ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

----------o0o----------

BÁO CÁO ĐỒ ÁN
MÔN KHAI THÁC DỮ LIỆU VÀ ỨNG DỤNG

PHIÊN BẢN GIẢI TRÌNH CHỈNH SỬA VÀ BỔ SUNG

ĐỀ TÀI:
PHÂN LOẠI KHẢ NĂNG TỐT NGHIỆP CỦA SINH VIÊN UIT

GVHD: Nguyễn Thị Anh Thư
Nhóm 9:

1. Nguyễn Văn Thành Đạt 20520436
2. Nguyễn Hoàng Gia 20520478
3. Nguyễn Thái Huy 20520547
4. Lê Ngọc Mỹ Trang 20520817
5. Nguyễn Thế Vinh 20520862

Thành phố Hồ Chí Minh, tháng 3 năm 2023

MỤC LỤC

I. Tiến hành xem xét lại bảng ‘diem_Thu’ để lấy thêm dữ liệu từ các khóa năm 2006 đến năm 2012

Kết quả:

Từ bảng ‘diem_Thu’ lọc được thêm 4121 sinh viên từ các khóa trên, gồm các thông tin bao gồm: ‘mssv’, số tín chỉ và điểm số liên quan tới các môn mà sinh viên đó đã học

B1: Thực hiện gán lại ‘sotc’ của các môn anh văn(mamh == ENG…) từ 0 -> 4

B2: Loại bỏ các môn giáo dục thể chất (mamh == PE…). Do các môn này không có ảnh hưởng đến điểm trung bình học kỳ hay trung bình tích lũy, cũng như tổng số tính chỉ tích lũy.

B3: Lọc ra các dòng có giá trị ‘trangthai’==1, vì theo quan sát nhận thấy đây là trạng thái các môn học đã được tính là qua môn, có số lượng điểm thành phần nhiều hơn các nhóm khác.

B4: Tạo một cột ‘giatri_monhoc’ = ‘sotc’ * ‘diem_hp’, để thuận tiện cho quá trình tính toán.

B5: Lọc ra các dòng có namhoc <=2012. Tiến hành lấy unique ‘mssv’ của các dòng đó ta sẽ được mssv của các sinh viên từ năm 2012 trở về trước.

B6: với mỗi giá trị mssv tìm được ta tính các giá trị sau:

- Tính tổng tính chỉ: bằng cách lọc ra các dòng có mssv tương ứng và lấy tổng (sum()) của ‘sotc’ các dòng đó.
- Tính năm nhập học: bằng cách lọc ra các dòng có mssv tương ứng và lấy tổng (min()) của ‘namhoc’ các dòng đó. Lấy min() là vì năm học của môn học đầu tiên của sinh viên cũng chính là năm nhập học của sinh viên.
- Tính điểm trung bình của từng học kỳ (8 kỳ), theo công thức:
    - Dtb_hk = tonggiatri_hk / tongtc_hk
    - Với:
        - tonggiatri_hk là tổng ‘giatri_monhoc’ của học kỳ tương ứng.
        - tongtc_hk là tổng ‘sotc’ của học kỳ tương ứng.

- Học kỳ tương ứng = ‘namhoc’ - năm nhập học (được tính bên trên) + ‘hocky’

- Tính điểm trung bình tích lũy tổng ‘giatri_monhoc’ / tổng tính chỉ (đã tính bên trên).

II. Tiến hành kết hợp với các bảng còn lại để lấy ra những thông tin quan trọng

Kết quả:

Trải qua quá trình kết hợp và xem xét các bảng còn lại, nhóm nhận ra không có bảng nào chứa thông tin liên quan khác tới nhóm sinh viên này, các thông tin có thể rút ra được bao gồm:

- ‘Mssv’
- ‘Sotc_tichluy’: được tính bằng tổng số tín chỉ các môn mà sinh viên đã học trong bảng diem_Thu
- ‘Khoahoc’: bằng cách lấy năm học bắt đầu của sinh viên trừ cho 2006 cộng 1
- Các thông tin liên quan đến điểm trung bình các học kỳ mà sinh viên đó đã học (từ thông tin trung bình môn học và số tín chỉ của từng môn)
- Từ điểm trung bình trên ta tính được đtb tích lũy và đtb toàn khóa

III. Tiến hành gán nhãn cho nhóm sinh viên mới:

Do đây là các khóa ban đầu của trường nên có thể chắc chắn là sinh viên đã không còn học nữa, bên cạnh đó cần phải lọc lại các điều kiện để xem xét coi sinh viên đó có tốt nghiệp hay chưa.

- Gán trước cho tất cả nhãn 1 tức là đã tốt nghiệp
- Lọc theo các thông tin hiện có như số tín chỉ tích lũy, do trong chương trình đào tạo của trường không có thông tin đào tạo từ khóa 6 trở về trước mà chỉ có khóa 7 nên nhóm chưa thể kết luận chính xác về tổng số tính chỉ của từng khóa đầu được, nên nhóm đã quyết định lấy 100 chỉ làm ngưỡng để quyết định, nếu số tín chỉ đã học nhỏ hơn hoặc bằng 100 thì gán nhãn 0

- Bên cạnh đó các yếu tố khác để xét xem sinh viên có tốt nghiệp hay không như đầu ra anh văn, bị xử lý học vụ cho thôi học, … đều không có nên nhóm không sử dụng chúng để làm căn cứ gán nhãn cho sinh viên

Kết quả đạt được gồm 3536 nhãn 1 và 585 nhãn 0

IV. Gộp các mẫu cũ và mẫu mới thành 1 dataframe

Tiến hành gộp dữ liệu mới gồm 2168 dòng với 4121 dòng ta được bộ dữ liệu mới gồm 6289 dòng gồm

- 922 nhãn 0
- 5367 nhãn 1

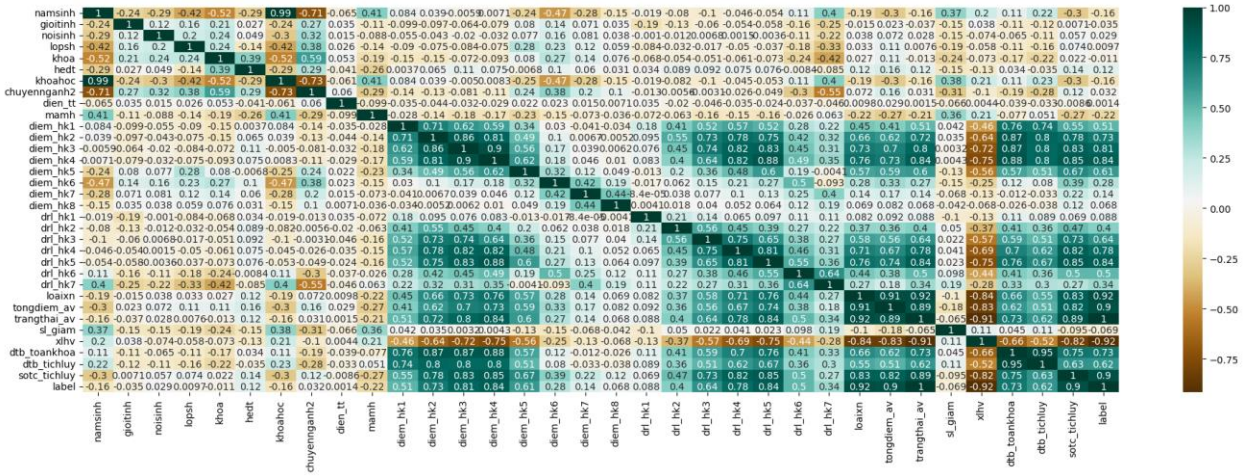
V. Điền khuyết các vị trí NaN của mẫu mới dựa trên mẫu cũ

- Ở cột ‘namsinh’, tính giá trị năm sinh của mỗi sinh viên bằng cách lấy giá trị ở cột ‘khoahoc’ của sinh viên đó trừ 1 cộng với 2006 và trừ cho 18 (mặc định 18 tuổi là vào đại học)
- Ở các cột như ‘gioitinh’, ‘noisinh’, ‘hedt’, ‘dien_tt’ nhóm tiến hành thay thế các giá trị null bằng các giá trị khác null xuất hiện thường xuyên nhất
    - Đối với cột ‘gioitinh’, do trường UIT chiếm số đông là nam nên giá trị này sẽ là 1
    - Đối với cột 'noisinh' là giá trị 'TP.HCM'
    - Đối với cột 'hedt' là giá trị 'CQUI'
    - Đối với cột 'dien_tt' là giá trị 'B'
    - Đối với cột ‘mamh’ là giá trị 'AVSC1'
- Cập nhật giá trị của cột ‘trangthai_av’:
    - Những sinh viên nào đã tốt nghiệp mà chưa được cập nhật chuẩn đầu ra ngoại ngữ (NULL) thì ta gán giá trị cho cột ‘trangthai_av’ bằng 1, tức là đã có bằng ngoại ngữ đạt chuẩn đầu ra
    - Những sinh viên nào không tốt nghiệp và chưa có thông tin liên quan đến chuẩn đầu ra ngoại ngữ (NULL) thì ta gán giá trị cho cột ‘trangthai_av’ bằng 0, tức là chưa có bằng ngoại ngữ đạt chuẩn đầu ra
- Cập nhật giá trị của cột ‘loaixn’:

    - Duyệt qua danh sách ‘loaixn’, trả về giá trị thuộc ‘loaixn’ khác NULL xuất hiện nhiều nhất
    - Những sinh viên nào có giá trị cột ‘trangthai_av’ bằng 1, thay thế những nơi mà giá trị ‘loaixn’ bị NULL thành giá trị xuất hiện nhiều nhất vừa tìm được ở trên (‘TOEIC)
    - Những sinh viên nào có giá trị cột ‘trangthai_av’ bằng 0, thay thế những nơi mà giá trị ‘loaixn’ bị NULL thành giá trị ‘Khong co’, tức là sinh viên này chưa có chứng chỉ ngoại ngữ nào
- Ở cột 'tongdiem_av'
    - Đối với một sinh viên có ‘trangthai_av’ bằng 1 nhưng giá trị ‘tongdiem_av’ bằng 0 hoặc NULL, thay thế giá trị đó bằng điểm số cao nhất dựa trên ‘loaixn’ của người đó và danh sách điểm số cao nhất ứng với mỗi ‘loaixn’ vừa tìm được ở trên
    - Đối với một sinh viên có ‘trangthai_av’ bằng 0 nhưng có điểm bằng NULL thì thay thế giá trị đó bằng 0
- Phần thông tin của cột 'sl_giam' sẽ được điền khuyết bằng giá trị 0 do chưa biết số lần mà người đó được giảm học phí cụ thể
- Cột 'xlhv', đối với những sinh viên có ‘label’ là 1, ta gán cho giá trị là 0 tức là chưa bị xử lý học vụ, còn đối với những sinh viên có ‘label’ là 0, ta trả về giá trị trung vị của cột ‘xlhv’ do trước khi bị cho thôi học thì sinh viên sẽ bị cảnh cáo xử lý học vụ
- Các cột như 'lopsh', 'khoa', 'chuyennganh2' có giá trị không thể thay thế bằng cách điền khuyết như trên được do một khóa bao giờ cũng có nhiều khoa nếu điền cùng 1 lớp thì không hợp lý, nên sẽ thay bằng giá trị 'Khong cung cap'
- Các cột điểm rèn luyện theo từng học kỳ đối với kỳ nào có ‘đtb_hk’ nếu bị NULL sẽ được điền bằng điểm trung bình của các giá trị đrl khác 0 thuộc cột đó, nếu kỳ đó đtb bằng -1, tức là sinh viên không học thì đrl bằng 0

- Cột ‘dtb_toankhoa’ cần thay thế các giá trị NULL bằng giá trị của cột ‘dtb_tichluy’ tương ứng, nếu trong trường hợp cả 2 cột bị NULL thì tức là sinh viên đó không hoàn thành bất cứ học kỳ nào, để giá trị 0

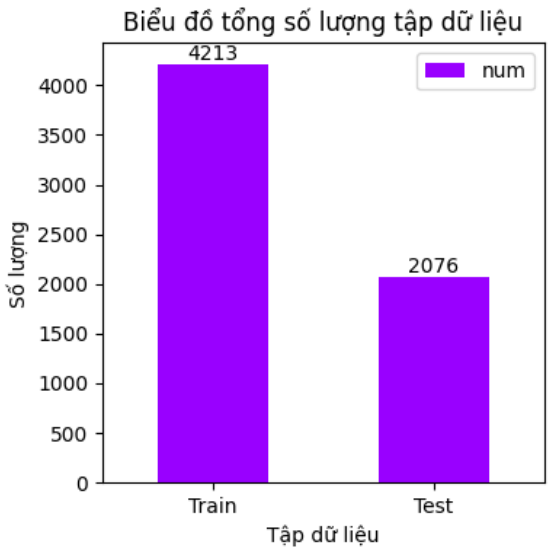
VI. Thống kê lại các thông tin đặc trưng của bảng dataset

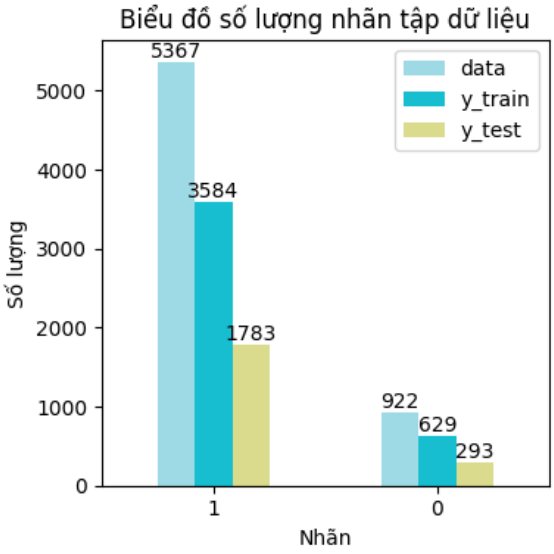


STT	Tên thuộc tính	Ý nghĩa	Phạm vi	Kiểu dữ liệu
1	namsinh	Năm sinh	1985, 2001	int64
2	gioitinh	Giới tính	0, 1	int64
3→ 10	mssv, noisinh, lopsh, khoa, hedt, chuyennganh2, mamh, loaixn	Mã số sinh viên, Nơi sinh, Lớp sinh hoạt, Khoa, Hệ đào tạo, Chuyên ngành, Xếp lớp anh văn, Loại bằng ngoại ngữ	_	object
11	khoahoc	Khóa học	1, 14	int64
12	dien_tt	Diện trúng tuyển	A, B, C	object
13	diem_hk1	Điểm trung bình học kỳ 1	-1.0, 9.7	float64

14	diem_hk2	Điểm trung bình học kỳ 2	-1.0, 9.36	float64
15	diem_hk3	Điểm trung bình học kỳ 3	-1.0, 9.66	float64
16	diem_hk4	Điểm trung bình học kỳ 4	-1.0, 9.9	float64
17	diem_hk5	Điểm trung bình học kỳ 5	-1.0, 10.0	float64
18	diem_hk6	Điểm trung bình học kỳ 6	-1.0, 10.0	float64
19	diem_hk7	Điểm trung bình học kỳ 7	-1.0, 10.0	float64
20	diem_hk8	Điểm trung bình học kỳ 8	-1.0, 9.0	float64
21→27	drl_hk1, drl_hk2, drl_hk3, drl_hk4, drl_hk5, drl_hk6, drl_hk7	Điểm rèn luyện học kỳ 1 đến 7	0, 100	int64
28	dtb_toankhoa	Điểm trung bình toàn khóa	0.0, 9.22	float64
29	dtb_tichluy	Điểm trung bình tích lũy	0.0, 9.22	float64
30	sotc_tichluy	Số tín chỉ tích lũy	0, 240	int64
31	tongdiem_av	Tổng điểm thi anh văn	0.0, 990.0	float64
32	trangthai_av	Trạng thái anh văn	0, 1	int64
33	sl_giam	Số lần được miễn giảm học phí	0, 6	int64
34	xlhv	Số lần bị xử lý học vụ	0, 6	int64
35	tinhtrang	Tình trạng học vụ của sinh viên	3, 9	in64

VII. Chia tập dữ liệu và huấn luyện



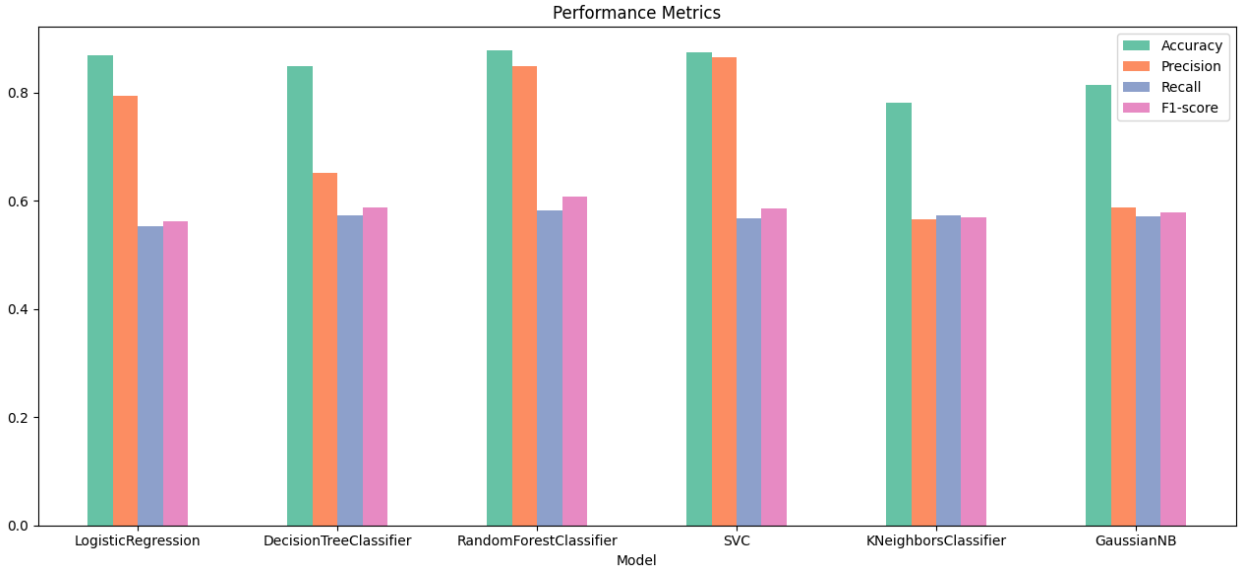


VIII. Kết quả

Dưới đây là danh sách kết quả đạt được với tập data mới

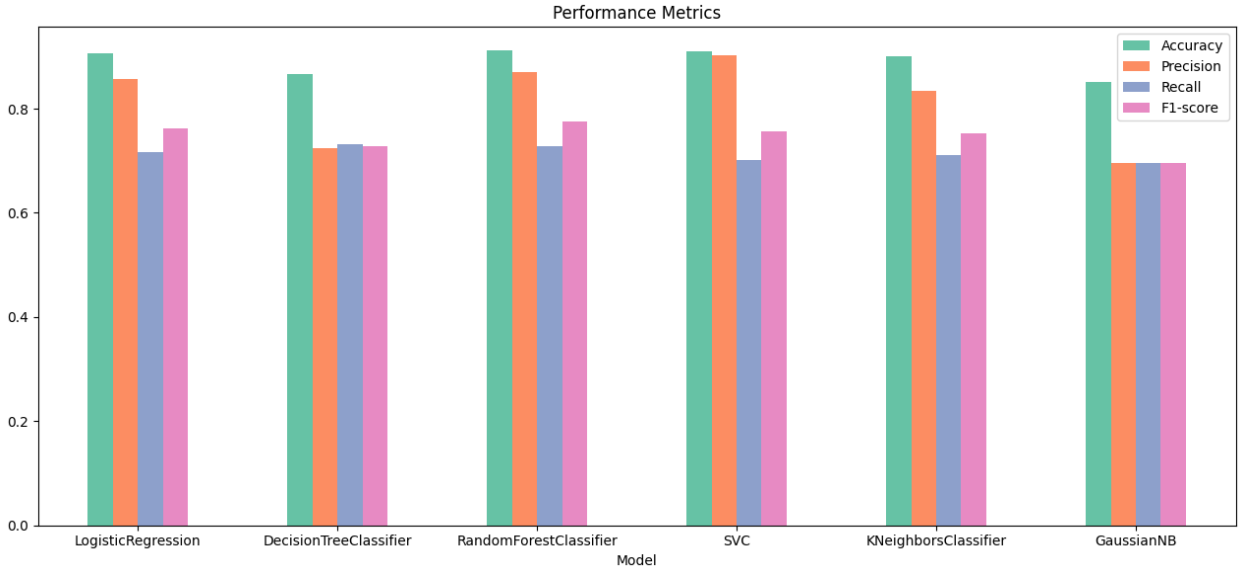
1. Nhóm 1

Model	Accuracy	Precision	Recall	F1-score
LogisticRegression	0.868497	0.794656	0.552668	0.561547
DecisionTreeClassifier	0.849711	0.652059	0.573105	0.588509
RandomForestClassifier	0.877649	0.849538	0.582240	0.607804
SVC	0.875241	0.864953	0.568003	0.586254
KNeighborsClassifier	0.780829	0.566456	0.572934	0.569226
GaussianNB	0.813584	0.587969	0.572038	0.578079



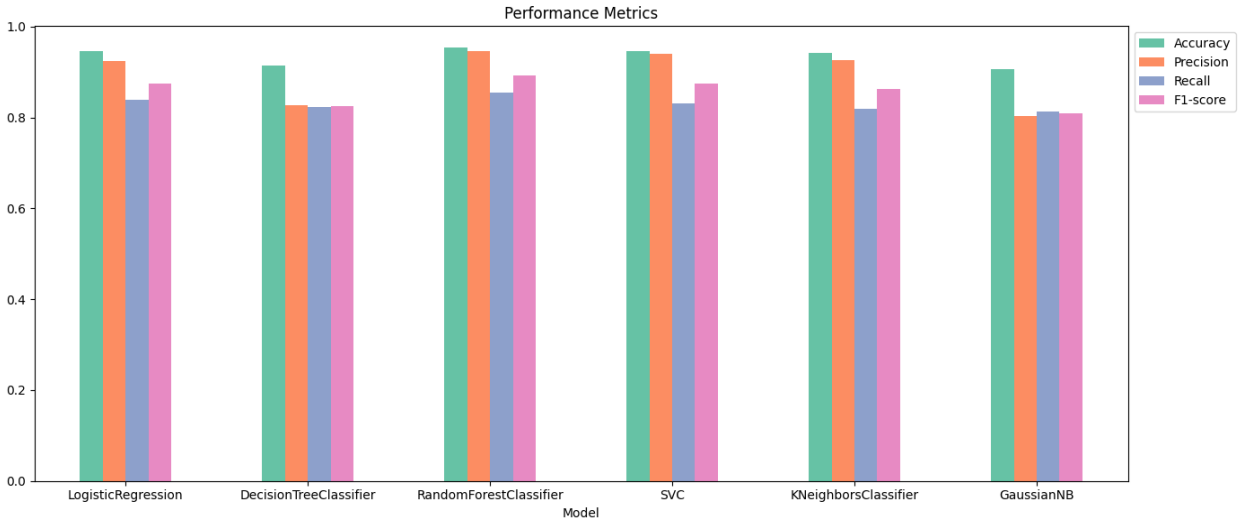
2. Nhóm 2

Model	Accuracy	Precision	Recall	F1-score
LogisticRegression	0.906551	0.857875	0.716002	0.761951
DecisionTreeClassifier	0.865607	0.723867	0.732095	0.727849
RandomForestClassifier	0.911368	0.870865	0.728789	0.775934
SVC	0.909923	0.903219	0.700852	0.755930
KNeighborsClassifier	0.901252	0.835102	0.710065	0.751770
GaussianNB	0.852119	0.695166	0.695722	0.695443



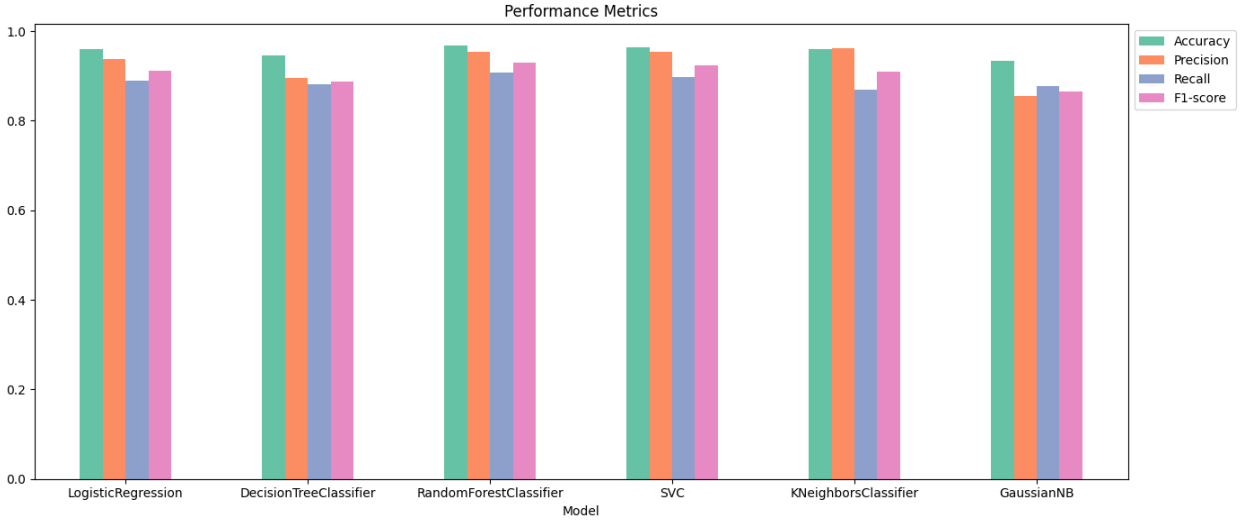
3. Nhóm 3

Model	Accuracy	Precision	Recall	F1-score
LogisticRegression	0.945087	0.924308	0.838260	0.874276
DecisionTreeClassifier	0.915222	0.826029	0.822300	0.824147
RandomForestClassifier	0.953276	0.946342	0.854436	0.892847
SVC	0.946532	0.940675	0.830545	0.874452
KNeighborsClassifier	0.941233	0.926961	0.818904	0.861774
GaussianNB	0.905588	0.803534	0.813839	0.808546



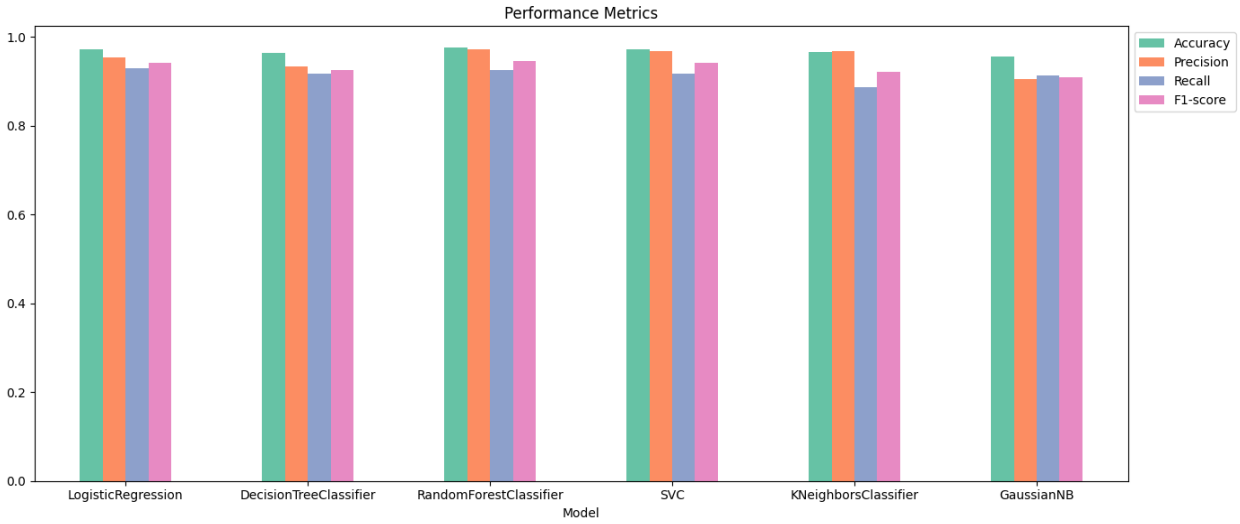
4. Nhóm 4

Model	Accuracy	Precision	Recall	F1-score
LogisticRegression	0.959538	0.938153	0.889455	0.911704
DecisionTreeClassifier	0.946532	0.895239	0.880457	0.887634
RandomForestClassifier	0.967245	0.953147	0.908202	0.928961
SVC	0.964836	0.954118	0.896817	0.922667
KNeighborsClassifier	0.960019	0.962031	0.869770	0.908615
GaussianNB	0.933044	0.855869	0.876883	0.865897



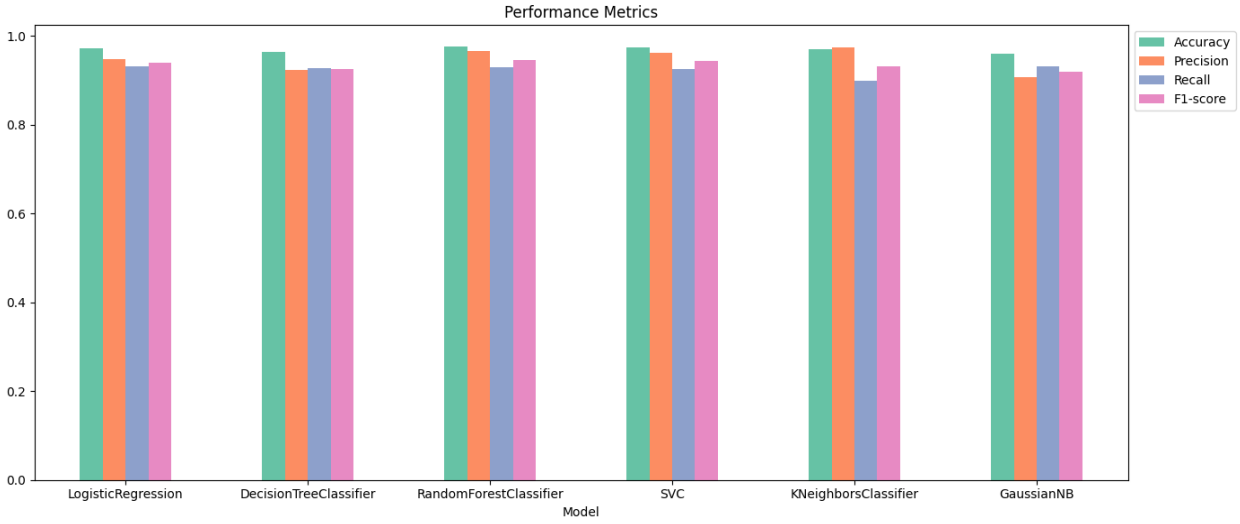
5. Nhóm 5

Model	Accuracy	Precision	Recall	F1-score
LogisticRegression	0.972062	0.953087	0.929545	0.940852
DecisionTreeClassifier	0.963873	0.932393	0.916221	0.924077
RandomForestClassifier	0.974952	0.970793	0.924097	0.945676
SVC	0.972543	0.967557	0.916991	0.940177
KNeighborsClassifier	0.964836	0.967127	0.885409	0.920671
GaussianNB	0.955202	0.904561	0.912600	0.908520



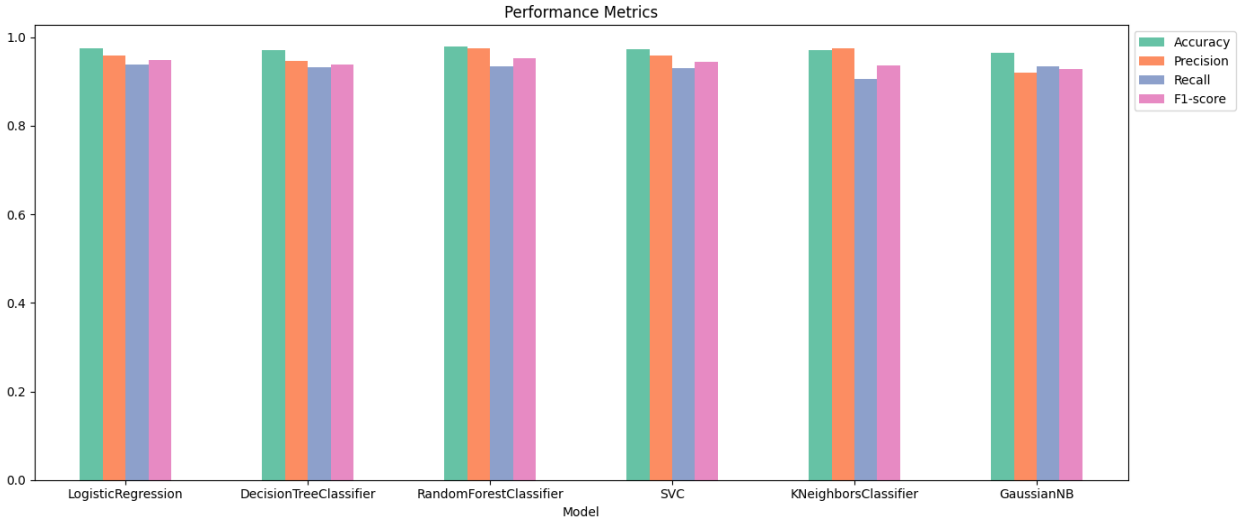
6. Nhóm 6

Model	Accuracy	Precision	Recall	F1-score
LogisticRegression	0.970617	0.946842	0.930130	0.938249
DecisionTreeClassifier	0.963391	0.922542	0.927350	0.924925
RandomForestClassifier	0.974952	0.966315	0.928375	0.946171
SVC	0.973025	0.960578	0.925828	0.942205
KNeighborsClassifier	0.969171	0.973481	0.897915	0.931010
GaussianNB	0.959056	0.906525	0.930530	0.917995



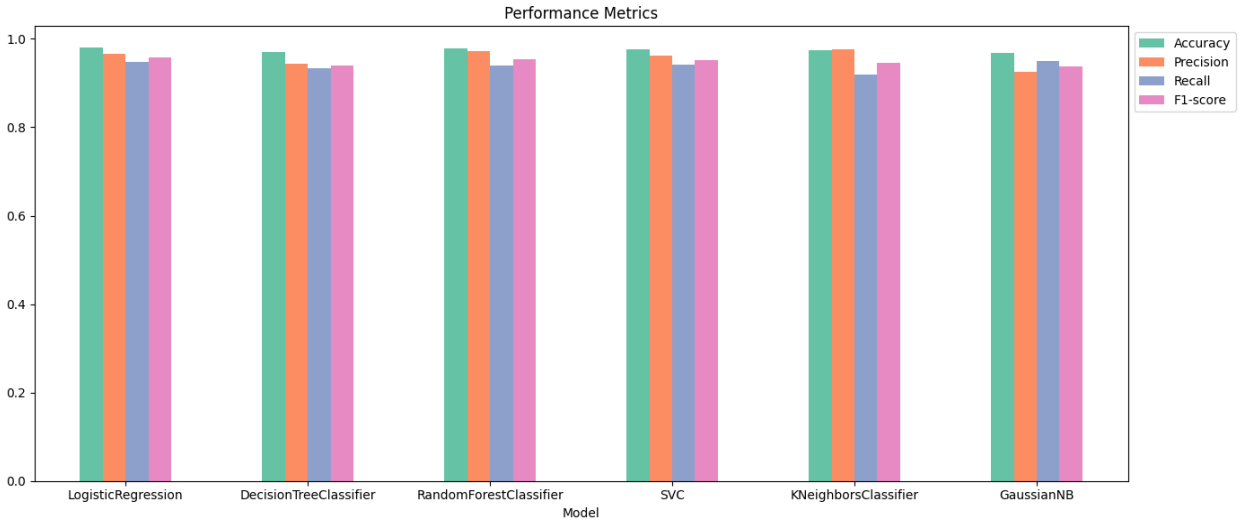
7. Nhóm 7

Model	Accuracy	Precision	Recall	F1-score
LogisticRegression	0.975434	0.958449	0.938638	0.948221
DecisionTreeClassifier	0.970617	0.945640	0.931556	0.938430
RandomForestClassifier	0.978324	0.974639	0.934617	0.953346
SVC	0.973507	0.958218	0.930386	0.943661
KNeighborsClassifier	0.971098	0.974709	0.904741	0.935733
GaussianNB	0.964355	0.920756	0.935041	0.927716



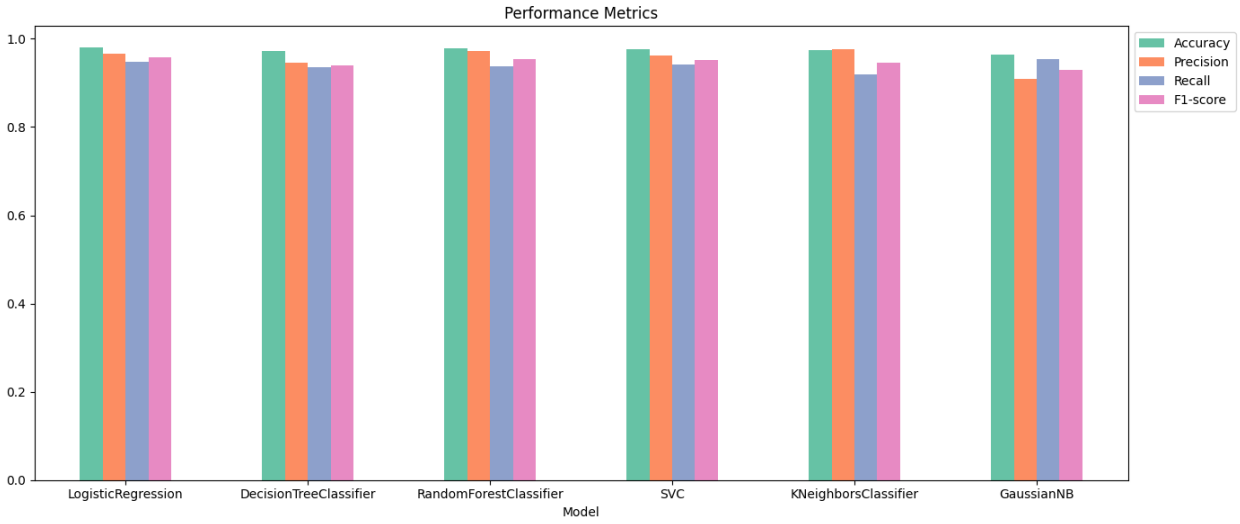
8. Nhóm 8

Model	Accuracy	Precision	Recall	F1-score
LogisticRegression	0.979769	0.967120	0.948292	0.957421
DecisionTreeClassifier	0.970617	0.943300	0.934408	0.938787
RandomForestClassifier	0.978805	0.972005	0.939175	0.954726
SVC	0.977360	0.962602	0.942612	0.952282
KNeighborsClassifier	0.974952	0.977163	0.918393	0.945000
GaussianNB	0.968690	0.925286	0.950399	0.937290



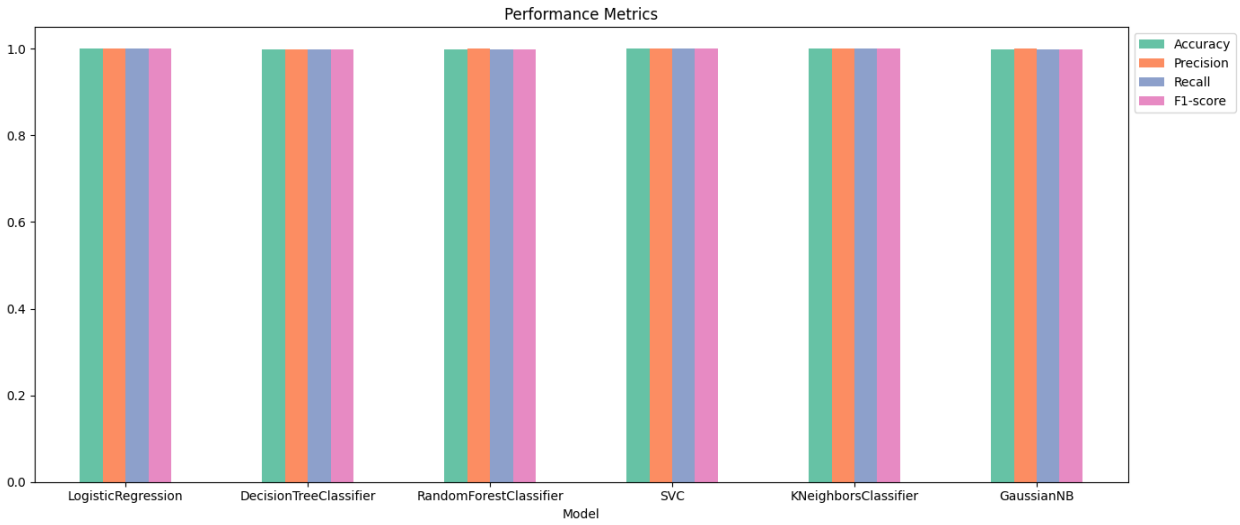
9. Nhóm 9

Model	Accuracy	Precision	Recall	F1-score
LogisticRegression	0.979769	0.967120	0.948292	0.957421
DecisionTreeClassifier	0.971580	0.946499	0.934969	0.940622
RandomForestClassifier	0.978324	0.971664	0.937469	0.953627
SVC	0.977360	0.962602	0.942612	0.952282
KNeighborsClassifier	0.974952	0.977163	0.918393	0.945000
GaussianNB	0.963873	0.908294	0.954725	0.929553



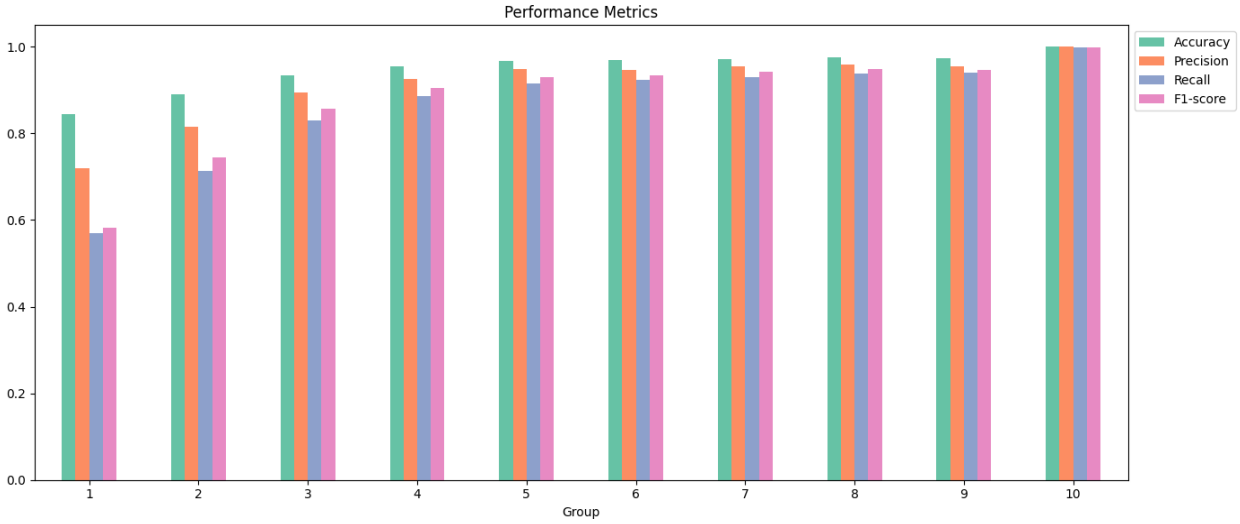
10. Nhóm 10

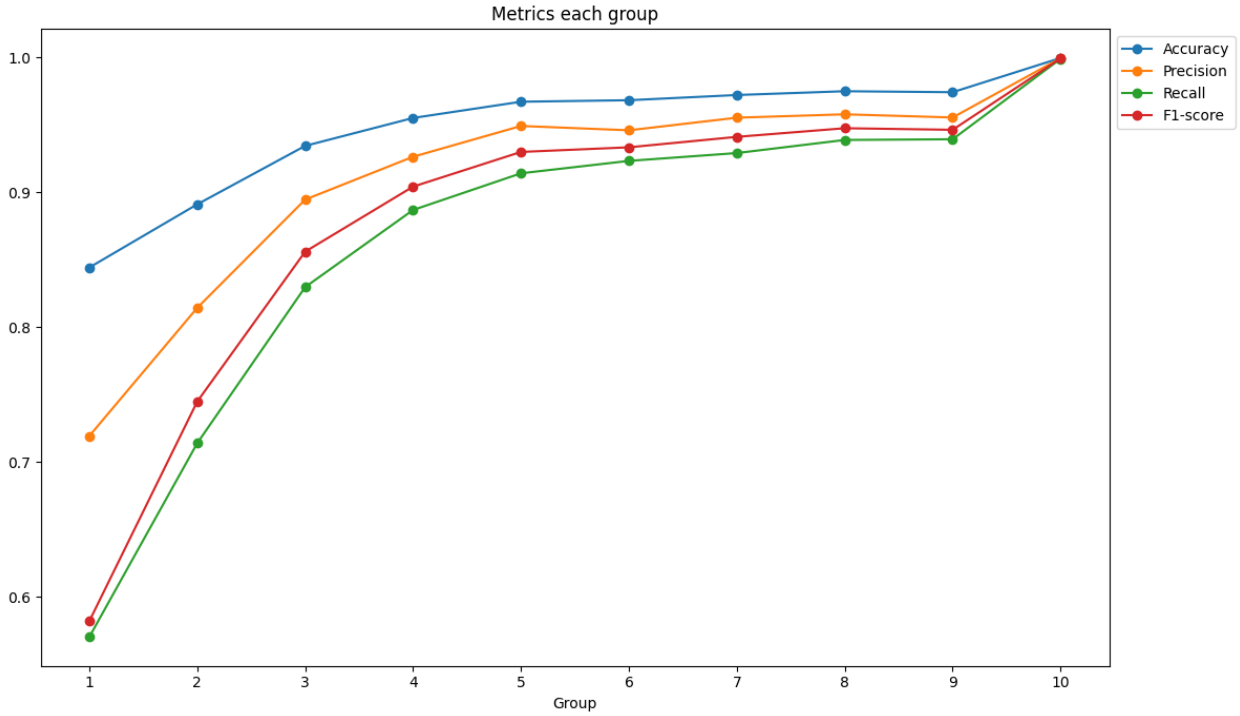
Model	Accuracy	Precision	Recall	F1-score
LogisticRegression	1.000000	1.000000	1.000000	1.000000
DecisionTreeClassifier	0.999037	0.998013	0.998013	0.998013
RandomForestClassifier	0.999518	0.999720	0.998294	0.999005
SVC	1.000000	1.000000	1.000000	1.000000
KNeighborsClassifier	1.000000	1.000000	1.000000	1.000000
GaussianNB	0.999518	0.999720	0.998294	0.999005



11. Trung bình kết quả theo nhóm:

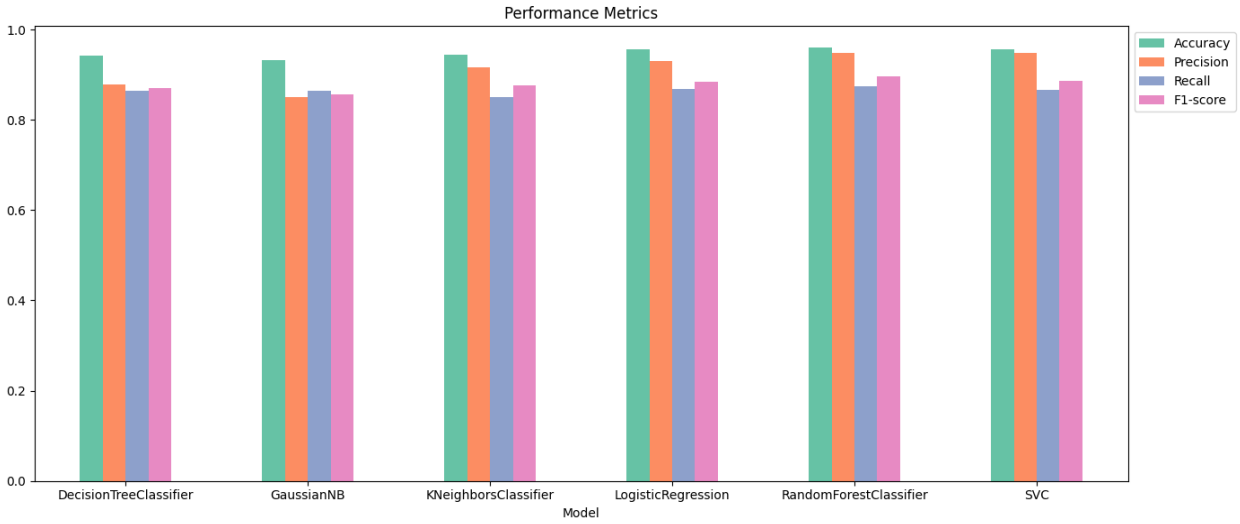
Group	Accuracy	Precision	Recall	F1-score
1	0.844252	0.719272	0.570165	0.581903
2	0.891137	0.814349	0.713921	0.744813
3	0.934489	0.894642	0.829714	0.856007
4	0.955202	0.926426	0.886931	0.904247
5	0.967245	0.949253	0.914144	0.929996
6	0.968369	0.946047	0.923355	0.933426
7	0.972222	0.955402	0.929163	0.941185
8	0.975032	0.957913	0.938880	0.947584
9	0.974310	0.955557	0.939410	0.946417
10	0.999679	0.999575	0.999100	0.999337

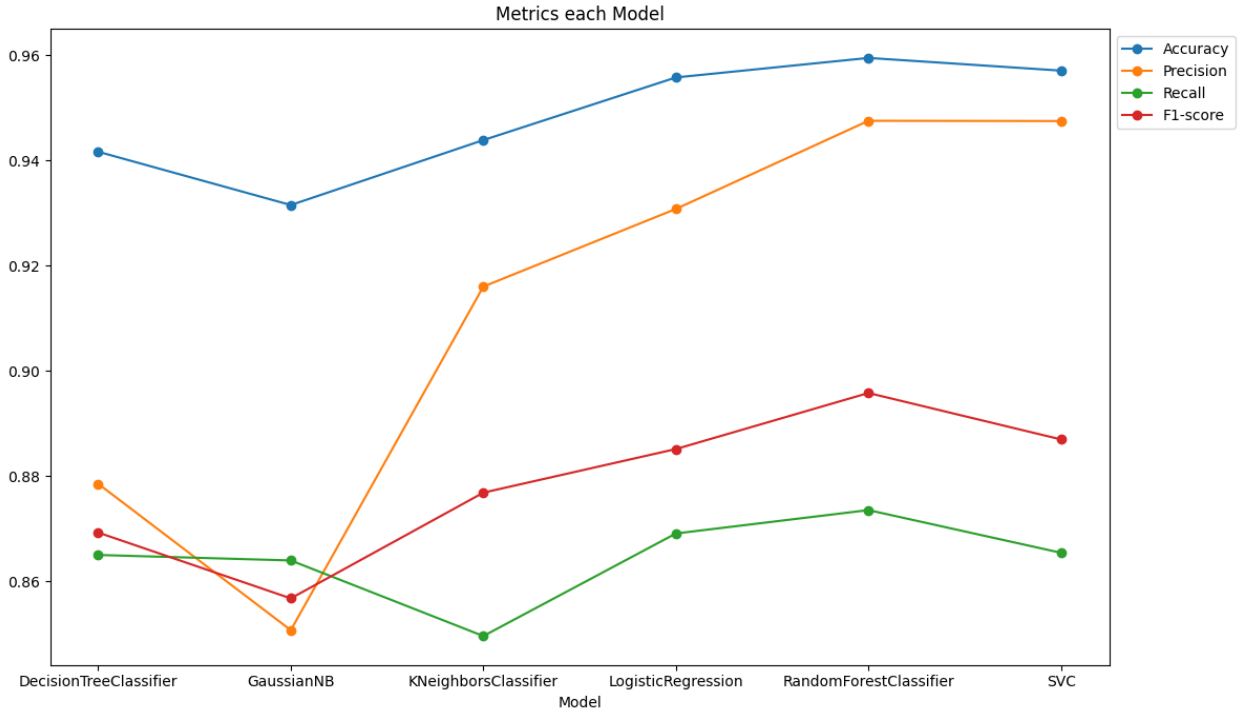




12. Trung bình kết quả theo phương pháp:

Model	Accuracy	Precision	Recall	F1-score
DecisionTreeClassifier	0.941618	0.878558	0.865047	0.869299
GaussianNB	0.931503	0.850768	0.864007	0.856804
KNeighborsClassifier	0.943834	0.916019	0.849652	0.876880
LogisticRegression	0.955732	0.930761	0.869128	0.885164
RandomForestClassifier	0.959441	0.947503	0.873569	0.895810
SVC	0.957033	0.947452	0.865465	0.886991


Performance Metrics
Accuracy
Precision
Recall
F1-score
1.0
0.8
0.6
0.4
0.2
0.0
DecisionTreeClassifier
GaussianNB
KNeighborsClassifier
LogisticRegression
RandomForestClassifier
SVC
Model


Metrics each Model
Accuracy
Precision
Recall
F1-score
0.96
0.94
0.92
0.90
0.88
0.86
DecisionTreeClassifier
GaussianNB
KNeighborsClassifier
LogisticRegression
RandomForestClassifier
SVC
Model