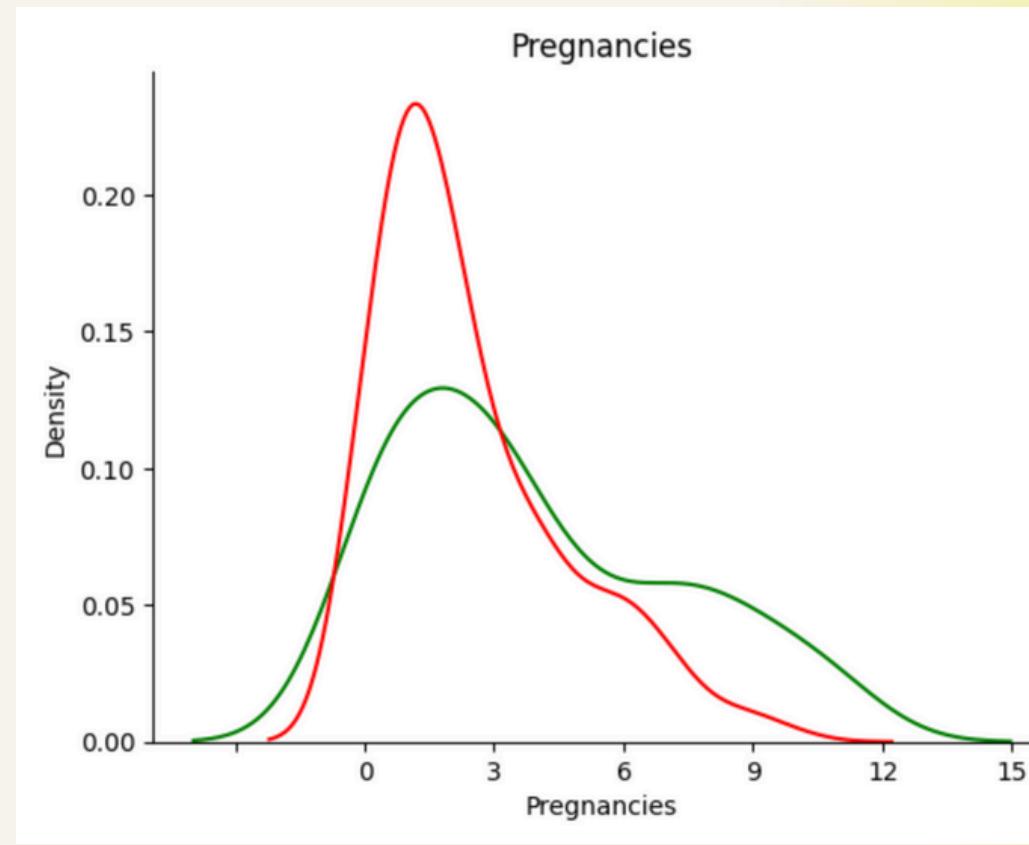
Age

Insulin

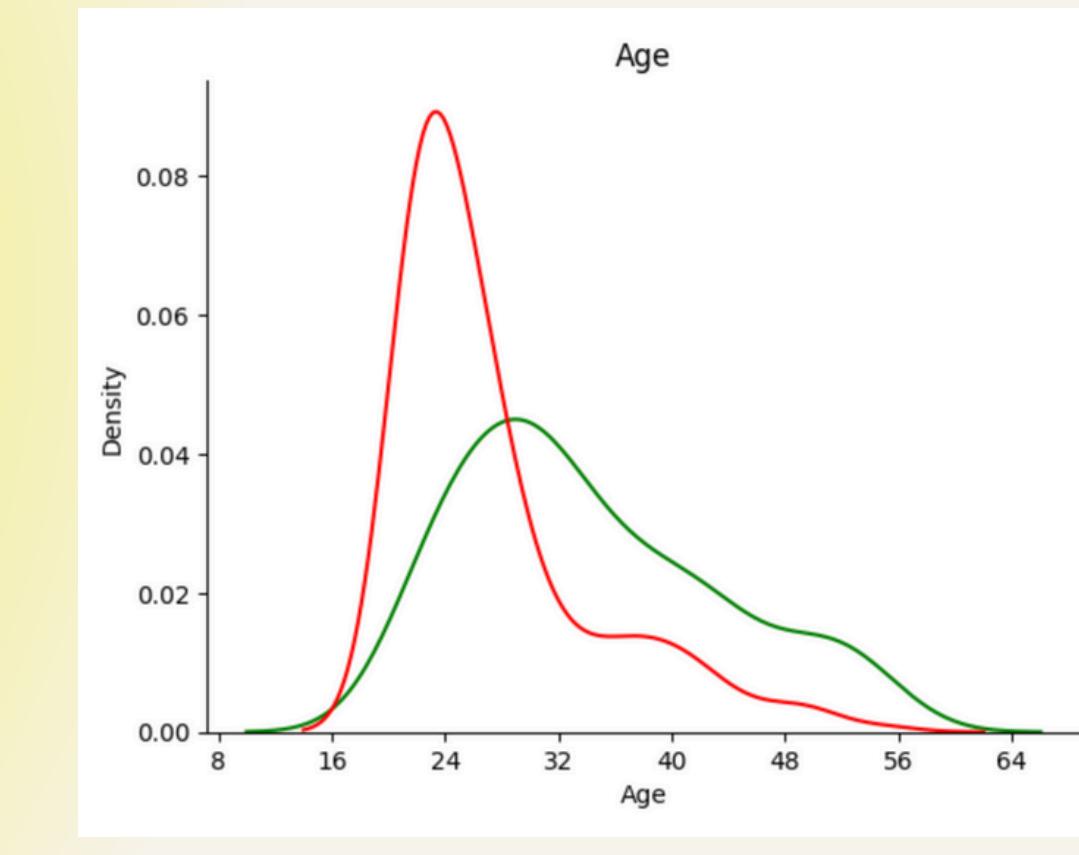
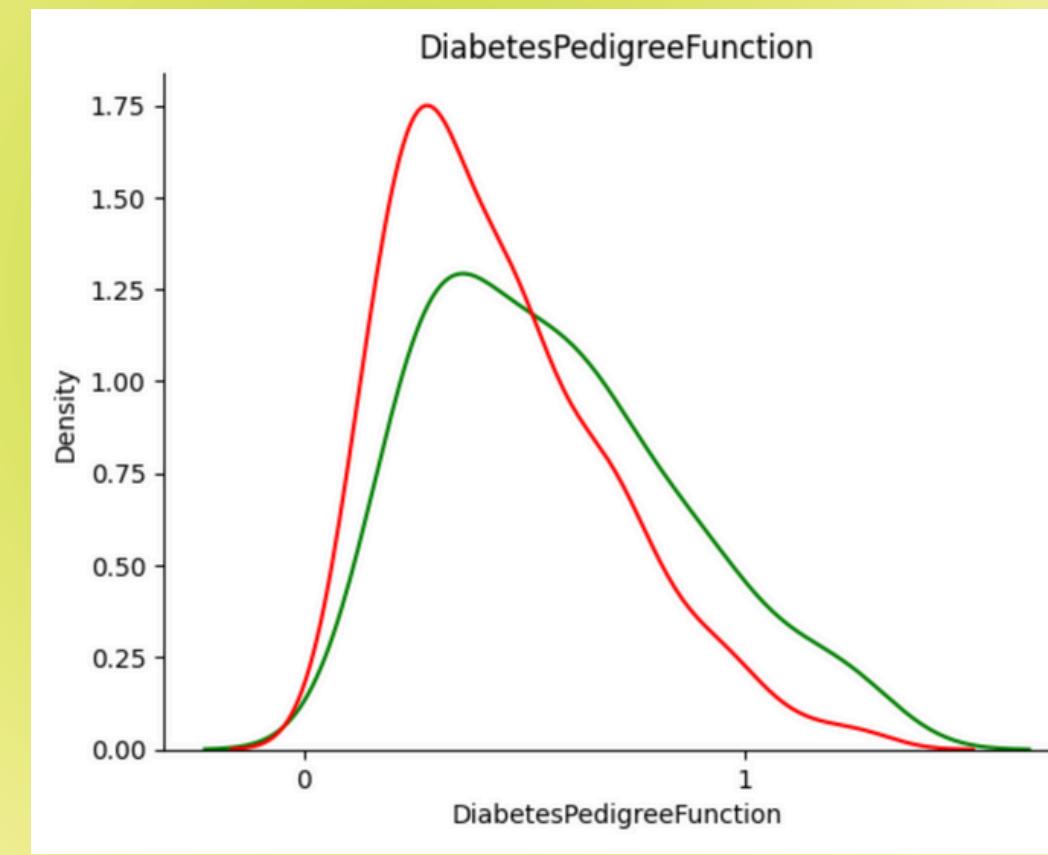
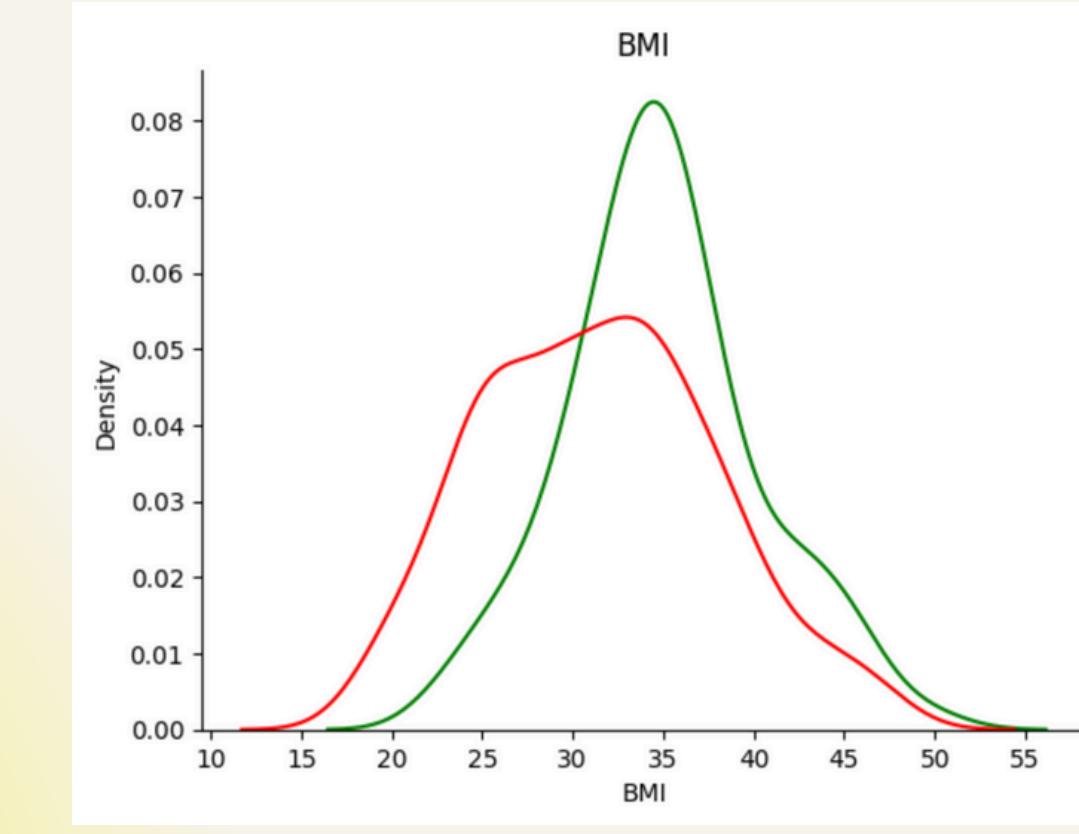
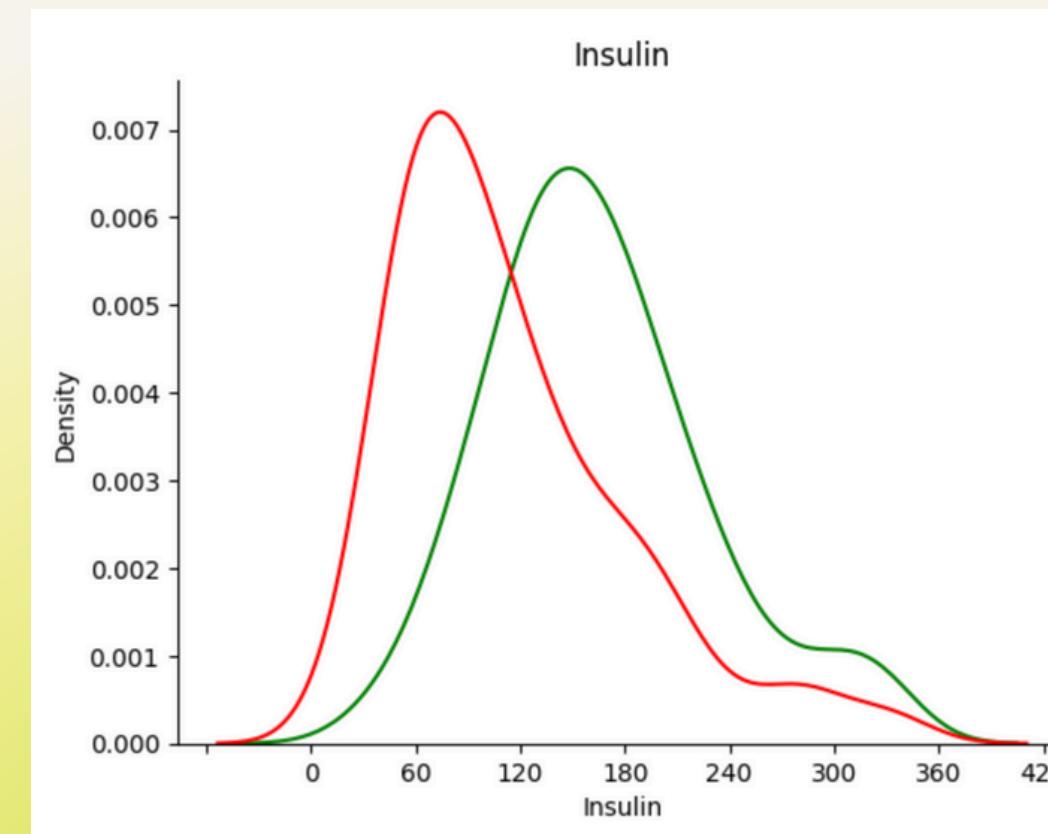
BMI Glucose SkinThickness
DiabetesPedigree Function



สร้างกราฟเพื่อดูการกระจายตัวของข้อมูล

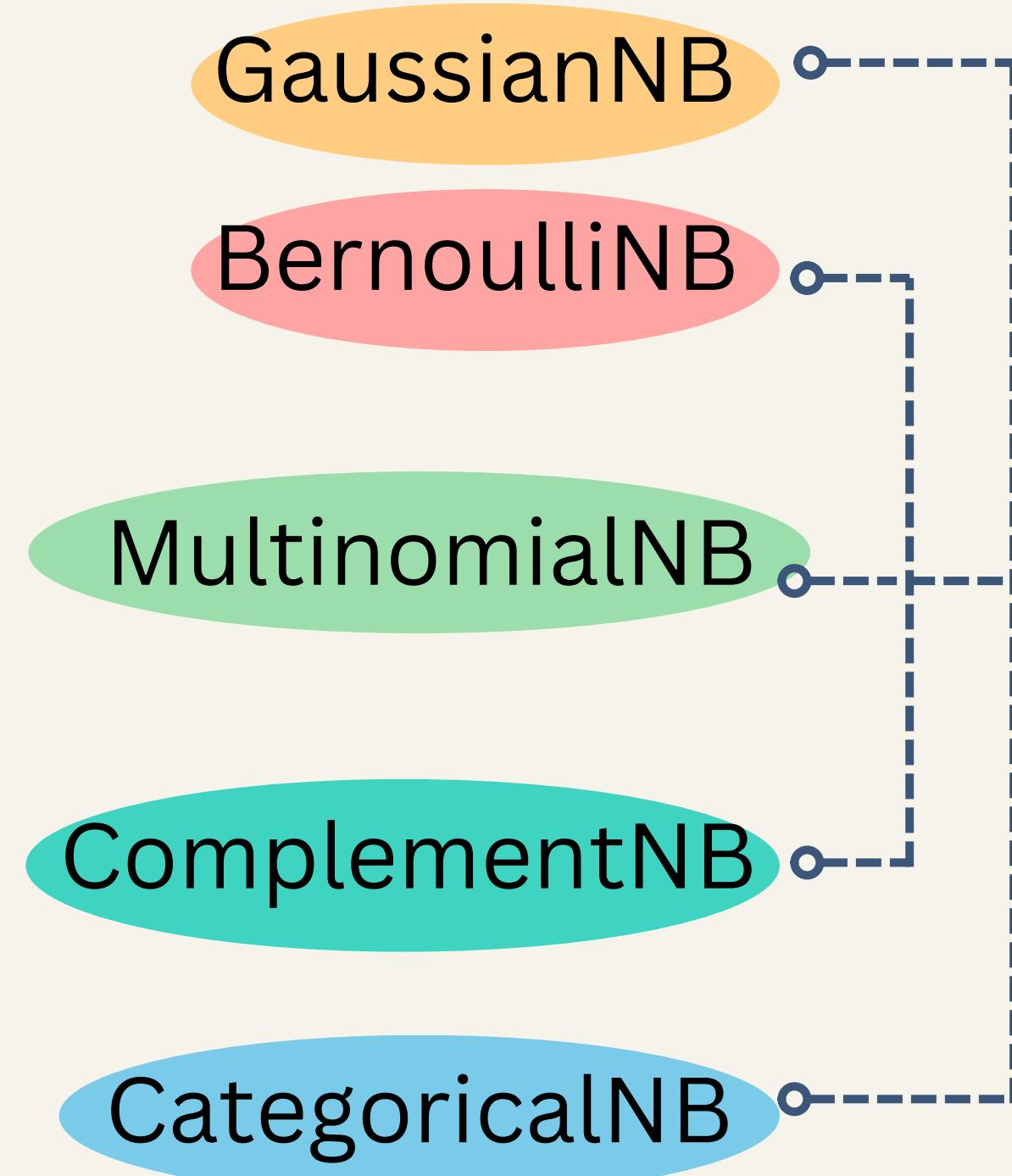


สร้างกราฟเพื่อดูการกระจายตัวของข้อมูล



TRAINING MODEL

NAÏVE BAYES



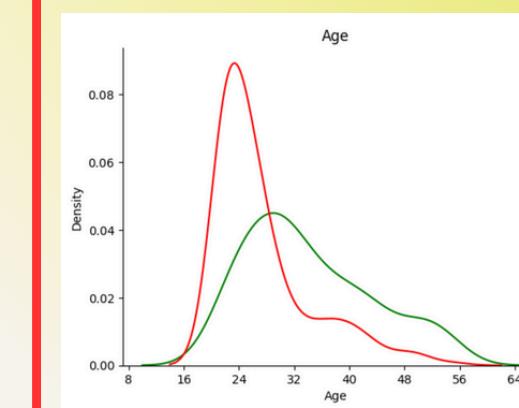
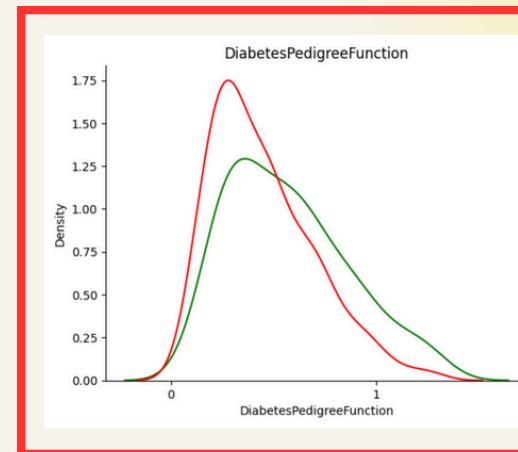
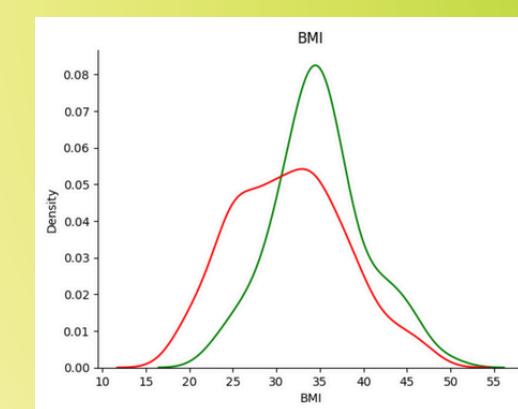
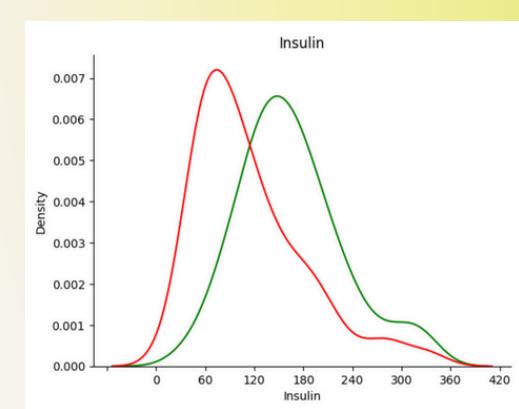
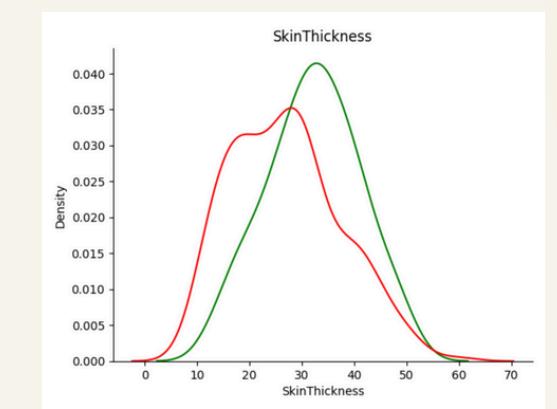
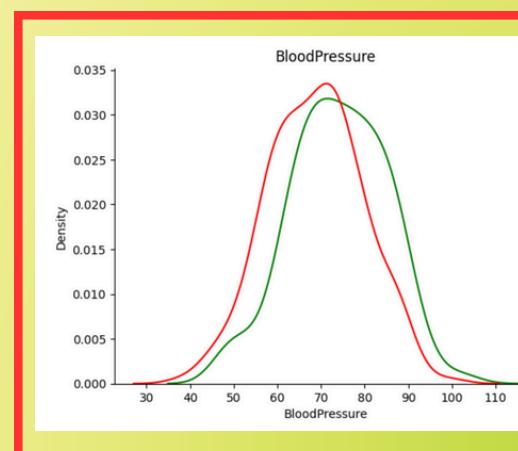
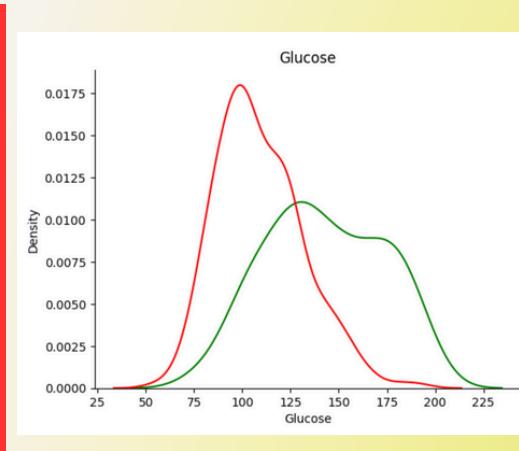
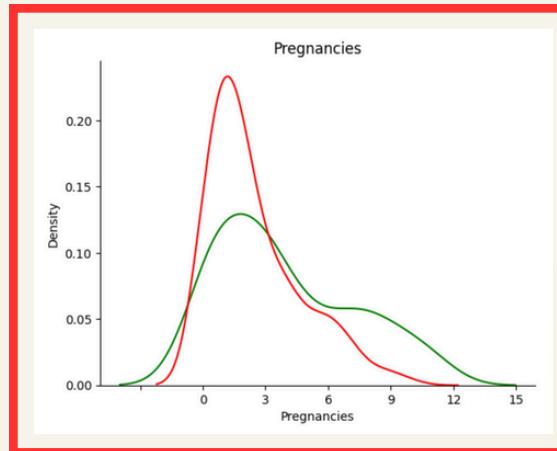
OTHER CLASS

- Random Forest
- K-Nearest Neighbors
- SVM
- Logistic Regression
- Decision Tree

ใช้คลาส GaussianNB ในการเรนโมเดล

STEP 1 เริ่มต้นจากการใช้ทั้งหมด 8 attributes

STEP 2 ตัด attribute ที่ละตัวที่ไม่สำคัญ 3 ลำดับแรก



สัดส่วนที่ทับซ้อนกันเรียงจากมากไปน้อย

BloodPressure: 0.9305

DiabetesPedigreeFunction: 0.8856

Pregnancies: 0.8833

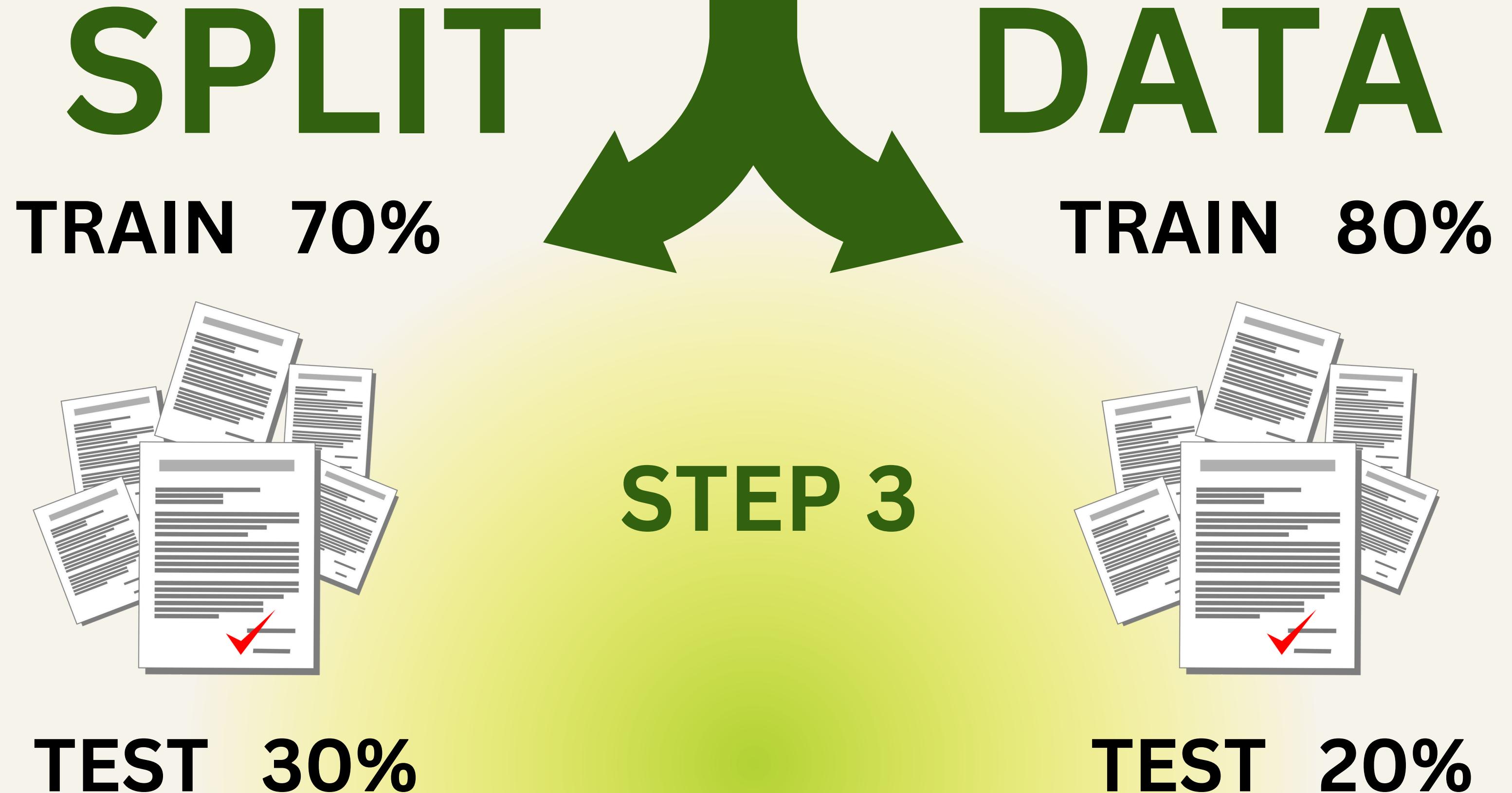
BMI: 0.8468

Glucose: 0.7293

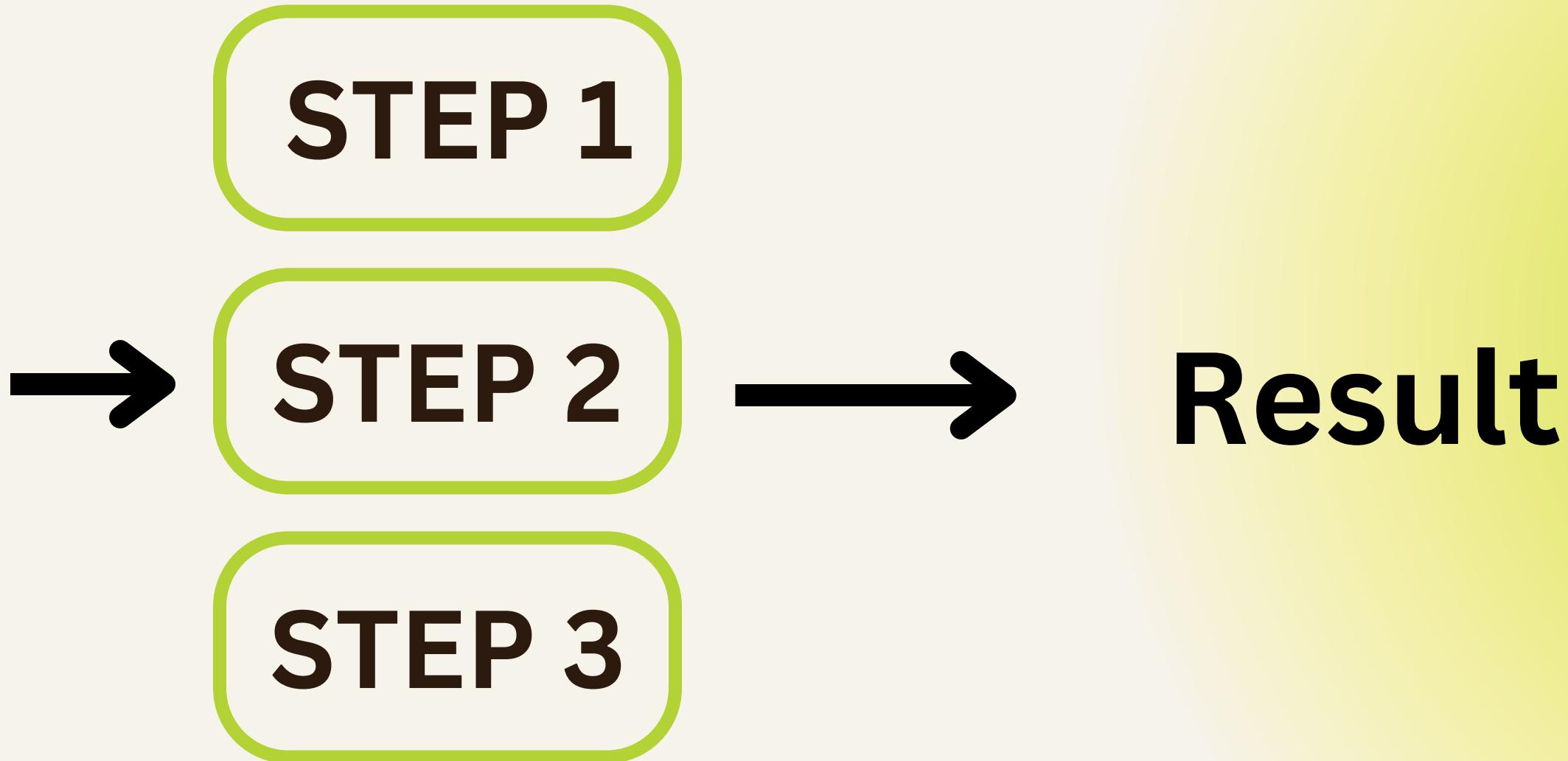
SkinThickness: 0.6551

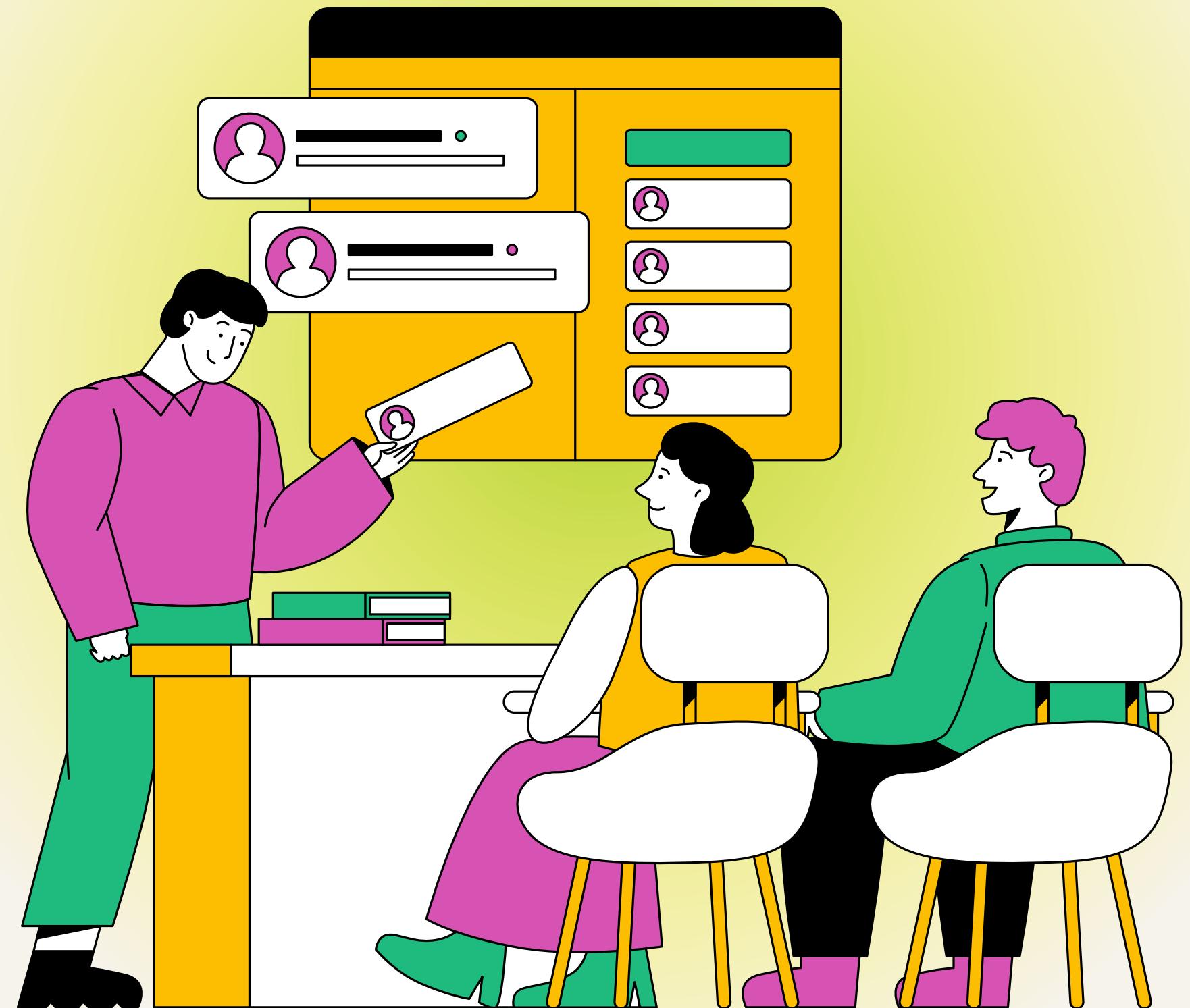
Insulin: 0.6545

Age: 0.6046



Bernoulli**NB**
Multinomial**NB**
Complement**NB**
Categorical**NB**

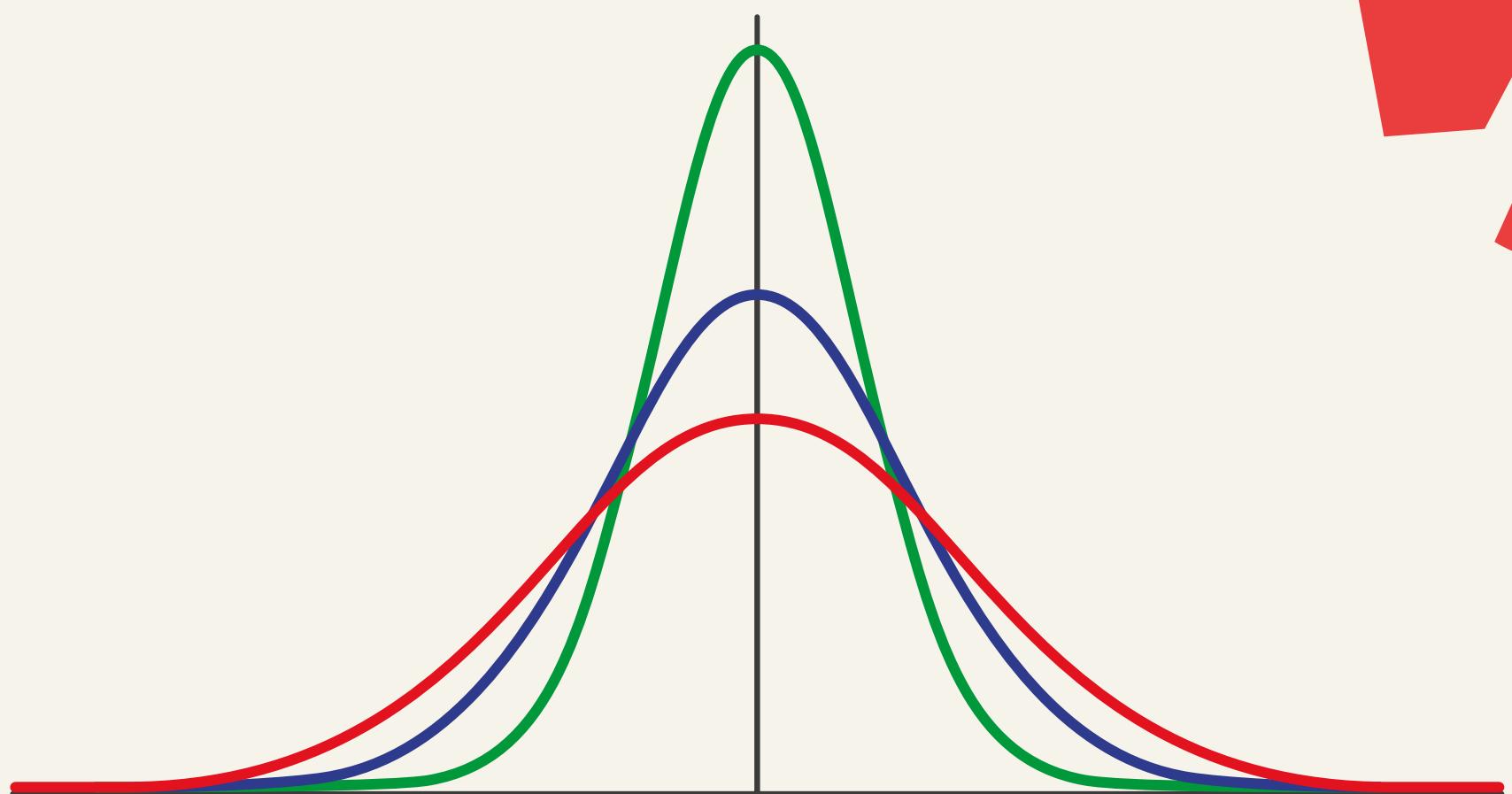




**ANALYZE
&
CONCLUSION**

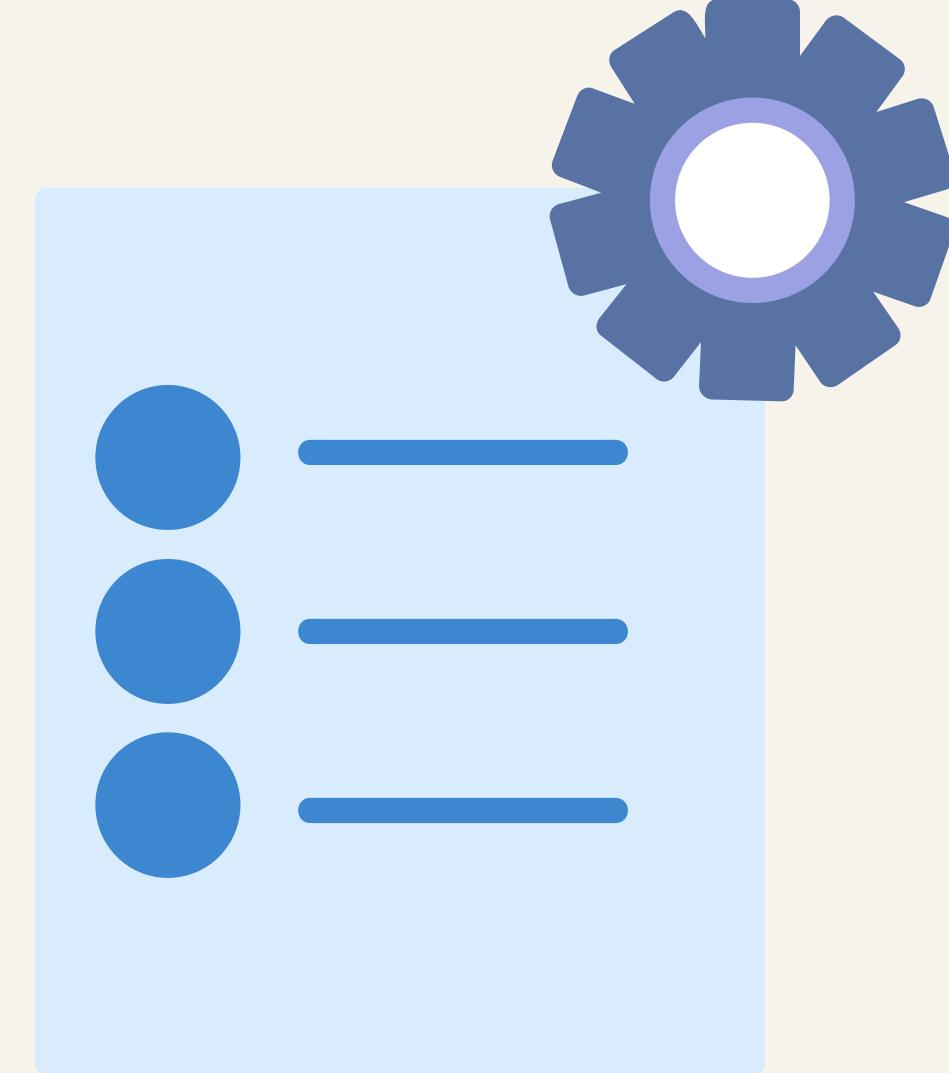
สรุปค่า ACCURACY ของแต่ละ MODEL ในแต่ละคลาสของ NAÏVE BAYES

8
Atributes

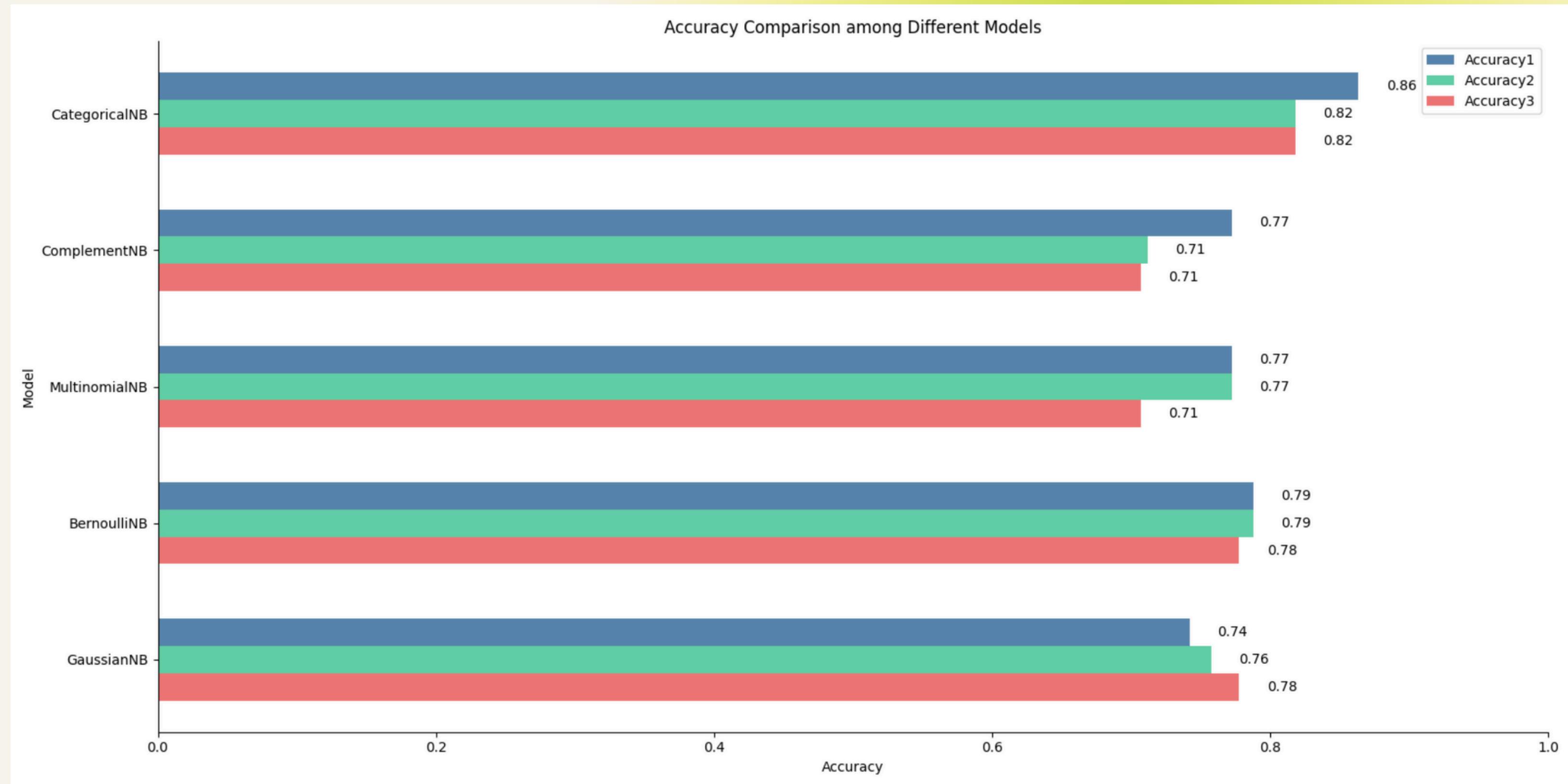


vs

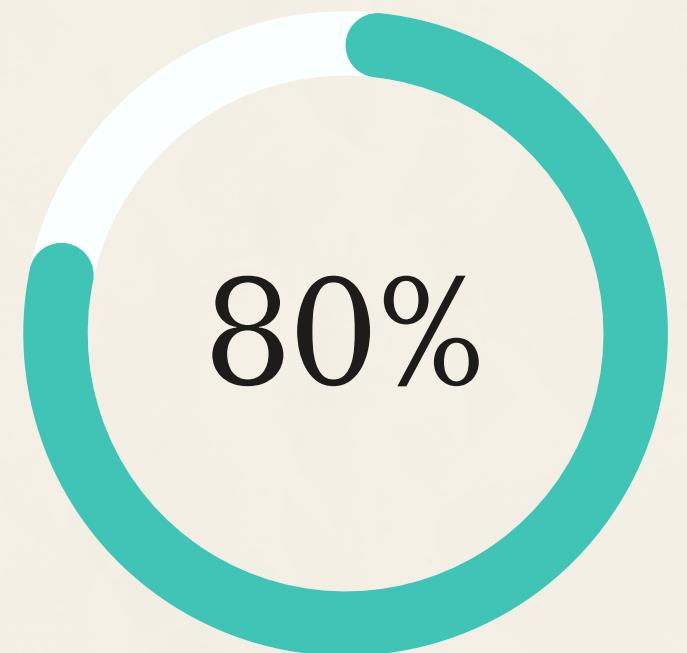
7,6,5
Atributes



สรุปค่า ACCURACY ของแต่ละ MODEL ในแต่ละคลาสของ NAÏVE BAYES



SPLIT DATA

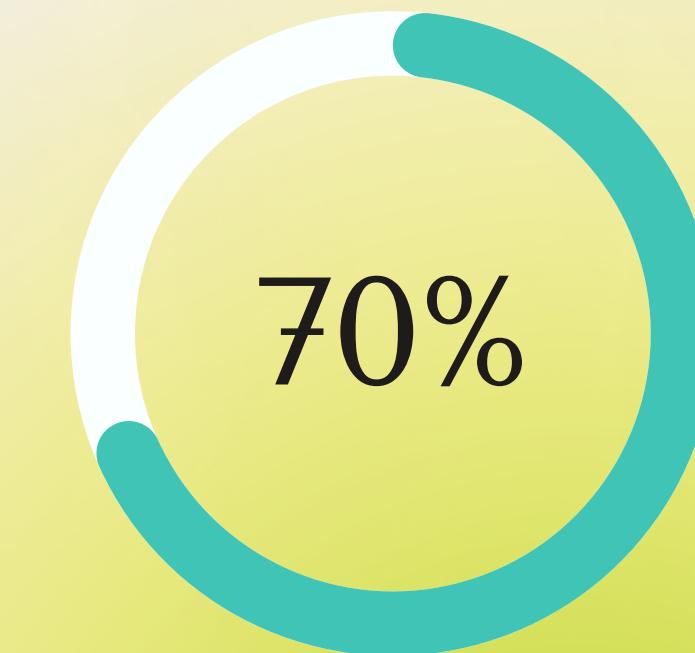


TRAINING MODEL

80:20



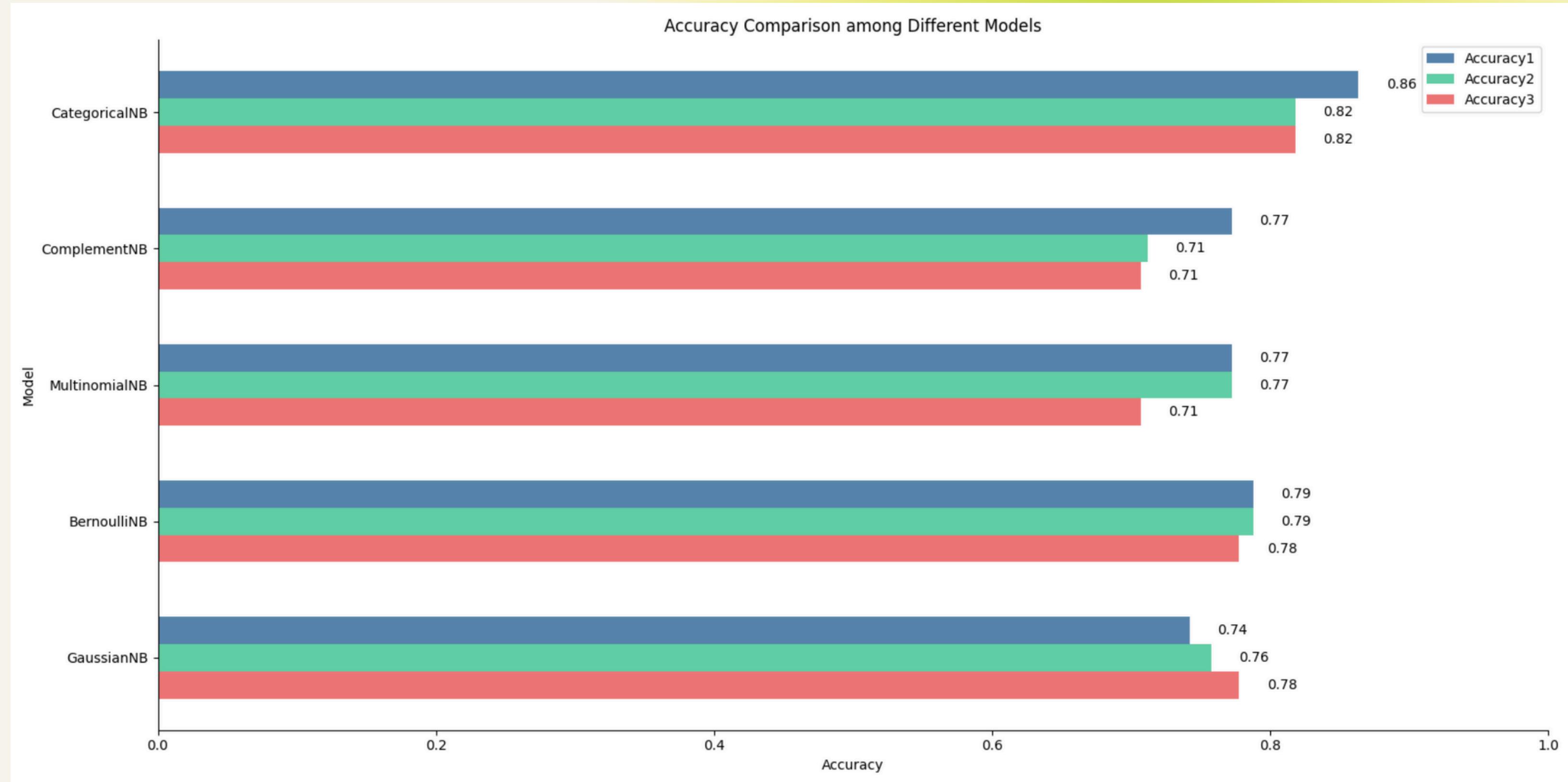
SPLIT DATA



TRAINING MODEL

70:30

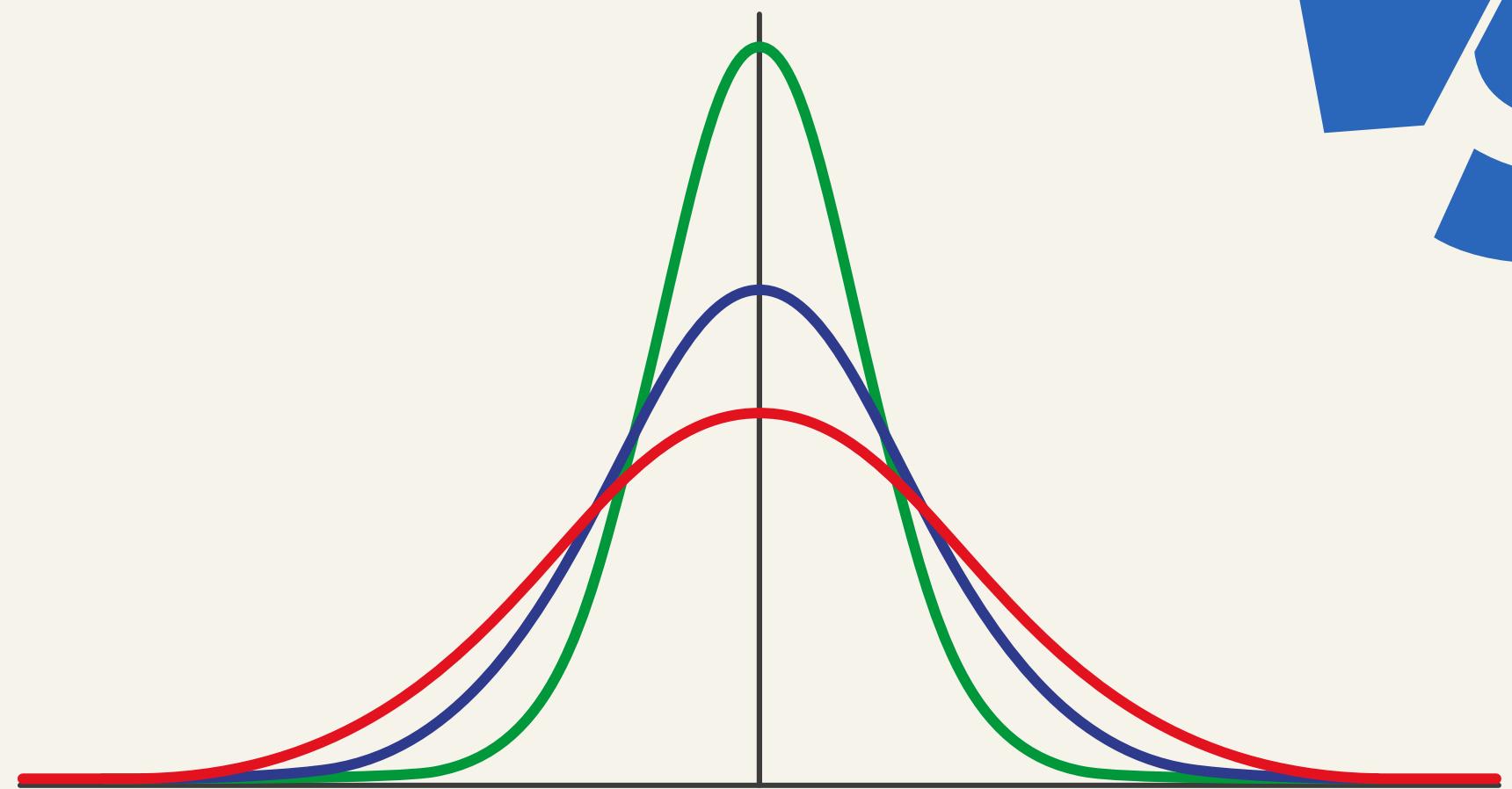
สรุปค่า ACCURACY ของแต่ละ MODEL ในแต่ละคลาสของ NAÏVE BAYES



**สรุปค่า ACCURACY ของแต่ละ MODEL
จากคลาสอื่น ๆ นอกจาก Navie Bayes**

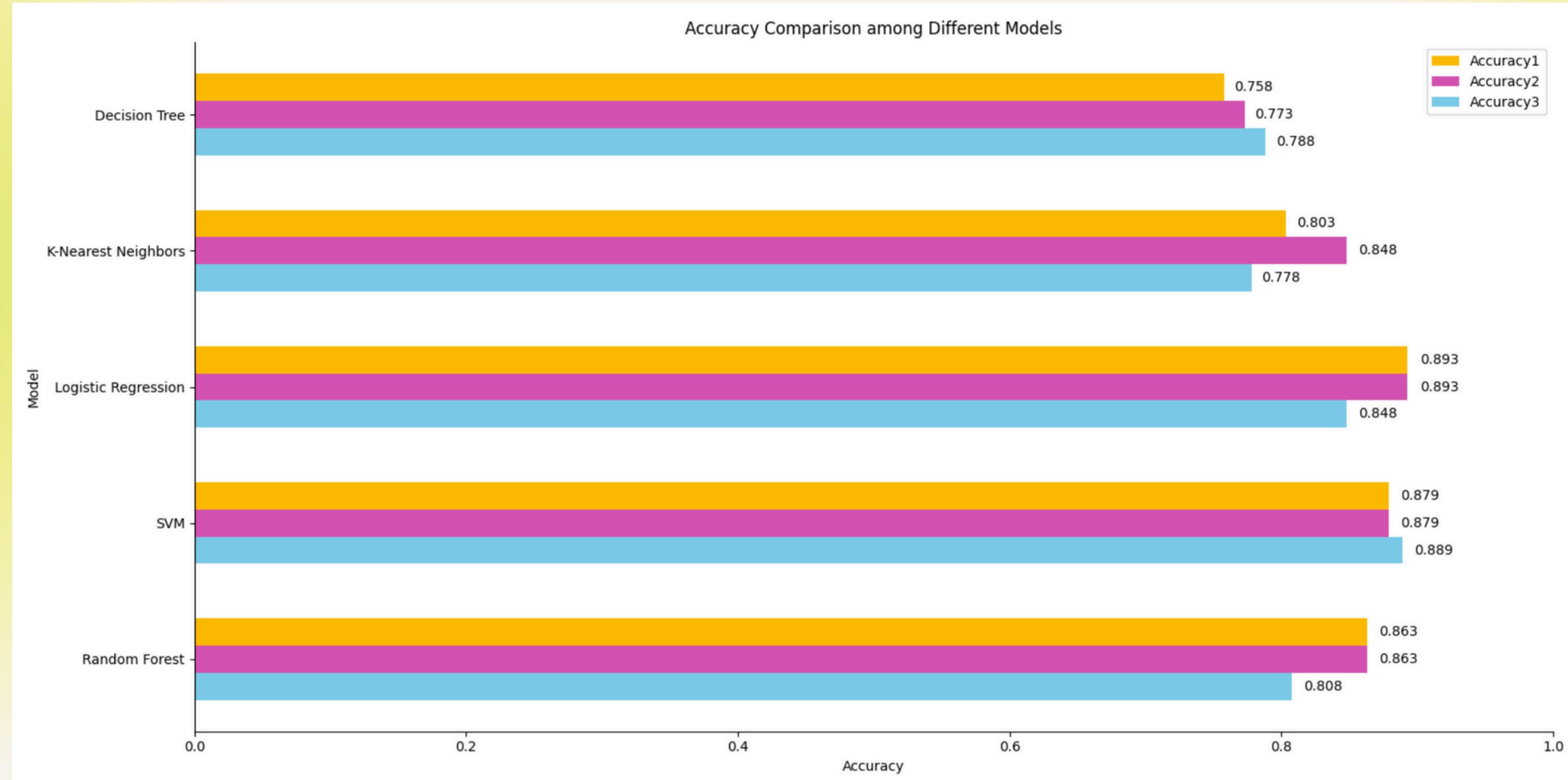
8
Atributes

vs

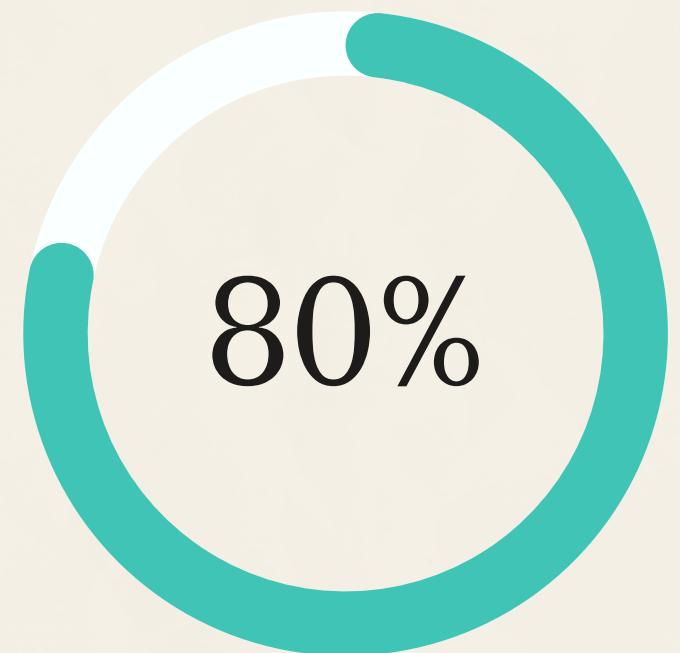


7,6,5
Atributes

สรุปค่า ACCURACY ของแต่ละ MODEL จากคลาสอื่น ๆ นอกจาก Navie Bayes

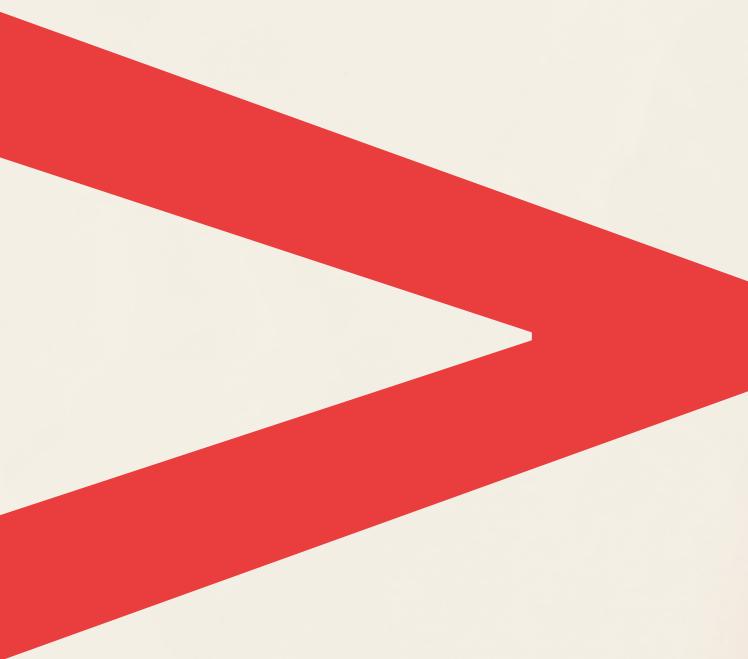


SPLIT DATA

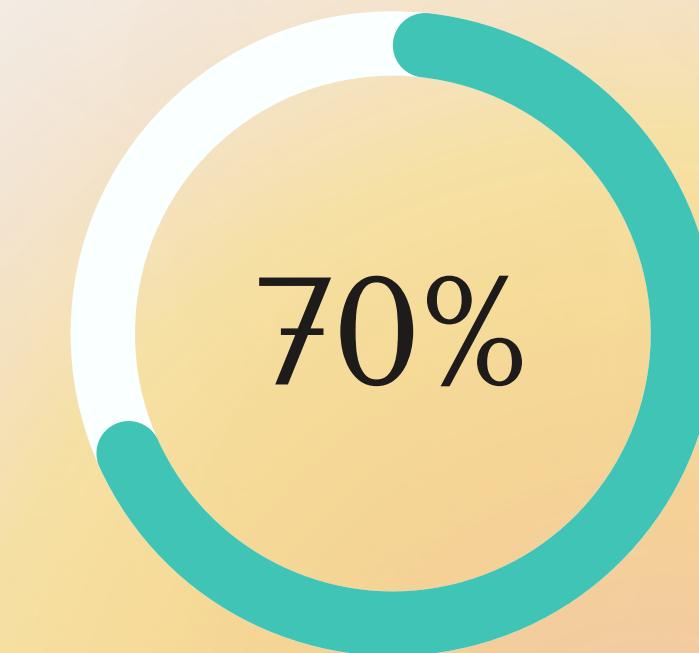


TRAINING MODEL

80:20



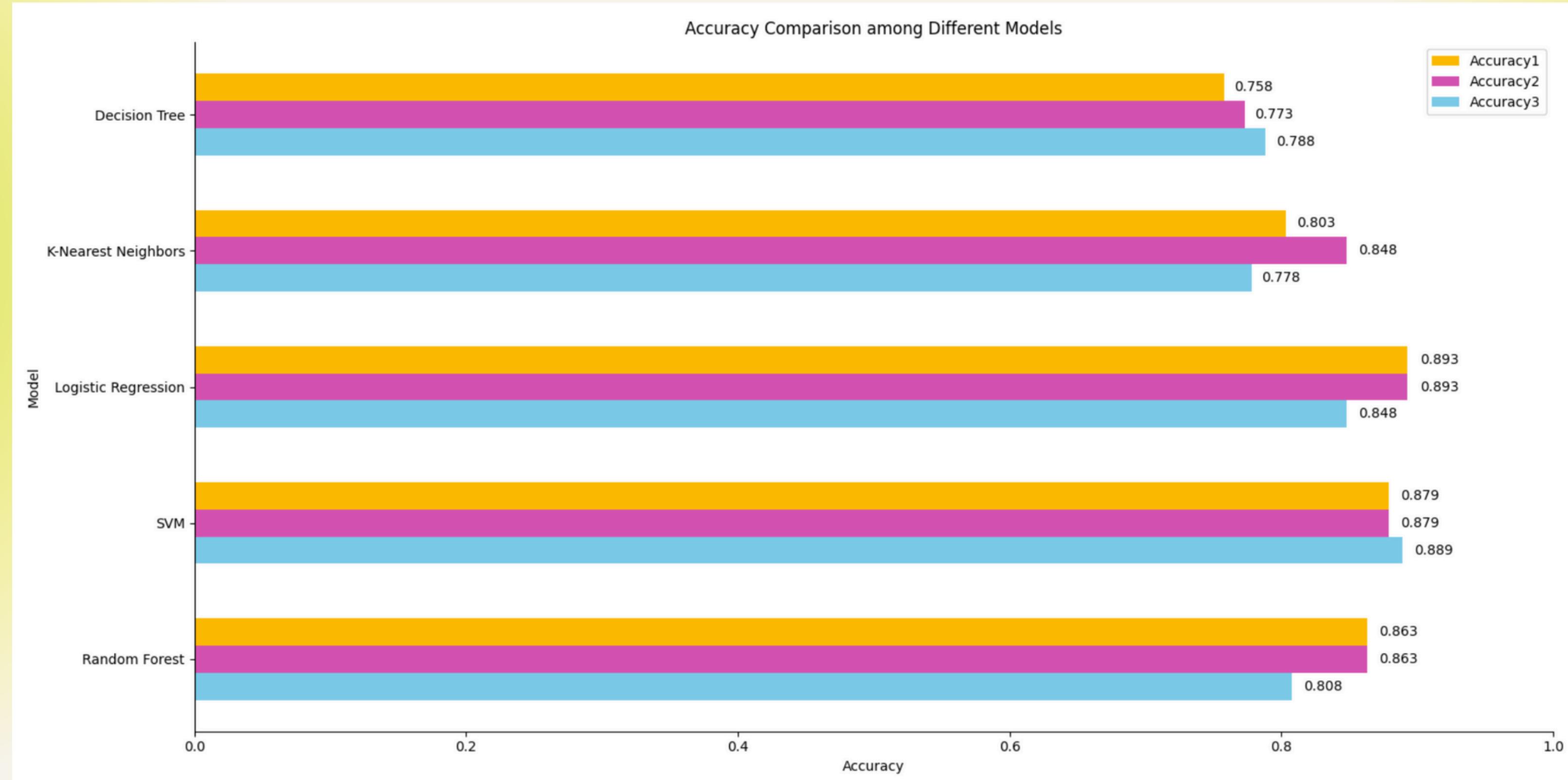
SPLIT DATA



TRAINING MODEL

70:30

สรุปค่า ACCURACY ของแต่ละ MODEL จากคลาสอื่น ๆ นอกจาก Navie Bayes



สรุปค่าการใช้ METRICS ในการวัดแต่ละโมเดล (ACCURACY, PRECISION, RECALL, F1 SCORE)

CONCLUSION

1

Logistic regression

โมเดลที่มีประสิทธิภาพ
มากที่สุด

2

Logistic regression

โมเดลที่หมายกับ Data set
ของเรามากที่สุด

3

Categorical NB

ค่าความแม่นยำ
มากที่สุด

5

Recall

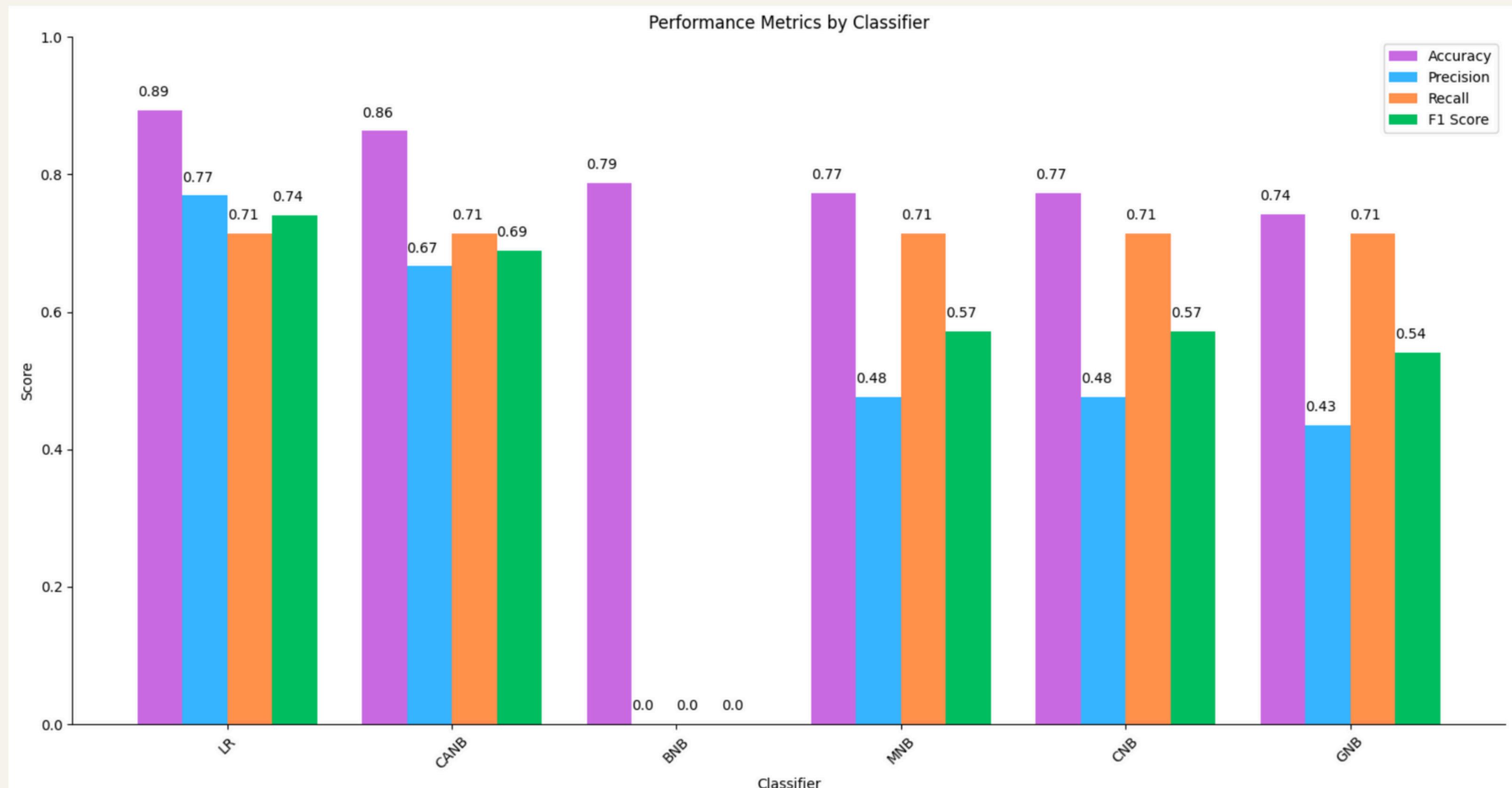
ค่า Recall
ของโมเดลทั้งหมด
(ยกเว้น BernoulliNB)
มีค่าเท่ากันอยู่ที่ 0.71

4

Bernoulli NB

ไม่น่าเชื่อถือ

สรุปค่าการใช้ METRICS ในการวัดแต่ละโมเดล (ACCURACY, PRECISION, RECALL, F1 SCORE)



CONCLUSION

1

Logistic regression

โมเดลที่มีประสิทธิภาพ
มากที่สุด

2

Logistic regression

โมเดลที่หมายกับ Data set
ของเรามากที่สุด

3

Categorical NB

ค่าความแม่นยำ
มากที่สุด

5

Recall

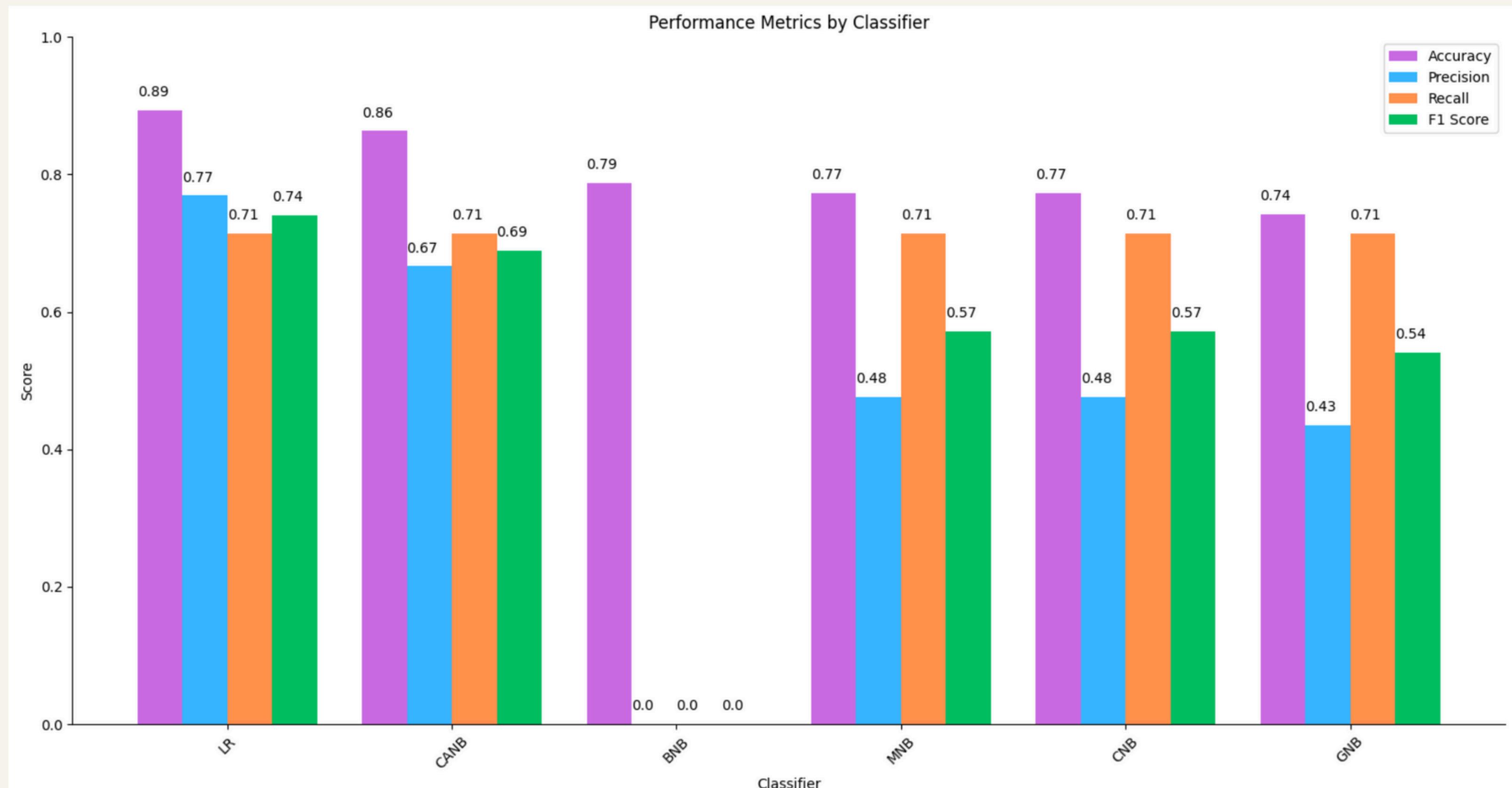
ค่า Recall
ของโมเดลทั้งหมด
(ยกเว้น BernoulliNB)
มีค่าเท่ากันอยู่ที่ 0.71

4

Bernoulli NB

ไม่น่าเชื่อถือ

สรุปค่าการใช้ METRICS ในการวัดแต่ละโมเดล (ACCURACY, PRECISION, RECALL, F1 SCORE)



CONCLUSION

1

Logistic regression

โมเดลที่มีประสิทธิภาพ
มากที่สุด

2

Logistic regression

โมเดลที่หมายกับ Data set
ของเรามากที่สุด

3

Categorical NB

ค่าความแม่นยำ
มากที่สุด

5

Recall

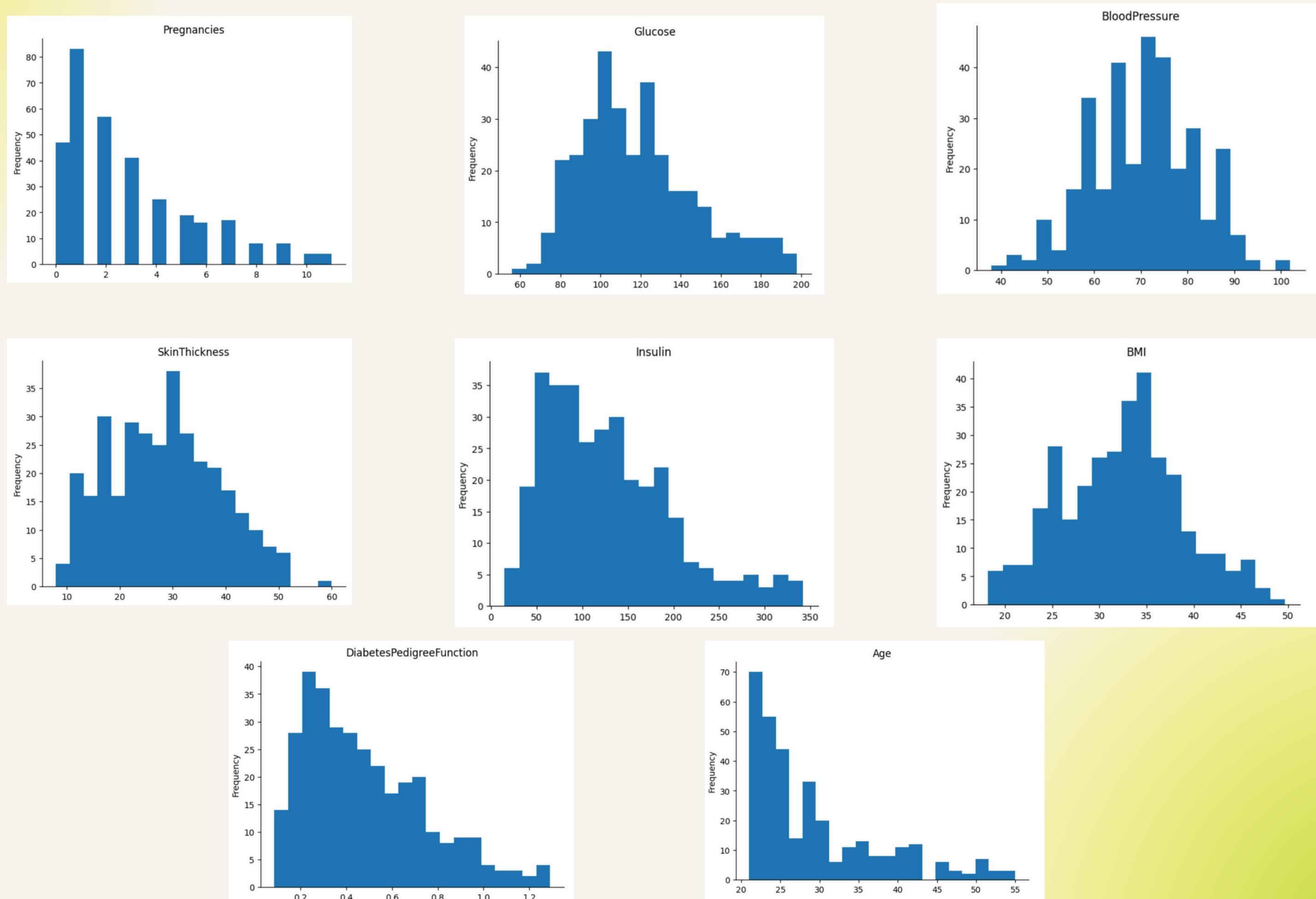
ค่า Recall
ของโมเดลทั้งหมด
(ยกเว้น BernoulliNB)
มีค่าเท่ากันอยู่ที่ 0.71

4

Bernoulli NB

ไม่น่าเชื่อถือ

| Outcome | |
|---------|--|
| 0 | |
| 1 | |
| 1 | |
| 1 | |
| 1 | |
| 1 | |
| ... | |
| 1 | |
| 1 | |
| 0 | |
| 1 | |
| 0 | |
| 0 | |



CONCLUSION

1

Logistic regression

โมเดลที่มีประสิทธิภาพ
มากที่สุด

2

Logistic regression

โมเดลที่หมายกับ Data set
ของเรามากที่สุด

3

Categorical NB

ค่าความแม่นยำ
มากที่สุด

5

Recall

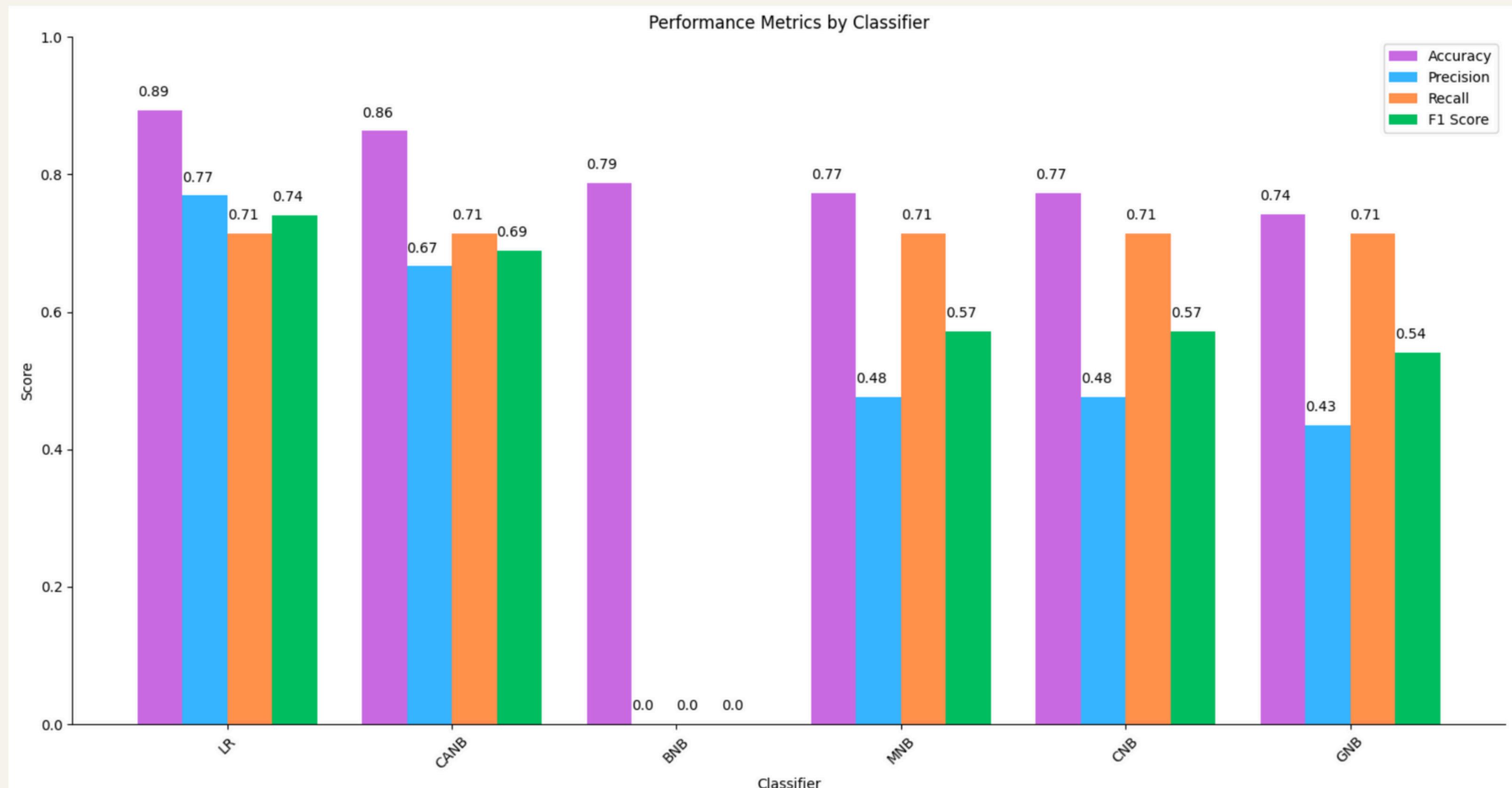
ค่า Recall
ของโมเดลทั้งหมด
(ยกเว้น BernoulliNB)
มีค่าเท่ากันอยู่ที่ 0.71

4

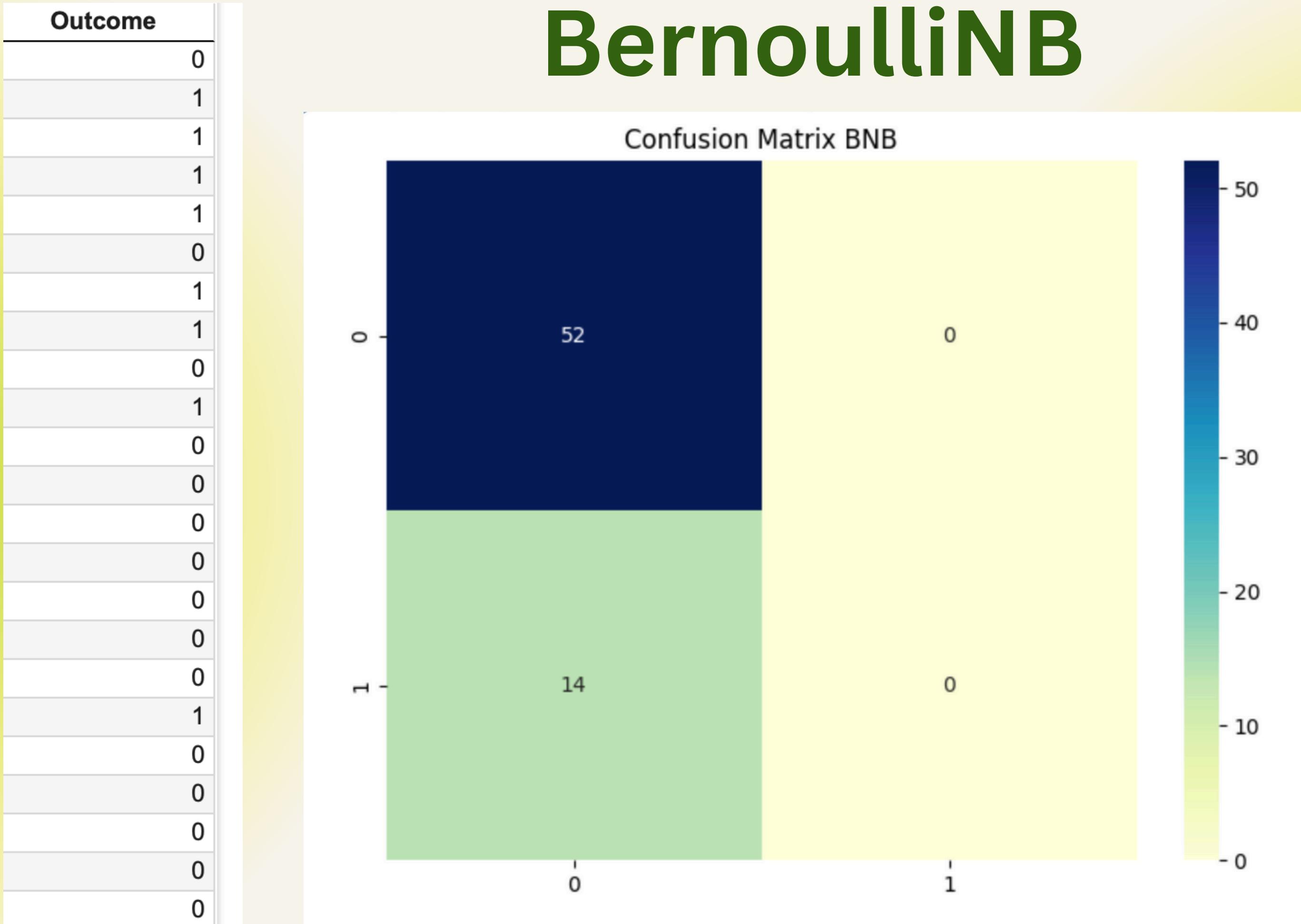
Bernoulli NB

ไม่น่าเชื่อถือ

สรุปค่าการใช้ METRICS ในการวัดแต่ละโมเดล (ACCURACY, PRECISION, RECALL, F1 SCORE)



BernoulliNB



CONCLUSION

1

Logistic regression

โมเดลที่มีประสิทธิภาพ
มากที่สุด

2

Logistic regression

โมเดลที่หมายกับ Data set
ของเรามากที่สุด

3

Categorical NB

ค่าความแม่นยำ
มากที่สุด

5

Recall

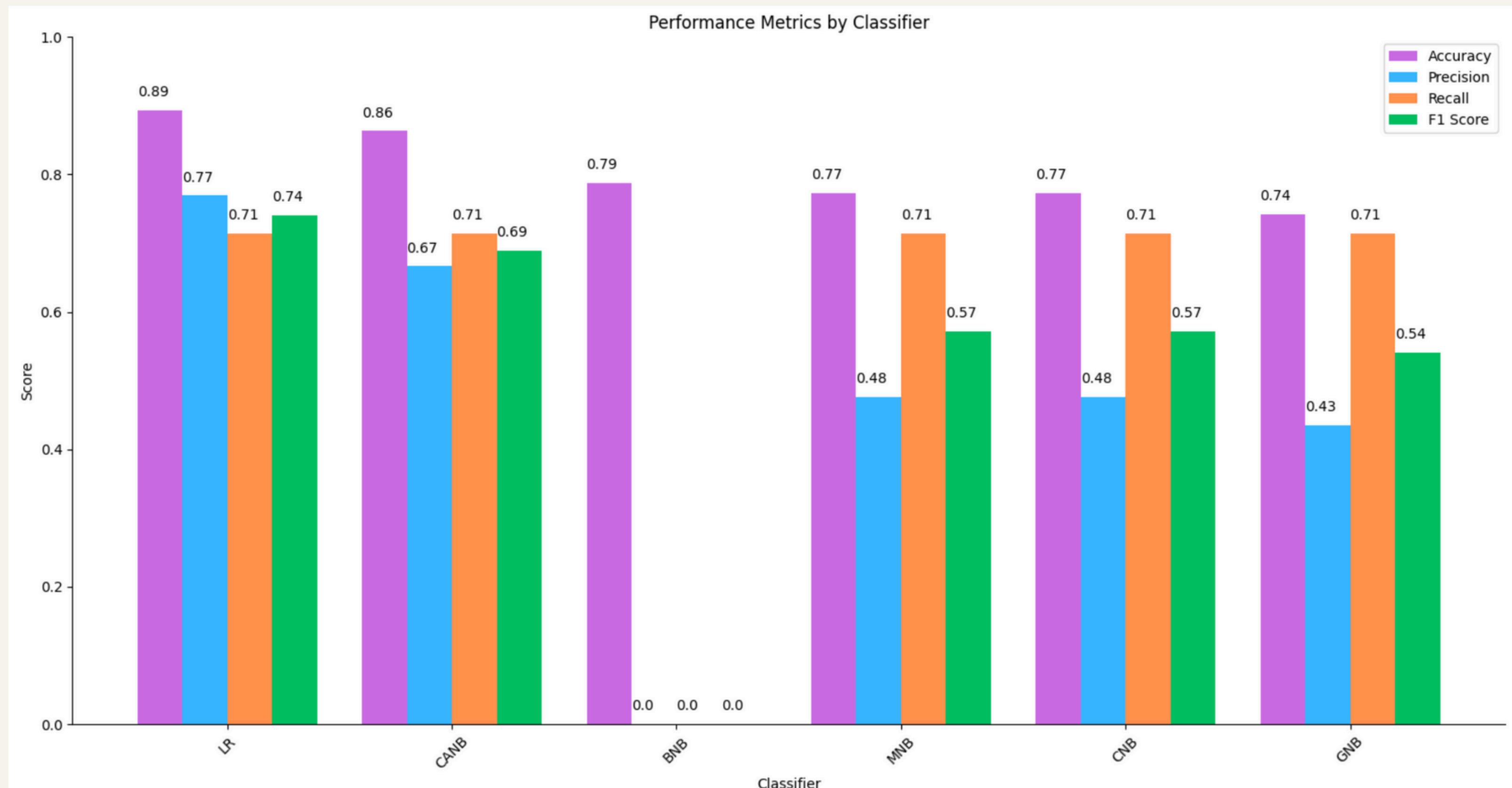
ค่า Recall
ของโมเดลทั้งหมด
(ยกเว้น BernoulliNB)
มีค่าเท่ากันอยู่ที่ 0.71

4

Bernoulli NB

ไม่น่าเชื่อถือ

สรุปค่าการใช้ METRICS ในการวัดแต่ละโมเดล (ACCURACY, PRECISION, RECALL, F1 SCORE)



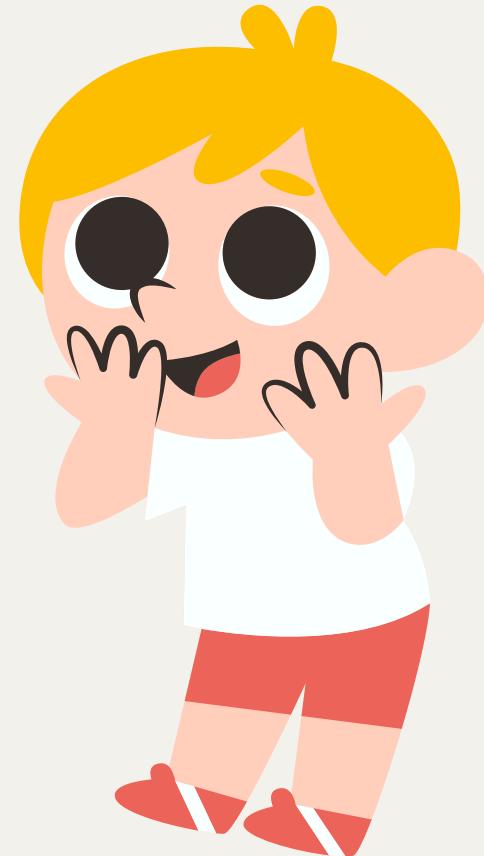
MEMBERS

Thank you

KANTIMA
6624650146



DUANGJAI
6624650245



THANAWAN
6624650252



CHADAPA
6624650187

