مقارنة خوارزميات التكييس والتعزيز في تعليم المجموعات (Ensemble learning) للتنبؤ بأمراض القلب

*أ.بدر نجيب عويدات

1. المستخلص: إن تعليم المجموعات (Ensemble learning) هو نهج عام لوصف التعلم الآلي يسعى إلى أداء تنبؤي أفضل من خلال الجمع بين التنبؤات من نماذج متعددة. يتضمن أسلوب تعليم المجموعات عددًا من الطرق منها أسلوب التكييس (Bagging) والتعزيز(Boosting) ويمتلك هذين الاسلوبين مجموعة من الخوارزميات منها خوارزمية الغابة العشوائية Random Force وخوارزمية التدرج التكيفي AdaBoost وخورزميات التدرج التعزيزي (Gradient Boosting). في هذا البحث، سنقارن بين الاسلوبين التكييس والتعزيز من حيث نسبة صحة خوارزمية التصنيف (Accuracy) ومقياس المثالية (Recall) ومقياس الدقة (Precision) معامل Cohen Kappa ومقياس F (F-measure) ومقياس الحساسية (Sensitivity) ومقياس النوعية (Specificity) ومستوى المنطقة الواقعة تحت منحنى ROC في التنبؤ بأمراض قصور القلب.

الكلمات المفتاحية: آلات تعزيز التدرج، آلة تعزيز التدرج الخفيف، تعزيز التدرج الأقصى، مقياس نسبة صحة، مقياس المثالية، مقياس الدقة، تعليم المجموعات.

2. المقدمة:

العديد من مشكلات التعلم الآلي معقدة للغاية بحيث لا يمكن حلها باستخدام نموذج أو خوارزمية واحدة. تدرب خوارزميات تعليم المجموعات (Ensemble learning) مجموعة من نماذج التعلم الآلي المتنوعة للعمل معًا لحل مشكلة ما. من خلال تجميع مخرجاتها، يمكن لنماذج المجموعات هذه تقديم نتائج غنية ودقيقة بمرونة. تتيح خوارزميات تعليم المجموعات إجراء تنبؤات قوية دون الحاجة إلى التعامل مع البيانات الضخمة وقوة المعالجة التي يتطلبها التعلم العميق. يقوم بتعيين نماذج متعددة للعمل على حل مشكلة ما، والجمع بين نتائجها للحصول على أداء أفضل من نموذج واحد يعمل بمفرده. يعمل نهج "حكمة الحشود" هذا على استخلاص المعلومات من عدة نماذج إلى مجموعة من النتائج عالية الدقة. [1]

يعد تعليم المجموعات أحد أكثر الطرق فعالية لبناء نموذج فعال للتعلم الآلي. ويمكن بناء نموذج تعليمي آلي جماعي باستخدام نماذج بسيطة للحصول على درجات عالية تتساوى مع النماذج التي تتطلب موارد كبيرة للبيانات مثل الشبكات العصبية. [3]

أمراض القلب والأوعية الدموية (CVDs) هي السبب الأول للوفاة على مستوى العالم، حيث تودي بحياة ما يقدر بنحو 17.9 مليون شخص كل عام، وهو ما يمثل 31 ٪ من جميع الوفيات في جميع أنحاء العالم. أربع من كل خمسة حالات الوفيات هي بسبب النوبات القلبية والسكتات الدماغية، وثلث هذه الوفيات تحدث قبل الأوان مع الأشخاص الذين تقل أعمارهم عن 70 عامًا. فشل القلب هو حدث شائع تسببه الأمراض القلبية الوعائية وقد استخدمنا في هذا البحث مجموعة بيانات تحتوي على 13 ميزة يمكن استخدامها للتنبؤ بأمراض القلب المحتملة.

3. مشكلة الدراسة:

تكمن مشكلة الدراسة في:

* كلية تقنية المعلومات، جامعة الزيتونة Bader_najep@yahoo.com

1.3. يعتبر تعليم المجموعات هو احد الأساليب التي تستخدم في التعلم الآلي لتقليل الأخطاء في تحليل البيانات التنبئي، لانه قد يتسبب نموذج التعلم الآلي المفرد في حدوث أخطاء في التنبؤ اعتمادًا على دقة مجموعة بيانات التدريب. ومن هنا فأن أسلوب تعليم المجموعات يحاول التغلب على هذه المشكلة من خلال تدريب عدة نماذج بالتتابع لتحسين دقة النظام ككل.

2.3. يعتبر أسلوب تعليم المجموعات من أفضل أساليب التنقيب عن البيانات ويمتلك العديد من الخوارزميات وهي تتناسب بشكل جيد مع البيانات الصغيرة والكبيرة على حدا سواء.

3.3. يعتبر التباين والتحيز من اكبر مشاكل خوارزميات تنقيب البيانات والتي تغلبت عليها أساليب تعليم المجموعات من خلال الية عمل محسنة بشكل جيد.

3.3. يعد التنبؤ بأمراض القلب Prediction Disease Heart وتشخيصها التحدي الأكبر في الصناعة الطبية ويعتمد على عدة عوامل في هذا البحث سوف نحاول التعرف عليها.

4. الدرسات السابقة:

1.4. جاءت دراسة مادوميتا بال وسميتا باريجا (2020) للكشف عن مرض قصور القلب باستخدام خوارزمية الغابة العشوائية (Random Forest) واستخدمت الدراسة ثلاثة معايير هي نسبة صحة خوارزمية التصنيف (Accuracy) و مقياس الحساسية (Sensitivity) و مقياس النوعية (Specificity) وقد اعطت Accuracy نسبة 86.9٪ و Sensitivity نسبة 90.6٪ و Specificity نسبة 82.7٪ واستخدم الباحث خصائص تشغيل جهاز الاستقبال ROC وأعطى نسبة 93.3٪.[17]

التعليق على الدراسة: تشابهت الدراستين في استخدام نفس مجموعة البيانات الطبية وكانت نتيجة خوارزمية الغابة العشوائية متقاربة في الدراستين. واختلفت الدراسة الحالية في استخدام عدد اكبر من معايير تقييم الخوارزميات وفي استخدام اكثر من خوارزمية واحدة.

2.4. قارنت دراسة كومبيلا سري شاران ، كولورو إس إن إس ماهيندراناث (2022) بين خمسة خوارزميات تنقيب بيانات هي خوارزمية شجرة القرار والغابة العشوائية والة دعم المتجهات SVM و خوارزمية التعزيز التكيفي AdaBoost وخوارزمية تعزيز التدرج وقد أظهرت خوارزمية الغابة العشوائية اعلى كفاءة بدرجة دقة 92.16٪ للتنبؤ بأمراض القلب. [18]

التعليق على الدراسة: تشابهت الدراسة الحالية في استخدام ثلاثة من خمسة خوارزميات مستخدمة في الدراسة السابقة وتشابهت أيضا في ان الدراسة الحالية كانت خوارزمية الغابة العشوائية أعطت نسبة كفاءة بدرجة 93.41% .

3.4. توصلت دراسة جيان يانغ وجينهان جوان (2022) الي ان خوارزمية XGBoost أعطت اعلى دقة في التنبؤ بامراض فشل القلب حيث أعطت Accuracy نسبة 93.44% و مقياس الدقة (Precision) بنسبة 92.66 و مقياس المثالية (Recall) بنسبة 97.16 و مقياس F (F-measure) بنسبة 94.86. وجاءت خوارزمية الغابة العشوائية في المرتبة الثانية من حيث النتائج واعطت Accuracy نسبة 91.15% و مقياس الدقة (Precision) بنسبة 90.26 و مقياس المثالية (Recall) بنسبة 96.15 و مقياس F (F-measure) بنسبة 90.37. [19]

التعليق على الدراسة: تشابهت الدراسة الحالية والدراسة السابقة في خوارزميتي الغابة العشوائية وخوارزمية التعزيز شديد التدرج (XGBoost) وكانت النتائج متقاربة جدا ولا توجد بينها فروقات تذكر، واختلفت الدراسة الحالية في تفوق الغابة العشوائية وخوارزمية التدرج الخفيف LightGBM على الخوارزميات الاخرى.

4.4. قام ميدل إسلام وآخرون (2022) ، بدراسة تسعة خوارزميات للتنبؤ بامراض القلب منها الغابة العشوائية والانحدار شديد التدرج XGBoost خوارزمية التعزيز التكيفي AdaBoost ومن ثم قام الباحثون باستخدام الخوارزميات التسعة كخوارزمية مكدس (Meta Classifier) استخدم الباحثون ثمانية معايير لتقييم عمل الخوارزميات وكانت مجموعة البيانات مكونة من 12 سمة و 1190 سجل وقد اعطت خوارزمية الغابة العشوائية اعلى معدل لمقياس الحساسية وجاءت في المرتبة الثانية خوارزمية XGBoost.[32]

التعليق على الدراسة: تشابهت الدراسة الحالية مع الدراسة السابقة في استخدام بعض الخوارزمية واغلب معايير تقييم الخوارزميات واختلفت معها في استخدام أسلوب المكدس الذي لم يستخدم في الدراسة الحالية واختلفت أيضا في مجموعة البيانات حيث استخدمت الدراسة السابقة مجموعة بيانات اكبر في عدد السجلات بشكل كبير وتشابهت معها في التركيز على مقياس الحساسية المعني بالاحصاء الطبي. وكانت خوارزمية الغابة العشوائية هي الاعلى نتيجة في الدراستين.

5.4. جاءت دراسة عصام الداود للعمل على مجموعة بيانات ائتمان المنازل التي تحتوي على 219 ميزة و 356251 سجلًا لمقارنة LightGBM و CatBoost و XGBoost وقد أعطت خوارزمية LightGBM اعلى دقة واسرع تنفيذ. [21]

التعليق على الدراسة: اختلفت الدراسة الحالية مع السابقة في مجموعة البيانات وتشابهت في تفوق خوارزمية LightGBM على الخوارزميات الاخرى.

5. المواد والطرق:

تم استخدام أربعة خوارزميات من خوارزميات تعليم المجموعات (Ensemble learning) في تنقيب البيانات للتنبؤ بفشل القلب، هما الغابة العشوائية (RF) وخوارزمية الانحدار شديد التدرج (XGBoost) وخوارزمية تعزيز التدرج الخفيف (LightGBM) وخوارزمية التعزيز التكيفي (AdaBoost). وتم تطبيق هذه الخوارزميات على مجموعة بيانات أمراض القلب مأخوذة من مستودع kaggle باستخدام Google Colab[i] ولغة البرمجة بايثون ومكتبة sklearn، وتحتوي مجموعة البيانات على 304 عينة (سجلات المرضى).

6. مجموعة البيانات:

في هذا البحث سوف نستخدم مجموعة بيانات أمراض القلب تحتوي على 13 سمة و 304 سجل وتحتوي على سمة هدف واحدة تتضمن احدى القيمتين (0,1) للدلالة على الاصابة من عدمها، وتم استخدام 5 سمات للتنبؤ بأمراض القلب. وفيما يلي توضيح للسمات المستخدمة في **الدراسة:**

تم استخدام 5 سمات من مجموع السمات ال13 وهي:

1. العمر (age): عمر المريض [سنوات].

2. ضغط الدم (trestbps): وتمثل ضغط الدم في الحالة الساكنة للشخص [مم زئبق].

3. الكوليسترول (chol): كوليسترول الدم [ملم / ديسيلتر].

4. معدل ضربات القلب (thalach): الحد الأقصى لمعدل ضربات القلب للشخص.

5. انحسارST (oldpeak): يتم قياس الانحسار عادةً بالمليمتر(mm) .

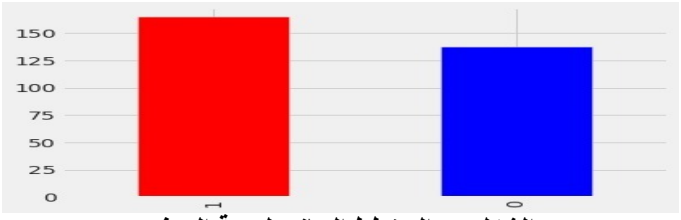
بالإضافة الي سمة target التي تمثل وجود مرض من عدمه.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
0	63	1	3	145	233	1	0	150	0	2.30	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.50	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.40	2	0	2	1

مخطط 1: مجموعة بيانات امراض القلب المستخدمة في الدراسة

7. تصوير البيانات:

مجموعة البيانات تحتوي على السمة الهدف ثنائية (مصاب او غير مصاب) بتعبير ثنائي (1,0) ويمكن تمثيلها مخطط الاعمدة لمعرفة توازن البيانات.

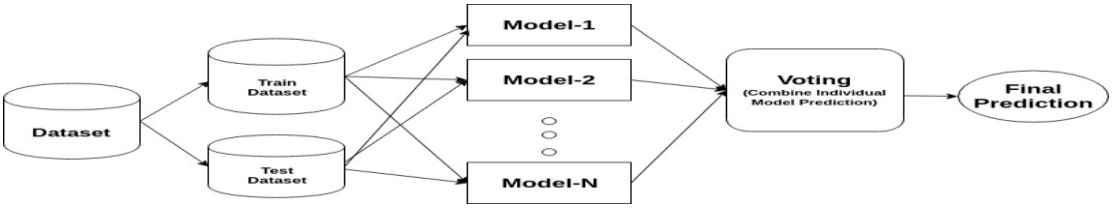


الشكل 1: المخطط البياني لسمة الهدف

لدينا 165 شخصًا يعانون من أمراض القلب و138 غير مصاب بأمراض القلب، لذا فإن البيانات متوازنة. ومجموعة البيانات مثالية لاستخدام حيث انها لا تحتوي على قيم فارغة.

8. التعليم الجماعي (Ensemble Learning):

إن تعليم المجموعات هو نموذج للتعلم الآلي حيث يتم تدريب نماذج متعددة (تسمى غالبًا "المتعلمين الضعفاء") على حل نفس المشكلة وجمعها للحصول على نتائج أفضل. الفرضية الرئيسية هي أنه عندما يتم دمج النماذج الضعيفة بشكل صحيح يمكننا الحصول على نماذج أكثر دقة و / أو قوة، فالغرض الرئيسي من التعليم الجماعي هو قدرته العالية في التعامل مع الإفراط في التجهيز الذي ينتج عنه مشاكل التحيز والتباين[ii].



الشكل 2: المخطط العام لخوارزميات تعليم المجموعات [2]

9. أنواع التعلم الجماعي (Types of Ensemble Learning):

ينقسم التعليم الجماعي الي ثلاثة أنواع هي التكييس (Bagging) والتعزيز(Boosting) والتراص (Stacking) وفي الجدول 2 يبين اهم الفروقات بين الأساليب الثلاثة.

جدول(1): مقارنة أساليب التعليم الجماعي.[5][6][7]

الخوارزمية	التكييس	التعزيز	التراص
تقسيم البيانات	عشوائي	إعطاء العينات المصنفة بشكل خاطئ أفضلية أعلى	مختلف بأختلاف الخوارزميات المستخدمة
الهدف المراد	تقليل التباين	زيادة قوة التنبؤ	على مبدا الاسلوبين
الطرق المستخدمة	فضاء جزئي عشوائي	انحدار متدرج	مزيج من الاسلوبين
الخوارزميات	RF	GBM, XGBoost, LightGBM, CatBoost, etc	خوارزمية او مجموعة من الخوارزميات

1.9. التكييس (Bagging): يعتمد مبدا التكييس على خوارزمية الغابة العشوائية. وتكون لبنة البناء الأساسية للغابة العشوائية مستوحاة من شجرة القرار(Classification and Regression Trees) ، وهي إحدى طرق تعلم الآلة لبناء نماذج تنبؤ من البيانات، إذ يتم الحصول على النماذج من خلال تقسيم البيانات وبناء نموذج بسيط للتنبؤ داخل كل قسم.

ويعتمد مبدا التقسيم في شجرات القرار على نوعين هما:

- مبدا مؤشر جيني (Gini) والتي يتم حسابه باستخدام الصيغة التالية:

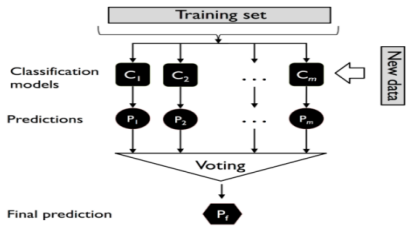
$$Gini\ Index = 1 - \sum_j p_j^2$$

حيث p_j هو احتمال الفئة j. يتم اختيار التقسيم الأمثل بواسطة الميزات ذات مؤشر جيني الأقل.

- مبدا الإنتروبيا (Entropy) والتي يتم حسابها باستخدام الصيغة التالية:

$$\text{Entropy} = -\sum_j p_j . \log_2 . p_j$$

حيث p_j هو احتمال الفئة j. يتم اختيار التقسيم الأمثل على غرار مؤشر جيني بواسطة الميزات ذات مؤشر الانتروبيا الأقل. ويكمن الاختلاف بينهما من الناحية الحسابية، تعتبر الإنتروبيا أكثر تعقيدًا لأنها تستخدم اللوغاريتمات، وبالتالي سيكون حساب مؤشر جيني أسرع من حيث المعالجة.



الشكل 3 : خوارزمية الغابة العشوائية

1.2.9. التعزيز: تجمع خوارزميات التعزيز بين كل متعلم ضعيف لإنشاء قاعدة تنبؤ واحدة قوية. لتحديد القاعدة الضعيفة، توجد خوارزمية تعلم أساسية (التعلم الآلي). عندما يتم تطبيق الخوارزمية الأساسية، فإنها تنشئ قواعد تنبؤ جديدة باستخدام عملية التكرار. بعد عدد معين من التكرار، فإنه يجمع كل القواعد الضعيفة لإنشاء قاعدة توقع واحدة.

الية عمل خوارزميات التعزيز:

تكون الخطوات الرئيسية لمبادئ التعزيز في تعليم الالة هي:

1.1.2.9. يتم ملاءمة شجرة القرار للبيانات $F1(x) = y$ حيث ان y هو القيمة المتوقعة او التسمية (Label).

2.1.2.9. ملاءمة شجرة القرار التالية مع Residual من الشجرة السابقة $h1(x) = y - F1$. مع العلم ان المتبقي لكل ملاحظة هو الفرق بين القيم المتوقعة لـ y (المتغير التابع) وقيم y المرصودة.

$$Residual = actual\ y\ value - predicted\ y\ value$$

3.1.2.9. أضافة الشجرة الجديدة إلى الخوارزمية $F2(x) = F1(x) + h1(x)$.

4.1.2.9. أضافة الشجرة الجديدة إلى الخوارزمية $F3(x) = F2(x) + h2(x)$

5.1.2.9. الاستمرار في هذه العملية حتى تخبرنا آلية ما بالتوقف (مثل التحقق المتقاطع... الخ).[18]

الخوارزمية بشكلها النهائي تعتبر كنموذج مضاف متدرج وبأعتبار ان b هي الأشجار الفردية

$$f(x) = \sum_{b=1}^{B} f^b(x) \quad [13]$$

10. أنواع خوارزميات التعزيز:

هناك ثلاثة أنواع من خوارزميات التعزيز وهي كما يلي:

1.10. التعزيز التكيفي (AdaBoost): كان التعزيز التكيفي من أوائل نماذج التعزيز التي تم تطويرها. حيث يتكيف ويحاول إجراء تصحيح ذاتي في كل تكرار لعملية التعزيز.

2.10. التّعزيز المتدرج (Gradient Boost): يشبه التعزيز التكيفي في التدريب بالتتابع. ويكمن الاختلاف بينهما في أن التّعزيز المتدرج لا يمنح العناصر المصنفة بشكل خاطئ وزنًا أكبر. وبدلاً من ذلك يحسن وظيفة الخسارة عبر إنتاج متعلمين أساسيين بالتتابع بحيث يكون المتعلم الأساسي الحالي أكثر فعالية من المتعلم السابق.

3.10. التعزيز شديد التدرج (XGBoost): يحسن التعزيز شديد التدرج التعزيز المتدرج لتحقيق السرعة والتوسع الحوسبي بطرق عديدة. يستخدم XGBoost أنوية متعددة على وحدة المعالجة المركزية بحيث يمكن للتعلم أن يحدث بالتوازي أثناء التدريب.

1.1.10. الية عمل خوارزمية التعزيز التكيفي AdaBoost:

يختلف أسلوب التدريب بناءً على نوع خوارزمية التعزيز ومع ذلك فإن الخوارزميات بشكل عام تتخذ الخطوات العامة الآتية لتدريب نموذج التعزيز:

1. تعين خوارزمية التعزيز وزنًا متساويًا لكل عينة من البيانات. وتغذي النموذج الأول والتي يطلق عليها الخوارزمية الأساسية بالبيانات لتجري تنبؤات لكل عينة من البيانات.

2. تقيم خوارزمية التعزيز تنبؤات النموذج وتزيد من قيمة وزن العينات التي صنفت بشكل خاطئ. كما تعين وزنًا بناءً على أداء النموذج. النموذج الذي يخرج تنبؤات جيدة سيكون له قدر عالٍ من التأثير في القرار النهائي.

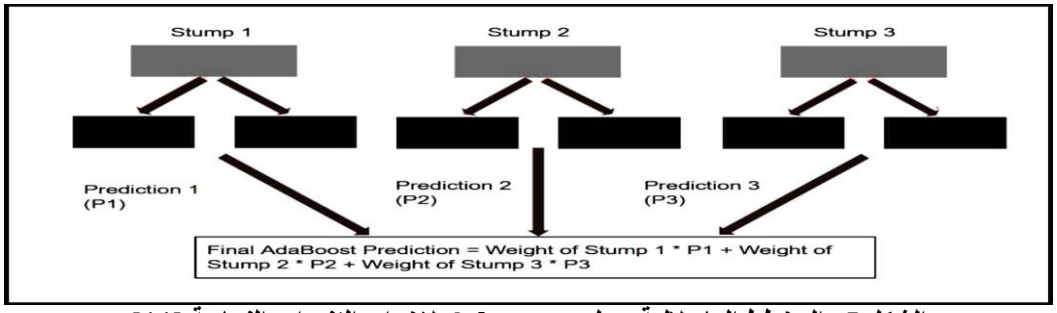
3. تمرر الخوارزمية البيانات الموزونة إلى شجرة القرار التالية.

4. تكرر الخوارزمية الخطوتين 2 و3 حتى تصبح مثيلات الأخطاء في التدريب أقل من مستوى معين. [13]

1. Basic Algorithm for Boosting:
Initialize: set all examples to have equal weights
2. For each t = 1, …,T ,
3.Learn a hypothesis h_t from weighted examples
4.Decrease weights of examples h_t classifies correctly
5.Calculateα_t, the weight of the current weak learner, h_t
6. $\boldsymbol{Return}\ h(x) = \sum_{t=1}^{T} \alpha_t\ h_t(x)$

الشكل 4: الخوارزمية الأساسية للتعزيز

الخوارزمية الأكثر ملاءمة والأكثر شيوعًا المستخدمة مع AdaBoost هي أشجار القرار بمستوى واحد. نظرًا لأن هذه الأشجار قصيرة جدًا وتحتوي على قرار واحد فقط للتصنيف، فغالبًا ما يطلق عليها اسم جذوع القرار (Decision Stump) وهي عقدة ذات ورقتين وكل جذع هي سمة او خاصية من خصائص بيانات التدريب.[14]



الشكل 5: المخطط العام لالية عمل AdaBoost لاخراج التنبؤات النهائية.[11]

الخطوة الأولى(إنشاء المتعلم الأساسي الأول): لإنشاء المتعلم الأول تأخذ الخوارزمية الميزة الأولى وتنشئ الجذع الأول F1. سيتم أنشاء عدد جذوع الأشجار بنفس عدد الميزات. تسمى هذه العملية بنموذج المتعلم الأساسي لجذوع الاشجار. يتم حساب Gini أو Entropy بنفس الطريقة التي يتم بها حسابهما لأشجار القرار. سيكون الجذع الأقل قيمة هو المتعلم الأساسي الأول. ويتم تهئية الاوزان بشكل مبدئي في الخطوة الاولى باستخدام الصيغة:

$$[10][15][16]\text{Weight}(x_i) = 1/n$$

حيث x_i بيانات التدريب و n هو عدد بيانات التدريب.

2. الخطوة الثانية (حساب الخطأ الإجمالي (Total Error)): إجمالي الخطأ هو مجموع كل الأخطاء في السجل المصنف بشكل خاطئ باستخدام وزن السجل.

$$\text{Error} = \text{Misclassification Count} * \text{Sample Weight}$$

$$[10][15][16]\ \text{Error} = \text{Misclassification Count} * \frac{1}{N}$$

3. الخطوة الثالثة (حساب أداء الجذع (Calculating Performance of Stump)): يمكن حساب أداء الجدع من خلال المعادلة:

$$\textbf{[10][15][16]}\alpha_t = \frac{1}{2}\log\frac{(1-\text{Total Error})}{\text{Total Error}}$$

حيث ان α_t هو Performance of Stump او ما يعرف بـ Amount of Say و log هو اللوغاريثم الطبيعي و Total Error هو إجمالي الخطأ. وتأتي أهمية هذه الخطوة لتحديث وزن العينة قبل الانتقال إلى النموذج أو المرحلة التالية لأنه إذا تم تطبيق نفس الوزن، فسيكون الناتج المستلم من النموذج الأول.

4. الخطوة الرابعة (تحديث الأوزان): بالنسبة للسجلات المصنفة بشكل غير صحيح، فإن صيغة تحديث الأوزان هي:

$$\text{New Sample Weight} = \text{Sample Weight} * e^{\text{Performance}}$$ [10][15] [16]

حيث ان وزن العينة الجديد = وزن العينة السابقة مضروب في الدالة الاسية لمقياس الأداء. بالنسبة للسجلات المصنفة بشكل صحيح، نستخدم نفس الصيغة مع كون قيمة الأداء سالبة. يؤدي هذا إلى تقليل وزن السجلات المصنفة بشكل صحيح مقارنة بالسجلات المصنفة بشكل غير صحيح. الصيغة هي:

$$\text{New Sample Weight} = \text{Sample Weight} * e^{-\text{Performance}}$$

5. الخطوة الخامسة (تطبيع الاوزان): يجب تسوية الاوزان (Normalize weights) وذلك بقسمة كل وزن محدث من الاوزان على مجموع الاوزان المحدثة عندما لا يتساوى مجموع الاوزان 1.

$$\text{Weight}_n = \frac{\text{Weight}}{\sum \text{Weight}} \text{ where Total Update Weight} < 1$$ [10][15][16]

حيث ان Update Weight(i) هو وزن العينة الحالية بعد التحديث.

7. كرر الخطوات من 2 إلى 5 ، حتى يصبح معدل الخطأ أقل من مستوى الحد أو الوصول الي نهاية التكرار T.

8. يتم الحصول على المنصف القوي من خلال الصيغة:

$$h(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$$ [15][16][9]

1.2.10. خوارزمية تعزيز التدرج (Gradient Boosting):

يمكن القول ان AdaBoost عبارة عن خوارزمية معززة تقوم بتعيين أوزان لنقاط البيانات بناءً على أخطاء المتعلمين الأساسيين، في حين أن Gradient Boosting عبارة عن خوارزمية معززة تدرب المتعلمين الأساسيين على التدرج السلبي لوظيفة الخسارة. ويمكن أن يكون المتعلم الأساسي في Gradient Boosting بمثابة شجرة قرار بأي عمق على عكس AdaBoost ذات شجرة بعمق واحد.

دالة الخسارة: يكون مبدا الانقسام في شجرات القرار غالبا مبني على مبدا الانتروبيا و ربح المعلومات (Entropy and information gain) ولكن من اهم عيوب شجرات القرار هو الانحياز للصفات التي تحتوي على مستويات أكثر للبيانات التي تتضمن متغيرات فئوية مختلفة المستويات. ومن هنا جاء التعزيز المتدرج لتحسين شجرات القرار باستخدام دالة الخسارة.

يتم تقييم مدى جودة أداء الشجرة في التعزيز المتدرج باستخدام دالة الخسارة. يعد الانتروبيا المتقاطع (Cross entropy) خيارًا شائعًا للتصنيف متعدد الفئات. وتكون الصيغة العامة لها بالشكل التالي حيث p هي التسمية (label) و q هي التنبؤ (Prediction).

$$\text{Loss}(p,q) = -\sum p(x) \log q(x) \ldots \text{where x equal Output Classes}$$ [12]

بشكل أساسي تكون الخسارة عالية عندما لا تتوافق التسمية (label) مع التنبؤ (Prediction) وتكون الخسارة صفرًا عندما يكونان في اتفاق تام. بعد الشجرة الأولى وحساب وظيفة الخسارة لتقييم النموذج، تضاف شجرة ثانية لتقليل الخسارة مقارنة بالشجرة الأولى. رياضيا، يتم إعطاء هذا من خلال المشتقة السلبية للخسارة (منحدر متدرج)iii فيما يتعلق بمخرجات النموذج السابق. وتكون الصيغة كالتالي:

$$\text{Boosted Ensemble} = \text{First Tree} + \eta * \text{Second Tree}$$

where η = Learing Rate and Loss(Boosted Ensemble) < Loss(First Tree) [12]

حيث η قيمة افتراضية معنية بمعدل التعلم لتقليص الخطوة المستخدمة في التحديث لمنع فرط التجهيز. فبعد كل خطوة من خطوات التعزيز يمكن الحصول على اوزان الميزات الجديدة مباشرة.[12]

ومن خلال النزول المتدرج وقانون تايلر ومنالصيغة الرياضية للمعادلة 1 يمكن الحصول على F(2) من الصيغة:

$$F(2) = F(1) + \eta * \text{Second Tree}$$

$$\text{where Second Tree} = -\frac{\partial L}{\partial(F(1)} = \frac{\partial \text{Loss}}{\partial \text{ Previous Model's Output}}) \quad [12]$$

وتكون الصيغة العامة لجميع حالات التعزيز هي:

$$F(m) = F(m-1) + \eta * \frac{\partial L}{\partial F(m-1)} \quad [12]$$

2.2.10. الانحدار شديد التدرج (XGBoost):

تكمن الاختلافات بين الخوارزميتين Gradient Boosting و XGBoost:

1.2.2.10. السرعة: يعد XGBoost أسرع من تعزيز التدرج التقليدي، خاصة على مجموعات البيانات الكبيرة، نظرًا لانه يستخدم الحوسبة المتوازية وخوارزميات محسّنة لاكتشاف الانقسام وبناء الأشجار.

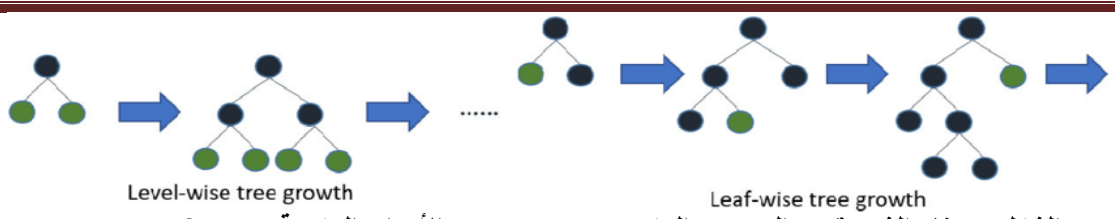
2.2.2.10. التنظيم: يحتوي XGBoost على تسوية مدمجة في شكل تسوية L1 و L2، والتي يمكن أن تساعد في منع الإفراط في التجهيز (overfit).

3.2.2.10. ميزات إضافية: يوفر XGBoost ميزات إضافية مثل التوقف المبكر والتحقق المتبادل، مما يسهل تدريب النماذج وضبط المعلمات الفائقة.

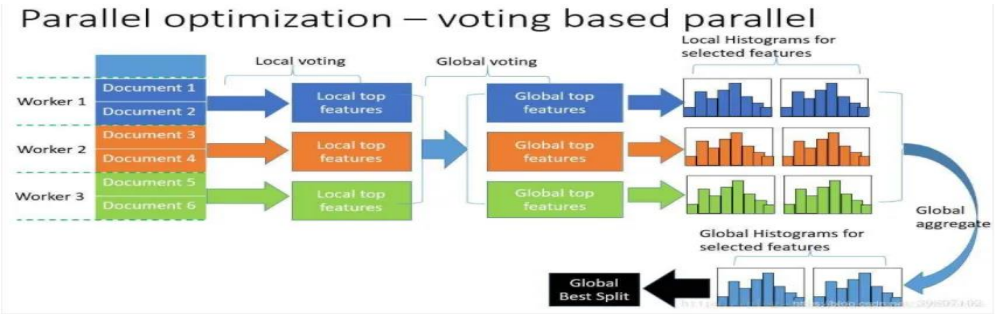
4.2.2.10. قابلية التوسع: XGBoost قابل للتطوير بدرجة كبيرة ويمكنه التعامل مع مجموعات البيانات الكبيرة، مما يجعله مثاليًا للحوسبة الموزعة.[27]

3.2.10. خوارزمية تعزيز التدرج الخفيف (LightGBM):

طورت الخوارزمية من قبل فريق Microsoft في أبريل 2017 لتقليل وقت التنفيذ. وهي متشابهة مع خوارزمية الانحدار الشديد XGBoost. ويكمن الاختلاف في ان خوارزمية LightGBM تعمل على مبدأ الأفق أولا (best-first) او الورقة الحكيمة (leaf-wise) على عكس خوارزمية XGBoost التي تعمل بمبدأ العمق أولا (depth-first) او المستوى الحكيم (level-wise). ونظرا لان الاسلوبين يعملان على نفس الشجرة. ولان الشجرة لا تزرع بالعمق الكامل و بتطبيق معايير التوقف المبكر وطرق التقليم يمكن أن يؤدي إلى أشجار مختلفة تمامًا. و لأن الأوراق تختار الانقسامات بناءً على مساهمتها في الخسارة الكلية وليس فقط الخسارة على طول فرع معين، فغالبًا (ليس دائمًا) ستتعلم الأشجار ذات الخطأ الأقل "أسرع" من المستوى الحكيم. وتعتبر خوارزمية LightGBM أسرع من XGBoost، خاصة على مجموعات البيانات الكبيرة، نظرًا لاستخدامها نهج تحسين الذاكرة المستند إلى الرسم البياني لميزات الحاوية المستمرة (histograms to bin continuous features) والتي لها دور كبير في سرعة الخوارزمية واستخدام اقل للذاكرة بدلا من استخدام كل القيم للسمات، مع استخدامها خوارزميات اكتشاف الانقسام الأسرع. بالاضافة الي استخدامها للمعالجة المتوازية GPU وتعدد الانوية. الشكل () يوضح اختلاف مبدأ المستوى الحكيم والاوراق الحكيمة.[24][21] [22] [23]



الشكل 6: بناء الشجرة من المستوى الحكيم (level-wise) والأوراق الحكيمة (leaf-wise). [21]



الشكل 7: نهج تحسين الذاكرة المستند إلى الرسم البياني لميزات الحاوية المستمرة [20]

11. معايير تقييم كفاءة خوارزميات التصنيف:

بعد بناء احد نماذج خوارزميات التصنيف يتم تطبيق عدة معايير لتاكد من صحة التصنيف ونتائج الخوارزمية ومنها:

1.11. نسبة صحة خوارزمية التصنيف (Accuracy):

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \frac{TP+TN}{P+N} \quad [27][30]$$

حيث ان:

TP = (True Positives) السجلات التي تم التنبؤ بفئتها بشكل صحيح.

TN = (True Negatives) السجلات التي تم التنبؤ بها لو رفضها ضمن الفئة بشكل صحيح.

FP = (False Positives) وهي السجلات السالبة التي تنبأ بها النموذج بشكل خاطئ والحقها بالفئة الموجبة. FN = (False Negatives) وهو عدد السجلات الموجبة التي تم اعتبارها بالخطأ سالبة مع أنها موجبة.[25] 2.11. مقياس المثالية (Recall): هي نسبة السجلات الموجبة التي تم التنبؤ بها وتصنيفها بشكل صحيح الي كل السجلات الموجبة.[1]

$$\text{Recall} = \frac{TP}{TP+FN} \quad [27][26]$$

3.11. مقياس الدقة (Precision):

$$\text{Precision} = \frac{TP}{TP+FP} \quad [31][26]$$

4.11. معامل Cohen Kappa: تستخدم نقاط Cohen Kappa لمقارنة الاصناف المتوقعة من نموذج بالتسميات الفعلية في البيانات. تتراوح النتيجة من −1 (أسوأ أداء ممكن) إلى 1 (أفضل أداء ممكن). تعني درجة كوهين كابا 0 أن النموذج ليس أفضل من التخمين العشوائي، والنتيجة 1 تعني أن النموذج مثالي. ويمكن تعريف معامل Cohen Kappa بشكل عام انه:

$$K = \frac{p_o - p_e}{1 - p_e} \quad [28]$$

حيث ان: $p_o = \frac{(TP+TN)}{N}$ [28]

نحتاج الان إلى حساب الاحتمال المتوقع بأن كلا المقيّمين متفقان بالصدفة. يتم حساب ذلك بضرب الاحتمال المتوقع بأن كلا المقيّمين متفقون على أن الفئات موجبة ، والفئات سالبة من خلال المعادلة التالية:[28]

$$pe = \left[\frac{pe(\text{rater 1 says Yes})}{N} * \frac{pe(\text{rater 2 says Yes})}{N} + \frac{pe(\text{rater 1 says No})}{N} * \frac{pe(\text{rater 2 says No})}{N}\right] \text{[28]}$$

5.11. مقياس F (F-measure):

$$\text{[25][26]}F1 = \frac{2*Precision*recall}{Precision*recall}$$

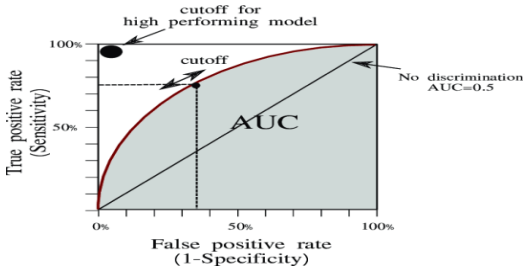
6.11. مقياس الحساسية (Sensitivity):

$$\text{[25]}Sensitivity = \frac{TP}{TP+FN}$$

7.11. مقياس النوعية (Specificity):

$$\text{[29]}Specificity = \frac{TN}{TN+FP}$$

8.11. المنطقة الواقعة تحت منحنى ROC: يوضح منحنى ROC العلاقة بين معدل إيجابي كاذب (ويعرف أيضًا باسم (FPR)و المعدل الإيجابي الحقيقي المعروف أيضًا باسم (TPR) عبر مختلف الحدود القصوى.



الشكل 8: المخطط العام لمنحنى ROC

12. النتائج: تم نقسيم البيانات الي 70% للتدريب و 30% للاختبار، في بيانات التدريب تم الحصول على نتائج التدريب حسب الجدول 2:

جدول 2: نتائج بيانات التدريب

TP	FP	FN	TN	Accuracy	Recall	Cohen's kappa(k)	F-measure	Precision	specificity	Sensitivity	ROC
Random Force Algorithm											
76	21	11	104	84.91	90.43	69.34	86.67	83.20	78.35	90.43	91.91
XGBoost Algorithm											
79	18	10	105	86.79	91.30	73.22	88.24	85.37	81.44	91.30	95.97
AdaBoost Algorithm											
95	2	3	112	97.64	97.39	95.25	97.82	98.25	97.94	97.39	99.56
LightGBM Algorithm											
81	16	10	105	87.78	91.30	75.17	88.98	86.78	83.51	91.30	93.62

وفي بيانات الاختبار تم الحصول على نتائج الاختبار حسب الجدول 3:

جدول 3: نتائج بيانات الاختبار

TP	FP	FN	TN	Accuracy	Recall	Cohen's kappa(k)	F-measure	Precision	specificity	Sensitivity	ROC
Random Force Algorithm											
36	5	1	49	93.41	98.00	86.56	94.23	90.74	87.80	98.00	95.85
XGBoost Algorithm											
33	8	2	48	89.01	96.00	77.51	90.57	85.71	80.49	96.00	94.22
AdaBoost Algorithm											
31	10	6	44	82.42	88.00	64.17	84.62	81.48	75.61	88.00	90.10
LightGBM Algorithm											
36	5	4	46	90.11	92.00	79.71	91.09	90.20	87.80	92.00	95.71

13. المناقشات

من خلال النتائج تبين ان النتائج تعتبر جيدة ولكن خوارزمية AdaBoost أعطت معدل خطأ في التدريب بمقدار 2.36% وهذا يعتبر تدريبا جيدا، ولكنها أعطت معدل خطأ كبير نسبيا في بيانات الاختبار بمقدار 17.85% مما يعني حدوث overfitting للنموذج وضعف النموذج في التعميم للتعامل مع البيانات الغير مرئية (الجديدة). في حين ان خوارزمية الغابة العشوائية أعطت مقدار خطا في بيانات التدريب بمقدار 15.1% واعطت في بيانات الاختبار مقدار خطأ 6.59% مما يعني ان النموذج جيد وقادر على التعميم.

تعتبر Accuracy بمثابة أدارك الخوارزمية وقد كانت خوارزمية الغابة العشوائية هي الاعلى بنسبة 93.41% وتعتبر هذه النسبة جيدة جدا بالمقارنة مع مستوى منحنى ROC الذي اعطى نسبة 95.85% وجاءت بعدها خوارزمية LightGBM بنسبة 90.11% مع مستوى لمنحنى ROC بنسبة 95.71%. وهذا التقارب بين مقياس الدقة (Accuracy) ومستوى منحى ROC يعطى دلالة ان النموذج مدرك بشكل ممتاز.

في التعليم الطبي (الإحصاء الطبي) يعد المقايسيين specificity و Sensitivity ذو الأهمية في تقييم دقة عمل الخوارزمية وقد كانت اعلى نسبة لخوارزمية الغابة العشوائية في مقياس الحساسية (Sensitivity) هي 98.00% وكان مقياس الخصوصية (Specificity) بنسبة 87.80% وهو الاعلى بين بقية الخوارزميات.

من الواضح ان خوارزمية الغابة العشوائية هي الامثل في التعامل مع البيانات الطبية وتشخيص الامراض، ولكن يظل هناك حاجة إلى مزيد من البحث لتأكيد هذه النتائج، ولاستكشاف استخدام هذه الخوارزميات في التطبيقات الطبية الأخرى.

14. الخاتمة:

في الختام، تقترح دراستنا أن LightGBM و RF هما خوارزميات فعالة للتنبؤ بأمراض القلب. تحقق هذه الخوارزميات دقة وأداء أعلى مقارنةً بـ AdaBoost و XGBoost، وقد توفر أداة مفيدة للكشف المبكر عن أمراض القلب والوقاية منها. هناك حاجة إلى مزيد من البحث لتأكيد هذه النتائج واستكشاف استخدام هذه الخوارزميات في التطبيقات الطبية الأخرى. ولكن في البيانات الطبية بقدر أهمية النتائج التي اعطها النموذج تكمن أهمية ادراك النموذج للبيانات بشكل جيد، لان البيانات الطبية لا يمكن تقبل الاخطاء لما تمثل من مخاطر على حياة الاخرين.

15. الاعمال المستقبلية:

لتحسين نتائج هذا البحث نوصي ببعض الاعمال المستقبلية وهي:

1.15. إجراء تجارب على مجموعة أكبر من البيانات لتحسين موثوقية النتائج.

2.15. استخدام تقنيات أخرى لتحسين نتائج التصنيف والتنبؤ بأمراض القلب، مثل تقنيات التعلم العميق والشبكات العصبية الاصطناعية.

3.15. استخدام متغيرات أخرى غير البيانات السريرية لتحسين نتائج التصنيف، مثل العوامل الوراثية والعوامل البيئية ونمط حياة الأشخاص.

4.15. يمكن استخدام أسلوب المكدس في البحوث المستقبلية لتحسين النتائج وزيادة الدقة والموثوقية في تنبؤ الأمراض.

5.15. يمكن استخدام أسلوب المكدس لدمج نتائج الخوارزميات المختلفة التي تم استخدامها في البحث الحالي، مثل Random Forest و XGBoost و AdaBoost وLightGBM ، واستخدام النتائج كمدخلات لخوارزمية المكدس لتحسين دقة التنبؤ بأمراض القلب.

Comparison of Ensemble learning algorithms for predicting heart disease

Bader N. Awedat
Iinformation Technology - Azzaytuna University
Bader_najep@yahoo.com

Abstract: Ensemble learning is a general approach to describing machine learning that seeks better predictive performance by combining predictions from multiple models. The group learning method includes a number of methods, including Bagging and Boosting, and these two methods have a set of algorithms, including Random Force algorithm, AdaBoost adaptive gradient algorithm, and Gradient Boosting algorithms. In this research, we will compare the two methods of conditioning and reinforcement in terms of the percentage of accuracy of the classification algorithm (Accuracy), the scale of perfection (Recall), the scale of accuracy (Precision), the Cohen Kappa coefficient, the F-measure, the Sensitivity scale, and the quality scale (Specificity) and the level of the area under the ROC curve in predicting heart failure disease.

Keywords: gradation enhancement machines, soft gradient enhancement machine, max gradient enhancement, accuracy ratio scale, ideality scale, precision scale, group marking.

16. المراجع:

1. Gautam Kunapuli, (2020), Ensemble Methods for Machine Learning, Manning Publications, ISBN-13: 9781617297137.

. Julia Gastinger, S´ebastien Nicolas, Du˘sica Stepi´c, Mischa Schmidt, Anett Sch¨ulke, 2 (2021), A study on Ensemble Learning for Time Series Forecasting and the need for Meta-Learning, International Joint Conference on Neural Network, DOI:10.1109/IJCNN52387.2021.9533378

. Unknown, (2022), Ensemble Learning and Ensemble Learning Techniques, Analytics 3 Vidhya, date access 26-9-2022, direct access: https://courses.analyticsvidhya.com/courses/ensemble-learning-and-ensemble-learning-techniques?utm_source=blog&utm_medium=boosting-algorithms-simplified.

4. Kyle D Peterso, (2018), Resting Heart Rate Variability Can Predict Track and Field Sprint Performance, OA Journal-Sports, Volume 1.

5. Unknown, (2022), Bootstrap aggregating, From Wikipedia, the free encyclopedia, Date Access 26-8-2022, direct access:

.https://en.wikipedia.org/wiki/Bootstrap_aggregating

6. Unknown, (2022), Optical character recognition, From Wikipedia, the free encyclopedia, Date Access 26-8-2022, direct access:

. https://en.wikipedia.org/wiki/Optical_character_recognition

7. Sajid Nagi, Dhruba Kr. Bhattacharyya, (2013), Classification of microarray cancer data using ensemble approach, Network Modeling Analysis in Health Informatics and Bioinformatics, volume 2, pages 159–173.

8- Aarshay Jain, (2022), Complete Machine Learning Guide to Parameter Tuning in Gradient Boosting (GBM) in Python, Analytics Vidhya, date Access 17-8-2022, direct access: https://www.analyticsvidhya.com/blog/2016/02/complete-guide-parameter-tuning-gradient-boosting-gbm-python/.

13- مجهول، (2022)، ما المقصود بالتعزيز، Amazon Web Services، تاريخ الوصول 26-9-2022، الرابط المباشر:

.https://aws.amazon.com/ar/what-is/boosting/

9- Satish Gunjal, (2020), Ensemble Learning: Bagging, Boosting & Stacking, Kaggle, Data Access 18-9-2022, Direct Link: https://www.kaggle.com/code/satishgunjal/ensemble-learning-bagging-boosting -stacking/notebook.

10- Jason Brownlee, (2021), Ensemble Learning Algorithms With Python: Make Better Predictions with Bagging, Boosting, and Stacking, Machine Learning Mastery.

11- Andrew William, (2021), A Comprehensive Mathematical Approach to Understand AdaBoost, Towards Data Science, Date Access 18-9-2022, Direct Access: https://towardsdatascience.com/a-comprehensive-mathematical-approach-to-understand-adaboost-f185104edced.

12. Cheshta Dhingra, (2020), A Visual Guide to Gradient Boosted Trees (XGBoost), Towards Data Science, Date Access 20-9-2022, Direct Access: https://towardsdatascience.com/a-visual-guide-to-gradient-boosted-trees-8d9ed578b33.

13. Bradley Boehmke, Brandon Greenwell, (2019), Hands-On Machine Learning with R, CRC Press. 1st Edition, New York.

14. Ashish Kumar, 2022, The Ultimate Guide to AdaBoost Algorithm, Great Learning, Date Access 23-9-2022, Direct Access: https://www.mygreatlearning.com/blog/adaboost-algorithm/.

15. Gajendra, AdaBoost Classifier: Understanding AdaBoost Classifier, Medium, Date Access 29-9-2022, Direct Access: https://medium.com/@gajendra.k.s/adaboost-classifier-e43bc88ecc07.

16. Peng Zhang, 2021, AN OPTIMIZED ADABOOST ALGORITHM BASED ON K-MEANS CLUSTERING, Journal of Physics Conference Series, first volume, DOI:10.1088/1742-6596/1856/1/012021.

17. Madhumita Pal, Smita Parija, 2020, Prediction of Heart Diseases using Random Forest, ICCIEA 2020 , IOP Publishing, Journal of Physics: Conference Series , doi:10.1088/1742-6596/1817/1/012009.

18. Kompella Sri Charan, Kolluru S S N S Mahendranath, (2022), Heart Disease Prediction Using Random Forest Algorithm, International Research Journal of Engineering and Technology (IRJET), Volume 9, Issue 3.

19. Jian Yang, Jinhan Guan, 2022, A Heart Disease Prediction Model Based on Feature Optimization and Smote-Xgboost Algorithm, Information, Volume 13, Number 475.

20. Summer Hu, 2021, Run through LightGBM Fast Training Techniques, date Access 15-1-2023, Link Access: https://medium.com/swlh/understand-lightgbm-fast-training-techniques-8dab16487cd5.

21. Essam Al Daoud, 2019, Comparison between XGBoost, LightGBM and CatBoost Using a Home Credit Dataset, World Academy of Science, Engineering and Technology, International Journal of Computer and Information Engineering, Vol:13, No:1

22. Liang, W., Luo, S., Zhao, G., & Wu, H. (2020), Predicting Hard Rock Pillar Stability Using GBDT, XGBoost, and LightGBM Algorithms. Mathematics, 8(5), 765. doi:10.3390/math8050765

23. Stephanie Bourdeau,2019, Deciding on How to Boost Your Decision Trees, Medium, date Access 15-1-2023, Link Access:

https://medium.com/@stephkendall/deciding-on-how-to-boost-your-decision-trees-.1ea5412c0fe7

24. Mingming Zhaoa, Jianguo Zhoub, Zifeng Wuc, Wenyu Pengd, Wei Zhoue, Yu Liang, 2020, Exploring the H2H genes in 3D v, IOP Conf. Series: Earth and Environmental Science 440 (2020) 042079, doi:10.1088/1755-1315/440/4/0420

25. Powers, David M W, Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation, Journal of Machine Learning Technologies, 2008.

26. Nabeela Ashraf1, Waqar Ahmad2, Rehan Ashraf3, A Comparative Study of Data Mining Algorithms for High Detection Rate in Intrusion Detection System, Annals of Emerging Technologies in Computing (AETiC) Vol.2, No.1, 2018.

27. Lidet Tefera, Precision and recall, Medium, Addis Ababa, 2020, date access 25-6-2022, direct link: https://medium.com/@lidetsal/precision-and-recall-30fd346cf90a.

28. Ajitesh Kumar, (2022), Cohen Kappa Score Python Example: Machine Learning, Data Analytics, Data Analytics, date access 1-6-2022, direct link:

https://vitalflux.com/cohen-kappa-score-python-example-machine-learning/

29. Ajitesh Kumar, (2022), Machine Learning – Sensitivity vs Specificity Difference , Machine Learning, Data Analytics, date access 1-6-2022, direct link:

https://vitalflux.com/ml-metrics-sensitivity-vs-specificity-difference/

30. Shruti Shishir Gosavi, (2018), A Comparison of Data Mining Classifiers in Weka, International Journal of Creative Research Thoughts (IJCRT), Volume 6, Issue 1, 2018 | ISSN: 2320-2882

31. Muhammad sakib khan inan, Istiakur rahman, (2022), integration of explainable artificial intelligence to identify significant landslide causal factors for extreme gradient boosting based landslide susceptibility mapping with improved feature selection, machine learning applied to geo-technical engineering, arxiv, v1.

32. Md. Maidul Islam, Tanzina Nasrin Tania, Sharmin Akter, Kazi Hassan Shakib, (2022), An Improved Heart Disease Prediction Using Stacked Ensemble Method, CC BY-NC-ND 4.0, DOI:10.13140/RG.2.2.16442.47044.

---

i يعد Google Colab نسخة سحابية محسنة من Jupyter Notebook المخصص لكتابة وتشغيل الشيفرات البرمجية ومستندات Notebook، وذلك من خلال بيئة برمجية متكاملة أو مستعرض ويب. يوفر Google Colab وصولًا مجانيًا إلى وحدات معالجة الرسومات GPU وTPU، والتي تستخدم لبناء نموذج التعلم الآلي أو التعلم العميق.

ii النحيز (Bias) :يقيس درجة الانحراف بين الناتج المتوقع لخوارزمية التعلم والنتيجة الحقيقية ، ويميز القدرة الملائمة للخوارزمية ، ويشير التحيز المرتفع إلى أن وظيفة التنبؤ مختلفة تمامًا عن النتيجة الحقيقية. والتبيان (Variance): يمثل الفرق بين "النماذج المدربة بواسطة مجموعات بيانات تدريبية مختلفة من نفس الحجم" و "قيم المخرجات المتوقعة لهذه النماذج". تؤدي تغييرات مجموعة التدريب إلى تغييرات في الأداء. يشير التباين العالي إلى أن النموذج غير مستقر للغاية. يمكن قياس الخطأ بين القيمة المتوقعة والقيمة الحقيقة، ثم جمعها وحساب متوسطها. وبذلك يمكن الوصول إلى إنحياز نموذج التنبؤ. فإذا كان الإنحياز يساوي 0، ذلك يعني أن النموذج يتنبؤ بكل شئ بشكل صحيح (بدون تباين) أو أن النموذج يتنبؤ بشكل خاطئ لكل القيم (تباين عالي). فالعلاقة بين التحيز والتباين علاقة عكسية.

iii يعتمد أصل التدرج على الملاحظة أنه إذا كانت الدالة متعددة المتغيرات F (x) معرفة وقابلة للاشتقاق في جوار النقطة a فإن F (x) ينخفض بسرعة أكبرإذا ذهبنا من a في اتجاه التدرج السلبي لـ F في $a, -\nabla F(a)$. ويتبع ذلك أنه اذا كان $a_{n+1} = a_n - \gamma \nabla F(a_n)$ فإنه لكل $\gamma \in R_+$ صغيرة بما يكفي $F(a_n) \geq F(a_{n+1})$. وبعبارة أخرى الحد $\nabla F(a_n)$ يطرح من a لأننا نريد التحرك ضد التدرج نحو الحد الأدنى المحلي. (من ويكيبيديا، الموسوعة الحرة).