



جمهورية السودان
وزارة التعليم العالي والبحث العلمي
جامعة شندى
كلية الدراسات العليا

بحث مقدم لنيل درجة الماجستير في علوم الاحصاء بعنوان:

دراسة مقارنة بين نموذج الإنحدار اللوجستي وخوارزميتي الغابة العشوائية والمتجه الداعم في التصنيف
(دراسة حالة بمركز الشهيد نوري – مستشفى عطبرة – للعام 2024م)

A Comparative Study between Logistic Regression Model and Random Forest
and Support Vector Machine Algorithms in Classification
(A Case Study at Al-Shahid Nuri Dialysis Center – Atbara Hospital – for the
year 2024)

إشراف:
د. رجاء مصطفى صالح الأمين

إعداد الطالبة:
ملاذ صلاح سرالختم ناصر

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

سورة البقرة

الحمد لله رب العالمين

سورة البقرة

الحمد لله رب العالمين

الآية

قال الله تعالى:

﴿وَقُلْ رَبِّ زِدْنِي عِلْمًا﴾

صدق الله العظيم

سورة طه الآية: 114

الإهداء

إلى قلبي الذي ينبض خارج صدري ... أمي الحبيبة (أ. هبه عثمان)، إلى
سندي الذي لا يتزعزع... أبي الغالي (د. صلاح سرالختم ناصر)، إلى من
جعلنا من التضحيات سلماً لطموحاتي وسقيا حبهم جعلتني أستطيع بلوغ النجوم ،
إليكما يا أغلى مافي الوجود أهدي أول إنجاز أكاديمي في حياتي.
إلى أخي وأختي (محمد ومزن) من شاركوني اللعب في الطفولة والذاكرة في
الشباب ، إلى من كانوا خير سند و عون لكم في قلبي مكان خاص لا يزول .
إلى منقذي الأكاديمي ومشرفتي الفاضله الدكتورة / رجاء مصطفى صالح الأمين
التي كانت خير مرشد ومعلم ، لم تبخل بعلمها ولا وقتها ولا نصائحها كنتي نعم
السند العلمي لهذا البحث .

إلى لجنة المناقشة الكرام وأساتذتي الأفاضل على مر السنين من التعليم
الأساسي حتى الدراسات العليا لكم جزيل الشكر وعظيم الإمتنان .
إلى مرضى الكلى في مستشفى عطبرة الذين كانوا مصدر إلهامي الأول وإلى
الكادر الطبي في مركز الشهيد نوري الذين ساهموا في جمع البيانات ودعموا
البحث هذا العمل من اجلكم و لأجلكم .
إلى زملاء الدراسة وإلى كل باحث شاركني النقاش العلمي لقد جعلتم الرحلة
أجمل والتحدي أهون .

إلى كل من دعى لي بظهر الغيب وكل من ساعدني ولو بكلمة و إلى كل من
صبر على وأنا غارقه في البحث أنتم شركاء النجاح ... والشكر لكم لا ينقطع .
وأخيراً أقدم هذا الجهد المتواضع لعله يكون لبنة في صرح العلم ، مدركه أنه لا
يخلو من الهفوات والنقص سائله المولى عز وجل أن ينفع به ويجعله خالصاً
لوجهه الكريم .

الشكر والتقدير

الحمد لله الذي بنعمته تتم الصالحات ، والصلاة والسلام على سيدنا محمد النبي الأمي الذي أرسله الله رحمة للعالمين . أبدأ شكري بحمد الله عز وجل الذي يسر لي سبل العلم والمعرفة وأعانني على إتمام هذا العمل فله الحمد أولاً وآخراً ، ظاهراً وباطناً .

أتقدم بجزيل الشكر والعرفان إلى مشرفتي الكريمة الدكتورة/ رجا مصطفى صالح الأمين التي لم تبخل على بتوجيهاتها القيمة ونصائحها الثمينة فكانت خير مرشد حيث صحت مسار البحث وأضاءت لي الطريق بعلمها الواسع وخبرتها الطويلة .

أشكر أعضاء لجنة المناقشة الأفاضل على تفضلهم بقبول مناقشة هذه الرسالة وما بذلوه من وقت وجهد في قراءتها وتقويمها ، سائله المولى أن يجعل هذا في ميزان حسناتهم .

وأقدم بالشكر لجميع أساتذتي الأفاضل في قسم الإحصاء الذين غرسوا في أساسيات العلم وفتحوا لي آفاق المعرفة .

أشكر إدارة مستشفى عطبرة و مركز الشهيد نوري لعلاج وغسيل الكلى على تعاونهم الكريم وتسهيلهم لجمع البيانات .

أشكر زملائي الباحثين في مرحلة الماجستير على ما تبادلناه من نقاشات علمية بناءه ، كما أشكر الأصدقاء الأعزاء الذين كانوا سنداً معنوياً في رحلة البحث .

شكري وإمتناني لا ينقطع لأسرتي الكريمة ، والدي وإخوتي الذين تحملوا وصبروا على إنشغالي وكانوا مصدر الدعم المعنوي الأكبر .

المستخلص

هدفت هذه الدراسة إلى تصنيف المرضى المصابين بأمراض الكلى من غير المصابين، باستخدام نماذج إحصائية حديثة وخوارزميات للتعلم الآلي، وذلك بناءً على بيانات معملية مُستقاة من مركز الشهيد نوري لعلاج وغسيل الكلى بمستشفى عطبرة بالسودان. تمثلت مشكلة البحث في تحديد الأسلوب الأكثر دقة بين النماذج الكلاسيكية والحديثة . هدفت الدراسة إلى مقارنة أداء ثلاثة نماذج رئيسية في تصنيف ثنائي تشتمل : نموذج الانحدار اللوجستي كأحد الأساليب الإحصائية التقليدية ، وخوارزمية الغابة العشوائية ، وخوارزمية المتجه الداعم كأساليب قائمة على الذكاء الاصطناعي والتعلم الآلي . إتمدت الدراسة على منهج وصفي تحليلي ، حيث تم جمع عينة عشوائية بلغ حجمها 260 حالة (130 مصاباً و130 غير مصاب) ، وتم تحليل ستة متغيرات مستقلة (الهيموجلوبين ، الكرياتينين ، الصوديوم ، البوتاسيوم ، الفسفور ، الكوليسترول) باستخدام برنامج بايثون . أظهرت نتائج التحليل أن جميع النماذج حققت أداءً جيداً . فقد سجلت خوارزمية الغابة العشوائية أعلى دقة تصنيفية بلغت 98.72% ، تليها خوارزمية المتجه الداعم بدقة 96.15% ، ثم نموذج الانحدار اللوجستي بدقة 93.59% . كما كشفت النتائج عن المتغيرات المؤثرة في التشخيص ؛ حيث كانت الكرياتينين و الهيموجلوبين و الفسفور من أهم المؤشرات الإحصائية المعبرة عن الإصابة . وأوصت الدراسة أن خوارزميات التعلم الآلي ، وخاصة الغابة العشوائية ، تتفوق على النموذج الإحصائي الكلاسيكي من حيث دقة التصنيف ، وأنها أدوات فعالة وقوية يمكن اعتمادها في دعم النظم التشخيصية الطبية لأمراض الكلى .

Abstract

This study aimed to classify patients with kidney diseases from non-patients , using modern statistical models and machine learning algorithms , based on laboratory data obtained from Al-Shahid Nouri Center for Kidney Treatment and Dialysis at Atbara Hospital in Sudan . The research problem was to determine the most accurate method among classical and contemporary models . The study aimed to compare the performance of three main models in binary classification : the **Logistic Regression** model as a traditional statistical method , the **Random Forest** algorithm , and the **Support Vector Machine** algorithm as artificial intelligence and machine learning-based approaches . The study adopted a descriptive-analytical methodology , Where a random sample of 260 cases (130 patients and 130 non-patients) was collected , and six independent variables (Hemoglobin , Creatinine , Sodium , Potassium , Phosphorus , Cholesterol) were analyzed using the Python programming language . The analysis results showed that all models performed well : the **Random Forest** algorithm achieved the highest classification accuracy at **98.72%** , followed by the **Support Vector Machine** with an accuracy of **96.15%** , and then the **Logistic Regression** model with an accuracy of **93.59%** . The results also revealed the influential variables in diagnosis ; **Creatinine** , **Hemoglobin** , and **phosphorus** were among the most significant statistical indicators of the disease . The study recommended that machine learning algorithms , especially Random Forest , outperform traditional statistical model in classification accuracy and are effective , robust tools that can be adopted to support medical diagnostic systems for kidney diseases .

فهرست الموضوعات

الرقم	الموضوع	رقم الصفحة
	الآية	أ
	الإهداء	ب
	الشكر والتقدير	ج
	المستخلص	د
	Abstract	هـ
	فهرس الموضوعات	و
	فهرس الجداول	ك
	فهرس الأشكال	ل
الفصل الأول : إجراءات البحث		
1-1	تمهيد	1
2-1	مشكلة البحث	2
3-1	أهمية البحث	2
4-1	أهداف البحث	2
5-1	فرضيات البحث	3
6-1	منهجية البحث	3
7-1	مجتمع البحث	3
8-1	حدود البحث	4
9-1	الدراسات السابقة	4
10-1	هيكلية البحث	10
الفصل الثاني : الجانب النظري		
نموذج الإنحدار اللوجستي		

11	تمهيد	1-2
11	الإنحدار اللوجستي	2-2
14	مجالات تطبيق الإنحدار اللوجستي	1-2-2
14	شروط تطبيق الإنحدار اللوجستي	2-2-2
15	أنواع الإنحدار اللوجستي	3-2-2
18	تقدير معلمات الإنحدار نموذج الإنحدار اللوجستي	4-2-2
21	الإختبارات المستخدمة في نموذج الإنحدار اللوجستي	5-2-2
التعلم الآلي		
24	تمهيد	3-2
24	تطبيقات التعلم الآلي	4-2
24	المصطلحات الشائعة في التعلم الآلي	1-4-2
25	تاريخ التعلم الآلي	2-4-2
27	التعلم بإشراف	3-4-2
28	فوائد التعلم الخاضع للإشراف	4-4-2
28	عيوب التعلم الخاضع للإشراف	5-4-2
29	التعلم بدون إشراف	6-4-2
29	فوائد التعلم غير الخاضع للإشراف	7-4-2
30	عيوب التعلم غير الخاضع للإشراف	8-4-2
30	التعلم المعزز	9-4-2
30	الفرق بين التعلم الإشرافي وغير الإشرافي والمعزز	10-4-2
31	خوارزميات التعلم الآلي	11-4-2
34	التصنيف الإحصائي باستخدام التعلم الآلي	12-4-2
35	طرق تصنيف البيانات	13-4-2

36	تصنيف البيانات الإحصائية حسب نوع البيانات والغرض من التصنيف	14-4-2
37	أهمية تصنيف البيانات الإحصائية	15-4-2
38	أنواع تصنيف البيانات	16-4-2
خوارزمية الغابة العشوائية		
39	أشجار القرار	1-3-2
39	مفهوم شجرة القرار	1-1-3-2
43	تعقيد شجرة إتخاذ القرار	2-1-3-2
43	آليات إيقاف توسع شجرة القرار	3-1-3-2
44	أنواع أشجار إتخاذ القرار	4-1-3-2
45	أشجار القرار العشوائية	5-1-3-2
45	بناء شجرة إتخاذ القرار	6-1-3-2
46	خطوات رسم شجرة القرارات	7-1-3-2
47	تحليل شجرة القرار	8-1-3-2
49	مميزات أشجار القرار	9-1-3-2
50	عيوب أشجار القرار	10-1-3-2
51	خوارزمية الغابة العشوائية	11-1-3-2
51	مفهوم الخوارزمية	12-1-3-2
51	تمهيد في الغابة العشوائية	13-1-3-2
52	خطوات عمل خوارزمية الغابة العشوائية	14-1-3-2
52	تاريخ الغابة العشوائية	15-1-3-2
53	ماهي الغابة العشوائية	16-1-3-2
56	خطوات بناء خوارزمية الغابة العشوائية	17-1-3-2
57	كيف تعمل خوارزمية الغابة العشوائية	18-1-3-2

58	مميزات خوارزمية الغابة العشوائية	19-1-3-2
58	عيوب خوارزمية الغابة العشوائية	20-1-3-2
خوارزمية المتجه الداعم		
60	تمهيد	2-3-2
60	آلة المتجه الداعم	5-2
61	مفهوم تقنية آلة المتجه الداعم	1-5-2
61	كيفية عمل تقنية آلة المتجه الداعم	2-5-2
62	إستخدامات تقنية آلة المتجه الداعم	3-5-2
65	دالة النواة	4-5-2
65	أهم أنواع دالة النواة	5-5-2
66	مميزات خوارزمية المتجه الداعم	6-5-2
66	عيوب خوارزمية المتجه الداعم	7-5-2
الفصل الثالث : الجانب التطبيقي		
67	تمهيد	1-3
67	لغة بايثون	2-3
68	تاريخ بايثون	1-2-3
68	مميزات لغة بايثون	2-2-3
69	إستخدامات لغة بايثون	3-2-3
70	لماذا لغة بايثون وليس غيرها	4-2-3
71	متغيرات الدراسة لمرض الكلى التي يتم التطبيق عليها	3-3
73	نتائج تحليل البيانات بإستخدام نموذج الانحدار اللوجستي بإستخدام برنامج بايثون	4-3
76	نتائج إختبار Hosmer-Lemeshow بإستخدام برنامج بايثون	1-4-3
77	نتائج إختبار Wald بإستخدام برنامج بايثون	2-4-3

82	نتائج تحليل البيانات باستخدام خوارزمية الغابة العشوائية باستخدام برنامج بايثون	5-3
88	نتائج تحليل البيانات باستخدام خوارزمية المتجه الداعم باستخدام برنامج بايثون	6-3
92	التحليل والتفسير الشامل للنتائج بعد التحليل	7-3
الفصل الرابع: النتائج والتوصيات		
94	تمهيد	1-4
94	النتائج	2-4
95	التوصيات	3-4
97	المراجع والمصادر	
الملاحق		
	ملحق (1)	
	ملحق (2)	

فهرس الجداول

رقم الصفحة	إسم الجدول	رقم الجدول
31	مقارنة بين التعلم الإشرافي وغير الإشرافي والمعزز	1-2
66	أنواع دوال النواة وصيغها الرياضية	2-2
72	مسميات متغيرات الدراسة	3-3
73	الإحصاءات الوصفية لمتغيرات الدراسة	4-3
74	مقاييس أداء نموذج الإنحدار اللوجستي في التصنيف الثنائي	5-3
77	معاملات الإنحدار المقدره ، إحصاء Wald ، القيمة الإحتمالية لإحصاء والد	6-3
79	تقديرات نموذج الإنحدار اللوجستي لمتغيرات الدراسة مع فترات الثقة 95%	7-3
80	تفسير المتغيرات ذات الدلالة الإحصائية	8-3
81	تفسير المتغيرات الغير دالة إحصائياً	9-3
82	الإحصاءات الوصفية لمتغيرات الدراسة	10-3
83	القيم المفقودة لكل متغير في مجموعة البيانات	11-3
84	مقاييس أداء خوارزمية الغابة العشوائية في التصنيف	12-3
86	الأهمية النسبية للمتغيرات	13-3
88	مقاييس أداء خوارزمية المتجه الداعم في التصنيف	14-3
91	الأهمية النسبية للمتغيرات	15-3

فهرس الأشكال

رقم الصفحة	إسم الشكل	رقم الشكل
34	خوارزميات التعلم الآلي	1-2
53	الغابة العشوائية	2-2
55	هيكلية الغابة العشوائية كأحد أنواع أشجار الإنحدار	3-2
56	الإطار العام لخوارزمية عمل الغابة العشوائية	4-2
64	متجهات الدعم الواقعة على الحد الفاصل بين البيانات في الفضاء	5-2

الفصل الأول

إجراءات الدراسة

الفصل الأول إجراءات الدراسة

1-1: تمهيد :-

يعتبر تصنيف البيانات من الخطوات المهمة في معالجة وتحليل البيانات ويساعد في تحليل وفهم الظواهر قيد التصنيف ، وهو عملية تمكننا من معرفة البيانات وخصائصها دون فرضيات مسبقه عما يجب أن تكون هذه المعرفة . وتوجد عدة طرق للتصنيف منها الدالة التمييزية والانحدار اللوجستي ، وظهرت في الأونة الأخيره طرق حديثة لتصنيف البيانات تعتمد على أساليب تنقيب البيانات والذكاء الاصطناعي منها خوارزميات وبرمجيات تستخدم في تصنيف البيانات مثل خوارزمية الغابة العشوائية وخوارزمية المتجه الداعم وخوارزمية NaiveBayes ، ويتم إستخدام أساليب التصنيف الكلاسيكية والحديثة في المجالات الطبية بكثرة خاصة في تصنيف الإصابة بالأمراض ، ومن هذه الأمراض مرض الكلى .

وفي الأونة الأخيرة إنتشرت أمراض الكلى بصورة كبيرة في السودان وهي من الأمراض التي تؤدي إلى الوفاة لذلك لأبد من وجود أساليب من خلالها يمكن معرفة مسببات المرض والمسببات الأكثر تأثيراً في المرض ومعرفة طريقه تجعلنا نصنف الشخص المصاب من الشخص الغير مصاب .

وتوجد عدة دراسات قامت بإستخدام النماذج الكلاسيكية في التصنيف بين المصابين وغير المصابين بمرض معين ، كما تطرقت بعض الدراسات لإستخدام بعض الخوارزميات في التصنيف ، لذلك سوف نقوم باستخدام خورزميات حديثة في التصنيف ونقوم بمقارنة نموذج الانحدار اللوجستي وخوارزميتي الغابة العشوائية (RF) و المتجه الداعم (SVM) في تصنيف المصابين وغير المصابين بأمراض الكلى .

2-1 : مشكلة البحث :-

تتمثل مشكلة البحث في إستكشاف إمكانية توظيف طرق إحصائية حديثة تعتمد على الذكاء الاصطناعي ، مثل خوارزميات الغابة العشوائية والمتجه الداعم ، لمساعدة مركز الشهيد نوري لعلاج الكلى في تصنيف البيانات الطبية للمرضى بدقة عالية ، ومن ثم مقارنة أداء هذه الخوارزميات بنموذج الإنحدار اللوجستي التقليدي ، بهدف تحديد الأسلوب الأكثر دقة وفعالية في دعم التشخيص داخل المركز .

3-1 : أهمية البحث :-

تتمثل أهمية البحث من جانبين جانب علمي وجانب عملي فمن الجانب العلمي فإن إستخدام طرق حديثة في التصنيف مثل خوارزميات الغابة العشوائية والمتجه الداعم ومقارنتها مع نموذج الإنحدار اللوجستي للوصول للأسلوب الأدق تصنيفاً يعتبر إضافة حقيقية وتطبيقية لعلم الإحصاء ، أما من الجانب العملي فإن وجود أسلوب إحصائي دقيق يمكننا من تصنيف المصابين وغير المصابين بمرض الكلى ومعرفة المتغيرات والعوامل التي تؤثر على الإصابة بالمرض يعتبر إضافة مهمة للقطاع الصحي بالبلاد .

4-1 : أهداف البحث :-

يهدف هذا البحث إلى الآتي :

1. التعرف على نموذج الإنحدار اللوجستي وصيغته الرياضية وإستخدامه في التصنيف .
2. التعرف على خوارزمية الغابة العشوائية وإستخدامها في التصنيف.
3. التعرف على خوارزمية المتجه الداعم وإستخدامها في التصنيف .
4. تطبيق خوارزميات الغابة العشوائية و المتجه الداعم ونموذج الإنحدار اللوجستي بإستخدام برنامج بايثون .
5. مقارنة نتائج التصنيف بين نموذج الإنحدار اللوجستي وخوارزميات الغابة العشوائية والمتجه الداعم ، للوصول إلى أدق أسلوب من حيث التصنيف .

5-1 : فرضيات البحث :-

يهدف البحث إلى إختبار الفرضيات :-

1. نموذج الانحدار اللوجستي المقدر له قدره على تصنيف بيانات البحث .
2. خوارزمية الغابة العشوائية (RF) تتصف بدقة تصنيف بيانات البحث .
3. خوارزمية المتجه الداعم (SVM) تتصف بدقة تصنيف بيانات البحث .
4. خوارزمية الغابة العشوائية أدق في تصنيف بيانات البحث من نموذج الانحدار اللوجستي و خوارزمية المتجه الداعم .
5. خوارزمية المتجه الداعم (SVM) أدق من نموذج الانحدار اللوجستي في تصنيف بيانات البحث .
6. خوارزمية الغابة العشوائية أدق أسلوب في تصنيف بيانات البحث .

6-1 : منهجية البحث :-

تتمثل منهجية البحث في الأسلوب الوصفي والأسلوب التحليلي ، حيث سنقوم بوصف بيانات الدراسة بإستخدام أساليب الإحصاء الوصفي ، ومن ثم سوف نقوم بتحليل بيانات الدراسة وتصنيفها بإستخدام نموذج الانحدار اللوجستي و خوارزميتي الغابة العشوائية والمتجه الداعم بإستخدام برنامج بايثون للوصول إلى الأسلوب الأكثر دقة في تصنيف المصابين وغير المصابين بمرض الكلى .

7-1 : مجتمع البحث :-

يتكون مجتمع الدراسة من جميع الأشخاص المصابين وغير المصابين بمرض الكلى من مركز الشهيد نوري بمستشفى عطبرة في العام 2024م ، حيث تم أخذ 130 عينة من المصابين و 130 عينة من غير المصابين ، وذلك بهدف إجراء المقارنات التصنيفية بين النماذج الإحصائية المختلفة .

8-1 : حدود البحث :-

الحدود المكانية : مركز الشهيد نوري - مستشفى عطبرة - عطبرة - ولاية نهر النيل - السودان .

الحدود الزمانية : من 1/يناير/2024م - 31/ديسمبر/2024م .

9-1 : الدراسات السابقة :-

1. في العام 2023م قام الباحث سمير فرج أبودحروج بعمل ورقة علمية منشوره بعنوان إستخدام نموذج SVM في التنبؤ بالسلاسل الزمنية . مجلة الدراسات الإستراتيجية والبحوث السياسية ، المجلد 2 ، العدد (1) .

هدفت الورقة العلمية إلى تحديد النموذج الأمثل من بين نماذج SVM وإستخدام النموذج الأمثل في التنبؤ بالأسعار العالمية الشهرية من الذهب مستقبلا ، وقد تم إستخدام منهجية أسلوب شعاع الدعم الألي للتنبؤ بأسعار الذهب العالمية وتم الإعتماد على البرنامج الإحصائي R في تحليل بيانات الدراسة ، وأشارت نتائج الدراسة إلى أن نموذج SVM هو النموذج الأفضل في التنبؤ وذلك لإمتلاكه أقل قيم لخطأ التدريب وعدد متجهات الدعم ، وأوصت الدراسة بإستخدام أساليب تقنية مثل الخوارزمية الجينية والمنطق المضبب ومقارنة نتائج هذه الأساليب مع نتائج نموذج SVM .

تتفق هذه الدراسة مع الدراسة الحالية في إستخدام خوارزميات حديثة تعتمد على الذكاء الإصطناعي في التنبؤ والتصنيف . وتختلف في أن على الرغم من أهمية دراسة أبو دحروج في إبراز كفاءة نموذج SVM في التنبؤ بالسلاسل الزمنية إلا أنها لم تتناول مجال التصنيف الطبي ولم تقارن أداء SVM بنماذج أخرى مثل الغابة العشوائية والانحدار اللوجستي . وتأتي هذه الدراسة لتملأ هذه الفجوة من خلال تطبيق هذه النماذج على بيانات طبية حقيقية ، وتقديم توصية عملية بأفضل خوارزمية يمكن إعتمادها في تشخيص أمراض الكلى بمركز الشهيد نوري مع توظيف إمكانيات بايثون الحديثة في التحليل .

2. في العام 2023م قامت الباحثة علياء علاء حسين ببحث ماجستير بعنوان
إستخدام خوارزمية الغابة العشوائية وخوارزمية الجندب الثنائي للكشف عن
البرامج الضارة لنظام الأندرويد . جامعة الكوفة - كلية علوم الحاسوب
والرياضيات .

هدف البحث إلى إختيار الخصائص ذات الصلة من تطبيقات أندرويد ، وتوصل
البحث إلى أن الجمع بين خوارزميتي الغابة العشوائية و الجندب الثنائي في الكشف
عن البرامج الضارة لنظام أندرويد يحقق نتائج فعالة وأداء أفضل .
تتفق هذه الدراسة مع الدراسة الحالية في إستخدام تقنيات حديثة في الكشف والتنبؤ .
وتختلف في المجال التطبيقي بشكل أساسي فبينما ركزت دراسة علياء حسين على
توظيف خوارزمية الغابة العشوائية في مجال أمن المعلومات للكشف عن البرامج
الضارة لتطبيقات الأندرويد ، فإن الدراسة الحالية تطبق نفس الخوارزمية في المجال
الطبي لتشخيص أمراض الكلى وتصنيف المرضى بناءً على بيانات كيميائية حيوية
حقيقية . كما أن الدراسة الحالية لم تكتف بإستخدام خوارزمية واحدة بل أجرت مقارنة
شاملة بين ثلاثة نماذج لتحديد الأكثر دقة في التصنيف الطبي وذلك بالإستعانة ببرنامج
بايثون وتحليل بيانات محلية من مركز الشهيد نوري بمستشفى عطبرة وهو مايضيف
بعداً تطبيقياً جديداً في السياق الصحي السوداني .

3. في العام 2022م قامت الدكتور سوزي فاروق النقودي بعمل ورقة علمية
بعنوان تقييم دقة أساليب التنقيب في البيانات DM في الحد من المخاطر
الإئتمانية وإنعكاسها على جودة القوائم المالية بالقطاع المصرفي . مجلة
البحوث المالية والتجارية ، المجلد 23 ، العدد (4) .

هدفت الورقة العلمية إلى تقييم دقة الأساليب المختلفة للتنقيب في البيانات DM
بإعتباره أحد التقنيات الحديثة في مجال الذكاء الإصطناعي AI في التنبؤ بمنح
القروض للحد من المخاطر الإئتمانية وإنعكاس ذلك على جوده القوائم المالية ، وقد تم
إستخدام المنهج التحليلي الوصفي بإستخدام تقنيات التنقيب عن البيانات وإستخدام

خوارزمية التصنيف والتجميع في تحليل البيانات واكتشاف المعرفة والتنبؤ بالبيانات عن طريق الأساليب المختلفة للتقريب في البيانات مثل أسلوب الإنحدار اللوجستي LR وشجرة القرار والغابات العشوائية RF وتوصلت الدراسة إلى أن التقريب في البيانات يساعد في التنبؤ باكتشاف المخاطر التي تهدد البنك وذلك من خلال تحديد الخصائص التي تشير إلى وجود مخاطر تحد من منح العميل القرض ، وأوصت الدراسة بمحاولة دمج أكثر من أسلوب من أساليب التقريب في البيانات للاستفادة من مزايا كل أسلوب والتغلب على عيوبه في التعامل مع البيانات .

تتفق هذه الدراسة مع الدراسة الحالية في استخدام تقنيات حديثة في مجال الذكاء الاصطناعي في التنبؤ . وتختلف في :

- هذه دراسة حالة بالقطاع المصرفي ومنح القروض أما الدراسة الحالية بمستشفى عطبره .

- في هذه الدراسة تم استخدام أسلوب الإنحدار اللوجستي وشجرة القرار والغابات العشوائية أما الدراسة الحالية تم المقارنة بين نموذج الإنحدار اللوجستي والغابات العشوائية والمتجه الداعم .

4. في العام 2022م قامت الدكتوراه الهام عبد حسين بعمل ورقة علمية بعنوان الضغط بين الإنحدار اللوجستي وشجرة التصنيف والإنحدار والغابة العشوائية في التنبؤ بأمراض الكلى . مجلة كلية التراث الجامعة ، المجلد 2 ، العدد (34) .

هدفت الورقة العلمية إلى التنبؤ بالمرضى المصابين بهذا النوع من المرض في المراحل المبكرة عن طريق الأساليب أو الطرق الإحصائية الحيوية مثل طريقة الإنحدار اللوجستي وشجرة التصنيف والغابة العشوائية ، وتوصلت الدراسة إلى أن طريقة الغابة العشوائية حققت الأفضلية من بين الطرق الثلاثة وإن أهم المتغيرات التي لها تأثير واضح على مرضى الكلى هي الألبومين واليوريا .

تتفق هذه الدراسة مع الدراسة الحالية في استخدام طرق إحصائية مثل طريقة الإنحدار اللوجستي والغابة العشوائية للتنبؤ بالمرضى المصابين بأمراض الكلى في المراحل المبكرة . وتختلف في أن بينما ركزت دراسة إلهام على المقارنة بين الإنحدار اللوجستي وشجرة التصنيف والغابة العشوائية باستخدام متغيري الألبومين واليوريا تأتي الدراسة الحالية لتقدم مقارنة موسعة تشمل أيضاً خوارزمية المتجه الداعم وتعتمد على ستة متغيرات حيوية أكثر تنوعاً مع الاستفادة من إمكانيات برنامج بايثون في التحليل الدقيق للبيانات كما أن تطبيق الدراسة على عينة من مركز الشهيد نوري في السودان يضيف بعداً محلياً جديداً ويسهم في إختبار مدى تعميم نتائج الدراسات السابقة على بيانات جغرافية وصحية مختلفة .

5. في العام 2021م قام الباحثان (Makala & Li) بعمل ورقة علمية بعنوان المقارنة بين نموذج SVM ونموذج ARIMA للتنبؤ بأسعار الذهب . مجلة الفيزياء ، المجلد 1767 ، العدد (36) .

هدفت الورقة العلمية إلى المقارنة بين نموذج SVM ونموذج ARIMA للتنبؤ بأسعار الذهب اليومية من عام 1979 إلى 2019 في التحليل حيث أظهرت نتائج الدراسة أن نموذج SVM هو النموذج الأنسب والملائم لتحليل بيانات الدراسة وتم استخدامه في التنبؤ بأسعار الذهب اليومية نظراً لدقته العالية .

تتفق هذه الدراسة مع الدراسة الحالية في التنبؤ باستخدام أساليب إحصائية حديثة . وتختلف في أن على الرغم من أهمية دراسة Makala & Li في إبراز كفاءة نموذج SVM في الأسواق المالية إلا أنها لم تتناول مجال التصنيف الطبي ولم تقارن أداء SVM مع نماذج أخرى شائعة في التعلم الآلي مثل الغابة العشوائية وتأتي الدراسة الحالية لتملاً هذه الفجوة من خلال تطبيق SVM في المجال الطبي لتشخيص أمراض الكلى وإجراء مقارنة شاملة بين SVM والإنحدار اللوجستي والغابة العشوائية و الاعتماد على بيانات محلية حديثة تعكس واقع المرضى في المنطقة ، واستخدام

برنامج بايثون بمكتباته المتخصصة في التعلم الآلي مما يتيح دقة تحليلية عالية (98.72% للغابة العشوائية) .

6. في العام 2020م قام الباحث زيدون وليد سلمان ببحث ماجستير بعنوان تقدير القيم الخالية بإستخدام خوارزمية الغابات العشوائية المعدلة عبر خوارزمية سرب الذكية . الجامعة المستنصرية - قسم علوم الحاسوب .

هدف البحث إلى معرفة البيانات الخالية التي تحدث بشكل متكرر في قواعد البيانات وتسبب إشكالية فيها ، وتوصلت الدراسة إلى أن إستخدام خوارزمية RF بشكل شائع للتنبؤ بالمعلومات وتوضيح الإستنتاجات يتجلى ذلك في قدره على إستخدام المعلومات الكبيرة والتفاعلات والعلاقات المعقدة ، كما أن إستخدام MCA يعد أفضل من الغابة العشوائية الكلاسيكية المقارنة تظهر تحسناً كبيراً في مقاييس الدقة والوقت .

7. في العام 2018م قام الباحث أحمد عبد الصمد حبيب بعمل ورقة علمية بعنوان المقارنة بين نموذجي آلة المتجه الداعم والانحدار اللوجستي في تصنيف مرضى السكري في مستشفى الموانئ العام في البصرة . مجلة الإقتصاد الخليجي - العدد (35) .

هدفت الورقة العلمية إلى تصنيف مرضى داء السكري بإستخدام طريقة آلة المتجه الداعم SVM وبيان دقتها ، وتم إستخدام المنهجية بالإعتماد على طريقه SVM في التصنيف للوصول إلى النتائج وإستخدام البرنامج الإحصائي R ، وتوصلت الدراسة إلى أن آلة المتجه الداعم تكون طريقة مثلى للإستخدام في مجال التصنيف وتمييز الأنماط وتحقيق دقة عالية في نسب التعرف الصحيح ، كما أوصت الدراسة بضرورة إستعمال آلة المتجه الداعم في الدراسات الإحصائية المستقبلية في مجال التمييز والتصنيف وكذلك في المجال الصحى وبالإمكان أيضاً إستعمالها لأغراض التنبؤ مثل التنبؤ بالأمراض وفي المجالات جميعها .

تتفق هذه الدراسة مع الدراسة الحالية في تصنيف مرضى السكري بالمقارنة بين نموذجي المتجه الداعم والانحدار اللوجستي . وتختلف في أن بينما ركزت دراسة

حبيب على المقارنة الثنائية بين SVM والانحدار اللوجستي في تشخيص مرض السكري بالإعتماد على برنامج R تأتي الدراسة الحالية لتقدم مقارنة أكثر شمولاً تضم ثلاثة نماذج في تشخيص أمراض الكلى مستفيدة من إمكانيات برنامج بايثون الحديثة كما أن تطبيق الدراسة على بيانات محلية في السودان يضيف بعداً جغرافياً جديداً ويسهم في إختبار مدى تعميم نتائج الدراسات السابقة على بيئات صحية مختلفة مع الوصول إلى دقة تصنيفية عالية 98.72% للغابة العشوائية تؤكد كفاءة هذه النماذج في السياق المحلي .

8. في العام 2016م قام الباحث أنور الزين بابكر مصطفى وآخرون بعمل ورقة علمية بعنوان إستخدام الانحدار اللوجستي لتحديد أهم العوامل المؤثرة على الإصابة بمرض القلب . مجلة السودان الأكاديمية للبحوث والعلوم ، المجلد 1 ، العدد (1) .

هدفت الورقة العلمية إلى بناء نموذج يمكن من التنبؤ بإحتمالات المصابين بمرض القلب وتحديد أهم العوامل التي تؤثر على الإصابة بأمراض القلب ، وقد تم إستخدام المنهج الوصفي والتحليلي حيث إعتد على الجانب الوصفي في إستخدام الجداول التكرارية والرسومات البيانية للبيانات أما الجانب التحليلي تم إستخدامه في بناء نموذج الانحدار اللوجستي وفحص جودة توفيقه للبيانات ، وتوصلت الدراسة إلى أن نموذج الانحدار اللوجستي الذي يوفق بيانات المتغيرات التي تؤثر على الإصابة بأمراض القلب له قدرة تفسيرية وتصنيفية عالية وبشكل دال إحصائياً .

تتفق هذه الدراسة مع الدراسة الحالية في التعرف على العوامل المؤثرة على الإصابة بمرض القلب . وتختلف في نطاق التحليل وأساليب المعالجة ، فبينما إعتدت دراسة أنور على نموذج الانحدار اللوجستي فقط في تحليل العوامل المؤثرة على أمراض القلب فإن الدراسة الحالية تقدم مقارنة شاملة بين ثلاثة نماذج مما يتيح تقييم أداء الأساليب التقليدية مقابل خوارزميات التعلم الآلي الحديثة كما أن الدراسة الحالية طبقت هذه النماذج على أمراض الكلى بدلاً من أمراض القلب وباستخدمة متغيرات

حيوية (الهيموجلوبين ، الكرياتينين ، الصوديوم ، البوتاسيوم ، الفسفور ، الكوليسترول) وليس فقط المتغيرات التقليدية إضافة إلى ذلك تم توظيف برنامج بايثون بمكتباته المتخصصة في التعلم الآلي مما أتاح دقة تصنيفية عالية بلغت 98.73% للغابة العشوائية وهو ما يمثل نقلة نوعية في دقة التحليل مقارنة بالأساليب الوصفية البحثية .

1-10 : هيكلية البحث :-

يتضمن هذا البحث على أربعة فصول ، الفصل الأول إجراءات الدراسة البحث وتشمل على (التمهيد ، مشكلة البحث ، أهداف البحث ، أهمية البحث ، فرضيات البحث ، منهجية البحث ، حدود البحث ، الدراسات السابقة و هيكلية البحث) . الفصل الثاني ويشمل على الإطار النظري وتحدثنا فيه عن : نموذج الانحدار اللوجستي ، التعلم الآلي ، خوارزمية الغابة العشوائية و خوارزمية المتجه الداعم . الفصل الثالث يحتوي على الجانب التطبيقي وفيه يتم تطبيق عملي للخوارزميات بإستخدام برنامج بايثون وعرض وشرح مصدر البيانات وبيئة العمل والأدوات المستخدمة في بايثون . أما الفصل الرابع يشمل على النتائج والتوصيات ، وأخيراً المصادر والمراجع .

الفصل الثاني الجانب النظري

الفصل الثاني نموذج الانحدار اللوجستي

1-2 : تمهيد :

الانحدار Regression : هو أسلوب إحصائي يختص بدراسة العلاقة بين متغير تابع ومتغير أو عدة متغيرات مستقلة ، ناتجاً عن ذلك معادلة رياضية حيث تمثل هذه العلاقة أفضل تمثيل قد تكون المعادلة خطية وقد تكون غير خطية ، بعد تحديد شكل هذه العلاقة نقوم بتقدير معالم النموذج التي تعبر عن مقدار تأثير التابع بالمتغيرات المفسره وذلك لغرض التفسير أو التنبؤ حسب طبيعة الدراسة . يعتبر تحليل الانحدار أحد الأساليب الإحصائية الهامة التي تستخدم في العديد من مجالات العلوم المختلفة ، وتهدف دراسة الانحدار إلى تقدير معالم المعادلة الرياضية التي تعبر عن العلاقة السببية بين المتغيرات . ولقد أصبحت أساليب الانحدار مكوناً أساسياً في أي تحليل للبيانات يتعلق بوصف العلاقة بين متغير الإستجابة ومتغير أو أكثر من المتغيرات التوضيحية . نموذج الانحدار اللوجستي هو نموذج الانحدار الأكثر استخداماً لتحليل بيانات المتغير التابع نوعية القياس . قبل البدء في دراسة شاملة لنموذج الانحدار اللوجستي ، من المهم أن نفهم أن الهدف من التحليل باستخدام هذا النموذج هو نفس هدف أي نموذج إندار آخر يستخدم في الإحصائيات أي لإيجاد أفضل ملائمة للنموذج و إيجاد نموذج قابل للتفسير لوصف العلاقة بين متغيرات النتائج (التابعة أو التفسيرية) ، وغالباً ماتسمى المتغيرات المستقلة ، المتغيرات المتغيرات المشتركة أو المتغيرات التوضيحية .

2-2 : الانحدار اللوجستي :

الانحدار اللوجستي Logistic Regression : هو أسلوب إحصائي لفحص العلاقة بين المتغير التابع النوعي القياس بمتغير واحد أو عدة متغيرات مستقلة (متغيرات مفسرة) .^[2] بحيث تكون المتغيرات المستقلة من أي نوع من مستويات القياس ، إلا أننا في الانحدار اللوجستي يكون هدفنا ليس تفسير التغير في قيم المتغير

التابع وإنما تفسير إحتمال حدوث وإحتمال عدم حدوث الظاهرة محل الدراسة وهو عبارة عن إنحدار غير خطي .

في علم الإحتمالات الإنحدار اللوجستي هو نموذج يستخدم للتنبؤ بإحتمالية وقوع حدث ما وذلك بملائمة البيانات على منحنى لوجستي^[3]. ويعرف أيضاً بنموذج اللوجستك ويقدم على أنه " الطريقة الإحصائية لنمذجة البيانات الثنائية " ، ويعد هذا النموذج من النماذج الخطية العامه المكون لنموذج ذى الحدين العشوائي أي أن نموذج اللوجستك كغيره من النماذج الخطية العامة يتكون من الجزء العشوائي (Random Part) الذي يشير إلى متغير الإستجابة والجزء النظامي (Systematic Part) الذي يحدد المتغيرات المستقلة المستخدمة في النموذج والجزء الرابط (Part Link) الذي يصف العلاقة الدالية بين الجزء النظامي ومتوسط الجزء العشوائي^[4]. ما يميز نموذج الإنحدار اللوجستي عن نموذج الإنحدار الخطي هو أن متغير النتيجة في الإنحدار اللوجستي ثنائي أو ثنائي التفرع . ينعكس هذا الفرق بين الإنحدار اللوجستي والإنحدار الخطي في شكل النموذج وإفتراضاته^[5]. تتبع الأساليب المستخدمة في التحليل بإستخدام الإنحدار اللوجستي إلى حد ما نفس المبادئ العامة المستخدمة في الإنحدار الخطي وبالتالي فإن التقنيات المستخدمة في تحليل الإنحدار الخطي تحفز نهجنا في الإنحدار اللوجستي .

يتميز الإنحدار اللوجستي بأنه النموذج الأكثر مرونة من نماذج الإنحدار التقليدي إذ أنه لا يشترط أن تكون العلاقة بين المتغيرات المستقلة والمتغير التابع خطية ، وأيضاً لا يشترط بأن تكون المتغيرات المستقلة من النوع المستمر أو المتقطع ولا أن يكون لها التوزيع الطبيعي ، وأيضاً بأن العلاقة بين المتغير التابع والمتغيرات المستقلة غير خطية ، ويمكن تحويلها إلى الشكل الخطي بإستعمال مايعرف بتحويل اللوجت^[6]. يستخدم الإنحدار اللوجستي عامة في التنبؤ بقيم المتغيرات التابعة الفئوية (Categorical Variable) اعتماداً على قيم مجموعة متغيرات مستقلة مختلطة (Mix Of Continuous And Categorical Variables) والتنبؤ بقيم المتغير

التابع يعبر عنها بصيغة إحصائية أى احتمال أن يأخذ المتغير التابع قيمة محددة – احتمال النجاح – بناءً على قيم المتغيرات المستقلة^[7]. يهدف الانحدار اللوجستي إلى تصنيف عناصر المجتمع المدروس إلى مجموعتين أو أكثر ، وذلك باستخدام التوزيع الاحتمالي اللوجستي المعروف بالعلاقة التالية^[8]:

$$E(y_i) = P_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik})}} \rightarrow (2-1)$$

$$i = 1, 2, \dots, n$$

$$J = 1, 2, \dots, k \quad ; \quad -\infty < x < +\infty$$

حيث :

$(K) \equiv$ عدد المتغيرات المستقلة .

$(p_i) \equiv$ القيمة الاحتمالية للمتغير التابع .

$(\hat{\beta}_j) \equiv$ المعالم المقدرة .

$(x_{ik}) \equiv$ المتغيرات المستقلة .

$E(y_i) \equiv$ القيمة المقدرة للمتغير التابع y .

يعتبر الطبيب والإحصائي جوزيف بيركسون أول من عرف مفهوم الانحدار اللوجستي وطبقه في مجال الإحصاء الحيوي سنة 1944م . ونُشرت أول ورقة بحثية يذكر فيها الانحدار اللوجستي بإسمة في مجلة الجمعية الإحصائية الأمريكية .

2-2-1: مجالات تطبيق الانحدار اللوجستي^[9]:

- الطب والإحصاء الحيوي : مثلاً إحتمال حدوث نوبة قلبية عند شخص ما خلال فترة زمنية معينة حسب المعرفة القبلية ببعض المعلومات الديمغرافية (عمر أو جنسة) أو الطبية (مؤشرات البدنية أو الصحية أو الغذائية) أو الوبائية (سلوكياتة كالتدخين) .
- الصيدلة : في تقدير رد الفعل والمقارنة بين نجاعة الأدوية .
- التأمينات : لفرز وتقسيم مجموعات العملاء حسب المخاطر ومدى قابلية جذبهم لمنتجات تأمين معينة .
- المجال البنكي : خصوصاً في تنقيط العملاء أثناء دراسة ملفات القروض (Credit Scoring) .
- التسويق : حساب توقعات ميل المستهلك إلى شراء منتج ما أو إمتناعه عن الشراء .
- سير الآراء والعلوم السياسية : للتنبؤ بقرار التصويت في الإنتخابات إعتقاداً على تنميط قبلي للمصوتين (مستواهم الإجتماعي ، توجهاتهم السياسية ، مستواهم التعليمي ،) .

2-2-2: شروط تطبيق الانحدار اللوجستي^[9]:

يرتكز الانحدار اللوجستي على مجموعة إفتراضات تعتبر شروطاً أساسية وجب توفيرها قبل التطبيق هي :

1. أن يكون المتغير التابع متغيراً إسمياً ثنائياً بالنسبة لصنف الانحدار اللوجستي ثنائي الحدين ، أو تصنيفياً بالنسبة للانحدار اللوجستي المتعدد الحدود أو رتبياً بالنسبة للانحدار الترتيبي .
2. أن يحتوي النموذج على أكثر من متغير مستقل ، يكون في المستوي التصنيفي أو النسبي أو الترتيبي أو الفئوي .

3. أن يحتمل المتغير التابع احتمالاً واحداً لكل فرد ، بمعنى لا يمكن أن يكون الفرد في فئتين في نفس الوقت Mutually Exclusive كأن ينتمي الفرد إلى مرضين في نفس الوقت .

4. يجب ألا يكون هنالك وجود للتعددية الخطية Multicollinearity ، والتي تعني عدم وجود إرتباطات قوية بين مجموعة المتغيرات المستقلة ، كما يمكن قياس تقدير التعددية الخطية من خلال إختبار الخطية في برنامج SPSS .

5. لا يجب أن يكون هناك قيم شاذة في المتغيرات المستقلة . ويمكن تحديد القيم الشاذة من خلال إختبار Mahalanobis المتوفر على برنامج SPSS .

2-2-3: أنواع الانحدار اللوجستي [9] :

ينقسم الانحدار اللوجستي إلى ثلاثة أنواع هي :

1. الانحدار اللوجستي الترتيبي Ordinal Logistic Regression :

يتميز هذا النوع بتفسير أثر المتغيرات المستقلة باختلاف مستويات قياسها على الإستجابات الرتبية بمعنى أن يكون المتغير التابع متغير ترتيبياً .

2. الانحدار اللوجستي المتعدد Multinomial Logistic Regression :

يتم إستخدامة في حالة كان المتغير التابع يتكون من أكثر من فئتين تصنيفيتين أو إسميتين .

شكل المعادلة يوضح كالآتي :

$$\hat{y} = E(y_i / x_{ij}) = p_i = \frac{e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}}{1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}}} \rightarrow (2-2)$$

$$J = 1, 2, 3, \dots, k$$

$$I = 1, 2, 3, \dots, n$$

حيث :

$E(y/x) = \hat{y}$ ≡ القيمة المتوقعة للمتغير التابع المقدر .

x_{ik} ≡ المتغير المستقل رقم k في المشاهدة رقم i .

$$\hat{\beta}_j \equiv \text{المعلمة المقدرة } j .$$

3. الانحدار اللوجستي ثنائي الحدين Binomial Logistic Regression :

يدعى أيضاً بالانحدار اللوجستي الثنائي ، حيث يعتبر أشهر أنواع الانحدارات اللوجستية ، ويستخدم الانحدار اللوجستي الثنائي في تفسير أثر المتغيرات المفسره على الإستجابات الثنائية ، بمعنى تفسير قدرة مجموعة من المتغيرات المستقلة " المنبئة " ذات المستويات المختلفة على التنبؤ بمتغير واحد تابع يكون ثنائي التفرع Dichotomous مثل (ذكر/ أنثي ، مريض/ معافي) لذلك فالقيم أما أن تكون موجودة (إيجابية) وتأخذ القيمة 1 أو قد تكون غير موجودة (سلبية) وتأخذ القيمة 0 ، ولكون الباحث الإحصائي يهتم بوجود (تحقق) الإستجابات عند كل مستوى من مستويات المتغير المستقل الذي يعتقد بعلاقتها بها ، لذلك فإن متغير نسبة الإستجابة المشاهدة هو نسبة الإستجابات المتحققة (الإيجابية) عند كل مستوى من مستويات المتغير التوضيحي ، وهو نسبة متغيرة من مستوى إلى آخر ويقتررب توزيعها كثيراً من التوزيع الطبيعي عندما يكون حجم العينة كبيراً .

وتأتي معادلة الانحدار اللوجستي ثنائي الحدين كالتالي :

$$Y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n \rightarrow (2-3)$$

حيث :

$$Y \equiv \text{المتغير التابع (الثنائي)} .$$

$$b_0 \equiv \text{الحد الثابت في معادلة الانحدار اللوجستي ثنائي الحدين} .$$

$$b_n x_n \equiv \text{قيمة اللوجيت بالنسبة للمتغيرات المستقلة} .$$

إذا رسمنا المنحني الذي يمثل العلاقة بين Y كمتغير تابع و X كمتغير مستقل فسنجد أن الدالة تأخذ شكل منحني .

إن قيم $E(y)$ تنحصر بين (الواحد والصفر) وبالتالي تتراكم بين هاتين القيمتين كما يعتمد شكل الدالة التزايدية والتناقصي على إشارة (b_j) والمتغير التابع في النموذج المقدر هو متغير برنولي يأخذ إحدى القيمتين (0 ، 1) .

دالة هذا المنحنى هي الدالة اللوجستية لهذا عندما يكون Y متغير ثنائي القيمة نفرض أن :

$$E(y/x) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}} \rightarrow (2-4)$$

تسمى المعادلة السابقة بالدالة اللوجستية الثنائية .

ومن خصائص الدالة اللوجستية أنه يمكن تحويلها لدالة خطية بتحويلة مناسبة فإذا وضعنا

$$E(y/x) = \pi(x)$$

وإستخدمنا التحويلة

$$G(x) = \ln \frac{\pi(x)}{1 - \pi(x)}$$

حيث نجد أن $\pi(x)$ هي احتمال حدوث الظاهرة بينما $1 - \pi(x)$ هي احتمال عدم حدوث الظاهرة .

$$G(x) = b_0 + b_1 x_1 \dots \rightarrow (2-5)$$

وتسمى $g(x)$ اللوجت Logit .

وترجع أهمية النموذج اللوجستي لكون المقدّر (e^{b_1}) يعطي ما يسمى بنسبة الخطر odd ratio . ولتوضيح فكرة نسبة الخطر أفرض أن Y يأخذ القيمة (1) إذا كان الشخص مصاب بمرض معين والقيمة (0) إذا لم يكن الشخص مصاب .

كذلك أفرض أن X يأخذ القيمة (1) إذا كان الشخص تنطبق عليه صفة معينة والقيمة (0) إذا لم تنطبق عليه الصفة .

مخاطر The odd أن يكون الشخص مصاباً من الأشخاص الذين عندهم $X=1$ أي تعرف ب :

$$\text{معامل الخطر (1)} = \frac{\text{عدد الذين عندهم المرض وسط الذين يحملون الصفة}}{\text{عدد الذين ليس عندهم المرض وسط الذين يحملون الصفة}}$$

كما أن مخاطر The odd أن يكون الشخص مصاباً بين الذين ليس عندهم الصفة أي $X=0$ تعرف ب :

معامل الخطر (0) = $\frac{\text{عدد الذين عندهم المرض وسط الذين لا يحملون الصفة}}{\text{عدد الذين ليس عندهم المرض وسط الذين لا يحملون الصفة}}$
وتعرف نسبة الخطر *Odd ratio* بأنها النسبة بين الأثنين أي :

$$OR = \frac{odds(1)}{odds(0)} \rightarrow (2-6)$$

تمثل هذه النسبة مخاطر الإصابة بالمرض للأشخاص الذين عندهم الصفة مقارنة بتلك التي للأشخاص الذين ليس عندهم المرض . ويمكننا تحليل الإنحدار اللوجستي مباشرة من معرفة نسبة الخطر .

4-2-2 : تقدير معلمات نموذج الإنحدار اللوجستي :

يبني نموذج الإنحدار اللوجستي على فرض أساسي هو أن المتغير التابع Y متغير الإستجابة الذي نهتم بدراسة هو متغير ثنائي يتبع توزيع Bernoulli يأخذ الرتبة 1 بإحتمال (P) والرتبة 0 بإحتمال $(q=1-p)$ أي حدوث الإستجابة وعدم حدوثها^[9].

يمكن التعبير عن نموذج الإنحدار اللوجستي في حالة إحتوائه على متغير مستقل واحد فقط بالصيغة الرياضية الآتية :

$$\pi(x) = E(y/x) = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}} \rightarrow (2-7)$$

حيث أن :

$\pi(x) = E(y/x)$ ≡ المتوسط المشروط conditional mean لمتغير الإستجابة y عند قيمة معينة ل (x) .

β_0, β_1 ≡ معلمات نموذج الإنحدار اللوجستي .

أما إذا كان نموذج الإنحدار اللوجستي يحتوي على أكثر من متغير مستقل واحد وليكن k من المتغيرات المستقلة عندئذ يعبر عنه بالصيغة الرياضية الآتية :

$$\pi(x) = E(y/x) = \frac{e^{b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k}}{1 + e^{b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k}} \rightarrow (2-8)$$

حيث أن :

$b_0, b_1, b_2, \dots, b_k \equiv$ معاملات نموذج الانحدار اللوجستي .

ولتحويل المعادلة (2-8) إلى الشكل الخطي يستخدم مايسمى بتحويل اللوجت logit transformation والتي يعبر عنها بالمعادلة (2-9) أدناه :

$$G(x) = \ln \frac{\pi(x)}{1 - \pi(x)} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k \rightarrow (2-9)$$

حيث أن :

$$p(y=1/x) = \pi(x) , \quad y=1$$

$p(y=1/x)$ يمثل الإحتمال الشرطي conditional probability بأن تكون $y=1$ عند قيمة معينة ل (x) .

$$p(y=0/x) = 1 - \pi(x)$$

$p(y=0/x)$ يمثل الإحتمال الشرطي conditional probability بأن تكون $y = 0$ عند قيمة معينة ل (x) .

أما لتقدير معاملات المعادلة (2-9) فتستخدم طريقة الإمكان الأعظم maximum likelihood method وكما يأتي :

$$p(y=1/x) = \pi(x) , \quad y=1 \quad \text{بما أنه}$$

$$p(y=0/x) = 1 - \pi(x) , \quad y=0$$

لذا عندما يكون عينة من المشاهدات المستقلة التي حجمها n فإن لكل زوج (x_i, y_i) ،
 $i=1, 2, \dots, n$.

حيث : y_i تمثل رتبة متغير الإستجابة الثنائي للمفردة i .
 x_i تمثل قيمة المتغير المستقل للمفردة i .

فإنه :

$$p(y_i = 1/x) = \pi(x_i)^{y_i}, \quad y_i = 1$$
$$p(y_i = 0/x) = [1 - \pi(x_i)^{1-y_i}], \quad y_i = 0$$

وعليه يمكن التعبير عن دالة الإمكان likelihood function بالصيغة الرياضية
الآتية :

$$l(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{1-y_i} \rightarrow (2-10)$$

حيث أن :

$$p(y_i = 1/x) = \pi(x_i)^{y_i}, \quad y_i = 1$$

حيث $p(y_i = 1/x)$ يمثل الإحتمال الشرطي conditional probability بأن تكون $y_i = 1$ عند قيمة معينة ل (x) .

$$p(y_i = 0/x) = [1 - \pi(x_i)^{1-y_i}], \quad y_i = 0$$

حيث $p(y_i = 0/x)$ يمثل الإحتمال الشرطي conditional probability بأن تكون $y_i = 0$ عند قيمة معينة ل (x) .

وبأخذ اللوغاريتم الطبيعي للطرفين نحصل على المعادلة (2-11) الآتية^[10]:

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^n \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \rightarrow (2-11)$$

ثم يتم اشتقاق المعادلة (2-11) بالنسبة للمعاملات المراد تقديرها (β) وجعلها مساوية للصفر فينتج عدد من المعادلات التي لا يمكن حلها إلا من خلال خوارزمية تكرارية تسمى خوارزمية المربعات الصغرى الموزونة التكرارية Iteratively

weighted least squares algorithm^[11]

5-2-2 : الإختبارات المستخدمة في نموذج الإنحدار اللوجستي :

بعد تقدير نموذج الإنحدار اللوجستي ، يعد تقييم جودة مطابقة النموذج للبيانات خطوة أساسية للتأكد من مدى ملائمة وموثوقية نتائجه . توجد عدة إختبارات إحصائية تساعد في تقييم جودة المطابقة ومن أبرزها :

(1) إختبار Hosmer – Lemeshow لجودة المطابقة :

لقياس جودة المطابقة Goodness of fit لنموذج الإنحدار اللوجستي يتم استخدام إختبار Hosmer – Lemeshow ، حيث يقوم هذا الإختبار بتجميع حالات العينة بناء على قيم الإحتمالات المتوقعة ، Hosmer – Lemeshow

إستخدام إحدى إستراتيجيتين للتجميع في هذا الإختبار هما :

- تجميع الحالات بناء على المؤنديات للإحتمالات المتوقعة .
- تجميع الحالات بناء على قيم ثابتة للإحتمالات المتوقعة .

وتفضل الإستراتيجية الأولى على الثانية خاصة عندما يكون هناك العديد من الإحتمالات المتوقعة صغيراً (أقل من 0.2) .

ووفقاً للإستراتيجية الأولى التي تقوم على تجميع الحالات بناء على المؤنديات للإحتمالات المتوقعة يتم توزيع الحالات (n) بعد ترتيبها تصاعدياً حسب القيم المتوقعة للإحتمالات على عشرة مجاميع ($g=10$) ، بحيث يكون عدد الحالات في كل مجموعة $n/10$ ، وبحيث توضع في المجموعة الأولى الحالات ذات أقل قيمة للإحتمالات المتوقعة ($n_1 = n/10$) ، وتوضع في المجموعة الأخيرة الحالات ذات القيم الأعلى للإحتمالات المتوقعة ($n_{10} = n/10$) وكذلك مع بقية المجاميع بالترتيب .

ويتم جمع القيم المشاهدة والمتوقعة للحالات وفقاً لقيمتي المتغير التابع الثنائي y (0,1) ، وذلك في كل فئة من المجاميع العشر . بعد ذلك يتم حساب إحصاء Hosmer – Lemeshow التي يرمز لها بالرمز H ، والتي يتم حسابها وفقاً لإحصاء مربع كاي χ^2 ، حيث أن الإحصاء H تتبع توزيع مربع كاي بدرجات حرية تساوي ($g-2$) ، وتحسب إحصاء H وفق الصيغة التالية :

$$H = \sum_{k=1}^g \frac{(o_k - n'_k p'_k)^2}{n'_k p'_k (1 - p'_k)} \rightarrow (2-12)$$

حيث أن :

n'_k هي العدد الكلى للحالات في المجموعة k .

$o_k = \sum_{i=1}^{n'_k} y_i$ أي أن o_k هي عدد الإستجابات ، $Y=1$.

$p'_k = \sum_{i=1}^{n'_k} \frac{p_i}{n'_k}$ وهي متوسط الإحتمالات المتوقعة للمجموعة k .

وقد أظهر Hosmer – Lemeshow أنه إذا كانت قيمة الإحصاءة H المحسوبة من جميع المئينيات هي الطريقة المعتمدة وبحيث كانت قيمة هذه الإحصاءة بدرجة حريه (g-2) عند مستوى معنوية أكبر من 0.05 فإن ذلك يعني أن النموذج مطابق للبيانات المشاهدة .

(2) إختبار Wald لمعنوية المعلمات :

لغرض إحتساب معنوية المعلمات المقدرة بإستخدام النموذج اللوجستي يتم إستخدام إحصاءه Wald ، لكل معامل من معاملات الإنحدار اللوجستي المقابل لكل متغير مستقل ، من أجل إختبار الفرضية الصفرية (فرضية العدم) التي تنص على (أن تأثير معامل لوجت ما يساوي صفراً)، ويتم حساب إحصاءة Wald وفق الصيغة التالية :

$$Wald = \left[\frac{b}{SE_b} \right] \rightarrow (2-13)$$

حيث أن :

$b \equiv$ قيمة معامل الإنحدار اللوجستي للمتغير المستقل .

$SE \equiv$ قيمة الخطأ المعياري لمعامل الإنحدار اللوجستي للمتغير المستقل .

علماً أن الإحصاءة Wald تتبع توزيع مربع كاي χ^2 . كما أن الإختبار هو إختبار من طرفين ، ويجن أن تكون معنوية المعلمات المناظرة لقبول أو رفض

فرضية العدم بإستخدام الإحتمالات أقل من 0.05 ، لكي يتم رفض فرضية العدم وقبول الفرضية البديلة التي تنص على (أن تأثير المعامل لا يساوي الصفر في المجتمع الذي سحبت منه العينة وهو معنوي عند مستوى معنوية 0.05 ، أي أن المتغير المستقل له تأثير في التنبؤ بقيمة المتغير التابع .

التعلم الآلي

2-3 : تمهيد :

يعتبر تعلم الآلة فرع من الذكاء الاصطناعي ، يمكن أنظمة تقانة المعلومات من التعرف على الأنماط والقوانين على أساس البيانات والخوارزميات الموجودة وتطوير الحلول بشكل مستقل . وهو مصطلح جماعي لتوليد المعرفة الصناعية من التجربة ، يمكن بعد ذلك تعميم المعرفة المكتسبة من البيانات وإستخدامها لحل المشكلات الجديدة وتحليل البيانات غير المعروفة سابقاً .

2-4 : تطبيقات التعلم الآلي :

التعلم الآلي هو نهج محتمل للغاية لتحقيق التكامل بين الإنسان والحاسوب يمكن تطبيقه في العديد من مجالات الكمبيوتر . التعلم الآلي ليس طريقة نموذجية لأنه يحتوي على العديد من الخوارزميات المختلفة التي تهدف إلى حل مهام التعلم الآلي المختلفة ، و يمكن لجميع خوارزميات التعلم مساعدة الحاسوب للتصرف كإنسان .

يتم تطبيق التعلم الآلي بالفعل في العديد من المجالات كالتعرف على الأنماط والذكاء الاصطناعي والرؤية الحاسوبية وإستخراج البيانات وتصنيف النص وما إلى ذلك حيث يوفر طريقة جديدة لتطوير ذكاء الآلات ، كما أنه يصبح طريقة أسهل لمساعدة الأشخاص على تحليل البيانات من مجموعة بيانات ضخمة^[12] .

2-4-1 : المصطلحات الشائعة في التعلم الآلي^[13] :

- خوارزمية التعلم الآلي : مجموعة من القواعد والتقنيات الإحصائية المستخدمة لتعلم الأنماط من البيانات وإستخلاص المعلومات المهمة منها .
- النموذج : المكون الرئيسي لتعلم الآلة ، يتم تدريب النموذج بإستخدام خوارزمية التعلم الآلي ، تقوم الخوارزمية بتخطيط جميع القرارات التي من المفترض أن يتخذها النموذج بناءً على المدخلات المحددة ، من أجل الحصول على المخرجات الصحيحة .

- **تغير التوقع :** ميزة أو مجموعة ميزات من البيانات يمكن إستخدامها للتنبؤ بالمخرجات .
- **تغير الإستجابة :** متغير الإخراج الذي يحتاج إلى التنبؤ بإستخدام متغير أو مجموعة من المتغيرات التوقع .
- **بيانات التدريب :** يتم بناء نموذج التعلم الآلي بإستخدام بيانات التدريب التي تساعد النموذج على تحديد الإتجاهات والأنماط الرئيسية الضرورية للتنبؤ بالمخرجات .
- **بيانات الإختبار :** تقوم بدور إختبار النموذج المدرب لتقييم مدى دقة النتيجة المتوقعة .^[13]

2-4-2 : تاريخ التعلم الآلي^[14] :

- يمكن تقسيم تاريخ التعلم الآلي إلى عدة مراحل وهي :
- قبل عشرينيات القرن الماضي : وضع توماس بايز وأندرية ماركوف وأدريان ماري ليجيندر وغيرهم من علماء الرياضيات الأساس لتقنيات التعلم الآلي .
- عام 1943م تم تقدير أول نموذج رياضي للشبكات العصبية في دراسة بحثية لوالتر بيتس ووارن ماكولوتش .
- عام 1949م تم نشر كتاب " تنظيم السلوك " لدونالد هب يبحث في كيفية إرتباط السلوك بنشاط الدماغ والشبكات العصبية .
- عام 1950م آلان تورينج وصف الذكاء الصناعي وتساءل عما إذا كانت الآلات لديها القدرة على التعلم .
- 1951م بني مارفن مينسكي ودين إدموندز أول شبكة عصبية إصطناعية .
- 1956م نظم جون مكارثي ومارفن مينسكي وناثانيال روتشستر وكلود شانون مؤتمر دارتموث الذي يعتبر الأساس في صياغة مصطلح " الذكاء الصناعي " .
- 1965م طور أليسكي إيفاخنيكو الذي لقب بأبو التعلم العميق وفالنتين لابا أول إدراك متعدد الطبقات .

- 1967م تم وضع خوارزمية الجار الأقرب .
- 1979م نشر عالم الحاسوب كونييهيكو فوكوشيما عمله على النيوكونيترون شبكة هرمية متعددة الطبقات تستخدم لإكتشاف الأنماط والتي ألهمت بدورها العالم نيوكوجنيترون في وضع الشبكات العصبية التلافيفية CNN .
- 1985م اخترع تيرينس سيجنوفسكي برنامج NETTALK لتعليم اللغة الإنجليزية بنفس الطريقة التي يتعلم بها الأطفال .
- 1995م نشر تين كام هو بحثة عن غابات القرار العشوائي .
- 1997م ديب بلو ، حاسوب آي بي إم للشطرنج ، يتفوق على غاري كاسباروف بطل العالم في الشطرنج .
- 2000م تمت صياغة مصطلح " التعلم العميق " لأول مره من قبل الباحث في الشبكات العصبية إيغور أيزنبرج .
- 2009م تم إطلاق ImageNet من قبل شركة Fei-Fei قاعدة بيانات ضخمة للصور تستخدم على نطاق واسع لأبحاث التعرف على الأشياء المرئية .
- 2011م طور مختبر X خوارزمية الذكاء الاصطناعي Google Brain ، في نفس العام فاز IBM Watson على منافسية من البشر في لعبة المهلومات Jeopardy .
- 2014م إيان جودفيلو وزملاؤه قاموا بتطوير شبكة التحيز التوليدية Generative Adversary Net GAN وفي نفس العام طور Facebook نظام التعرف العميق DeepFace للتعرف على الوجوه يمكنه إكتشاف الوجوه البشرية في الصور بدقة تصل إلى 97% ثم قدمت Google نظاماً للتعلم الآلي يسمى Sibyl .
- 2015م أصبح AlphaGo أول ذكاء إصطناعي يتفوق على لاعب محترف في لعبة Go .

- 2020م أعلنت Open AI عن GPT-3 خوارزمية قوية لمعالجة اللغة الطبيعية مع القدرة على إنشاء نص يشبه الإنسان [14].

تلعب الخوارزميات دوراً مركزياً في تعلم الآله ، وهى مسؤولة عن التعرف على الأنماط وتوليد الحلول . يمكن تصنيفها وفقاً لنماذج التعلم المختلفة إلى [15] :

- التعلم المشرف عليه .
- التعلم غير المشرف عليه .
- التعلم شبه الخاضع للإشراف .
- التعلم المعزز .
- التعلم الفعال .

3-4-2: التعلم بإشراف (التعلم الخاضع للرقابة أو المشرف عليه) :

هو خوارزمية تعلم آلة شائعة الاستخدام في العديد من المجالات المختلفة لعلوم الحاسب . في طريقة التعلم بإشراف يمكن للكمبيوتر إنشاء نموذج تعليمي يعتمد على مجموعة بيانات التدريب وفقاً لنموذج التعلم هذا يمكن للحاسوب استخدام الخوارزمية للتنبؤ بالمعلومات الجديدة وتحليلها وذلك باستخدام خوارزميات خاصة . يمكن للكمبيوتر العثور على أفضل توقع وتقليل معدل الخطأ بنفسه . يستخدم التعلم بإشراف بشكل رئيسي لنمطين مختلفين : التصنيف والإنحدار .

في التعلم بإشراف عندما يعطي المطور للحاسوب بعض العينات يتم إرفاق كل عينة ببعض معلومات التصنيف سيقوم الكمبيوتر بتحليل هذه العينات للحصول على تجارب التعلم بحيث يتم تقليل معدل الخطأ عندما يقوم المصنف بالتعرف على كل الأنماط ، كل مصنف يحتوي على خوارزمية تعلم آلي مختلفة . فعلى سبيل المثال خوارزمية أشجار القرار وخوارزمية أقرب الجيران تتناسب مع مصنفين مختلفين ، لديهم مزايا وعيوب خاصة بهما حتى يتمكنوا من تحقيق أهداف التعلم المختلفة .

2-4-4: فوائد التعلم الخاضع للإشراف :

- التعلم الخاضع للإشراف مفيد لإستخلاص النتائج من الخبرة السابقة أو المعرفة السابقة .
- يمكن حل أنواع مختلفة من المشاكل الحسابية العملية .
- النتيجة أدق من طريقة التعلم غير الخاضع للإشراف .
- قبل تقديم البيانات للتدريب ، أنت تعرف بالضبط عدد الفئات الموجودة .

2-4-5: عيوب التعلم الخاضع للإشراف :

- يستغرق التدريب وقتاً طويلاً في الحساب .
- إذا كانت مجموعة البيانات التجريبية مختلفة عن مجموعة بيانات التدريب ، فإن خوارزمية التعلم الخاضع للإشراف ستواجه مشاكل عند توقع النتائج .
- يعد وضع علامات على البيانات مضيعة للوقت ومكلفاً ، وفي بعض الأحيان لا يمكن جمع مجموعة بيانات مصنفة بشكل كافٍ .
- التعلم الخاضع للإشراف محدود للغاية بحيث لا يمكن تنفيذ بعض مهام التعلم الآلي المعقدة بإستخدامه .
- لا يمكن أن يوفر التعلم الخاضع للإشراف معلومات غير معروفة من البيانات التعليمية مثل التعلم غير خاضع للإشراف .
- إذا كانت مسألة تصنيف ، إذا قدمنا مدخلات ليست جزءاً من أى من الفئات في مجموعة بيانات التدريب ، فقد يكون ناتج تسمية الفئة غير صحيح . على سبيل المثال ، إفتراض أنك قمت بتدريب مصنف الصور بإستخدام بيانات القطط والكلاب . ثم أعطيت صورة زرافة ، فقد يكون الناتج قطعة أو كلباً ، وهذا غير صحيح .

2-4-6: التعلم بدون إشراف (التعلم غير المشرف عليه) :

يستخدم أيضاً لتصنيف البيانات الأصلية ، حيث يهدف المصنف إلى العثور على معلومات التصنيف للعينات غير المصنفة . الهدف من التعلم الغير خاضع للرقابة هو ترك جهاز الحاسوب يتعلم بنفسه أى يفترض أن يقوم الحاسوب بتحليل العينات المعطاه

في هذا النوع من التعلم لا يتمكن الكمبيوتر من العثور على أفضل نتيجة يمكن أخذها ولا يعرف أيضاً إذا كانت هذه النتيجة صحيحة أم لا . عندما يتلقى الكمبيوتر البيانات الأصلية يمكنه العثور على التنظيم المحتمل ضمن البيانات تلقائياً ثم سيبنى الكمبيوتر هذا التنظيم في الحالة الجديدة وهذا هو الفرق بين التعلم بإشراف والتعلم بلا إشراف .

في بعض الحالات تكون هذه الطريقة أقوى من التعلم تحت إشراف وذلك لأنه لا توجد حاجة لإجراء تصنيف للعينات مقدماً ، في بعض الأحيان قد لا تكون طريقة التصنيف لدينا هي الأفضل ومن ناحية أخرى قد يكتشف الكمبيوتر أفضل طريقة للتصنيف بعد أن يتعلمها من العينات مراراً وتكراراً .

2-4-7: فوائد التعلم غير الخاضع للإشراف :

- يتطلب وضع العلامات على البيانات الكثير من العمل والمال . التعلم غير الخاضع للإشراف يحل هذه المشكلة من خلال التعلم من البيانات غير المسماه (غير معلمة) .
- إنه مفيد جداً في العثور على أنماط البيانات التي لا يمكن العثور عليها باستخدام الطرق التقليدية .
- يتم تقليل حجم البيانات بسهولة باستخدام هذا النوع من التعلم .

2-4-8: عيوب التعلم غير الخاضع للإشراف :

- قد تكون النتيجة أقل دقة من طريقة التعلم الإشرافي . لأننا لا نملك أي تسميات للبيانات ويجب أن يتعلم النموذج بالمعرفة المكتسبة من البيانات الأولية .
- كلما زادت الميزات ، زادت تعقيدها .
- إنها عملية تستغرق وقتاً طويلاً . لأن مرحلة تعلم الخوارزمية قد تستغرق الكثير من الوقت لتحليل وحساب جميع الاحتمالات .

2-4-9: التعلم المعزز :

في التعلم المعزز ، يحاول الوكيل حل مشكلة عن طريق التجربة والخطأ من خلال التفاعل مع بيئته طبيعتها غير معروفة للوكيل . يمكن للوكيل تغيير حالة البيئته من خلال أفعاله أثناء تلقي تعليقات فورية من البيئته . هدف الوكيل هو حل المشكلة من خلال إيجاد سلسلة الإجراءات المثلى . على الرغم من أن التعلم المعزز هو أحد مجالات التعلم الآلي ، إلا أنه يختلف إختلافاً جوهرياً عن أساليب التعلم الآلي القياسية (الإشرافي وغير الإشرافي) من نواحي كثيرة :

- لا يعتمد التعلم المعزز على تعلم البيانات . بدلاً من ذلك في التعلم المعزز يتعلم الوكيل من خبرة المكتسبة أثناء التفاعل مع البيئته ولا يعتمد على المشرف .
- يركز التعلم المعزز على إيجاد السياسة المثلى بدلاً من تحليل البيانات .

2-4-10: الفرق بين التعلم الإشرافي وغير الإشرافي والمعزز :

يحدث التعلم الإشرافي عندما يتم تغذية الخوارزمية بمجموعة من الأمثلة التعليمية بالإجابات الصحيحة (الأهداف) وتستخدم الخوارزمية هذه الإجابات كدليل للمساعدة في حل المشكلة . في المقابل لا يتطلب التعلم غير الإشرافي بيانات معنونة ، والنموذج نفسه هو الذي يحل المشكلة بمفرده ، بدون مراقب خارجي من خلال إيجاد أوجه التشابه بين المدخلات وإكتشاف الأنماط المخفية . على عكس هذين النهجين ، لا يتطلب التعلم المعزز مجموعة بيانات ويتفاعل الجهاز أو الوكيل مع بيئته للحصول على أفضل إجراء لحل المشكلة عن طريق التجربة والخطأ وتلقي المكافآت

من البيئه . التعلم المعزز ليس إشرافياً تماماً ، لأنه لايعتمد كلياً على مجموعات بيانات التدريب (المسماه) . في الواقع يعتمد التعلم المعزز على القدرة على مراقبة رد الفعل على الإجراءات المتخذة وقياسها بالمكافأه . كما أنها ليست غير إشرافية ، لأننا نعرف بالفعل متى نمثل " المتعلم " وهي المكافأه المتوقعة . بإختصار في التعلم الخاضع للإشراف الهدف هو إنشاء صيغة تعتمد على قيم المدخلات والمخرجات . في التعلم غير الخاضع للإشراف تم العثور على إتصال بين قيم الإدخال وتجميعها . في التعلم المعزز يتعلم الوكيل من خلال التفاعل مع البيئه .
بناءً على ذلك يمكن رؤية الفرق بين هذه الأساليب الثلاثة للتعلم الآلي في الجدول التالي^[16]:

جدول (1-2) : مقارنة بين التعلم الإشرافي وغير الإشرافي والمعزز^[16]

التعلم الإشرافي	التعلم غير الإشرافي	التعلم المعزز	التعريف
يتعلم من خلال مجموعات البيانات ذات العلامات	بدون توجيه ، يتم تدريبه من خلال البيانات غير المسماه	يتفاعل مع بيئة العمل	
البيانات المصنفة	البيانات غير مصنفة	لا يوجد تعريف للبيانات	نوع البيانات
التصنيف والتوقع	قواعد الارتباط والتجميع	على أساس المكافآت	نوع المسألة
مشرف إضافي	بدون مشرف	بدون مشرف	المشرف
تعيين بيانات الإدخال	إكتشاف النمط	تعلم سلسلة من الإجراءات	الهدف

11-4-2: خوارزميات التعلم الآلي^[17] :

يمكن تصنيف خوارزميات التعلم الآلي على النحو :

1. خوارزميات التعلم الآلي الخاضع للإشراف :

يتم تقسيم مجموعة بيانات الإدخال إلى مجموعة بيانات السلسلة والاختبار . تحتوي مجموعة بيانات السلسلة على متغير إخراج يحتاج إلى توقعه أو تصنيفه .

تتعلم الخوارزميات من أنماط مجموعة بيانات التدريب وتطبيقها على مجموعة بيانات الإختبار للتنبؤ أو التصنيف .

بناءً على المبدأ السابق ، تطبق عدة خوارزميات في إطار التعلم الخاضع للإشراف ، منها :

أ- **شجرة القرار Decision Tree** : الهدف الرئيسي لإستخدامها التصنيف ، يتم جمع السمات وفرزها بحسب قيمها حيث تتكون كل شجرة من عقد وفروع تمثل كل عقدة سمات في مجموعة يتم تصنيفها ويمثل كل فرع قيمة يمكن أن تأخذها العقدة .

ب- **نظرية بايز Naïve Bayes** : الهدف الرئيسي لإستخدامها التصنيف والتجميع تعتمد على الإحتمال الشرطي . يخلق الأشجار بناءً على إحتمالية حدوثها .

ت- **متجهات الدعم الآلي Support Vector Machine** : من الخوارزميات الأكثر إستخداماً على نطاق واسع ، تعتمد على مبدأ حساب الهامش حيث يتم رسم الهوامش بطريقة تجعل المسافة بين الهامش والفئات هي الحد الأقصى .

2. خوارزميات التعلم الآلي غير الخاضع للإشراف :

تستخدم بشكل أساسي في التجميع ، حيث تستخدم الميزات التي تم تعلمها مسبقاً للتعرف على فئة البيانات .

تتنوع خوارزميات التعلم غير الخاضع للإشراف لتشمل عدة تقنيات أساسية ، من أبرزها:

أ- **التجميع K-Mean Clustering** : تقوم على فرز العناصر المتشابهة في مجموعات حيث يكون متوسط القيمة في مجموعة معينة هو مركز تجمع البيانات .

ب- **تحليل العنصر الرئيسي Principle Component Analysis PCA** :

تستخدم في تقليل أبعاد البيانات لجعل العمليات الحسابية أسرع وأسهل .

3. خوارزميات التعلم شبة الخاضع للإشراف :

نهج للتعلم الآلي يجمع بين كمية صغيرة من البيانات المصنفة وكمية كبيرة من البيانات غير المصنفة أثناء التدريب ، وهو يقع بين التعلم الخاضع للإشراف مع بيانات تدريب مصنفة وغير الخاضع للإشراف بدون بيانات التدريب المصنفة ويساهم في فرز البيانات غير المصنفة بهدف إعادة تصنيفها .

يمكن تقسيم هذه الخوارزميات وفقاً للمهمة الرئيسية التي تؤديها ، ومنها :

أ- **النماذج التوليدية Generative Models** : تعتبر من أقدم طرق التعلم شبه الخاضع للإشراف ، يقوم على الإكتشاف التلقائي وتعلم الإنتظام والضبط في بيانات الإدخال من خلال إستخدام نموذج لإنشاء بيانات جديدة مماثلة يتم إستخلاصها من مجموعة البيانات الأصلية .

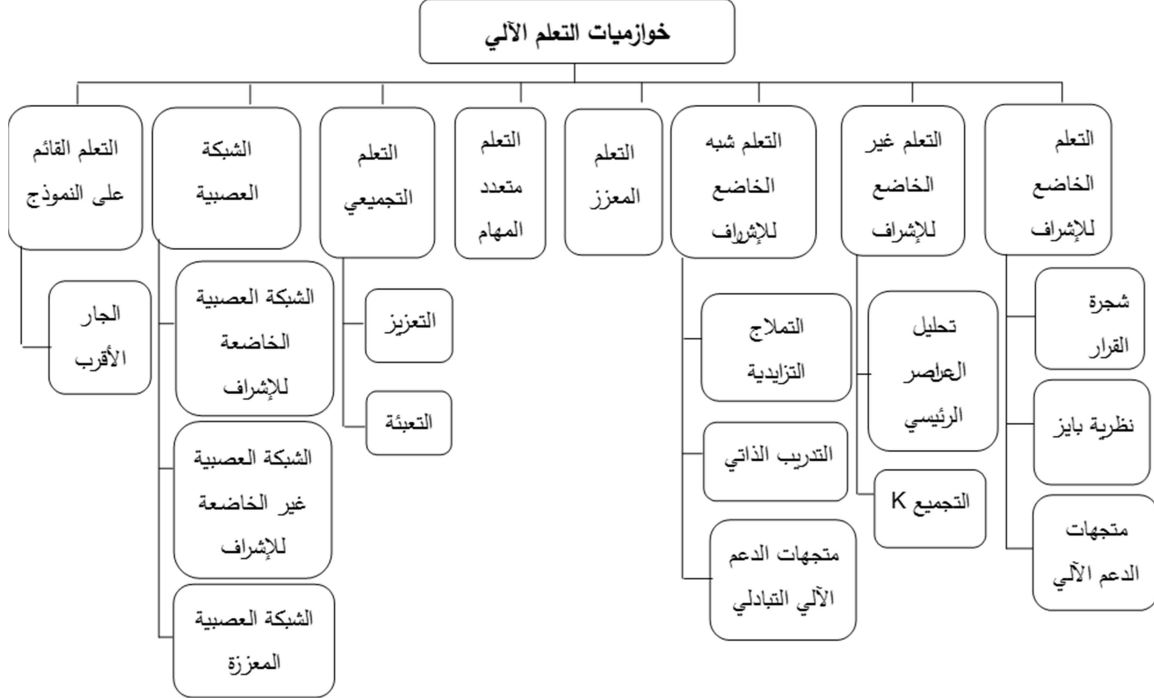
ب- **التدريب الذاتي Self-Training** : خوارزمية تعليمية شبه خاضعة للإشراف يقوم المتعلم بوضع العلامات على الأمثلة غير المصنفة وإعادة تدريب نفسه على مجموعة تدريب موسعة مصنفة ، حيث أن الحصول على بيانات مصنفة بالكامل أمراً صعباً .

ت- **متجهات الدعم الآلي التبادلي Support Vector Machine TSVM** : يستخدم بهدف تصنيف البيانات غير المصنفة من خلال جعل الهامش هو الحد الأقصى بين البيانات المصنفة وغير المصنفة .

4. **خوارزميات التعلم المعزز Reinforcment Learning** :

نوع من التعلم يتخذ القرارات بناء على إجراءات يتم تنفيذها بحيث تكون النتيجة أكثر إيجابية دون معرفة المتعلم الإجراءات التي يجب تنفيذها بشكل مسبق ، ويعتمد التعلم المعزز على معيارين فقط هما التجربة والبحث عن الخطأ والنتيجة المتأخره.

الشكل (1-2) : خوارزميات التعلم الآلي [17]



12-4-2 : التصنيف الإحصائي باستخدام التعلم الآلي [18] :

التصنيف الإحصائي هو عبارة عن عملية إحصائية يتم فيها توزيع بيانات معينة لمجتمع إحصائي على مجموعات مختلفة بناء على معلومات كمية تستند إلى واحدة أو أكثر من الخواص الأساسية لهذه البيانات أو أعضاء المجتمع الإحصائي . أو هو عملية تحديد البيانات وتصنيفها إلى فئات أو مجموعات مختلفة بناء على خصائص معينة . تتضمن هذه العملية تحليل البيانات وتنظيمها بطريقة هادفة ومفيدة لتسهيل إسترجاعها وتحليلها . يمكن تصنيف البيانات بناء على سمات مختلفة مثل النوع والتنسيق والحساسية والمصدر وغيرها . الهدف من تصنيف البيانات هو تسهيل إدارة البيانات وتخزينها وحمايتها ومساعدة المستخدمين في العثور على المعلومات التي يحتاجون إليها بسرعة وكفاءة .

تستند عملية التصنيف هذه على خاصيات أصلية في العناصر التي قد تكون (رموزاً أو متغيرات) وتستند على مجموعة تدريب من هذه العناصر .

من أمثلة خوارزميات التصنيف : بيرسيبترون ، تعلم شجرة القرار ، شبكة عصبونية إصطناعية ، المصنف بايز ساذج ، كي أقرب جار ، غابة عشوائية^[18] .

2-4-13: طرق تصنيف البيانات^[19]:

هناك عدة طرق لتصنيف البيانات الإحصائية ، وتشمل :

- تصنيف حسب المتغيرات : يمكن تصنيف البيانات حسب المتغيرات التي تتم دراستها ، مثل المتغيرات الكمية مثل (العمر والوزن) والمتغيرات النوعية (مثل الجنس والتعليم) .
- تصنيف حسب الحالة : يمكن تصنيف البيانات حسب الحالة التي تتم دراستها ، مثل البيانات الإقتصادية والبيانات الصحية والبيانات الإجتماعية .
- تصنيف حسب المصدر : يمكن تصنيف البيانات حسب مصدرها ، مثل البيانات الحكومية والبيانات الأكاديمية والبيانات المتاحة للجمهور .
- تصنيف حسب الطريقة المستخدمة في الجمع : يمكن تصنيف البيانات حسب الطريقة المستخدمة في جمعها ، مثل البيانات الإستقصائية والبيانات الواردة من مصادر ثانوية .
- تصنيف حسب الزمن : يمكن تصنيف البيانات حسب الفترة الزمنية التي تم جمعها فيها ، مثل البيانات السنوية والبيانات الشهرية والبيانات الأسبوعية^[19] .

2-4-14: يمكن تصنيف البيانات الإحصائية بعدة طرق حسب نوع البيانات والغرض من التصنيف . فيما يلي بعض الطرق الشائعة لتصنيف البيانات الإحصائية^[20] :

- **البيانات الفئوية** : يمكن تصنيف هذا النوع من البيانات بناءً على فئات أو مجموعات محددة مسبقاً . على سبيل المثال ، الإستطلاع الذي يسأل الأشخاص عن جنسهم وعمرهم ودخلهم ومستوى تعليمهم سيتم تصنيف البيانات بناءً على هذه الفئات .
- **البيانات الكمية** : يمكن تصنيف هذا النوع من البيانات بناءً على قياساتها . على سبيل المثال ، سيتم تصنيف مجموعة البيانات التي تحتوي على معلومات حول الطول والوزن والعمر لمجموعة من الأشخاص بناءً على قياساتهم .
- **البيانات الإسمية** : يمكن تصنيف هذا النوع من البيانات عادة بناءً على فئات غير مرتبة . على سبيل المثال ، سيتم تصنيف مجموعة البيانات التي تحتوي على معلومات حول ألوان السيارات في ساحة إنتظار السيارات بناءً على ألوان السيارات غير المرتبة .
- **البيانات الترتيبية** : يمكن تصنيف هذا النوع من البيانات عادة بناءً على الفئات المرتبة . على سبيل المثال ، مجموعة البيانات التي تحتوي على معلومات حول مستوى الرضا ، والذي يتم ترتيبه .
- **البيانات المنفصلة** : يمكن تصنيف هذا النوع من البيانات عادة بناءً على قيم غير مستمرة . على سبيل المثال ، سيتم تصنيف مجموعة البيانات التي تحتوي على معلومات حول عدد الأطفال في الأسرة بناءً على عدد الأطفال ، وهو عدد منفصل .
- **البيانات المستمرة** : يمكن تصنيف هذا النوع من البيانات عادة بناءً على القيم المستمرة . على سبيل المثال ، سيتم تصنيف مجموعة البيانات التي تحتوي على معلومات حول درجة الحرارة والرطوبة والضغط في موقع معين بناءً على هذه القيم المستمرة .

ويمكن إستخدام كل من هذه الطرق لتصنيف البيانات الإحصائية حسب نوع البيانات والغرض من التصنيف [20].

2-4-15: أهمية تصنيف البيانات الإحصائية وإمكانية إستخدامها في الأبحاث الإحصائية [21] :

تصنيف البيانات الإحصائية يعد من الخطوات الأساسية في معالجة وتحليل البيانات الإحصائية . فهو يمكننا من تجميع البيانات الإحصائية المتشابهة في مجموعات وتحديد العلاقات والتفاوتات بين هذه المجموعات ، وذلك بغرض تحضيرها للتحليلات الإحصائية .

تصنيف البيانات الإحصائية يمكن إستخدامها في الأبحاث الإحصائية بالعديد من الطرق ، منها :

- تمييز الخصائص الأساسية لمجموعات البيانات ، وذلك بتحديد المتغيرات الرئيسية والثانوية في البيانات وفق نوع الأسئلة المستخدمة في إستقصاء البيانات .

- تقدير المعدلات والانحرافات المعيارية لمجموعات البيانات ، وذلك بغرض التحقق من عدالة عينة البيانات وتقدير نسب الخطأ الإحصائي .

- التمييز بين مختلف المجموعات والفروق بينها ، وذلك بتحليل البيانات بإستخدام الإحصائيات المناسبة لتمثيل هذه العلاقات والفروق .

- تطوير المؤشرات الإحصائية للتعبير عن ظواهر مختلفة ، وذلك بتحليل البيانات لتحديد العلاقات بين مختلف المتغيرات وتطوير المؤشرات الإحصائية المناسبة للتعبير عن هذه العلاقات .

بشكل عام يمكن القول بأن تصنيف البيانات الإحصائية يمثل خطوة مهمة في معالجة وتحليل البيانات ، ويساعد على تحليل وفهم الظواهر الإحصائية وتطوير المؤشرات الإحصائية المناسبة للتعبير عن هذه الظواهر [21] .

2-4-16: أنواع تصنيف البيانات^[22] :

هنالك عدة أنواع من تصنيف البيانات ، ومن أهمها :

- **تصنيف ثنائي Binary Classification** : يتم فيه تصنيف البيانات إلى فئتين رئيسيتين فقط ، مثل التصنيف إلى "نعم" أو "لا" ، أو "إيجابي" أو "سلبي" .
 - **تصنيف متعدد Multi-class Classification** : يتم فيه تصنيف البيانات إلى أكثر من فئة ، مثل تصنيف الصور إلى فئات مختلفة مثل "كلب" أو "قطه" أو "طائر" .
 - **تصنيف تسلسلي Sequential Classification** : يتم فيه تصنيف البيانات التي تأتي في تسلسل معين ، مثل تصنيف الكلمات في جملة إلى أجزاء مختلفة مثل "فعل" أو "إسم" أو "حرف جر" .
 - **تصنيف تجميعي Hierarchical Classification** : يتم فيه تصنيف البيانات في تسلسل هرمي ، حيث تكون هناك فئات رئيسية وفئات فرعية تتدرج تحتها ، مثل تصنيف الحيوانات إلى فئات رئيسية مثل "ثدييات" أو فئات فرعية مثل "كلب" أو "قطه" تتدرج تحت فئة الثدييات .
- هذه بعض الأمثلة الشائعة لأنواع تصنيف البيانات ، وهناك أنواع أخرى أيضاً تعتمد على طبيعة البيانات ونوع المشكلة المطلوب حلها^[22].

خوارزمية الغابة العشوائية

2-3-1: أشجار القرار Decision Trees :

في عالم خوارزميات تعلم الآلة ، تعد الغابة العشوائية Random Forest واحده من أنجح النماذج وأكثرها موثوقية في حل مسائل التصنيف والانحدار . لكن سر نجاحها لا يكمن في فكره معقدة غامضة ، بل في مفهوم جماعي بسيط وقوي : تجميع العديد من المتنبئين الضعفاء لإنشاء متنبئ قوي . والمتنبئ الضعيف هنا هو شجرة القرار Decision Tree .

لذا تعتبر شجرة القرار هي بالفعل القلب النابض واللبنه الأساسية في بناء خوارزمية الغابة العشوائية لذا سنتطرق إليها في بدايه هذا المبحث أولاً .

2-3-1-1 : مفهوم شجرة القرار :

تعد عملية إتخاذ القرار جوهر العملية الإدارية بحكم تركيزها على تقديم الحلول للمشكلات الإدارية والتنظيمية والسلوكية والتطويرية . فهو العملية المبنية على الدراسة والتفكير الموضوعي الواعي للوصول إلى إختيار أحد البدائل المطروحة الذي يتناول حل مشكلة قائمة أو إقتناص فرصة متوقعة للحصول عليها في المستقبل^[23] فالقرارات قد تتخذ بشكل متتابع أى قد تكون متسلسلة حيث يتوقف بعضها على البعض الآخر وأن ناتج القرار يؤثر في الذي يليه ، لذا جاءت شجرة إتخاذ القرار كأداة يهدف إلى تسهيل عملية صنع وإتخاذ القرار . وبهذا يمكن تعريف شجرة إتخاذ القرارات على أنها تمثيل بياني لعملية صنع وإتخاذ القرار تعرض فيه الإستراتيجيات والحالات الطبيعية والعوائد لمساعدة متخذ القرار في إختيار القرار الصائب^[24] . إذ ترجع جذور أسلوب شجرة إتخاذ القرارات إلى مدخل النظم في إتخاذ القرارات والذي يقوم على التفاعل بين الأدوات والوسائل المستخدمة لإتخاذ القرار وبين البيئة المحيطة بإتخاذ القرار . كما يفترض هذا المدخل أن هناك سلسلة من التأثيرات تؤثر في عملية إتخاذ القرارات . بمعنى أن إتخاذ القرار في نظام فرعي ينتج تأثيراً يكون له ردود فعل تنتشر في سلسلة متعاقبة في النظام وفي بيئته . ومن

هنا يجب على متخذ القرار أن يأخذ في الحسبان سلسلة التأثيرات هذه بقدر الإمكان عند إتخاذ القرار [25]. وسميت شجرة إتخاذ القرار بذلك لأنها تشبه الشجرة على الرغم من أنها تكون أفقية وليست عمودية أي رأسية وأساس هذه الشجرة هو نقطة القرار ، أما قراراتها فتتمدد من نقطة القرار والأحداث ، وكل حادثة تنتج أو تؤدي إلى نتيجة أو أكثر والتي قد تؤدي بعضها إلى أحداث أخرى وقرارات تالية [26]. وتتكون شجرة إتخاذ القرار من ثلاثة متغيرات هي البدائل المطروحة لحل المشكلة موضوع القرار والإحتمالات التي تمثل الكسب المتوقع أو الفشل والقيم التي تمثل إجمالي العوائد المتوقعة خلال فترة محددة . فإذا كان هناك مثلاً مشكلة يواجهها المدير وكان أمامه بديلان وقرر إختيار أحدهما فإنه سينتقل إلى حالة جديدة قد يواجهها فيها ثلاثة بدائل للإختيار . وطبقاً لهذا الأسلوب فإن تحديد ما إذا كان الحدث المتوقع حدوثه مناسباً أم لا يتطلب من المدير متخذ القرار أن يحلل نتائج القرار (القيم) وإحتمالاته من نهاية الشجرة إلى بدايتها حتى يصل إلى إختيار البديل المناسب على ضوء معايير وإعتبارات أهمها : النتائج المتوقعة من كل بديل ، الإمكانيات والموارد المتاحة لتنفيذ البديل ، ودرجة المخاطرة ودرجة النجاح المتوقعة من البديل [25].

أشجار القرار هي طريقة تعلم خاضع للإشراف ، تستخدم للتصنيف والإنحدار (Regression) على حد سواء . الهدف منها خلق نموذج للتنبؤ بقيمة معينة عن طريق تعلم قواعد بسيطة مستنتجة من خصائص البيانات . تطبق عملية التصنيف عن طريق مجموعة من القواعد أو الشروط التي تحدد المسار الذي سيتبع ابتداء من عقدة الجذر وإنتهاء بإحدى العقد النهائية التي تمثل القرار النهائي . وينبغي عند كل العقد غير النهائية إتخاذ قرار حول مسار العقد التالية . والقرار في حد ذاته هو إختيار حل من بين عدة حلول لمشكلة معينة . وعليه فإن إتخاذ القرار هو إختيار أحد البدائل المتاحة ، لذلك فعملية إتخاذ القرار هي مجموعة متتالية من الخطوات والإجراءات التي تؤدي في نهايتها إلى إختيار أفضل الحلول البديلة . يتميز هذا الأسلوب من تمكين متخذ القرار من رؤية البدائل المتاحة والأخطار والنتائج المتوقعة لكل منها

بوضوح . ويستعمل أسلوب شجرة القرار في حل المشكلات ذات البدائل المتعددة ، وكذلك الحالات المتعددة المحتمل مواجهتها ، خاصة عندما تكون المشكلة متعلقة بعنصر المخاطره وعدم التأكد . تعرف شجرة القرار أيضا بأنها مجموعة محددة من العقد بحيث أن هنالك عقدة مميزة تقع في أعلى الشجرة تدعى بجذر الشجرة ، وكل شجرة من المجاميع أسفلها عبارة عن شجرة تدعى الأشجار الفرعية (Subtree) للجذر . القاعدة الأساسية في بناء شجرة القرار هي إيجاد أفضل سؤال عند كل فرع من فروع الشجرة بحيث يقسم هذا السؤال البيانات إلى قسمين ، القسم الأول منها ينطبق عليهم السؤال والقسم الثاني لا ينطبق ، وهكذا يتم من خلال سلسلة من الأسئلة بناء شجرة القرار بفروعها المتسلسلة [27].

1. نقاط الإنبثاق : Nodes

وهي نقاط إنطلاق يمكن تصنيفها إلى :

- نقاط قرار : Decisions Node وهي النقاط التي يتعين عندها إتخاذ قرار معين (تبني أحد البدائل أو الإستراتيجيات) ويرمز لها بمربع صغير □
- نقاط الأحداث Chance Nodes وهي نقاط تشير إلى الأحداث أو الظروف (حالات طبيعية) ويتم تمثيلها بشكل دائره صغيرة ○

2. الفروع المنبثقة من النقاط : Braches وهي أشبه بأغصان الشجرة تنبثق من النقاط المختلفة وكالاتي :

- فروع قرار Decisions Branches وهي فروع تنبثق من نقاط القرار Decisions Node وتمثل الإستراتيجيات أو الخيارات المتاحة لمتخذ القرار (تمثل في بعض الأحيان بخطين متوازيين لتمييزها عن الفروع الأحداث) .
- فروع الأحداث : Chance Branches وتمثل الحالات الطبيعية أو الظروف الخارجية .
- الفروع النهائية : Terminal Branches وهي الفروع التي لاتنتهي بنقطة حدث أو قرار .

3. إحتتمالات حصول حالات طبيعية : Probabilities يتم تحديدها من واقع

الخبره الشخصية لمتخذ القرار أو السجلات التاريخية التي يحتفظ بها .

4. العوائد : Outcomes وهي النتيجة التي تحصل عند تبني إستراتيجية معينة

وحصول حالة طبيعية محددة وقد تكون موجبة أو سالبة [24].

يدور التعلم الآلي أساساً حول التنبؤ بالمستقبل على أساس الماضي ويعد مفهوم أشجار إتخاذ القرار من أول الأساليب التي إستخدمت في تعلم الآلة . تعتبر شجرة القرار نموذجاً كلاسيكياً وطبيعياً للتعلم . أنه أسلوب يستخدم هيكل الأشجار لتمثيل عدد من مسارات إتخاذ القرارات المحتملة ومخرجات كل مسار منها . إن شجرة القرار تتميز بالسهولة في فهمها وتتبعها وتفسيرها . تنقسم أشجار القرار في تعلم الآلة إلى ثلاثة أشجار تمثل البيانات الرقمية ، وآخرى تمثل بيانات الفئات والأخيرة تمثل البيانات المفقودة .

تتكون أشجار القرار من عقد ، وتكون محدده بعقده رئيسية تسمى الجذر . والتي ليس لها أي مدخلات بينما كل العقد الأخرى لها مدخل واحد . والعقد التي لها مخرج تسمى داخلية أو إحتمالية . وكل منها تنقسم إلى عقدتان تمثل كل منها دالة بقيم خصائص . وتتكون العقد من بيانات رمزية أو رقمية .

يمكن أن تستخدم أشجار التصنيف البيانات الكمية والفئوية لنمذجة النتائج الفئوية ، تستخدم أشجار الإنحدار أيضاً البيانات الكمية والفئوية ولكن بدلاً من ذلك تقوم بنمذجة النتائج الكمية . تبدأ أشجار القرار بعقدة الجذر . تعمل كنقطة بداية (في الأعلى) ويتبعها إنقسامات تنتج فروعاً ترتبط الفروع بعد ذلك بالأوراق والتي تعرف أيضاً بإسم العقد والتي تشكل نقاط إتخاذ القرار . يتم إنتاج التصنيف النهائي عندما لا تولد الورقة أى فروع جديدة وتؤدي إلى ما يعرف بالعقدة الطرفية وبالتالي فإن أشجار القرار لاتتفكك وتشرح فقط كيفية صياغة التصنيف أو الإنحدار ولكنها تنتج أيضاً مخططاً إنسيابياً مرئياً يمكنك إظهاره للآخرين . سهولة التفسير هي ميزة قوية لإستخدام أشجار القرار ويمكن تطبيقها على مجموعة واسعة من حالات الإستخدام .

2-1-3-2 : تعقيد شجرة إتخاذ القرار :

كلما زاد تعقيد شجرة إتخاذ القرار كلما قلت كفاءتها وإهتم الباحثون بما اسموه " آليات التوقيف " لإيقاف نمو الشجرة ويكون ذلك بآليات التشذيب ويقاس تعقيد الشجرة بالعوامل التالية :

- العدد الكلي للعقد فيها .
- العدد الكلي للأوراق النهائية .
- عمق الشجرة .
- عدد الخصائص المستخدمة .

إن إشتقاق شجرة القرار هي خوارزمية تبنى إعتماداً على مجموعة من البيانات المعرفة وسيكون هدفنا الوصول إلى أفضل شجرة إتخاذ قرار وذلك من خلال تقليل العقد والعمق والتكرار وتنشأ بطريقتين :

- الأولى : من أعلى إلى أسفل وهي الأكثر إستخداماً .
- الثانية : من أسفل إلى أعلى .

2-1-3-3 : آليات إيقاف توسع شجرة القرار :

- يجب أن تنتمي كل عناصر مجموعة التدريب لقيمة واحدة .
- الوصول إلى الحد الأعلى من عمق الشجرة .
- أن لا تتجاوز العتبة المحددة من العقد .

التشذيب :

هي مجموعة من الآليات التي تقوم بحذف الخيارات الفرعية التي تقلل من دقة شجرة إتخاذ القرار . وتحسن بالتالي من أدائها وتربط تلك الآليات شجرة إتخاذ القرار بمفهوم " الدقة ترفع من البساطة " وهي تفترض بأن دقة الشجرة ترتفع عند الإقتصار على الشجرة الأساسية .

وهناك العديد من الطرائق الشائعة لتشذيب أشجار القرار ومنها :

- طريقة تشذيب الرابط الأضعف .
- طريقة تشذيب الأخطاء المكررة .
- طريقة تشذيب الأخطاء .
- طريقة تشذيب الخيارات المتشائمة .
- طريقة تشذيب المبنى على حذف الأخطاء .
- طريقة تشذيب التحسين .
- طريقة تشذيب تقليل وصف الطول .

4-1-3-2 : أنواع أشجار إتخاذ القرار :

- شجرة (ID3) : تعد هذه الخوارزمية الأبسط والأكثر بدائية بين أنواع أشجار القرار وهي لاتطبق أي من آليات التشذيب أو إيقاف النمو وهي لاتتعامل مع الخصائص الرقمية ولاتطبق آلية معالجة فاقد البيانات .
- شجرة (C4.5) : وهي تطوير لشجرة (ID3) تستطيع إضافة خصائص البيانات الرقمية وتستخدم للتشذيب طريقة المبنى على حذف الأخطاء .
- شجرة (CART) : وتسمى أيضاً بالأشجار الثنائية ، وذلك لأنه يجب أن يتفرع من كل عقدة فيها خياران فقط وهي الأنسب للتصنيف والانحدار، تستخدم للتشذيب طريقة (حذف الرابط الأضعف) وهي الأفضل في خلق أشجار الانحدار والتي تنتبأ أوراقها بأرقام حقيقية وليست صفوفاً .
- شجرة (CHAID) : طبق الإحصائيون منذ السبعينات آليات الإحصاء على أشجار القرار ومنها شيد وهي إختصار (التحديد والتفاعل الآلي مع مربع كاي) وإستخدمت لتحديد الخصائص الرمزية غالباً تطبق شيد معادلات إحتساب فاقد البيانات. ولا تستخدم أي من آليات التشذيب .

- شجرة (QUEST) : وهو اختصار (أشجار إحصائية تتميز بالفعالية والمرونة والسرعة) تستخدم إختبارات نوبا وليفان للخصائص المتصلة ، ومربع كاي للخصائص الرمزية .

2-3-1-5 : أشجار القرار العشوائية :

بدلاً من السعي لتحقيق الإنقسام الأكثر فاعلية في كل جولة من التقسيم التكراري ، تتمثل التقنية البديلة في إنشاء عدة أشجار والجمع بين تنبؤاتها لتحديد المسار الأمثل للتصنيف أو التنبؤ يتضمن ذلك إختيار عشوائياً للأسئلة الثنائية للنمو العديد من أشجار القرار المختلفة والمعروفة بإسم الأشجار العشوائية ، تتم مقارنة النتائج من كل شجرة والتصويت عليها لإنشاء شجرة مثالية لإنتاج النموذج النهائي، والمعروف بإسم الفئة النهائية .

2-3-1-6 : بناء شجرة اتخاذ القرار :

يتم إنشاء أشجار القرار من خلال تقسيم البيانات أولاً إلى مجموعتين ثم تتكرر عملية الإنقسام الثنائي هذه في كل فرع (طبقة) والهدف هو تحديد السؤال الثنائي الذي يقسم البيانات بشكل أفضل إلى مجموعتين متجانستين في كل فرع من فروع الشجرة ، بحيث يقلل من مستويات البيانات في المجموعة التالية. الأنتروبيا Entropy هو مصطلح يعني معدل محتويات البيانات .

ما يجب أن تذكره عند إستخدام أشجار القرار هو قابليتها للإفراط في الحجم ، سبب فرط الحجم في هذه الحالة هو بيانات التدريب مع الأخذ في الإعتبار الأنماط الموجودة في بيانات التدريب الخاصة بك ، تكون شجرة القرار دقيقة في تدريب الجولة الأولى من البيانات ومع ذلك قد تفشل شجرة القرار نفسها بعد ذلك في التنبؤ ببيانات الإختبار، حيث قد تكون هناك قواعد لم تواجهها بعد أو لأن بيانات التدريب أو الإختبار لم تكن ممثلة لمجموعة البيانات بأكملها [28].

2-3-1-7 : خطوات رسم شجرة القرارات :

يبدأ رسم شجرة القرار من اليسار إلى اليمين وفي كل مربع هناك سيطرة لإتخاذ القرار وفي كل دائرة يأخذ المصير بخصوص عمله والنقاط التالية وصف لكل خطوه من خطوات بناء شجرة القرارات (step-by-step) وهي :

1. تبدأ من خلال رسم مربع صغير، ومن ثم خط خارج إلى اليمين لكل بديل من بدائل القرار المحتملة ، ينبغي التأكد من وجود مسافة كافية بين الخطوط وتسمية كل خط للبديل الذي يمثلها .

2. في نهاية كل خط من اليمين ، يشير إلى نوع من النتائج المحتملة ، إذا كانت النتيجة وجود بديل محدد غير مؤكد وبالتالي إذن يرتبط في نهاية الخط بدائرة صغيرة ، إذا كانت النتيجة وجود بديل يكون قرار آخر، إذن يرتبط في نهاية الخط بمربع صغير وبتعبير آخر، أي المربعات تمثل دائماً القرارات والدوائر تمثل عادة العوائد غير المؤكدة .

3. ومن الدوائر، رسم خطوط باتجاه اليمين لعوائد محتملة ، لكل تسمية وتضم الإحتمالية لحدوث العائد ، والإحتمالات للعوائد المنبثقة من الدائرة يجب أن يكون مجموعها واحد (1.0) .

ومن المربعات، رسم خطوط باتجاه اليمين يمثل كل واحد من هذه الخطوط خيار من خيارات القرار، ينبغي تسمية جميع الخطوط لمعرفة ما تعنيه ، إذا كانت هناك لا توجد تقريباً قرارات أو بدائل فتترك نهاية الخط من جهة اليمين فارغة .

4. تكرار هذه العملية لغاية رسم الشجرة أفقياً والتي تمثل جميع القرارات المحتملة وجميع العوائد المحتملة .

5. تحديد الإيراد أو القيمة النقدية لكل عائد محتمل ، العوائد تكون في نهاية كل خط من جهة اليمين .

2-3-1 : تحليل شجرة القرار :

بعد رسم شجرة القرار، تُحلل الشجرة من خلال إيجاد النتائج من اليمين إلى اليسار، يسمح تحليل شجرة القرار لإدارة وتنظيم مشكلة القرار عن طريق تحديد جميع النتائج المحتملة وكذلك يكون مدخلاً يساعد الإدارة لتحديد أي بديل لأي خيار محدد وإتخاذ الخيار الأعلى ربحاً .

تعد القيمة النقدية المتوقعة (EMV) المعيار الأكثر إستخداماً لتحليل شجرة القرارات والخطوة الأولى للتحليل هي رسم شجرة القرار وتحديد النتائج النقدية لجميع العوائد لمشكلة محددة ونجد القيمة النقدية المتوقعة (EMV) لتحليل شجرة القرار بإستعمال الطي الخلفي لشجرة القرار وتدعى أيضاً الرجوع للخلف .

يمكن تحديد القيمة النقدية المتوقعة لكل بديل حيث تُحسب في المعادلة الآتية :

$$\text{القيمة النقدية (EMV) المتوقعة للبديل (i)} = (\text{عائد حالة الطبيعة الأولى}) \times (\text{إحتمال حالة الطبيعة الأولى}) + (\text{عائد حالة الطبيعة الثانية}) \times (\text{إحتمال حالة الطبيعة الثانية}) + \dots + (\text{العائد لحالة الطبيعة الأخير}) \times (\text{إحتمال لحالة الطبيعة الأخير}) .$$

نبدأ بالعائد أي عند العقدة النهائية للفروع الموجودة في أقصى الطرف الأيمن للشجرة والتوجه للخلف إلى عقدة القرار الأولى (initial) وتحديد بعض قيم التراجع إلى الخلف لكل عقدة .

تستعمل الخطوات التالية لتحليل شجرة القرار وهي:

- تحديد المشكلة .
- بناء أو رسم شجرة القرار .
- تحديد الإحتمالات لحالات الطبيعة .
- تقدير العائد لكل مكون محتمل من بدائل القرار وحالات الطبيعة .
- حل المشكلة من خلال حساب القيمة النقدية المتوقعة لكل عقدة من حالات الطبيعة .

يحسب العائد المتوقع لكل عقدة على النحو الآتي :

- في العقدة النهائي (Node Terminal) : قيمة الرجوع للخلف يساوي القيمة النهائية .

- في عقدة الحدث (Node Event) : تحديد قيمة الرجوع للخلف لمتخذ القرار يكون التحديد بإستعمال قيمة العائد المتوقعة من خلال ضرب العائد لكل حدث فرعي في إحتمالية الحدث ، ثم تُجمع النتائج ليتم الحصول على قيمة العائد المتوقع لعقدة الحدث .

في عقدة القرار تحدد قيمة الرجوع للخلف مساوياً لأعلى قيمة للرجوع وبتعبير آخر في عقدة القرار تكون القيمة المتوقعة (EV) مساوية للقيمة المتوقعة الأمثل في العقدة أي تكون (الحد الأدنى أو الحد الأعلى لكل معيار أمثل مطلوب) .

نقوم بتقليم الشجرة وذلك لعمل الحسابات وذلك بإزالة جميع الفروع من كل عقدة أو نقطة قرار ماعدا الفرع الذي له أعلى قيمة للعائد والفروع التي لم تُختار يوضع فوقها علامة خطين مائلين (//) (lines Slash) أي إختيار عقدة القرار للعائد المتوقع يكون مرتبطاً بفرع غير مقلم (unprune) باقي لمفرده وهذه العملية تستمر إلى عقدة القرار الأول وبذلك تُحل مشكلة القرار .

يبدل الفرع غير المقلم إلى من هو أفضل بديل وذلك للمتابعة ، وإذا كانت الشجرة تتضمن قرارات متعددة المراحل (عدة قرارات) ينبغي إنتظار الأحداث اللاحقة قبل إتخاذ القرار وذلك لما يجب أن تعمله للتالي وإذا تم الحصول على إحتمال جديد أو تقدير للعائد ، عندئذ تُكرر هذه العملية [29].

2-3-1 : مميزات أشجار القرار^[30]:

- سهولة الفهم والتفسير ويمكن ذلك من خلال تصور الأشجار . وينتج عنها مجموعة من القواعد .
- قدرة على التعامل مع البيانات العددية والفئوية .
- يمكن التحقق من صحة النموذج بإستخدام الإختبارات الإحصائية . وهذا يجعل من الممكن حساب كفاءه النموذج بدقة عالية .
- تتبع هذه الخوارزمية نفس النهج الذي يتبعه البشر بشكل عام أثناء إتخاذ القرارات
- عدد المعلومات الفائقة المراد ضبطها تكاد تكون خالية .
- مقاومة للخطأ ، وإذا كانت بيانات التدريب تحتوي على أخطاء ، فإن خوارزميات شجرة القرار ستكون أكثر ملاءمة لمعالجة مثل هذه المشكلات .
- تنتج قواعد مفهومة وبديهية للغاية .
- يمكنها إدارة المتغيرات المستمرة والمتقطعة .
- طريقة عملها بسيطة للغاية ويمكن شرحها بسهولة لأي شخص .
- ليس لديها أي إفتراضات حول خطية البيانات وبالتالي يمكن إستخدامها عندما تكون المعاملات غير الخطية ذات صلة .
- توفر وقت إعداد البيانات ، لأنها ليست حساسة للقيم المفقودة والقيم البعيدة .
- ليست هناك حاجة لتوحيد وتحجيم البيانات .
- تعتبر مفهوم شجرة القرار أكثر دراية للمبرمجين وأسهل في الفهم من الخوارزميات المماثلة الأخرى .

2-3-10 : عيوب أشجار القرار^[30] :

- تنبؤات أشجار القرار ليست سلسلة ولا مستمرة ، ولكنها تقديرات تقريبية ثابتة متعددة المستويات . لذلك فهذه الخوارزمية لا تجيد الإستقراء .
- هناك مفاهيم يصعب تعلمها لأن أشجار القرار لا تعبر عنها بسهولة . مثل مشاكل XOR أو التكافؤ أو متعدد الإرسال .
- هناك إحتمال كبير لظهور مشاكل الفرط في التناسب Overfitting في أشجار القرار .
- عرضة للأخطاء في مشاكل التصنيف مع الفصول الكبيرة وعدد صغير نسبياً من الأمثلة التعليمية .
- بالنسبة لشجرة القرار ، قد يكون الحساب أحياناً أكثر تعقيداً من الخوارزميات الأخرى .
- غالباً ما يستغرق تدريب النموذج مزيداً من الوقت .
- يمكن أن تؤدي إضافة نقطة بيانات جديدة إلى إعادة إنتاج الشجرة ، ويجب إعادة حساب جميع العقد وإنشائها .
- غالباً ما تكون شجرة القرار المنفردة متعلماً ضعيفاً ، لذلك من أجل التنبؤ الأفضل ، نحتاج إلى مجموعة من أشجار القرار لإنشاء غابة عشوائية .
- عرضة للضبط الزائد من أجل ملاءمة البيانات (حتى بيانات الضوضاء) ، تستمر في إنشاء عقد جديدة ، وفي النهاية تصبح الشجرة معقدة للغاية بحيث لا يمكن تفسيرها بهذه الطريقة يفقد قابليته للتعميم .
- بسبب الضبط الزائد ، هناك إحتمال كبير للتباين (من أجل تحقيق صفر تحيز ، مما يؤدي إلى تباين كبير) في المخرجات ، مما يؤدي إلى العديد من الأخطاء في التقدير النهائي ويظهر دقة منخفضة في النتائج^[30] .

2-3-11 : خوارزمية الغابة العشوائية Random Forest :

2-3-12 : مفهوم الخوارزمية :

إن التعريف البسيط لكلمة خوارزمية (Algorithm) إنها طريقة أو خطة أو قاعدة للوصول إعتباراً من معطيات إلى نتائج ، ونستطيع صياغة تعريف آخر أكثر دقة كالتالي هي عبارة عن مجموعة من الخطوات المتسلسلة التي تصف بصورة مضبوطة وبدون أى غموض جميع الخطوات الرياضية والمنطقية اللازمة لحل مسألة ما ، وقد تطور هذا المفهوم وأصبح يعني طريقة أو خطة شاملة وعامة لحل مسألة ما ، وتقوم بوصف كافة الخطوات بشكل مفصل [31].

أو هي مجموعة من البرامج توضع بشكل عام وبقواعد عامة لمعالجة البيانات المدخلة بكافة الأشكال وتقوم بإيجاد العلاقات والأنماط في البيانات عن طريق تطبيق معادلات إحصائية ورياضية حيث تتسم كل خوارزمية بصفات ومخرجات معينة لتستطيع تمثيل البيانات بطرق مختلفة أو التنبؤ بمخرجات البيانات الجديدة بناءً على العلاقات والأنماط المستنتجة من البيانات المدخلة .

2-3-13 : تمهيد في الغابة العشوائية [32].

في السنوات الأخيرة زاد إهتمام وشعبية نموذج الغابات العشوائية زيادة كبيرة ، ويعود ذلك في معظمه إلى حقيقة أنه يمكن تطبيقه بسرعة على أي مشكلة في علم البيانات تقريباً . مجموعة من النتائج القياسية في مجموعة متنوعة من المشاكل أظهرت الغابات العشوائية قوة لاتصدق مراراً وتكراراً ، لكنها في نفس الوقت مريحة وعملية للغابة

تعتبر Random Forest أحد الأساليب الشائعة في تعلم الآلة والذكاء الاصطناعي ، وتستخدم بشكل واسع في تصنيف البيانات وتوقع القيم . وهي مجموعة من أشجار القرار Decision Trees التي تستخدم بشكل جماعي للحصول على تنبؤات دقيقة ، تقوم خوارزمية RF على مبدأ إنشاء عدة أشجار قرار مستقلة ثم دمج نتائجها للحصول على التنبؤ النهائي .

2-3-14 : تعمل خوارزمية الغابة العشوائية من خلال عدة خطوات^[32] :

1. بناء الأشجار وإختيار العينة العشوائية : يتم بناء كل شجرة من خلال أخذ عينة عشوائية من البيانات التدريبية ، بحيث يمكن أن يكون لبعض العينات أن تظهر أكثر من مره .

أ- إنشاء شجرة القرار : يتم بناء شجرة القرار من خلال تقسيم البيانات بإستخدام أفضل ميزات ، وتستمر عملية التقسيم حتى تصل الشجرة إلى عمق محدد أو يتم تلبية شروط التوقف الأخرى .

ب- إختيار الميزات العشوائية : عند كل تقسيم في الشجرة ، يتم إختيار مجموعة عشوائية من الميزات (وليس جميعها) لتحديد أفضل تقسيم . هذا يعزز تنوع الأشجار في الغابة ويقلل التباين بينهم .

2. التصنيف أو التنبؤ .

أ- التصنيف : عند تصنيف مثال ، يتم تمرير هذا المثال عبر جميع الأشجار في الغابة وكل شجرة تصدر تصنيفاً . والتصنيف النهائي يتم تحديده بناءً على التصويت الأكثر شيوعاً من جميع الأشجار .

ب- التنبؤ : في حالة التنبؤ بالقيم ، يتم حساب متوسط التنبؤات من جميع الأشجار لتحديد القيمة النهائية .

2-3-15 : تاريخ الغابة العشوائية :

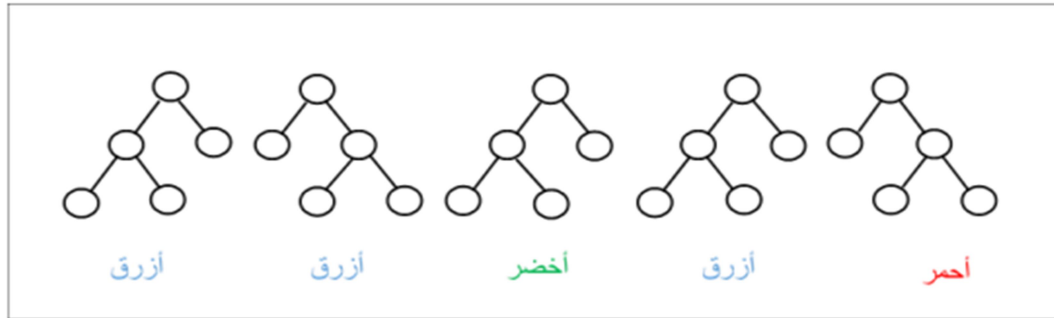
عندما نتحدث عن توليد وتطوير خوارزمية الغابات العشوائية ، يجب أن نعود إلى الثمانينات . يمكن القول إن هذه الخوارزمية هي تبلور للجهود المضنية التي بذلها ليوبريمان وأديل كاتلر وهوتين كام وديتريخ وأميت وجيمان ، وقد ساهم كل منهم إسهاماً مهماً في التطوير المبكر لخوارزمية الغابات العشوائية . إقترح Leo براءات الإختراع الخاصة بهما . قدم كل من Amit و Gemen و Ho Tim Kam فكرة إختيار عشوائي للمميزات بشكل مستقل ، وإستخدموا فكرة Breiman في التعبئة لبناء مجموعة من أشجار القرارات . التي تتحكم في التباين . بعد ذلك قدم

Deitterich فكرة تحسين العقدة العشوائية في النموذج لزيادة تحسين Senry العشوائي .

2-3-16 : ماهي الغابة العشوائية :

تعتبر Random Forest كمجموعة من أشجار القرار ، يتم تقسيم مجموعة بيانات الإدخال إلى أجزاء فرعية وتغذيتها في أشجار القرار المختلفة ، يتم أخذ متوسط المخرجات من جميع أشجار القرار في الاعتبار ثم إستخراج النتيجة النهائية . توفر الغابات العشوائية تصنيفاً أكثر دقة وقوة مقارنة بخوارزمية شجرة القرار على سبيل المثال : خمسة أشجار قرار لتصنيف الألوان ، 3 أشجار من أصل خمسة تنتبأ باللون الأزرق وإثنتان لهما مخرجات مختلفة ، وهما الأخضر والأحمر ، نأخذ متوسط جميع النواتج ، فيكون اللون الأزرق هو الأعلى وزناً .

الشكل (2-2) : الغابة العشوائية Random Forest^[32]



وتعتبر أيضاً خوارزمية الغابات العشوائية من خوارزميات التعلم الآلي تم تطويرها بناءً على مجموعة من أشجار القرار ، تستخدم هذه الخوارزمية للتصنيف ، الإنحدار ومهام أخرى . عادة ما تتمتع خوارزمية الغابة العشوائية بدقة أفضل مقارنة بشجرة القرار^[32] . وهي عبارة عن خوارزمية تعليمية متعددة الإستخدامات قادره على أداء مهام الإنحدار والتصنيف في الوقت نفسه ، إنها أيضاً طريقة للحد من أبعاد

البيانات وتستخدم للتعامل مع القيم المفقودة ، والقيم المتطرفة وغيرها من الخطوات المهمة في إكتشاف البيانات وحقت نتائج جيدة بالإضافة إلى ذلك تعمل أيضاً كطريقة هامة في التعلم الجماعي ، حيث تظهر قوتها عند دمج عدة نماذج غير فعالة في نموذج واحد فعال .

في الغابات العشوائية سنقوم بإنشاء الكثير من أشجار القرار وليس فقط توليد الأشجار الفريدة كما هو الحال في نموذج CART عند تصنيف كائن جديد وتمييزه بناءً على سمات معينة ، ستعطي كل شجرة في الغابات العشوائية خيار التصنيف الخاص بها و " التصويت " وفقاً لذلك ، وسيكون الناتج الكلي للغابة هو أكبر عدد من الأصوات . خيار التصنيف ؛ في مشكلة الانحدار ، سيكون ناتج الغابة العشوائية هو متوسط جميع مخرجات شجرة القرار .

كما أن خوارزمية الغابة العشوائية Random Forest تعتبر إحدى خوارزميات التعلم الخاضعة للإشراف Supervised أي أن مخرجات الغابة العشوائية يجب أن تتطابق مع متغيرات الهدف وبمقارنتها تنتج أخطاء التنبؤ وتعتمد على مبدأ تقنيات أشجار التصنيف والانحدار ومن مميزاتها أنها دقيقة حسابياً وتعمل بسرعة وذلك عبر بيانات كبيرة نسبياً وهي من التقنيات الحديثة حيث يتم إستخدامها في العديد من التطبيقات في مجالات متنوعة لإعتمادها على مبدأ التصنيف والانحدار فهي عبارة عن مخطط لمجموعة أشجار تستخدم لبناء النموذج يعطى تنبؤات من خلال أوراقها الناتجة عن مساحات وتفرعات مختاره عشوائياً من البيانات بمبدأ مشابه لبديهيات أشجار الانحدار .

الشكل (2-3) : هيكلية الغابة العشوائية كأحد أنواع أشجار الإنحدار^[33]



كل تفرع في الشجرة في الشكل يمثل نقطة قرار تم إتخاذها على أساس جملة شرطية وهكذا تستمر التفرعات لحين الوصول إلى القرارات النهائية المتمثلة بعقد الأوراق حيث أن كل ورقة تعتبر كعقدة منفصلة من قرار منفصل عن باقي الأوراق وإن هذه الأشجار تعطي تطابق أمثل بين المخرجات المتمثلة بالتنبؤات بالمقارنة مع المتغير الأصلي الذي تم إعتبره كمتغير هدف ، أي سيتم تطوير أسلوب التنبؤ والحصول على تنبؤات مثلى بأقل أخطاء للتنبؤ عند إستخدام أسلوب RF كأحد تقنيات أشجار التصنيف والإنحدار مقارنة بالأساليب التقليدية للتنبؤ . توفر نمذجة السلاسل الزمنية بإستخدام الغابات العشوائية قدرة تنبؤية معززة وأكثر دقة مقارنة بنماذج السلاسل الزمنية التقليدية للتنبؤ خصوصاً ببيانات الإرصاء الجوية وبيانات أخرى كثيرة على العموم . الشكل السابق يوضح مبدأ عمل خوارزمية الغابة العشوائية كأحد أشجار التصنيف والإنحدار الذي يستخدم في التنبؤ أما الإطار الأشمل الذي يتميز به أسلوب الغابة العشوائية فهو أكثر تعميماً عن أشجار الإنحدار والتصنيف وذلك لإعتماده على مبدأ تقسيم عينة بيانات الدراسة إلى عدة عينات فرعية Bootstrap Sample Sets وذلك لأخذ جميع الأنماط السلوكية لعينة الدراسة في جميع الفترات المختلفة والحصول على شجرة إنحدار لكل عينة فرعية ومن ثم فإن مجاميع هذه

الأشجار سووية سوف تمثل مايسمى بالغابة العشوائية RF وإن القرار النهائي يكون مستنبطاً من خلال غالبية عقد الأوراق لجميع أشجار الإنحدار.

الشكل (2-4) : الإطار العام لخوارزمية عمل الغابة العشوائية (RF) [33]



هنالك إحتمال أن تكون الأشجار في الغابة العشوائية مترابطة فيما بينها بحسب الشكل فإنها عائدة إلى نفس نوع البيانات وكذلك تم إعتداد مبدأ التعبئة Bagging Principle الذي أساسه هو عملية المعاينة التمهيدية Bootstrap Sampling إذ تعمل طريقة التعبئة على تحسين أداء أشجار التصنيف والإنحدار وتجعل RF أكثر حصانة عند تجميعها مع بعضها . يتم معالجة ذلك بجعل الأشجار في الغابة العشوائية غير مترابطة مع بعضها (مختلفة) لذلك فقد قدم (Breiman , 2001) مقترحاً لأن تنمو كل شجرة بشكل منفصل وكذلك بشكل عشوائي وبإجتماع هذين المبدأين ستحدد ملامح وعدد مجموعات العينات الفرعية المشار إليها في الشكل بعد تحويل الأشجار في الغابة العشوائية من مترابطة إلى غير مترابطة (مختلفة) مما سيضمن زيادة ملحوظة في دقة تنبؤ الغابة العشوائية .

2-3-17: يتم بناء خوارزمية الغابة العشوائية باستخدام الخطوات الثلاثة أدناه :

1. من بيانات التدريب يتم إستخراج B من العينات التمهيدية والتي هي في الأصل مترابطة فيما بينها إذ أن B تمثل حجم الغابة أو عدد الأشجار المتعددة المشار إليها في الشكل (2-4) .

2. لكل مجموعة من مجموعات البيانات B فإن نمو الشجرة T_b سيتم بإتباع

خطوات متسلسلة في كل عقده من عقد الشجرة لحين الوصول إلى $n_{\min}$

والتي تمثل الحد الأدنى من أوراق الأشجار أو عدد العقد وكما يلي :

أ- إختيار m والتي تمثل العدد المختار عشوائياً من التنبؤات في كل قسم

من العدد الكلي للمتغيرات p .

ب- إختيار أفضل التنبؤات من التنبؤات المختارة في (أ) وقد تم الإشارة

إليها بالرمز m مع إختيار القسم العائده إليه بهدف تقليل قيمة Mse

للتنبؤات المختارة في (أ) .

ت- فصل العقدة إلى عقدتين فرعيتين تبعاً للمعيار المستخدم أو القيم

التنبؤية الأفضل التي تم إختيارها في (ب) .

3. إستخلاص المخرجات من جميع الأشجار من خلال إيجاد المجموعة $\{T_b\}_1^B$

وأخيراً فإنه عند نقطة معينة X فإن التنبؤ ممكن أن يكون حسب المعادلة التالية

[33].

$$RF = \frac{1}{B} \sum_{b=1}^B T_b(x) \rightarrow (2-14)$$

2-3-18 : كيف تعمل خوارزمية الغابات العشوائية :

في الغابة العشوائية تكون قواعد " الزراعة " و " النمو " لكل شجرة قرار كما

يلي :

لنفترض أننا قمنا بتعيين عدد العينات في مجموعة التدريب على N ، ثم

حصلنا على عينات N هذه من خلال تكرار عدة عينات مع إعادة التعيين ، وسوف

يتم إستخدام نتائج أخذ العينات هذه كمجموعة تدريب لدينا لإنشاء شجرة قرار .

في حالة وجود متغيرات M للإدخال ، سنختار كل عقده متغيرات m

($m < M$) بشكل عشوائي ، ثم نستخدم متغيرات m هذه لتحديد أفضل نقطة تقسيم .

أثناء إنشاء شجرة القرار ، تظل قيمة m دون تغيير سوف تنمو كل شجرة قرار بأكبر

قدر ممكن دون التقليل . توقع البيانات الجديدة من خلال جمع كل أشجار القرار (تصويت الأغلبية للتصنيف ومتوسط الإنحدار) .

خوارزمية الغابة العشوائية هي مجموعة من أشجار القرار ، فشجرة قرار واحدة هي جزء من خوارزمية الغابة العشوائية . مايميز الغابة العشوائية هو أن كل شجرة قرار فيها تعتمد على عينة من البيانات وعلى طريقة في التصنيف مختلفة عن الشجرة الأخرى وبعد أن يتم إعطاء قرار من كل شجرة في الغابة يتم التصويت على أفضل شجرة قرار وتعتمد نتيجة تلك الشجرة [34].

2-3-19 : مميزات خوارزمية الغابة العشوائية [35]:

- تقلل من فرط التخصيص في أشجار القرار و تساعد على تحسين الدقة والكفاءة .
- الخوارزمية تعتبر مرنة لكل من مسائل التصنيف والتنبؤ .
- تعمل بشكل جيد مع كل من القيم الفئوية والقيم المستمرة .
- تقوم بأتممة القيم المفقودة الموجودة في البيانات .
- ضبط البيانات غير مطلوب في هذه الخوارزمية لأنها تستخدم نهجاً قائماً على القواعد .

2-3-20 : عيوب خوارزمية الغابة العشوائية [35]:

- الإفتقار إلى القابلية للتفسير : تعتبر الغابات العشوائية بشكل عام نماذج الصندوق الأسود ، مما يعني أنه قد يكون من الصعب فهم كيفية قيام هذه الخوارزمية بتنبؤاتها . قد يكون هذا عيباً إذا كنت بحاجة إلى شرح الأسباب الكامنة وراء توقع معين لأصحاب المصلحة .
- التعقيد الحسابي : يمكن أن تكون الغابات العشوائية مكثفة من الناحية الحسابية ، خاصة إذا كان لدينا عدد كبير من الميزات أو الأشجار في الغابة الخاصة بنا . هذا يمكن يجعلهم أبطأ في التدريب ويصعب توسيع نطاقهم لمجموعات البيانات الكبيرة جداً .

- الإفراط في التخصيص : كان من ضمن الإيجابيات هو تقليل فرط التخصيص لأشجار القرار لكن يمكن أن تكون الغابات العشوائية عرضة للإفراط في التخصيص ، خاصة إذا كنت تستخدم عدداً كبيراً من الأشجار أو إذا كان لديها ميزات عالية الأبعاد . يمكن أن يؤدي هذا إلى نماذج تعمل بشكل جيد على بيانات التدريب ولكن بشكل سيئ على بيانات الإختبار .
- التحيز إتجاه الميزات ذات المستويات المتعددة : تميل الغابات العشوائية إلى تفضيل الميزات ذات المستويات أو الفئات المتعددة ، مما قد يؤدي إلى زيادة التخصيص على هذه الميزات وعدم ملائمة الآخرين .
- صعوبة التعامل مع البيانات غير المتوازنة : يمكن أن تعاني الغابات العشوائية مع مجموعات البيانات غير المتوازنة ، حيث تكون إحدى الفئات أكثر شيوعاً من الفئات الأخرى . يمكن أن يؤدي هذا إلى نماذج ذات دقة عالية في فئة الأغلبية ولكن دقة ضعيفة في فئات الأقليات [35] .

خوارزمية المتجه الداعم

2-3-2 : تمهيد :

ينقسم أسلوب أو تقنية آلة المتجه الداعم إلى قسمين الأول آلة المتجه الداعم للتصنيف Support Vector Machine for Classification ، والقسم الثاني هو آلة المتجه الداعم للانحدار Support Vector Machine for Regression .

2-5: آلة المتجه الداعم Support Vector Machine :

ظهرت تقنية آلة المتجه الداعم في أوائل التسعينات في عام 1993م ، حيث قدمت من قبل الباحثين (Guyon & Boser) وهي عبارة عن مجموعة أساليب مرتبطة بطرق ومناهج تعلم الآلة وتستخدم في تصنيف البيانات ، وتنتمي إلى مجموعة من المصنفات الخطية للبيانات والملفات الإحصائية ، وهي أداة للتنبؤ وتستخدم نظرية التعلم كأداة لتحقيق أقصى قدر من الدقة التنبؤية مع تجنب الإفراط تلقائياً في معالجة البيانات ، حيث يمكن تعريفها بأنها أنظمة تستخدم مساحة فرضية مبنية على الوظائف الخطية في مساحة متعددة الأبعاد وفق الخوارزمية التي تنفذ التمييز بين البيانات المستمدة من نظرية التعلم الإحصائية^[36].

كما تعد تقنية آلة المتجه الداعم من أشهر طرق التصنيف المبني على أنظمة الحاسب الآلي ، والتي تعتمد على إيجاد منحنى أو مستوى معين يفصل العينات المدخلة عن بعضها البعض ، وتتميز بقدرتها على تصنيف المسائل ذات الفئات الثنائية ، حيث تقوم هذه الخوارزمية بحساب سطح أو مجموعة أسطح في بعد آخر يختلف طوله عن طول متجه الخصائص^[37] ، كما تتحدد دقة الخوارزم بقدرته على الفصل بين الخصائص بحيث تكون أقرب عينة من كلا الخصائص أبعد ما يكون عن بعضها البعض وهو ما يسمى بالهامش ، وبصفه عامه كلما زاد هامش الفصل كلما قل الخطأ في حالة التعميم لجزء البيانات الغير خاص بمرحلة التكرار والتعلم ، وبالرغم مما تبدو عليه المشكلة من سهولة ، إلا أنه في أغلب الأحيان لا يمكن الفصل بين الخصائص خطياً ، وحينها نلجأ لتحويل محاور متجهات الخصائص لبعد أعلى بحيث

يتم الفصل بينهم بسطح ، ويراعى في هذا المنظور العبء الحسابي فيتم حساب الضرب القياسي للمتجهات بواسطة دالة المصفوفة^[38] .

2-5-1 : مفهوم تقنية آلة المتجه الداعم (SVM) :

هي إحدى تقنيات الذكاء الاصطناعي التي يمكن أن تساعد في حل مشاكل تصنيف البيانات الضخمة (Big Data) ، وهي عبارة عن أسلوب يجمع بين النظرية الإحصائية والتعلم الخاضع للإشراف (Supervised Mechine Learning) ، طورت من قبل الباحثة (Vapnik) عام 1998م ، حيث تعتمد فكرة تقنية آلة المتجه الداعم على البحث على أفضل طريقة لتقسيم البيانات إلى مجموعتين بوضع مستوى فائق (Hyper Plane) بينهما بغض النظر عن طبيعة البيانات سواء كانت قابله للفصل الخطي أو لا ، وهنا تكمن مصدر قوتها .

2-5-2 : كيفية عمل تقنية آلة المتجه الداعم (SVM) :

تقنية آلة المتجه الداعم عبارة عن منهج تدريب ثنائي ، تستخدم لتصنيف بيانات التدريب (Training Data) إلى صنفين (موجب ، سالب) مثلا ، حيث يتم تدريبها بالإعتماد على مبدأ تقليل الخطأ إلى الحد الأدنى وذلك بين القيم المتوقعة والقيم الفعلية ، كما أن ذلك سيققل احتمال حدوث الأخطاء عند تطبيق الخوارزمية على النماذج الجديد ، وعند إستخدام تقنية آلة المتجه الداعم في عمليات التصنيف ، تستطيع تلك التقنية إيجاد أفضل سطح فاصل وذلك حتى يتم فصل الأمثلة الموجبه عن الأمثلة السالبة ، حيث يتم إختبار السطح الفاصل لكي تكون المسافة متواجده بين أقرب مثال موجب وسالب ، كما أن هذا السطح الفاصل يكون أكبر مايمكن ، ولهذا تكون عملية التصنيف لبيانات الإختبار صحيحه، وأثبتت تلك التقنيه نجاحها في التغلب على الكثير من المشكلات مثل معالجة الصور (Image Processing) ، تصنيف النصوص (Text Classification) ، وتحليل الأسواق المالية^[39] .

حيث يكون السطح الفاصل معرف بمجموعة من النقاط نتاج ضربهم القياسي مع متجه فى الإحداثيات الجديد (ذات البعد الأعلى) والتي تكون ثابتة ، إن عملية

تصنيف البيانات تعد من أكثر عمليات المحاكاه شيوعاً ، بوجود نقاط البيانات التي تتميز بعدة خصائص ، ويكون الهدف هو تصنيف نقطه جديده وتحديد إلى أي من الخصائص تنتمي ، حيث ينظر لنقطة البيان على أنها متجه له عدد من الخصائص ، وإذا تم الفصل بسطح بعده ينقص عنه بواحد يكون التصنيف خطياً ويكون غير خطياً ماعدا ذلك ، وفي حالة توافر أكثر من فاصل خطي يتم إختيار الفاصل الذي يضمن هامش أوسع بين أقرب نقطتين من الخصائص المختلفه لبعضها ، وهو مايسمى بالمستوى ذو الهامش الأكبر^[40].

تم إستخدام أسلوب تعلم الآلة المستند إلى تقنية آلة المتجه الداعم في الغالب للتعرف على الأنماط في العديد من المجالات ، حيث أثبتت هذه التقنية كفاءتها من أي تقنية أخرى للتعرف على الأنماط الأكثر إستخداماً ، مثل الشبكات العصبية الاصطناعية (ANN) وتصنيف (Basiean Classification) والأداء الذي تقدمه تقنية آلة المتجه الداعم أعلى نسبياً إذا كان يحتوي على مجموعة بيانات كبيرة (Large dataset) لتعميم المشكلة ، وتتمثل القوة الرئيسية لتقنية آلة المتجه الداعم في أن تدريب البيانات سهل نسبياً ، حيث تقيس هذه التقنية بشكل جيد نسبياً البيانات عالية الأبعاد (High dimensional data) ، ويمكن التحكم صراحة في المفاضلة بين تعقيد التصنيف والخطأ ، وهناك الكثير من النطاق المستقبلي لهذه التقنية في كافة المجالات لتحسين دقة التنبؤ^[41].

3-5-2 : إستخدامات تقنية آلة المتجه الداعم (SVM) :

تستخدم هذه التقنية في تطبيقات تصنيف وتجزئة الصور (Image Classification and Segmentation) ، وتصنيف النصوص (Assignment Category) ، وتصنيف البريد الإلكتروني فيما إذا كان مزيف أو حقيقي (Ham Email) ، كذلك تحليل الآراء (Sentiment Analysis) وغيرها ، كما تستخدم في التطبيقات البيولوجية وغيرها من العلوم^[42].

تعد تقنية آلة المتجه الداعم (SVM) من أهم التقنيات المستخدمة في تصنيف البيانات ، إذ تعتمد على عوامل متعددة ومتغيرات تؤثر على نحو مباشر أو غير مباشر على إيجاد الحل النهائي ، فمثلاً نموذج آلة المتجه الداعم (SVM) يعتمد على عدد من المعلمات الأساسية مثل المستوى الفاصل (Hyperplane) ، ومضاريب الكرانج (Multipliers Lagrange) ، التي تؤثر على نحو كبير على دقة عملية التصنيف ، إذ أن البيانات الأساسية في فضاء الإدخال (Space Input) تصنف على وفق النموذج الرياضي الآتي^[43] :

$$\begin{aligned} W^T x_i + b \geq +1 & \rightarrow (2-15) & \text{for } d_i = +1 & , i=1,2,\dots,N \\ W^T x_i + b \leq -1 & \rightarrow (2-16) & \text{for } d_i = -1 & , i=1,2,\dots,N \end{aligned}$$

حيث أن :

W : متجه الأوزان Wight .

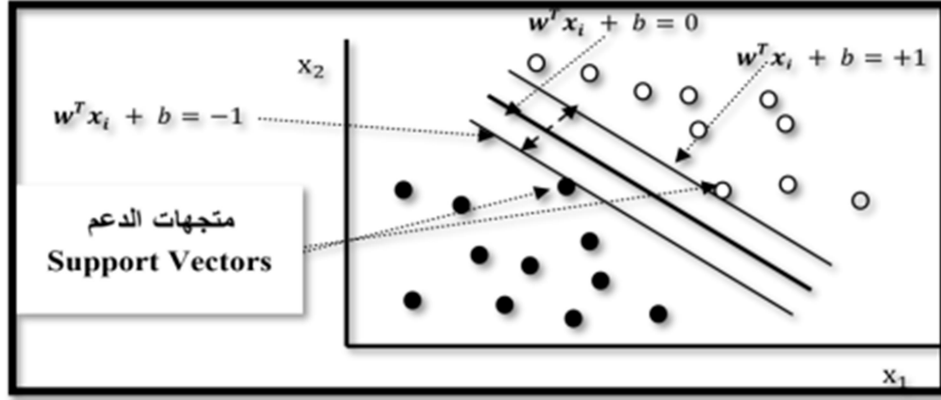
x : متجه الإدخال Input Vector .

b : قيمة التحيز Bias .

d : قيمة الإخراج .

كما يمكن ملاحظة معادلات الحدود في المستوى من خلال الشكل الآتي :

الشكل (5-2) : متجهات الدعم الواقعة على الحد الفاصل بين البيانات في الفضاء^[44]



من الشكل اعلاه نلاحظ أن معادلة المستوى الفاصل (plane Hyper) تكتب بالشكل الآتي :

$$W^T x_i + b = 0 \rightarrow (2-17)$$

إن البيانات القريبة أو التي تقع على حدود الحد الفاصل تسمى متجهات الدعم أو المساندة Support Vector .

كما يمكن حساب المسافة بين النقاط في المستوى الفاصل من خلال العلاقة الآتية :

$$d(W, b, x_i) = \frac{|W^T x_i + b|}{\|W\|} \rightarrow (2-18)$$

وبعد عدد من الإجراءات والتحويلات الرياضية ، يتم إيجاد قيم كل من متجه الأوزان المثالي (W^*) والتحيز المثالي (b^*) ومن بعدها تحسب دالة التصنيف الآتية:

$$f(x) = I(W^* \bullet x + b^*) \rightarrow (2-19)$$

إذ أن :

(W^*) : تمثل الوزن المثالي .

(b^*) : تمثل قيمة التحيز المثاليه .

SGN : تمثل القرار النهائي لإنتماء (x) لأحد الأصناف .

إن عملية التصنيف في أغلب التطبيقات تعتمد أسلوب الفصل غير الخطي (Nonlinear) لأن البيانات التي صُنفت تكون عموماً غير خطية ، لذلك يعتمد على هذه الآلية في معظم تطبيقات تقنية آلية المتجه الداعم التي تأخذ إجراءات رياضية في تكوين دالة الفصل النهائية ، ومن أهم هذه الإجراءات دالة النواة (Kernel trick) التي تقوم بتحويل البيانات من فضاء منخفض البعد (Low dimensional) إلى فضاء عالٍ البعد (Hight dimensional) من أجل تصنيفها^[44] .

4-5-2 : دالة النواة Kernel function :

ويكون تطبيق دالة النواة في حالة البيانات غير الخطية ، ويرمز لها بالرمز $K(X_i, X_j)$ وتعرف بالنواة الخطية (Linear Kernel) وهي عبارة عن مجموعة دوال النواة تكون متألّفة من متغيرات ، لأنها تكون معتمدة على حساب المضروبات الداخلية (dot product) من متجهين إثنين ، وذلك يعني أن الدوال يمكن إعادة صياغتها إلى فضاء عالي الأبعاد عن طريق مخطط دالة الخواص (Feature Mapping Function) التي تكون غير خطية^[45] .

$$X \rightarrow \phi(x)$$

حيث أن $\phi(x)$: دالة تحويل مدخلات البيانات إلى فضاء خواص جديد ذات أبعاد أعلى ودالة الهدف في هذه الحالة تكون :

$$Y = f(x) = w' * \phi(x) + b \rightarrow (2-20)$$

5-5-2 : أهم أنواع دالة النواة :

ومن أهم أنواع دوال النواة (Kernel trick) التي تستخدم في مسائل التصنيف تعرض كما في الجدول التالي :

جدول (2-2) : أنواع دوال النواة وصيغها الرياضية^[46]

نوع الدالة	الصيغة الرياضية
Gaussian (RBF) kernel	$K(x_i, x_j) = \exp \left\langle -\frac{\ x_i - x_j\ ^2}{2\sigma^2} \right\rangle$
Polynomial kernel	$K(x_i, x_j) = (1 + x_i' x_j)^d$
Linear kernel	$K(x_i, x_j) = x_i' x_j$
Monomial	$K(x_i, x_j) = [x_i' y]^m$

2-5-6: مميزات خوارزمية المتجه الداعم^[16]:

- إنها فعالة للغاية حتى مع البيانات الكبيرة .
- يمكن إدارة البيانات غير الخطية بشكل فعال باستخدام تكتيكات النواة .
- يمكن استخدامها لحل مشاكل التصنيف والنحدار .

2-5-7: عيوب خوارزمية المتجه الداعم :

- علينا أن نختار نواه مثالية لـ SVM ، وهو أمر صعب .
- في مجموعة البيانات الكبيرة ، يستغرق التدريب وقتاً أطول نسبياً .
- آلة المتجه الداعمه ليست نموذجاً احتمالياً ، لذلك لا يمكننا شرح التصنيف من حيث الإحتماليه .
- نموذج SVM (مقارنة بشجرة القرار) يصعب فهمه وتفسيره^[16] .

الفصل الثالث الجانب التطبيقي

الفصل الثالث الجانب التطبيقي

1-3 : تمهيد :

يعرف التحليل على أنه العملية المنهجية لتقسيم موضوع أو مادة معقدة إلى عناصرها الأولية ، بهدف تفسيرها وفهمها بشكل أعمق .
وتبرز أهمية تحليل البيانات كمرحلة محورية في هذا النوع من الدراسات ، حيث تهدف هذه الوحدة إلى تطبيق إطار منهجي متكامل . يشمل ذلك تحديد آلية البحث ، وإختيار الخوارزميات الملائمة ، وإستعراض أدوات التحليل المستخدمة ، ووصف مجموعة البيانات قيد الدراسة . كما سيتم بناء النموذج التحليلي المناسب وتقييم أدائه من خلال قياس دقة التصنيف .

2-3 : لغة بايثون :

بايثون لغة سهلة القراءة للغاية ومتنوعة ومتعددة الإستخدامات ، وإسمها مستوحى من مجموعة كوميدية بريطانية بإسم (Monty Python) وكان أحد الأهداف الأساسية لفريق تطوير بايثون هو جعل اللغة مرحة وسهلة الإستخدام ، وإعدادها بسيط ، وطريقة كتابتها مباشرة وتعطيك تقريراً مباشراً عند حدوث أخطاء، وهي خيار ممتاز للمبتدئين والوافدين الجدد على البرمجة . لغة بايثون هي لغة متعددة الاستعمالات ، وتدعم مختلف أنماط البرمجة مثل كتابة السكريبتات والبرمجة كائنية التوجه (Object – Oriented) وهي مناسبة للأغراض العامة ، وإستعمالها يتزايد في سوق العمل إذ تعتمد عليها منظمات مثل (United Space Alliance) " شركة في مجال إرسال مركبات فضائية وتتعاقد معها ناسا " و (Industrial Light & Magic " أستديو للتأثيرات السينمائية وللرسوم المتحركة " ، وتوفر بايثون قدرات كثيرة لمن يريد تعلم لغة برمجة جديدة .

طورت اللغة في نهاية الثمانينات من القرن الماضي ، ونشرت أول مره في عام 1991م ، طورت بايثون من قبل Guido van Rossum ، وهو عضو نشط

للغاية في المجتمع . وتعد بايثون على أنها بديل عن لغة ABC ، وأول إصدار منها كان يتضمن التعامل مع الإستثناءات (Exception Handling) والدوال والأصناف (Classes) مع إمكانية الوراثة فيها . وعندما أنشئ منتدى محادثة في Usenet بإسم comp.lang.python في 1994م ، فبدأت قاعدة مستخدمي بايثون بالنمو ، مما مهد الطريق لها لتصبح واحدة من أكثر لغات البرمجة شيوعاً وخصوصاً لتطوير البرمجيات مفتوحة المصدر .

3-2-1 : تاريخ بايثون :

ظهرت بايثون في أواخر الثمانينات على يد غيدو فان روسوم (Guido van Rossum) ، وقد عدت خليفة للغة ABC . كما إستفادت بايثون من الكثير من اللغات السابقة لها ، مثل Modula-3 و C و C++ و Algol-68 و SmallTalk ، وغيرها من اللغات .

نشر الإصدار 2.0 من لغة بايثون عام 2000 ، وقد قدم العديد من الميزات الجديدة ، مثل القوائم الفهمية (List Comprehensions) ، ونظام كنس المهملات (Garbage Collection) . وظهر في عام 2008 الإصدار بايثون 3.0 ، والذي شكل طفره في اللغة ، بين أنه لم يكن متوافقاً تماماً مع الإصدارات السابقة ، لذلك قرر فريق التطوير الإستمرار في دعم إصدار أخير من سلسلة بايثون 2.x ، وهو بايثون 2.7 حتى عام 2020 .

3-2-2 : مميزات لغة بايثون :

تتميز بايثون بعدة أمور عن غيرها من لغات البرمجة ، منها :

- سهولة التعلم : يسهل تعلم لغة بايثون ، إذ تتألف من عدد قليل من الكلمات المفتاحية ، وتتميز بصياغة بسيطة وواضحة .
- المقروئية : شيفرة لغة بايثون واضحة ومنظمة وسهلة القراءة .
- سهولة الصيانة : شيفرة بايثون سهلة الصيانة إلى حد بعيد .

- مكتبة قياسية واسعة : تحتوي مكتبة بايثون القياسية على عدد كبير من الحزم المحمولة التي تتوافق مع أنظمة يونكس وويندوز وماك .
- الوضع التفاعلي : تدعم بايثون الوضع التفاعلي ، مما يتيح إمكانية تنفيذ الشيفرات مباشرة على سطر الأوامر وتلقيها .
- متعددة المنصات : يمكن تشغيل لغة بايثون على طيف واسع من المنصات والأجهزة ، مع الاحتفاظ بنفس الواجهه على جميع تلك المنصات .
- التوسعية : من أهم مميزات بايثون ، هو توفرها على عدد هائل من الوحدات ، التي يمكنها توسيع قدرات اللغة في كل مجالات التطوير ، مثل تحليل البيانات والرسومات ثنائية وثلاثية الأبعاد ، وتطوير الألعاب ، والأنظمة المدمجة ، والبحث العلمي ، وتطوير المواقع وغيرها من المجالات
- قواعد البيانات : توفر بايثون واجهات لجميع قواعد البيانات الأساسية .
- الرسوميات : تدعم بايثون التطبيقات الرسومية .
- دعم البرامج الكبيره : بايثون مناسبة للبرامج الكبيره والمعقده .

3-2-3 : استخدامات لغة بايثون :

تستخدم لغة بايثون في كل المجالات ، فهي لغة برمجة متعددة الأغراض ، ومن مجالات إستخدامها : تحليل البيانات ، والروبوتات ، وتعلم الآلة ، وتطبيقات REST ، وتطوير المواقع والألعاب ، والرسوم ثلاثية الأبعاد ، والأتمتة وبرمجة الأنظمة المدمجة ، والكثير من المجالات الأخرى التي لايسعنا حصرها هنا .

تستخدم الكثير من المواقع والشركات العملاقه لغة بايثون ، ومنها Spotify و Google و Amazon إضافة إلى Facebook التي تستخدم بايثون لمعالجة الصور . وفي كل يوم تتحول شركات جديده إلى إستخدام بايثون ، مثل Instagram التي قررت مؤخراً إستخدامها وفضلتها على PHP . تستخدم بايثون أيضاً من قبل بعض الجهات العلمية والبحثية ، مثل وكالة الفضاء الأمريكية ناسا ، والتي لها مستودع خاص بالمشاريع المطوره ببايثون .

4-2-3 : لماذا لغة بايثون وليس غيرها :

تحديد أفضل لغة برمجة للتعلم قد يكون مهمة صعبة . لو سألت عدة أشخاص عن لغة البرمجة التي يجب تعلمها ، فستحصل على عدة إجابات ، ويكفي أن تدخل على جوجل وتكتب أفضل لغة برمجة ، وستجد آراء مختلفة ، وجدالاً لا ينتهي هذا الموضوع .

ولذلك يمكن أن نقدم بعض الحجج لماذا نرى أن لغة بايثون هي لغة المستقبل ، وأن تعلم لغة بايثون مثالي للمبتدئين الذين يريدون دخول عالم البرمجة وعلوم الحاسوب .

- الشعبية : بحسب إستطلاع موقع Stackoverflow ، بايثون هي لغة البرمجة العامة الأسرع نمواً ، كما تمكنت سنة 2019م من التفوق على جافا بعدها أكثر لغة برمجة متعددة الأغراض إستخداماً ، كما أنها رابع أكثر لغة تكنولوجيا برمجة إستخداماً وراء JavaScript و CSS/HTML و SQL ، كما أنها أكثر لغة برمجة محبوبة من قبل المبرمجين .

بايثون ليست لغة قوية وفعالة فحسب ، كما يدل ذلك حقيقة أنها أكثر لغة برمجة متعددة الأغراض إستخداماً ، بل هي فوق ذلك محبوبة من قبل المبرمجين ، وهذا مؤشر على سهولتها ، ثم فوق كل ذلك جميعاً ، فإن مستقبلها يبدو مشرقاً ، لأنها الأسرع نمواً .

- طلب سوق العمل : يستخدم بايثون بعض أكبر الشركات في مجال التكنولوجيا ، مثل Uber و PayPal و Google و Facebook و Instagram ، و Netflix و Dropbox و Reddit . إضافة إلى هذا ، تستخدم بايثون بكثافة في مجال الذكاء الاصطناعي والتعلم الآلي وتحليل البيانات وأنظمة المراقبة وغير ذلك .

يقدر موقع Stackoverflow ، الدخل السنوي لمطوري بايثون المحترفين بحوالي 63 ألف دولار ، وهو مبلغ كبير ، ويدل على أن هناك طلباً كبيراً على لغة بايثون في سوق العمل .

- الدعم : نظراً لشعبيتها الكبيرة ، تتمتع لغة بايثون بدعم جيد على جميع المستويات تقريباً . وبعدها اللغة المفضلة للمبتدئين ، فهناك قناطر من المراجع والمواد والدورات التعليمية التي تشرح مفاهيم البرمجة الأساسية ، إضافة إلى صياغة اللغة وتطبيقاتها [48].

3-3: متغيرات الدراسة لمرض الكلى التي يتم التطبيق عليها :

تمثل الكلى إحدى الركائز الأساسية لصحة الجسم ، فهي تعمل كمرشح ذكي ينقي الدم من الفضلات والسموم ، وينظم توازن السوائل والأملاح ، ويساهم في ضبط ضغط الدم ، وينتج هرمونات حيوية تدعم إنتاج خلايا الدم الحمراء وصحة العظام . أي خلل في وظائفها قد يؤدي إلى عواقب صحية خطيرة ، تصل إلى الفشل الكلوي الذي يهدد الحياة [47].

في إطار هذا البحث تم جمع بيانات مرضى الكلى بعناية شديدة من مستشفى عطره من مركز الشهيد د/ نوري لعلاج وغسيل الكلى بمدينة عطره وتحتوي هذه البيانات على 6 متغيرات مستقلة تؤثر بالإصابة بالمرض (S.Na ، Hb ، S.K ، S.Phosphorus ، S.cholesterol ، S.creatinine) . بالإضافة للمتغير التابع Y

تتميز هذه البيانات بأنها جمعت من مركز متخصص في أمراض الكلى ، مما يضمن دقة التشخيص وصحة المعلومات ، كما أنها تضمنت 6 متغيرات مستقلة مؤثرة في تشخيص الإصابة مما يتيح تحليلاً متعدد الأبعاد للمشكلة ، كما أنها لم تسجل أي قيم مفقودة في جميع الحقول مما يقلل من احتمالية التحيز ويعزز موثوقية النتائج وهذا يدل على نظافة البيانات ، بالإضافة إلى أن هذه البيانات غطت فترة زمنية محددة (2024م) وموقعاً جغرافياً واضحاً مما يساهم في

تحديد نطاق الدراسة بدقة . لذلك تعتبر بيانات الدراسة ملائمة تماماً للتطبيق الإحصائي والتعلم الآلي ، وتمهد الطريق بإجراء مقارنة دقيقة وموضوعية بين النماذج الإحصائية والتعلم الآلي في تصنيف مرضى الكلى .

تكونت بيانات الدراسة من 260 حالة ، تم ترميزها ثنائياً وفقاً للحالة الصحية للفرد ، حيث أعطيت القيمة (0) للأشخاص غير المصابين (الأصحاء) ، في حين أعطيت القيمة (1) للأشخاص المصابين بأمراض الكلى . وقد تم استخدام هذا الترميز لتسهيل عملية التصنيف في النماذج الإحصائية وخوارزميات التعلم الآلي.

جدول (3-3) : مسميات متغيرات الدراسة

المتغير	المسمى
Hb	الهيموجلوبين
S.Na	الصوديوم
S.K	البوتاسيوم
S.PhosPhorus	الفسفور
S.cholesterol	الكولسترول
S.creatinine	الكرياتين في الدم
Y	المتغير التابع (المتغير الهدف)

4-3 : نتائج تحليل البيانات باستخدام نموذج الإنحدار اللوجستي باستخدام برنامج بايثون :

جدول (4-3) : الإحصاءات الوصفية لمتغيرات الدراسة

	Hb	S.creatinine	S.sodium	S.potassium	S.phosphorus	S.cholesterol	Y
count	260.000000	260.000000	260.000000	260.000000	260.000000	260.000000	260.000000
mean	10.668077	1.963077	136.465385	4.653846	4.663462	150.238462	0.500000
std	2.437754	0.759815	5.112478	0.785432	1.612345	71.158891	0.500964
min	4.600000	0.600000	120.000000	3.300000	1.800000	54.000000	0.000000
25%	8.900000	2.000000	135.000000	4.200000	3.300000	115.750000	0.000000
50%	11.400000	2.000000	137.000000	4.600000	4.600000	166.000000	0.500000
75%	12.700000	2.200000	138.000000	5.100000	5.900000	171.000000	1.000000
max	14.800000	5.000000	193.000000	8.100000	10.600000	1112.000000	1.000000

المصدر : من إعداد الباحثه من تحليل البيانات باستخدام برنامج بايثون إصدار 3.0

التفسير :

من جدول (4-3) تظهر الإحصاءات الوصفية لعينة مكونة من 260 مريضاً تبايناً واضحاً في مؤشرات وظائف الكلى حيث يبلغ متوسط الكرياتينين 1.96 mg/dl والفسفور 4.66 mg/dl مما يعكس إعتلالاً كلياً مزمنياً ، إلى جانب انخفاض متوسط الهيموجلوبين إلى 10.67 mg/dl المؤشر على فقر الدم المصاحب ، كما تكشف القيم المتطرفة خاصة في متغير الكوليسترول عن حالات شاذة ، في حين يتسم المتغير التابع Y بتوزيع متوازن تماماً (وسط 0.5) يسهل عملية النمذجة التنبؤية .

• أداء النموذج :

مصفوفة الارتباك confusion Matrix :

$$\begin{bmatrix} 38 & 1 \\ 4 & 35 \end{bmatrix}$$

التفسير :

تعتبر 38 تصنيفات صحيحة لفئة الغير مصابين. وتعتبر 35 تصنيفات صحيحة لفئة المصابين .

أما 4 تعتبر تصنيفات خاطئة لفئة الغير مصابين (تم تصنيفها مصابين) . و 1 تعتبر تصنيفات خاطئة لفئة المصابين (تم تصنيفها غير مصاب) .

• مقاييس أداء نموذج الإنحدار اللوجستي في التصنيف الثنائي:

جدول (3-5) : مقاييس أداء نموذج الإنحدار اللوجستي في التصنيف الثنائي

	Precision	Recall	f1-score	Support
غير مصابين	0.90	0.97	0.94	39
مصابين	0.97	0.90	0.93	39
Micro avg	0.94	0.94	0.94	78
macro avg	0.94	0.94	0.94	78
weighted avg	0.94	0.94	0.94	78

المصدر: من إعداد الباحثه من تحليل البيانات بإستخدام برنامج بايثون إصدار 3.0

التفسير من جدول (3-5) :

- الدقة Precision :

من خلال النتائج الواردة في الجدول (3-5) ، تبين أن أداء النموذج في التصنيف كان قوياً جداً وكفوياً بشكل عام ، فبالنسبة لفئة الغير مصابين بلغت دقة النموذج 0.90 مما يعني أن 90% من الحالات التي تنبأ النموذج بها بأنها تنتمي إلى هذه الفئة كانت توقعات صحيحة .

أما فيما يخص فئة المصابين فقد سجل النموذج دقة متميزة بلغت 0.97 مما يعكس أنه عندما يتنبأ النموذج بعينة على أنها من فئة المصابين ، فإن هذه التوقعات تكون صحيحة بنسبة 97% .

- الإستدعاء Recall :

يظهر مقياس الإستدعاء قدرة النموذج على إكتشاف الحالات الفعلية لكل فئة بشكل صحيح .

فبالنسبة لفئة الغير مصابين بلغت قيمة الإستدعاء 0.97 مما يعني أن النموذج نجح في التعرف على 97% من جميع العينات التي تنتمي حقاً إلى هذه الفئة وتصنيفها بشكل صحيح مما يشير إلى أداء ممتاز في تغطية هذه الفئة .

أما بالنسبة لفئة المصابين فقد سجلت قيمة الإستدعاء 0.90 وهذا يعني أن النموذج إستطاع أن يحدد ويصنف بشكل صحيح 90% من العينات الفعلية التي تنتمي إلى الفئة 1 مما يظهر أداء قوياً وجيداً في إستدعاء حالات هذه الفئة أيضاً ، وإن كان أقل بقليل من أدائه مع فئة الغير مصابين .

• الدقة الكلية Accuracy :

Accuracy_score	0.9358974358974359
----------------	--------------------

التفسير:

نسبة التصنيفات الصحيحة بشكل عام تساوي 0.9359 (93.59%) وهذا يدل على أن نموذج الإنحدار اللوجستي ذو دقة عالية ونموذج جيد في التصنيف .

1-4-3 : نتائج إختبار Hosmer-Lemeshow باستخدام برنامج بايثون :

Hosmer-Lemeshow: 1.6970

p-value: 0.9889

dof: 8

التفسير :

قيمة إحصائية الإختبار تساوي (1.6970) وهذه القيمة تمثل مدى الإختلاف بين القيم المتوقعة والقيم الملاحظة وكلما كانت أقرب من الصفر كان التوافق أفضل .
قيمة p_value تساوي (0.9889) وهي مرتفعة جداً وأكبر من 0.05 وهذا يدل على أن النموذج يصف البيانات بشكل جيد .
درجات الحرية $dof = 8$.

2-4-3 : نتائج إختبار Wald باستخدام برنامج بايثون :

جدول(3-6) : معاملات الانحدار المقدره ، إحصاء Wald ، القيمة الإحتمالية لإحصاء

Wald

	result.params	wald_stats	p_values
Const	0.921907	0.002315	0.961624
Hb	-1.898964	7.052501	0.007915
S.creatinine	18.651297	7.565800	0.005949
S.Na	-0.108049	1.912411	0.166696
S.K	0.628703	0.186844	0.665556
S.PhosPhorus	1.902692	5.586596	0.018098
S.cholesterol	-0.079954	1.494966	0.221447

المصدر: من إعداد الباحثه من تحليل البيانات باستخدام برنامج بايثون إصدار 3.0

التفسير من جدول (3-6) :

من الجدول نلاحظ أن قيمة الثابت تساوي 0.92190 ، وبلغت قيمة إحصاء والد 0.002315 ، كما تجدر الإشارة إلى أن قيمة الإحتمال p_value المرتبطة بهذا المعامل تساوي 0.961624 وهي أكبر من مستوى الدلالة 0.05 وهذا يعني أن هذا المعامل لا يحقق الدلالة الإحصائية عند مستوى ثقة 95% . وحقق متغير الهيموجلوبين معاملاً -1.89894 أي له تأثيراً سلبياً على إحتمالية النتيجة $y=1$ ، وبلغت قيمة إحصاء والد 7.052501 ، والقيمة الإحتمالية 0.007915 وهي أقل من مستوى الدلالة 0.05 وهذا يعني أن المتغير دال إحصائياً ، مما يعني أن زيادة الهيموجلوبين تقلل من إحتمالية النتيجة $y=1$. كما حقق متغير الكرياتينين معامل قيمته تساوي 18.6515 أي له تأثير إيجابي كبير هذا يعني أن زيادة خطر الإصابة بنسبة 97% لكل وحدة زيادة في الكرياتينين ، وقد بلغت إحصاء والد 7.565800 ،

وكانت القيمة الإحتمالية تساوي 0.005949 وهي أقل من مستوى الدلالة 0.05 وهذا يعني أن المتغير دال إحصائياً ، مما يعني أن زيادة الكرياتينين تزيد إحتمالية النتيجة $y=1$ بشكل كبير . أما متغير الصوديوم كان معامل قيمته -0.108049 مما يشير إلى تأثير سلبي طفيف نظرياً على إحتمالية النتيجة $y=1$ ، وقيمة إحصاء والد تساوي 1.912411 ، وكانت قيمة الإحتمال المرتبطة بهذا المتغير تساوي 0.166696 وهي أعلى من 0.05 مما يدل على أن المتغير غير دال إحصائياً . وحقق متغير البوتاسيوم معامل قيمته 0.628703 مما يشير إلى تأثير إيجابي طفيف نظرياً على إحتمالية النتيجة $y=1$ ، وقيمة إحصاء والد تساوي 0.186844 ، وكانت قيمة الإحتمال المرتبطة بهذا المتغير تساوي 0.66555 وهي أعلى من 0.05 مما يدل على أن المتغير غير دال إحصائياً . وحقق متغير الفسفور معامل 1.902692 أي له تأثيراً إيجابياً على إحتمالية النتيجة $y=1$ ، وبلغت قيمة إحصاء والد 5.586596 ، والقيمة الإحتمالية 0.018098 وهي أقل من مستوى الدلالة 0.05 وهذا يعني أن المتغير دال إحصائياً ، مما يعني أن زيادة الفسفور تزيد من إحتمالية النتيجة $y=1$. أما متغير الكوليسترول معامل قيمته -0.079954 مما يشير إلى تأثير سلبي طفيف نظرياً على إحتمالية النتيجة $y=1$ ، وقيمة إحصاء والد تساوي 1.494966 ، وكانت قيمة الإحتمال المرتبطة بهذا المتغير تساوي 0.221447 وهي أعلى من 0.05 مما يدل على أن المتغير غير دال إحصائياً

- المتغيرات ذات الدلالة الإحصائية ($P < 0.05$) :

• الهيموجلوبين Hb

معامل سلبي -1.899

كل زيادة وحدة في الهيموجلوبين تقلل إحتمالية $y=1$

تأثير قوي ودال إحصائياً ($p=0.0079$)

• الكرياتينين S.creatinine

معامل موجب كبير 18.65

تأثيره الأقوى في النموذج

كل زيادة وحدة في الكرياتينين تزيد احتمالية $y=1$ بشكل كبير ($p=0.0059$)

• الفسفور S.phosPhorus

معامل موجب 1.90

تأثير إيجابي ودال إحصائياً ($p=0.0181$)

يزيد احتمالية $y=1$

- المتغيرات غير الدالة إحصائياً ($p>0.05$)

الكوليسترول S.cholesterol و البوتاسيوم S.K و الصوديوم S.Na

الثابت Const أيضاً غير دال إحصائياً

جدول (3-7) : ملخص تقديرات نموذج الإنحدار اللوجستي لمتغيرات الدراسة مع

فترات الثقة 95%

	Coef	std err	Z	P> z	[0.025	0.975]
Const	0.9219	19.160	0.048	0.962	-36.631	38.475
Hb	-1.8990	0.715	-2.656	0.008	-3.300	-0.497
S.creatinine	18.6513	6.781	2.751	0.006	5.361	31.941
S.Na	-0.1080	0.078	-1.383	0.167	-0.261	0.045
S.K	0.6287	1.454	0.432	0.666	-2.222	3.479
S.PhosPhorus	1.9027	0.805	2.364	0.018	0.325	3.480
S.cholesterol	-0.0800	0.065	-1.223	0.221	-0.208	0.048

المصدر: من إعداد الباحثه من تحليل البيانات باستخدام برنامج بايثون إصدار 3.0

التفسير من جدول (7-3) :

الثابت غير دال إحصائياً أي ليس له تأثير حيث بلغت قيمته الإحتمالية 0.962 . أما متغير الهيموجلوبين هو متغير مؤثر وذي دلالة إحصائية عالية في النموذج حيث بلغ معامل السلبى قيمة -1.8990 مع قيمة دلالة إحصائية (p_value) قدرها 0.008 ، فتره الثقة [-3.300 , -0.497] ، أما الكرياتينين هو متغير قوي وذي تأثير إيجابي كبير ، حيث بلغ معامل الموجب 18.651 وهو دال إحصائياً بقيمة معنوية عالية $p=0.006$ فترة الثقة [5.361 , 31.941] موجب بالكامل وتأثيره قوي وإيجابي ، ومتغير الصوديوم غير دال إحصائياً بقيمة معنوية $p=0.167$ ، ومتغير البوتاسيوم غير دال إحصائياً بقيمة معنوية $p=0.666$ ، أما متغير الفسفور هو متغير ذو دلالة إحصائية وتأثير إيجابي حيث بلغ معامل الموجب 1.9027 مع قيمة دلالة إحصائية $p=0.018$ ، فترة الثقة [0.325 , 3.480] موجبة بالكامل وتأثيره إيجابي ، أما متغير الكوليسترول غير دال إحصائياً بقيمة معنوية $p=0.221$.

- تفسير المتغيرات المؤثرة (ذات دلالة إحصائية) :

جدول (8-3) : تفسير المتغيرات ذات الدلالة الإحصائية

المتغير	المعامل	القيمة p	التفسير
Hb	-1.8990	0.008	كل زيادة وحدة في الهيموجلبين تقلل إحتمالية $y=1$ بنسبة كبيرة
S.creatinine	18.6513	0.006	التأثير الأقوى في النموذج
S.phosPhorus	1.9027	0.018	كل زيادة وحدة في الفسفور تزيد إحتمالية $y=1$

المصدر: من إعداد الباحثه من تحليل البيانات بإستخدام برنامج بايثون إصدار 3.0

- تفسير المتغيرات غير المؤثرة (بدون دلالة إحصائية) :
جدول (3-9) : المتغيرات الغير دالة إحصائياً

المتغير	المعامل	القيمة p	التفسير
Const	0.9219	0.962	الثابت غير ذو دلالة إحصائية
S.Na	-0.1080	0.167	الصوديوم لا يؤثر بشكل معنوي
S.K	0.6287	0.666	البوتاسيوم لا يؤثر بشكل معنوي
S.cholesterol	-0.0800	0.221	الكوليسترول لا يؤثر بشكل معنوي

المصدر: من إعداد الباحثه من تحليل البيانات باستخدام برنامج بايثون إصدار 3.0

5-3 : نتائج تحليل البيانات باستخدام خوارزمية الغابة العشوائية (Random Forest) باستخدام برنامج بايثون :

جدول(3-10) : الإحصاءات الوصفية لمتغيرات الدراسة

	Hb	S.creatinine	S.sodium	S.potassium	S.phosphorus	S.cholesterol	Y
count	260.000000	260.000000	260.000000	260.000000	260.000000	260.000000	260.000000
mean	10.668077	1.963077	136.465385	4.653846	4.663462	150.238462	0.500000
std	2.437754	0.759815	5.112478	0.785432	1.612345	71.158891	0.500964
min	4.600000	0.600000	120.000000	3.300000	1.800000	54.000000	0.000000
25%	8.900000	2.000000	135.000000	4.200000	3.300000	115.750000	0.000000
50%	11.400000	2.000000	137.000000	4.600000	4.600000	166.000000	0.500000
75%	12.700000	2.200000	138.000000	5.100000	5.900000	171.000000	1.000000
max	14.800000	5.000000	193.000000	8.100000	10.600000	1112.000000	1.000000

المصدر: من إعداد الباحثه من تحليل البيانات باستخدام برنامج بايثون إصدار 3.0

التفسير :

من جدول (3-10) تظهر الإحصاءات الوصفية لعينة مكونة من 260 مريضاً تبايناً واضحاً في مؤشرات وظائف الكلى حيث يبلغ متوسط الكرياتينين 1.96 mg/dl والفسفور 4.66 mg/dl مما يعكس إعتلالاً كلوياً مزمنياً ، إلى جانب إنخفاض متوسط الهيموجلوبين إلى 10.67 mg/dl المؤشر على فقر الدم المصاحب ، كما تكشف القيم المتطرفة خاصة في متغير الكوليسترول عن حالات شاذة ، في حين يتسم المتغير التابع Y بتوزيع متوازن تماماً (وسط 0.5) يسهل عملية النمذجة التنبؤية .

• القيم المفقودة :

جدول (3-11) : عدد القيم المفقودة لكل متغير في مجموعة البيانات

Hb	0
S.creatinine	0
S.Na	0
S.K	0
S.PhosPhorus	0
S.cholesterol	0
Y	0

المصدر: من إعداد الباحثه من تحليل البيانات باستخدام برنامج بايثون إصدار 3.0

من جدول (3-11) نلاحظ أنه لا توجد قيم مفقوده في أي من الأعمدة .

• دقة النموذج accuracy_score

Accuracy_score	98.72
----------------	-------

التفسير: الدقة العالية 98.72% تشير إلى أن نموذج الغابة العشوائية قام بتصنيف العينات بشكل صحيح جداً .

- مقاييس أداء خوارزمية الغابة العشوائية في التصنيف :
- جدول (12-3) : مقاييس أداء خوارزمية الغابة العشوائية في التصنيف

	precision	Recall	f1-score	Support
غير مصابين	0.97	1.00	0.99	39
مصابين	1.00	0.97	0.99	39
Micro avg	0.99	0.99	0.99	78
macro avg	0.99	0.99	0.99	78
weighted avg	0.99	0.99	0.99	78

المصدر: من إعداد الباحثه من تحليل البيانات بإستخدام برنامج بايثون إصدار 3.0

التفسير من جدول (12-3) :

- فئة الغير مصابين :

بناءً على تحليل البيانات الواردة في الجدول (12-3) نلاحظ أن أداء النموذج في تصنيف فئة الغير مصابين كان جيداً بشكل ملحوظ حيث بلغت الدقة قيمة عالية تساوي 0.97 مما يشير إلى أن معظم التوقعات التي صنفها النموذج على أنها تنتمي لهذه الفئة كانت صحيحة ، كما أن الإستدعاء كان مرتفعاً أيضاً وكانت قيمته 1.00 مما يعني أن النموذج إستطاع التعرف وتصنيف الغالبية العظمى من العينات الفعلية التي تنتمي إلى فئة الغير مصابين بشكل صحيح .

إنعكس هذا التوازن بين الدقة والإستدعاء في نتيجة F1 المرتفعة وكانت 0.99 والتي تظهر كفاءة النموذج بشكل عام في التعامل مع هذه الفئة ، وقد تم حساب هذه المقاييس بناءً على دعم (Support) مقداره 39 عينة مما يمنح النتائج مصداقية إحصائية معقولة لحجم هذه الفئة في مجموعة البيانات .

- فئة المصابين :

بناءً على تحليل البيانات الواردة في الجدول (3-12) يظهر أداء النموذج في تصنيف فئة المصابين بأنه ممتاز وقوي جداً حيث تشير قيمة الدقة العالية وهي 1.00 إلى أن النسبة الكبيرة من التوقعات التي صنفها النموذج على أنها تنتمي لهذه الفئة كانت صحيحة ، كما أن الإستدعاء كان مرتفعاً أيضاً وكانت قيمته 0.97 توضح قدرة النموذج الفائقة على التعرف وتصنيف الغالبية العظمى من العينات الفعلية التي تنتمي إلى فئة المصابين بشكل صحيح .

إنعكس هذا الأداء المتميز والتوازن العالي بين الدقة والإستدعاء في نتيجة F1 المرتفعة وكانت 0.99 مما يشير إلى كفاءة شبه كاملة للنموذج في التعامل مع هذه الفئة ، وقد تم حساب هذه المقاييس بناءً على دعم (Support) مقداره 39 عينة مما يمنح النتائج مصداقية إحصائية ويعطي مؤشراً واضحاً على قدرة النموذج العالية في تمييز هذه الفئة بدقة كبيرة .

- المتوسطات :

بعد تحليل أداء النموذج على الفئتين الغير مصابين والمصابين بشكل منفصل ، يمكن تقييم أدائه الإجمالي من خلال النظر إلى المتوسطات الحسابية للمقاييس ، يبلغ المتوسط الحسابي البسيط (Macro avg) النتيجة 0.99 مما يعني أن متوسط أداء النموذج عبر الفئتين دون الأخذ في الاعتبار عدد العينات في كل فئة هو أداء ممتاز وشبه كامل .

أما المتوسط المرجح (Weighted avg) والذي يأخذ في الاعتبار عدد العينات في كل فئة (حيث كان الدعم في لكل فئة 39 عينة) فقد بلغ أيضاً 0.99 وهذا يؤكد أن الأداء المتميز للنموذج ليس ناتجاً عن إختلال في توزيع البيانات بل هو أداء قوي ومتسق عبر الفئتين حتي عند ترجيح النتائج بناءً على حجم كل فئة في مجموعة البيانات مما يعزز من مصداقية النموذج وموثوقيته في التصنيف بشكل عام .

• مصفوفة الارتباك :

$$\begin{bmatrix} 39 & 0 \\ 1 & 38 \end{bmatrix}$$

التفسير :

- الصف الأول (فئة الغير مصابين) :

39 : عينة تم تصنيفها بشكل صحيح (True Positive) لفئة الغير مصابين .

0 : تم تصنيفها خطأ (False Negative) لفئة المصابين .

- الصف الثاني (الفئة 1) :

1 : عينة واحدة تم تصنيفها خطأ (False Positive) لفئة الغير مصابين .

38 : عينة تم تصنيفها بشكل صحيح (True Negative) لفئة المصابين .

• الأهمية النسبية للمتغيرات :

جدول(3-13) : الأهمية النسبية للمتغيرات

	Feature	Importance
1	S.cholesterol	0.406626
2	S.PhosPhorus	0.232888
3	Hb	0.220826
4	S.creatinine	0.102819
5	S.Na	0.026204
6	S.K	0.010637

المصدر: من إعداد الباحثه من تحليل البيانات باستخدام برنامج بايثون إصدار 3.0

التفسير من جدول (3-13) :

تظهر نتائج تحليل الأهمية النسبية للمتغيرات أن الكوليسترول في الدم (S.cholesterol) هو العامل الأكثر تأثيراً في النموذج للتنبؤ بالمتغير المستهدف ، حيث تبلغ أهميته النسبية حوالي 0.407 ، مما يعني أنه يساهم بأكثر من 40% في

قدرة النموذج على التنبؤ . ويليه في الأهمية الفسفور في الدم (S.phosphorus)
بنسبة 0.233 ، ثم الهيموجلوبين (Hb) بنسبة 0.221 ، مما يشير إلى دور هذه
العوامل الثلاثة الرئيسي في النموذج .
أما الكرياتينين في الدم (S.creatinine) فيحتل المرتبة الرابعة بأهمية نسبية تقدر
ب 0.103 ، في حين تبدو مستويات الصوديوم (S.Na) و البوتاسيوم (S.K)
أقل تأثيراً ، حيث تبلغ أهميتهما 0.026 و 0.011 على التوالي ، مما يعكس دوراً
محدوداً لهما في هذا السياق التنبؤي مقارنة بالعوامل الأخرى .

6-3 : نتائج تحليل البيانات باستخدام خوارزمية المتجه الداعم Support Vector (Machine) باستخدام برنامج بايثون :

• دقة النموذج accuracy_score

Accuracy_score	96.15
----------------	-------

التفسير:

هذا يعني أن النموذج دقيق جداً ويمكن الإعتماد عليه في المساعدة على التشخيص الطبي ، حيث يصنف 96 من كل 100 حالة بشكل صحيح .

• مقاييس أداء خوارزمية المتجه الداعم في التصنيف :

جدول (3-14) : مقاييس أداء خوارزمية المتجه الداعم في التصنيف

	precision	Recall	f1-score	Support
مصابين	0.95	0.97	0.96	39
غير مصابين	0.97	0.95	0.96	39
Micro avg	0.96	0.96	0.96	78
macro avg	0.96	0.96	0.96	78
weighted avg	0.96	0.96	0.96	78

المصدر: من إعداد الباحثه من تحليل البيانات باستخدام برنامج بايثون إصدار 3.0

التفسير من جدول (3-14) :

- الدقة Precision :

يظهر الجدول (3-14) أداءً قوياً للنموذج في التصنيف ، بالنسبة لفئة الغير مصابين بلغت الدقة 0.95 مما يعني أن 95% من التوقعات التي صنفها النموذج على أنها تنتمي إلى فئة الغير مصابين كانت صحيحة .

أما بالنسبة لفئة المصابين فقد سجلت الدقة قيمة أعلى بلغت 0.97 مما يعكس أن 97% من التوقعات الخاصة بهذه الفئة كانت صحيحة . تشير هذه النتائج إلى موثوقية عالية للنموذج في تصنيف كلتا الفئتين مع أداء متميز قليلاً في تمييز فئة المصابين .

- الإستدعاء Recall :

يُظهر مقياس الإستدعاء في الجدول قدرة عالية للنموذج على التعرف على العينات الفعلية لكل فئة وتصنيفها بشكل صحيح .

بالنسبة لفئة الغير المصابين سجل الإستدعاء قيمة تساوي 0.97 مما يعني أن النموذج نجح في إكتشاف وتصنيف 97% من جميع العينات التي تنتمي حقاً إلى هذه الفئة وهذا يشير إلى أداء ممتاز في إسترجاع حالات فئة الغير مصابين .

أما بالنسبة لفئة المصابين فقد بلغت قيمة الإستدعاء 0.95 وهو ما يعني أن النموذج إستطاع أن يحدد 95% من العينات الفعلية لفئة المصابين بشكل صحيح ، يُظهر هذا أداءً قوياً وجيداً جداً أيضاً وإن كان أقل قليلاً من أدائه مع فئة الغير مصابين .

- نتيجة F1 :

تُظهر نتيجة F1 التي تبلغ 0.96 لكل من فئة الغير مصابين وفئة المصابين توازناً ممتازاً وقوة شاملة في أداء النموذج ، مقياس F1 هو متوسط بين الدقة والإستدعاء وقيمتها المرتفعة والمتساوية لكلا الفئتين تعني أن النموذج لم يحقق دقة عالية فحسب في توقعاته (تجنب الأخطاء الإيجابية الخاطئة) بل كان فعالاً بنفس القدر في إستدعاء العينات الفعلية (تجنب الأخطاء السلبية الخاطئة) ، وهذا يشير إلى كفاءته العالية وموثوقيته في تصنيف كلتا الفئتين بشكل متسق وفعال .

- المتوسطات :

يبلغ المتوسط الصغير (Micro Average) قيمة تساوي 0.96 مما يعني أن الأداء الإجمالي للنموذج عند حساب جميع التصنيفات الصحيحة والخاطئة مباشرة كان متميزاً

كذلك يُظهر المتوسط الكبير (Macro Average) والذي يحسب ببساطة متوسط أداء كل فئة على حدة دون مراعاة حجمها نفس القيمة 0.96 هذا يشير إلى أن الأداء الممتاز للنموذج مستقر عبر الفئتين دون تحيز لفئة على حساب أخرى .

والأهم من ذلك أن المتوسط المرجح أو المتوسط الموزون (Weighted Average) الذي يأخذ في الاعتبار عدد العينات في كل فئة ليعطي صورة أكثر واقعية عن الأداء على مجموعة البيانات الفعلية يسجل أيضاً 0.96 هذا التأكيد النهائي يؤكد أن التميز في الأداء ليس حصيلة إختلال في توزيع العينات بل هو نتاج كفاءة حقيقية ومتوازنة للنموذج في التعامل مع جميع الفئات بغض النظر عن حجمها النسبي .

• مصفوفة الارتباك :

$$\begin{bmatrix} 38 & 1 \\ 2 & 37 \end{bmatrix}$$

التفسير :

تعتبر 38 تصنيفات صحيحة لفئة الغير مصابين . وتعتبر 37 تصنيفات صحيحة لفئة المصابين .

أما 2 تعتبر تصنيفات خاطئة لفئة الغير مصابين (تم تصنيفها مصابين) . و 1 تعتبر تصنيفات خاطئة لفئة المصابين (تم تصنيفها غير مصاب) .

• الأهمية النسبية للمتغيرات :

جدول (3-15) : الأهمية النسبية للمتغيرات

	Feature	Importance
1	S.creatinine	2.010524
2	Hb	1.938159
3	S.PhosPhorus	1.144592
4	S.cholesterol	0.699010
5	S.Na	0.260441
6	S.K	0.255082

المصدر: من إعداد الباحثه من تحليل البيانات باستخدام برنامج بايثون إصدار 3.0

التفسير من جدول (3-15) :

تظهر نتائج تحليل الأهمية النسبية للمتغيرات أن الكرياتينين في الدم (S.creatinine) هو العامل الأكثر تأثيراً في النموذج للتنبؤ بالمتغير المستهدف ، حيث تبلغ أهميته النسبية حوالي 2.01 ، يليه في الأهمية الهيموجلوبين (Hb) بأهمية عالية جداً وقريبة منه 1.94 ، ثم الفسفور (S.phosphorus) بأهمية 1.14 .
على عكس النتائج السابقة إنخفضت الأهمية النسبية للكوليسترول (S.cholesterol) بشكل كبير ليحتل المرتبة الرابعة بأهمية 0.70 . وما زالت أهمية الصوديوم (S.Na) و البوتاسيوم (S.K) محدودة حيث تبلغ 0.26 و 0.25 على التوالي .

7-3 : التحليل والتفسير الشامل للنتائج بعد التحليل :

بعد تطبيق نموذج الإنحدار اللوجستي وخوارزميتي الغابة العشوائية والمتجه الداعم على بيانات مرضى الكلى ، تم الحصول على النتائج التالية :

المعيار / النموذج	الإنحدار اللوجستي	الغابة العشوائية	المتجه الداعم
الدقة الكلية	93.95%	98.72%	96.15%
التوازن العام (الدقة / الإستدعاء)	جيد (حساسية عالية لفئة الغير مصابين)	ممتاز ومتوازن لكلا الفئتين	متوازن ومرتفع
المتغير الأكثر أهمية	الكرياتينين (تأثير موجب قوي 18.65)	الكوليسترول (الأهمية : 40.66%)	الكرياتينين
متغيرات ذات تأثير بارز	الكرياتينين (معامل : 18.65) الهيموجلوبين (معامل : - 1.90) الفسفور (معامل : 1.90)	الكوليسترول الفسفور (23.29%) الهيموجلوبين (22.08%)	الكرياتينين الهيموجلوبين الفسفور
إختبارات جودة النموذج	- Hosmer-lemeshow : P=0.9889 (مطابقة جداً للبيانات) - Wald : أكد دلالة المتغيرات الرئيسية		
التقييم النهائي	نموذج قوي وتفسيري ، ذو دلالة إحصائية واضحة للمتغيرات	النموذج الأفضل أداءً من حيث الدقة والتوازن عبر جميع المقاييس	نموذج قوي جداً وأدائه قريب من الغابة العشوائية

ونستنتج من خلال مقارنة نتائج دقة التصنيف بين الأساليب المستخدمة أن خوارزمية الغابة العشوائية هي الأفضل من حيث الدقة التصنيفية الكلية (98.72% مقابل 96.15% لخوارزمية المتجه الداعم و93.59% للانحدار اللوجستي) وكذلك من حيث توازن الأداء عبر جميع مقاييس التقييم الإحصائية .

الفصل الرابع

النتائج والتوصيات

الفصل الرابع

النتائج و التوصيات

1-4 : تمهيد :

يهدف هذا الفصل إلى عرض النتائج التي توصلت إليها الدراسة ومناقشتها ، إلى جانب التوصيات المستمدة منها . فقد إعتمدت الدراسة على تقييم أداء نموذج الانحدار اللوجستي وخوارزميتي الغابة العشوائية والمتجه الداعم وذلك في تصنيف بيانات مرضى الكلى وتحديد دقة كل نموذج . كما تم تحليل أهم المؤشرات المخبرية المساهمة في عملية التنبؤ بالمرض .

2-4 : النتائج :

1. نموذج الانحدار اللوجستي له قدرة على تصنيف بيانات الدراسة حيث حقق دقة تصنيف بلغت (93.59%) .
2. خوارزمية الغابة العشوائية (RF) تتصف بدقة تصنيف البيانات حيث سجلت أعلى دقة بين النماذج الثلاثة بلغت (98.72%) .
3. خوارزمية المتجه الداعم (SVM) تتصف بدقة تصنيف البيانات حيث حققت دقة تصنيف بلغت (96.15%) .
4. خوارزمية الغابة العشوائية أدق في التصنيف من نموذج الانحدار اللوجستي وخوارزمية المتجه الداعم ، كانت دقة خوارزمية الغابة العشوائية (98.72%) أعلى من دقة خوارزمية المتجه الداعم التي كانت (96.15%) ودقة الانحدار اللوجستي (93.59%) .
5. خوارزمية المتجه الداعم أدق من نموذج الانحدار اللوجستي في تصنيف البيانات حيث تفوقت دقة خوارزمية المتجه الداعم (96.15%) على دقة اللوجستي (93.59%) .

6. تعتبر خوارزمية الغابة العشوائية الأفضل في تصنيف مرضى الكلى بناءً على البيانات المستخدمة .

7. الكرياتينين والهيموجلوبين والفسفور هم أهم المؤشرات المخبرية للتنبؤ بأمراض الكلى ، كما أثر متغير الكوليسترول تأثيراً متوسطاً في التنبؤ بأمراض الكلى ومتغير الصوديوم و البوتاسيوم كان تأثيرهما ضئيلاً إحصائياً في التنبؤ .

8. نموذج الانحدار اللوجستي وخوارزميتي الغابة العشوائية والمتجه الداعم جميعها قابلة للتطبيق في المجال الطبي ، مع تفوق الخوارزميات القائمة على التعلم الآلي (الغابة العشوائية RF و المتجه الداعم SVM) من حيث الدقة .

3-4 : التوصيات :

1. نوصي بإعتماد خوارزمية الغابة العشوائية كحل أساسي في الأنظمة التشخيصية لأمراض الكلى ، مع دمج خوارزمية المتجه الداعم كطبقة تحقق إضافية ، والإحتفاظ بنموذج الانحدار اللوجستي كأداة تفسيرية مساندة تقدم رؤى مباشرة وقابلة للفهم للمتخصصين في المجال الصحي .

2. تطبيق نموذج الانحدار اللوجستي وخوارزميتي الغابة العشوائية والمتجه الداعم على مجموعات بيانات أكبر من مستشفيات متعددة ومناطق جغرافية مختلفة لضمان عمومية النتائج .

3. دراسة تأثير حجم العينة على دقة النماذج الثلاثة ومقارنة أدائها مع زيادة البيانات .

4. إختبار خوارزميات أخرى متقدمة مثل الشبكات العصبية أو XGBoost أو Naïve Bayes لمقارنة الأداء .

5. تطبيق النماذج على أمراض أخرى مثل : السكري ، أمراض القلب ، السرطان ، الأمراض العصبية .

6. إضافة متغيرات مثل العمر والجنس والتاريخ العائلي والعوامل الوراثية لتحسين دقة النماذج والخوارزميات .
7. بناء نماذج وخوارزميات تركز على الكشف المبكر عن أمراض الكلى قبل تطورها ، بإستخدام بيانات طويلة المدى .

المراجع والمصادر

المراجع والمصادر

1. القرآن الكريم : سورة طه ، الآية (114) .
2. Abass HJ. Steps for sample selection in social research .
Iraqi Journal of Science . (2019) ; 60(3) .
3. عباس ع. إستخدام نموذج الإنحدار اللوجستي في التنبؤ بالدوال ذات المتغيرات الإقتصادية التابعة النوعية . جامعة كركوك للعلوم الإدارية والإقتصادية ؛ (2012) .
4. عمران النصر اوي ن. إستخدام نموذج البوتستراب في تحليل النماذج المعلمية وشبة المعلمية والمقارنة بينها . جامعة كربلاء . كلية الإدارة والإقتصاد ؛ (2017) .
5. محمد على ه ، على ط . تعظيم كفاءة التحليل لمنحنى اللوجستيك بإستخدام التحويل بالقوى . المجلة العراقية للعلوم الإحصائية . (2008) (14) .
6. سعيد ر . إستخدام نموذج الإنحدار اللوجستي في دراسة العوامل المساعدة على تشخيص حالات الإصابة بسرطان المثانة . جامعة بغداد . كلية الإدارة والإقتصاد . قسم الإحصاء . (2015) .
7. دعيش م ، ساري م . نموذج الإنحدار اللوجستي ، مفهومه ، خصائصه ، تطبيقاته . مجلة السراج في التربية وقضايا المجتمع . (2015) (1) .
8. مبروك إ ، سعيد ع . إستخدام الإنحدار اللوجستي في تقدير معدلات البطالة في الدول العربية . الجزائر جامعة محمد بوضياف بالمسيلة (2018-2019) .
9. العلي إ . أسس التحليل الإحصائي متعدد المتغيرات . سوريا . جامعة تشرين . كلية الإقتصاد . (2020) .
10. محمود ع ، الجاعوني ، خليل ف . إستخدام تقنية الإنحدار اللوجستي الثنائي الإستجابة في دراسة أهم المحددات الإقتصادية والإجتماعية لكفاية دخل

- الأسرة . مجلة جامعة دمشق للعلوم الاقتصادية والقانونية . (2011) (27) (1)
- .
11. David W . & Lemeshow S . Applied logistic regression . newyork . wiley . (2013) .
12. قاسم ب . تحليل أثر بعض المتغيرات في الإصابة بمرض اللثة باستخدام نموذج الانحدار اللوجستي . مجلة العلوم الاقتصادية . (2011) (7) (27) .
13. حبيب ك ، بشارة إ . تصنيف زهرة القرحية . سوريا . جامعة تشرين كلية الهندسة المعلوماتية . قسم الذكاء الصناعي . (2020) .
14. Zeleha S, Sharabi M, Choen R. The secret of digital pedagogical competence: the mediation of teacher self – efficacy between personality traits and digital competence . In: Daniela I, editor. Human, technologies and quality of education, 2020. Riga: University of Latvia Press; 2020 .
15. Al-Joubi, A. (2021) . The Historical Development of Machine Learning : A Conceptual Analysis . Journal of Artificial Intelligence Research, 2(4) .
16. Larsen–Freeman D. Techniques and principles in language teaching . 3rd ed. Oxford: Oxford University Press; 2013 . Jiang N, Fang F, Zhang C. How concreteness and congruence affect vocabulary learning and processing in Chinese learners of English . lang Learn Technol . 2017; 21 (3) .

17. وزان م ، عطية ع . تعلم الآلة وعلم البيانات . العراق . جامعة القدس
كلية علوم الحاسوب وتكنولوجيا المعلومات .

18. Dey A. Machine learning algorithms: a review
[Internet].2016 [cited 2024 Nov 15] .Available from :
<https://www.semanticscholar.org/paper/Machine-Learning-Algorithms-A-Review-Dey/56e8863838b4dcc4790108cd1e7e680a104a7c30> .

19. <https://www.ejaba.com/tag/%D9%85%D8%A7-%D9%87%D9%8A-%D8%B7%D8%B1%D9%82-%D8%AA%D8%B5%D9%86%D9%86%D9%8A%D9%81-%D8%A7%D9%84%D8%A8%D9%8A%D8%A7%D9%86%D8%A7%D8%AA-%D8%A7%D9%84%D8%A5%D8%AD%D8%B5%D8%A7%D8%A6%D9%8A%D8%A9>.

20. <https://www.ejaba.com/tag/%D9%83%D9%8A%D9%81-%D9%85%D9%83%D9%86-%D8%AA%D8%B5%D9%86%D9%8A%D9%81-%D8%A7%D9%84%D8%A8%D9%8A%D8%A7%D9%86%D8%A7%D8%AA-%D8%A7%D9%84%D8%A5%D8%AD%D8%B5%D8%A7%D8%A6%D9%8A%D8%A9>.

21. <https://www.ejaba.com/tag/%D9%85%D8%A7-%D9%87%D9%8A-%D8%A3%D9%87%D9%85%D9%8A%D8%A9>

%D8%AA%D8%AA%D8%B5%D9%86%D9%8A%D9%81-
%D8%A7%D9%84%D8%A8%D9%8A%D8%A7%D9%86%D
8%A7%D8%AA-
%D8%A7%D9%84%D8%A5%D8%AD%D8%B5%D8%A7%
D8%A6%D9%8A%D8%A9-
%D9%88%D9%83%D9%8A%D9%81-
%D9%8A%D9%85%D9%83%D9%86-
%D8%A7%D8%B3%D8%AA%D8%AE%D8%AF%D8%A7%
D9%85%D9%87%D8%A7%D9%81%D9%8A-
%D8%A7%D9%84%D8%A3%D8%A8%D8%AD%D8%A7%
D8%AB-
%D8%A7%D9%84%D8%A5%D8%AD%D8%B5%D8%A7%
D8%A6%D9%8A%D8%A9.

22. <https://www.ejaba.com/tag/%D9%85%D8%A7-%D9%87%D9%8A-%D8%A3%D9%86%D9%88%D8%A7%D8%B9-%D8%AA%D8%B5%D9%86%D9%8A%D9%81-%D8%A7%D9%84%D8%A8%D9%8A%D8%A7%D9%86%D8%A7%D8%AA>.

23. <https://acaiaworld.com/blog/classification-and-its-most-famous-algorithms/>.

24. العنزي ، على س . إنعكاسات السلوكيات الوظيفية للعاملين في أداء المنظمات . دار ومكتبة عدنان للطباعة والنشر والتوزيع . بغداد . (2017)
(1) .

25. العامري ص ، الغالب ط . الإدارة والأعمال . دار وائل للنشر والتوزيع . الأردن . عمان . (2011) (3) .
26. كنعان ن . إتخاذ القرارات الإدارية بين النظرية والتطبيق . مكتبة دار الثقافة للنشر والتوزيع . الأردن . عمان . (1995) (3) .
27. عامر ط ، المصري إ . صناعة وإتخاذ القرار . مؤسسة طيبة للنشر والتوزيع . القاهرة . (2015) (1) .
28. موسى م . تحليل الأداء الأكاديمي للطلاب بإستخدام تقنيات تعلم الآلة . مجلة جامعة الملك عبدالعزيز . الرياض . معهد الإدارة العامة . قطاع تقنية المعلومات . (2021) (10) (1) .
29. محمود أ . تطبيقات أسلوب أشجار إتخاذ القرار في علم التعلم الآلي . مجلة إكليل للدراسات الإنسانية . الجامعة المستنصرية . (2022) (10) .
30. الحديثي أ ، الخفاجي ل . المفاضلة بين قرارات الطاقة بإستخدام شجرة القرارات . مجلة التقني . (2013) (26) (7) .
31. <https://sciences24.com/%D8%AE%D9%88%D8%A7%D8%B1%D8%B2%D9%85%D9%8A%D8%A9-%D8%A3%D8%B4%D8%AC%D8%A7%D8%B1-%D8%A7%D9%84%D9%82%D8%B1%D8%A7%D8%B1-decision-trees/> .
32. https://ar.m.wikipedia.org/wiki/%D8%AA%D8%B5%D9%86%D9%8A%D9%81_%D8%A5%D8%AD%D8%B5%D8%A7%D8%A6%D9%8A.
33. ساعي م ، اسبر ج ، جولحة أ . تحليل خوارزميات العقدة والتصنيف للتنبؤ بإستهلاك الطاقة الكهربائية . مجلة جامعة تشرين . للعلوم الهندسية . (2022) (44) (5) .

34. الجبوري ع ، الحنون أ . إستخدام نماذج ARIMA و الغابة العشوائية للتنبؤ ببيانات الأنواء الجوية . المجلة العراقية للعلوم الإحصائية . العراق . جامعة الموصل . كلية علوم الحاسوب والرياضيات . قسم الإحصاء المعلوماتية . (2022) (19) (2) .
35. <https://sciences24.com/D8%B4%D8%B1%D8%AD-%D9%85%D8%A8%D8%B3%D8%B7-%D9%84%D8%AE%D9%88%D8%A7%D8%B1%D8%B2%D9%85%D9%8A%D8%A9-%D8%A7%D9%84%D8%BA%D8%A7%D8%A8%D8%A9-%D8%A7%D9%84%D8%B9%D8%B4%D9%88%D8%A7%D8%A6%D9%8A%D8%A9-random-fo/>.
36. <https://sciences24.com/%D8%AE%D9%88%D8%A7%D8%B1%D8%B2%D9%85%D9%8A%D8%A9-%D8%A3%D8%B4%D8%AC%D8%A7%D8%B1-%D8%A7%D9%84%D9%82%D8%B1%D8%A7%D8%B1-decision-trees/> .
37. Liu H An application of support vector machine for evaluating credit risk of bank.In: Procceeding of the 7th International Conference on Innovation and Management;2010 .
38. Jakkula V. Tutorial on support vector machine (SVM) .Pullman :School of EECS , Washington State University ; 2010 .

39. Yin HC . A novel machine learning model for risk management , In :Proceeding of the first Asia Pacific Conference on Global Business , Economics , Finance and Social Science (AP14 Singapore Conference) ; 2014 .
40. Saif S, et al. Finding rules for audit opinions prediction through data mining method . Eur Online J NatSoc Sci . 2010; 1(2) .
41. Ma Y , Guo G . Support vector machines applications . Cham: Springer International Publishing ; 2014 .
42. Pradhan A . Support Vector machine – A survey . Int J Emerg Technol Adv Eng . 2012 ; 2(8) .
43. Danenas P , Garsva G . Svm and xbrl based decision support system for credit risk evaluation . 2011 .
44. Min S , Lee J , Han I . Hybrid genetic algorithms and support vector machines for bankruptcy prediction . Expert Syst Appl. 2006 ; 31(3) , Steinwart I , Christmann A. Support vectr machine . New York : Springer Science & Business Media ; 2008 .
45. Danenas P , Garsva G . Selection of support vector machines –based classifiers for credit risk domain . Expert Syst Appl . 2015; 42(6) , Bellotti T , Crook J. Supportvector machines for credit scoring and discovery of significant features. Expert Syst Appl. 2009 ; 36(2) .
46. <https://www.webteb.com> .

47. Huang C, Chen M , Wangv C. Credit scoring with a data a data mining approach based on support vector machines . Expert Syst Appl . 2007; 33(4) .
48. ساني د . حافظ على كليتك . مؤسسة سمربان للكلية . جامعة عين شمس . القاهرة . (2014) (1) .
49. تاغليفييري ل . البرمجة بلغة بايثون . أكاديمية حسوب . (2020) .

الملاحق

(1) الملحق

Hb	S.creatinine	S.Na	S.K	S.PhosPhorus	S.cholesterol	Y
7.7	4	134	4.7	4.8	131	1
12.2	3	130	4.1	6.6	135	1
10.6	4	135	4.6	5.4	152	1
9.4	4	132	4.3	5.8	111	1
6.7	3	145	5.2	4.3	115	1
9.8	2	134	4	5.4	102	1
10.6	3	135	3.8	7	96	1
9.6	2	129	3.6	5.3	103	1
9.3	2	138	4.3	4.9	165	1
9.4	2.2	173	3.6	5.6	107	1
12.1	2.1	132	4.3	5.6	142	1
10.9	2.2	137	6.3	4.9	116	1
6.6	3	133	4.3	5.4	98	1
13.8	2.3	140	5	5.7	155	1
11.1	2.4	133	4.1	6.3	113	1
12.4	2.4	134	5.3	6.3	176	1
10.6	2	134	3.6	6	130	1
9.6	2.4	128	4	8.7	98	1
11.4	2	134	5.6	6.6	163	1
9.4	2	129	4.1	6.6	102	1
10.3	2	137	5.4	5.5	134	1

9.4	2.2	123	4.2	6.6	134	1
12.1	2	130	6	6.6	154	1
9.5	2	131	4.7	6.1	117	1
9.1	2	135	4.9	5.9	129	1
11.5	3	128	4.9	5.7	162	1
8.5	2	137	3.8	4	83	1
11.3	3	132	3.9	5.5	161	1
7.9	2	137	4.5	5.5	100	1
12.6	2	133	5.1	6.8	142	1
9.8	2.2	129	4.8	5.1	133	1
9.7	2.1	134	5.1	4.9	136	1
9.9	2	132	5.3	5.3	134	1
10.5	1.8	132	5.7	5.6	135	1
4.8	1.9	130	4	5.5	80	1
8.9	1.6	135	5.7	5.3	1112	1
9.2	1.6	138	4.2	5.6	130	1
8.9	2	134	4	5.6	100	1
10.3	2	134	4.6	4.9	126	1
10.2	2	135	4.5	6.6	134	1
9.2	2	142	3.9	8.1	94	1
10.7	2	137	4.3	7	112	1
10	3	136	4.6	5.6	116	1
9.8	3	136	4.6	7	126	1

9.2	3	137	4.6	7.8	113	1
8.6	3.2	132	4.2	6.6	412	1
7.5	4	139	4.5	4.1	159	1
7.3	5	145	5	5.5	76	1
6.5	4	132	4.2	6.1	108	1
11.8	4	138	4.7	4.2	95	1
7.2	3	140	4	5.7	159	1
6.7	2	136	5.5	5.7	92	1
8.1	3	137	4.4	6.6	108	1
7.8	2	136	5.7	5.9	104	1
7.7	2	137	5.4	5.9	112	1
8.2	2.2	139	5.8	6.7	86	1
7.5	2.1	135	5.2	4.3	114	1
4.9	2.2	140	5.2	6.6	84	1
5.6	2.2	136	7.1	7.5	83	1
8.2	2.3	132	3.9	7.9	87	1
6.7	2.4	120	4.7	4.3	69	1
6.3	2.4	135	5.2	5.9	151	1
6.7	2	135	4.1	5.1	79	1
7.9	2.4	135	5.1	6.4	73	1
5.6	2	138	4.5	5.1	177	1
9.2	3	135	4.7	6.2	87	1
8.2	2	133	4.2	7.2	80	1

10.1	2.2	131	8.1	6	106	1
7.8	2	135	6.8	3.4	194	1
5	2	132	4	5.5	142	1
10.1	2	125	4.3	7	94	1
13	3	135	4.5	4.5	79	1
9.1	2	134	5	2.4	146	1
5.3	3	140	4.1	4.3	82	1
5.5	4	132	4.8	4.3	147	1
6	4	137	4.9	6.2	135	1
5.3	4	127	4.7	4	135	1
6	4	135	3.5	4	180	1
8.1	3	130	5.1	3.9	131	1
6.1	2	137	4.6	4.5	205	1
10.1	3	136	4.5	3.4	140	1
6.2	2	135	5.6	4.6	132	1
8.8	2	135	5.5	4.3	148	1
9.3	2.2	136	4.4	6.7	121	1
7.2	2.1	134	5.4	6.5	158	1
7	2.2	136	4.8	8.1	109	1
5.2	2.2	136	4.4	6.3	147	1
5.3	2.3	136	4.6	6.3	134	1
6.9	2.4	140	5.5	7	124	1
7.3	2.4	140	4.4	5.9	110	1

4.6	2	143	4.9	5.9	104	1
9.5	2.4	140	5.6	3.5	92	1
8.8	2	137	4.1	4.9	256	1
10.2	2	139	4.2	4.3	110	1
11.8	2	133	4.6	4.6	60	1
10.5	2.2	136	5	5.9	104	1
11	2	133	5.1	3.1	80	1
8.8	2	133	6.5	10.6	58	1
9.3	2	134	5.2	8.1	75	1
8.9	3	136	4.3	6.5	125	1
11.4	2	138	4.8	6.5	72	1
7.2	3	131	4.5	5.2	106	1
8.9	2	138	4.5	4.7	115	1
8.8	2	131	6.2	2.2	54	1
11.3	2.2	134	3.4	3.2	137	1
7.6	2.1	193	3.4	3.2	137	1
8.1	2.3	149	4.1	3.3	113	1
10.3	2.4	134	4.9	2.1	80	1
8.7	2.4	135	4.8	3.9	133	1
9.5	2	134	4.6	2.3	157	1
12.1	2.4	136	3.8	3.7	196	1
12.2	2	132	4.5	3.7	187	1
12.8	4	131	4.9	8	88	1

7.3	2	140	4.4	4	137	1
7.7	2.2	131	4.3	2.4	125	1
7.3	2	136	4.3	4.5	149	1
7.6	2.2	139	6.7	3.3	79	1
6.7	2.1	142	6.1	3.5	111	1
10	2.2	134	4.4	3.5	138	1
7.7	2.2	140	6	4.2	70	1
10.6	2.3	130	3.9	4	123	1
9.8	2.4	128	4.6	4	189	1
6.7	2.4	137	7.7	4.2	98	1
7	2	140	6	5.6	113	1
12.7	2.4	133	4.1	5.6	208	1
8.3	2	139	6.3	4.1	102	1
10.7	2	138	5.5	9.7	108	1
8.5	2	145	5	6.6	62	1
8	2	134	6.4	5	88	1
9.7	2	137	5.5	7.4	113	1
7	0.7	138	4	3	173	0
12.9	0.7	139	4.1	3.1	167	0
12.3	0.7	139	4	2.8	165	0
14.2	0.7	137	4.2	2.9	172	0
12.8	0.8	139	4.6	3	169	0
14.8	1.2	138	4.5	2.6	175	0

13.2	0.8	139	4.5	3.3	171	0
12.2	0.8	138	4.2	2.6	171	0
11.4	0.7	137	3.9	3.4	174	0
12	0.8	139	4.6	3.1	172	0
11.6	0.9	137	4.3	3.2	167	0
12	2	136	4.1	3	173	0
13	0.6	137	5	3.6	170	0
11.3	0.9	137	5.4	3.5	173	0
12.9	0.7	136	4	3.7	171	0
12.4	0.7	138	4.2	3.6	165	0
11.9	0.8	139	4.2	3.4	170	0
12.9	0.8	137	3.9	2.9	174	0
11.3	2	135	4.5	3.8	167	0
11.2	0.9	136	4.5	2.7	168	0
12.8	0.7	138	4.2	3.9	176	0
12.6	0.9	137	3.9	3.2	170	0
12.6	1.2	135	3.9	2.4	168	0
11.2	0.9	137	4.8	3	165	0
13.2	1.3	136	5	2	169	0
12.2	0.9	137	3.7	2.5	167	0
11.7	0.9	138	3.8	3.6	170	0
13.8	2	138	4.8	2.5	172	0
12.7	0.8	138	4.8	3.5	164	0

11.7	0.7	138	4.1	3.5	172	0
13.6	1.3	137	4.7	3.6	167	0
14.1	1.2	137	4.4	3.8	166	0
11.5	1.2	136	4.4	2.9	168	0
13	1.2	137	4.8	3.8	170	0
11.8	0.7	137	4	1.8	166	0
12.3	0.7	137	5.2	2.9	173	0
11	1	138	4.4	3	171	0
10.9	1	137	4.5	2.5	168	0
12	0.8	137	4.7	2.6	169	0
12.8	0.8	136	4.3	4.1	170	0
12.9	2	137	4.5	2.7	172	0
12.9	2	137	5.6	3.1	172	0
12.8	2	138	4.6	2.9	172	0
11.9	2	137	4.7	3.8	169	0
12.9	2	135	3.3	3.3	171	0
13.4	2	137	4.1	3.3	169	0
12.9	2	138	3.9	3.5	170	0
13	2	139	4.8	2.4	163	0
12.1	2	135	4.6	3.4	168	0
12.7	1.3	137	4.6	4.3	168	0
12.8	2	138	3.7	3.4	166	0
14.2	2	135	5.4	2.6	170	0

13.4	0.8	138	4.6	2.8	172	0
13.1	2	136	4.5	3.2	176	0
12.8	2	137	3.6	2.6	168	0
12.6	2	138	5.1	3.1	174	0
12.8	0.7	138	4.7	2.2	171	0
10.6	2	139	4.4	2.8	170	0
12.5	2	138	4.8	3.2	173	0
12	1.3	137	4.6	2.2	168	0
12.6	2	136	3.6	2.9	170	0
11	2	136	3.8	3	172	0
13.4	2	138	4.4	2.9	171	0
11.7	1.3	137	4.4	3.6	168	0
12.9	2	139	3.4	3.1	176	0
12.5	2	136	5.2	2.4	164	0
12.9	2	137	4.1	4.1	172	0
14.8	2	137	4.6	2.7	167	0
12	0.7	137	4.3	2.7	172	0
11.9	2	137	4.5	4.2	167	0
12.9	2	135	4.4	2.4	171	0
13.4	2	137	4.4	2.9	171	0
12.5	2	137	4.2	2.9	171	0
13.9	2	137	3.9	2.3	169	0
10.9	2	137	4.5	2.4	172	0

12.4	0.7	139	3.5	3.3	173	0
11.4	2	137	5.4	3.5	169	0
13.1	2	137	4.4	3.5	169	0
12.6	2	137	3.6	2.8	173	0
13.2	2	137	3.9	2.9	170	0
12.5	2	137	4.3	2.9	168	0
12.5	0.9	137	5.7	3.1	170	0
14.4	2	136	4.9	2.1	165	0
12.7	2	138	3.9	3.1	169	0
11.8	2	137	4.6	3.1	172	0
11.6	2	135	3.9	3.2	173	0
13.6	2	136	4.1	2.9	166	0
13.1	2	137	4.3	2.7	168	0
12.3	2	137	4.9	3.1	166	0
12	2	137	4.4	2.5	169	0
11.3	2	136	4.6	2.9	169	0
14.4	2	138	4.7	2.7	165	0
12.8	2	136	4.2	3	170	0
13.5	2	136	4.3	3.1	167	0
13.9	2	138	4.6	2.5	170	0
11.9	2	136	4.1	3.2	173	0
11.7	2	137	3.8	2.9	172	0
12.3	2	137	4.5	3.4	171	0

12.2	2	137	4.1	3.7	171	0
9.8	2	137	4.2	3.9	170	0
10.9	0.8	137	4.9	3.7	165	0
11.4	2	137	4.4	3.6	170	0
11.5	2	136	4.8	2.3	171	0
13	2	138	4.2	3.7	169	0
13	2	138	4.7	3.4	173	0
12.4	2	138	3.8	3.8	171	0
12.2	0.8	135	4.7	3.1	171	0
13.5	0.8	138	3.6	2.5	174	0
13	0.9	137	4.2	3.2	172	0
12.9	0.7	139	4.8	3	175	0
12.8	2	137	4.5	3.6	170	0
13.2	0.8	137	4.3	3.9	172	0
13	1	138	5.4	2.5	175	0
12.8	0.7	137	5.2	3.3	170	0
11.9	1	136	4.5	2	171	0
12.9	0.7	137	4.4	3	170	0
11.2	1.2	137	4.5	3	170	0
12	1.1	137	4.3	2.6	174	0
14	2	136	4	2.5	165	0
12.7	2	135	5	2.9	169	0
13.2	2	137	4.6	2.3	173	0

14	2	137	3.1	3	171	0
11.4	2	137	4	2.9	172	0
13	2	139	4.5	2.9	169	0
12.5	2	137	5	3.2	168	0
11.4	1.2	137	4	3.1	172	0
13.4	2	137	4	3.1	172	0
11.7	2	138	4.6	3.7	170	0
12.7	2	135	4.7	2.3	172	0
12.7	2	135	4.7	2.3	172	0

الملحق (2)

مخرجات تحليل البيانات الإحصائية باستخدام برنامج بايثون :

- تحليل نموذج الإنحدار اللوجستي باستخدام برنامج بايثون :

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import
classification_report, confusion_matrix, accuracy_score
data=pd.read_csv('E:\\Data1.csv', sep=';')
print(data.head())
print(data.info())
print(data.describe())
X = data.drop('y', axis=1)
y = data['y']
X_train, X_test, y_train, y_test =
train_test_split(X,y,test_size=0.3, random_state=42)
logreg = LogisticRegression(max_iter=1000)
logreg.fit(X_train,y_train)
y_pred = logreg.predict(X_test)
cm = confusion_matrix(y_test,y_pred)
print("confusion Matrix:\n", cm)
report = classification_report(y_test,y_pred)
print("classification Report:\n", report)
accuracy = accuracy_score(y_test, y_pred)
print("accuracy_score:\n", accuracy)
```

• مخرجات تحليل نموذج الإنحدار اللوجستي باستخدام برنامج بايثون :

	Hb	S.creatinine	S.Na	S.K	S.PhosPhorus	S.cholesterol	y
0	7.7	4.0	134	4.7	4.8	131	1
1	12.2	3.0	130	4.1	6.6	135	1
2	10.6	4.0	135	4.6	5.4	152	1
3	9.4	4.0	132	4.3	5.8	111	1
4	6.7	3.0	145	5.2	4.3	115	1

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 260 entries, 0 to 259
```

```
Data columns (total 7 columns):
```

```
Hb                260 non-null float64
```

```
S.creatinine      260 non-null float64
```

```
S.Na              260 non-null int64
```

```
S.K               260 non-null float64
```

```
S.PhosPhorus      260 non-null float64
```

```
S.cholesterol     260 non-null int64
```

```
y                260 non-null int64
```

```
dtypes: float64(4), int64(3)
```

```
memory usage: 14.3 KB
```

```
None
```

	Hb	S.creatinine	...	S.cholesterol	y
count	260.000000	260.000000	...	260.000000	260.000000
mean	10.668077	1.963077	...	150.238462	0.500000

std	2.437754	0.759815	...	71.158891	0.500964
min	4.600000	0.600000	...	54.000000	0.000000
25%	8.900000	2.000000	...	115.750000	0.000000
50%	11.400000	2.000000	...	166.000000	0.500000
75%	12.700000	2.200000	...	171.000000	1.000000
max	14.800000	5.000000	...	1112.000000	1.000000

[8 rows x 7 columns]

C:\ProgramData\Anaconda3\lib\site-packages\sklearn\linear_model\logistic.py:433: FutureWarning: Default solver will be changed to 'lbfgs' in 0.22. Specify a solver to silence this warning.

FutureWarning)

confusion Matrix:

[[38 1]

[4 35]]

classification Report:

	precision	recall	f1-score	support
0	0.90	0.97	0.94	39
1	0.97	0.90	0.93	39
micro avg	0.94	0.94	0.94	78
macro avg	0.94	0.94	0.94	78
weighted avg	0.94	0.94	0.94	78

accuracy_score:

0.9358974358974359

- إختبار Hosmer-Lemeshow باستخدام برنامج بايثون :

```
import pandas as pd
import numpy as np
from scipy.stats import chi2
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
data=pd.read_csv('E:\\Data1.csv',sep=';')
X = data.drop('y', axis=1)
y = data['y']
X_train, X_test, y_train, y_test =
train_test_split(X,y,test_size=0.3, random_state=42)
model = LogisticRegression(max_iter=1000)
model.fit(X_train,y_train)
y_pred_proba = model.predict_proba(X_test)[:, 1]
def hosmer_lemeshow_test(y_true, y_pred, groups=10):
    df = pd.DataFrame({'y_true': y_true, 'y_pred': y_pred})
    df['decile'] = pd.qcut(df['y_pred'], groups, labels=False)
    observed = df.groupby('decile')['y_true'].sum()
    expected = df.groupby('decile')['y_pred'].sum()
    n = df.groupby('decile').size()
    hl_stat = ((observed - expected)**2 / (expected * (1 -
expected/n))) .sum()
```

```

dof = groups - 2
p_value = 1 - chi2.cdf(hl_stat, dof)
return hl_stat, p_value, dof
hl_stat, p_value, dof = hosmer_lemeshow_test(y_test,
y_pred_proba)
print (f"Hosmer-Lemeshow: {hl_stat:.4f}")
print(f"p-value: {p_value:.4f}")
print(f"dof: {dof}")

```

• مخرجات نتائج إختبار Hosmer–Lemeshow بإستخدام برنامج بايثون :

```

C:\ProgramData\Anaconda3\lib\site-
packages\sklearn\linear_model\logistic.py:433: FutureWarning:
Default solver will be changed to 'lbfgs' in 0.22. Specify a
solver to silence this warning.
FutureWarning)
Hosmer-Lemeshow: 1.6970
p-value: 0.9889
dof: 8

```

- إختبار Wald بإستخدام برنامج بايثون :

```

import pandas as pd
import numpy as np
import statsmodels.api as sm
from scipy import stats
data=pd.read_csv('E:\\Data1.csv', sep=';')
Q1 = data['S.cholesterol'].quantile(0.25)
Q3 = data['S.cholesterol'].quantile(0.75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
data = data[(data['S.cholesterol'] >= lower_bound) &
(data['S.cholesterol'] <= upper_bound)]
X = data.drop('y' , axis=1)
y = data['y']
X = sm.add_constant(X)
logit_model = sm.Logit(y,X)
result = logit_model.fit()
wald_stats = (result.params / result.bse)**2
p_values = stats.chi2.sf(wald_stats, 1)
wald_result = pd.DataFrame({
    'result.params' : result.params,
    'wald_stats' : wald_stats,
    'p_values' : p_values
})
print(wald_result)
print(result.summary())

```

• مخرجات نتائج إختبار Wald بإستخدام برنامج بايثون :

	result.params	wald_stats	p_values
const	0.921907	0.002315	0.961624
Hb	-1.898964	7.052501	0.007915
S.creatinine	18.651297	7.565800	0.005949
S.Na	-0.108049	1.912411	0.166696
S.K	0.628703	0.186844	0.665556
S.PhosPhorus	1.902692	5.586596	0.018098
S.cholesterol	-0.079954	1.494966	0.221447

Logit Regression Results

=====

Dep. Variable:	y	No. Observations:	257
Model:	Logit	Df Residuals:	250
Method:	MLE	Df Model:	6
Date:	Mon, 12 Jan 2026	Pseudo R-squ.:	0.9523
Time:	10:51:38	Log-Likelihood:	-8.5014
converged:	True	LL-Null:	-178.12
		LLR p-value:	3.148e-70

=====

	coef	std err	z	P> z	[0.025	0.975]
const	0.9219	19.160	0.048	0.962	-36.631	38.475
Hb	-1.8990	0.715	-2.656	0.008	-3.300	-0.497
S.creatinine	18.6513	6.781	2.751	0.006	5.361	31.941
S.Na	-0.1080	0.078	-1.383	0.167	-0.261	0.045
S.K	0.6287	1.454	0.432	0.666	-2.222	3.479
S.PhosPhorus	1.9027	0.805	2.364	0.018	0.325	3.480
S.cholesterol	-0.0800	0.065	-1.223	0.221	-0.208	0.048

=====

- تحليل خوارزمية الغابة العشوائية بإستخدام برنامج بايثون :

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import
classification_report, confusion_matrix, accuracy_score
data = pd.read_csv('E:\\Data1.csv', sep=';')
print(data.head())
print(data.info())
print(data.describe())
print(data.isnull().sum())
x = data.drop('y', axis=1)
y = data['y']
x_train, x_test, y_train, y_test =
train_test_split(x,y,test_size=0.3, random_state=42)
rf_model = RandomForestClassifier(
    n_estimators=100,
    max_depth=10,
    random_state=42
)
rf_model.fit(x_train, y_train)
```



```

y_pred = rf_model.predict(x_test)
print(f"\n accuracy_score: {accuracy_score(y_test,
y_pred)*100:.2f}")
print("\n classification_report:")
print(classification_report(y_test, y_pred))
cm = confusion_matrix(y_test, y_pred)
print("confusion matrix")
print(cm)
feature_importances = pd.Series(rf_model.feature_importances_,
index=x.columns).sort_values(ascending=False)
importance_df = pd.DataFrame({'Feature':
feature_importances.index, 'Importance':
feature_importances.values})
print(importance_df)

```

• مخرجات تحليل خوارزمية الغابة العشوائية باستخدام برنامج بايثون :

	Hb	S.creatinine	S.Na	S.K	S.PhosPhorus	S.cholesterol	y
0	7.7	4.0	134	4.7	4.8	131	1
1	12.2	3.0	130	4.1	6.6	135	1
2	10.6	4.0	135	4.6	5.4	152	1
3	9.4	4.0	132	4.3	5.8	111	1
4	6.7	3.0	145	5.2	4.3	115	1

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 260 entries, 0 to 259
```

```
Data columns (total 7 columns):
```

```
Hb                260 non-null float64
```

```
S.creatinine      260 non-null float64
```

```
S.Na              260 non-null int64
```

```
S.K               260 non-null float64
```

```
S.PhosPhorus      260 non-null float64
```

```
S.cholesterol     260 non-null int64
```

```
y                260 non-null int64
```

```
dtypes: float64(4), int64(3)
```

```
memory usage: 14.3 KB
```

```
None
```

	Hb	S.creatinine	...	S.cholesterol	y
count	260.000000	260.000000	...	260.000000	260.000000
mean	10.668077	1.963077	...	150.238462	0.500000
std	2.437754	0.759815	...	71.158891	0.500964
min	4.600000	0.600000	...	54.000000	0.000000
25%	8.900000	2.000000	...	115.750000	0.000000
50%	11.400000	2.000000	...	166.000000	0.500000
75%	12.700000	2.200000	...	171.000000	1.000000
max	14.800000	5.000000	...	1112.000000	1.000000

```
[8 rows x 7 columns]
```

Hb	0
S.creatinine	0
S.Na	0
S.K	0
S.PhosPhorus	0

```
S.cholesterol    0
y                0
dtype: int64
```

```
accuracy_score: 98.72
```

```
classification_report:
```

	precision	recall	f1-score	support
0	0.97	1.00	0.99	39
1	1.00	0.97	0.99	39
micro avg	0.99	0.99	0.99	78
macro avg	0.99	0.99	0.99	78
weighted avg	0.99	0.99	0.99	78

```
confusion matrix
```

```
[[39  0]
 [ 1 38]]
```

	Feature	Importance
0	S.cholesterol	0.406626
1	S.PhosPhorus	0.232888
2	Hb	0.220826
3	S.creatinine	0.102819
4	S.Na	0.026204
5	S.K	0.010637

- تحليل خوارزمية المتجه الداعم بإستخدام برنامج بايثون :

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.metrics import
classification_report, confusion_matrix, accuracy_score
data = pd.read_csv('E:\\Data1.csv', delimiter=';')
print(data.head())
print(data.info())
print(data['y'].value_counts())
X = data.drop('y', axis=1)
y = data['y']
X_train, X_test, y_train, y_test =
train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)
def manual_scaler(X_train, X_test):
    mean = X_train.mean()
    std_dev = X_train.std()
    std_dev = std_dev.replace(0, 1)
    X_train_scaled = (X_train - mean) / std_dev
    X_test_scaled = (X_test - mean) / std_dev
    return X_train_scaled, X_test_scaled
X_train_scaled, X_test_scaled = manual_scaler(X_train, X_test)
```

```

print("manual_scaler")
svm_linear = SVC(kernel='linear', random_state=42)
svm_linear.fit(X_train_scaled, y_train)
y_pred = svm_linear.predict(X_test_scaled)
accuracy = accuracy_score(y_test, y_pred)
print(f"\n accuracy_score: {accuracy:.4f}")
print("\n classification_report:")
print(classification_report(y_test, y_pred))
print("\n confusion_matrix:")
cm = confusion_matrix(y_test, y_pred)
print(cm)
importance = svm_linear.coef_[0]
feature_importance = pd.DataFrame({
    'feature': X.columns,
    'importance': abs(importance)
}).sort_values('importance', ascending=False)
print(feature_importance)

```

- مخرجات تحليل خوارزمية المتجه الداعم باستخدام برنامج بايثون :

	Hb	S.creatinine	S.Na	S.K	S.PhosPhorus	S.cholesterol	y
0	7.7	4.0	134	4.7	4.8	131	1
1	12.2	3.0	130	4.1	6.6	135	1
2	10.6	4.0	135	4.6	5.4	152	1
3	9.4	4.0	132	4.3	5.8	111	1
4	6.7	3.0	145	5.2	4.3	115	1

```
<class 'pandas.core.frame.DataFrame'>
```

```
RangeIndex: 260 entries, 0 to 259
```

```
Data columns (total 7 columns):
```

```

Hb                260 non-null float64
S.creatinine      260 non-null float64
S.Na              260 non-null int64
S.K               260 non-null float64
S.PhosPhorus      260 non-null float64
S.cholesterol     260 non-null int64
y                 260 non-null int64

```

```
dtypes: float64(4), int64(3)
```

```
memory usage: 14.3 KB
```

```
None
```

```
1      130
```

```
0      130
```

```
Name: y, dtype: int64
```

```
manual_scaler
```

```
accuracy_score: 0.9615
```

```
classification_report:
```

	precision	recall	f1-score	support
0	0.95	0.97	0.96	39
1	0.97	0.95	0.96	39
micro avg	0.96	0.96	0.96	78

macro avg	0.96	0.96	0.96	78
weighted avg	0.96	0.96	0.96	78

confusion_matrix:

```
[[38  1]
 [ 2 37]]
```

	feature	importance
1	S.creatinine	2.010524
0	Hb	1.938159
4	S.PhosPhorus	1.144592
5	S.cholesterol	0.699010
2	S.Na	0.260441
3	S.K	0.255082