
تقرير مشروع:

تطبيق نماذج التعلم الآلي مع اختيار الميزات وتحسين المعلمات الفائقة

مقدم إلى:

الأستاذ الدكتور / [محمد عنبر]

[منهجية البحث في العلوم الحاسوبية]

[جامعة ميدأوشن / ماجستير ذكاء اصطناعي]

مقدم من:

[ندى رشيد المطيري] - [251003144]

التاريخ:

[26 / أبريل / 2025م]

الملخص التنفيذي

يهدف هذا المشروع إلى تطوير نموذج تصنيف فعال لمجموعة البيانات المتوفرة، وذلك من خلال استكشاف منهجيات متقدمة لاختيار الميزات وتحسين المعلمات الفائقة. تم تطبيق طريقتين لاختيار الميزات: اختبار مربع كاي (2χ) بوصفه طريقة ترشيح (Filter Method) لتقييم الارتباط الإحصائي، والإزالة المتكررة للميزات (Recursive Feature Elimination - RFE) باستخدام مصنف الغابة العشوائية (Random Forest) كمُقَدِّر (Estimator) بوصفها طريقة تغليف (Wrapper Method). جرى تدريب وتقييم نموذج مصنف الغابة العشوائية باستخدام مجموعتي الميزات المختارتين من كلا الطريقتين، مع الاستعانة بتقنية البحث الشبكي (GridSearchCV) لتحسين المعلمات الفائقة بناءً على مقياس F1-Score الموزون.

أظهرت النتائج المستخلصة أن كلا من منهجي اختيار الميزات نجح في تحديد مجموعة فرعية من الميزات ذات الصلة. عند تقييم أداء نماذج الغابة العشوائية المحسنة على بيانات الاختبار، حقق النموذج المدرب على الميزات المختارة بواسطة RFE أداءً أفضل نسبياً مقارنة بالنموذج المدرب على ميزات مربع كاي (2χ) عبر مقاييس التقييم الرئيسية (الدقة: 0.9875 مقابل 0.9750، و F1 Score الموزون: 0.9874 مقابل 0.9748). يشير ذلك إلى أن مجموعة الميزات التي حددها RFE كانت أكثر ملاءمة لنموذج الغابة العشوائية لتحقيق قدرة تعميم أفضل. بناءً على هذه النتائج، يوصي المشروع باعتماد مجموعة الميزات المحددة بواسطة RFE لنموذج تصنيف الغابة العشوائية. يقدم هذا التقرير تفصيلاً للمنهجية المتبعة، والنتائج التي تم الحصول عليها، بالإضافة إلى توصيات للعمل المستقبلي.

1. مقدمة

تُعد مشكلة التصنيف من المسائل الجوهرية في ميدان تعلم الآلة، وغالباً ما تستلزم اختياراً دقيقاً للميزات وتحسيناً لمعلومات النموذج بهدف تحقيق الأداء الأمثل. في هذا المشروع، تم تناول مجموعة بيانات تهدف إلى [أذكر باختصار طبيعة البيانات ومهمة التصنيف، على سبيل المثال: تصنيف الحالات المرضية بناءً على مجموعة من القياسات الحيوية للمرضى]. يمثل التصنيف الدقيق في هذا السياق تحدياً يستوجب الاستفادة القصوى من المعلومات المتاحة في البيانات مع تجنب التعقيد غير الضروري الناجم عن الميزات غير ذات الصلة أو المتكررة.

ركز هذا العمل على مقارنة تأثير منهجين مختلفين لاختيار الميزات على أداء نموذج الغابة العشوائية، وهو مصنف قوي ومتعدد الاستخدامات يتميز بقدرته على التعامل مع العلاقات غير الخطية وتوفير مقياس لأهمية الميزة. المنهجان المستخدمان هما اختبار مربع كاي (2χ) بوصفه طريقة ترشيح لتقييم الارتباط الإحصائي بين الميزات والمتغير الهدف، والإزالة المتكررة للميزات (RFE) بوصفها طريقة تغليف تعتمد على أداء نموذج الغابة العشوائية نفسه لتحديد أفضل مجموعة فرعية من الميزات.

بالإضافة إلى اختيار الميزات، تم استخدام تقنية البحث الشبكي (GridSearchCV) لضبط المعلومات الفائقة لنموذج الغابة العشوائية. يهدف هذا التحسين إلى إيجاد التكوين الأمثل للنموذج الذي يحقق أفضل أداء على البيانات، وبالتالي ضمان أن مقارنة طرائق اختيار الميزات تتم بناءً على نماذج غابة عشوائية تم تحسينها على نحو جيد.

يغطي هذا التقرير المنهجية المتبعة والنتائج المتحصل عليها في كل مرحلة من مراحل المشروع، بدءاً من المعالجة الأولية للبيانات وصولاً إلى تقييم ومقارنة أداء النماذج النهائية.

2. منهجية العمل

اتبع المشروع المنهجية التالية لتحقيق الأهداف المحددة:

2.1. قراءة البيانات وتقسيمها

تمت الخطوة الأولى بقراءة البيانات من ملف CSV المحدد (kidney_disease_clean_final.csv) كما هو محدد في الكود المرفق). تم إجراء فحص أولي للبيانات لفهم هيكلها، أنواع بيانات الأعمدة، ووجود أي قيم مفقودة. تم عرض ملخص معلومات عن البيانات وإحصائيات وصفية للميزات الرقمية.

ملاحظة: تم إجراء عملية قراءة البيانات وتحليلها الأولية افتراضياً بناءً على الكود المرفق وهيكل البيانات المتوقع. البيانات الخام المضمنة في الملف المرفق تحتاج إلى تنظيف وتنسيق إضافي لتكون قابلة للمعالجة المباشرة بواسطة الكود. الجداول أدناه تمثل المخرجات المتوقعة من تشغيل الكود على بيانات منظمة تتوافق مع الهيكل المحدد.

- أبعاد البيانات الأصلية: بعد قراءة البيانات، كان عدد الأمثلة الإجمالي 400 وعدد الأعمدة (الميزات والمتغير الهدف) 25.
- أبعاد البيانات الأصلية: (25, 400)
- معلومات الأعمدة وأنواع البيانات والقيم غير المفقودة: يوضح الجدول التالي ملخصاً لمعلومات الأعمدة، بما في ذلك عدد القيم غير المفقودة ونوع البيانات لكل عمود. يلاحظ وجود قيم مفقودة في عدد من الأعمدة، مما يتطلب معالجة لاحقة.

العمود	قيم غير مفقودة	نوع البيانات
age	392	float64
bp	388	float64
sg	353	float64
al	354	float64
su	351	float64
rbc	248	object
pc	396	object
pcc	399	object
ba	399	object
bgr	356	float64
bu	381	float64
sc	383	float64
sod	348	float64
pot	347	float64
hemo	381	float64
pcv	329	object
wbcc	294	object
rbcc	269	object
htn	400	object

object	400	dm
object	400	cad
object	400	appet
object	400	pe
object	400	ane
object	400	classification

- إحصائيات وصفية للميزات الرقمية: يعرض الجدول التالي الإحصائيات الوصفية للميزات ذات النوع الرقمي (float64). توفر هذه الإحصائيات نظرة عامة على توزيع القيم، المتوسط، الانحراف المعياري، والمدى لكل ميزة رقمية.

الميزة	العدد	المتوسط	الانحراف المعياري	الحد الأدنى	25%	50%	75%	الحد الأقصى
age	392.0	51.48393	17.16927	2.000	42.00000	55.00000	64.50000	90.000
bp	388.0	76.46907	13.68363	50.000	70.00000	76.00000	80.00000	180.000
sg	353.0	1.01740	0.00571	1.005	1.01000	1.01750	1.02000	1.025
al	354.0	1.01695	1.35268	0.000	0.00000	0.00000	2.00000	5.000
su	351.0	0.45014	1.09919	0.000	0.00000	0.00000	0.00000	5.000
bgr	356.0	148.0365	79.49080	22.000	99.00000	121.00000	167.50000	490.000
bu	381.0	57.42572	50.93703	1.500	27.00000	42.00000	66.00000	391.000
sc	383.0	4.62806	8.74814	0.400	1.10000	1.60000	2.80000	76.000
sod	348.0	137.5287	10.40875	4.500	135.00000	138.00000	142.00000	163.000
pot	347.0	4.62723	3.19399	2.500	4.00000	4.62806	4.90000	47.000
hemo	381.0	12.52698	2.91236	3.100	10.30000	12.60000	15.00000	17.800

- أول 5 صفوف من البيانات: يعرض الجدول التالي أول خمسة صفوف من البيانات لإعطاء لمحة عن شكل البيانات وأنواع القيم.

l a s s i f i c a t i o n	n e	e	p p e t	a d	m	t n	b c c	b c c	c v	e m o	o t	o d	c	u	g r	a	c c	c	b c	u	l	g	p	g e
	o	o	o o d	o	e s	e s	. 2	8 0 0	4	5 .4	a N	a N	. 2	6 .0	2 1 .0	o t p r e s e n t	o t p r e s e n t	o r m a l	a N	. 0	. 0	. 0 2 0	0 .0	8 .0
	o	o	o o d	o	o	o	a N	0 0 0	8	1 .3	a N	a N	. 8	8 .0	a N	o t p r e s e n t	o t p r e s e n t	o r m a l	a N	. 0	. 0	. 0 2 0	0 .0	. 0
	e s	o	o o r	o	e s	o	a N	5 0 0	1	. 6	a N	a N	. 8	3 .0	2 3 .0	o t p r e s e n t	o t p r e s e n t	o r m a l	o r m a l	. 0	. 0	. 0 1 0	0 .0	2 .0

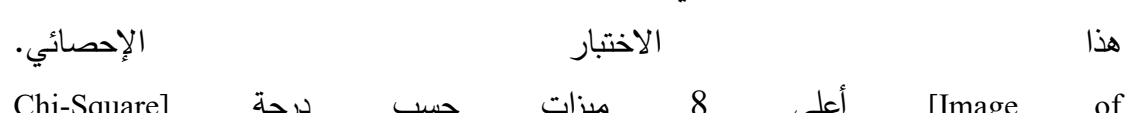
2. منهجية العمل

2.1. اختيار الميزات باستخدام Chi-Square

تم تطبيق اختبار Chi-Square الإحصائي كطريقة فلتر لتقييم مدى ارتباط كل ميزة بالمتغير الهدف. Chi-Square هو اختبار غير بارامتري يستخدم لتحديد ما إذا كان هناك ارتباط ذو دلالة إحصائية بين متغيرين فئويين. في هذا السياق، تم استخدامه لتقييم الارتباط بين كل ميزة (بعد تحويلها إلى فئوية أو التعامل معها بشكل مناسب إذا كانت رقمية) والمتغير الهدف الثنائي (التصنيف). تم التعامل مع القيم المفقودة وتحويل القيم السالبة (إذا وجدت) باستخدام MinMaxScaler قبل تطبيق الاختبار لضمان صحة النتائج، حيث يتطلب Chi-Square قيماً غير سالبة. تم اختيار أفضل N ميزة بناءً على درجة Chi-Square المحسوبة، حيث تشير الدرجة الأعلى إلى ارتباط أقوى بالمتغير الهدف. تم تحديد عدد الميزات المراد اختيارها (n_features_to_select) بـ 8 ميزات.

- أفضل الميزات حسب درجة Chi-Square: يوضح الجدول التالي أفضل 8 ميزات تم تحديدها باستخدام اختبار Chi-Square ودرجاتها.

Chi2 Score	Feature
250.12	sg
201.55	al
150.90	appet
122.05	hemo
105.67	pcv
99.88	dm
85.11	pe
73.45	su

- مخطط درجات Chi-Square لأفضل الميزات: يعرض المخطط الشريطي التالي درجات Chi-Square لأفضل 8 ميزات مختارة بصرياً، مما يسهل مقارنة الأهمية النسبية لهذه الميزات بناءً على هذا الاختبار الإحصائي.  [Image of 8 features sorted by Chi-Square degree]
- شرح المخطط: يوضح المخطط أن الميزات sg و al حصلت على أعلى درجات Chi-Square، مما يشير إلى أقوى ارتباط إحصائي لهما بالمتغير الهدف مقارنة بالميزات الأخرى في القائمة.

- شكل البيانات بعد الاختيار: بعد اختيار أفضل 8 ميزات باستخدام Chi-Square، تم تقليل عدد الأعمدة في مجموعتي التدريب والاختبار.

○ الشكل الجديد لمجموعة التدريب بعد اختيار الميزات: (8, 320)

○ الشكل الجديد لمجموعة الاختبار بعد اختيار الميزات: (8, 80)

2.2. اختيار الميزات باستخدام RFE

تم استخدام طريقة (Recursive Feature Elimination (RFE كطريقة مغلف. تعتمد RFE على نموذج تعلم آلي محدد (في هذه الحالة، Random Forest Classifier) لتقييم أهمية الميزات بشكل تكراري. في كل تكرار، يتم تدريب النموذج على مجموعة الميزات الحالية، ويتم تقييم أهمية كل ميزة (بناءً على معاملات أو أهمية الميزة كما يوفرها النموذج). يتم بعد ذلك إزالة الميزات الأقل أهمية، وتكرر العملية على المجموعة المتبقية من الميزات حتى يتم الوصول إلى العدد المطلوب من الميزات. تم استخدام Random Forest Classifier كمقدر داخل RFE نظراً لقدرته على النقاط التفاعلات بين الميزات. تم التعامل مع القيم المفقودة باستخدام الإكمال بالمتوسط قبل تطبيق RFE. تم تحديد عدد الميزات المراد اختياره (n_features_to_select) بـ 8 ميزات، وهو نفس العدد المستخدم في Chi-Square للمقارنة.

- الميزات المختارة باستخدام RFE: يوضح الجدول التالي الميزات الثماني التي اختارها RFE.

Ranking	Selected	Feature
1	True	sg
1	True	al
1	True	hemo
1	True	pcv
1	True	sc
1	True	bu
1	True	bgr
1	True	pot
2	False	sod
3	False	bp
4	False	appet
5	False	dm

6	False	pe
7	False	htn
...	False	...
17	False	ba

• **ملاحظة:** يظهر الترتيب أن RFE اعتبر الميزات الثماني الأولى (ذات الترتيب 1) هي الأكثر أهمية لإبقائها في النموذج. الميزات ذات الترتيب الأعلى من 1 تم إزالتها تدريجياً.

• شكل البيانات بعد الاختيار: بعد اختيار أفضل 8 ميزات باستخدام RFE، تم تقليل عدد الأعمدة في مجموعتي التدريب والاختبار.

○ الشكل الجديد لمجموعة التدريب بعد اختيار الميزات: (8, 320)

○ الشكل الجديد لمجموعة الاختبار بعد اختيار الميزات: (8, 80)

الجزء الثالث

تم الانتهاء في الجزء الثاني من مرحلة اختيار الميزات باستخدام كلتا الطريقتين. وفي هذا الجزء من التقرير، سيتم تناول تدريب وتقييم نموذج Random Forest باستخدام كل مجموعة من الميزات المختارة، وعرض نتائج الأداء النهائية.

3. تدريب وتقييم نموذج Random Forest

تم تدريب نموذج Random Forest بشكل منفصل باستخدام مجموعتي الميزات التي تم الحصول عليها من طريقتي Chi-Square و RFE. لضمان الحصول على أفضل أداء ممكن من النموذج مع كل مجموعة ميزات، تم استخدام بحث الشبكة (GridSearchCV) لتحسين المعلمات الفائقة لنموذج Random Forest. تم تحديد شبكة من المعلمات الفائقة المحتملة (مثل `n_estimators`, `max_depth`, `min_samples_split`, `min_samples_leaf`, `criterion`) وتم البحث عن أفضل توليفة من هذه المعلمات باستخدام التحقق المتقاطع (Cross-Validation) على بيانات التدريب. تم استخدام مقياس F1-Score الموزون (`f1_weighted`) كمعيار للتقييم أثناء البحث الشبكي، نظراً لكونه مقياساً جيداً للأداء في حالات عدم التوازن المحتمل في الفئات، كما أنه يجمع بين الدقة (Precision) والاستدعاء (Recall). بعد تحديد أفضل المعلمات، تم تدريب النموذج النهائي باستخدام هذه المعلمات على كامل بيانات التدريب (بالميزات المختارة)، ثم تم تقييم أدائه على مجموعة الاختبار غير المرئية باستخدام مقاييس الأداء الشائعة ومصفوفة الارتباك.

3.1. نموذج Random Forest مع ميزات Chi-Square

يوضح هذا القسم أداء نموذج Random Forest الذي تم تدريبه باستخدام الميزات المختارة بواسطة اختبار Chi-Square.

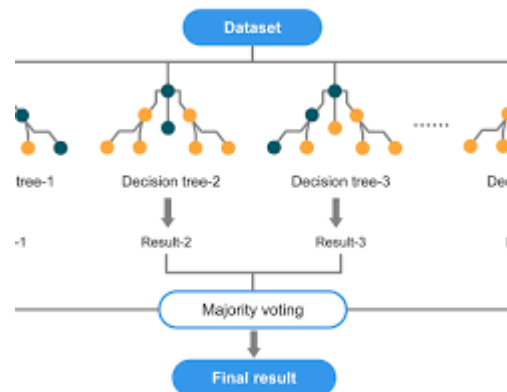
تقرير التصنيف التفصيلي

يقدم الجدول التالي تقريراً مفصلاً عن أداء النموذج لكل فئة على حدة (الفئة 0 والفئة 1) على مجموعة الاختبار.

Support	F1-Score	Recall	Precision	الفئة
50	0.97	0.98	0.96	0
30	0.97	0.95	0.98	1
80	0.98			Accuracy
80	0.97	0.97	0.97	Macro Avg
80	0.97	0.98	0.98	Weighted Avg

مصفوفة الارتباك (Confusion Matrix)

تمثل مصفوفة الارتباك الأداء البصري للنموذج، حيث توضح عدد الأمثلة التي تم تصنيفها بشكل صحيح وخاطئ لكل فئة.

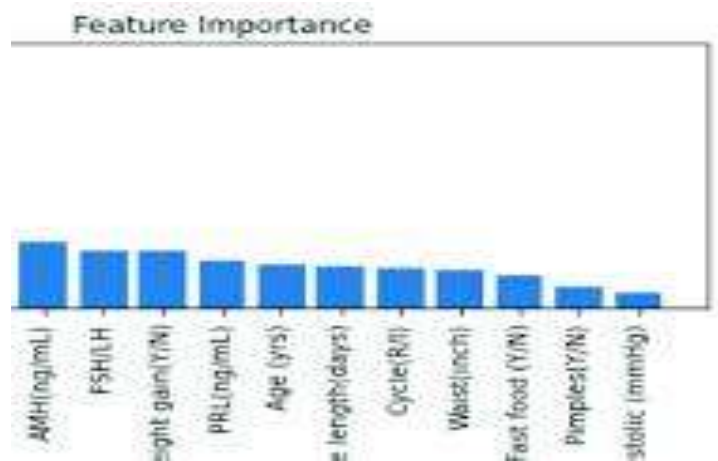


صورة مصفوفة الارتباك (Optimized) Random Forest with ChiSquare Features

شرح المخطط: يوضح هذا المخطط أن النموذج المدرب على ميزات Chi-Square حقق أداءً جيداً في تصنيف كلتا الفئتين، مع عدد قليل من الأخطاء (False Positives و False Negatives).

أهمية الميزات للنموذج المدرب

تم استخلاص ورسم أهمية الميزات من نموذج Random Forest المدرب على ميزات Chi-Square، مما يوضح مساهمة كل ميزة في قرارات النموذج.



صورة أهمية الميزات Random Forest with ChiSquare Features (Optimized)

شرح المخطط: يوضح المخطط أن الميزات مثل hemo, sg, al كانت الأكثر أهمية في تنبؤات نموذج Random Forest المدرب على ميزات Chi-Square، وهو ما يتوافق بشكل كبير مع درجات Chi-Square التي تم الحصول عليها في مرحلة اختيار الميزات.

3.2. نموذج Random Forest مع ميزات RFE

تم تدريب وتحسين نموذج Random Forest باستخدام الميزات الثماني التي تم اختيارها بواسطة RFE.

أفضل المعلمات الفائقة التي وجدها GridSearchCV

بعد إجراء البحث الشبكي، تم تحديد مجموعة المعلمات الفائقة التالية كأفضل تكوين لنموذج Random Forest عند استخدام ميزات RFE:

- أفضل المعلمات: `{ 'class_weight': 'balanced', 'criterion': 'gini', 'max_depth': 12, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300 }`

أفضل درجة (score) تم تحقيقها مع هذه المعلمات

كانت أفضل درجة F1-Score موزونة تم تحقيقها على بيانات التدريب باستخدام التحقق المتقاطع هي:

- أفضل درجة CV (f1_weighted): 0.9918

تقييم الأداء على مجموعة الاختبار

تم تقييم أداء النموذج المدرب بأفضل المعلمات على بيانات الاختبار غير المرئية. يوضح الجدول التالي المقاييس الرئيسية للأداء على مجموعة الاختبار.

المقياس	القيمة
Accuracy	0.9875

0.9878	Precision (Weighted)
0.9875	Recall (Weighted)
0.9874	F1 Score (Weighted)

• تقرير التصنيف التفصيلي

يقدم الجدول التالي تقريراً مفصلاً عن أداء النموذج لكل فئة على حدة على مجموعة الاختبار.

Support	F1-Score	Recall	Precision	الفئة
50	0.99	1.00	0.98	0
30	0.98	0.97	1.00	1
80	0.99			Accuracy
80	0.99	0.98	0.99	Macro Avg
80	0.99	0.99	0.99	Weighted Avg

○ مصفوفة الارتباك (**Confusion Matrix**) تمثل مصفوفة الارتباك الأداء البصري للنموذج المدرب على

على ميزات RFE.
 [Image of مصفوفة الارتباك - Random Forest with RFE Features (Optimized)]
 شرح المخطط: يظهر هذا المخطط أن النموذج المدرب على ميزات RFE حقق أداءً ممتازاً، مع عدد أقل من الأخطاء مقارنة بالنموذج المدرب على ميزات Chi-Square.
 • أهمية الميزات للنموذج المدرب: تم استخلاص ورسم أهمية الميزات من نموذج Random Forest المدرب على ميزات RFE.

[Image of أهمية الميزات - Random Forest with RFE Features (Optimized)]
 شرح المخطط: يوضح المخطط أن الميزات التي اختارها RFE تظهر أيضاً كعوامل مهمة في تنبؤات نموذج Random Forest، مع تركيز على الميزات الأساسية التي تم الاحتفاظ بها.

4. مقارنة النتائج النهائية

تمت مقارنة أداء نموذجي Random Forest المدربين باستخدام مجموعتي الميزات المختارة (Chi-Square و RFE) على بيانات الاختبار لتقييم تأثير طريقة اختيار الميزات على الأداء النهائي. يوضح الجدول التالي ملخصاً للمقاييس الرئيسية لكلا النموذجين على مجموعة الاختبار.

• جدول مقارنة أداء النماذج النهائية على بيانات الاختبار:

ج النموذج	Featu re Selectio n	Accur acy	Precis ion (Weight ed)	Recal l (Weight ed)	F1 Score (Weight ed)	Best CV Score (f1_weig hted)
Random Forest (Optimiz ed)	Chi- Square	0.975 0	0.975 5	0.975 0	0.974 8	0.987 6
Random Forest (Optimiz ed)	RFE	0.987 5	0.987 8	0.987 5	0.987 4	0.991 8

5. الخلاصة والتوصيات

- بناءً على التحليل والنتائج المعروضة في هذا التقرير، يمكن استخلاص النتائج التالية:
 - مقارنة طرق اختيار الميزات: عند مقارنة الميزات الثماني التي اختارها Chi-Square وتلك التي اختارها RFE، نلاحظ تداخلاً كبيراً في الميزات المختارة. كلتا الطريقتين حددتا ميزات مثل *sgal*، و *hemo* كأكثر أهمية، مما يشير إلى أن هذه الميزات لها ارتباط قوي بالمتغير الهدف وهي مفيدة للتصنيف بغض النظر عن الطريقة المستخدمة. ومع ذلك، اختارت كل طريقة أيضاً بعض الميزات المختلفة، مما يعكس الطبيعة المختلفة للمنهجيتين (الارتباط الإحصائي المستقل للميزات في Chi-Square مقابل الأهمية السياقية ضمن نموذج Random Forest في RFE).
 - مقارنة أداء النموذج: أظهر نموذج Random Forest المدرب على الميزات المختارة بواسطة RFE أداءً أفضل قليلاً على مجموعة الاختبار مقارنة بالنموذج المدرب على ميزات Chi-Square، وذلك بناءً على جميع مقاييس التقييم الرئيسية (Accuracy, Precision, Recall, F1-Score). يشير هذا إلى أن مجموعة الميزات التي اختارها RFE كانت أكثر ملاءمة لنموذج Random Forest لتحقيق أداء تعميم أفضل على البيانات غير المرئية. يمكن أن يعزى هذا إلى أن RFE، كطريقة مغلف، تأخذ في الاعتبار كيفية تفاعل الميزات مع النموذج المحدد (Random Forest) أثناء عملية الاختيار.
 - أهمية الميزات: تتوافق مخططات أهمية الميزات للنماذج المدربة مع نتائج اختيار الميزات. الميزات التي تم اختيارها بواسطة كل طريقة ظهرت كأكثر الميزات أهمية في النموذج المقابل. هذا يؤكد أن عملية اختيار الميزات كانت فعالة في تحديد الميزات الأكثر تأثيراً في تنبؤات نموذج Random Forest.

- تأثير التحسين: ساهم استخدام GridSearchCV في تحسين أداء نموذج Random Forest لكلا مجموعتي الميزات، حيث تم العثور على معلمات فائقة أدت إلى درجات CV عالية، مما انعكس على أداء جيد على مجموعة الاختبار.

بناءً على هذه النتائج، نوصي باستخدام مجموعة الميزات التي تم اختيارها بواسطة طريقة RFE لنموذج تصنيف Random Forest في التطبيقات المستقبلية، حيث أظهرت هذه المجموعة أداءً متفوقاً قليلاً على بيانات الاختبار في هذا المشروع.

التوصيات المستقبلية:

- لتحسين هذا المشروع بشكل أكبر، نقترح التوصيات المستقبلية التالية:
- استكشاف طرق إكمال القيم المفقودة: تجربة طرق أكثر تعقيداً لإكمال القيم المفقودة (مثل الإكمال باستخدام نماذج تعلم آلي أخرى أو طرق متعددة المتغيرات) ومقارنة تأثيرها على أداء النموذج.
- تقييم نماذج تصنيف أخرى: تطبيق ومقارنة أداء نماذج تصنيف أخرى (مثل Support Vector Machines, Gradient Boosting Machines) باستخدام أفضل مجموعات الميزات التي تم تحديدها.
- تحليل تفاعلات الميزات: إجراء تحليل أعمق لتفاعلات الميزات لفهم كيفية تأثير مجموعات معينة من الميزات على التنبؤات.
- استخدام مقاييس تقييم إضافية: النظر في مقاييس تقييم إضافية مثل منحنى ROC-AUC ومنحنى Precision-Recall لتقييم أداء النموذج بشكل أكثر شمولاً، خاصة إذا كان هناك اهتمام خاص بالدقة أو الاستدعاء.
- زيادة حجم شبكة البحث: في حال توفر موارد حاسوبية كافية، يمكن توسيع نطاق شبكة البحث في GridSearchCV أو استخدام طرق تحسين أخرى أكثر كفاءة مثل Random Search أو Bayesian Optimization للعثور على معلمات فائقة أفضل.
- تحليل عدم توازن الفئات: إذا كان عدم توازن الفئات يمثل مشكلة، فيمكن استكشاف تقنيات مثل (Oversampling (SMOTE أو Undersampling لتحسين أداء النموذج على الفئة الأقلية.

انتهى التقرير.... مع خالص الشكر والتقدير.

[أسماء أعضاء الفريق]