

الجمهورية العربية السورية
المعهد العالي للعلوم التطبيقية والتكنولوجيا
قسم النظم المعلوماتية

نموذج دعم قرار للتنبؤ بمستوى الخطورة لمرضى القلب
**Decision support model for predicting the risk
for patients with heart disease**

أعدت هذه الأطروحة لنيل درجة الماجستير في نظم المعلومات ودعم القرار

إعداد

م. أنس عيسى

إشراف

د. محمد الشايطة

د. مصطفى دقاق

المعهد العالي للعلوم التطبيقية والتكنولوجيا

Higher Institute for Applied Sciences and Technology

Institut Supérieur des Sciences Appliquées et de Technologie

مؤسسة سورية حكومية للتعليم العالي أحدثت في عام 1983، بهدف إعداد أطر متميزة مؤهلة للبحث العلمي والتطوير في مجال العلوم التطبيقية والتقانة، لتساهم بفاعلية في التنمية العلمية والصناعية والاقتصادية في القطر. يشكل التأهيل الهندسي والدراسات العليا في المعهد العالي محور عمليّة إعداد الأطر المتخصّصة. يخرّج المعهد العالي مهندسين متميزين، بعد دراسة لمدة خمس سنوات، في اختصاصات الاتصالات والمعلوماتية والنظم الإلكترونية والميكاترونكس وهندسة الطيران وعلوم وهندسة المواد. كما يمنح المعهد العالي درجة الماجستير الأكاديمي، ماجستير بحثي يمتد على سنتين، من خلال مجموعة من برامج الماجستير في نظم الاتصالات وفي التحكم والروبوتيك وفي علوم المواد وفي نظم المعلومات واتخاذ القرار وفي نظم معالجة المعطيات الكبيرة. وأخيراً يمنح المعهد العالي درجة الدكتوراه في عدة اختصاصات موازية لما ذكر في برامج الماجستير. يعتمد المعهد العالي في تميزه على تركيزه على النوع وليس الكم، فهو ينتقي في المرحلة الهندسية شريحة الطلاب المتفوقين في شهادة الدراسة الثانوية السورية من الفرع العلمي أو من في حكمهم. أما في مرحلة الماجستير فيقبل المعهد العالي حملة الشهادات الجامعية الموازية للماجستير المطلوب، وذلك على أساس مفاضلة خاصة لاختيار الأفضل. كما يقدم المعهد للطلاب جواً متميزاً للدراسة والبحث بدءاً من كوادره المتفرغة عالية التأهيل ومناهجه المواكبة للتطورات العلمية، وانتهاءً بإمكانيات مخبراته المتميزة في القطر وبنيتها التحتية الموازية من صالات حواسب وورش ومقدرات مادية وشبكات تعاون مع الصناعة والهيئات الأكاديمية داخل وخارج القطر. كل ذلك في وسائل للراحة والترفيه من سكن طلابي مطعم وصالات رياضية وملاعب ونشاطات طلابية متنوعة. بالإضافة إلى نشاطه الأكاديمي، يضم المعهد العالي أقساماً علمية ومخابر متنوعة ومراكز تكنولوجية، كمخبر الدراسات البيئية ومركز تقانات اللحام ومركز الحوسبة عالية الأداء. تقدم هذه الفعاليات خدمات واستشارات للقطاعين العام والخاص، بالإضافة إلى المشاريع التطويرية والنشاطات البحثية والدورات التدريبية وتنظيم ورش العمل والمؤتمرات العلمية.

ﺗﺼﺮﯨﺢ

أنا الموقع أدناه أنس مالك عيسى، معدّ أطروحة الماجستير التي تحمل العنوان:

نموذج دعم قرار للتنبؤ بمستوى الخطورة لمرضى القلب

Decision support model for predicting the risk for patients with heart disease

أصّرّح بأن:

الأعمال والنتائج المعروضة في هذه الأطروحة، هي نتيجة مجهودي الشخصي، وتحت إشراف كامل من قبل المشرفين، وقد نسب ما عدا ذلك من معلومات ونتائج متعلقة بالبحث إلى مصادرها، وأشير إليها بشكل صريح في متن النص وقائمة المراجع.

كلمة شكر

من أيّ أبواب الثناء سندخل وبأي أبيات القصيد نعبر، وفي كل لمسة من جودكم
وأكفكم للمكرّمات أسطر.

إلى الأيادي المعطاءة، إلى السادة المشرفين د. مصطفى الدقاق و د. محمد الشايطة
أتوجه إليكم بخالص الشكر والعرفان لما بذلتموه من جهود في توجيهي وتعليمي في
مراحل هذا البحث.

كما أتوجه بخالص الامتنان إلى أساتذتي الأفاضل في ماجستير نظم المعلومات ودعم
القرار، لما قدموه لنا من علم ومعرفة.

والشكر موصول إلى المعهد العالي للعلوم التطبيقية والتكنولوجيا بجميع كوادره من
مدرسين وإداريين لعطائهم الذي لا ينضب.

م. أنس عيسى

الإهداء

إلى معلمي الأول، قدوتي وسندي، صاحب أعظم قلب أعرفه.
والدي العظيم: م. مالك عيسى.

إلى أيقونة المحبة والحنان، التي رافقتي دعاؤها وعطاؤها في كل خطوة في حياتي.
أمي الغالية: سناء القطريب

إلى معنى السعادة البهجة، الذين ينثرون عطر وجودهم الجميل في أيامي.
أخواتي: ميس ويارا

إلى رفاق الدرب وأصدقاء الرحلة، الذين يحلو كل شيء معهم.
أصدقائي

الملخص

تعتبر أنظمة دعم القرار الطبي القائمة على أساس خوارزميات تعلم الآلة والتعلم العميق من أكثر فروع الذكاء الصناعي التي تتلقى الاهتمام حالياً، نظراً لدورها الكبير في تقديم المساعدة للطبيب والمريض على حدٍ سواء. ويعتبر تصميم أنظمة دعم قرار طبي لتخمين درجة خطورة الإصابة بالفشل القلبي لمرضى القلب أمراً مهماً جداً وملحاً كونها تحفظ حياة المرضى وتساعد الأطباء في تحديد مسار العلاج.

نقدم في البحث الحالي نظام دعم قرار طبي لمرضى الفشل القلبي لمساعدة الطبيب في تحديد درجة الخطورة باستخدام مجموعة بيانات طبية مؤلفة من 167 عمود (معلومة طبية) و2008 حالة طبية.

يتم في المرحلة الأولى معالجة مجموعة البيانات ثم يتم تطبيق خوارزمية موازنة مقترحة من قبلنا لموازنة نسب أصناف مجموعة البيانات تعتمد على مرحلتين الأولى هي العنقدة لتحديد نقاط التداخل بين الأصناف وتقليلها والثانية تتضمن زيادة العينات الممثلة للأصناف غير المسيطرة بشكل ديناميكي يعتمد على درجة الارتباط مع عمود الهدف. وفي المرحلة التالية يتم تطبيق خوارزمية اختيار سمات لتحديد الأعمدة المشاركة في عملية التقييم وحذف الأعمدة غير الضرورية. يتم في المرحلة التالية تدريب ثلاث أنظمة تعلم آلة هي أشجار القرار وغابات القرار والانحدار المنطقي وبناء وتدريب نموذج تعلم عميق يعتمد طبقات Dense Layers. تم بناء هذه النماذج مرة بدون تحسين ومرة أخرى بعد تطبيق خوارزمية تحسين لضبط بارامترات التدريب واختيار الأفضل منها.

أما في المرحلة الأخيرة يتم تقييم أداء الأنظمة المدربة سواء قبل أو بعد الموازنة وقبل أو بعد اختيار السمات وقبل أو بعد التحسين.

تبين النتائج العملية أن نظام LR هو أفضل نموذج تعلم آلة حيث حقق دقة 92.67% بعد تطبيق التحسين في حين حقق نموذج التعلم العميق المحسن دقة 98.61%. أدت عملية اختيار السمات إلى اختيار 47.7% فقط من كامل السمات مع انخفاض طفيف في الأداء وتحسين في الزمن بمقدار 0.83% لكل خطوة من خطوات تكرارات التدريب.

تم التحقق من خوارزمية موازنة البيانات المقترحة من خلال تطبيقها على مجموعة البيانات المستخدمة في البحث وعلى مجموعة بيانات طبية أخرى حيث تحسن أداء نماذج تعلم الآلة بعد تطبيق الموازنة في كلا الحالتين.

كلمات مفتاحية:

نظم دعم القرار الطبي، كشف الفشل القلبي، تعلم الآلة، التعلم العميق، موازنة البيانات.

Abstract

Medical decision support-based machine learning and deep learning system is one of the most artificial intelligence fields that are receiving attention in the current time, due to its essential role to introduce help for physician and patients. Design of medical decision support system to predict the risk of heart failure is very essential part since it saves patients' lives and helps physicians to guide the path of cure. In the current research, we introduce a medical decision support system for heart failure patients to identify the risk degree using a medical dataset consisting of 167 columns (medical information) and 2008 medical cases. In the first step, the dataset is preprocessed and a proposed balancing algorithm is applied to balance the classes ratios. In the second step, a feature selection algorithm is applied to define the participating columns in the final prediction and remove the redundant ones. In the next step, 3 different machine learning classifiers (decision trees, random forests, logistic regression) are trained and one proposed dense-layers based deep learning model is also build and trained. These models are built with and without optimization (the optimization is used to tune the training parameters). In the final step, the trained models are evaluated (before and after balancing, before and after feature selection and before and after optimization). The experiments show that the LR model is the best ML model with 92.67% accuracy, while the deep learning model achieves 98.61% accuracy. The feature selection process leads to only 47.7% of the entire features. However, it caused a little bit performance degradation but it decreased the training time by 0.84 MS for each training step.

The balancing algorithm has been validated via applying it on two different medical datasets; the one which is used in our research and another medical one. The results proved that the performance of the ML models was enhanced after applying the balancing algorithms in both datasets.

Keywords: Medical support systems, Hear Failure Detection, Machine Learning, Deep Learning, Data balancing.

قائمة المحتويات

1	الفصل الأول: مقدمة
3	1.1. تمهيد
3	2.1. إشكالية ودوافع البحث
5	3.1. فكرة الحل المقترح
6	4.1. تحديات البحث
7	5.1. المساهمات الأساسية في البحث
10	6.1. مخطط البحث
11	2. الفصل الثاني: الدراسة النظرية
13	1.2.1. مقدمة
13	2.2. ما هو الفشل القلبي
14	3.2. أنظمة دعم القرار الطبي
15	1.3.2. نظام دعم القرار الطبي لمرض الفشل القلبي
17	2.3.2. تقنيات تعلم الآلة المستخدمة في أنظمة كشف الفشل القلبي
18	3.3.2. مجموعات البيانات الطبية لكشف الفشل القلبي
19	4.2. الخلاصة
21	3. الفصل الثالث: الدراسة المرجعية
23	1.3.1. مقدمة

2.3. الدراسات المرجعية في مجال تصميم أنظمة دعم القرار الطبي لتشخيص مرض الفشل القلبي.....	23
3.3. الخلاصة.....	29
4. الفصل الرابع: الدراسة التجريبية والنماذج المقترحة.....	31
1.4. مقدمة.....	33
2.4. المخطط المقترح للبحث.....	33
3.4. مجموعة البيانات المستخدمة في البحث.....	35
1.3.4. خطوات المعالجة المسبقة لمجموعة البيانات.....	35
2.3.4. موازنة مجموعة البيانات Dataset Balancing.....	43
3.3.4. خوارزمية موازنة مجموعة البيانات المعتمدة على تقليل التداخل بين الأصناف.....	47
4.4. خطوات عملية تدريب نماذج تعلم الآلة مع التحسين.....	48
5.4. نموذج التعلم العميق المقترح مع طريقة التحسين المقترحة.....	50
6.4. طريقة تقييم الأداء (معاملات قياس الأداء).....	53
6.4. الخلاصة.....	53
5. الفصل الخامس: الاختبارات والنتائج.....	55
1.5. مقدمة.....	57
2.5. سيناريوهات التدريب المقترحة.....	57
1.2.5. التدريب على مجموعة البيانات الأصلية.....	57

60	2.2.5. التدريب مع الموازنة باستخدام SMOTE
66	3.2.5. تدريب مصنفات تعلم الآلة بدون خوارزمية التحسين
67	4.2.5. تدريب مصنفات تعلم الآلة مع خوارزمية التحسين
69	5.2.5. تدريب نموذج التعلم العميق بدون تحسين
71	6.2.5. التدريب باستخدام إعادة التشكيل (SMOTE):
75	7.2.5. إعادة تشكيل مجموعة البيانات بالاعتماد على السمات الأكثر ارتباطاً بالهدف
79	8.2.5. التدريب مع الموازنة عبر إعادة تشكيل مجموعة البيانات باستخدام نموذج تقدير كثافة النواة.....
83	9.2.5. تدريب نموذج التعلم العميق مع تحسين
85	10.2.5. نموذج التعلم العميق مع اختيار السمات
89	11.2.5. إعادة التدريب بعد تطبيق خوارزمية الموازنة MUDSFC
92	12.2.5. إعادة تدريب واختبار نماذج تعلم الآلة النسخة المحسنة من خوارزمية الموازنة MUDSFC بإضافة مرحلة كشف العينات المتداخلة بين الأصناف
94	13.2.5. التدريب مع الموازنة بالطريقة المقترحة من قبلنا
103	14.2.5. التدريب مع الموازنة بالطريقة المقترحة من قبلنا والتحسين
107	3.5. مقارنة الدراسة الحالية مع الدراسات السابقة.....
110	6. الفصل السادس: الخاتمة والآفاق المستقبلية.....
112	1.6. الخاتمة.....

114	2.6. الآفاق المستقبلية
116	7. الملاحق.
118	1.7. الملحق أ: وصف مجموعة البيانات.
125	8. المراجع.

قائمة الأشكال

- الشكل (2-1) الخطوات العامة لأي نظام دعم قرار طبي لكشف الفشل القلبي..... 18
- الشكل (1-4) مخطط النظام المقترح..... 34
- الشكل (2-4) خطوات المعالجة المسبقة لمجموعة البيانات..... 37
- الشكل (3-4) المخطط الانسيابي الموضح لخوارزمية MUDSFC..... 46
- الشكل (4-4) التعديل في خوارزمية MUDSFC بإضافة مرحلة تقليل التداخل..... 47
- الشكل (5-1) الموازنة باستخدام SMOTE قبل الأمثلة..... 61
- الشكل (5-2) الموازنة باستخدام SMOTE مع الأمثلة..... 64
- الشكل (3-5) مصفوفات التشتت CONFUSION MATRIXES لنماذج تعلم الآلة الثلاثة المدربة بدون تحسين..... 66
- الشكل (4-5) مصفوفات التشتت CONFUSION MATRIXES للنماذج الثلاثة المدربة مع تحسين..... 68
- الشكل (5-5) ناتج عملية التدريب: منحنى الدقة ACCURACY ومنحنى الخسارة Loss..... 70
- الشكل (6-5) مصفوفة التشتت لنموذج التعلم العميق قبل تطبيق خوارزمية التحسين RANDOM SEARCH..... 70
- الشكل (5-7) نتائج التدريب بإعادة التشكيل باستخدام SMOTE قبل الأمثلة..... 72
- الشكل (5-8) نتائج التدريب بإعادة التشكيل باستخدام SMOTE بعد الأمثلة..... 74
- الشكل (5-9) نتائج إعادة التشكيل باستخدام MUDFSCS قبل الأمثلة..... 76
- الشكل (5-10) نتائج إعادة التشكيل باستخدام MUDFSCS بعد الأمثلة..... 78
- الشكل (5-11) نتائج إعادة التشكيل باستخدام KDE قبل الأمثلة..... 81
- الشكل (5-12) نتائج إعادة التشكيل باستخدام KDE بعد الأمثلة..... 82
- الشكل (13-5) مصفوفة التشتت لنموذج التعلم العميق بعد تطبيق خوارزمية التحسين RANDOM SEARCH..... 84
- الشكل (14-5) منحنيات الدقة والخسارة مع عدد التكرارات للنموذج المدرب على السمات المختارة عند: (أ) عتبة 0.02، (ب) عتبة 0.05، (ج) عتبة 0.08..... 86
- الشكل (15-5) مصفوفة التشتت عند عتبة اختيار سمات: (أ) 0.02، (ب) 0.05، (ج) 0.08..... 88
- الشكل (5-16) الموازنة بالطريقة المقترحة قبل الأمثلة..... 98
- الشكل (5-17) الموازنة بالطريقة المقترحة بعد سبع دورات من الحذف..... 100
- الشكل (5-18) الموازنة بالطريقة المقترحة بعد الأمثلة..... 104
- الشكل (19-5) توزيع نقاط بيانات مجموعة البيانات المستخدمة قبل وبعد تطبيق مرحلة حذف النقاط المتداخلة اعتماداً على خوارزمية DBSCAN..... 105
- الشكل (20-5) توزيع نقاط بيانات قبل وبعد تطبيق مرحلة حذف النقاط المتداخلة اعتماداً على خوارزمية العنقدة DBSCAN..... 106

قائمة الجداول

9	الجدول (1-1) العمود الهدف
19	الجدول (1-2) أهم مجموعات البيانات الطبية لمرض الفشل القلبي
38	الجدول (1-4) تحويل البيانات الفئوية إلى بيانات رقمية
39	الجدول (2-4) السمات غير ذات الصلة
41	الجدول (3-4) السمات منخفضة التباين
53	الجدول (4-4) معاملات قياس الأداء
58	الجدول (5-1) التحقق المتقاطع لمصنف الغابة العشوائية
58	الجدول (5-2) نتائج اختبار على مجموعة البيانات المعالجة بدون موازنة أو تحسين
59	الجدول (5-3) نتائج تدريب المصنفات مع الأمثلة
62	الجدول (5-4) نتائج التحقق على مجموعة البيانات الموازنة باستخدام SMOTE
63	الجدول (5-5) شبكة المعلمات الفائقة
65	الجدول (5-6) نتائج التحقق باستخدام SMOTE مع الأمثلة
66	الجدول (7-5) نتائج اختبار مصنفات تعلم الآلة المدربة بدون تحسين
67	الجدول (8-5) نتائج اختبار نماذج تعلم الآلة المدربة بعد تطبيق عملية التحسين
71	الجدول (9-5) نتائج اختبار نموذج التعلم العميق المدرب قبل تطبيق عملية التحسين OPTIMIZATION
73	الجدول (5-10) نتائج التدريب بإعادة التشكيل باستخدام SMOTE قبل الأمثلة
77	الجدول (5-11) نتائج إعادة التشكيل باستخدام MUDFSCS قبل الأمثلة
84	الجدول (12-5) نتائج اختبار نموذج التعلم العميق المدرب بعد تطبيق عملية التحسين OPTIMIZATION
87	الجدول (13-5) مقارنة أداء نظام التعلم العميق المقترح في ظل عدة عتبات لاختيار السمات
	الجدول (14-5) مقارنة بين نتائج مصنفات تعلم الآلة ومصنف التعلم العميق قبل وبعد عملية التحسين وباستخدام طريقة OVER SAMPLING BALANCING ومن ثم باستخدام الخوارزمية MUDSFC المصممة من قبلنا
91	
107	الجدول (5-15) نتائج تقييم أداء مصنفات تعلم الآلة بدون ومع خوارزمية موازنة السمات المقترحة في البحث
108	الجدول (16-5) مقارنة الدراسة الحالية مع الدراسات السابقة

قائمة الاختصارات والمصطلحات

المصطلح	المعنى باللغة الإنكليزية	المعنى باللغة العربية
AI	Artificial intelligence	الذكاء الصناعي
CDSS	Clinical Decision Support System	أنظمة دعم القرار السريري
CVD	Cardiovascular Disease	أمراض القلب والأوعية الدموية
DBSCAN	Density-based spatial clustering of applications with noise	التجميع المكاني القائم على الكثافة للتطبيقات ذات الضوضاء

التعلم العميق	Deep Learning	DL
أشجار القرار مصنف	Decision Tree	DT
الصورة البيانية الكهربائية للقلب	Electrocardiogram	ECG
السجل الصحي الإلكتروني	Electronic Health Record	HER
مصنف الانحدار اللوجستي	Logistic Regression	LR
مصنف الغابة العشوائية	Random Forest	RF
الاستبعاد العودي للسمات	Recursive Feature Elimination	REF
تابع الوحدة الخطية المصححة	Rectified Linear Unit	RELU
تقنية توليف عينات أقلية	Synthetic Minority Over-sampling Technique	SMOTE
مستودع بيانات جامعة كاليفورنيا	the University of California Irvine machine learning repository	UCI

1. الفصل الأول: مقدمة

يقدم هذا الفصل تعريف بالإطار العام ودوافع وإشكالية البحث.

1.1.1. تمهيد

أدى التطور الكبير في تكنولوجيا المعلومات إلى تطور المجالات الأخرى بنفس الدرجة نظراً للأهمية البالغة في استخدام تقنيات المعالجة وخوارزميات الحاسب في العديد من تطبيقات الحياة العملية.

كان للطب الحصة الأكبر في هذه التطبيقات نظراً للحاجة الماسة لتطبيقات الحاسب في دعم القرار الطبي وتقديم المساعدة الطبية وتسهيل العمل الطبي.

إن أنظمة دعم القرار الطبي تمثل دوراً هاماً في المجال الطبي نظراً لما توفره من عناء وجهد ووقت على الطبيب والمريض بنفس الوقت كما أنها تساهم في الكشف عن العديد من المعلومات الطبية التي تساهم في الحفاظ على سلامة المرضى وتعزيز معلومات الطبيب.

تعتمد أنظمة دعم القرار الطبي على أمرين أساسيين هما مجموعات البيانات الطبية الضخمة وخوارزميات تعلم الآلة القوية التي تستطيع معالجة البيانات الضخمة واتخاذ القرار الطبي بشكل سريع وفعال.

نحاول في البحث الحالي تقديم عدة مساهمات بحثية جديدة في مجال تصميم أنظمة دعم القرار الطبي بدءاً من استخدام مجموعة بيانات طبية جديدة تتضمن تحدي متعلق باستخدام عدد كبير من الخصائص Predictors وتتطلب موازنة في الأصناف إضافة لحاجتها للعديد من عمليات المعالجة، أما التحدي الثاني فيمكن في تصميم نظام موازنة خاص لمجموعة البيانات وتصميم نظام تعلم عميق خاص بها.

2.1. إشكالية ودوافع البحث

يُعتبر مرض القلب أحد أكثر الأمراض البشرية تعقيداً وفتناً بالحياة حول العالم. في هذا المرض يكون القلب عادةً غير قادر على دفع الكمية المطلوبة من الدم إلى أجزاء أخرى من الجسم

لأداء الوظائف الطبيعية، ونتيجة لذلك يحدث قصور القلب [1]. هذا يجعل من عملية تصميم أنظمة دعم قرار طبي للتحقق من خطورة الفشل القلبي Medical Decision Support Systems for Heart Failure Prediction أمراً حتمياً، حيث تساعد هذه الأنظمة الأطباء في اتخاذ القرار الطبي سواء لإعطاء الأدوية أو لإجراء عمليات جراحية مستعجلة للمرضى عندما تشير هذه الأنظمة إلى أن خطورة الفشل القلبي وفقاً للوضع الحالي للمريض هي عالية. تساعد مثل هكذا أنظمة الطبيب على السرعة في اتخاذ القرار كما أنها تقلل من مخاطر الفشل القلبي لاحقاً كونها بمثابة تنبيه مبكر للمريض والطبيب، وبسبب الكلفة العالية للفحوصات الدورية والتحاليل المستمرة فإن مثل هذه الأنظمة توفر المال كما أنها توفر الجهد [2] [3].

لحسن الحظ فإن خوارزميات علوم الحاسب بما فيها خوارزميات تعلم الآلة Machine Learning (ML) وخوارزميات التعلم العميق (DL) Deep Learning قادرة على معالجة البيانات الضخمة التي تأتي من الملفات الطبية الخاصة بمرضى القلب، كما أنها قادرة على اتخاذ القرار الطبي حول درجة خطورة الفشل القلبي، وهذا الأمر كان من غير الممكن قبل ظهور هذه التقنيات.

هناك العديد من نظم دعم القرار الطبي الخاصة بتشخيص أو تخمين الفشل القلبي لكن معظمها تتضمن العديد من المشاكل مثل استخدام مجموعات بيانات صغيرة الحجم أو قليلة الأبعاد. هناك بعض هذه الأنظمة التي استطاعت تشخيص الإصابة بالفشل القلبي لكن دون تحديد المدة التي يمكن أن يقضيها المريض قبل حدوث الفشل القلبي.

تعاني الأنظمة السابقة أيضاً من مشكلة اختيار الخصائص المناسبة لاتخاذ القرار الطبي ولحل هذه المشكلة لجأت بعض الأبحاث لطريقة اختيار السمات لكنها كانت على حساب الدقة.

فيما يلي اقتباس من مجلة (Frontiers)، حول مرض القلب وضرورة دراسته ضمن سياق من الأعراض والتاريخ المرضي، وليس كجزيرة معزولة، (تظل أمراض القلب والأوعية الدموية

السبب الرئيسي للوفاة والاعتلال المشترك في جميع أنحاء العالم. حتى الآن، تتضمن الاستراتيجية الطبية الشائعة لمواجهة الأمراض وجهة نظر متخصصة، تركز على تشخيص ودراسة وعلاج كل حالة مرضية، بما في ذلك أمراض القلب والأوعية الدموية، باعتبارها كيانًا منفردًا و متميزًا. وبالتالي، يهتم اختصاصيو الرعاية الصحية عمومًا بمرض واحد محدد، وفقًا لتخصصهم. ومع ذلك، يتصرف جسم الإنسان بشكل مختلف فيما يتعلق بهذا الرأي، وتشير البيانات المتزايدة إلى أن العديد من الحالات المرضية، التي تعتبر تقليديًا غير مرتبطة (مثل أمراض القلب والأوعية الدموية، والأمراض العصبية التنكسية، والسرطان، وهشاشة العظام)، تتفاعل فعليًا وتتشارك في العديد من المسارات الخلوية الأساسية والمؤشرات الحيوية الكيميائية الحيوية والجزئية. قد تؤدي بعض عوامل الخطر إلى أمراض القلب والأوعية الدموية في موضوع واحد، ولكنها قد تؤدي إلى حالة أخرى في شخص آخر، أو حتى المزيد من الأمراض في نفس المرضى. قد تكون المعرفة الأفضل للاختلافات / أوجه التشابه بين الأمراض، والتي يُفكر فيها عمومًا وتُدار ككيانات منفصلة، مهمة لفهم الحالات البشرية الفيزيولوجية المرضية وعلاجها بشكل أفضل. في هذا السياق، من الضروري اتباع نهج شامل لتقدير مخاطر المرض بشكل أفضل، مع الأخذ في الاعتبار أنه لا ينبغي النظر إلى كل عضو على أنه "جزيرة"، وأن الفهم الأفضل لتعقيد الأسباب الكامنة وراء الظروف المختلفة يجب أن يعيد توجيه انتباه الأطباء والباحثين للعمل على مستويات متعددة، من أجل تطوير استراتيجيات وقائية وعلاجية ممكنة فعالة لحالات متعددة، من أجل رعاية المريض الشخصية، ما يدفعنا إلى محاولة استكشاف مجموعات بيانات جديدة، أكبر وأكثر تنوعًا.

3.1. فكرة الحل المقترح

في البحث الحالي لا نقدم فكرة تخمين الإصابة بالفشل القلبي المسبب للوفاة مباشرة وإنما في ثلاثة مراحل هي الفشل القلبي خلال 28 يوم ثم خلال 3 أشهر ثم خلال 6 أشهر من تاريخ إدخال المريض للمشفى.

يتم في البحث الحالي أيضاً حل مشاكل مجموعات البيانات الطبية من خلال تطبيق طريقة اختيار السمات (الخصائص) مرتبطة بالخاصية الهدف فضلاً عن موازنة مجموعة البيانات بطريقة مبتكرة وجديدة بدون التأثير على دقة التشخيص وبدون حذف معلومات مهمة.

من أجل أخذ كل مشكلات أنظمة التشخيص الطبي لمرض الفشل القلبي تم استخدام مجموعة بيانات طبية جديدة تتضمن كل المشكلات التي يمكن أن تواجه أنظمة دعم القرار الطبي في العالم الحقيقي مع محاولة حل هذه المشاكل سواء بعمليات المعالجة المسبقة، أو اختيار السمات المرتبطة بالهدف أو موازنة السمات بالاعتماد على طريقة مبتكرة تحافظ على المعلومات الطبية دون خلل.

4.1. تحديات البحث

يكن التحدي الأساسي في البحث الحالي في معالجة مجموعة البيانات الطبية الخاصة بمرض الفشل القلبي حيث تتضمن مجموعة البيانات المشكلات الآتية:

أ. عدم الموازنة بين أصناف عمود الهدف: حيث الصنف 0 والذي يمثل الحالة الطبيعية أو حالة عدم حدوث فشل قلبي خلال فترة قضاء المريض ضمن المشفى يمتلك أكبر عدد من العينات ضمن مجموعة البيانات مما يؤثر في عملية التصنيف على باقي الأصناف وهي:

- الصنف رقم 1 وتمثل الموت نتيجة الفشل القلبي ضمن 28 يوماً من تاريخ قبول المريض.
- الصنف رقم 2 وتمثل الموت نتيجة الفشل القلبي ضمن 3 أسابيع من تاريخ قبول المريض.
- الصنف رقم 3 وتمثل الموت نتيجة الفشل القلبي ضمن 6 أسابيع من قبول المريض.

ب. وجود عدد ضخم من الأعمدة أو الخصائص والتي تمثل السمات التي يجب استخدامها لتخمين الفشل القلبي وعددها 167 عمود.

ج. وجود العديد من القيم الفارغة والقيم المفقودة ضمن مجموعة البيانات والتي يجب معالجتها.

5.1. المساهمات الأساسية في البحث

نُحَصِّص المساهمات الأساسية المقدمّة في هذا البحث بالنقاط التالية:

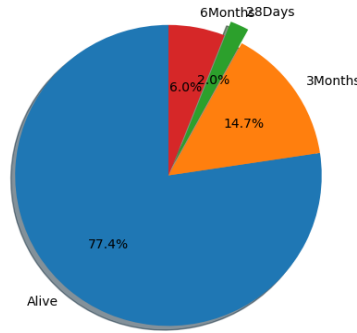
تم في البحث استخدام مجموعة بيانات طبية لتشخيص الفشل القلبي وما يميز مجموعة البيانات المستخدمة أنها تتضمن تحدي خاص يتمثل في العدد الكبير للخصائص (الأعمدة) Predictors المستخدمة لإنجاز التشخيص النهائي وهذه المسألة تسمى مسألة تصنيف ضمن فضاء معقد الأبعاد Complexity Dimensional Classification Problem، حيث يبلغ عدد أعمدة مجموعة البيانات 167 عمود.

المساهمة الأساسية في البحث الحالي تكمن في بناء نموذج دعم قرار قادر على اتخاذ قرار تحديد درجة خطورة المريض (هل سيموت المريض خلال 28 يوماً؟ أم هل ستتضاعف الخطورة خلال ثلاثة أشهر ويتعرض للموت؟ أم هل ستكون خطورة الموت خلال الستة أشهر القادمة؟ أم أن المريض سيتحسن ويبقى على قيد الحياة؟). كل الأبحاث السابقة كانت تصمم أنظمة دعم القرار على أساس اتخاذ قرار واحد هل سيعيش المريض أم سيموت؟ أما في البحث الحالي سنقوم بتقييم درجة خطورة الإصابة أولاً خلال 28 يوم ثم خلال 3 أشهر ثم خلال 6 أشهر من تاريخ قبول المريض في المشفى ونحدد وفقاً لتحليل بيانات المريض مستوى الخطوة لديه). الجدير بالذكر أنّ مجموعة البيانات المستخدمة لم يتم تطبيق أي نظام تعلم آلة أو نظام تعلم عميق عليها في أي بحث سابق وهذا يعتبر تحدي للبحث الحالي في إمكانية بناء نظام دعم قرار بالاعتماد على مجموعة البيانات المقترحة.

إنّ مجموعة البيانات بشكلها الحالي لا يمكن استخدامها مباشرة لبناء نظام دعم قرار أبداً نظراً لأنها تتضمن العديد من المشاكل التي اقترحنا حلها وفق الآتي:

1. تكمن الخطوة الأولى المهمة في معالجة مجموعة البيانات وتقليل الأبعاد واختصار عدد الأعمدة بحيث نبقى على الأعمدة المهمة للمشاركة في بناء نظام دعم القرار.

2. تكمن المشكلة الثانية في أن مجموعة البيانات غير متوازنة فالصنف الذي يمثل حالة الوفاة فيها يتضمن عدد عينات قليل جداً، موزعة بالشكل التالي: (5:3, 15:2, 37:1) في حين أن الصنف الممثل للبقاء على قيد الحياة هو المسيطر حيث يبلغ عدده 1951 حالة بقاء على قيد الحياة، لذا لا بد من عملية Balancing لإعادة توازن مجموعة البيانات حيث تم الموازنة باستخدام طريقة SMOTE، بحيث نضاعف عدد أصناف الأقلية 10 مرات لكل منها، بحيث يصبح توزيع العينات بالشكل التالي:



3. تكمن المشكلة الثالثة في Dimension Complexity أي عدد الأعمدة الكبير المشارك في اتخاذ القرار النهائي ولهذا السبب نقترح تطبيق عملية اختيار سمات وفق عدة اعتبارات لاختيار السمات الأكثر ترابطاً مع عمود الهدف (حالة المريض وفاة أو بقاء على قيد الحياة)، ولهذا السبب سنقوم ببناء عدة سيناريوهات تدريب واختبار لمعرفة الأفضل منها.

4. هناك مشكلة أيضاً في عمود الهدف حيث أنه لدينا في مجموعة البيانات 3 أعمدة تتضمن حالة الوفاة للمريض خلال 28 يوم وخلال 3 أشهر وخلال 6 أشهر، ولهذا السبب قمنا بدمج هذه الأعمدة ضمن عمود هدف واحد وبحيث رمزنا كل حالة برقم خاص فيها

الرقم	الحالة
0	البقاء على قيد الحياة
1	الوفاة خلال 28 يوم
2	الوفاة خلال 3 أشهر
3	الوفاة خلال 6 أشهر

الجدول (I-1) العمود الهدف

5. تكمن المساهمة التالية في البحث في اختيار النموذج الأفضل من نماذج تعلم الآلة والتعلم العميق بحيث نحصل على أفضل أداء ممكن وفي البحث الحالي وبعد تطبيق كل عمليات المعالجة لحل المشاكل السابقة من 1 إلى 4، اقترح البحث استخدام مبدأ الأمثلة أو التحسين Optimization أي أننا لن نطبق نظام التعلم العميق أو تعلم الآلة مباشرة وإنما سنقوم بتغليفه ضمن خوارزمية أمثلة تجعل عملية تعلم المصنف تتكرر عدداً من المرات بحيث يتم في كل مرة اختيار بارامترات تدريب مختلفة وتدريب المصنف حتى الوصول لأفضل بارامترات تحقق أفضل نتيجة وأقل خطأ وبهذه الحالة نكون وصلنا لأفضل نموذج تعلم آلة أو تعلم عميق. في نماذج تعلم الآلة سنقوم باستخدام خوارزمية الأمثلة Grid Search Optimization، وسنقوم بتدريب المصنفات مرة باستخدام عملية التحسين أو الأمثلة ومرة بدونها لنقارن الأداء، أما في خوارزميات التعلم العميق، سنقوم بتطبيق خوارزمية Keras Tuner التي تعتمد البحث العشوائي Random Search للحصول على أفضل أداء من نموذج التعلم العميق المقترح.

6. المساهمة الأساسية هي في تصميم خوارزمية Balancing أو موازنة لمجموعة البيانات جديدة من قبلنا لا تعتمد على زيادة عينات الأصناف غير المسيطرة فقط أو إنقاص عينات الصنف المسيطر فقط وإنما على دمج المبدئين معاً باستخدام طريقة تراعي خصوصية مجموعة البيانات من خلال الاعتماد على السمات الأكثر ارتباطاً مع مجموعة البيانات وبحيث نزيد أو ننقص الأسطر المتعلقة بهذه الأعمدة. نقترح الخوارزمية أخذ الأعمدة الأكثر ارتباطاً مع عمود الهدف ثم توليف نقاط جديدة من الأسطر التي تحقق معياراً أو قيماً

Mixed Up Down Sampling based الخوارزمية، تم تسمية الأعمدة، هذه الأعمدة، تم تسمية الخوارزمية Mixed Up Down Sampling based Feature Correlation (MUDSFC).

7. أما المساهمة الأساسية الأخرى في البحث فتكمن في بناء نموذج تعلم عميق مؤلف من تسع طبقات منها طبقة دخل وطبقة خرج. أما طبقة الدخل فتم تصميمها باستخدام طبقة Dense Layer دخلها هو عبارة عن شعاع سمات الدخل، وعدد عصبوناتها تكون بحجم شعاع سمات الدخل تماماً أما تابع التفعيل فقد تم استخدام تابع Relu وهو تابع لا خطي ومناسب جداً للاستخدام في حالتنا كونه أسرع في عمليات الحساب من تابع Sigmoid وسيوفر لنا وقت في عملية التدريب. أما الطبقات الستة التالية فهي تمثل الطبقات المخفية وتم تصميمها بحيث تكون Dense Layers كل منها يتضمن 50 عصبون وتابع تفعيل Relu أما الطبقة الأخيرة فهي طبقة الخرج وتم تصميمها بحيث تكون Dense Layer مؤلفة من أربع عصبونات وتمثل المخارج الخاصة بنظامنا وهي درجة خطورة الإصابة بالفشل القلبي وهي أربع حالات (حالة البقاء على قيد الحياة، حالة الوفاة خلال 28 يوم، والوفاة خلال ثلاثة أشهر والوفاة خلال ستة أشهر). أما تابع التفعيل فقد تم استخدام تابع Softmax وليس تابع Relu لأننا نريد الحصول على احتماليات لكل صنف من الأصناف الأربعة وتابع Softmax يقوم بهذا الأمر حيث يتم اختيار الصنف ذو الاحتمالية الأعلى ليكون نتيجة التخمين النهائية (القرار النهائي).

6.1. مخطط البحث

يتم تقسيم الأطروحة إلى ستة فصول. نعرض في الفصل الأول مشكلة البحث والحلول المقترحة، أما في الفصل الثاني فيتم تقديم الدراسة النظرية التي تتضمن ملخصاً عن الأفكار النظرية الأساسية في البحث في حين يعرض الفصل الثالث أهم الدراسات المرجعية ذات الصلة. يتضمن الفصل الرابع الدراسة العملية والنماذج والخوارزميات المقترحة أما الفصل الخامس فيتضمن النتائج العملية التي تم التوصل لها ومناقشة تفصيلية لهذه النتائج وسيناريوهات التدريب والاختبار، في حين أن الخاتمة والتوصيات المستقبلية موجودة في الفصل السادس.

2. الفصل الثاني :الدراسة النظرية

نستعرض في هذا الفصل الإطار النظري للبحث.

1.2.1. مقدمة

نستعرض في هذا الفصل مفهوم أنظمة دعم القرار الطبي في مجال أمراض القلب وتشخيص الإصابة بالفشل القلبي، كما سنستعرض أهم خوارزميات تعلم الآلة والتعلم العميق المستخدمة في أنظمة دعم القرار الطبي وكيفية استخدامها في تشخيص الفشل القلبي.

2.2. ما هو الفشل القلبي

يعني قصور القلب أو فشل القلب أن القلب غير قادر على ضخ الدم في جميع أنحاء الجسم بشكل صحيح. يحدث هذا عادةً لأن القلب أصبح ضعيفًا جدًا أو متصلبًا.

يحدث فشل القلب - المعروف أحيانًا باسم قصور القلب الاحتقاني - عندما لا تضخ عضلة القلب الدم كما ينبغي. عندما يحدث هذا الأمر، غالبًا ما يتراجع الدم ويمكن أن تتراكم السوائل في الرئتين، مما يتسبب في ضيق التنفس [3].

تؤدي بعض أمراض القلب، مثل تضيق الشرايين في القلب (مرض الشريان التاجي) أو ارتفاع ضغط الدم، إلى ضعف القلب أو تيبسه تدريجيًا بحيث يتعذر معه ملء وضخ الدم بشكل صحيح [3].

يمكن أن يحسن العلاج المناسب علامات وأعراض قصور القلب وقد يساعد بعض الأشخاص على العيش لفترة أطول. يمكن أن تؤدي التغييرات في نمط الحياة - مثل فقدان الوزن وممارسة الرياضة وتقليل الملح (الصوديوم) في النظام الغذائي والتحكم في الإجهاد - إلى تحسين حالة المريض. ومع ذلك، يمكن أن يكون قصور القلب مهددًا للحياة. قد يعاني الأشخاص المصابون بفشل القلب من أعراض شديدة، وقد يحتاج البعض إلى زراعة قلب أو جهاز مساعدة بطيني [4].

تتمثل إحدى طرق الوقاية من قصور القلب في الوقاية من الحالات التي يمكن أن تسببه والسيطرة عليها، مثل مرض الشريان التاجي وارتفاع ضغط الدم والسكري والسمنة.

إنّ التشخيص الدقيق والسليم لمدى خطورة الفشل القلبي لدى المرضى هو أمر ضروري لتقليل المخاطر المرتبطة بمشاكل القلب الحادة وتحسين مستوى سلامة القلب [3]. وفقاً للجمعية الأوروبية لمرض القلب، فقد تم تشخيص 26 مليون بالغ في جميع أنحاء العالم بأمراض القلب و3.6 مليون شخص يتم تشخيصهم كل عام. ما يقرب من 50% من مرضى القلب الذين يعانون من مرض القلب يموتون في غضون 1-2 سنة، وتبلغ تكاليف إدارة أمراض القلب حوالي 3% من الميزانية المالية للرعاية الصحية [5].

3.2. أنظمة دعم القرار الطبي

أنظمة دعم القرار السريري CDSS هي أنظمة كمبيوتر "تزود الأطباء والموظفين والمرضى أو غيرهم من الأفراد بالمعرفة والمعلومات الخاصة بالشخص، ويتم تصنيفها بذكاء أو تقديمها في الأوقات المناسبة، لتعزيز الصحة والرعاية الصحية" [8]. تم تصميم CDSS لمجموعة متنوعة من الأغراض مثل التشخيص، وتنبؤ الاستجابة للعلاج، والتوصية بالعلاج (التخصيص)، والتشخيص، وتحديد أولويات رعاية المرضى وفقاً لمستوى مخاطرتهم. يمكن أن تكون هذه الأنظمة مفيدة في الممارسات السريرية باعتبارها "مجموعة ثانية من العيون" للأطباء، وتجمع بين معارفهم البشرية و "المعرفة" المضمنة في النظام. يمكن أن تساعد CDSS في تحسين سلامة المرضى وجودة الرعاية وكفاءة الرعاية الصحية بالإضافة إلى تقليل تكاليف الرعاية الصحية [9]. يمكنهم تحسين سلامة المرضى ليس فقط عن طريق تقليل الأخطاء الطبية ولكن أيضاً من خلال التذكير بالأدوية أو الأحداث الطبية الأخرى للمرضى أو الأطباء [10]. بالإضافة إلى ذلك يمكن أن تكون أنظمة CDSS مفيدة في الأماكن منخفضة الموارد حيث يكون عدد المؤسسات الطبية والمعدات والأطباء المؤهلين محدوداً.

يتمتع التعلم الآلي والذكاء الاصطناعي على نطاق أوسع بإمكانيات فورية ومستقبلية كبيرة لتحويل جميع جوانب الطب تقريباً، ومع ذلك في العديد من التطبيقات حتى خارج الطب أصبح الافتقار إلى الشفافية في تطبيقات الذكاء الاصطناعي مشكلة متزايدة. يحدث ذلك بشكل خاص حيث يحتاج المستخدمون إلى تفسير مخرجات أنظمة الذكاء الاصطناعي. يوفر الذكاء الاصطناعي القابل للتفسير الأساس المنطقي الذي يسمح للمستخدمين بفهم سبب إنتاج النظام لمخرجات معينة مثلاً أن ينتج نظام تشخيص الفشل القلبي خرجاً مساوياً للقيمة (0) يعني أن المريض بحالة طبيعية أما القيمة (1) فتعني مثلاً أن المريض يعاني من قصور القلب.

يعتبر مجال دعم القرارات الطبية (السريرية) أحد المجالات التي هي في حاجة ماسة إلى الذكاء الاصطناعي القابل للتفسير. تدعم هذه الأنظمة الأطباء في اتخاذ قراراتهم في العيادة وفي حالة عدم وجود قابلية للتفسير قد يؤدي إلى مشاكل الاعتماد المفرط أو النقص. سيسمح تقديم التفسيرات لكيفية التوصل إلى التوصيات للممارسين باتخاذ قرارات أكثر دقة، وفي بعض الحالات تكون هذه التوصيات منقذة للحياة [11].

1.3.2. نظام دعم القرار الطبي لمرض الفشل القلبي

أنظمة دعم القرار السريري (CDSS) هي أنظمة معلومات قائمة على الكمبيوتر مصممة لمساعدة مقدمي الرعاية الصحية على تنفيذ الإرشادات السريرية في نقطة الرعاية. تستخدم CDSS بيانات المريض لإنشاء تقييمات مخصصة للمرضى وتوصيات علاج قائمة على الأدلة لمقدمي الرعاية الصحية للنظر فيها. يتم إدخال معلومات المريض يدوياً أو تلقائياً من خلال نظام السجل الصحي الإلكتروني (Electronic Health Record EHR). تشمل أنظمة CDSS للوقاية من أمراض القلب والأوعية الدموية (CVD) واحداً أو أكثر من البيانات الآتية:

1. تذكير بخدمات الوقاية من الأمراض القلبية الوعائية المتأخرة (مثلاً فحص عوامل الخطر مثل ارتفاع ضغط الدم والسكري وارتفاع الكوليسترول).
2. تقييمات مخاطر إصابة المرضى بالأمراض القلبية الوعائية بناءً على تاريخهم الطبي وأعراضهم ونتائج الاختبارات السريرية.
3. توصيات للعلاجات القائمة على الأدلة للوقاية من الأمراض القلبية الوعائية، بما في ذلك تكثيف العلاج.
4. توصيات لتغيير السلوك الصحي لمناقشتها مع المرضى (مثلاً الإقلاع عن التدخين، وزيادة النشاط البدني، وتقليل تناول الملح المفرط).
5. تنبيهات عندما لا تكون مؤشرات عوامل خطر الأمراض القلبية الوعائية في الهدف.

غالباً ما يتم دمج نظم دعم القرار في أنظمة السجلات الصحية الإلكترونية ويتم دمجها مع وظائف الكمبيوتر الأخرى التي تقدم تقارير موجزة عن رعاية المرضى، وردود الفعل على مؤشرات الجودة، والمعايير. يمكن أيضاً ربط أنظمة إدارة المعرفة التي توفر الوصول إلى المقالات العلمية والخوارزميات المصممة للوقاية من الأمراض القلبية.

أظهرت الأبحاث الحديثة أن أنظمة دعم القرار السريري المخصصة يمكنها تحسين التوصيات الموجهة للوصفات الطبية لعلاج مرضى القلب إضافةً إلى المشاركة الإلكترونية للمرضى والتي يمكنها أيضاً تحسين وصف الأدوية بشكل إيجابي [12].

في العديد من أنظمة دعم القرار الطبي لتخمين الفشل القلبي تعد البيانات التي تم الحصول عليها أثناء الممارسة السريرية الروتينية المصدر الرئيسي لبيانات الإدخال. يتم إنشاء البيانات السريرية من قبل مقدمي الرعاية في السجلات الصحية الإلكترونية EHR. تعتمد جودة نظم دعم القرار الطبي على البيانات "الخام" والتي تحتاج العديد من المعالجات قبل بناء نظام دعم

القرار الطبي فهناك معلومات شخصية ومعلومات تتعلق بتاريخ دخول المشفى ومعلومات مفقودة وغيرها العديد من البيانات التي تحتاج لمعالجة [13].

2.3.2. تقنيات تعلم الآلة المستخدمة في أنظمة كشف الفشل القلبي

شهد العقد الماضي التطور السريع للتعلم الآلي وتقنيات التعلم العميق في إدارة الحالات الطبية المتنوعة مثل مرض السكري وأمراض الكبد وأمراض القلب [14]. تستخدم تقنيات التعلم الآلي البيانات الطبية الموجودة (بما في ذلك الفحوص السريرية أو نتائج الاختبار أو التصوير) لاكتساب نظرة متكاملة لإدارة حالات مثل قصور القلب. توفر البيانات الطبية المتعلقة بفشل القلب وتوافر أجهزة وبرامج حاسوبية قوية للغاية للباحثين فرصة مثالية لتطبيق خوارزميات التعلم الآلي الأكثر تعقيدًا والمكثفة من الناحية الحسابية لتحليل هذا المرض بشكل نقدي.

التعلم الآلي هو أحد مجالات علوم الكمبيوتر التي تستخدم الذكاء الاصطناعي لتعلم العلاقات أو الأنماط من البيانات دون الحاجة إلى تعريفها مسبقًا [14].

التعلم العميق هو نوع خاص من التعلم الآلي، والذي يؤدي بشكل أفضل من مناهج التعلم الآلي التقليدية بحكم قدرته على تعلم تسلسل من المعلومات ذات المغزى من بيانات الإدخال باستخدام الشبكات العصبية مع عدد كبير من الطبقات المخفية.

هناك العديد من الأبحاث التي توضح فائدة طرق التعلم الآلي في تحليل الصور الطبية للكشف عن الهياكل التشريحية والتجزئة والكشف بمساعدة الكمبيوتر والتشخيص بمساعدة الكمبيوتر.

فمثلا في دراسة [15] استخدمت طرق التعلم الآلي للتشخيص وتقدير الشدة والتنبؤ بالأحداث العكسية في مرضى فشل القلب بناءً على بيانات طبية امتدت من 2005 إلى 2016. وعلى الرغم من أن هذه الدراسة توفر أساسًا جيدًا إلا أن التركيز اقتصر على خوارزميات التعلم الآلي التقليدية المرتبطة بفشل القلب وليس أساليب التعلم العميق المعاصرة وبالتالي الحد من تأثيرها.

لحسن الحظ فإن خوارزميات علوم الحاسب بما فيها خوارزميات تعلم الآلة (ML) وخوارزميات التعلم العميق (DL) قادرة على معالجة البيانات الضخمة التي تأتي من الملفات الطبية الخاصة بمرضى القلب، كما أنها قادرة على اتخاذ القرار الطبي حول درجة خطورة الفشل القلبي، وهذا الأمر كان من غير الممكن قبل ظهور هذه التقنيات.

هناك العديد من خوارزميات تعلم الآلة التي يمكن استخدامها في تصميم أنظمة تشخيص خطورة الفشل القلبي ومنها خوارزميات أشجار القرار (Decision Trees (DT)، غابات القرار العشوائية (Random Forests (RF)، الانحدار المنطقي (Logistic Regression (LR)، أشعة متجهات الدعم (Support Vector Machines (SVM)، مصنف الجار الأقرب K-Nearest Neighbour (KNN)، مصنف بايس للقرار Naïve Bayes وغيرها الكثير من خوارزميات تعلم الآلة [3] [16] [17].

يوضح الشكل (1-2) الخطوات العامة لأي نظام دعم قرار طبي لكشف الفشل القلبي.



الشكل (1-2) الخطوات العامة لأي نظام دعم قرار طبي لكشف الفشل القلبي.

3.3.2. مجموعات البيانات الطبية لكشف الفشل القلبي

يتم الاحتفاظ بكمية مفصلة للغاية ودقيقة من البيانات المتعلقة بمرضى قصور القلب في البلدان ذات المرافق الطبية الحديثة. وتشمل هذه البيانات زيارات المستشفيات والإجراءات ونتائج

التحاليل المخبرية والتصوير وسجلات الصيدلة. يتم جمعها في الغالب من قبل المتخصصين في الرعاية الصحية في المستشفيات ومنتشآت الرعاية الصحية أو في بعض الأحيان بواسطة الأجهزة الرقمية.

يمكن أن تكون البيانات في شكل نصي (ملاحظات سريرية) أو صور (الموجات فوق الصوتية والتصوير المقطعي) أو إشارات تخطيط القلب. استخدام مجموعة متنوعة من البيانات المتعلقة بفشل القلب في الدراسات السابقة، ويرد وصف موجز لعدد قليل منها في الجدول (1-2). وتشمل مجموعات البيانات التركيبية السكانية للمريض والتاريخ الطبي وفحوصات المريض والنتائج المخبرية والتشخيص ورموز الإجراءات. وتجدر الإشارة إلى أن مجموعات بيانات PhysioNet و Holter تحتوي على بيانات في شكل إشارات ECG، بينما تحمل مجموعات البيانات الأخرى بيانات نصية.

عدد السمات	عدد الحالات	مجموعة البيانات
13	303	Cleveland
11	294	Hungarian
12	123	Switzerland
12	200	Long Beach
13	270	Heart

الجدول (1-2) أهم مجموعات البيانات الطبية لمرض الفشل القلبي.

4.2. الخلاصة

قدّمنا في هذا الفصل دراسة نظرية حول أنظمة دعم القرار الطبي وخصوصاً المتعلقة بأنظمة تشخيص وتخمين الفشل القلبي من خلال تحليل بيانات مجموعات البيانات الطبية العالمية (مجموعات البيانات الطبية).

تم في الفصل عرض أهم طرق الذكاء الصناعي مثل تعلم الآلة والتعلم العميق المستخدمة في نظم دعم القرار الطبي لتشخيص الفشل القلبي، كما تم عرض معلومات عن أهم مجموعات البيانات الطبية المستخدمة في هذا المجال.

3. الفصل الثالث: الدراسة المرجعية

نستعرض في هذا الفصل دراسة مرجعية حول أهم الأبحاث في مجال أنظمة دعم القرار الطبي لمرض القلب

1.3.1 مقدمة

نستعرض في هذا الفصل دراسة مرجعية حول أهم الأبحاث في مجال أنظمة دعم القرار الطبي لتشخيص أمراض القلب بشكل عام وحالة الفشل القلبي بشكل خاص.

2.3. الدراسات المرجعية في مجال تصميم أنظمة دعم القرار الطبي

لتشخيص مرض الفشل القلبي

إن الدراسات التي قدمت في مجال تشخيص أمراض القلب أو تشخيص الفشل القلبي أو تخمين الوفاة بسبب أمراض القلب كثيرة جداً وسنعرض في الآتي أهم هذه الدراسات في السنوات الأخيرة.

بعض الدراسات استخدمت تقنيات تعلم الآلة الشهيرة مثل Random Forests [24] [25] و [3] Logistic Regression و XGBoost و [26] [16] Decision Trees و [26] [14] و [3] Naive Bayes و [26] [16] [27] [3] Support Vector Machines (SVM) و [26] [16] [27] [3] وغيرها الكثير من تقنيات تعلم الآلة. ومنها من اعتمدت تقنيات التعلم العميق [28] [29]. في العام 2018، صمم Haq وآخرون [3] نظاماً لكشف أمراض القلب باستخدام تقنيات تعلم الآلة. استخدم الباحثون ست تقنيات تعلم آلة معروفة هي:

(K-NN, SVM, ANN, NB, DT, LR)

اعتمدوا في بناء نظامهم على مجموعة بيانات Cleveland Dataset هي مجموعة بيانات Cleveland heart disease dataset والمؤلفة من 303 سجل (303 مريض) و 76 خاصية أو سمة. استخدم البحث أيضاً خوارزمية Relief المعروفة لاختيار السمات من أجل تقليل عدد السمات الـ 76. توصل البحث إلى أن أفضل دقة كشف كانت 89% والتي تم الحصول عليها باستخدام مصنف LR مع خوارزمية اختيار السمات.

قدم الباحث Ali وآخرون [27] نظام دعم قرار طبي مقاد بالسّمات لكشف الفشل القلبي وذلك بالاعتماد على نموذج إحصائي وعلى مصنف بايس الغوصي Gaussian NB. تم الاعتماد في عملية التخمين على 13 سمة مرضية، وفي مرحلة التصنيف تم الاستعانة بمصنف NB من أجل اتخاذ القرار. تم تدريب المصنف على مجموعة بيانات مرضية مؤلفة من 297 شخص وتم التوصل لدقة تخمين مقدارها 93.33%، وتكمن المشكلة الأساسية في البحث في صغر حجم مجموعة البيانات المستخدمة.

لاحقاً، قام الباحث Shah وآخرون [16] ببناء نظام لتخمين أمراض القلب باستخدام تقنيات تعلم الآلة.

تم اعتماد مجموعة بيانات مرضية مؤلفة من 303 سجل (معلومات من 303 مريض) و76 سمة أو خاصية. تم استخدام 14 سمة فقط من بين السمات الـ 76. أشارت النتائج إلى أنّ مصنف K-NN حقق الأداء الأفضل بين المصنّفات الأربعة. وكما هي حالة الدراستين السابقتين، فإن هذه الدراسة أيضاً اعتمدت مجموعة بيانات صغيرة الحجم.

في دراسة أخرى للباحث Rani وآخرون [26]، تم تصميم نظام دعم قرار طبي هجين لتخمين أمراض القلب بالاعتماد على البيانات السريرية المأخوذة من مرضى القلب. تم استخدام طريقة اختيار سمات تعتمد على الخوارزميات الجينية Genetic Algorithms (GA) وحذف السمات المتكررة Recursive Feature Elimination (RFE).

تم في مرحلة التصنيف استخدام مصنّفات (adabost,SVM,NB,LR,RF)، حيث حقق مصنف RF الدقة الأفضل في عملية تخمين الإصابة بأمراض القلب حيث بلغت 86.6% على مجموعة بيانات Cleveland heart disease dataset.

حاول الباحثون في البحث [30] الحصول على أفضل دقة ممكنة من مصنّفات تعلم الآلة لكشف خطر الإصابة بمرض القلب. درب الباحثون مصنّفات (K-NN,DT,RF) على

مجموعة بيانات مأخوذة من موقع Kaggle مؤلفة من 1025 سجل و 14 سمة (عمود). حقق مصنف RF دقة 100% وفقاً لما أورده الباحثون في حين حققت المصنفات الأخرى دقة تتراوح بين 89.62% و 97.95%.

اقترحت عملية تقييم السمات Feature Evaluation في البحث المقدم من قبل الباحث Reddy وآخرين [31]، حيث تم استخدام تقنيات تعلم الآلة على نسختين من مجموعة بيانات Cleveland heart disease dataset تتضمن النسخة الأولى كل السمات 76 في حين تتضمن النسخة الثانية السمات المنتقاة باستخدام خوارزمية Relief .

حققت خوارزمية sequential minimal optimization (SMO) الدقة الأفضل في التخمين حيث بلغت 85.14% تليها خوارزمية LR بدقة وصلت لـ 84.18%.

في العام 2021، قدم الباحث Grag وآخرون [32] نظاماً لكشف خطر الإصابة بأمراض القلب بالاعتماد على تعلم الآلة، حيث استخدم الباحثون خوارزميتي (K-NN,RF) حيث تم تصميم النظام بالاعتماد على مجموعة بيانات Heart disease UCI وهي مؤلفة من 13 سمة و 303 سجل فقط.

حقق مصنف RF دقة تخمين مقدارها 81.96% في حين كانت دقة مصنف K-NN هي 86.88%. كانت مجموعة البيانات المستخدمة صغيرة جداً كما لم يطبق الباحثون أي عمليات اختيار سمات أو تحسين.

مجدداً تم استخدام مجموعة بيانات Cleveland heart disease dataset من قبل الباحث Kavitha وآخرين [33]. تم اعتماد نظام هجين مؤلف من مصنفي (RF,DT) للحصول على أفضل أداء، حيث تم التوصل لدقة تخمين 88.7% من النموذج الهجين.

في دراسة أجريت في العام 2021 من قبل الباحث Bharti وآخرين [28]، تم استخدام تقنيات تعلم الآلة والتعلم العميق من أجل بناء نظام تخمين أمراض القلب. تم تدريب الأنظمة على مجموعة بيانات Heart Disease UCI المؤلفة من 13 سمة و303 سجل. تم عزل بعض السمات من مجموعة البيانات باستخدام تقنية Isolation Forests (IS) أما نموذج التعلم العميق المستخدم فكان طبقة Dense وطبقة Dropout وطبقة Flatten وبنتيجة الاختبارات تم التوصل لدقة تخمين 94.2%.

لتشخيص الإصابة بالفشل القلبي قبل حدوثه، صمم Plati وآخرون [34] نظاماً يعتمد تقنيات تعلم الآلة مدرباً على مجموعة بيانات مؤلفة من 422 حالة طبية (سجل). تم استخدام مصنفات DT و NB و RF و SVM و K-NN و Logistic Model Tree (LMT) و Rotation Forest (ROT) حيث تم التوصل لدقة 91.23% باستخدام مصنف ROT باستخدام كل سمات مجموعة البيانات.

مؤخراً في العام 2022، قدم الباحث Khiabani وآخرون [35] نظام دعم قرار طبي لتشخيص الفشل القلبي باستخدام الشبكات العصبونية والعمليات الإحصائية. تم تجميع مجموعة البيانات الطبية للفشل القلبي من 175 سجل طبي. تم استخدام شبكة Perceptron كمصنف لتخمين الإصابة بالفشل القلبي. تم استخدام 20% من مجموعة البيانات للاختبار و20% للتحقق. حقق النظام المقترح دقة 81.25%.

في العام 2022 أيضاً استخدم Nagavelli وآخرون [29] تقنيات التعلم الآلي المختلفة بناءً على تحليل موجز للكشف عن أمراض القلب. أولاً، تم استخدام Naïve Bayes مع طريقة الترجيح للتنبؤ بأمراض القلب.

بعد ذلك ووفقاً لخصائص المجال الترددي والمجال الزمني ونظرية المعلومات تم استخدام نموذج تعلم آلة تلقائي لتحليل الأعراض واكتشاف أمراض القلب الإقفارية. تم استخدام مصنفين

أساسيين في الدراسة هما مصنف SVM ومصنف XGBoost مع أفضل أداء للتصنيف في الحالتين. أما النموذج الثالث فيمثل التعرف التلقائي على قصور القلب باستخدام خوارزمية SVM المحسنة.

أخيراً، بالنسبة لنظام دعم القرار السريري، تم استخدام نموذج فعال للتنبؤ بأمراض القلب حيث تم الاستفادة من خوارزمية العنقدة DBScan لتحليل ترابط العينات. تم تطبيق الاختبارات العملية على مجموعة بيانات طبية مؤلفة من 227 حالة مرضية و347 حالة طبيعية.

أثبتت النتائج تفوق مصنف XGBoost على مصنف SVM في كشف الفشل القلبي حيث حقق مصنف XGBoost دقة تصنيف مقدارها 95.9% في حين كانت قيم Precision و Recall هي 97.1% و 94.67% على التوالي.

درس الباحث Nishat وآخرون [36] ست خوارزميات تعلم آلي خاضعة للإشراف وتطبيقها لتحليل مجموعة بيانات مؤلفة من 299 حالة طبية من مستودع UCI للتعلم الآلي. حيث تم التنبؤ ببقائهم على قيد الحياة بعد حدوث حالة قصور القلب. تم اتباع ثلاث طرق متميزة باستخدام خوارزميات Decision Tree Classifier، و Logistic Regression، و Gaussian Naïve Bayes، و Random Forest Classifier، و K-Nearest Neighbors، و SVM. تم إجراء تحجيم البيانات كخطوة معالجة مسبقة باستخدام طريقة Max-Min. كما تم استخدام تقنيات التحسين Grid Search و Random Search لتحسين أداء الخوارزميات.

بالإضافة إلى ذلك تم استخدام تقنية زيادة عدد عينات الأقلية وتقنية تقليل عينات البيانات المجاورة (SMOTE-ENN المعدلة، وتمت مقارنة أداء جميع الأساليب المستخدمة لتحديد الأفضل. تشير النتائج التجريبية بوضوح إلى أن Random Forest Classifier (RF)

يتفوق على جميع النماذج الأخرى بدقة اختبار تبلغ 90% عند استخدامه مع SMOTE-ENN وتقنية القياس القياسية.

يمكن القول أنّ كل الدراسات السابقة استخدمت مجموعات بيانات صغيرة الحجم باستثناء الدراسة [9] التي تضمنت 1025 سجل طبي، لكنها لم تعالج سوى 14 سمة فقط. وكذلك في معظم الأبحاث كان عدد السمات قليل حيث كان أعلاها في مجموعة بيانات Cleveland heart disease التي تتضمن 76 سمة.

بعض الدراسات السابقة لم تطبق أي عملية معالجة لمجموعات البيانات وبعضها الآخر لم تطبق عمليات اختيار سمات بل استخدمت مجموعة البيانات كاملةً مما أدى لانخفاض دقة التشخيص أو التخمين.

في البحث الحالي نستخدم مجموعة بيانات طبية لمرضى فشل القلب مؤلفة من 2008 سجل و167 سمة (عمود) وهي أضخم من أي مجموعة بيانات مستخدمة سابقاً في تشخيص الفشل القلبي، كما أن الدراسة الحالية هي أول دراسة تستخدم مجموعة البيانات هذه وتصمم نظام تعلم آلة ونظام تعلم عميق بالاعتماد عليها.

يقدم البحث الحالي أيضاً طريقة جديدة لموازنة مجموعة البيانات Mixed Up Down Sampling based Feature Correlation (MUDSFC) من أجل موازنة عينات الأصناف المسيطرة والأصناف الثانوية. يقدم البحث أيضاً عملية تخمين خطورة الإصابة بالفشل القلبي ضمن فترة زمنية تمتد بين 28 يوم و6 أشهر.

يقدم البحث أيضاً تحسيناً للخوارزمية MUDSFC من خلال إضافة مرحلة عنقدة باستخدام خوارزمية DBSCAN من أجل تجميع عينات البيانات ودراسة الترابط بين الأصناف بحيث يتم إهمال عينات الصنف المسيطر Majority Class Points المتداخلة مع عينات الأصناف

غير المسيطرة (الثانوية) Minority Class Points والتي تشكل مشكلة في عملية التصنيف ثم يتم مضاعفة عينات الأصناف غير المسيطرة (الثانوية) باستخدام خوارزمية MUDSFC. كما أنّ البحث يستخدم تقنيات تعلم الآلة والتعلم العميق ويطبق خوارزمية تحسين على كل منها للحصول على أفضل أداء ممكن.

يقترح البحث أيضاً بناء النظام التعلم العميق بحالتين الأولى باعتماد الأعمدة التي يمكن استخدامها من مجموعة البيانات والثانية بتطبيق خوارزمية اختيار سمات تعتمد على مبدأ الترابط Correlation based feature selection.

يستخدم البحث معاملات قياس الأداء لمقارنة كل سيناريوهات التدريب المختلفة.

3.3. الخلاصة

قدّمنا في هذا الفصل دراسة مرجعية موسّعة عن أهم الدراسات السابقة والخوارزميات المستخدمة في مجال تصميم أنظمة دعم قرار طبي لكشف الإصابة بالفشل القلبي.

4. الفصل الرابع: الدراسة التجريبية والنماذج المقترحة

نستعرض في هذا الفصل الدراسة التجريبية التي قمنا بها في البحث

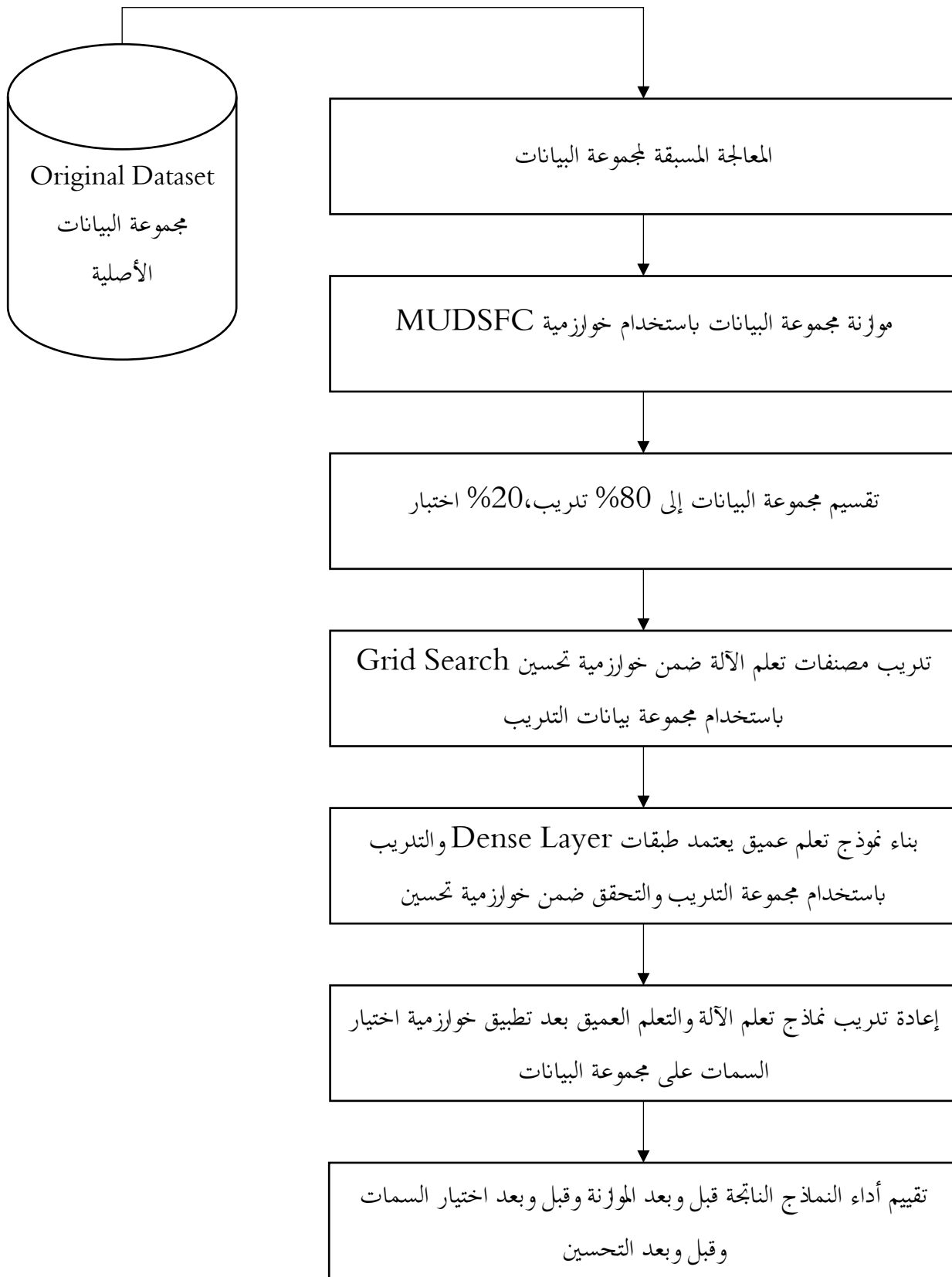
1.4.1. مقدمة

نستعرض في هذا الفصل الدراسة التجريبية التي قمنا بها في البحث الحالي إضافة للخوارزميات المقترحة والخوارزميات الجديدة والنماذج المقترحة من قبلنا لكل مرحلة من مراحل البحث.

2.4. المخطط المقترح للبحث

يوضح الشكل (1-4) المخطط الصندوقي للنظام المقترح. حيث سيتم بناء النظام المقترح وفق عدة سيناريوهات.

1. معالجة مجموعة البيانات المستخدمة كونها تتضمن عدة مشاكل.
2. تطبيق خوارزمية موازنة مصممة من قبلنا عليها من أجل تحسين توزيع العينات.
3. تقسيم مجموعة البيانات إلى مجموعة تدريب 80%، ومجموعة اختبار 20%.
4. تدريب عدة أنظمة تعلم عميق باستخدام مجموعة التدريب بدون تحسين ثم مع استخدام التحسين المطبق باستخدام خوارزمية Grid Search.
5. تصميم نظام تعلم عميق مؤلف من طبقات Dense Layer فقط ويتم تدريب النموذج بدون ومع خوارزمية تحسين Random Search.
6. يتم مقارنة أداء كل تلك السيناريوهات باستخدام مقاييس تقييم الأداء.



الشكل (4-1) مخطط النظام المقترح

3.4. مجموعة البيانات المستخدمة في البحث

تم في البحث استخدام مجموعة بيانات طبية لتشخيص الفشل القلبي تم استخدام مجموعة البيانات التالية المصممة من قبل مؤسسة PhysioNet. [23] [37].

Hospitalized patients with heart failure: integrating electronic healthcare records and external outcome data.

وهي متاحة بشكل مباشر [38] للباحثين للعمل عليها. ما يميز مجموعة البيانات المستخدمة أنها تتضمن تحدي خاص يتمثل في العدد الكبير للخصائص (الأعمدة) Predictors المستخدمة لإنجاز التشخيص النهائي وهذه المسألة تسمى مسألة تصنيف ضمن فضاء معقد الأبعاد Complexity Dimensional Classification Problem، حيث يبلغ عدد أعمدة مجموعة البيانات 167 عمود.

1.3.4. خطوات المعالجة المسبقة لمجموعة البيانات

نجد في الملحق (أ) وصف لمجموعة البيانات المستخدمة ومنه نستنتج ما يلي حول السمات التي تمثلها.

أولاً: من حيث الدلالة

تضم مجموعة البيانات معلومات متنوعة عن كل حالة مرضية، يمكن تصنيفها كما يلي:

1. معلومات شخصية: مثل العمر، الجنس، العمل.
2. معلومات إدارية: مثل رقم المريض، قسم المشفى الذي تعامل مع الحالة.
3. المؤشرات الحيوية للمريض: مثل ضغط الدم، التنفس.
4. الأمراض المصاحبة: مثل القرحة الهضمية، السكري.

5. نتائج الاستشفاء: مثل الوفاة خلال 28 يوم من دخول المستشفى، العودة إلى قسم الطوارئ خلال 3 أشهر من دخول المستشفى.

ثانياً: من حيث طبيعة البيانات

نجد أنواع البيانات التالية:

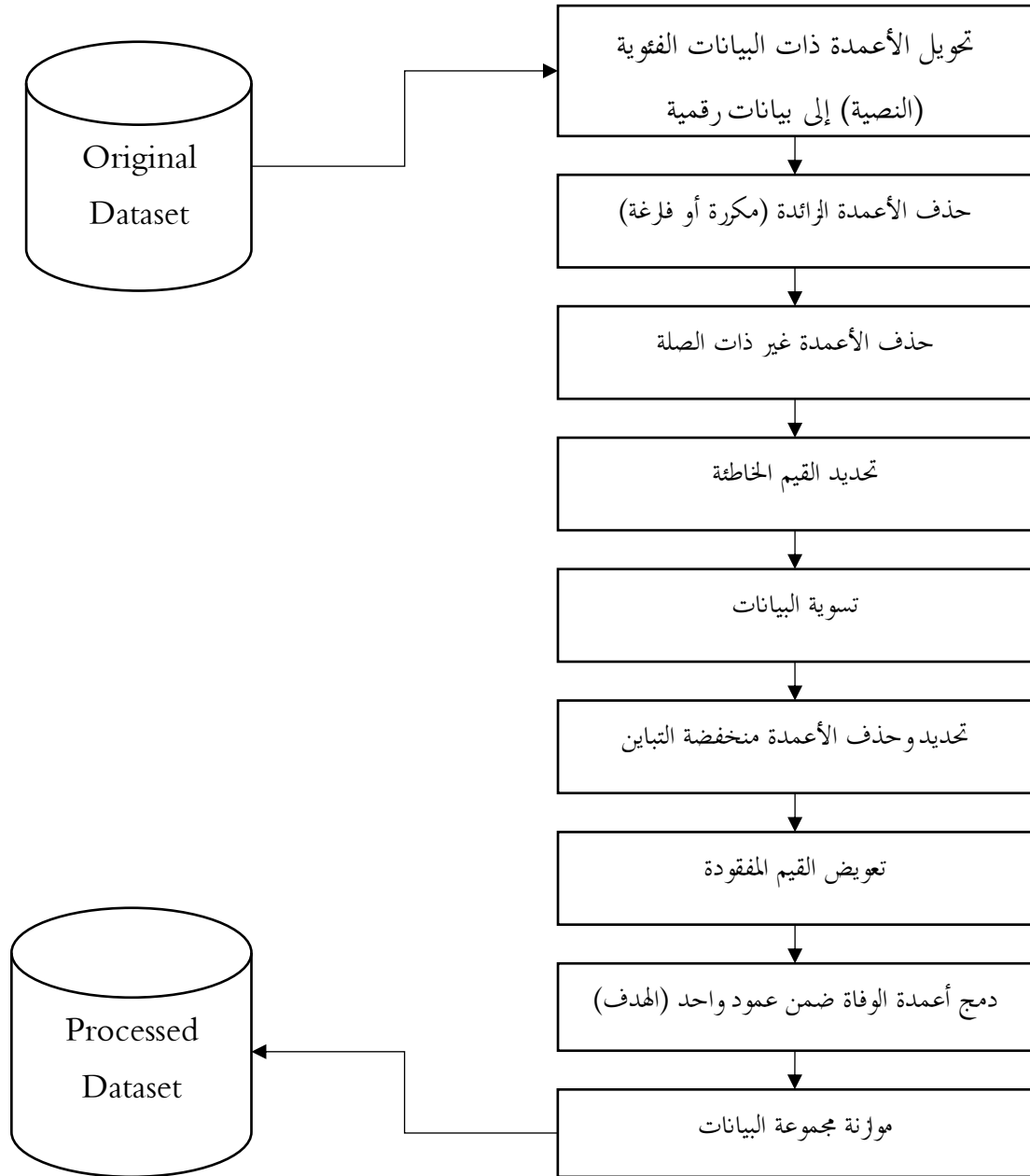
1. البيانات الفئوية: وهي نوعين:

- بيانات اسمية: مثل جنس المريض (Male, Female).
- بيانات ترتيبية: مثل استجابة حذقة العين، حيث تأخذ القيم التالية (1,2,3,4) بحيث يكون للترتيب معنى.

2. البيانات الرقمية: وهي نوعين:

- بيانات رقمية مستمرة: مثل وزن المريض.
- بيانات رقمية متقطعة: مثل عدد مرات دخول المريض للمشفى.
- بيانات رقمية ثنائية: مثل مرض السكري.

يبين الشكل (2-4) خطوات المعالجة المسبقة المقترحة من قبلنا والتي سيتم تطبيقها على مجموعة البيانات.



الشكل (2-4) خطوات المعالجة المسبقة لمجموعة البيانات

1. تحويل البيانات الفئوية إلى بيانات رقمية:

باستخدام طريقة Label Encoding، حيث من أجل كل سمة فئوية يتم إسناد عدد صحيح فريد لكل تسمية.

[0,1,2,3]	I, II, III, IV	NYHA.cardiac.function.classification
[0,1,2,3]	I, II, III, IV	Killip.grade
[0,1]	TypeII,NonTypeII	type.II.respiratory.failure
[0,1,2,3]	Nonresponsive , ResponsiveToPain , ResponsiveToSound , Clear	Consciousness
[0,1,2]	None, IMV, NIMV	respiratory.support.
[0,1]	OxygenTherapy , AmbientAir	oxygen.inhalation
[0,1,2,3,4,5,6,7]	[21,29], [29,39] [39,49], [49,59] [59,69], [69,79] [79,89], [89,110]	ageCat
[0,1]	Male, Female	Gender

الجدول (I-4) تحويل البيانات الفئوية إلى بيانات رقمية

2. حذف العمود cholinesterase كونه فارغ

3. حذف السمات التي ليست على صلة بموضوع البحث.

هي الأعمدة التي تعبر عن المعلومات الإدارية الخاصة بالمشفى، إضافة إلى الأعمدة التي تعبر عن نتائج الاستشفاء، إذ لا معنى للتنبؤ اعتماداً على رقم المريض، كما أن

نتائج الاستشفاء لن تكون متوفرة في حالة المريض الجديد الذي نرغب بتقييم خطورة حالته، تم في هذه المرحلة حذف من الأعمدة التالية:

Column1	inpatient.number	death.within.28.days
death.within.3.months	death.within.6.months	re.admission.within.28.days
re.admission.within.28.days	re.admission.within.6.months	dischargeDay
visit.times	time.of.death..days.from.admission.	return.to.emergency.department.within.6.months
time.of.death..days.from.admission	re.admission.time..days.from.admission.	

الجدول (2-4) السمات غير ذات الصلة

4. القيم الخاطئة

هي قيم موجودة كقيم عددية في مجموعة البيانات، أي أنها ليست معلمة كـ Nan أو '?'، أو فراغ، لكنها خاطئة بحسب المعرفة بدلالاتها ضمن طبيعة المسألة المطروحة، وفي حالة مجموعة البيانات التي بين أيدينا فقد تم ملاحظة وجود قيم خاطئة في الأعمدة التالية:

في حال كانت نتيجة الاستشفاء ليست وفاة، تم تحديد القيم الخاطئة التالية:

- القيم الصفرية لعمود Pulse الذي يمثل ضغط الدم.
- القيم الصفرية والسالبة لعمود respiration الذي يمثل عدد مرات التنفس في الدقيقة.
- القيم الصفرية والسالبة لعمودي systolic. blood.pressure و diastolic.blood.pressure اللذين يمثلان ضغط الدم الانقباضي وضغط الدم الانبساطي.
- القيم الصفرية والسالبة لعمود map الذي يمثل متوسط الضغط الشرياني.

- القيم الصفرية والسالبة لعمودي height و weight اللذين يمثلان وزن وطول المريض.

تم تحديد القيم الخاطئة في مجموعة البيانات، وتعليمها كقيم مفقودة (Nan)، قبل الانتقال إلى الخطوة التالية، تعويض القيم المفقودة.

5. تسوية البيانات

نظراً لأن السمات الممثلة في مجموعة البيانات متنوعة من حيث مقاييس كل سمة، على سبيل المثال، كمية البوتاسيوم في الدم تتراوح بين 3.5 إلى 4.5 ميلي مول/لتر، في حين تتراوح قيم الصوديوم بين 136 إلى 146 ميلي مول/لتر. وبالتالي نحن بحاجة عملية إزاحة (scaling) للسمات في مجموعة البيانات، بهدف توحيد المقاييس. عملية الإزاحة للسمات يمكن أن تتم بإحدى طريقتين:

- التسوية (Normalization): تعتبر جيدة، عندما لا يوجد افتراضات مسبقة عن التوزيع الاحتمالي الذي تتبعه البيانات.
- التقييس (Standardization): يمكن أن تكون مفيدة في الحالات التي تتبع فيها البيانات توزيعاً احتمالياً طبيعياً.

تم عمل الإزاحة للسمات عبر تطبيق Min-Max Scaler، ويعرّف بالشكل التالي:

$$y = (x - \min) / (\max - \min)$$

حيث من أجل كل سمة رقمية، مستمرة، تعتبر أصغر قيمة مبدأ القياس وتعطي القيمة 0، وتعطي القيمة الأكبر قيمة نهاية القياس وهي 1، على سبيل المثال أحد السمات تتراوح قيمها بين ال 20 وال 50، وبالتالي القيمة 35 سوف تصبح 0.5، وبالتالي نكون قد وحدنا مقاييس جميع القيم الرقمية لتصبح ضمن المجال [0,1].

6. حذف السمات منخفضة التباين

وهي السمات التي تكون فيها جميع القيم ثابتة، أو يكون تغير القيم منخفضاً، باعتبار هذه السمات لا تقدّم معلومات مفيدة لأي نموذج قد يتم تطبيقه.

لتحقيق ذلك قمنا بتطبيق variance threshold filter لتحديد السمات التي لا تحقق قيمها عتبة تغير معينة، تم بعد ذلك تطبيق variance threshold filter، واختيار عتبة تباين 0.1، بحيث نحدد السمات التي تحوي المعلومات الأقل، وتم تحديد وحذف السمات التالية:

pH	connective.tissue.disease	Hemiplegia
malignant.lymphoma	AIDS	acute.renal.failure
leukemia	Hematocrit	basophil.ratio
basophil.count	eosinophil.ratio	body.temperature.blood.gas
free.calcium	hydroxybutyrate.dehydrogenase.to.lactate.dehydrogenase	

الجدول (3-4) السمات منخفضة التباين

7. القيم المفقودة

إحدى المشاكل التقليدية التي تعاني منها مجموعات البيانات الطبية في العالم الحقيقي، هي القيم المفقودة، وهذا أمر طبيعي عند الأخذ بعين الاعتبار ظروف تسجيل هذه البيانات، والعدد الكبير من المعلومات المطلوب تسجيلها، سيما وأنّ كثير من هذه المعلومات قد تتطلب فحوصات موسّعة، أو إجراء تحاليل معينة، قد تتطلب وقتاً، مع الأخذ بعين الاعتبار أنه، وفيما يتعلق بمرض القلب تحديداً، فإن كثيراً من الحالات يتم استقبالها في قسم الطوارئ، بصورة حالة طارئة، الأمر الذي لا يتيح الكثير من الوقت للتوسع في إجراء كافة التحاليل والفحوصات، مثلاً في حالة مجموعة البيانات التي نعمل عليها، فإن 956 حالة من أصل 2008 قد تم استقبالها في قسم الطوارئ.

هناك العديد من المقاربات التي يمكن استخدامها عند الحديث عن تعويض القيم المفقودة، بدءاً بالمقاربات البسيطة، التي تقوم بتعويض القيمة المفقودة بأحد الواصفات الإحصائية، مثل التعويض بمتوسط قيم العمود، أو القيمة الأكثر تكراراً.

يمكننا مرة أخرى، استغلال معرفتنا بموضوع ومجال مجموعة البيانات، لتعويض القيم المفقودة، باستخدام خوارزمية الجار الأقرب KNN، والسبب وراء هذا الاختيار هو أن الخبرة الطبية، تخبرنا أن المرضى الذين يعانون من نفس المرض يميلون لإظهار أعراض متشابهة، ومؤشرات حيوية متقاربة، أضف إلى ذلك ارتباط عدد من هذه المؤشرات ببعضها البعض بحكم الطبيعة الفيزيولوجية لجسم الإنسان، على سبيل المثال يترافق ضغط الدم المرتفع بزيادة في النبض وزيادة التنفس.

استخدام خوارزمية KNN لتعويض القيم المفقودة: تفترض خوارزمية الجار الأقرب أن الكائنات المتشابهة تتواجد على مقربة من بعضها البعض، وهي بالتالي خوارزمية تعتمد معيار المسافة كمقياس للتقارب.

الفكرة وراء التعويض بخوارزمية الجار الأقرب هي تحديد النقاط الأكثر قرباً، وبالتالي الأكثر تشابهاً مع النقطة المراد تعويض قيمها المفقودة، ومن ثم استخدام هذه النقاط الأكثر تشابهاً لتقدير القيم المفقودة، باستخدام أحد الواصفات الإحصائية، كالمتوسط، أو المنوال.

في حالة مجموعة البيانات التي نعمل عليها، اعتمدنا معيار المسافة الإقليدية، وتم إيجاد مصفوفة المسافات، وهي مصفوفة مربعة، يمثل فيها العنصر $d(i,j)$ المسافة الإقليدية بين النقطة (i) والنقطة (j).

من أجل كل سمة، يتم تطبيق خوارزمية الجار الأقرب:

الدخل:

• K: عدد الجيران الأقرب

- مصفوفة المسافات
- طريقة التصويت: mean بالنسبة للقيم الحقيقية، Mode بالنسبة للقيم الثنائية.
- عتبة الفقد: يمثل أقل عدد ممكن من النقاط الجارة وليست قيم مفقودة، وفي حال تجاوز عدد القيم المفقودة هذه العتبة، تبقى القيمة المفقودة دون تعويض، وتعلم بـ Nan.
- تم اختيار قيمة K اعتماداً على القاعدة التجريبية التي تخبرنا أنه من بين أفضل القيم هو أقرب عدد فردي للجذر التربيعي لعدد النقاط، وعليه تم اختيار قيمة $K = 43$.

8. دمج أعمدة الوفاة ضمن عمود واحد

بالنسبة لعمود الهدف، تتضمن مجموعة البيانات المستخدمة 3 أعمدة منفصلة كل عمود منها يوضح أحد الحالات الممكنة للوفاة (الوفاة خلال 28 يوم من قبول المريض، الوفاة بعد 3 أشهر والوفاة بعد 6 أشهر). لهذا السبب تم إنشاء عمود جديد يمثل عمود الهدف Death ويتضمن كل حالات الخطورة الممكنة الناتجة عن أمراض القلب والتي تؤدي للفشل القلبي وهي حالات الوفاة الثلاثة وتم ترميزها بالأرقام 1 و2 و3 أما الرقم 0 فيعني أن المريض بقي على قيد الحياة.

2.3.4 موازنة مجموعة البيانات Dataset Balancing

لدى دراستنا وتحليلنا لمجموعة البيانات تبين أن القسم الأعظمي منها يتضمن حالات بدون وفاة (بقاء على قيد الحياة)، حيث هناك نسبة ما يزيد عن 98% تتبع للصنف 0 أو حالة عدم الوفاة وهناك 2% فقط تتبع لصنف 1 أو الوفاة مما يجعل المصنف دوماً يصنف العينات على أنها تتبع للصنف 0 ونحصل على دقة عالية للتدريب لكنها دقة وهمية حيث أن المصنف يصنف 98% من العينات على أنها 0 وبالتالي لو أخطأ في تصنيف كل عينات الصنف 1 ستبقى دقة التصنيف 98%.

لحل هذه المشكلة سوف نعتمد مقارنة إعادة تشكيل مجموعة البيانات عبر استخدام تقنية SMOTE حيث نقوم بزيادة عدد عينات الأصناف الأقلية (1،2،3) التي تمثل حالات الخطورة

بمضاعفة كل منها عشر مرات وفي هذه الحالة يمكن تدريب المصنفات دون الحصول على عملية تدريب وهمية.

تعمل خوارزمية SMOTE بالشكل التالي:

- يتم في البداية اختيار نقطة عشوائية من الصنف الأقلية
- تحديد أقرب K جار من النقطة المختارة (القيمة الافتراضية $K=5$)
- باستخدام أحد هذه الجيران يتم تحديد شعاع (متجهة) بين النقطة الأقلية المختارة، والجار لها.
- يتم ضرب المتجه بقيمة عشوائية من المجال $[0,1]$
- يتم إضافة النقطة الأقلية الجديدة التي تم توليفها.

ما تقوم به هذه التقنية هو تحريك نقطة البيانات قليلاً في اتجاه جارتها. هذا يضمن أن نقطة البيانات التركيبية الخاصة بك ليست نسخة طبق الأصل من نقطة بيانات موجودة، مع التأكد أيضاً من أنها ليست مختلفة جداً عن الملاحظات المعروفة في فئة الأقلية لديك.

بالنهاية حصلنا على مجموعة بيانات معالجة موازنة مؤلفة من 139 عمود بينها عمود الهدف .Death

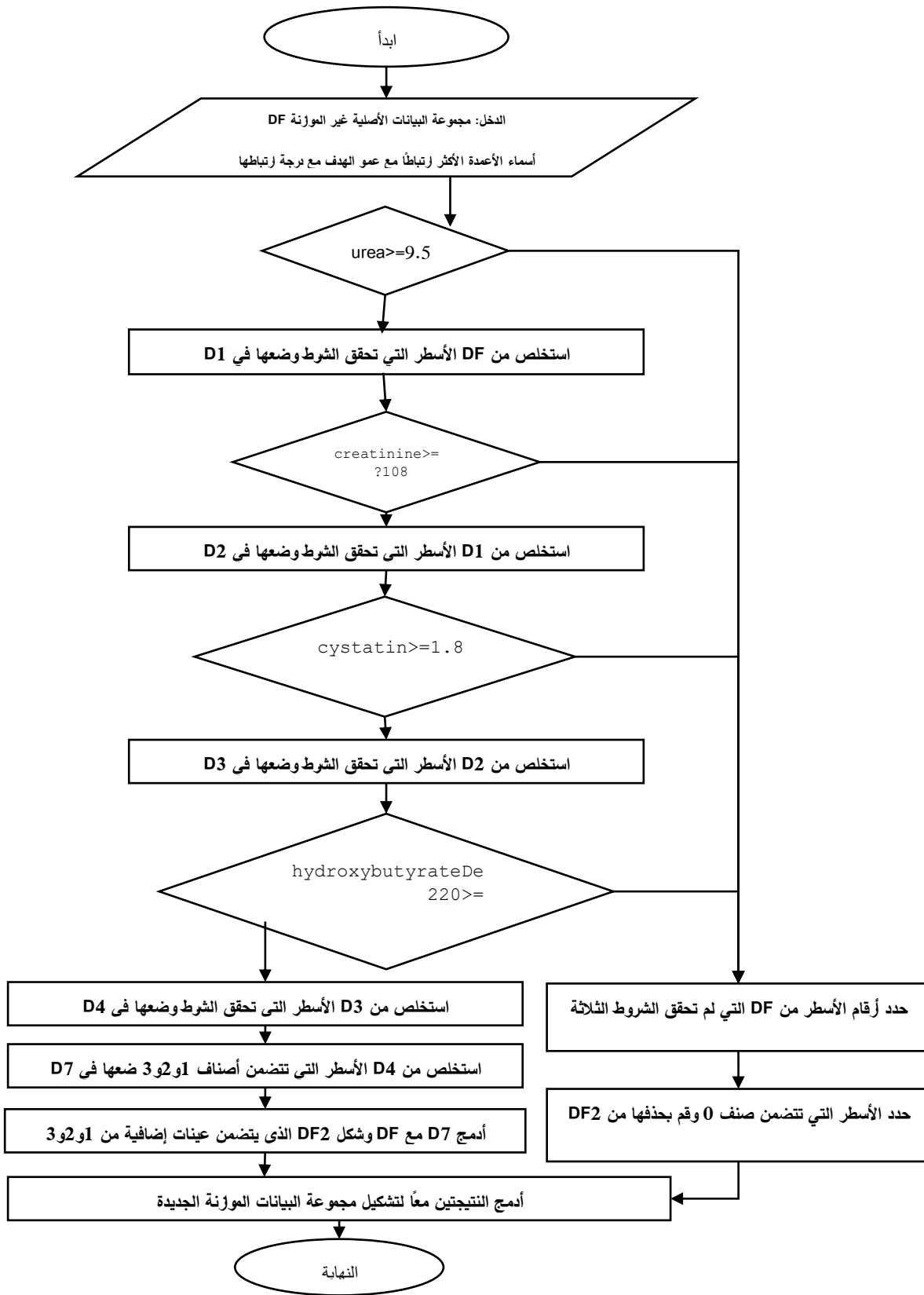
وتم قمنا بتقديم طريقة جديدة للموازنة مقترحة من قبلنا وهي خوارزمية الموازنة Mixed Up Down Sampling based Feature Correlation أو اختصاراً MUDSFC.

تم تطوير هذه الخوارزمية من خلال إضافة خطوة عنقدة Clustering الهدف منها تقليل التداخل Overlapping بين عينات الصنف 0 وعينات باقي الأصناف خصوصاً الصنفين (1,2).

تعتمد خوارزمية MUDSFC على زيادة عدد عينات الصنف غير المسيطر الصنف (1,2,3) وإنقاص عدد عينات الصنف المسيطر (0) ولكن بطريقة ديناميكية تعتمد على مدى ارتباط العينات مع عمود الهدف بحيث يتم مضاعفة العينات التي ترتبط بشكل كبير مع عمود الهدف وإلغاء العينات المرتبطة بشكل أقل مع عمود الهدف بالاعتماد على نموذج كثافة النواة لتوليد نقاط جديدة بهدف موازنة مجموعات البيانات من أجل إنجاز هذه المرحلة.

لدى تطبيق عملية اختيار السمات وجدنا أن السمات الأكثر ترابطاً مع عمود الهدف هي urea hydroxybutyrate.dehydrogenase ،cystatin،creatinine enzymatic.method وجميعها تملك معدل ارتباط أعلى من (0.3). تم استخدام هذه الأعمدة كأساس لاختيار الأسطر كالاتي، سنقوم بمضاعفة النقاط الأقلية التي تحقق قيمها أكبر من المتوسط لقيم العمود، وبالمقابل سنقوم بتقليل النقاط الأكثرية التي تحقق قيمها أصغر من المتوسط:

1. الأسطر التي تكون قيم عمود Urea لها أكبر من 9.5 يتم أخذها.
2. من نتيجة 1 يتم أخذ الأسطر التي تكون قيم عمود creatinine enzymatic.method لها أعلى من 108 والتي تنتمي للأصناف 1 و2 و3.
3. من نتيجة 2 يتم أخذ الأسطر التي تكون قيم عمود cystatin لها أعلى من 1.84 والتي تنتمي للأصناف 1 و2 و3.
4. من نتيجة (3) يتم أخذ الأسطر التي تكون قيم عمود hydroxybutyrate.dehydrogenase أعلى من (220) والتي تنتمي للأصناف (1,2,3).
5. يتم تطبيق عكس الشروط من 1 إلى 4 من أجل انتقاء عينات الصنف المسيطر (0).
6. نقوم بتوليف عينات جديدة من العينات التي تحقق الخطوات من 1 إلى 4 واحذف العينات التي تحقق الخطوة 5.
7. يتم توليف عينات من الصنف الأكثرية بحيث يتم الوصول إلى عدد العينات في حالة الموازنة بطريقة SMOTE وبهذا يتم الوصول إلى مجموعة البيانات الموازنة.

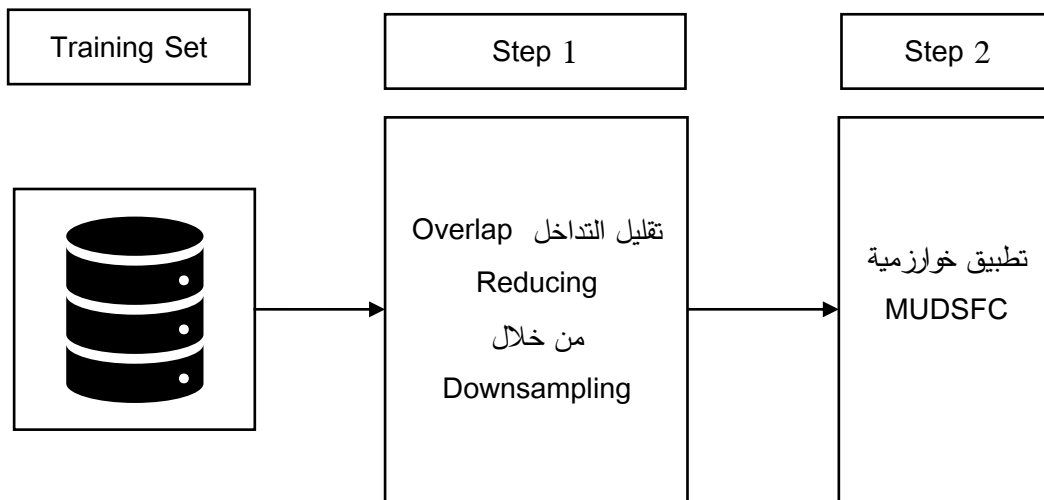


الشكل (3-4) المخطط الانسيابي الموضح لخوارزمية MUDSFC

3.3.4. خوارزمية موازنة مجموعة البيانات المعتمدة على تقليل التداخل بين

الأصناف

استطاعت الخوارزمية MUDSFC المقترحة تحقيق موازنة لمجموعة البيانات بشكل جيد، لكن دون الاعتماد على مبدأ تقليل التداخل بين الأصناف. في التطوير على هذه الخوارزمية قمنا بإضافة مرحلة قبل عملية الموازنة وهي مرحلة العنقدة Clustering التي تهدف إلى تجميع عينات مجموعة البيانات في عناقيد مترابطة معاً بحيث ينتهي لنا معرفة العينات المرتبطة ببعضها البعض بين الأصناف وبالتالي إمكانية إلغائها وفقاً لعتبة إلغاء محددة. يوضح الشكل (4-4) التعديل في بنية خوارزمية الموازنة من خلال مرحلة تقليل التداخل.



الشكل (4-4) التعديل في خوارزمية MUDSFC بإضافة مرحلة تقليل التداخل

يتم في مرحلة تقليل التداخل تطبيق خوارزمية العنقدة DBScan وفق المراحل الآتية:

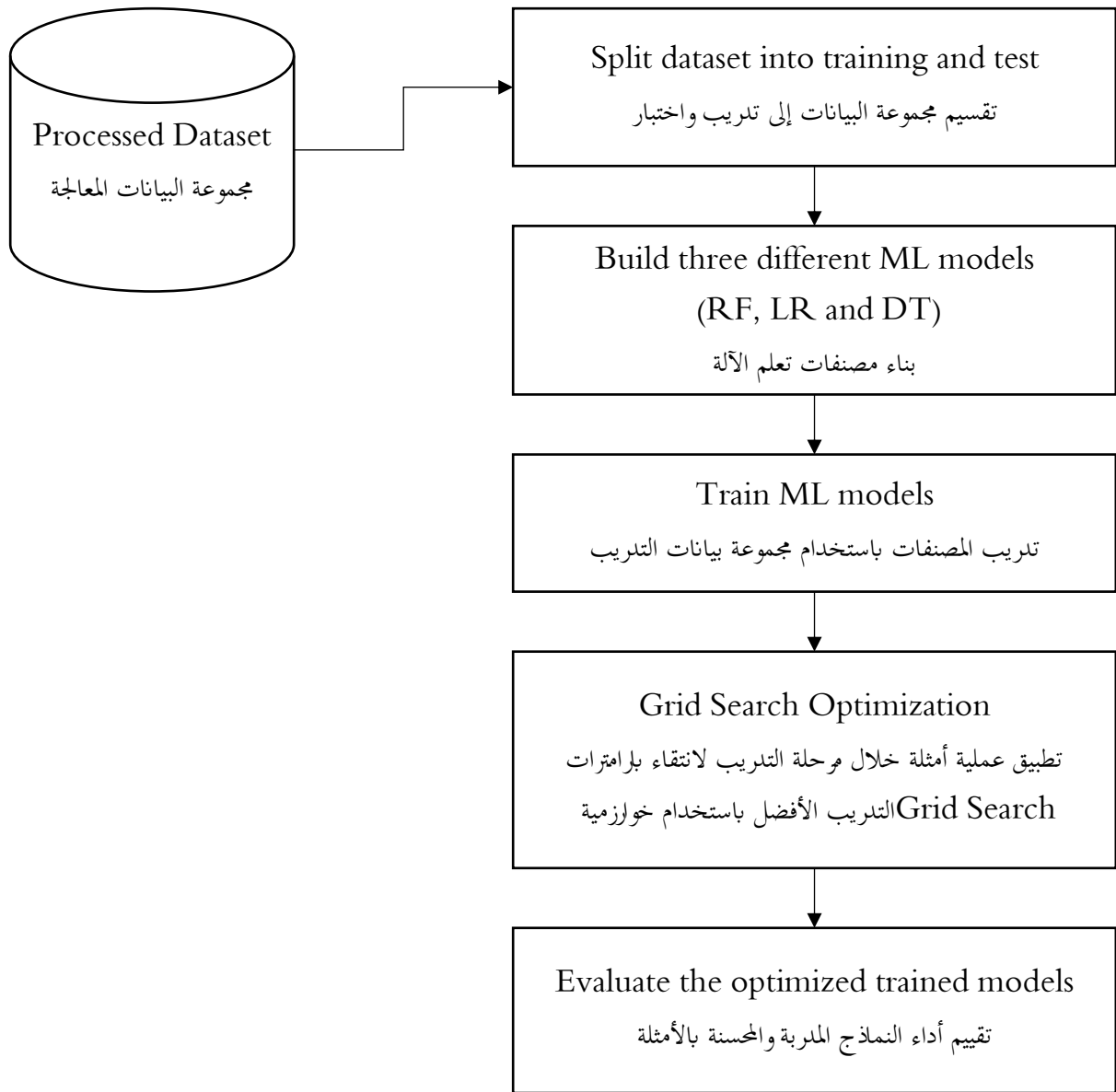
- تطبيق خوارزمية العنقدة DBScan
- تحديد نقاط الأكثرية Majority Points والاحتفاظ بالقسم الأكبر منها
- تحديد نقاط الأقلية والأكثرية Majority and Minority وحذف النقاط الأقرب منها إلى النقاط الأقلية (النقاط التي تمثل التداخل بين الصنف الأساسي أو عينات Majority وبقية الأصناف أو عينات Minority).

في نهاية هذه المرحلة نكون قد أَلغينا عينات الصنف 0 Majority الأكثر ارتباطاً بالأصناف الأخرى والذي من شأنه تقليل التداخل وتحسين دقة التصنيف.

- تطبيق خوارزمية MUDSFC لزيادة عدد العينات Upsampling فقط.
- يوضح الشكل (4-5) خطوات تقليل تداخل الأصناف المستخدمة في الخوارزمية المعدلة من خوارزمية MUDSFC.

4.4. خطوات عملية تدريب نماذج تعلم الآلة مع التحسين

يتضمن الشكل (4-5) الخطوات المقترحة لعملية تدريب نماذج تعلم الآلة مع استخدام خوارزمية التحسين (الأمثلة).



الشكل (4-5) الخطوات المقترحة لعملية تدريب نماذج تعلم الآلة مع استخدام خوارزمية التحسين (الأمثلة).

تقوم خوارزمية Grid Search بتحسين أداء شبكة من بارامترات نماذج تعلم الآلة معاً في كل تكرار من تكرارات التدريب بحيث يتم الوصول مع الوقت إلى أفضل تركيبة لتقييم هذه البارامترات تحقق أفضل أداء. تقسم خوارزمية Grid Search فضاء البارامترات إلى شبكة من القيم ثم تقوم بتجريب كل تركيبة ممكنة من هذه البارامترات معاً حتى تصل إلى أفضل تركيبة تحقق الأداء الأفضل (مما يعني زمن تدريب أطول لكن أداء أفضل) [39].

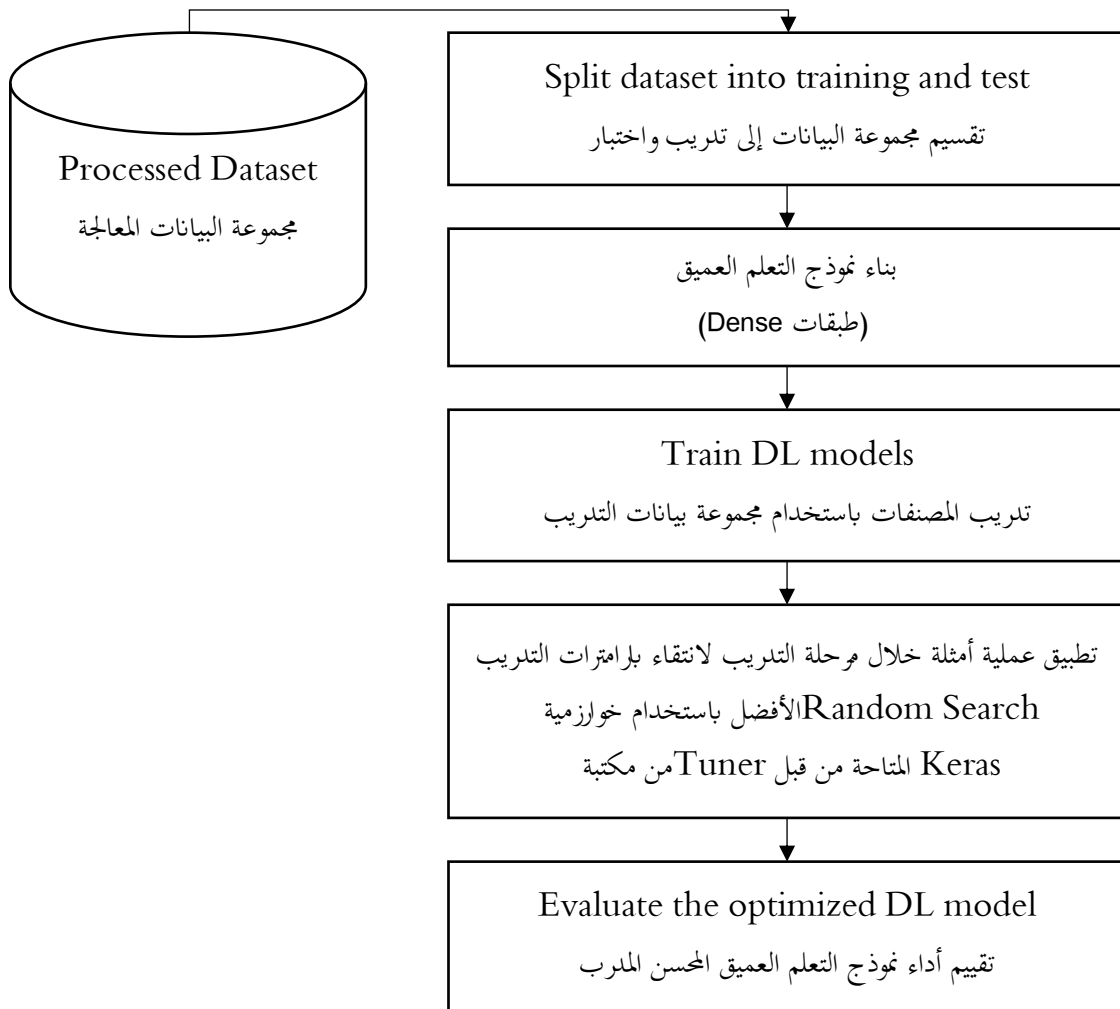
باستخدام خوارزمية الأمثلة تم تدريب ثلاث مصنفات تعلم آلة هي:

- مصنف غابات القرار (RF) Random forests
- مصنف أشجار القرار (DT) Decision Trees
- مصنف الانحدار المنطقي (LR) Logistic Regression

5.4. نموذج التعلم العميق المقترح مع طريقة التحسين المقترحة

يقترح البحث استخدام نموذج تعلم عميق مبني على أساس شبكة تعلم عميق قائم على طبقات التخفيض Dense Layer، يتألف النموذج المقترح استخدامه من 5 طبقات Dense Layers تمثل الأولى طبقة الدخل Input layer وهي تستقبل السمات والتي تمثل عناصر سطر واحد من أسطر من أسطر مجموعة البيانات. أما الطبقات الأربعة التالية فهي تمثل الطبقات المخفية حيث تم استخدام 50 عصبون في كل منها. أما طبقة Dense الأخيرة فهي طبقة الخرج وتمثل عدد عصبوناتها (4) عدد الأصناف في مجموعة البيانات (حالة البقاء على قيد الحياة وحالات الوفاة الثلاثة). تم استخدام طبقات Dense في كل النموذج لسهولة تدريبها كونها طبقات كاملة الاتصال وكونها ديناميكية حيث يمكن استخدامها طبقات دخل أو خفية أو خرج. تم استخدام تابع التنعيل Relu في الدخل والطبقات الخفية لسهولة حسابه ولكونه لا يأخذ زمناً طويلاً في الحسابات أما في الخرج فقد تم استخدام تابع Softmax لكونه يقوم بحساب

احتماليات المخارج كقيم في المجال من 0 إلى 1 وبالتالي يكون الخرج هو الصنف الذي حقق أعلى احتمالية. يوضح الشكل (4-6) نموذج التعلم العميق المقترح بدون عملية التحسين.



الشكل (4-6) نموذج التعلم العميق المقترح بدون عملية التحسين

يتضمن الشكل (4-7) شكل نموذج التعلم العميق المقترح والطبقات ضمنه وعدد البارامترات اللازم تعليمها حيث تم تصميم النموذج مع كامل البارامترات من قبلنا.

تم استخدام خوارزمية التحسين Random Search [40] من أجل تحسين تدريب نموذج التعلم العميق من خلال اختيار عدة بارامترات تدريب عشوائية وتحسين قيمها مع كل تكرار

والسبب في اختيارها كونها تشبه لحد كبير خوارزمية Grid Search التي استخدمناها في نماذج تعلم الآلة. مكتبة Random Search متاحة من قبل مكتبة Keras.

في مرحلة Compile للنموذج تم استخدام طريقة Adam ك Optimizer لعملية التدريب وتم اختيار طريقة Adam كونها تناسب حالات High Dimensional data أي البيانات ذات عدد الخصائص أو السمات الكبير (وهي حالتنا المدروسة تماماً).

كما تم اختيار Categorical Cross Entropy كتابع خسارة والسبب أن مسألة التصنيف هنا تتضمن تصنيف 4 أصناف وليس صنف واحد (0 أو 1) بالتالي لا نستطيع استخدام Binary Cross Entropy، تم تدريب النموذج 100 تكرار وتم استخدام GPU الخاص بـ Google Colab من أجل عملية التدريب حيث تم اعتماد Patch Size = 50 وبالتالي سيقوم المعالج بإعطاء النموذج 50 عينة تدريب كل دفعة حتى تنتهي كامل عينات التدريب (وذلك في كل تكرار من تكرارات الخوارزمية).

```
Model: "sequential_1"
```

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 50)	6950
dense_1 (Dense)	(None, 50)	2550
dense_2 (Dense)	(None, 50)	2550
dense_3 (Dense)	(None, 50)	2550
dense_4 (Dense)	(None, 50)	2550
dense_5 (Dense)	(None, 50)	2550
dense_6 (Dense)	(None, 50)	2550
dense_7 (Dense)	(None, 50)	2550
dense_8 (Dense)	(None, 4)	204

```

Total params: 25,004
Trainable params: 25,004
Non-trainable params: 0

```

الشكل (4-7) معمارية نموذج التعلم العميق المصمم من قبلنا

6.4. طريقة تقييم الأداء (معاملات قياس الأداء)

تم استخدام معاملات قياس الأداء التالية، والتي تعطي علاقتها وفق الآتي [41]:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1-score} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

المتغير	الدلالة
TP	عدد العينات التي صنفتها الحقيقي هو 1 (فشل قلبي أدى للوفاة) وتم تخمين أنها تنتمي للصنف 1 بشكل صحيح.
TN	عدد العينات التي تمثل الأشخاص الذين لم يعانون من فشل قلبي ووفاة (بقاء على قيد الحياة) وتم تخمين ذلك من قبل النموذج.
FP	عدد العينات التي خمنها النموذج على أنها تنتمي للصنف 1 (الوفاة نتيجة مرض القلب) إلا أنها تنتمي في الحقيقة للصنف (0) أو البقاء على قيد الحياة.
FN	عدد العينات التي خمنها النموذج على أنها تنتمي للصنف 0 (بقاء على قيد الحياة) إلا أنها تنتمي في الحقيقة للصنف (1) أو الوفاة.

الجدول (4-4) معاملات قياس الأداء

6.4. الخلاصة

قدّمنا في هذا الفصل شرحاً عن الدراسة العملية التي قمنا بها في البحث، وتم عرض الخوارزميات المقترحة والخوارزميات المصممة من قبلنا وطرق التعلم العميق وتعلم الآلة المستخدمة في البحث.

5. الفصل الخامس: الاختبارات والنتائج

نستعرض في هذا الفصل الاختبارات والنتائج التي قمنا بها على النماذج المقترحة

1.5.1 مقدمة

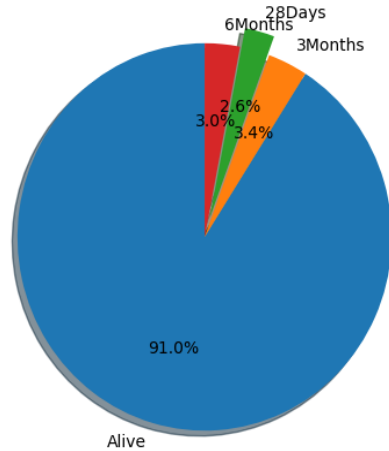
نستعرض في هذا الفصل الاختبارات والنتائج التي قمنا بها على النماذج المختلفة التي تم اقتراحها في الدراسة. كما نستعرض سبعة سيناريوهات تدريب مختلفة مع نتائج الاختبارات الخاصة بها.

2.5. سيناريوهات التدريب المقترحة

تم تنفيذ التجارب العملية على مجموعة البيانات المعالجة والموازنة بعد تقسيمها إلى 80% تدريب و20% اختبار. تم اقتراح 6 سيناريوهات تدريب مختلفة مع اختبار هذه السيناريوهات وتقييمها باستخدام معاملات تقييم الأداء، وهذه السيناريوهات هي:

1.2.5. التدريب على مجموعة البيانات الأصلية.

تم في هذه المرحلة تدريب مصنفات تعلم الآلة على مجموعة البيانات المعالجة دون أن تخضع لعملية الموازنة، ودون تحسين لأي من النماذج، الشكل التالي يظهر توزيع الأصناف في مجموعة البيانات الأصلية:



تم تقسيم مجموعة البيانات إلى مجموعتي تدريب واختبار بنسبة 80:20، وبالتالي حجم مجموعة بيانات الاختبار 402 عينة.

يظهر الجدول (1-5) نتائج التحقق عبر تطبيق Stratified K-Fold Cross Validation على مصنف الغابة العشوائية (RF).

Fold	Accuracy	Precision	Recall	F1_score	AUC
0	0.981	0.975	0.444	0.975	0.000
1	0.981	0.975	0.444	0.975	0.000
2	0.987	0.981	0.555	0.983	0.000
3	0.981	0.969	0.375	0.974	0.854
4	0.975	0.950	0.250	0.962	0.909
5	0.975	0.956	0.250	0.965	0.949
6	0.981	0.975	0.444	0.975	0.000
7	0.975	0.975	0.444	0.975	0.000
8	0.975	0.950	0.333	0.962	0.000
9	0.981	0.975	0.444	0.975	0.000
Mean	0.980	0.968	0.398	0.972	0.271
Std	0.0037	0.0108	0.0917	0.0063	0.4150

الجدول (1-5) التحقق المتقاطع لمصنف الغابة العشوائية

فيما يظهر الجدول (2-5) نتائج التحقق للمصنفات الثلاثة التي تم تدريبها على مجموعة البيانات بدون موازنة.

Model/Metrics	Accuracy	Precision	Recall	F1_score	AUC
RF	0.989	0.981	0.569	0.985	0.286
LR	0.964	0.961	0.35	0.962	0.165
DT	0.986	0.988	0.714	0.976	0.249

الجدول (2-5) نتائج اختبار على مجموعة البيانات المعالجة بدون موازنة أو تحسين

حيث يتضح من الجدول (1-5) أن النماذج الثلاثة لا يمكن الاعتماد عليها في حالة التدريب على مجموعة البيانات بدون موازنة بسبب القيم المنخفضة لمقيم AUC والتي تعني عدم قدرة هذه المصنفات في هذه الحالة على التمييز بين الأصناف الأقلية.

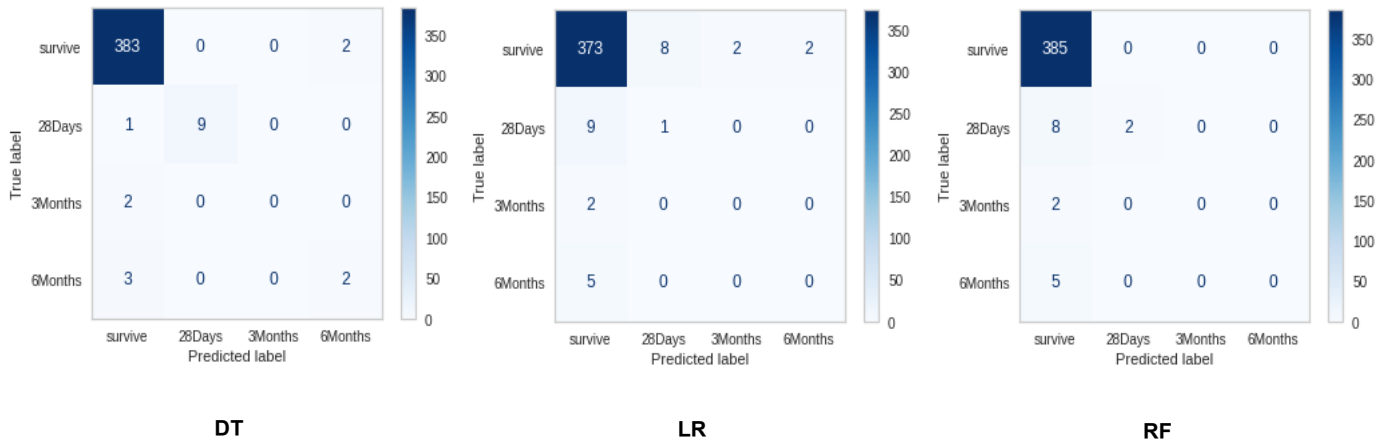
ثم قمنا بإعادة تدريب النماذج السابقة مع أمثلة للمعلمات الفائقة (Hyperparameter)، وكانت النتائج التالية:

Model/Metrics	Accuracy	Precision	Recall	F1_score	AUC
RF	9130.9	810.98	76280.	940.98	5940.2
LR	6640.9	230.96	6820.3	400.96	7730.1
DT	9070.9	730.98	2290.7	8870.9	2940.2

الجدول (3-5) نتائج تدريب المصنفات مع الأمثلة

فيما يظهر الشكل التالي نتائج مصفوفات التشتت للنماذج المدربة، مع وبدون أمثلة المعلمات الفائقة.

قبل الأمثلة



بعد الأمثلة

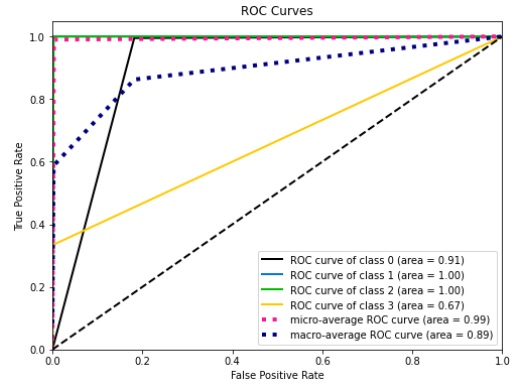
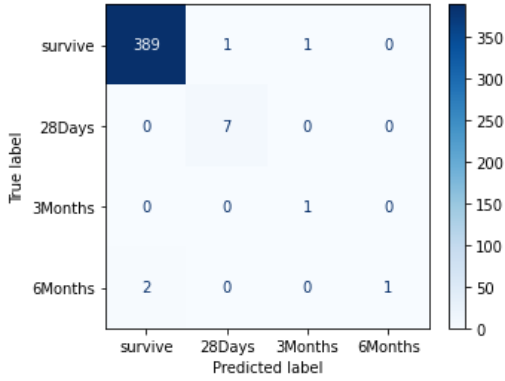


2.2.5. التدريب مع الموازنة باستخدام SMOTE

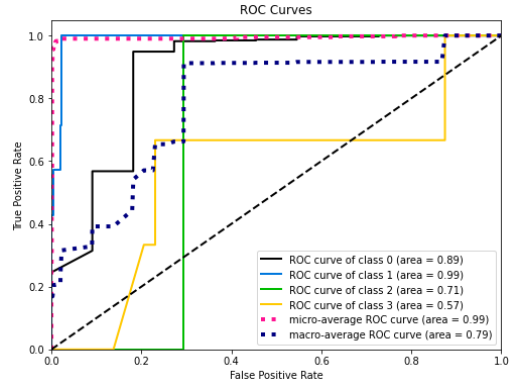
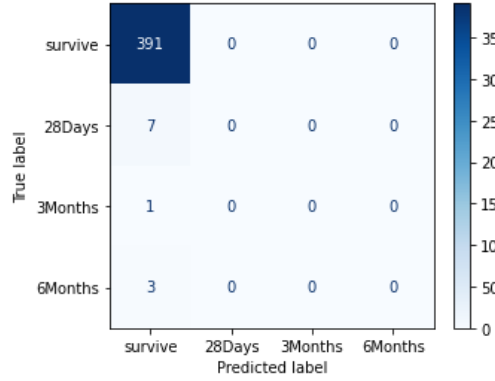
تم تطبيق الموازنة باستخدام تقنية SMOTE، حيث تم بداية تقسيم مجموعة البيانات إلى مجموعتي تدريب واختبار بواقع 80:20، تم بعد ذلك تطبيق عملية الموازنة باستخدام تقنية SMOTE على مجموعة بيانات التدريب، بحيث يتم مضاعفة عدد عينات الأصناف الأقلية 10 مرات لكل منها.

يظهر الشكل (3-5) مخططات AUC، ومصفوفات التشتت للنماذج المدربة على مجموعة التدريب الموازنة.

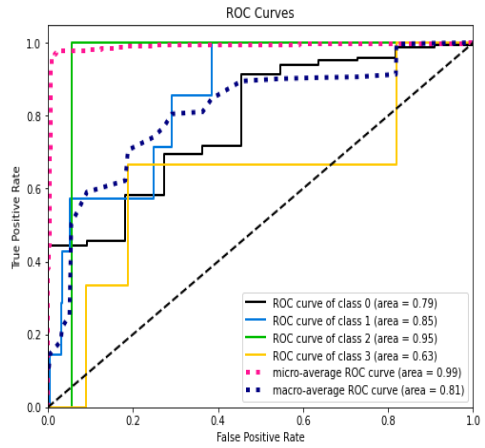
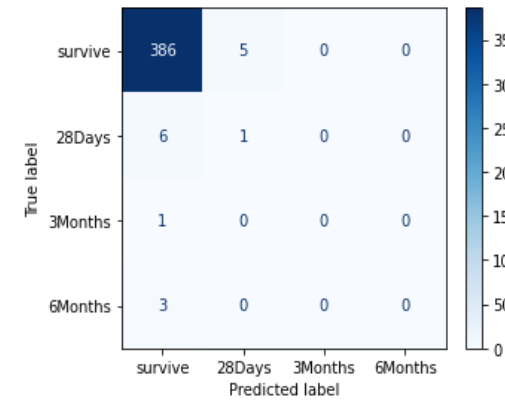
DT



RF



LR



الشكل (5-1) الموازنة باستخدام SMOTE قبل الأمتلة

فيما يظهر الشكل التالي نتائج تدريب النماذج السابقة (classification report)

	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.97	1.00	0.99	391	0	0.97	0.99	0.98	391
1	0.00	0.00	0.00	7	1	0.17	0.14	0.15	7
2	0.00	0.00	0.00	1	2	0.00	0.00	0.00	1
3	0.00	0.00	0.00	3	3	0.00	0.00	0.00	3
accuracy			0.97	402	accuracy			0.96	402
macro avg	0.24	0.25	0.25	402	macro avg	0.29	0.28	0.28	402
weighted avg	0.95	0.97	0.96	402	weighted avg	0.95	0.96	0.96	402

RF

LR

	precision	recall	f1-score	support
0	0.99	0.99	0.99	391
1	0.88	1.00	0.93	7
2	0.50	1.00	0.67	1
3	1.00	0.33	0.50	3
accuracy			0.99	402
macro avg	0.84	0.83	0.77	402
weighted avg	0.99	0.99	0.99	402

DT

فيما يظهر الجدول (4-5)، نتائج التحقق للمصنفات الثلاثة المدربة على المجموعة الموزنة.

Model/Metrics	Accuracy	Precision	Recall	F1_score
RF	0.9726	0.9460	0.9726	0.9591
LR	96280.	95090.	96260.	95670.
DT	9000.9	9160.9	9000.9	8930.9

الجدول (4-5) نتائج التحقق على مجموعة البيانات الموزنة باستخدام SMOTE

نلاحظ أن النماذج ما زالت غير كفؤة بالنظر إلى مصفوفات التشتت، وتقارير التصنيف، ولا يمكن الاعتماد عليها، والملاحظ أنها فشلت في التعرف على عينات الاختبار الأقلية، أي التي تمثل حالات الوفاة، وقامت بتصنيف معظمها من الصنف 0 أي البقاء على قيد الحياة.

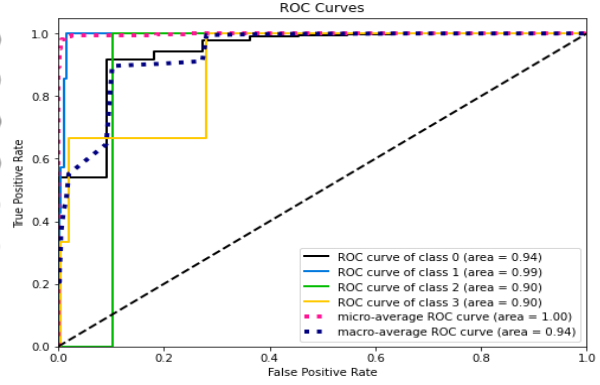
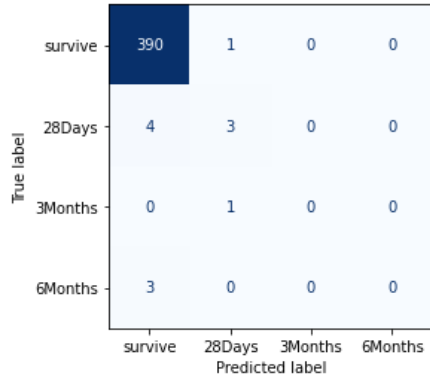
قمنا بعد ذلك بإعادة تدريب النماذج السابقة مع أمثلة المعلمات الفائقة، وفق شبكة البارامترات التالية:

RF, DT	Max-depth: [2,3,4,5,10,15,20,30] Criterion: [Gini, entropy] Class-weight: [Balanced, none]
LR	Solver: [sag, saga, newton-cg, lbfgs] Class-weight: [balanced, none] C: [0.001,0.01,1,10,100] Penalty: [11,12]

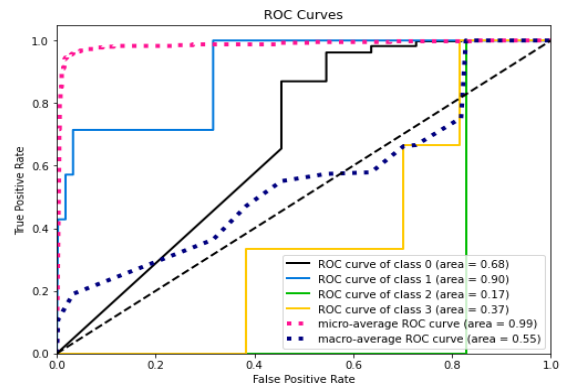
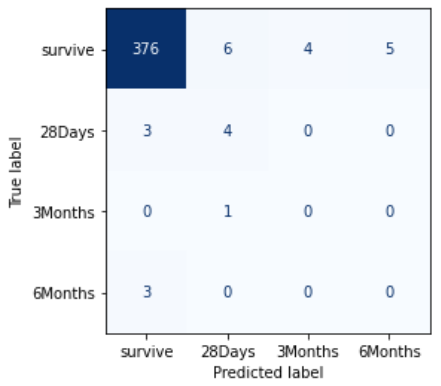
الجدول (5-5) شبكة المعلمات الفائقة

يظهر الشكل (4-5) مخططات AUC، ومصفوفات التشتت للنماذج المدربة على مجموعة التدريب الموازنة مع أمثلة المعلمات الفائقة.

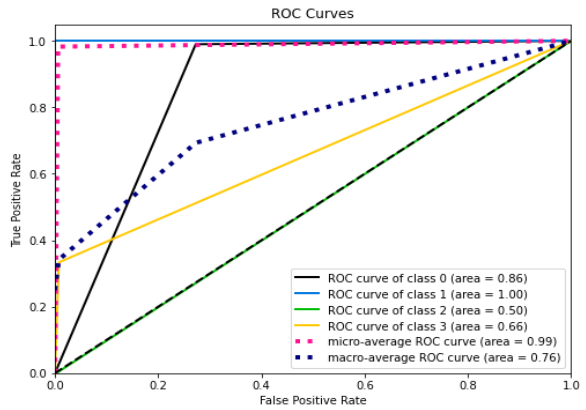
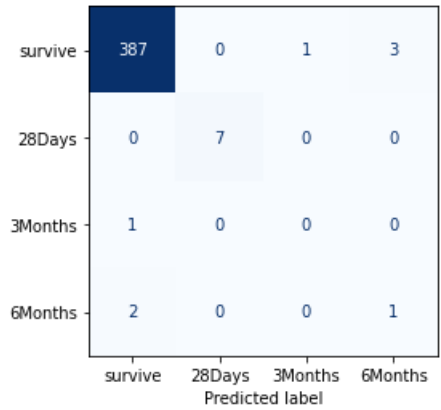
RF



LR



DT



الشكل (5-2) الموازنة باستخدام SMOTE مع الأمثلة

فيما يظهر الشكل التالي نتائج تدريب النماذج السابقة (classification report)

	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.98	1.00	0.99	391	0	0.98	0.96	0.97	391
1	0.60	0.43	0.50	7	1	0.36	0.57	0.44	7
2	0.00	0.00	0.00	1	2	0.00	0.00	0.00	1
3	0.00	0.00	0.00	3	3	0.00	0.00	0.00	3
accuracy			0.98	402	accuracy			0.95	402
macro avg	0.40	0.36	0.37	402	macro avg	0.34	0.38	0.35	402
weighted avg	0.97	0.98	0.97	402	weighted avg	0.96	0.95	0.95	402

RF

LR

	precision	recall	f1-score	support
0	0.99	0.99	0.99	391
1	1.00	1.00	1.00	7
2	0.00	0.00	0.00	1
3	0.25	0.33	0.29	3
accuracy			0.98	402
macro avg	0.56	0.58	0.57	402
weighted avg	0.98	0.98	0.98	402

DT

من النتائج السابقة نلاحظ أن النماذج ما زالت تعاني في تمييز الأصناف الأقلية من الصنف 2 و3، وذلك بسبب العدد القليل من العينات الموجودة في مجموعة البيانات، لذلك سنلجأ إلى مقارنة إعادة تشكيل البيانات.

Model/Metrics	Accuracy	Precision	Recall	F1_score
RF	0.9761	0.9659	0.9776	0.9714
LR	94520.	96360.	94520.	95390.
DT	8250.9	8440.9	8250.9	8340.9

الجدول (5-6) نتائج التحقق باستخدام SMOTE مع الأمثلة

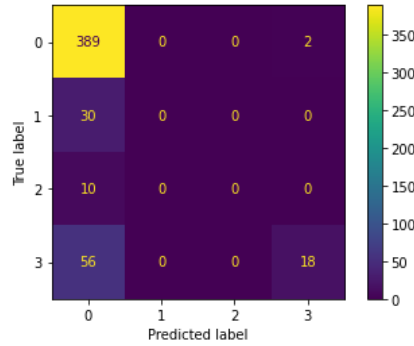
3.2.5. تدريب مصنفات تعلم الآلة بدون خوارزمية التحسين

تم في هذا النموذج تدريب مصنفات تعلم الآلة الثلاثة المقترحة بدون استخدام خوارزمية تحسين والنتائج موضحة في الجدول (5-1).

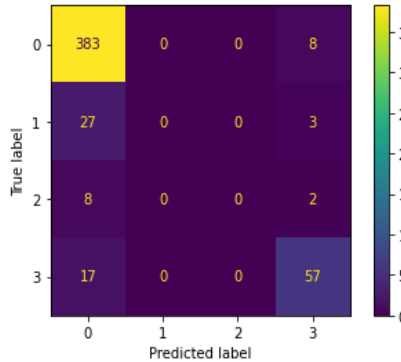
Model/Metrics	Accuracy %	Precision %	Recall %	F1_score %
RF	80.5	75.2	80.5	77.7
LR	80.9	73.7	80.9	76.49
DT	87.1	80.1	87.1	83.4

الجدول (5-7) نتائج اختبار مصنفات تعلم الآلة المدربة بدون تحسين

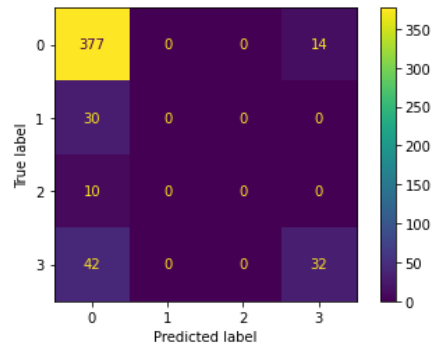
يوضح الشكل (5-5) ناتج اختبار أنظمة تعلم الآلة بدون تحسين في Google Colab مع مصفوفات التشتت Confusion matrix لكل منها.



RF



DT



LR

الشكل (5-3) مصفوفات التشتت Confusion Matrixes لنماذج تعلم الآلة الثلاثة المدربة بدون تحسين

حيث يتضح من الشكل (5-5) أن مصفوفة التشتت للنماذج الثلاثة نجحت في اتخاذ القرار للصنفين الأول والرابع فقط (حالة عدم الوفاة وحالة الوفاة بعد 6 أشهر) وفشلت في تحديد باقي الحالات، وتبين الاختبارات أن جميع النماذج (غير المحسنة) لم تتمكن من تمييز عينات الصنفين (2,3)

4.2.5. تدريب مصنفات تعلم الآلة مع خوارزمية التحسين

تم في هذا السيناريو تدريب مصنفات تعلم الآلة الثلاثة المقترحة مع استخدام خوارزمية التحسين Grid Search وتم التوصل للنتائج الموضحة ضمن الجدول (5-2).

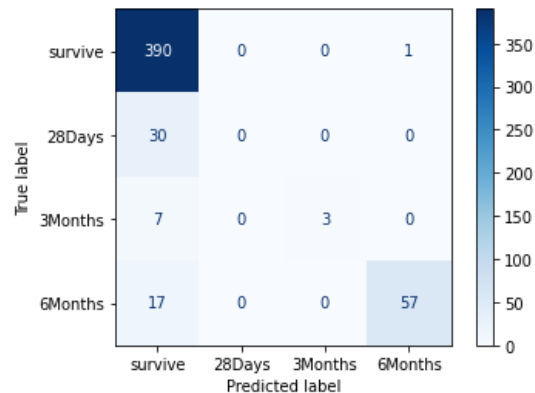
RF	89.1	84.39	89.1	85.89
LR	92.67	94.89	92.67	93.16
DT	85.55	83.82	85.54	84.67

الجدول (5-8) نتائج اختبار نماذج تعلم الآلة المدربة بعد تطبيق عملية التحسين

أما مصفوفات التشتت مع نتائج قياسات الأداء الخاصة بنماذج تعلم الآلة بعد التحسين موضحة في الشكل (5-6).

	precision	recall	f1-score
0	0.88	1.00	0.93
1	0.00	0.00	0.00
2	1.00	0.30	0.46
3	0.98	0.77	0.86
accuracy			0.89
macro avg	0.72	0.52	0.56
weighted avg	0.84	0.89	0.86

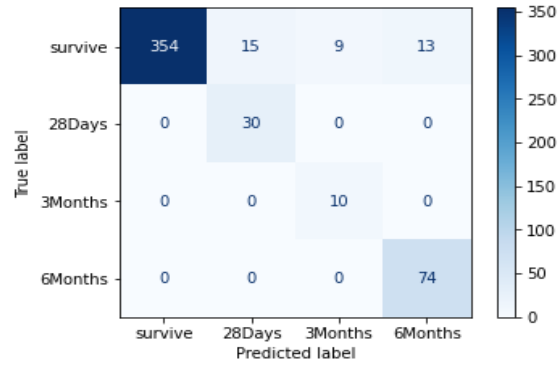
قياسات الأداء لمصنف RF المحسن



مصفوفة التشتت لمصنف RF المحسن

	precision	recall	f1-score
0	1.00	0.91	0.95
1	0.67	1.00	0.80
2	0.53	1.00	0.69
3	0.85	1.00	0.92
accuracy			0.93
macro avg	0.76	0.98	0.84
weighted avg	0.95	0.93	0.93

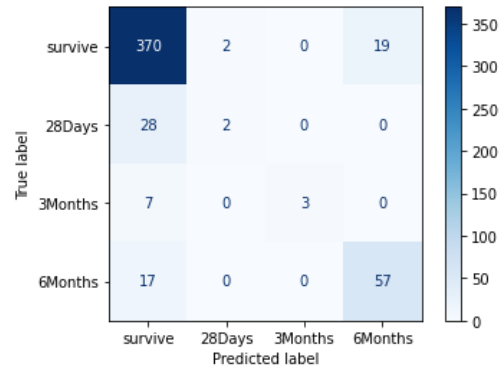
قياسات الأداء لمصنف LR المحسن



مصفوفة التشتت لمصنف LR المحسن

	precision	recall	f1-score
0	0.88	0.95	0.91
1	0.50	0.07	0.12
2	1.00	0.30	0.46
3	0.75	0.77	0.76
accuracy			0.86
macro avg	0.78	0.52	0.56
weighted avg	0.84	0.86	0.83

قياسات الأداء لمصنف DT المحسن



مصفوفة التشتت لمصنف DT المحسن

الشكل (4-5) مصفوفات التشتت *Confusion Matrixes* للنماذج الثلاثة المدربة مع تحسين

يمكن الملاحظة من النتائج الموضحة في الشكل (5-6) أن عملية التحسين (الأمثلة) أدت للحصول على نماذج أكثر دقة وذات أداء أعلى لكل من LR, DT, RF، والمشكلة الوحيدة في عملية الأمثلة هي زمن التدريب الذي يستغرق وقتاً طويلاً للعثور على النموذج الأفضل. وكان نموذج LR هو الأكثر تحسناً بين النماذج الثلاثة.

من حيث دقة الاختبار، نلاحظ أن نموذج RF تحسنت فيه الدقة بمقدار 8.6% بعد استخدام عملية التحسين (الأمثلة)، ونموذج LR قد تحسن بمقدار 11.77% في حين أن نموذج DT لم يتحسن أدائه باستخدام الأمثلة.

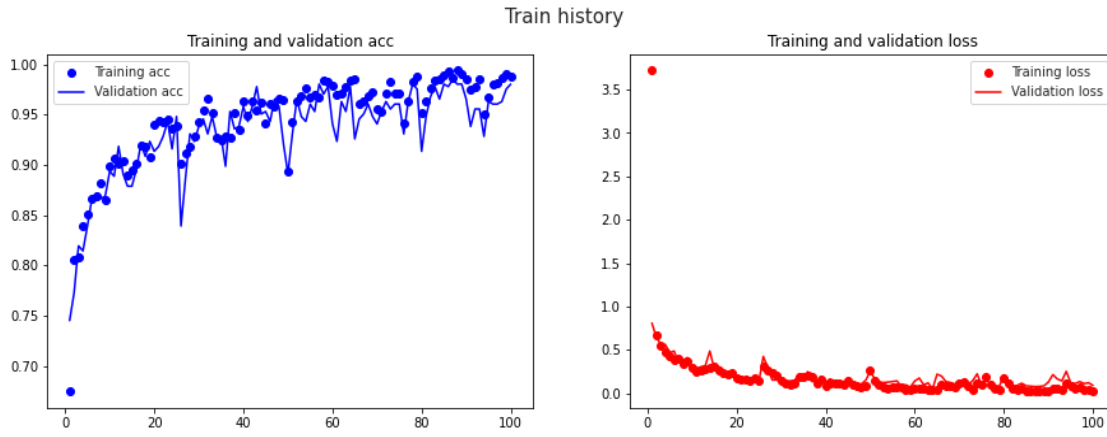
من حيث Precision تحسن أداء نموذج RF بمقدار 9.19%، في حين تحسين أداء نموذج LR بمقدار 21.19%. أما من حيث Recall وهو مؤشر يدل على معدل القبول الصحيح في نظام دعم القرار، فقد تحسن الأداء بمقدار 8.6% لنموذج RF و11.77% لنموذج LR. ومن حيث F1-Score تحسن الأداء بمقدار 11.59% و16.67% لنموذجي RF وLR على الترتيب.

تبين مصفوفات التشتت وقياسات الأداء أنّ نموذج الانحدار المنطقي LR هو النموذج الأفضل بعد عملية التحسين ونلاحظ من مصفوفة التشتت أنّ هناك مشكلة في تصنيف الأصناف 1 و2 في كل المصنفات باستثناء مصنف LR المحسن.

كنتيجة نهائية يمكن القول أنّ نموذج Optimized LR حقق الأداء الأفضل من حيث كل بارامترات الأداء بعد تطبيق عملية التحسين.

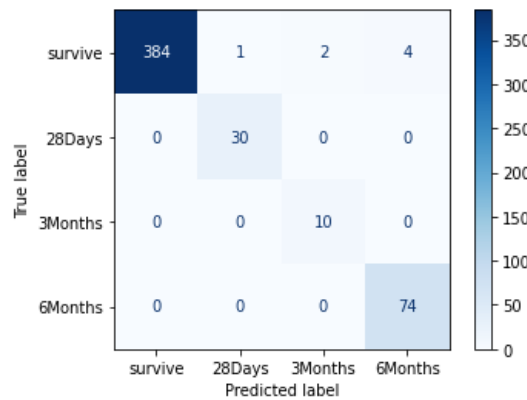
5.2.5. تدريب نموذج التعلم العميق بدون تحسين

يتضمن الشكل (5-7) ناتج عملية تدريب نموذج التعلم العميق حيث يبين الشكل منحنى الدقة Accuracy مقابل عدد التكرارات ومنحنى الخسارة (الخطأ) مقابل عدد التكرارات، وكما هو واضح أنّ منحنى الدقة يرتفع تدريجياً ليلبغ قيمته الأعظمية في حين أنّ منحنى الخسارة يتناقص مع زيادة عدد التكرارات ليلبغ قيمته الأدنى.



الشكل (5-5) ناتج عملية التدريب: منحني الدقة *Accuracy* ومنحني الخسارة *Loss*

كنتيجة حصلنا على دقة تدريب 99.07% ودقة تحقق 98.02% ودقة اختبار 98.61% . يوضح الشكل (5-8) مصفوفة التشتت لنموذج التعلم العميق حيث يتضح أن الأصناف الثاني والثالث والرابع لا تتضمن أية أخطاء في حين أن صنف *Survive* أو البقاء على قيد الحياة يتضمن 384 عينة صنف بشكل صحيح و 7 عينات صنف بشكل خاطئ. لمعرفة النتائج التفصيلية يتضمن الجدول (5-9) النتائج التفصيلية لعملية تدريب نموذج التعلم العميق قبل تطبيق خوارزمية تحسين *Random Search*.



الشكل (5-6) مصفوفة التشتت لنموذج التعلم العميق قبل تطبيق خوارزمية التحسين *Random Search*

Type of calculation	Validation Accuracy %	Test Accuracy %	Precision %	Recall %	F1_score %
Class 0			100	98	99
Class 1			97	100	98
Class 2			83	100	91
Class 3			95	100	97
Real Values	98.02	98.61	98.72	98.61	98.63
Macro Average	98.02	98.61	94	100	96
Weighted Average	98.02	99	99	99	99

الجدول (5-9) نتائج اختبار نموذج التعلم العميق المدرب قبل تطبيق عملية التحسين Optimization

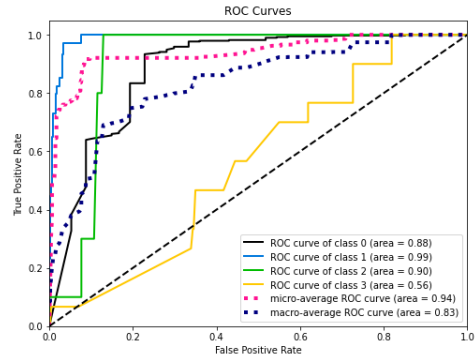
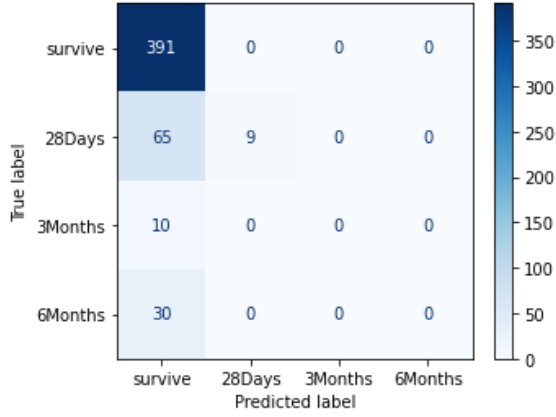
يبين الجدول (5-9) أنَّ أفضل صنف في عملية التصنيف من حيث الدقة كان الصنف (0) وهذا طبيعي كون أنَّ عدد عيناته كان أكبر مقارنة بباقي الأصناف.

6.2.5. التدريب باستخدام إعادة التشكيل (SMOTE):

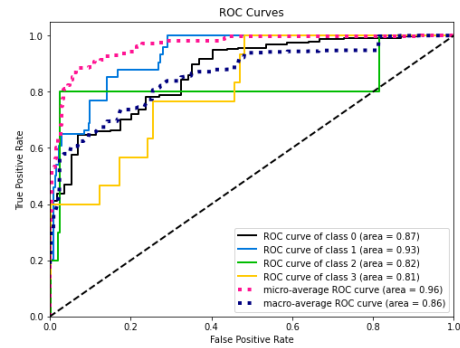
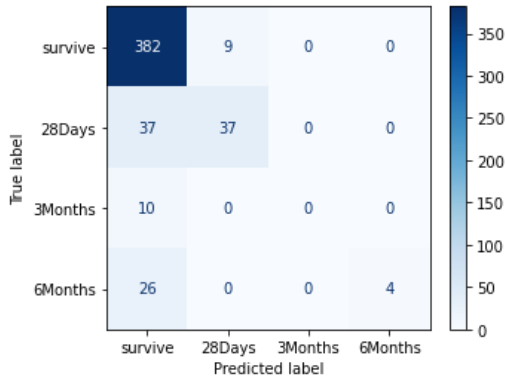
حيث سنقوم في هذه السيناريو بإعادة تشكيل مجموعة البيانات بحيث يتم زيادة عدد عينات الأقلية بمقدار 10 مرات لكل صنف قبل التدريب، ثم تتم عملية الأمثلة للمعلمات الفائقة.

النتائج قبل الأمثلة:

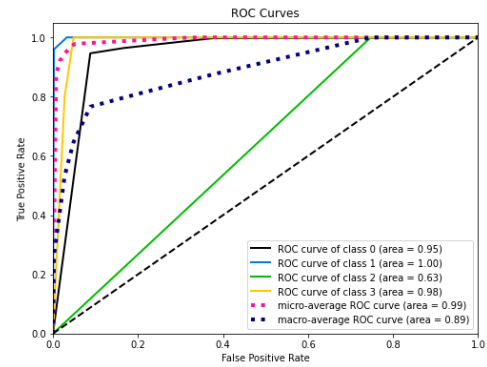
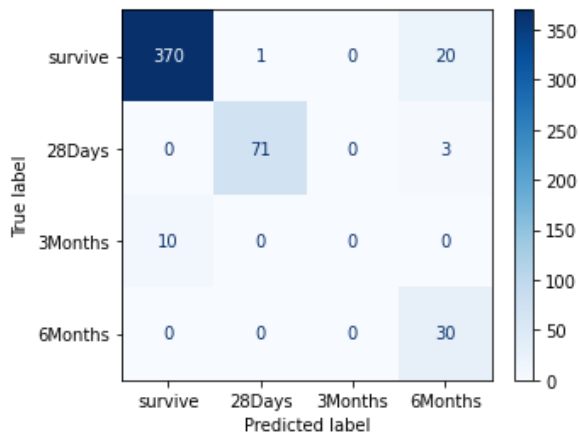
RF



LR



DT



الشكل (5-7) نتائج التدريب بإعادة التشكيل باستخدام SMOTE قبل الأمثلة

	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.79	1.00	0.88	391	0	0.84	0.98	0.90	391
1	1.00	0.12	0.22	74	1	0.80	0.50	0.62	74
2	0.00	0.00	0.00	10	2	0.00	0.00	0.00	10
3	0.00	0.00	0.00	30	3	1.00	0.13	0.24	30
accuracy			0.79	505	accuracy			0.84	505
macro avg	0.45	0.28	0.27	505	macro avg	0.66	0.40	0.44	505
weighted avg	0.76	0.79	0.71	505	weighted avg	0.83	0.84	0.80	505

RF

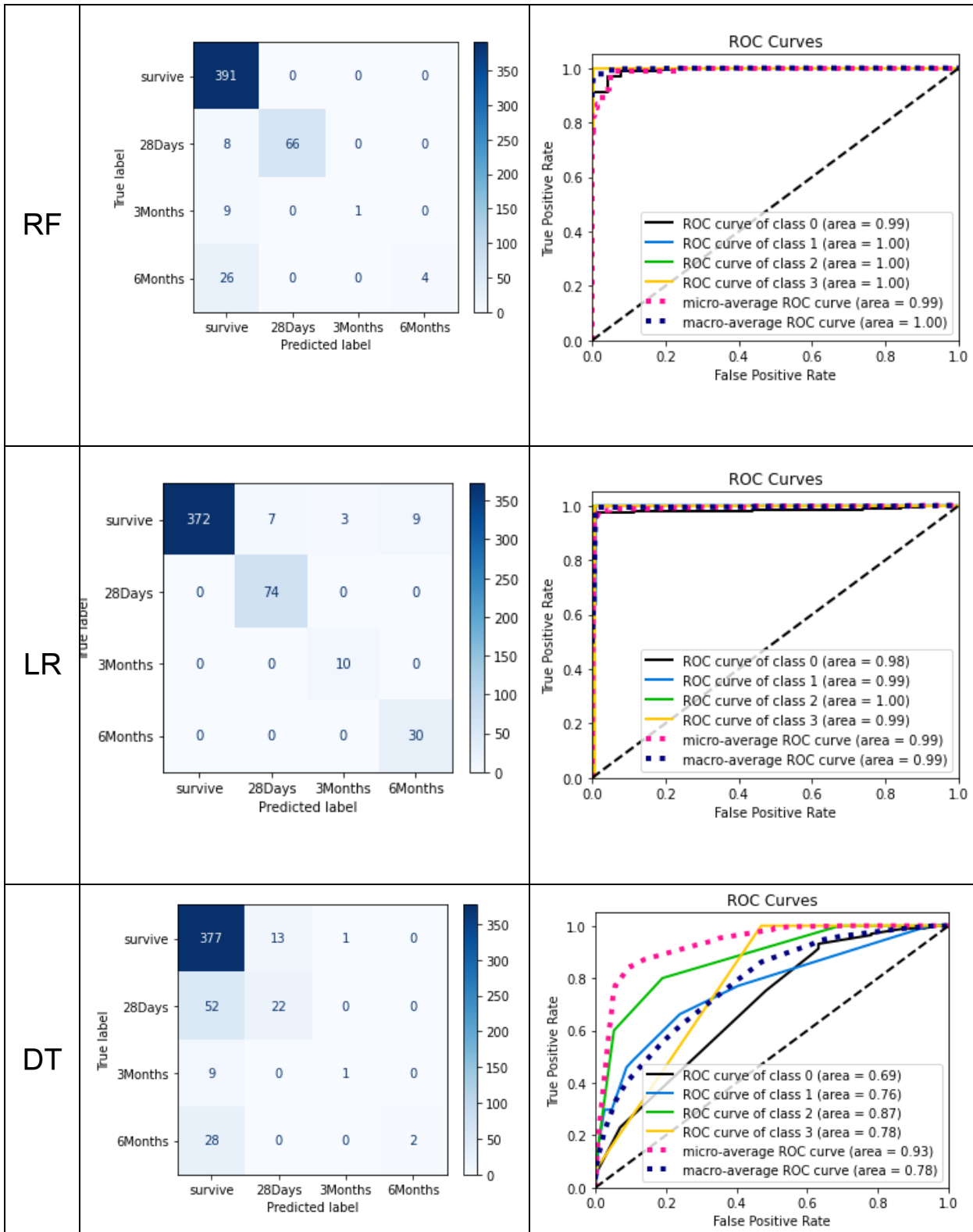
LR

	precision	recall	f1-score	support
0	0.97	0.95	0.96	391
1	0.99	0.96	0.97	74
2	0.00	0.00	0.00	10
3	0.57	1.00	0.72	30
accuracy			0.93	505
macro avg	0.63	0.73	0.66	505
weighted avg	0.93	0.93	0.93	505

DT

Model/Metrics	Accuracy	Precision	Recall	F1_score
RF	0.7920	0.7568	0.7920	0.7143
LR	0.8376	0.8273	0.8376	0.8035
DT	0.9326	0.9320	0.9326	0.9285

الجدول (5-10) نتائج التدريب بإعادة التشكيل باستخدام SMOTE قبل الأمثلة



الشكل (5-8) نتائج التدريب بإعادة التشكيل باستخدام SMOTE بعد الأمثلة

	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.90	1.00	0.95	391	0	1.00	0.95	0.98	391
1	1.00	0.89	0.94	74	1	0.91	1.00	0.95	74
2	1.00	0.10	0.18	10	2	0.77	1.00	0.87	10
3	1.00	0.13	0.24	30	3	0.77	1.00	0.87	30
accuracy			0.91	505	accuracy			0.96	505
macro avg	0.98	0.53	0.58	505	macro avg	0.86	0.99	0.92	505
weighted avg	0.92	0.91	0.89	505	weighted avg	0.97	0.96	0.96	505

RF **LR**

	precision	recall	f1-score	support
0	0.81	0.96	0.88	391
1	0.63	0.30	0.40	74
2	0.50	0.10	0.17	10
3	1.00	0.07	0.12	30
accuracy			0.80	505
macro avg	0.73	0.36	0.39	505
weighted avg	0.79	0.80	0.75	505

DT

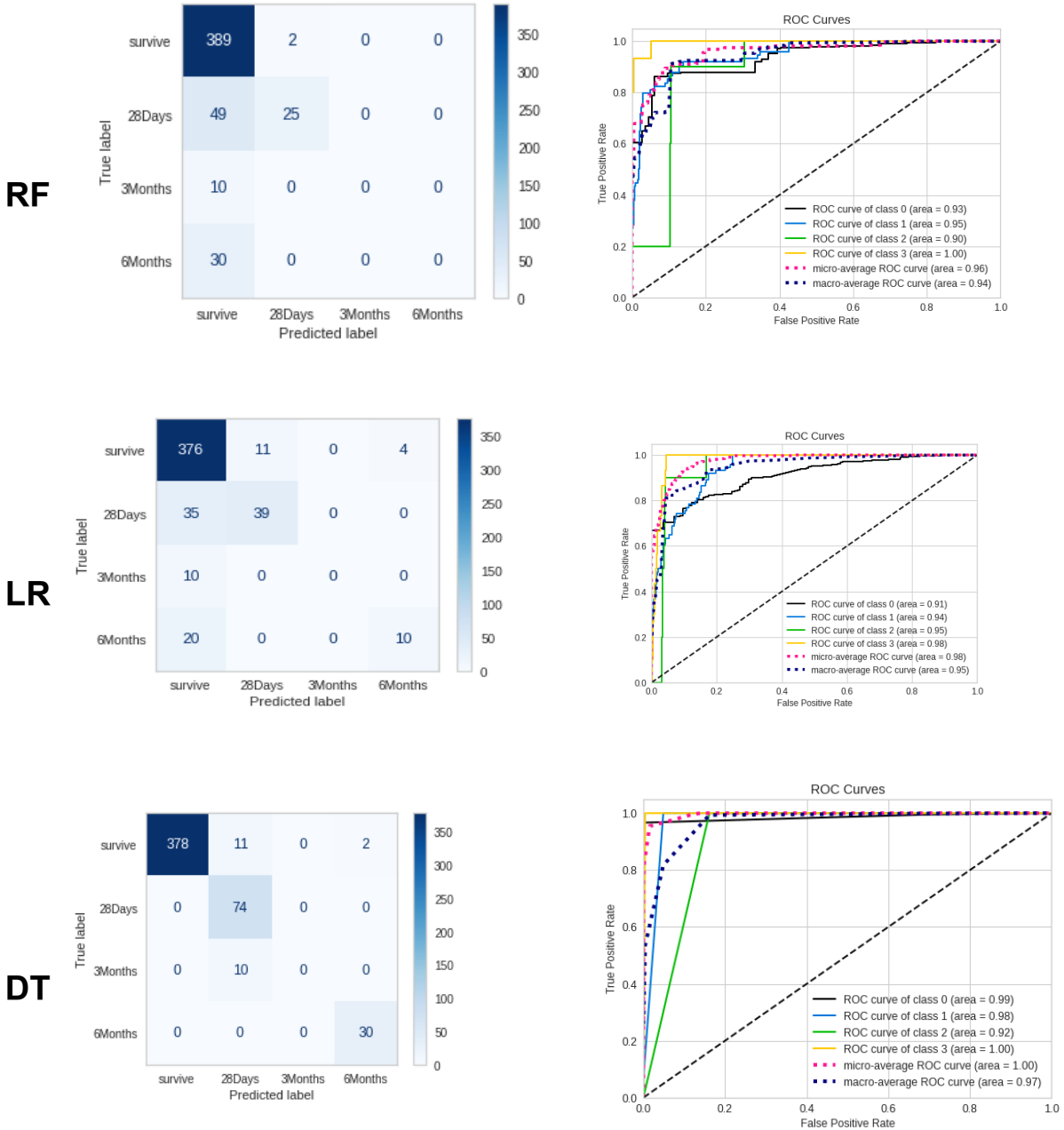
7.2.5. إعادة تشكيل مجموعة البيانات بالاعتماد على السمات الأكثر ارتباطاً

بالهدف

سنقوم في هذا السيناريو بإعادة تشكيل البيانات، عبر زيادة عدد العينات الأقلية لكن مع الأخذ بعين الاعتبار مدى الارتباط بين السمات و بين الهدف، من خلال حساب معامل الارتباط الخطي، ومن ثم زيادة أصناف الأقلية الأكثر ارتباطاً مع الهدف والتي تحقق معياراً معيناً، حيث تم اعتماد معيار المتوسط الحسابي (mean)، و بالتالي سيتم زيادة العينات الأقلية التي تحقق قيمها عند السمات الأكثر ارتباطاً بالهدف أنها أكبر من المتوسط، و بالمقابل سيتم إنقاص عدد عينات الصنف الأكثرية التي تحقق الشرط المعاكس، أي الأسطر التي تحقق

قيمتها أصغر من المتوسط لسمات الأكثر ارتباطاً، ثم يتم بعد ذلك تطبيق الموازنة عبر استخدام (SMOTE) للوصول إلى نفس حجم مجموعة البيانات في السيناريو السابق.

النتائج قبل الأمثلة:



الشكل (5-9) نتائج إعادة التشكيل باستخدام MUDFSCS قبل الأمثلة

	precision	recall	f1-score	support		precision	recall	f1-score	support
0	0.81	0.99	0.90	391	0	0.85	0.96	0.90	391
1	0.93	0.34	0.50	74	1	0.78	0.53	0.63	74
2	0.00	0.00	0.00	10	2	0.00	0.00	0.00	10
3	0.00	0.00	0.00	30	3	0.71	0.33	0.45	30
accuracy			0.82	505	accuracy			0.84	505
macro avg	0.43	0.33	0.35	505	macro avg	0.59	0.46	0.50	505
weighted avg	0.77	0.82	0.77	505	weighted avg	0.82	0.84	0.82	505

RF

LR

	precision	recall	f1-score	support
0	1.00	0.97	0.98	391
1	0.78	1.00	0.88	74
2	0.00	0.00	0.00	10
3	0.94	1.00	0.97	30
accuracy			0.95	505
macro avg	0.68	0.74	0.71	505
weighted avg	0.94	0.95	0.95	505

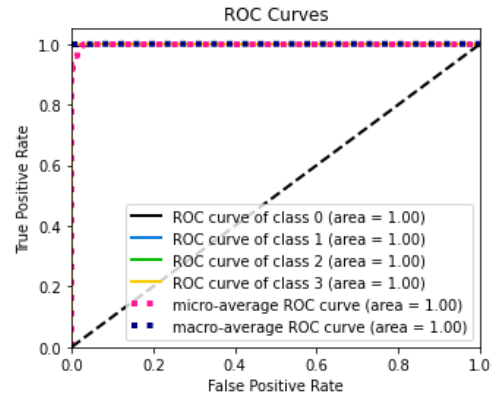
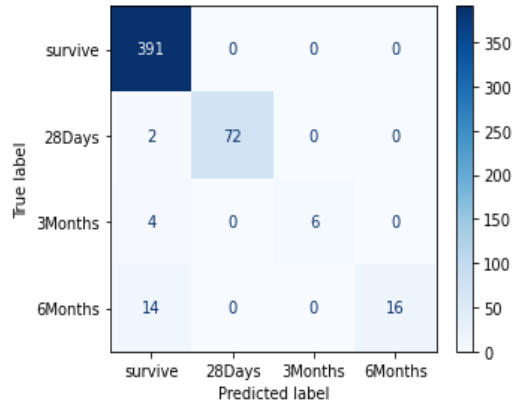
DT

Model/Metrics	Accuracy	Precision	Recall	F1_score
RF	0.8189	0.7657	0.8189	0.7657
LR	0.8415	0.8186	0.8415	0.8189
DT	0.9544	0.9440	0.9544	0.9469

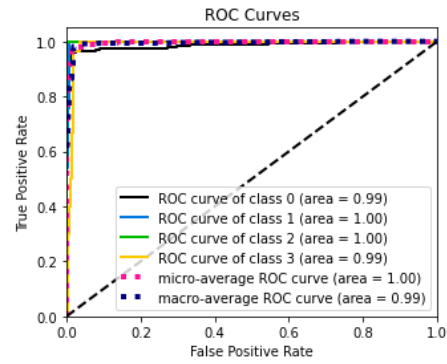
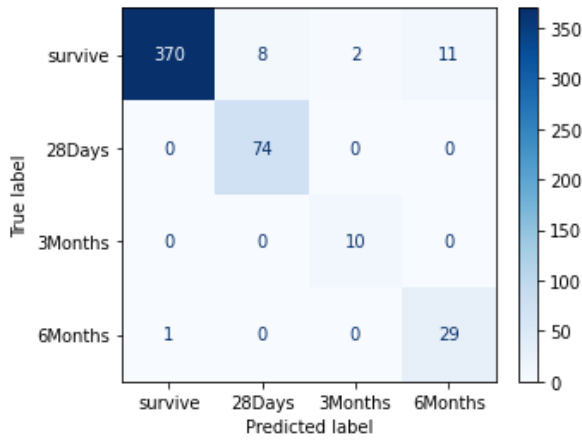
الجدول (11-5)) نتائج إعادة التشكيل باستخدام MUDFSCS قبل الأمثلة

النتائج بعد الأمثلة:

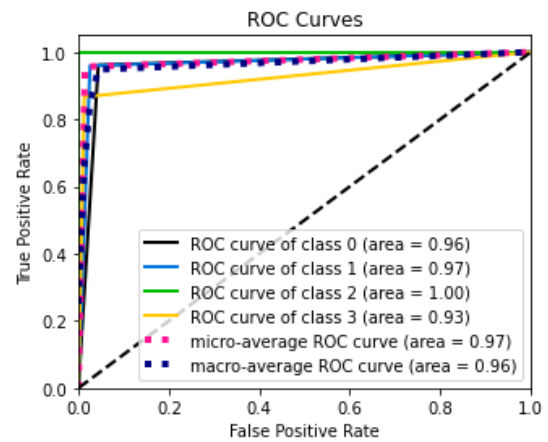
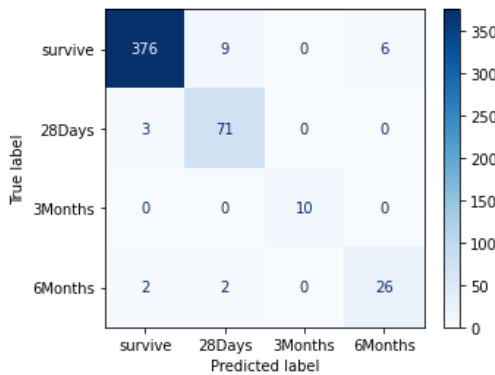
RF



LR



DT



الشكل (5-10) نتائج إعادة التشكيل باستخدام MUDFSCS بعد الأمثلة

8.2.5. التدريب مع الموازنة عبر إعادة تشكيل مجموعة البيانات باستخدام

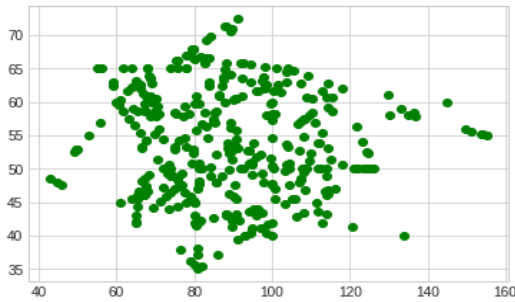
نموذج تقدير كثافة النواة

تم في البداية زيادة عدد العينات الأقلية الأكثر ارتباطاً بالهدف كما في السيناريو السابق، لكن بدلاً من استخدام تقنية SMOTE لموازنة مجموعة البيانات سنستخدم نموذج تقدير كثافة النواة (Kernel Density Estimation)، محاولين بذلك تجاوز أحد جوانب النقد الموجهة للموازنة باستخدام تقنية SMOTE والمتمثلة في أن العينات المولدة ستبدو وكأنها مرتبطة خطياً، وذلك بسبب الآلية التي يتم فيها توليد العينات.

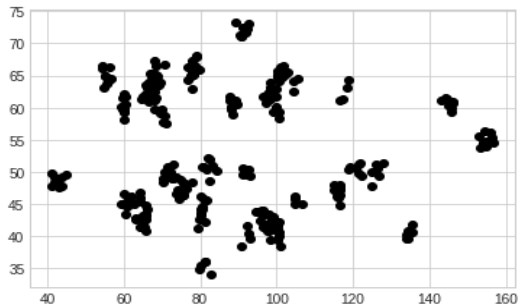
ما نقوم به هو تطبيق نموذج تقدير كثافة النواة لكل مجموعة من النقاط الأقلية، على سبيل المثال النقاط التي تمثل حالات الوفاة خلال 3 أشهر، ومن ثم سنحصل على تابع الكثافة الاحتمالية لمجموعة النقاط هذه، والنقاط المولدة ستكون تتبع لنفس تابع الكثافة الاحتمالية لمجموعة النقاط الأساسية.

الشكل التالي يوضح الفرق بين النقاط المولدة باستخدام SMOTE والنقاط المولدة باستخدام

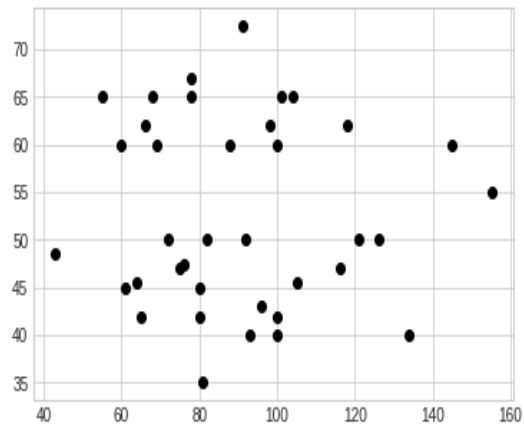
نموذج كثافة النواة



SMOTE samples



KDE samples

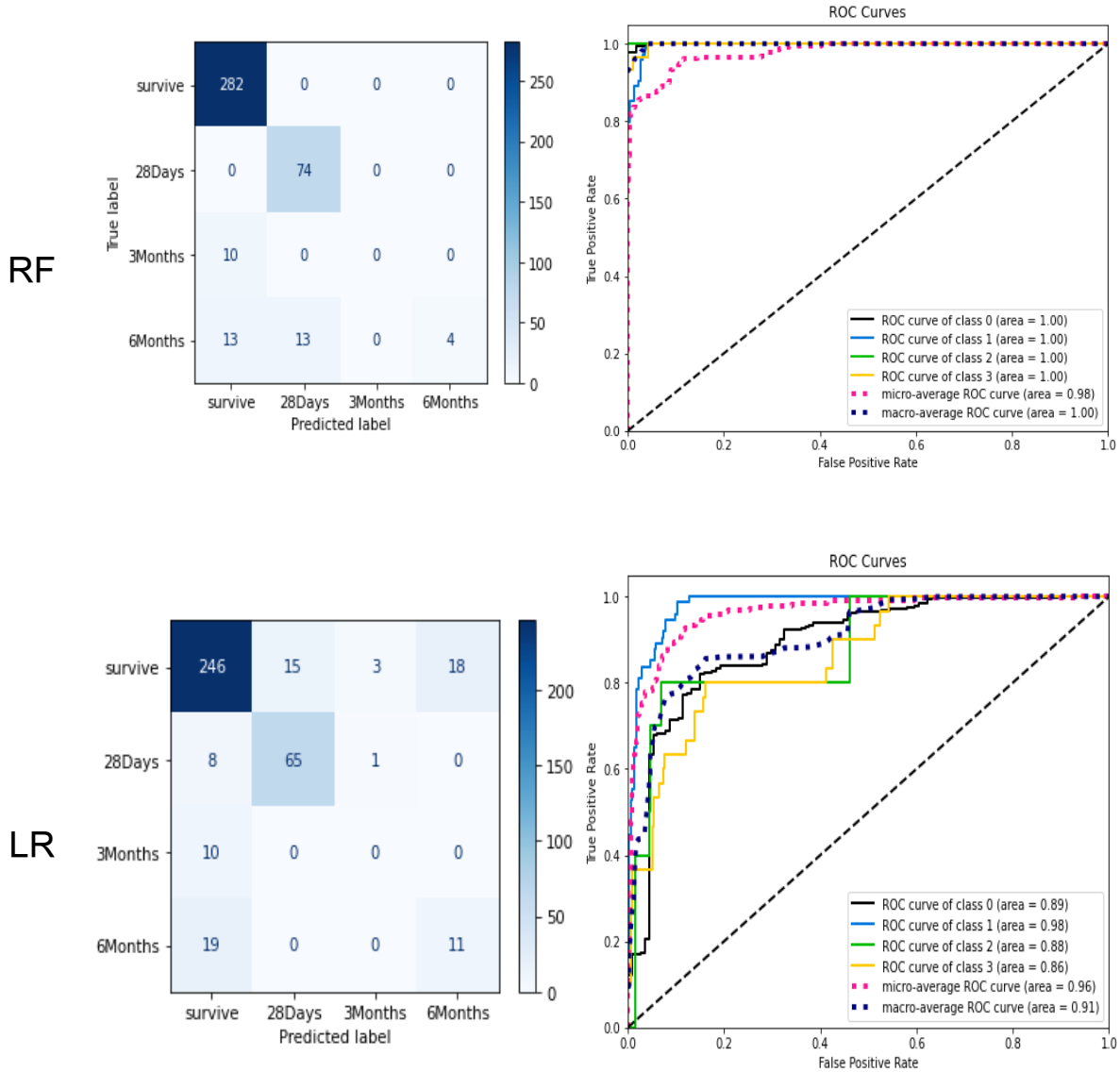


Original samples

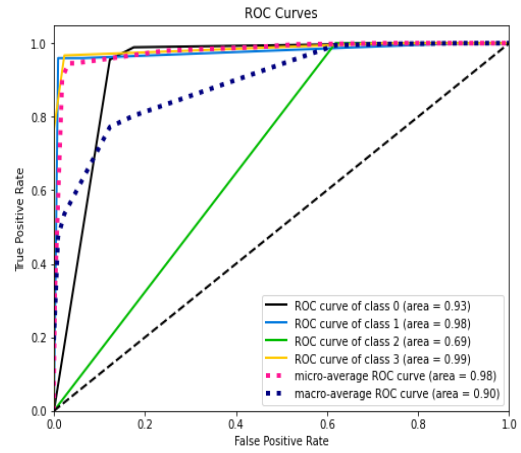
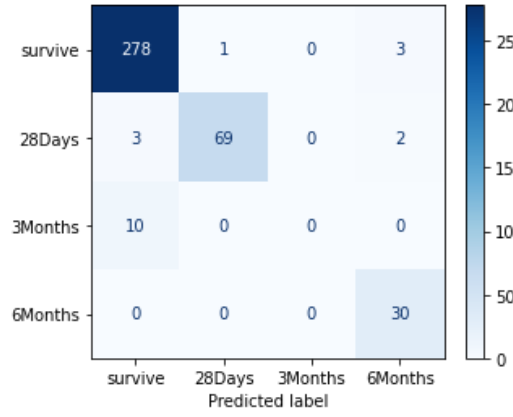
ثم سنقوم بخطوة تقليل التداخل من خلال استبعاد نقاط الأكثرية الأقرب لنقاط الأقلية.

بعد ذلك سنكمل كما في سيناريوهات التدريب السابقة من حيث تدريب نماذج تعلم الآلة بدون أمثلة ومع أمثلة للمعلمات الفائقة.

النتائج قبل الأمثلة



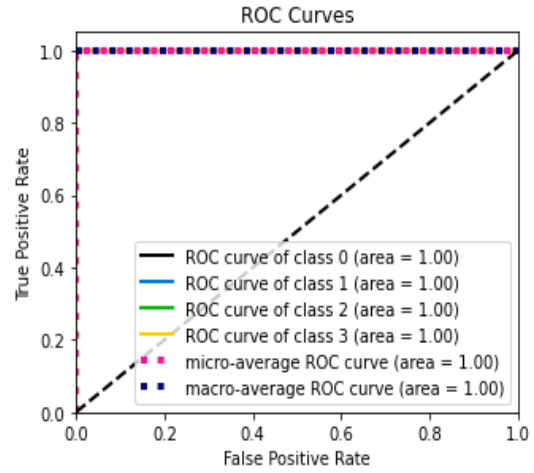
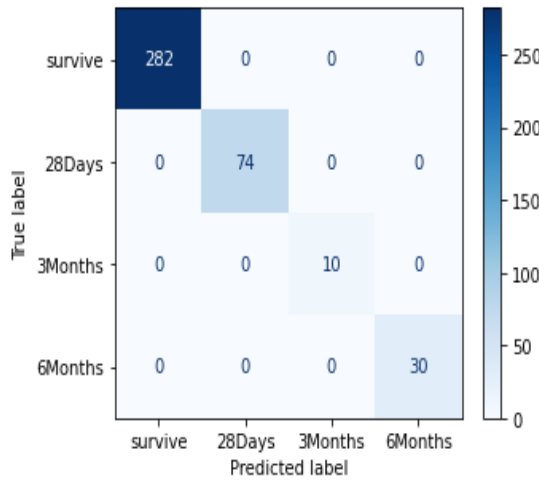
DT



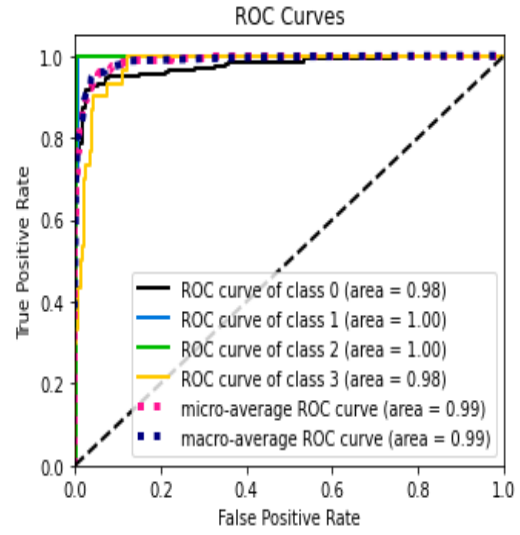
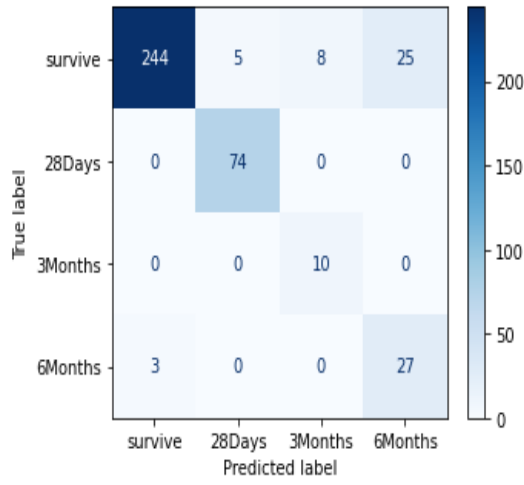
الشكل (5-11) نتائج إعادة التشكيل باستخدام KDE قبل الأمثلة

النتائج بعد الأمثلة

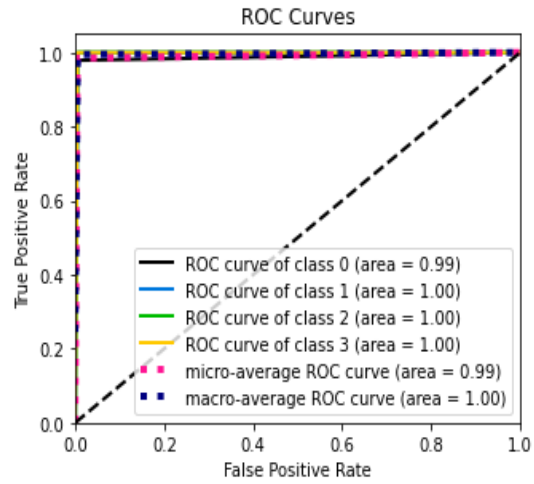
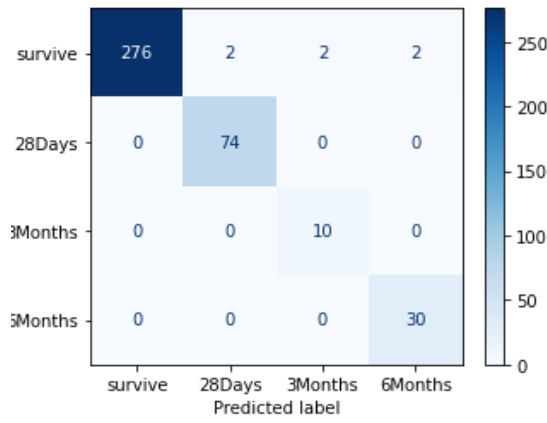
RF



LR



DT



الشكل (5-12) نتائج إعادة التشكيل باستخدام KDE بعد الأمثلة

9.2.5. تدريب نموذج التعلم العميق مع تحسين

تم تنفيذ عملية التحسين باستخدام البحث العشوائي حيث تم تطبيق عملية البحث 10 مرات متتالية وباعتماد دقة التحقق كمعيار لاختبار تحسين أداء البارامترات حيث يتم في كل تكرار تحديد عدد التكرارات الأفضل ومعدل التعلم الأفضل وخوارزمية التعلم ويتم تكرار ذلك 10 مرات ثم الاحتفاظ بأفضل نتيجة تدريب.

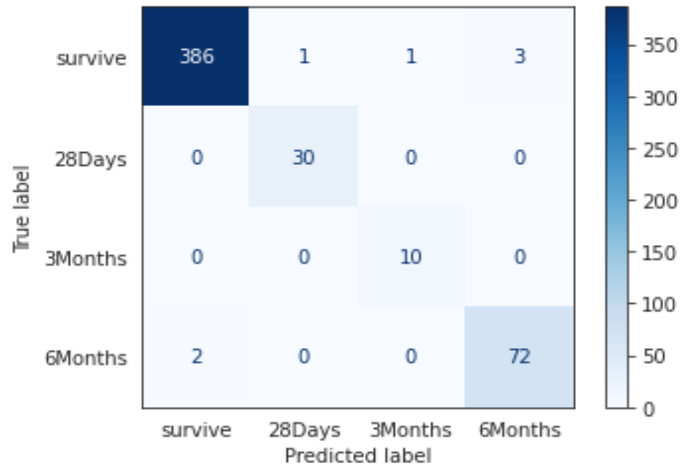
يوضح الشكل التالي معمارية نموذج التعلم العميق المحسن المقترح

```
Model: "sequential"
-----
Layer (type)                Output Shape                Param #
-----
dense (Dense)                (None, 128)                 17792
dense_1 (Dense)              (None, 128)                 16512
dense_2 (Dense)              (None, 128)                 16512
dense_3 (Dense)              (None, 128)                 16512
dense_4 (Dense)              (None, 4)                   516
-----
Total params: 67,844
Trainable params: 67,844
Non-trainable params: 0
```

يتضمن الشكل (5-13) مصفوفة التشتت لنتيجة اختبار نموذج التعلم العميق المحسن، ويبين الجدول (5-12) نتائج الاختبار التفصيلية للأصناف الأربعة والنتيجة النهائية لكل من بارامترات الأداء.

بمقارنة الجدولين 2 و3 يتضح أن عملية التدريب باستخدام التحسين قد حسنت الأداء بشكل طفيف وبمقدار 0.98% لدقة التحقق في حين بقيت دقة الاختبار كما هي 98.62% في كلتا

الحالتين، مما يمكن ملاحظته أن توزع الأخطاء يختلف بشكل طفيف بين الحالتين مع تحسن بسيط.



الشكل (5-13) مصفوفة التشتت لنموذج التعلم العميق بعد تطبيق خوارزمية التحسين Random Search

Type of calculation	Validation Accuracy %	Test Accuracy %	Precision %	Recall %	F1_score %
Class 0			99	99	99
Class 1			97	100	98
Class 2			91	100	95
Class 3			96	97	97
Real Values	99	98.61	98.64	98.61	98.62
Macro Average	99	98.61	96	99	97
Weighted Average	99	99	99	99	99

الجدول (5-12) نتائج اختبار نموذج التعلم العميق المدرب بعد تطبيق عملية التحسين Optimization

يبين الجدول (5-12) الخاص بالنموذج المدرب وفقاً لعملية التحسين أنّ أخطاء الأصناف الفردية قد قلت بالمقارنة مع حالة تدريب النموذج بدون تحسين.

10.2.5. نموذج التعلم العميق مع اختيار السمات

(هذا السيناريو سيتم تنفيذه بعدة سيناريوهات فرعية لاختيار السمات).

كون أنّ نموذج التعلم العميق أعطى نتائجاً أفضل من نماذج تعلم الآلة سيتم تدريب هذا النموذج بعد تنفيذ عملية اختيار السمات على مجموعة البيانات وسيتم تنفيذ عملية الاختيار وفق الآتي:

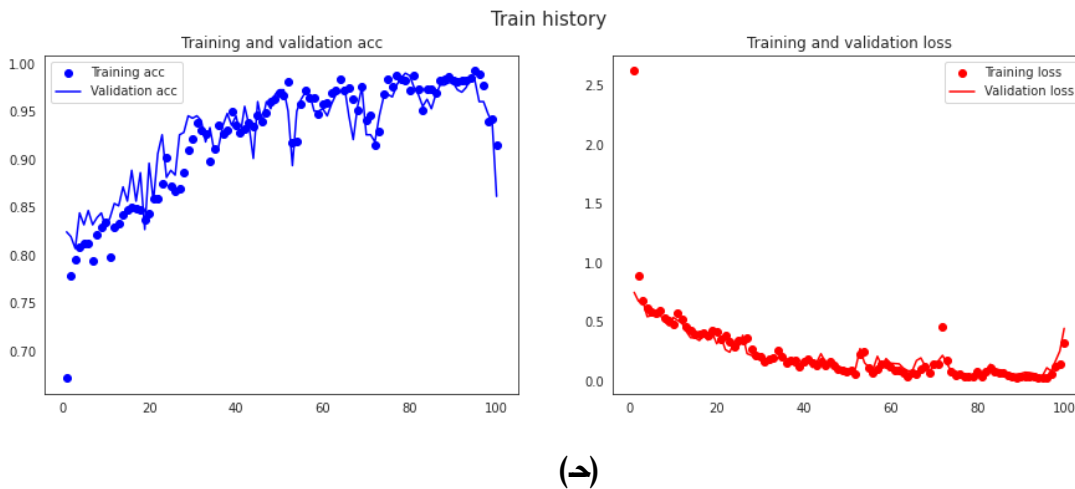
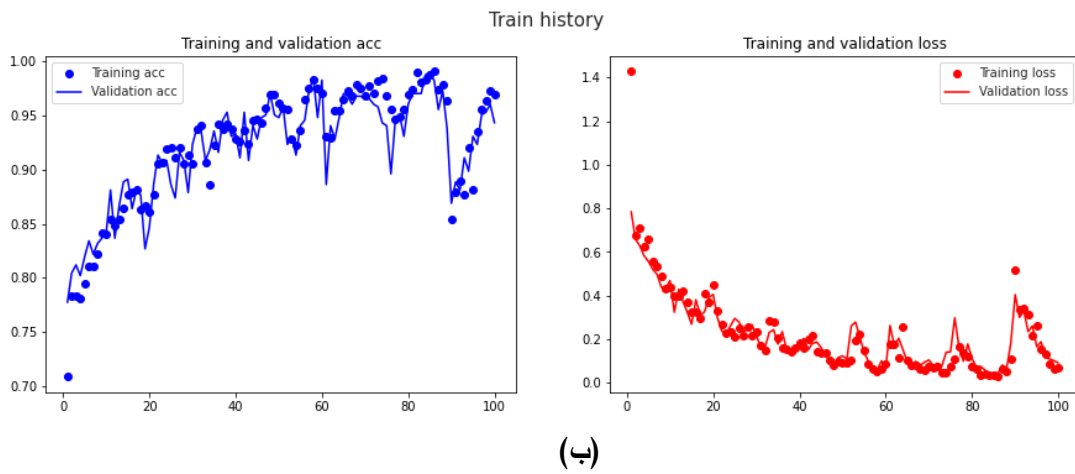
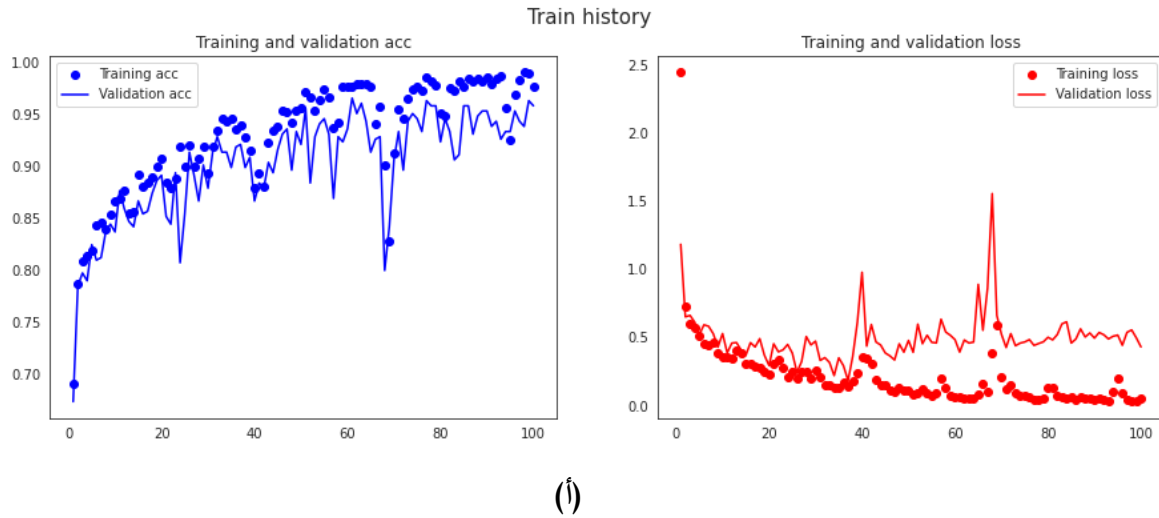
1. يتم حساب الارتباط Correlation بين جميع أعمدة مجموعة البيانات وبين عمود الهدف death.

2. يتم تطبيق عتبة اختيار تبدأ من 0.05 ثم 0.08 ثم 0.02.

3. يتم تدريب نموذج التعلم العميق بالسمات المختارة وفق الخطوة رقم 2.

تم بدايةً العمل عند عتبة 0.05، حيث تم الحصول بنتيجة تطبيق هذه العتبة على 31 سمة (خاصية) من أصل 139، ثم تم بعد ذلك تدريب نموذج التعلم العميق باستخدام هذه السمات وبنفس درجة التقسيم 20% من مجموعة التدريب للتحقق وبنفس بنية نموذج التعلم العميق ونفس بارامترات التدريب تماماً. ثم تمت إعادة نفس التجربة لكن مع استخدام عتبة 0.02 ثم 0.08. عند عتبة 0.02 حصلنا على (55) سمة فقط وعند عتبة 0.08 حصلنا على (16) سمة فقط.

يبين الشكل (5-14) ناتج عملية التدريب حيث يظهر منحنى الدقة مع التكرارات ومنحنى الخسارة مع التكرارات وبمقارنة بسيطة مع النموذج الأصلي يظهر أنّ الأداء قد انخفض قليلاً.



الشكل (5-14) منحنيات الدقة والخسارة مع عدد التكرارات للنموذج المدرب على السمات المختارة عند: (أ) عتبة 0.02، (ب) عتبة

0.05، (ج) عتبة 0.08

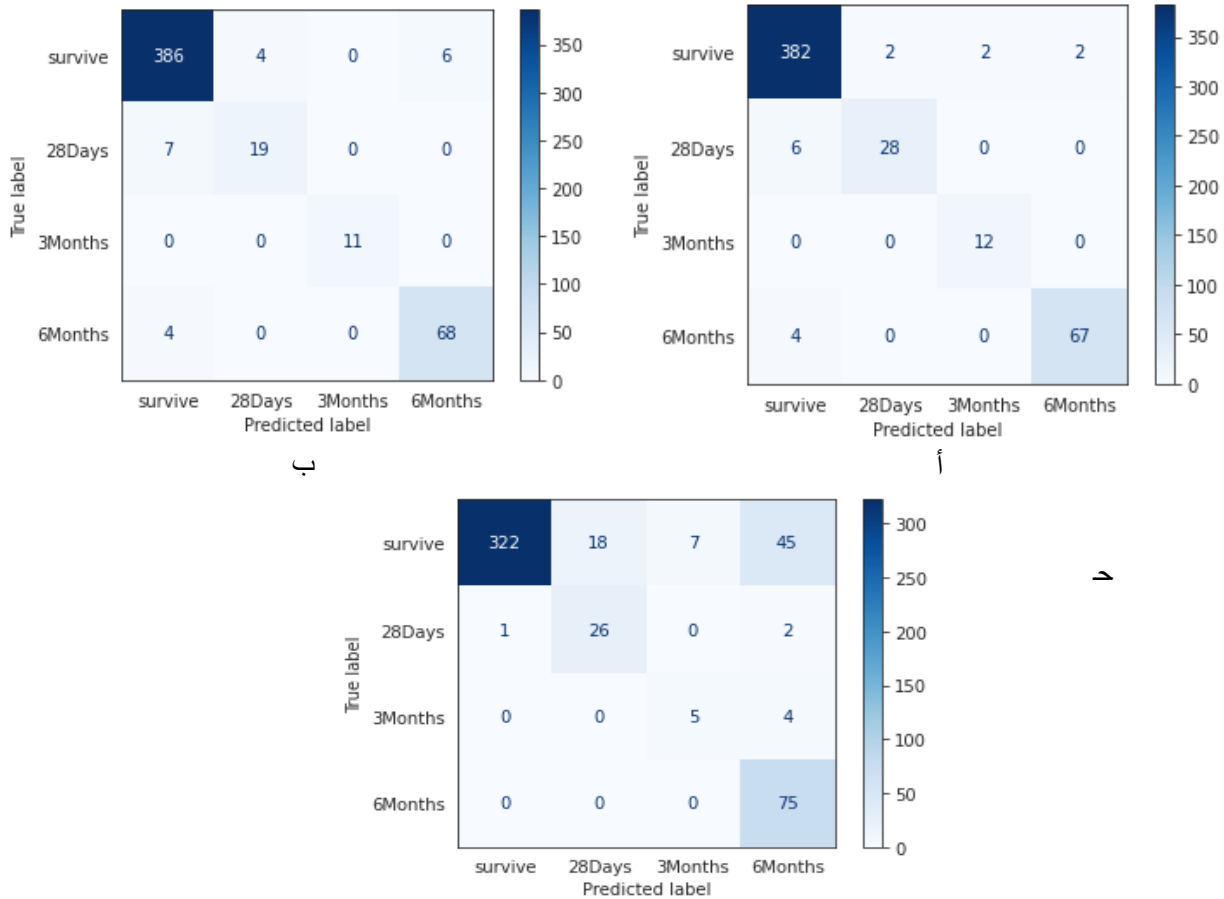
أما بالنسبة لنتيجة اختبار مجموعة الاختبار على النموذج المدرب، فيبين الجدول (5-13) مقارنة في الأداء بين النموذج في حال استخدام عتبة 0.02 ثم 0.05 ثم 0.08. في حين يوضح الشكل (5-15) مصفوفات التشتت لنموذج التعلم العميق عند عدة عتبات لاختيار السمات.

Model/Metrics	NoF	Accuracy %	Precision %	Recall %	F1_score %	Training time
Original Dense model	138	98.6	98.72	98.61	98.63	4.55 ms/step
With FS/ Th=0.02	55	96.8	96.84	96.83	96.79	3.72 ms/step
With FS/ Th=0.05	31	95.8	95.77	95.84	95.79	3.71 ms/step
With FS/ Th=0.08	16	84.7	90.35	84.75	85.9	3.69 ms/step

الجدول (5-13) مقارنة أداء نظام التعلم العميق المقترح في ظل عدة عتبات لاختيار السمات

NoF: number of features, Th: correlation threshold, FS: feature selected.

يبين الجدول (5-13) أنَّ اختيار 55 سمة من أصل 138 سمة بمعدل اختيار 39.8% يخفض دقة الاختبار بمقدار 1.8% فقط في حين يوفر زمن تدريب مقداره 0.83 ميلي ثانية في الخطوة الواحدة من التدريب. إن خفض معدل اختيار السمات إلى 31 سمة فقط (أي ما يعادل 22.4%) يخفض الأداء بمقدار 2.8% ولدى تخفيضه إلى 16% فقط ستخفض دقة الاختبار بمقدار 13.9%.



الشكل (5-15) مصفوفة التشتت عند عتبة اختيار سمات: (أ) 0.02، (ب) 0.05، (ج) 0.08

يبين الجدول (5-13) أنّ اختيار السمات بعتبة 0.02 أدى لانخفاض طفيف في الدقة، في حين أن رفع العتبة لمستوى 0.05 قلل الدقة بمقدار 2.8% وهو أيضاً قيمة قليلة، لكن مع تخفيض السمات المختارة بعتبة 0.08 تنخفض الدقة بمقدار 13.9% مما يعني أن بعض السمات (الأعمدة) المهمة من مجموعة البيانات قد تم حذفها.

يوصي البحث الحالي باستخدام عتبة اختيار سمات 0.02 لأنها تحافظ على أداء قريب جداً من الأداء الأصلي وتقلل زمن التدريب بمقدار 0.83 ميلي ثانية كل خطوة تدريب.

من حيث معدل القبول الصحيح Recall كان تحديد درجة الخطورة (الموت خلال 28 يوم والموت خلال 3 أشهر) هي الأدق في التخمين يليهما حالة عدم وجود خطورة بمعدل 99%

أما درجة الخطورة (الموت بعد 6 أشهر) فكانت الأقل دقة من حيث التخمين بمعدل قبول صحيح 96%.

لكن بشكل عام ومن حيث كل قياسات الدقة كان تحديد درجة الخطورة لكل الأصناف (الحالات) دقيقاً حيث بلغت الدقة الوسطية للأصناف بين 96% و 99%.

11.2.5. إعادة التدريب بعد تطبيق خوارزمية الموازنة MUDSFC

في هذا السيناريو سنقوم بإعادة تنفيذ كل سيناريوهات التدريب والاختبار السابقة لنرى مدى فعالية خوارزمية MUDSFC المصممة من قبلنا وأي الحالات التي سنوصي فيها باستخدام الخوارزمية.

- الهدف هو الوصول لمجموعة بيانات متوازنة بنفس حجم مجموعة البيانات التي وزاها بطريقة SMOTE.
- تم بداية حساب قيم ارتباط السمات مع العمود الهدف، باستخدام تابع corr الذي توفره مكتبة pandas، الذي يقوم بحساب قيم الارتباط بين الأعمدة، بالاعتماد على معامل ارتباط بيرسون، كونه الطريقة الأكثر شيوعاً لقياس الارتباط الخطي حيث يمثل الارتباط بين كل سمة والسمة الهدف برقم من المجال $[-1, 1]$ ، يقيس قوة واتجاه العلاقة بين متغيرين.
- تم اختيار أكثر 4 سمات مرتبطة بالهدف، لكل منها معامل ارتباط أكبر من 0.3.
- يتم في هذه المرحلة زيادة عدد عينات الأقلية في مجموعة بيانات التدريب، بتقنية مشابهة لما تقوم به SMOTE، حيث نقوم بتوليف نقاط أقلية جديدة.

B: مجموعة النقاط الأقلية المرشحة لتوليف نقاط أقلية جديدة، $B = []$.

F: مصفوفة السمات الأكثر ارتباطاً بالهدف، $F = [F1, F2, F3, F4]$.

M: مجموعة نقاط الأقلية في مجموعة البيانات:

من أجل كل نقطة $m \in M$:

من أجل كل سمة $f \in F$:

إذا تحقق الشرط: $M(F_i) > F_i.mean$

أضف m إلى B .

حيث: $M(F_i)$ ، قيمة نقطة الأقلية m في السمة F_i .

$F_i.mean$: المتوسط الحسابي لقيم السمة F_i .

- يتم عكس الشروط السابقة، بالنسبة لنقاط الأكثرية، أي النقاط الأكثرية التي تحقق قيمها أصغر من المتوسط الحسابي للسمة الأكثر ارتباطاً بالهدف ونقوم بحذفها.
- يتم توليد نقاط جديدة باستخدام نموذج كثافة النواة (Kernel Density) الذي توفره مكتبة Sklearn، يتم بداية عزل السمات ذات الطبيعة الفئوية، أي التي تكون قيمها قيم صحيحة، مثل الجنس الذي يأخذ إحدى قيمتين $[0,1]$ ، قبل تطبيق النموذج، و من ثم يتم تعويضه حسب واصف المنوال أو القيمة الأكثر تكراراً، و يأخذ نموذج كثافة النواة معامل ال Bandwidth الذي يمثل عرض نطاق النواة، حيث يتم تحديد أفضل قيمة لهذا المعامل باستخدام GridSearchCV، المعامل الثاني لنموذج كثافة النواة هو Kernel، والقيمة الافتراضية له هي gaussian بعد أن يتم إيجاد أفضل قيم لمعامل عرض النطاق، يتم استخدام التابع Sample الذي يقوم بدوره بإرجاع النقاط المؤلفة الجديدة.

يوضح الجدول (5-14) مقارنة بين نتائج مصنفات تعلم الآلة ومصنف التعلم العميق قبل وبعد عملية التحسين وباستخدام طريقة SMOTE من ثم باستخدام الخوارزمية MUDSFC المصممة من قبلنا.

Model/Metrics	Accuracy %	Precision %	Recall %	F1_score %
SMOTE optimized RF model	89.1	84.39	89.1	86.89
MUDSFC optimized RF model	88	90.6	88.96	89.39
SMOTE optimized DT model	85.55	83.82	85.54	84.22
MUDSFC optimized DT	82.17	79.69	82.17	80.9
SMOTE optimized LR model	92.67	94.89	92.67	93.16
MUDSFC optimized LR model	95	95.74	95.05	95.1
SMOTE optimized Dense model	98.6	98.72	98.61	98.63
MUDSFC optimized Dense model	98.68	98.73	98.68	98.68

الجدول (5-14) مقارنة بين نتائج مصنفات تعلم الآلة ومصنف التعلم العميق قبل وبعد عملية التحسين وباستخدام طريقة *over sampling*

Balancing ومن ثم باستخدام الخوارزمية *MUDSFC* المصممة من قبلنا.

الجدول (5-14) يوضح أن استخدام خوارزمية *MUDSFC* لموازنة مجموعة البيانات إلى تحسين الأداء في نظام RF المحسن حيث ارتفع معامل Precision بمقدار 6.21% إلا أن مصنف DT انخفض أداؤه والسبب هو أن مصنف DT فشل في تصنيف عينات الصنفين 1 و2 وخوارزمية *MUDSFC* عندما طبقت قامت بزيادة عدد عينات هذين الصنفين لذلك زادت الأخطاء والمشكلة هنا تكمن في المصنف لذا لا نوصي باستخدامه.

أما التحسين الأكبر فكان من نصيب نموذج LR المحسن وكون أن هذا النموذج بالأصل كان الأفضل بين نماذج تعلم الآلة المحسنة لدى استخدام طريقة الموازنة *Oversampling* الأصلية، فإنه لدى استخدام طريقتنا *MUDSFC* فقد تحسنت الدقة بمقدار 2.33% وتحسنت

معاملات Precision و Recall و F1-score بمقدار 0.85% و 2.38% و 1.94% على التوالي.

أما من ناحية استخدام نموذج التعلم العميق المصمم من قبلنا مع خوارزمية MUDSFC المقترحة من قبلنا فقد تحسنت الدقة بمقدار 0.08% وتحسن معامل Recall بمقدار 0.01% ومعامل Precision بمقدار 0.07% ومعامل F1-score بمقدار 0.05% عن الحالة السابقة.

إن استخدام الخوارزمية MUDSFC لموازنة مجموعة البيانات حسن الأداء في المصنفات الجيدة بشكل واضح وذلك نظراً لأن زيادة العينات وفق هذه الطريقة كانت من طرف العينات الأكثر ترابطاً مع عمود الهدف Death وإنقاص العينات كانت من طرف العينات الأقل ترابطاً مع عمود الهدف (عملية موازنة ذكية).

12.2.5. إعادة تدريب واختبار نماذج تعلم الآلة النسخة المحسنة من خوارزمية الموازنة MUDSFC بإضافة مرحلة كشف العينات المتداخلة بين الأصناف

في هذا السيناريو سيتم إعادة تدريب مصنفات تعلم الآلة كونها حققت أداء أقل من نموذج التعلم العميق، وسنسعى في هذا السيناريو للوصول لأفضل دقة تشخيص من خلال تحسين خوارزمية موازنة السمات المقترحة حيث تم إضافة مرحلة هي عملية العنقدة Clustering التي تهدف لكشف نقاط البيانات المتداخلة بين الصنف المسيطر (0) والأصناف الثانوية غير المسيطرة (1،2،3). ثم يتم تقليل نقاط بيانات الصنف المسيطر بحذف نقاط البيانات المتداخلة.

تقوم هذه المرحلة بإنقاص نقاط الأكثرية الأكثر قرباً من نقاط الأقلية، والتي من شأنها أن تكون سبباً في أخطاء التصنيف، حيث يتم تطبيق خوارزمية DBSCAN، وهي من طرق التعلم غير الخاضعة للإشراف، والتي تحدد المجموعات المميزة في البيانات، تعمل

الخوارزمية عن طريق الانتقاء التعسفي لنقطة في مجموعة البيانات، حتى يتم زيارة جميع النقاط، إذا كان هناك ما لا يقل عن (minPoint) نقطة داخل دائرة نصف قطرها (epsilon) نعتبر هذه النقاط من نفس المجموعة، يتم تقدير المعامل epsilon عبر تطبيق خوارزمية الجار الأقرب وحساب المسافة الوسطية بين كل نقطة و أقرب K جار لها، و ثم ترتيب هذه المسافات تصاعدياً ثم رسمها لتحديد نقطة المنعطف.

في حالتنا اخترنا $\epsilon=100$, $\text{minPoints}=15$ ، وقد نتج عن عملية العنقدة، 4 عناقيد، من أجل كل عنقود نميز إحدى حالتين إما أن العنقود هو فقط من نقاط الأكثرية، أو أن العنقود هو مزيج من نقاط الأقلية والأكثرية، بالنسبة للحالة الأولى، وضوحاً لا يوجد تداخل بين العينات الأقلية والأكثرية، بالنسبة للحالة الثانية،

- نقوم بتحديد النقاط الأقلية في العنقود
- نقوم بحساب مصفوفة المسافات بين نقاط العنقود الأقلية والأكثرية (الأسطر نقاط الأقلية، الأعمدة نقاط الأكثرية)
- نقوم بالتحويل من مصفوفة المسافات إلى مصفوفة من عمودين العمود الأول يمثل نقاط الأقلية، والعمود الثاني هو أصغر قيمة في السطر الخاص به في مصفوفة المسافات، بهذا نكون حددنا نقاط الأكثرية الأقرب إلى نقاط الأقلية.
- نقوم بأخذ العناصر الفريدة من العمود الثاني، حيث هناك حالات تكون فيها نقطة الأكثرية مكررة، بمعنى أنها الأقرب لأكثر من نقطة أقلية.
- أخيراً نحذف النقاط الأكثرية التي حددناها في الخطوة السابقة.

13.2.5. التدريب مع الموازنة بالطريقة المقترحة من قبلنا

تم بداية حساب قيم ارتباط السمات مع العمود الهدف، باستخدام تابع corr الذي توفره مكتبة pandas، الذي يقوم بحساب قيم الارتباط بين الأعمدة، بالاعتماد على معامل ارتباط بيرسون، كونه الطريقة الأكثر شيوعاً لقياس الارتباط الخطي حيث يمثل الارتباط بين كل سمة والسمة الهدف برقم من المجال $[-1,1]$ ، يقيس قوة واتجاه العلاقة بين متغيرين.

تم اختيار أكثر 4 سمات مرتبطة بالهدف، لكل منها معامل ارتباط أكبر من 0.3.

المرحلة الأولى Hybrid Oversampling UnderSampling:

يتم في هذه المرحلة زيادة عدد عينات الأقلية في مجموعة بيانات التدريب، بتقنية مشابهة لما تقوم به SMOTE، حيث نقوم بتوليف نقاط أقلية جديدة.

B: مجموعة النقاط الأقلية المرشحة لتوليف نقاط أقلية جديدة، $B = []$.

F: مصفوفة السمات الأكثر ارتباطاً بالهدف، $F = [F_1, F_2, F_3, F_4]$.

M: مجموعة نقاط الأقلية في بيانات التدريب، و هي بالشكل التالي:

$M = [1:\{45\}, 2:\{39\}, 3:\{52\}]$

من أجل كل نقطة $m \in M$:

من أجل كل سمة: $f \in F$

إذا تحقق الشرط: $M(F_i) > F_i.mean$

أضف m إلى B.

حيث: $M(F_i)$ ، قيمة نقطة الأقلية m في السمة F_i .

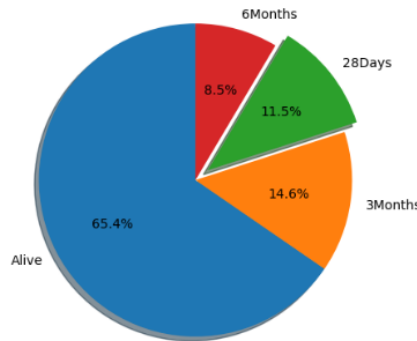
$F_i.mean$: المتوسط الحسابي لقيم السمة F_i .

يتم عكس الشروط السابقة، بالنسبة لنقاط الأكثرية، أي النقاط الأكثرية التي تحقق قيمها أكبر من المتوسط الحسابي للسمات الأكثر ارتباطاً بالهدف، حيث نحذف منها ما لا يزيد عن IR: imbalance Ratio، وهو النسبة بين فئة الأقلية إلى الأكثرية، وكمثال من حالتنا، إذا تم تحديد نقطة من الفئة الأقلية 1 والتي يبلغ عدد نقاطها 52 عينة، كنقطة مرشحة لتوليف نقاط أقلية جديدة، فإننا بالمقابل سنحذف من نقاط الأكثرية في مجموعة التدريب البالغ عددها 1365، بما لا يزيد عن 26 نقطة كشرط للتوقف.

يتم توليد نقاط أقلية جديدة باستخدام نموذج كثافة النواة (Kernel Density) الذي توفره مكتبة Sklearn، حيث يتم بداية عزل السمات ذات الطبيعة الفئوية، أي التي تكون قيمها قيم صحيحة، مثل الجنس الذي يأخذ إحدى قيمتين [0,1]، قبل تطبيق النموذج، ومن ثم يتم تعويضه حسب واصف المنوال أو القيمة الأكثر تكراراً،

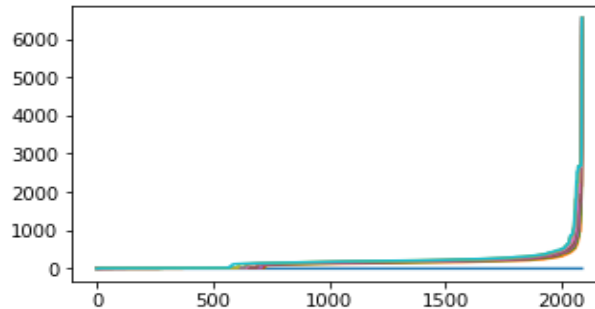
حيث يأخذ نموذج كثافة النواة معامل ال Bandwidth الذي يمثل عرض نطاق النواة، حيث يتم تحديد أفضل قيمة لهذا المعامل باستخدام GridSearchCV، المعامل الثاني لنموذج كثافة النواة هو Kernel، والقيمة الافتراضية له هي gaussian. بعد أن يتم إيجاد أفضل قيم لمعامل عرض النطاق، يتم استخدام التابع Sample الذي يقوم بدوره بإرجاع النقاط المؤلفة الجديدة.

بنهاية هذه الخطوة أصبحت مجموعة التدريب على الشكل التالي:



المرحلة الثانية UnderSample overlapped Majority Points:

الهدف من هذه المرحلة إنقاص نقاط الأكثرية الأكثر قريباً من نقاط الأقلية، و التي من شأنها أن تكون سبباً في أخطاء التصنيف، حيث يتم تطبيق خوارزمية DBSCAN ، وهي من طرق التعلم غير الخاضعة للإشراف، والتي تحدد المجموعات المميزة في البيانات، تعمل الخوارزمية عن طريق الانتقاء التعسفي لنقطة في مجموعة البيانات، حتى يتم زيارة جميع النقاط، إذا كان هناك ما لا يقل عن (minPoint) نقطة داخل دائرة نصف قطرها (epsilon) نعتبر هذه النقاط من نفس المجموعة، يتم تقدير المعامل epsilon عبر تطبيق خوارزمية الجار الأقرب وحساب المسافة الوسطية بين كل نقطة و أقرب K جار لها، و ثم ترتيب هذه المسافات تصاعدياً ثم رسمها لتحديد نقطة المنعطف الشكل().

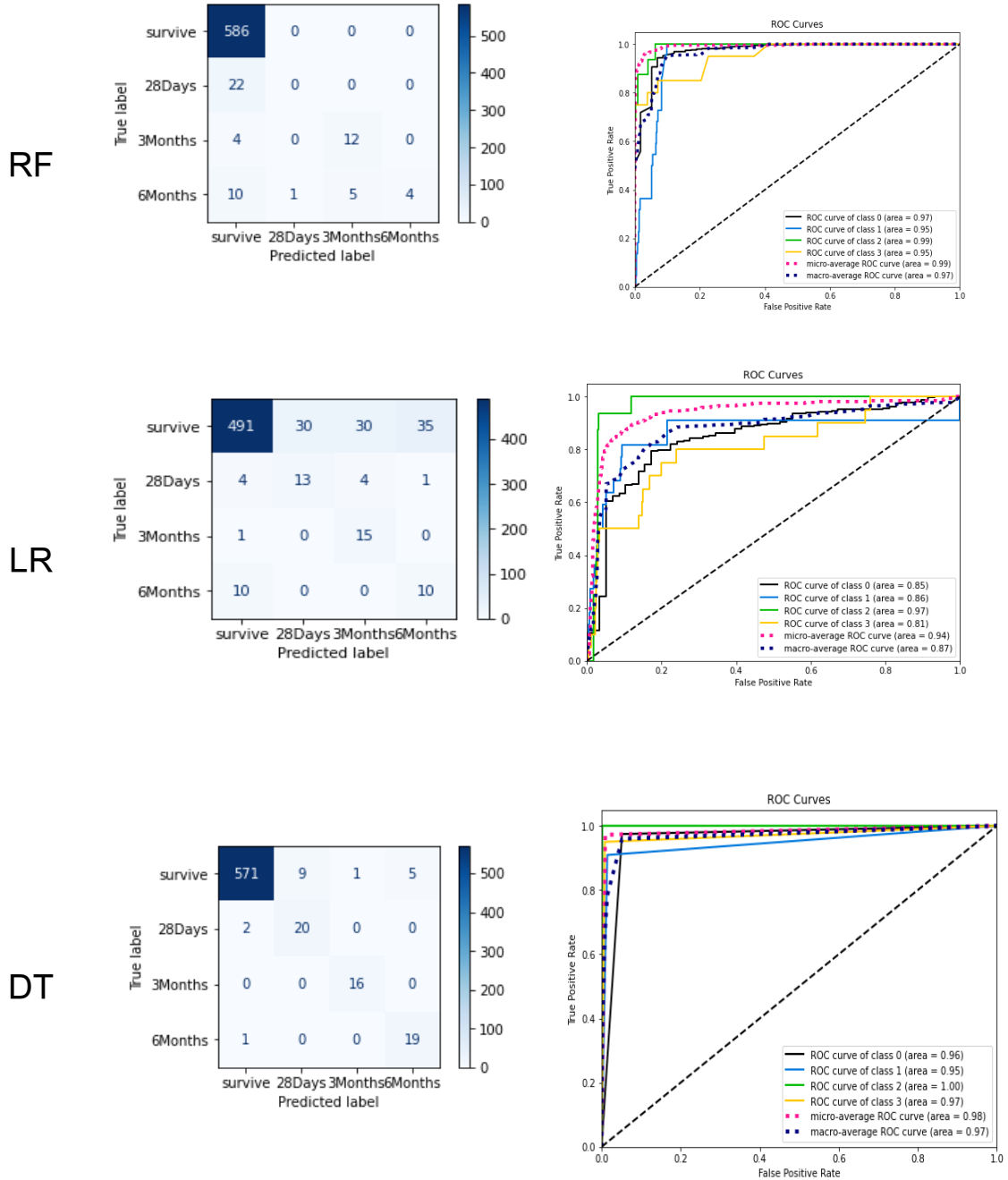


في حالتنا اخترنا $\epsilon=100$, $\text{minPoints}=15$ ، وقد نتج عن عملية العنقدة، 4 عناقيد، من أجل كل عنقود نميز إحدى حالتين إما أن العنقود هو فقط من نقاط الأكثرية، أو أن العنقود هو مزيج من نقاط الأقلية والأكثرية، بالنسبة للحالة الأولى، وضوحاً لا يوجد تداخل بين العينات الأقلية والأكثرية، بالنسبة للحالة الثانية،

- نقوم بتحديد النقاط الأقلية في العنقود
 - نقوم بحساب مصفوفة المسافات بين نقاط العنقود الأقلية والأكثرية (الأسطر نقاط الأقلية، الأعمدة نقاط الأكثرية)
 - نقوم بالتحويل من مصفوفة المسافات إلى مصفوفة من عمودين العمود الأول يمثل نقاط الأقلية، والعمود الثاني هو أصغر قيمة في السطر الخاص به في مصفوفة المسافات، بهذا نكون حددنا نقاط الأكثرية الأقرب إلى نقاط الأقلية.
 - نقوم بأخذ العناصر الفريدة من العمود الثاني، حيث هناك حالات تكون فيها نقطة الأكثرية مكررة، بمعنى أنها الأقرب لأكثر من نقطة أقلية.
 - أخيراً نحذف النقاط الأكثرية التي حددناها في الخطوة السابقة.
- في حالتنا كان عدد العناقيد قليل (4 عناقيد فقط، اثنان منها فقط فيها تداخل) لذلك عالجتنا كل عنقود على حدة، وبعد كل دورة من حذف نقاط الأكثرية كنا نقوم بإعادة تدريب نماذج تعلم الآلة الثلاثة (RF-DT-LR).
- مجموعة بيانات التدريب التي بدأنا منها: 2087 عينة، الجدول () يوضح النقاط الأكثرية المستبعدة في كل دورة.

النقاط المستبعدة من العنقود الثاني	النقاط المستبعدة من العنقود الأول	الدورة
15	51	1
-	54	2
-	55	3
-	58	4
-	54	5
-	58	6
-	60	7

نستعرض فيما يلي نتائج الاختبار بعد ثلاث دورات، حيث أصبح حجم مجموعة بيانات التدريب 1912 عينة، يوضح الشكل (5-16) النتائج التي حصلنا عليها.



الشكل (5-16) الموازنة بالطريقة المقترحة قبل الأمثلة

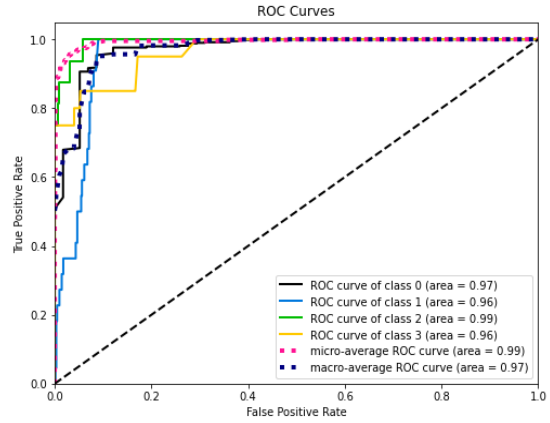
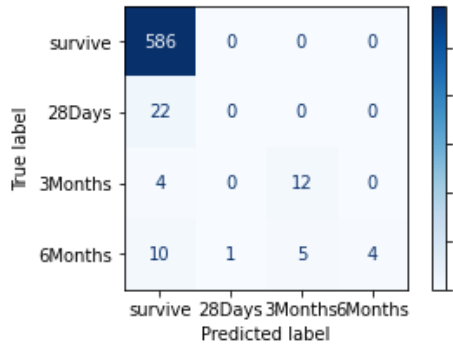
Model/Metrics	Accuracy	Precision	Recall	F1_score
RF	0.934	0.905	0.934	0.911
LR	0.821	0.905	0.821	0.852
DT	0.972	0.976	0.972	0.973

حول النتائج السابقة، نلاحظ أن المصنفات الثلاثة قد حققت نتائج أفضل من حالة تدريبها على البيانات الموازنة بطريقة SMOTE، من ناحية القدرة على تصنيف عينات الأقلية كما أنها كانت أفضل في تصنيف حالات الأقلية.

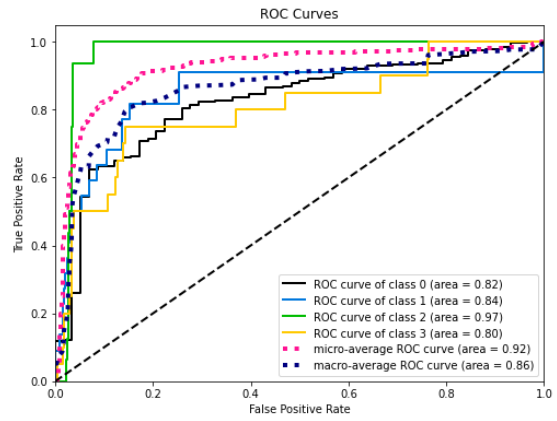
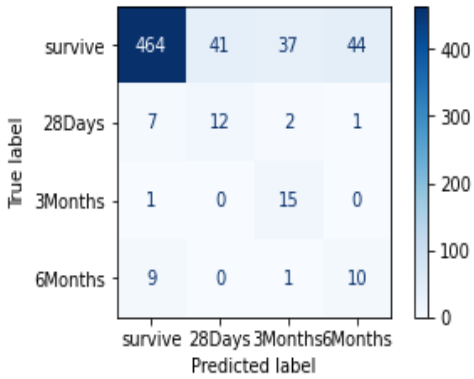
الملاحظة الثانية كانت تفوق مصنف أشجار القرار على مصنف الغابة العشوائية، ولدى البحث في هذه النقطة، تبين أن التدريب بالمعاملات الافتراضية، قد يكون السبب وراء هذه النتيجة، حيث أن النماذج المجمعة بشكل عام ليست أفضل من نظيراتها المفردة بشكل مطلق، خاصة في حالة التدريب بالمعاملات الافتراضية.

قمنا بعد ذلك بمتابعة أربع دورات إضافية من حذف النقاط الأكثرية، يوضح الشكل (5-17) النتائج التي تم الوصول إليها:

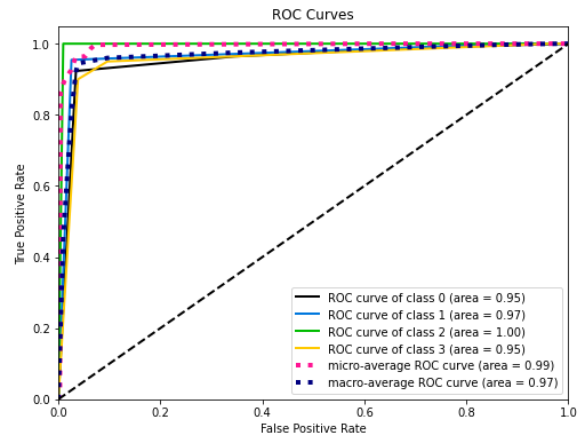
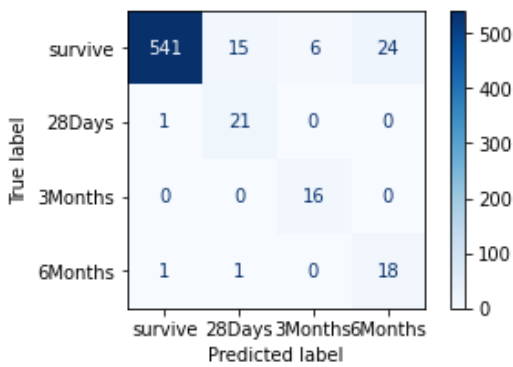
RF



LR



DT

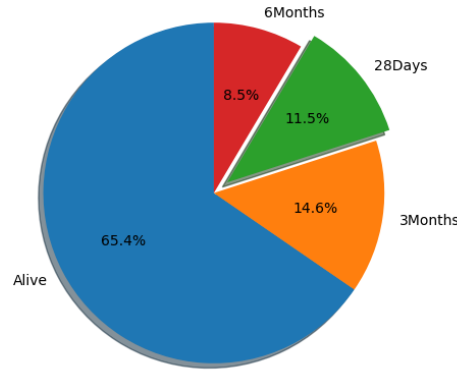


الشكل (17-5) الموازنة بالطريقة المقترحة بعد سبع دورات من الحذف

ونستعرض في الجدول التالي باقي المقيّمات للنماذج المطبقة أخيراً:

Model/Metrics	Accuracy	Precision	Recall	F1_score
RF	0.934	0.905	0.934	0.911
LR	0.777	0.897	0.777	0.821
DT	0.925	0.957	0.925	0.935

من النتائج السابقة نلاحظ أن قد بالغنا في حذف النقاط الأكثرية في مجموعة بيانات التدريب، التي أصبحت بحجم 1864 عينة، موزعة على الشكل التالي:



هذا الإفراط في الحذف أدى إلى عكس نزعة المصنفات في الأخطاء، بحيث أصبحت تخطئ في تصنيف الصنف الأكثرية، بتصنيفه كأصناف أقلية، هنا يجدر التنويه أن الحكم على النتيجة السابقة، متعلق بطبيعة المسألة المطروحة ودلالاتها، في حالتنا هذه، نعتبر هذا النمط من أخطاء التصنيف مقبولاً، باعتبار أنه من الأفضل أن أخطئ في تصنيف حالة ليست بخطر

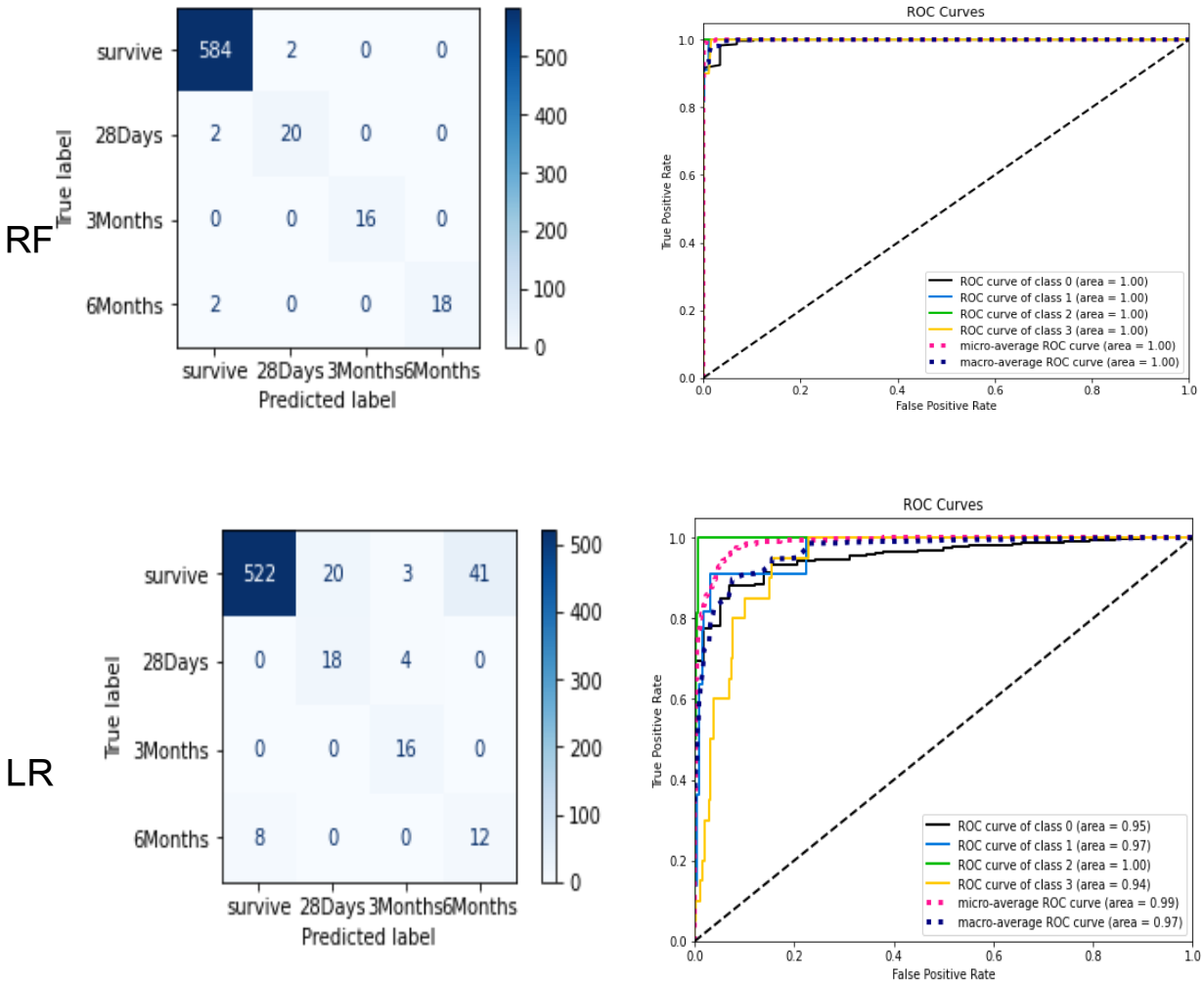
في الحقيقة على أنها حالة خطرة تستوجب المزيد من الاهتمام و الفحوص، على النوع الثاني من أخطاء التصنيف، أي تصنيف الحالة الخطيرة في الحقيقة على أنها ليست خطيرة، و نكرر أن الحكم في هذا النوع من الأخطاء، رهن بالسياق الدلالي للمسألة المطروحة.

بكل الأحوال سنعمد مجموعة التدريب التي وصلنا إليها عند التكرار 3 مرات، كونها أدت إلى نتائج أفضل.

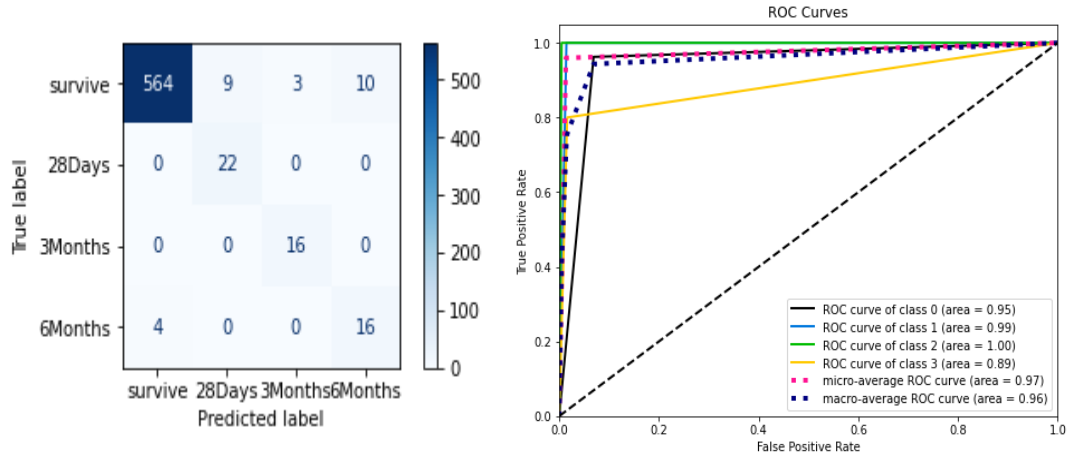
14.2.5. التدريب مع الموازنة بالطريقة المقترحة من قبلنا والتحسين

سنسعى في هذا الجزء إلى تحسين النتائج التي وصلنا إليها عبر أمثلة المعلمات الفائقة للنماذج الثلاث المطبقة، عبر تطبيق البحث الشبكي (GridSearchCV)، الذي تقدمه لنا مكتبة Sklearn، حيث يتم تمرير النماذج الثلاث المراد توليف معلماتها الفائقة، كما يتم تمرير شبكة المعلمات التي نريد إعادة التدريب وفقها، والمعامل الثالث هو عدد الطيات التي نريد التحقق من النتائج من خلالها عبر التحقق المتقاطع (Cross validation).

يوضح الشكل التالي النتائج التي حصلنا عليها:



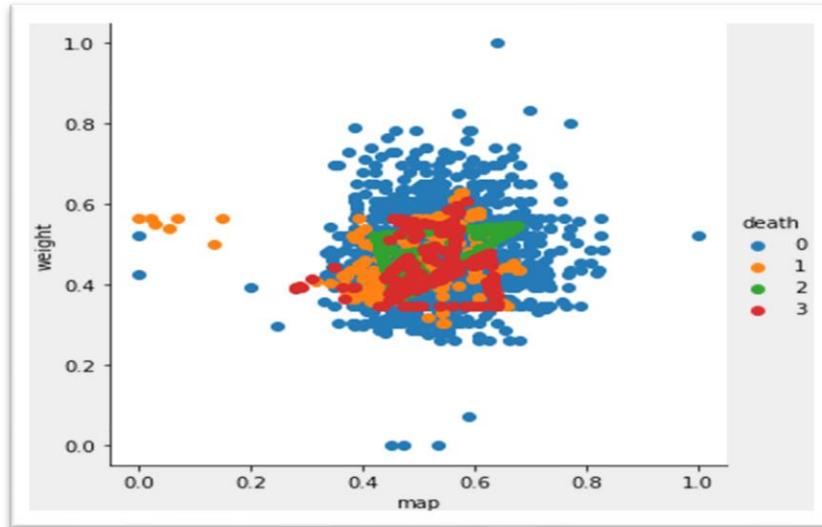
DT



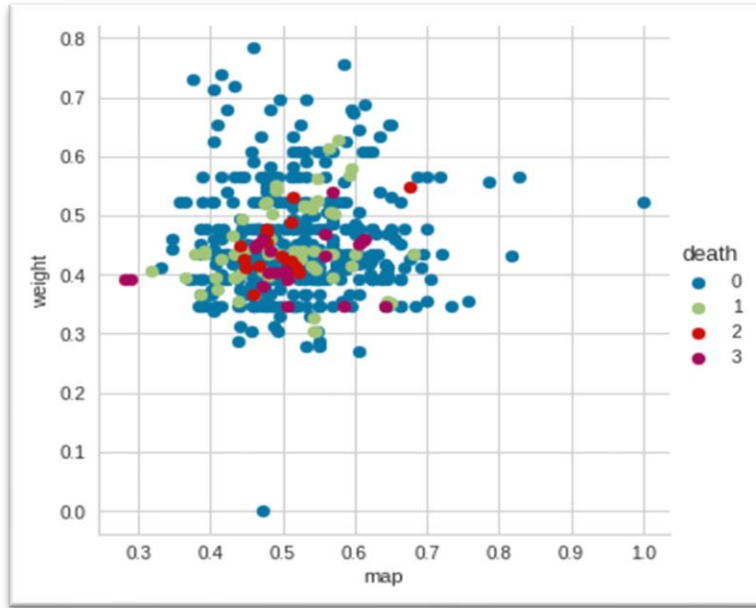
الشكل (5-18) الموازنة بالطريقة المقترحة بعد الأمثلة

نلاحظ أن النتائج قد تحسنت بصورة جيدة، وبالتالي يمكن القول إنه يمكن الاعتماد على طريقتنا في الموازنة مع تطبيق أمثلة لاختيار المعلمات الفائقة.

يوضح الشكل (5-19) توزيع بيانات مجموعة البيانات قبل وبعد عملية العنقدة وحذف النقاط المتداخلة.



قبل حذف النقاط المتداخلة

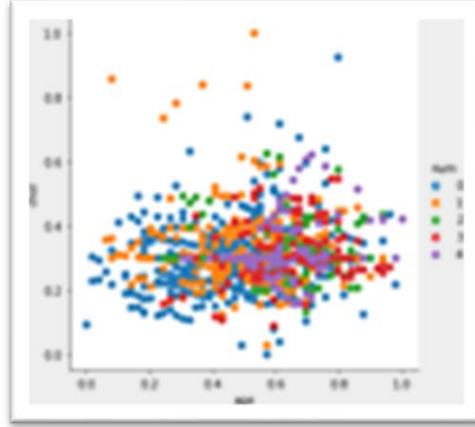


بعد حذف النقاط المتداخلة

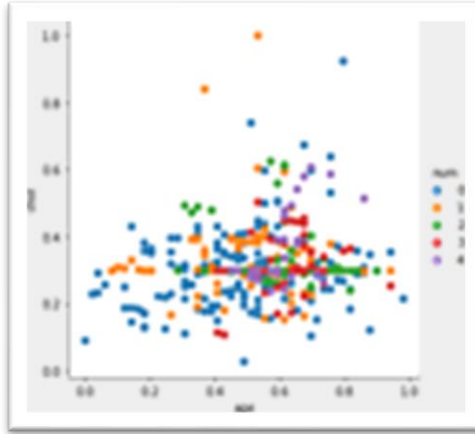
الشكل (5-19) توزع نقاط بيانات مجموعة البيانات المستخدمة قبل وبعد تطبيق مرحلة حذف النقاط المتداخلة اعتماداً على خوارزمية *DBScan*

من أجل التحقق من رصانة الخوارزمية المقترحة، تم إجراء عملية تحقق Validation بتطبيق الخوارزمية على مجموعة بيانات heart_disease_uci وهي مجموعة بيانات خاصة بكشف مرض الفشل القلبي اعتماداً على 15 خاصية (عمود). يمثل عمود num عمود الهدف وله خمس قيم أما القيمة 0 فتعني أن القلب طبيعي، والقيم من 1 إلى 4 فتعبر عن درجة الفشل القلبي فكلما كان الرقم أكبر يدل على أن احتمالية الفشل القلبي أكبر، مجموعة البيانات متاحة من خلال الرابط [42].

يوضح الشكل (5-20) توزيع نقاط البيانات في العنقود الأول قبل وبعد حذف النقاط المتداخلة.



قبل الحذف



بعد الحذف

الشكل (5-20) توزيع نقاط بيانات قبل وبعد تطبيق مرحلة حذف النقاط المتداخلة اعتماداً على خوارزمية العنقدة *DBScan*

يبين الجدول (5-12) الفرق بين أداء خوارزميات DT و RF المطبقة بشكل مباشر على مجموعة البيانات وبين أداء ذات الخوارزميات باستخدام طريقة موازنة البيانات المقترحة.

يتضح من الشكل أن أداء الخوارزميات قد تحسن بشكل ملحوظ لدى تطبيق طريقة الموازنة حيث ارتفع أداء مصنف DT من حيث معيار الدقة Accuracy بمقدار 1.72%، في حين زادت دقة مصنف RF بمقدار 13.03%.

القياسات الأصلية مأخوذة من ناتج تطبيق الخوارزميات ذاتها بدون استخدام موازنة [43] حيث تم مقارنة هذه القياسات مع نتائج اختبار وتقييم ذات المصنفات باستخدام خوارزمية الموازنة المقترحة من قبلنا في البحث.

نتائج تقييم أداء المصنفات بدون موازنة بيانات

Model/Metrics	Accuracy	Precision	Recall	F1_score
RF	0.51	0.45	0.51	0.47
DT	48.0	0.52	0.48	0.50

نتائج تقييم أداء المصنفات مع الموازنة باستخدام طريقة MUDSFC المحسنة

Model/Metrics	Accuracy	Precision	Recall	F1_score
RF	0.6944	0.7225	0.6853	0.6914
DT	0.6114	0.6292	0.5853	0.6070

الجدول (5-15) نتائج تقييم أداء مصنفات تعلم الآلة بدون ومع خوارزمية موازنة السمات المقترحة في البحث

3.5. مقارنة الدراسة الحالية مع الدراسات السابقة

من أجل معرفة موقع الدراسة الحالية بين الدراسات السابقة، تم مقارنة الدراسة الحالية مع أهم وآخر الدراسات المشابهة في مجال أنظمة دعم القرار الطبي الخاصة بمرض القلب وتم تلخيص النتائج في الجدول (5-16).

النتائج	اختيار السمات؟	حجم مجموعة البيانات	الطرق المتبعة	الدراسة
أفضل دقة لمصنف LR 89%	نعم	303 سجل 76 عمود	K-NN, SVM, ANN, NB, DT, LR	Haq وآخرون [3]
دقة 93.33%	لا	297 حالة طبية	نموذج إحصائي وعلى مصنف بايس الغوصي	Ali وآخرون [7]
حقق مصنف K-NN الأداء الأفضل	نعم	303 سجل 76 عمود	DT و K-NN و NB و RF.	Shah وآخرون [5]

دقة مصنف RF كانت الأفضل %86.6	لا	303 سجل 76 عمود	الخوارزمية الجينية، RFE، مصنفات SVM و NB و LR و RF و Adabost	Rani وآخرون [8]
دقة مصنف RF 100%، المصنفات الأخرى بين 89.62% و 97.95%	لا	1025 سجل و 14 عمود	RF و DT و K-NN	Ali وآخرون [9]
دقة 85.14%	نعم	303 سجل 76 عمود	SMO	Reddy وآخرون [10]
دقة مصنف K-NN 86.88%	لا	303 سطر و 13 عمود	RF و K-NN	Grag وآخرون [11]
الدقة 94.2%	لا	303 سطر و 13 عمود	IS, Deep Dense Layers	Bharti وآخرون [13]
دقة مصنف ROT كانت الأفضل %91.23	نعم	422 سجل	SVM و RF و NB و DT و K-NN و (LMT) و (ROT)	Plati وآخرون [14]
دقة 81.25%	لا	175 سجل	Perceptron	Khiabani وآخرون [15]
أفضل دقة تم تسجيلها كانت 99%	نعم	2008 سجل 167 سمة	MUDSFC LR, RF, DT Dense Deep Model	الدراسة الحالية

الجدول (5-16) مقارنة الدراسة الحالية مع الدراسات السابقة

يبين الجدول (5-16) تفوق الدراسة الحالية على الدراسات السابقة ليس من حيث الأداء وحسب وإنما في نوعية مجموعة البيانات المستخدمة (عدد الأعمدة أو السمات وعدد السجلات أو الحالات).

6. الفصل السادس: الخاتمة والآفاق المستقبلية

1.6.1. الخاتمة

تم في البحث الحالي تصميم نظام دعم قرار طبي لكشف خطورة حدوث الفشل القلبي لدى مرضى القلب باستخدام مجموعة بيانات طبية معيارية تتضمن معلومات نصية (167 عمود) تمثل قياسات وتحاليل طبية ومعلومات سريرية عن حالة 2008 حالة طبية (سجل طبي).

تم تطبيق عمليات معالجة مسبقة وموازنة لمجموعة البيانات حيث تم تصميم خوارزمية جديدة في مجال موازنة مجموعات البيانات تقوم على أساس الترابط بين عمود الهدف وباقي أعمدة مجموعة البيانات لتطبيق عمليتي موازنة مختلفتين تتضمن الأولى زيادة عينات الأصناف غير المسيطرة في حين تمثل الثانية تقليل عينات الصنف المسيطر.

تم اقتراح استخدام خوارزمية اختيار سمات قائمة على اختيار السمات الأكثر ترابطاً مع عمود الهدف باستخدام 3 عتبات مختلفة.

تم اقتراح 7 سيناريوهات تدريب مختلفة تتضمن تدريب 3 أنظمة تعلم آلة ونموذج تعلم عميق مقترح من قبلنا. تم تدريب الأنظمة بدون ومع اختيار السمات. تم تدريبها أيضاً باستخدام موازنة SMOTE ثم باستخدام خوارزمية MUDSFC المقترحة من قبلنا ثم باستخدام النسخة المسحنة من خوارزمية MUDSFC باستخدام تقليل تداخل الأصناف اعتماداً على العنقدة Clustering.

تم كذلك تدريب النماذج بدون تحسين وبعد تطبيق التحسين.

لم يتحسن نماذج تعلم الآلة DT و RF قبل أو بعد تطبيق خوارزمية التحسين.

تحسن أداء نموذج LR بمقدار 11.7% من حيث الدقة. كما أنه تحسن بمقدار 9.19% من حيث Precision وبمقدار 11.77% من حيث Recall.

بالنسبة لنموذج التعلم العميق، حسنت عملية التدريب باستخدام التحسين الأداء بشكل طفيف وبمقدار 0.98% لدقة التحقق في حين بقيت دقة الاختبار كما هي 98.62% في كلتا الحالتين. ومن الملاحظ تحسن دقة تصنيف الأصناف الفرعية 1 و2 و3.

إن عملية اختيار السمات بعتبة 0.02 وفرت زمن تدريب مقداره 0.83 ميلي ثانية لكل خطوة ضمن التكرار الواحد من التدريب مع انخفاض طفيف جداً في الدقة مقداره 1.8% رغم أن نسبة السمات التي نتجت بعد الاختيار كانت فقط 47.7% من كامل سمات مجموعة البيانات.

أثبتت خوارزمية الموازنة اعتماداً على ارتباط السمات المقترحة من قبل البحث فعالية جيدة في موازنة مجموعات البيانات حيث تم تطبيقها على مجموعة البيانات المستخدمة في البحث وعلى مجموعة بيانات أخرى تتضمن حالة عدم موازنة بين الأصناف وحققت الهدف المرجو منها أيضاً حيث تحسنت دقة المصنفات لدى تطبيق الخوارزمية عن الحالة التي تم فيها تدريب المصنفات بدون موازنة.

قدم البحث الحالي الإسهامات الآتية في مجال البحث العلمي:

1. تحديد مستوى خطورة الإصابة بالفشل القلبي مع الأخذ بعين الاعتبار الفترة الزمنية الممكن حدوث الوفاة فيها.
2. اختبار ومقارنة نماذج تعلم الآلة والتعلم العميق.
3. اقتراح خوارزمية جديدة لموازنة البيانات.
4. بناء نموذج تعلم عميق وتحسينه عبر تطبيق السمات.
5. إعداد ورقة بحثية بالنتائج التي تم الوصول إليها.

2.6. الآفاق المستقبلية

طرح هذا البحث العديد من النماذج الجديدة سواء في مجال التعلم العميق أو خوارزميات موازنة السمات سواء المعتمدة على زيادة عينات الأصناف غير المسيطرة وتقليل عينات الصنف المسيطر أو عن طريق تقليل العينات المتداخلة بين الصنف المسيطر وبقية الأصناف الثانوية. يوصي البحث باستخدام مصنف LR مع مجموعات البيانات متعددة السمات، واستخدام مبدأ اختيار السمات.

يوصي البحث بتطبيق عملية موازنة السمات باستخدام خوارزمية Enhanced MUDSFC من أجل زيادة عينات الأصناف الثانوية بشكل ديناميكي وتقليل عينات الصنف المسيطر المتداخلة مع عينات الأصناف الثانوية غير المسيطرة.

يقترح البحث استخدام تقنيات اختيار سمات أخرى والمقارنة مع الطريقة المتبعة.

تم في هذا البحث اتباع مقاربات تقليدية بما يخص موازنة البيانات سواء لبيانات التدريب أو إعادة تشكيل مجموعة البيانات، يقترح البحث اعتماد مقاربات أخرى مستقبلاً مثل تقنيات (FSL: Few-Shot learning) التي من شأنها أن تمكن نموذج تم تدريبه سابقاً من التعميم على فئات جديدة من البيانات باستخدام عدد قليل من العينات المصنفة لكل فصل.س

7. الملاحق

1.7. الملحق أ: وصف مجموعة البيانات.

النوع	الوصف	السمة
رقمي، مستمر	المعرف الخاص بالمريض	inpatient.number
Home, Unknown, Died, Healthcare Facility	وجهة المريض بعد الخروج من المشفى	DestinationDischarge
Cardiology, General Ward, ICU, Others	قسم القبول في المشفى	admission.ward
NonEmergency, Emergency	طريقة القبول في المشفى	admission.way
Officcer, Worker, Others	مهنة المريض	Occupation
Cardiology, General Ward, ICU, Others	القسم الذي تخرج منه المريض	discharge.department
من 1 إلى 5	عدد مرات دخول المريض للمشفى	visit.times
Male, Female	جنس المريض	Gender
[36.5,38.9]	درجة حرارة جسم المريض	body.temperature
المتوسط: 85.72	نبض المرض (دقة/دقيقة)	Pulse
المتوسط: 19.09	التنفس (نفس/ دقيقة)	Respiration
المتوسط: 131.62	ضغط الدم الانقباضي (مليمتر زئبقي) -mmHg	systolic.blood.pressure
المتوسط: 76.70	ضغط الدم الانبساطي (مليمتر زئبقي) -mmHg	diastolic.blood.pressure
المتوسط: 94.89	متوسط الضغط الشرياني (مليمتر زئبقي) -mmHg	Map
المتوسط: 52.55	الوزن (كيلوغرام)	Weight
المتوسط: 1.56	الطول (متر)	Height
المتوسط: 21.79	مؤشر كتلة الجسم (كغ/متر مربع)	BMI
Left, Right, Both	نوع الفشل القلبي	type.of.heart.failure
Class I, Class II, Class IV	تصنيف جمعية القلب في نيويورك	NYHA.cardiac.function.classification

	النوع	الوصف	السمة
Class I, Class II, Class III, Class IV	فئوي، اسمي	تصنيف كليليب	Killip.grade
	رقمي، ثنائي	احتشاء عضلة القلب	myocardial.infarction
	رقمي، ثنائي	قصور القلب الاحتقاني	congestive.heart.failure
	رقمي، ثنائي	أمراض الأوعية الدموية الطرفية	peripheral.vascular.disease
	رقمي، ثنائي	أمراض دماغية وعائية	cerebrovascular.disease
	رقمي، ثنائي	مرض عقلي	dementia
	رقمي، ثنائي	انسداد رئوي مزمن	Chronic.obstructive.pulmonary.disease
	رقمي، ثنائي	التهاب النسيج الضام	connective.tissue.disease
	رقمي، ثنائي	القرحة الهضمية	peptic.ulcer.disease
	رقمي، ثنائي	مرض السكري	Diabetes
	رقمي، ثنائي	مرض كلوي مزمن	moderate.to.severe.chronic.kidney.disease
	رقمي، ثنائي	الشلل النصفي	Hemiplegia
	رقمي، ثنائي	سرطان الدم	leukemia
	رقمي، ثنائي	سرطان غدد لمفاوية	malignant.lymphoma
	رقمي، ثنائي	الأورام الصلبة	solid.tumor
	رقمي، ثنائي	أمراض الكبد	liver.disease
	رقمي، ثنائي	مرض الإيدز	AIDS
من 1 إلى 6	رقمي، منقطع	مؤشر تشارلسون للأمراض المصاحبة	CCI.score
	رقمي، ثنائي	فشل الجهاز التنفسي من النمط الثاني	type.II.respiratory.failure
Nonresponsive, ResponsiveToPain, ResponsiveToSound, Clear	فئوي، نصي	تقييم وعي المريض	consciousness
[1,2,3,4]	فئوي، رقمي	استجابة حدقة العين	eye.opening
[1,2,3,4,5]	فئوي، رقمي	الاستجابة اللفظية	verbal.response
[1,2,3,4,5,6]	فئوي، رقمي	الاستجابة الحركية	movement

النوع	الوصف	السمة
IMV,NIMV,None	دعم التنفس (تهوية غازية، غير غازية، بدون تهوية)	respiratory.support.
OxygenTherapy, AmbientAir	الايوكسجين	oxygen.inhalation
المتوسط: 32.67	رقمي، ثنائي	نسبة الاوكسجين المستنشق
المتوسط: 47.70	رقمي، ثنائي	فشل كلوي حاد
المتوسط: 47.70	رقمي، مستمر	نسبة الكسر البطني الأيسر
المتوسط: 53.76	رقمي، مستمر	القطر الانبساطي للبطين الأيسر (مليمتر)
المتوسط: 2.85	رقمي، مستمر	السرعة القصوى لموجة الصمام التاجي E (متر ثانية)
المتوسط: 1.73	رقمي، مستمر	السرعة القصوى لموجة الصمام التاجي A (متر ثانية)
المتوسط: 1.22	رقمي، مستمر	نسبة E/A
المتوسط: 3.66	رقمي، مستمر	سرعة رجوع الصمام ثلاثي الشرف (متر ثانية)
المتوسط: 34.37	رقمي، مستمر	ضغط الصمام ثلاثي الشرف (مليمتر زئبقي)
Alive,Dead DischargeAgainstOrder	فئوي، نصي	نتيجة الاستشفاء في المشفى
	رقمي، ثنائي	الوفاة خلال 28 يوم
	رقمي، ثنائي	إعادة القبول في المشفى خلال 28 يوم
	رقمي، ثنائي	الوفاة خلال 3 أشهر
	رقمي، ثنائي	إعادة القبول في المشفى خلال 3 أشهر
	رقمي، ثنائي	الوفاة خلال 6 أشهر

	النوع	الوصف	السمة
	رقمي، ثنائي	إعادة القبول في المشفى خلال 6 أشهر	re.admission.within.6.months
المتوسط: 29.52	رقمي، مستمر	عدد الأيام من القبول للوفاة	time.of.death..days.from.admission.
المتوسط: 126.71	رقمي، مستمر	عدد الأيام من القبول الأولي إلى إعادة القبول	re.admission.time..days.from.admission.
	رقمي، ثنائي	العودة إلى قسم الطوارئ خلال 6 أشهر	return.to.emergency.department.within.6.months
المتوسط: 126.73	رقمي، مستمر	عدد الأيام من القبول الأولي إلى العودة إلى قسم الطوارئ	time.to.emergency.department.within.6.months
المتوسط: 108.91	رقمي، مستمر	الكرياتينين (ميكرومول/لتر)	creatinine.enzymatic.method
المتوسط: 9.55	رقمي، مستمر	اليوريا (ميلي مول/لتر)	Urea
المتوسط: 483.42	رقمي، مستمر	حمض البول (ميكرو مول/لتر)	uric.acid
المتوسط: 68.62	رقمي، مستمر	معدل الترشيح الكبيبي	glomerular.filtration.rate
المتوسط: 1.84	رقمي، مستمر	سيستاتين (ميلي غرام / لتر)	cystatin
المتوسط: 7.32	رقمي، مستمر	تعداد الخلايا البيضاء (3^{10} خلية/لتر)	white.blood.cell
المتوسط: 0.068	رقمي، مستمر	معدل وحيدات الخلية في الدم	monocyte.ratio
المتوسط: 0.47	رقمي، مستمر	تعداد وحيدات الخلية في الدم (3^{10} لتر)	monocyte.count
المتوسط: 3.86	رقمي، مستمر	تعداد الكريات الحمراء (12^{10} كرية/لتر)	red.blood.cell
المتوسط: 14.93	رقمي، مستمر	معامل التباين للكريّة الحمراء (%)	coefficient.of.variation.of.red.blood.cell.distribution.width
المتوسط: 48.68	رقمي، مستمر	الانحراف المعياري للكريّة الحمراء	standard.deviation.of.red.blood.cell.distribution.width

الوصف	النوع	السمعة
متوسط حجم الكرية الحمراء (فيتمو لىتر)	رقمى؁ مستمر	mean.corpuscular.volume
نسبة الكرىات الحمر فى الدم (%)	رقمى؁ مستمر	hematocrit
تعداد الخلىا للمفاوىة (9^{10} التر)	رقمى؁ مستمر	lymphocyte.count
متوسط كمىة الهىموغلوبىن (بىكو غرام)	رقمى؁ مستمر	mean.hemoglobin.volume
متوسط تركىز الهىموغلوبىن (غرامالتر)	رقمى؁ مستمر	mean.hemoglobin.concentration
متوسط كمىة الصفىحات (فىمو لىتر)	رقمى؁ مستمر	mean.platelet.volume
معدل الخلىا القاعدىة	رقمى؁ مستمر	basophil.ratio
عدد الخلىا القاعدىة (9^{10} لتر)	رقمى؁ مستمر	basophil.count
معدل الحمضات	رقمى؁ مستمر	eosinophil.ratio
عدد الحمضات (9^{10} التر)	رقمى؁ مستمر	eosinophil.count
الهىموغلوبىن (غرامالتر)	رقمى؁ مستمر	hemoglobin
تعداد الصفىحات الدموىة (9^{10} التر)	رقمى؁ مستمر	platelet
عرض توزع الصفىحات (فىمو لىتر)	رقمى؁ مستمر	platelet.distribution.width
نسبة الصفىحات الدموىة (%)	رقمى؁ مستمر	platelet.hematocrit
نسبة العدلات (%)	رقمى؁ مستمر	neutrophil.ratio
عدد العدلات (9^{10} التر)	رقمى؁ مستمر	neutrophil.count
بروتىن دىمر (غرامالتر)	رقمى؁ مستمر	D.dimer
معدل التجلط (%)	رقمى؁ مستمر	international.normalized.ratio
وقت التفعىل الجزئى للثرومبوبلاستىن (ثانىة)	رقمى؁ مستمر	activated.partial.thromboplastin.time
زمن الثرومبىن (ثا)	رقمى؁ مستمر	thrombin.time
نشاط البروثرومبىن (%)	رقمى؁ مستمر	prothrombin.activity
معدل وقت التخنتر	رقمى؁ مستمر	prothrombin.time.ratio

الوصف	النوع	السمعة
المتوسط: 3.21	رقمي، مستمر	fibrinogen
المتوسط: 0.67	رقمي، مستمر	high.sensitivity.troponin
المتوسط: 173.81	رقمي، مستمر	myoglobin
المتوسط: 23.80	رقمي، مستمر	carbon.dioxide.binding.capacity
المتوسط: 2.29	رقمي، مستمر	Calcium
المتوسط: 3.98	رقمي، مستمر	potassium
المتوسط: 101.7	رقمي، مستمر	chloride
المتوسط: 138.23	رقمي، مستمر	Sodium
المتوسط: 1.12	رقمي، مستمر	Inorganic.Phosphorus
المتوسط: 0.87	رقمي، مستمر	serum.magnesium
المتوسط: 0.20	رقمي، مستمر	creatin.kinase.isoenzyme.to.creatin.kinase
المتوسط: 0.80	رقمي، مستمر	hydroxybutyrate.dehydrogenase.to.lactate.dehydrogenase
المتوسط: 220.64	رقمي، مستمر	hydroxybutyrate.dehydrogenase
المتوسط: 66.68	رقمي، مستمر	glutamic.oxaloacetic.transaminase
المتوسط: 139.19	رقمي، مستمر	creatin.kinase
المتوسط: 20.13	رقمي، مستمر	creatin.kinase.isoenzyme
المتوسط: 287.13	رقمي، مستمر	lactate.dehydrogenase
المتوسط: 1279.23	رقمي، مستمر	brain.natriuretic.peptide
المتوسط: 22.91	رقمي، مستمر	high.sensitivity.protein
المتوسط: 4.59	رقمي، مستمر	nucleotidase
	رقمي، مستمر	cholinesterase
المتوسط: 62.79	رقمي، مستمر	glutamyltranspeptidase

الوصف	النوع	السمعة
جلوتاميك بروفيك (وحدة/ لتر)	رقمي، مستمر	glutamic.pyruvic.transaminase
معدل اوكسيبلاتين	رقمي، مستمر	glutamic.oxaliplatin
بيليروبين (ميكرو مول لتر)	رقمي، مستمر	indirect.bilirubin
فوسفات قلوي (وحدة/لتر)	رقمي، مستمر	alkaline.phosphatase
غلوبولين (غرام/لتر)	رقمي، مستمر	Globulin
بيليروبين مباشر (ميكرو مول/لتر)	رقمي، مستمر	direct.bilirubin
بيليروبين كلي (ميكرو مول/لتر)	رقمي، مستمر	total.bilirubin
حمض الصفراء (ميكرو مول/لتر)	رقمي، مستمر	total.bile.acid
البروتين الكلي (غرام/لتر)	رقمي، مستمر	total.protein
معدل الترسيب (ميلي مترا ساعة)	رقمي، مستمر	erythrocyte.sedimentation.rate
الكوليسترول (ميلي مول/لتر)	رقمي، مستمر	cholesterol
كوليسترول شحمي منخفض الكثافة (ميلي مول/لتر)	رقمي، مستمر	low.density.lipoprotein.cholesterol

8. المراجع

- [1] A. L. T. B. H. a. G. C. F. Bui ، "Epidemiology and risk profile of heart failure "-2011
- [2] M. Durairaj , N. Ramasamy. ، "A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate"-2016
- [3] A. U. Haq ،J. P. Li ،M. H. Memon ،S. Nazir , R. Sun ، "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms" -2018
- [4] Mayo Clinic Staff ، "Heart failure mayoclinic" -2022
- [5] J. López-Sendón ، "The heart failure epidemic"-2011
- [6] K. Sullivan ،B. S. Doumouras ،B. T. Santema ،M. N. Walsh ،P. S. Douglas ،A. A. Voors , H. G. V. Spall. ، "Sex-specific differences in heart failure: pathophysiology, risk factors, management, and outcomes"-2021
- [7] Y. Lai ،L. Wang , M. T. Gladwin ، "Insights into the pulmonary vascular complications of heart failure with preserved ejection fraction"-2019

- [8] J. A. Osheroff ,J. M. Teich ,B. Middleton ,E. B. Steen ,A. Wright , D. E. Detmer , "A roadmap for national action on clinical decision support"-2007
- [9] A. Belard ,T. Buchman ,J. Forsberg ,B. K. Potter ,C. J. Dente ,A. Kirk , E. Elster , "Precision diagnosis: a view of the clinical decision support systems (CDSS) landscape through the lens of critical care"-2017
- [10] R. T. Sutton ,D. Pincock ,D. C. Baumgart ,D. C. Sadowski ,R. N. Fedorak , K. I. Kroeker , "An overview of clinical decision support systems: benefits, risks, and strategies for success"-2020
- [11] A. M. Antoniadis ,Y. Du ,Y. Guendouz ,L. Wei ,C. Mazo ,B. A. Becker , C. Mooney , "Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: a systematic review"-2021
- [12] A. Silva Almodóvar , M. C. Nahata , "Implementing clinical decision support tools and pharmacovigilance to reduce the use of potentially harmful medications and health care costs in adults with heart failure"-2021
- [13] T. K. J. Groenhof ,Z. H. Rittersma ,M. L. Bots ,M. Brandjes ,J. J. L. Jacobs ,D. E. Grobbee ,W. W. v. Solinge ,F. L. J. Visseren ,S. Haitjema , F. W. Asselbergs , "A computerised decision support system for cardiovascular risk management ‘live’ in the electronic health record environment: development, validation and implementation—the Utrecht Cardiovascular Cohort Initiative"-2019

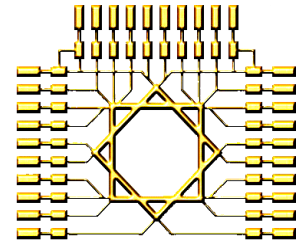
- [14] S. E. Awan ,F. Sohel ,F. M. Sanfilippo ,M. Bennamoun , G. Dwivedi ,
"Machine learning in heart failure: ready for prime time"-2018
- [15] E. E. Tripoliti ,T. G. Papadopoulos ,G. S. Karanasiou ,K. K. Naka , D. I.
Fotiadis , "Heart failure: diagnosis, severity estimation and prediction of
adverse events through machine learning techniques"-2017
- [16] D. Shah ,S. Patel , S. K. Bharti , "Heart disease prediction using machine
learning techniques. SN Computer Science"-2020
- [17] V. V. Ramalingam ,A. Dandapath , M. K. Raja , "Heart disease prediction
using machine learning techniques: a survey"-2018
- [18] E. Choi ,A. Schuetz ,W. F. Stewart , J. Sun , "Using recurrent neural
network models for early detection of heart failure onset"-2017
- [19] D. J. Cornforth , H. F. Jelinek , "Detection of congestive heart failure
using Renyi entropy"-2016
- [20] Y. Isler , "Discrimination of systolic and diastolic dysfunctions using
multi-layer perceptron in heart rate variability analysis"-2016
- [21] P. C. Austin ,J. V. Tu ,J. E. Ho ,D. Levy , D. S. Lee , "Using methods
from the data-mining and machine-learning literature for disease
classification and prediction: a case study examining classification of
heart failure subtypes"-2013
- [22] Y. Kang ,M. D. McHugh ,J. Chittams , K. H. Bowles , "Utilizing home
health care electronic health records for telehomecare patients with heart
failure: a decision tree approach to detect associations with
rehospitalizations"-2016

- [23] Z. Y. Z. Zhang ,L. Cao ,Z. Xu ,R. Chen ,L. Lv , P. Xu , "Hospitalized patients with heart failure: integrating electronic healthcare records and external outcome data"-2020
- [24] M. G. El-Shafiey ,A. Hagag ,E.-S. A. El-Dahshan , M. A. Ismail , "A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest"-2022
- [25] W. A. W. A. Bakar ,N. L. N. B. Josdi ,M. B. Man , Y. S. Triana , "An Evaluation of Artificial Neural Networks and Random Forests for Heart Disease Prediction"-2022
- [26] P. Rani ,R. Kumar ,N. M. Ahmed , A. Jain , "A decision support system for heart disease prediction based upon machine learning"-2021
- [27] L. S. U. K. N. A. G. I. Y. I. Q. A. N. a. R. N. Ali , "A feature-driven decision support system for heart failure prediction based on statistical model and Gaussian naive bayes"-2019
- [28] R. Bharti ,A. Khamparia ,M. Shabaz ,G. Dhiman ,S. Pande, P. Singh , "Prediction of heart disease using a combination of machine learning and deep learning"-2021
- [29] U. Nagavelli ,D. Samanta, P. Chakraborty , "Machine Learning Technology-Based Heart Disease Detection Models"-2022
- [30] M. M. Ali ,B. K. Paul ,K. Ahmed ,F. M. Bui ,J. M. Quinn , M. A. Moni , "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison"-2021

- [31] K. V. V. Reddy ,I. Elamvazuthi ,A. A. Aziz ,S. Paramasivam ,H. N. Chua , S. Pranavanand , "Heart disease risk prediction using machine learning classifiers with attribute evaluators"-2021
- [32] A. Garg ,B. Sharma , R. Khan , "Heart disease prediction using machine learning techniques"-2021
- [33] M. Kavitha ,G. Gnaneswar ,R. Dinesh ,Y. R. Sai , R. S. Suraj , "Heart disease prediction using hybrid machine learning model"-2021
- [34] D. K. Plati ,E. E. Tripoliti ,A. Bechlioulis ,A. Rammos ,I. Dimou ,L. Lakkas , C. Watson , "A machine learning approach for chronic heart failure diagnosis"-2021
- [35] S. J. Khiabani ,A. Batani, E. Khanmohammadi , "A hybrid decision support system for heart failure diagnosis using neural networks and statistical process control"-2022
- [36] M. Nishat ,F. F. Mirza ,I. J. Ratul ,A. Al-Monsur ,A. Mohammad ,S. M. Nasrullah ,M. T. Reza, M. R. H. Khan , "A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset"-2022
- [37] A. Goldberger ,L. Amaral ,L. Glass ,J. Hausdorff ,P. Ivanov ,R. Mark ,J. Mietus ,G. Moody ,C. Peng , H. Stanley , "Components of a new research resource for complex physiologic signals"-2020
- [38] physionet , "Heart Failure Dataset,"-2020
<https://physionet.org/content/heart-failure-zigong/1.2/>.

- [39] T. Yu , H. Zhu , "Hyper-parameter optimization: A review of algorithms and applications"-2020
- [40] J. Bergstra , Y. Bengio , "Random search for hyper-parameter optimization"-2012
- [41] R. Yacouby, D. Axman , "Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models"-2020
- [42] R. Detrano , A. Janosi , W. Steinbrunn , M. Pfisterer , J.-J. Schmid , S. Sandhu , K. H. Guppy , S. Lee , V. Froelicher , "International application of a new probability algorithm for the diagnosis of coronary artery disease"-2019
- [43] D. W. Aha , D. Kibler , "Instance-based prediction of heart-disease presence with the Cleveland database " ,
"<https://archive.ics.uci.edu/ml/datasets/Heart+Disease>.
- [44] "heart disease multi class classification"
<https://www.kaggle.com/code/arnoldyanga/heart-disease-multi-class-classification/notebook>.

**Syrian Arab Republic
Higher Institute for Applied Sciences and Technology
Informatics Department**



HIAS

Decision support model for predicting the risk for patients with heart disease

Submitted in Fulfillment of the Requirements for Master's Degree in
Decision Support System

By

Eng. Anas Issa

Supervised By

PhD. Mohammad Al-chaita

PhD. Mustapha Al-Dakkak

December 2022