المجلة الدولية للهندسة وتقنية المعلومات
journal homepage:www.ijeit.misuratau.edu.ly

توقع تخصص الطالب الجامعي من بياناته باستخدام خوارزميات تعلم الآلة

عبدالله علي المسلاتي
كلية التقنية الصناعية - مصراتة
elmissallati@gmail.com
abdallah.emasallati@cit.edu.ly

إسحاق يوسف الأرباح
كلية التقنية الصناعية - مصراتة
Issaclarbah@gmail.com

أحمد عبد القادر جحا
كلية التقنية الصناعية - مصراتة
goha_99@yahoo.com

الملخص— **هدفت هذه الدراسة إلى تقييم إمكانية توقع تخصص الطالب في الكلية بناءً على بياناته الأكاديمية باستخدام خوارزميات تعلم الآلة. تم جمع بيانات 570 طالبًا من كلية التقنية الصناعية - مصراتة، وتم استخدام 7 نماذج من نماذج تعلم الآلة للتنبؤ بتخصص الطالب. حققت الشبكة العصبية والآلة الداعمة للمتجهات أعلى دقة في التنبؤ، بينما كانت دقة النماذج الأخرى جيدة إلى أقل من جيدة.**
أوصت الدراسة بجمع المزيد من البيانات من كليات أخرى، واستخدام المزيد من البيانات الشخصية للطلبة، واستخدام المزيد من الميزات لتحسين دقة التنبؤ. كما أظهرت الدراسة أن خوارزميات تعلم الآلة يمكن أن تكون أداة مفيدة للتنبؤ بتخصص الطالب في الكلية، لكن هناك حاجة إلى مزيد من البحث لتحسين دقة التنبؤ.

الكلمات المفتاحية— **تعلم الآلة، التصنيف، التوقع، التنبؤ، البيانات الأكاديمية.**

1. المقدمة

إن اختيار التخصص العلمي المناسب للطالب في الكلية من أصعب عمليات الاختيار، حيث يحدد الاختيار المناسب للتخصص العلمي للطالب النجاح والتقدم في هذا المجال. في العادة، يكون اختيار التخصص بناء على رغبة الطالب وميوله العلمية، وأحيانًا يكون الاختيار بناء على نصيحة من أقرانه الذين سبقوه بالدراسة، وأحيانًا أخرى قليلة يكون الاختيار بناءً على احتياجات سوق العمل.

إن الاعتماد على الرغبة والميول أحيانًا لا يكونا ذوي جدوى، خاصة عندما يكتشف الطالب أن أدائه ضعيف في التخصص بعد فوات الأوان، وهذا ما يجعله يغير تخصصه أو يترك الدراسة. إن الاعتماد على البيانات الأكاديمية للطالب، خاصة في فصول الدراسة الأولى، قد يساهم في اتخاذ قرار مستنير لاختيار تخصص علمي مناسب.

في إسبانيا، يترك طالب واحد من بين 3 طلبة الدراسة الجامعية وفقًا لدراسة إسبانية [1]، وهذا ما يعد خسارة فادحة للنظام التعليمي في إسبانيا. في ليبيا، لا توجد إحصائيات رسمية لعدد الذين يغادرون مقاعد الدراسة دون أن يحصلوا على شهادة تخرج من الكلية.

لقد أصبح الذكاء الاصطناعي مساهمًا فعالًا في جميع مناحي الحياة، ويمكن استخدام تقنياته والاستفادة منها في شتى المجالات. ووفقًا لدراسة كريستوفر روميرو وآخرون [2]، شهد مجال التنقيب التعليمي (EDM) وتحليلات التعلم (LA) نموًا سريعًا خلال العقد الماضي، ويتناول هذا المجال تطبيق تقنيات استخراج البيانات على البيانات التعليمية بهدف تحسين وفهم عملية التعلم، وذلك من خلال تحليل البيانات وتحويلها إلى رؤى قابلة للتطبيق تفيد الطلاب والمعلمين والإداريين. في هذه الورقة، سنستخدم تقنيات تعلم الآلة، وهي فرع من فروع الذكاء الاصطناعي، لتوقع تخصص الطالب الذي يفضل أن يلتحق به بناءً على بياناته الأكاديمية. سنقوم بإنشاء نماذج تعلم آلة مختلفة (باستخدام خوارزميات مختلفة) والمقارنة بين نتائجها. وقد اعتمدت على بيانات طلبة كلية التقنية الصناعية - مصراتة في تدريب هذه النماذج.

تناولت هذه الورقة استخدام خوارزميات تعلم الآلة المختلفة، والمعتمدة على الإشراف، لاستكشاف ما إذا كان هناك نمط معين يتبعه الطلاب في اختيار تخصصهم العلمي بناءً على بعض من بياناتهم الأكاديمية والشخصية.

استلمت الورقة بالكامل في 12 ابريل 2024 وروجعت في 02 مايو 2024
وقبلت للنشر في 15 مايو 2024
ونشرت ومتاحة على الشبكة العنكبوتية في 08 أغسطس 2024.

في حالة وجود ذلك النمط أو العلاقة، فإنه بإمكاننا التنبؤ أو المساعدة في التنبؤ بتخصص الطالب العلمي المناسب من خلال تدريبها على تلك البيانات.

2. الدراسات السابقة

شكلت القوة الحاسوبية المتاحة حاليًا دفعة قوية للذكاء الاصطناعي وخوارزمياته المختلفة. ولذلك، ظهرت العديد من الدراسات والبحوث التي وظفت هذه القوة في مجال التعليم، مستخدمةً البيانات الدراسية للطلبة. ركزت أغلب هذه الدراسات على التنبؤ بالأداء العلمي للطالب أو توقع الطلبة الذين سيتركون مقاعد الدراسة بناءً على أدائهم.

أثبتت ليا بي ماكفادين إمكانية استخدام بيانات نظام إدارة التعلم (LMS) للتنبؤ بأداء الطلبة الأكاديمي وتقديم الدعم المبكر لمن يحتاجه [3]. أكد بحثها ما اقترحته دراسات سابقة، وزادت عليه من خلال تحليل بيانات مشروع بحثي دولي. قام فرشيد المربوطي وآخرون بالاستفادة من البيانات الأكاديمية إلى جانب نتائج الاختبارات والمشاركة في الأنشطة التي تم الحصول عليها من نظام إدارة التعلم بجامعة الغرب الأوسط الأمريكية لإنشاء نموذج تنبؤي يبدأ العمل من الأسبوع الخامس من الفصل الدراسي الأول[4] .

نشر أنطونيو خيسوس وآخرون، في دراستهم ملخصًا لدراسات مختلفة، ومقارنة فيما بينها من حيث استخدام البيانات من أنظمة إدارة التعلم عبر الإنترنت (LMS)، واستخدام البيانات الشخصية، واستخدام البيانات الأكاديمية، والفترة التي استهدفت بياناتها للتنبؤ [5]. يوضح (الجدول1) هذا الملخص:

جدول 1. مقارنة بين الدراسات المختلفة التي قامت بنمذجة أنظمة ذكية لتوقع أداء الطلبة خلال الفصل الدراسي حسب دراسة أنطونيو خيسوس [5].

العمل البحثي	بيانات LMS	بيانات شخصية	البيانات الأكاديمية	توقيت التنبؤ
ماكفادين وآخرون. (2010)	✓			التسجيل والسنة الأولى
المربوطي وآخرون. (2016)	✓		✓	الأسبوع الخامس من الفصل الدراسي الأول
جراي وبيركنز (2019)	✓		✓	من الأسبوع 3 إلى الأسبوع 12 من السنة الأولى
ميجويس وآخرون. (2018)		✓	✓	في نهاية السنة الأولى
تشن وآخرون. (2020)	✓	✓	✓	من الشهر 1 إلى الشهر 6 من الفصل الدراسي الأول
هلال وآخرون. (2018)	✓	✓	✓	التسجيل والسنة الأولى
منجوش (2020)			✓	التسجيل
هوفيت وشينز (2017)			✓	التسجيل
فرناندیز غارسیا وآخرون. (2020)			✓	في نهاية السنة الأولى
اقتراحنا			✓	التسجيل وفي نهاية كل فصل من الفصول الأربعة الأولى

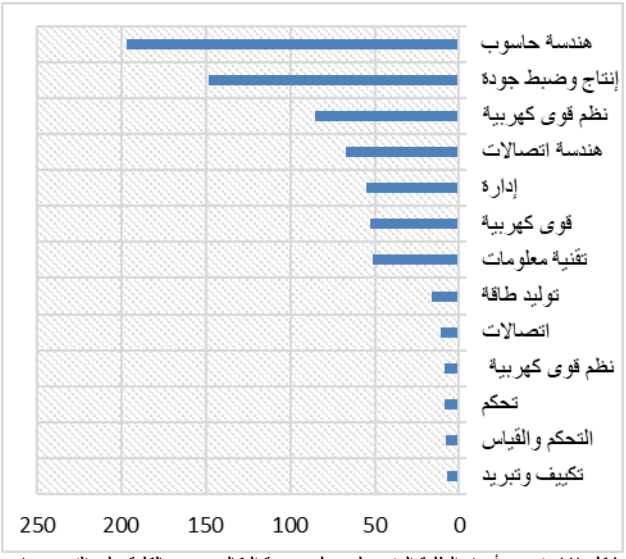
يلاحظ أن أغلب الدراسات اعتمدت على البيانات الأكاديمية للطالب، بغض النظر أن بعضها اعتمدت على مزيج منها. وفي هذه الورقة اعتمدنا على

بيانات الطالب الأكاديمية وميزة واحدة من البيانات الشخصية تمثلت في جنس الطالب، وكان أغلب البيانات مأخوذة في الفصلين الأولين للدراسة.

3. الحالة الدراسية المستخدمة

تم اختيار كلية التقنية الصناعية - مصراتة كحالة دراسية في هذه الدراسة. تم جمع البيانات من منظومة الدراسة والتسجيل بالكلية. تأسست الكلية عام 1989م تحت مسمى "المعهد العالي للصناعة"، كانت تتبع في بادئ الأمر اللجنة الشعبية للصناعة والمعادن (سابقا)، ثم نقلت تبعيتها إلى اللجنة الشعبية العامة للتعليم العالي (سابقاً)، ثم إلى وزارة التعليم التقني والفني، ،اخيراً أصبحت تبعيتها لوزارة الصناعة والمعادن.

حتى فصل ربيع 2023/2022، تم تسجيل (3228) طالبًا في الكلية. تخرج منهم (716) طالبًا بدرجة بكالوريوس، و(289) طالبًا بدرجة دبلوم عالٍ، ليصبح إجمالي عدد الخريجين 1005 خريجين من أصل 3228 طالبًا وتُمثل هذه النسبة 31% تقريبًا من إجمالي عدد الطلاب المسجلين بالكلية منذ افتتاحها، وبعبارة أخرى، من بين كل 10 طلاب سجلوا في الكلية، لم يتمكن 6 طلاب من إكمال الدراسة لأسباب مختلفة، ويُعد هذا العدد كبيرًا، وربما يرجع أحد أسبابه إلى عدم اختيار التخصص العلمي المناسب. دُرس في الكلية 13 تخصصاً مختلفاً، لم يعد يدرس منها الآن سوى تخصص: هندسة الاتصالات، وهندسة الحاسوب، وتقنية المعلومات، وهندسة الكهروميكانيكا (نظم القوى الكهربية والتحكم)، الهندسة الصناعية. ويبين (الشكل 1) توزيع الطلاب الخريجين بدرجة البكالوريوس على التخصصات العلمية في الكلية.



شكل (1). توزيع أعداد الطلبة المتحصلين على درجة البكالوريوس بالكلية على التخصصات

4. الأدوات المستخدمة

تم استخدام برنامج أورانج (Orange) الإصدار 3.36 في هذه الورقة لتدريب النماذج المختلفة. أورانج هو حزمة أدوات مفتوحة المصدر لتعلم الآلة والتنقيب في البيانات مبنية على لغة البرمجة بايثون (Python). تم تطوير البرنامج بواسطة مجموعة من الباحثين والمطورين في جامعة ليوبليانا في سلوفينيا [6].

يوفر البرنامج مجموعة واسعة من الميزات، بما في ذلك:

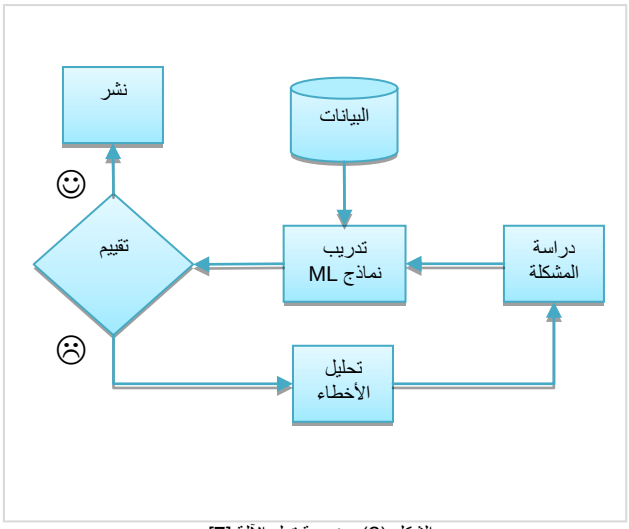
- إدارة البيانات: يوفر البرنامج مجموعة من الأدوات لتحميل البيانات وتحويلها وتنظيفها.
- التحليل الاستكشافي للبيانات: يوفر البرنامج مجموعة من الأدوات لتحليل البيانات واكتشاف الأنماط فيها.
- بناء نماذج التعلم الآلي: يوفر البرنامج مجموعة من النماذج الإحصائية والذكاء الاصطناعي لتعلم الآلة، مثل الخوارزميات التصنيفية والتنبؤية.
- التنقيب في البيانات: يوفر البرنامج مجموعة من الأدوات للتنقيب في البيانات، مثل استخراج قواعد الارتباط واكتشاف الأنماط.

تتكون واجهة المستخدم الرسومية لبرنامج أورانج من بيئة عمل قائمة على المكونات، حيث يمكن للمستخدمين إنشاء مشاريع التحليل باستخدام مجموعة من المكونات الجاهزة. تتميز الواجهة بتصميم بسيط وسهل الاستخدام، مما يجعلها مناسبة للمستخدمين من جميع المستويات. كما يوفر البرنامج دعمًا كاملاً للبرمجة النصية بلغة بايثون، مما يسمح للمستخدمين بإنشاء مشاريع التحليل المخصصة. تتميز البرمجة النصية في أورنج بالمرونة، مما يسمح للمستخدمين بإنشاء مشاريع تحليل معقدة.

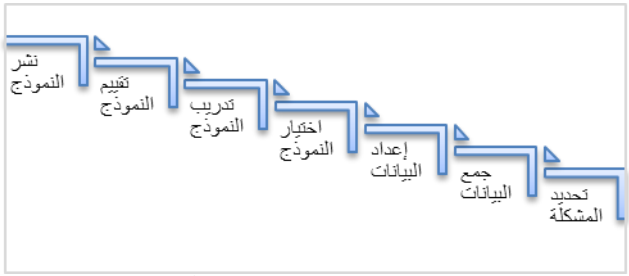
5. المنهجية

أستخدمت المنهجية الشائعة لتعلم الآلة [7]، ويبين (الشكل 2) ماهية هذه المنهجية، وفي العادة تتكون هذه المنهجية من عدة مراحل متتالية كما تظهر في (الشكل 3)، وهي:

1. تحديد المشكلة: تحديد المشكلة التي نريد معالجتها بدقة ووضوح.
2. جمع البيانات: جمع بيانات عالية الجودة تمثل المشكلة.
3. معالجة البيانات: تنظيف البيانات وإزالة الأخطاء وتحويلها إلى تنسيق مناسب وتقسيمها إلى مجموعات.
4. اختيار النموذج: اختيار نموذج تعلم الآلة المناسب للمشكلة.
5. تدريب النموذج: تدريب النموذج على مجموعة التدريب من البيانات.
6. تقييم النموذج: تقييم قدرة النموذج على التعميم على بيانات جديدة.
7. نشر النموذج: نشر النموذج للاستخدام في العالم الحقيقي.



الشكل (2). منهجية تعلم الآلة [7].



الشكل (3). تسلسل مراحل منهجية تعلم الآلة.

أ. تحديد المشكلة:

تحديد ماهية المشكلة التي نريد حلها هي الخطوة الأولى في منهجية تعلم الآلة. في هذه الورقة، نسعى للإجابة على السؤال التالي:

هل يمكننا توقع التخصص العلمي المناسب للطالب في الكلية بناءً على بياناته الأكاديمية والشخصية باستخدام خوارزميات تعلم الآلة؟

والإجابة عن هذا السؤال هي محور هذا البحث. ستُدرب نماذج مختلفة من نماذج تعلم الآلة لتحديد ما إذا كانت هناك علاقة بين اختيار الطالب للتخصص العلمي وبياناته الأكاديمية في كلية التقنية الصناعية – مصراتة كحالة دراسية.

ب. تجميع البيانات:

تم الحصول على البيانات من مصدر رسمي، وهو نظام التسجيل في كلية التقنية الصناعية - مصراتة. تم جمع بيانات السجل التاريخي للطلبة الناجحين الذين حصلوا على شهادة البكالوريوس من الكلية، والذين بلغ عددهم 570 طالبًا، مع ملاحظة أنه أُستبعدت بيانات طلبة الإستدراكي لعدم

فائدتها. تضمنت البيانات تفاصيل المقررات التي درسها الطلبة في كل فصل دراسي، بما في ذلك درجاتهم في هذه المقررات. بالإضافة إلى ذلك، تم استخدام ملف البيانات الشخصية للطلبة. استخرج ملف إكسل من هذين المصدرين يضم البيانات المراد التدريب عليها، وتم تسميته "مقررات عامة للطلبة المتخرجين.xlsx"، وقد رشحت 10 ميزات لاستخدامها في تدريب النماذج، وأغلبها درجات المقررات الدراسية التي درسوها في الفصلين الأولين، وجُل قيمها رقمية (numeric)، وميزتان منهن قيمها فئوية (categorical) وكانت الميزة department القسم أو التخصص هي الميزة الهدف (target) المطلوب توقعها. يوضح (الجدول 2) هذه البيانات ونوعها ووصفها.

جدول 2.الميزات المختارة للتدريب

ت	الميزة	نوعها	قيمتها	دورها	وصفها
1	Gender	فئوية	1، 2	ميزة	الجنس
2	Math1	رقمية		ميزة	رياضة 1
3	Math2	رقمية		ميزة	رياضة 2
4	Computer1	رقمية		ميزة	حاسوب1
5	Computer2	رقمية		ميزة	حاسوب2
6	Physics1	رقمية		ميزة	فيزياء1
7	Physics2	رقمية		ميزة	فيزياء2
8	Numerical-Analysis	رقمية		ميزة	تحليل عددي
9	Average	رقمية		ميزة	المعدل
10	Department	فئوية	Telecom, IT, Electromech-anical, Computer, Industrial	هدف	القسم أوالتخصص

ج. إعداد البيانات:

تُعَدّ مرحلة إعداد البيانات من أهم مراحل نهج تعلم الآلة، فهي تُمثّل الأساس الراسخ لعملية التدريب وضمان تحقيق النتائج المرجوة. أُختيرت 10 ميزات لتدريب النماذج، كما هو موضّح في الجدول (2). كان الهدف من هذه المرحلة هو توقع التخصص أو القسم الذي يتوجب للطالب الإلتحاق به.

تضمنت الخطوة التالية من إعداد البيانات إجراء تحليلات إحصائية دقيقة على الميزات، ويساعد ذلك على فهم نوعية الخصائص وجودتها بشكل أفضل، حيث يُوضّح (الجدول 3) التوزيع التكراري للخصائص، والمتوسطات، وكذلك نسبة البيانات المفقودة، وغيرها من العمليات الإحصائية ذات الصلة.

جدول 3.إحصائيات الميزات المختارة للتدريب

الميزة	التوزيع	المتوسط	المنوال	الوسيط	الثلثات	أدنى قيمة	أعلى قيمة	الفقد
Gender			1		0.37			0 (0 %)
Math1		50	50	50	0.41	0	98	8 (1 %)
Math2		50	50	50	0.44	0	99	0 (0 %)
Computer1		59	50	60	0.34	0	98	2 (0 %)
Computer2		62	50	64	0.31	0	100	2 (0 %)
Physics1		53	50	53	0.34	0	99	7 (1 %)
Physics2		58	50	58	0.27	0	95	4 (1 %)
Numerical-Analysis		55	50	55	0.37	0	100	29 (5 %)
Average		68	61	67	0.10	56	92	0 (0 %)
Department			Computer		1.48			0 (0 %)

يلاحظ من خلال التحليل وجود بعض الخصائص التي تحتوي على بيانات مفقودة، وتمّ معالجة هذه البيانات بتعويضها بالمتوسط الحسابي لتعويض القيم المفقودة [7]، دون التأثير على دقة النتائج.

رُتبت الخصائص بناءً على أهميتها وارتباطها بقيمة الهدف، وساعدنا ذلك على تحديد أكثر الخصائص تأثيراً على عملية التنبؤ، وذلك باستخدام طرق إحصائية متقدمة لحساب انتروبيا المعلومات مثل طريقة إكتساب المعلومات (information gain) [8]، ومعدل المعلومات (information rate) [9]، ومعامل جيني (Gini) [10]. يُوضّح (الجدول 4) ترتيب أهمية الخصائص بالنسبة للميزة الهدف.

جدول 4.تقييم الميزات الأكثر أهمية بالنسبة للهدف باستخدام طرق تقييم مختلفة

الميزة	Info. gain	Gain ratio	Gini	أهميتها
Gender	0.174	0.323	0.031	**1**
NumericalAnalysis	0.179	0.09	0.039	**2**
Math2	0.132	0.066	0.036	**3**
Physics2	0.087	0.044	0.023	**4**
Math1	0.083	0.041	0.028	**5**
Computer1	0.078	0.039	0.026	**6**
Computer2	0.065	0.033	0.022	**7**
Average	0.063	0.032	0.016	**8**
Physics1	0.052	0.026	0.015	**9**

تمّت معالجة الخصائص الفئوية التي تحتوي على قيم نصية، مثل ميزة القسم "Department"، وتمّ تحويل هذه القيم إلى قيم رقمية، وذلك لتوافقها مع متطلبات خوارزميات تعلم الآلة التي تعتمد على العمليات الحسابية.

تمّت معالجة الميزات الرقمية باستخدام طريقة تحجيم الميزات (Feature Scaling) تهدف هذه الطريقة إلى وضع الميزات الرقمية المختلفة في نطاق متشابه، مما يُحسّن من أداء خوارزميات تعلم الآلة ويُقلّل من تأثير الاختلافات الكبيرة في مقاييس الميزات، واعتمدت طريقة التطبيع (Min-max scaling) في هذه الدراسة، نظراً لفعاليتها في معالجة مجموعات البيانات المختلفة.وقد استخدمت جميع الميزات في التدريب ولم تستبعد أي منها.

د. *اختيار النماذج:*

اختيرت في هذه الدراسة 7 من أشهر خوارزميات تعلم الآلة، وهي:

- أشجار القرار (Decision trees): نهج متعدد الاستخدامات لتصنيف وتقدير البيانات المعقدة، مشابهة لآلات دعم المتجهات (SVMs) في إمكانية إجراء مهام التصنيف والانحدار والتنبؤ متعدد المخرجات، تتمتع بقوة كبيرة في معالجة مجموعات البيانات المعقدة، وتشكل اللبنات الأساسية للغابات العشوائية، وهي من أقوى خوارزميات التعلم الآلي المتاحة اليوم[7].
- آلة دعم المتجهات (SVM): نموذج قوي ومرن في تعلم الآلة، قادر على إجراء تصنيف خطي أو غير خطي، وتقدير الانحدار، وكشف القيم المتطرفة، من أكثر النماذج شيوعًا في مجال تعلم الآلة، مناسبة بشكل خاص للتصنيف في مجموعات البيانات المعقدة ولكن الصغيرة أو متوسطة الحجم [7].
- الانحدار اللوجستي (Logistic regression): خوارزمية تصنيف ثنائية لتقدير احتمال انتماء عينة معينة إلى فئة معينة، يُستخدم لتقدير احتمال أن تكون رسالة إلكترونية بريدًا عشوائيًا، على سبيل المثال، يستخدم دالة لوجستية لتحويل قيمة الاحتمال إلى قيمة بين 0 و 1، يمكن استخدامه لتقدير الاحتمالات حتى عندما تكون البيانات غير خطية [7].
- الغابات العشوائية (Random forests): مجموعة من أشجار القرار يتم تدريبها باستخدام طريقة التغليف (bagging) أو أحيانًا طريقة اللصق (pasting)، تُستخدم بشكل شائع في مهام التصنيف والانحدار، أكثر ملاءمة ومحسّنة لأشجار القرار [7].
- الشبكات العصبية الاصطناعية (ANNs): نماذج تعلم آلة مستوحاة من هيكل الدماغ البشري، مرنة وقوية وقابلة للتطوير، ومثالية لحل المهام المعقدة والواسعة مثل تصنيف الصور والتعرّف على الكلام وتوصيات المنتجات، تتكون من مجموعة من العقد تُعرف أيضًا باسم العصبونات، ترتبط ببعضها البعض بوصلات، يتم تدريبها على مجموعة بيانات من الأمثلة لتعلم كيفية ربط المدخلات بالمخرجات، هناك العديد من أنواع الشبكات العصبية المختلفة، بما في ذلك شبكات الإدراك متعدد الطبقات (MLPs) والشبكات العصبية التوليدية (GANs) [7].
- التدرج المعزز (Gradient Boosting): عائلة من خوارزميات تعلم الآلة التي تستخدم طريقة التعزيز المتسلسلة لتحسين أداء نموذج التصنيف أو الانحدار، تبدأ بنموذج بسيط، ثم تضيف نماذج إضافية

واحدة تلو الأخرى، بحيث يتعلم كل نموذج من أخطاء النموذج السابق [11].

- ك أقرب الجيران (k-Nearest Neighbors): من أبسط أساسيات طرق التصنيف في تعلم الآلة، خيار شائع عندما تكون المعرفة المسبقة بتوزيع البيانات محدودة أو معدومة، يعتمد على حساب المسافة بين البيانات الجديدة ونقاط البيانات الموجودة في مجموعة التدريب، يتم تصنيف البيانات الجديدة بناءً على فئة أقرب k من نقاط البيانات [12].

ه. تدريب النموذج:

تم تدريب النماذج المختلفة باستخدام برنامج أورانج، ويتيح هذا البرنامج إجراء التدريب لأكثر من نموذج دفعة واحدة وعلى نفس بيانات التدريب والاختبار، كما يصدر تقييماً لها. ويبين (الشكل 4) الخوارزميات المذكورة قيد التدريب في برنامج أورانج. لقد استخدمت طريقة التحقق المتقاطع (Cross-validation) لتدريب النماذج [13]، وهذه العملية تقسم مجموعة البيانات إلى مجموعات فرعية لاختبار أداء نموذج تعلم الآلة. يتم ذلك عن طريق تقسيم البيانات إلى مجموعات فرعية، تُعرف أحيانًا بالطيات (folds)، حيث يتم تدريب النموذج على بعض الطيات واختباره على الطيات الأخرى. تتلخص الخطوات الاساسية لتنفيذ عملية التحقق المتقاطع: في أولاً: تقسيم البيانات حيث تقسم مجموعة البيانات إلى عدة مجموعات فرعية بنفس الحجم، مثلاً، إذا كانت هناك 10 طيات، يمكن تقسيم البيانات إلى 10 مجموعات فرعية. تليه: التدريب والتقييم حيث يتم تدريب النموذج على طيات معينة واختباره على الطيات الأخرى، ويتم تكرار هذه العملية حتى يتم تدريب النموذج على جميع الطيات واختباره على جميع الطيات. أخيرا: تقييم الأداء حيث يتم حساب متوسط أداء النموذج على جميع الاختبارات التي تمت عبر الطيات.

و. تقييم النموذج:

وفي هذه المرحلة يتم تقييم جودة النموذج المدرب، وبالاستعانة بمصفوفة الإرتباك (confusion matrix) يمكننا رصد عدد قيم التوقعات الصحيحة من قيم الحقيقية للنماذج المستخدمة، وتبين (الأشكال 5-أ، 5-ب، 5-ج، 5-د، 5-ه، 5-و، 5-ز) مصفوفات الإرتباك لجميع الخوارزميات المستخدمة. لقد استخدمت العديد من المقاييس لتقييم أداء النماذج والمقارنة بينها، وهذه المقاييس هي:

- دقة التصنيف (accuracy): وهي نسبة التوقعات الصحيحة إلى إجمالي التوقعات، وتُعطي قيمة MCC رقمًا بين -1 و 1، وكلما اقتربت القيمة من 1 كان ذلك أفضل. والمعادلة (1) تبين كيفية حسابها [14].
- الدقة (Precision): أو (القيمة التنبؤية الإيجابية) وهي نسبة التوقعات الإيجابية الصحيحة إلى إجمالي التوقعات الإيجابية. والمعادلة (2) تبين كيفية حسابها [15].
- الاستدعاء (Recall): أو (الحساسية) وهي نسبة الحالات الإيجابية التي تم التنبؤ بها بشكل صحيح إلى إجمالي الحالات الإيجابية. والمعادلة (3) تبين طريقة حسابها [15].
- المعامل F1: يمثل مقياس توازن بين الدقة والاستدعاء. والمعادلة (4) تبين كيفية حساب هذا المقياس [16].
- معامل ارتباط ماثيو (MCC): وهو مقياس للقوة والاتفاق بين التنبؤات والقيم الحقيقية. والمعادلة (5) تبين كيفية حسابه [17].

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (1)$$

$$Precision = \frac{tp}{tp + fp} \quad (2)$$

$$Recall = \frac{tp}{tp + fn} \quad (3)$$

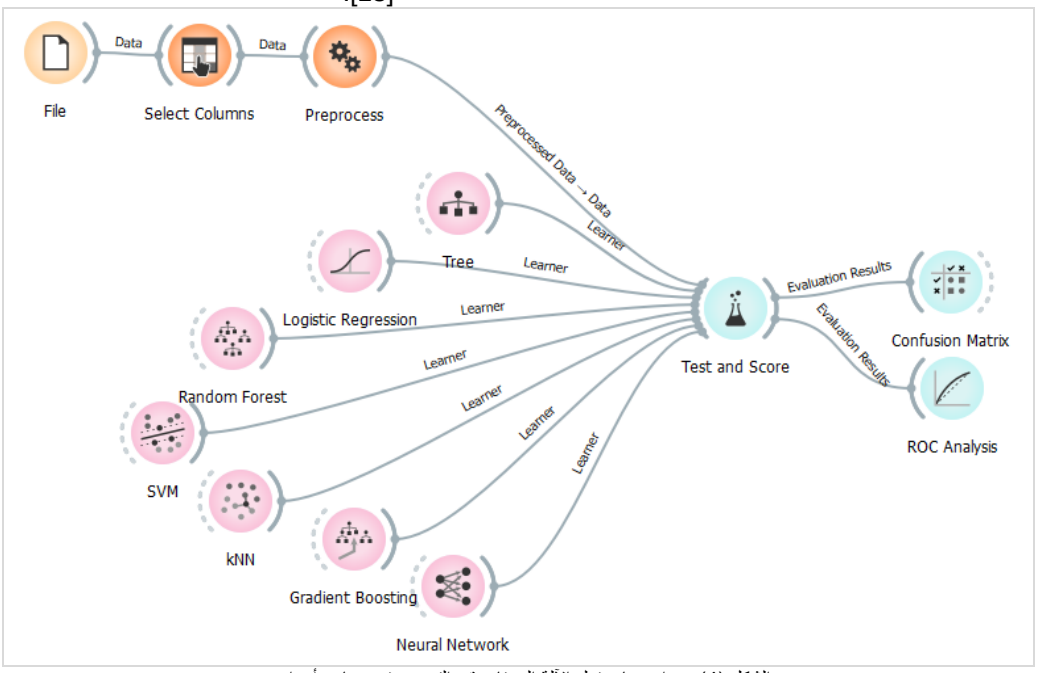
$$F1 = \frac{2tp}{2tp + fp + fn} \quad (4)$$

$$MCC = \frac{tp \cdot tn - fp \cdot fn}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \quad (5)$$

حيث:

tp عدد *التوقعات الصحيحة الإيجابية*
fp عدد *التوقعات الخاطئة الإيجابية*
tn عدد *التوقعات الصحيحة السلبية*
fn عدد *التوقعات الخاطئة السلبية*

- المساحة تحت منحنى (ROC): أو (AUC) وهو مقياس لأداء نموذج التصنيف الثنائي، و ROC هو منحنى يمثل العلاقة بين معدل اكتشاف الحالات الإيجابية (TPR) ومعدل الإنذار الخاطئ (FPR) لنموذج التصنيف، وكلما زادت مساحة AUC، كان ذلك أفضل، وتشير القيمة 1 أن النموذج يمكنه التمييز بشكل مثالي بين الحالات الإيجابية والسالبة، وبينما تشير القيمة 0.5 إلى أن نموذج التصنيف لا يمكنه التمييز بين الحالات الإيجابية والحالات السلبية بشكل أفضل من الصدفة [18].



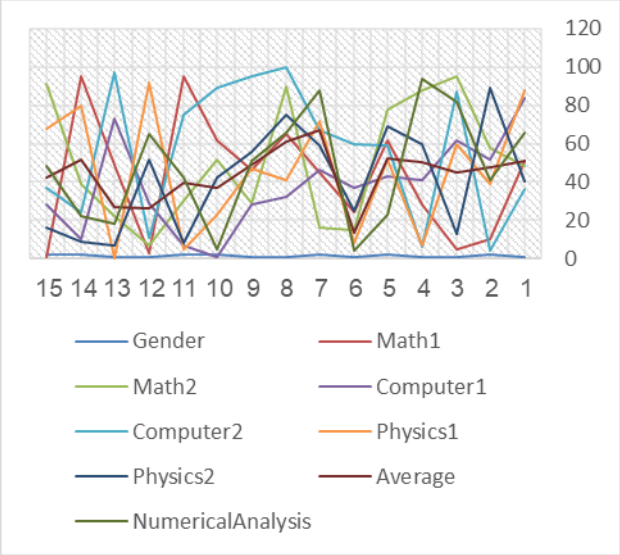
الشكل (4). خوارزميات تعلم الآلة المختارة قيد التدريب في برنامج أورانج.

على بيانات طلبة عشوائية كما تظهر في (الشكل 6)، يمكن ملاحظة توقع النماذج لتخصصات الطلبة المقترحة من قبل النماذج المختلفة كما هي مبينة في (الجدول 5)، أكثر التوقعات تسير نحو قسم الحاسوب كما قلنا آنفا وأقلها إلى تخصص تقنية المعلومات، ربما بسبب العدد الكبير لبيانات تخصص الحاسوب مقارنة بتخصص تقنية المعلومات.

لتحسين عمل النماذج، نوصي باستخدم بيانات أكثر ومن كليات أخرى كحالات دراسية، وذات جودة أفضل، كذلك استخدام المزيد من البيانات الشخصية للطلبة، وأيضا استخدام عدد أكبر من الميزات، ونوصي كذلك بالتركيز على نموذجين أو ثلاثة وتحسين معاملاتها للحصول على أفضل النتائج.

جدول 5.نتائج تقييم لخوارزميات تعلم الآلة المختلفة

MCC	Recall	Prec	F1	**CA**	AUC	النموذج
0.222	0.414	0.413	0.413	0.414	0.631	Tree
0.278	0.468	0.445	0.431	0.468	0.727	Logistic Regression
0.247	0.439	0.422	0.425	0.439	0.696	Random Forest
0.314	0.493	0.437	0.448	0.493	0.733	SVM
0.221	0.419	0.408	0.407	0.419	0.665	kNN
0.273	0.460	0.420	0.435	0.460	0.742	Neural Network
0.303	0.479	0.465	0.469	0.479	0.727	Gradient Boosting



شكل(6). بيانات عشوائية لـ 15 طالباً

جدول 6.توقعات النماذج المختلفة للتخصصات للـ 15 طالباً

ت	شجرة القرار	الإنحدار اللوجستي	الغابات العشوائية	الآلة الداعمة للمتجهات	k أقرب الجيران	التدرج المعزز	الشبكات العصبية
1	اتصا	حا	كهرو	كهرو	صنا	اتصا	حا
2	حا	حا	حا	حا	صنا	صنا	حا
3	كهرو	حا	كهرو	كهرو	كهرو	اتصا	حا
4	كهرو	اتصا	حا	كهرو	كهرو	اتصا	كهرو
5	حا	حا	حا	حا	حا	حا	حا
6	صنا	حا	صنا	كهرو	صنا	صنا	حا
7	اتصا	ت.م	حا	ت.م	كهرو	حا	حا
8	حا	حا	اتصا	حا	حا	حا	حا
9	صنا	حا	حا	حا	صنا	صنا	حا
10	حا	حا	حا	كهرو	كهرو	اتصا	حا
11	اتصا	حا	حا	كهرو	صنا	اتصا	اتصا
12	كهرو	حا	صنا	كهرو	صنا	صنا	كهرو
13	صنا	حا	حا	كهرو	صنا	حا	حا
14	ت.م	ت.م	صنا	حا	صنا	اتصا	اتصا
15	حا	حا	حا	حا	حا	صنا	حا

9. المراجع

Actual \ Predicted	Computer	Electromechanical	IT	Industrial	Telecom	∑
Computer	85	30	1	33	27	176
Electromechanical	33	42	2	32	19	128
IT	4	0	25	2	1	32
Industrial	39	36	1	70	12	158
Telecom	31	14	1	16	14	76
∑	192	122	30	153	73	570

شكل(5-أ). مصفوفة الإرتباك لشجرة القرار

Actual \ Predicted	Computer	Electromechanical	IT	Industrial	Telecom	∑
Computer	114	20	4	34	4	176
Electromechanical	46	35	1	45	1	128
IT	13	1	15	3	0	32
Industrial	32	23	2	101	0	158
Telecom	52	11	0	11	2	76
∑	257	90	22	194	7	570

شكل(5-ب). مصفوفة الإرتباك للإنحدار اللوجستي

Actual \ Predicted	Computer	Electromechanical	IT	Industrial	Telecom	∑
Computer	90	23	6	44	13	176
Electromechanical	38	39	1	39	11	128
IT	4	1	23	3	1	32
Industrial	39	26	1	88	4	158
Telecom	33	16	2	15	10	76
∑	204	105	33	189	39	570

شكل(5-ج). مصفوفة الإرتباك للغابة العشوائية

Actual \ Predicted	Computer	Electromechanical	IT	Industrial	Telecom	∑
Computer	125	10	4	35	2	176
Electromechanical	48	35	1	42	2	128
IT	10	1	18	3	0	32
Industrial	37	17	0	103	1	158
Telecom	52	12	0	12	0	76
∑	272	75	23	195	5	570

شكل(5-د). مصفوفة الإرتباك لآلة دعم المتجهات

Actual \ Predicted	Computer	Electromechanical	IT	Industrial	Telecom	∑
Computer	100	25	8	29	14	176
Electromechanical	45	41	1	30	11	128
IT	7	0	21	3	1	32
Industrial	53	30	1	68	6	158
Telecom	37	18	2	10	9	76
∑	242	114	33	140	41	570

شكل(5-هـ). مصفوفة الإرتباك لـ k أقرب الجيران

Actual \ Predicted	Computer	Electromechanical	IT	Industrial	Telecom	∑
Computer	101	22	6	36	11	176
Electromechanical	41	43	1	37	6	128
IT	5	1	21	4	1	32
Industrial	28	30	2	95	3	158
Telecom	39	20	2	13	2	76
∑	214	116	32	185	23	570

شكل(5-و). مصفوفة الإرتباك للشبكة العصبية

Actual \ Predicted	Computer	Electromechanical	IT	Industrial	Telecom	∑
Computer	93	20	2	43	18	176
Electromechanical	27	54	1	35	11	128
IT	0	4	24	4	0	32
Industrial	37	22	1	93	5	158
Telecom	39	14	2	10	11	76
∑	196	114	30	185	45	570

شكل(5-ز). مصفوفة الإرتباك للتدرج المعزز

لقد حُسبت قيم المعايير السابقة لجميع الخوارزميات المستخدمة في الدراسة، وكانت النتائج كما هي موضحة في (الجدول 6).

8. الاستنتاجات والتوصيات

تم تقييم النماذج باستخدام مجموعة متنوعة من المقاييس، بما في ذلك AUCو CA و F1 و Precision و Recall وMCC ، وحققت الشبكة العصبية والآلة الداعمة للمتجهات أعلى القيم فيما يخص المقياس AUC، وثالث أعلى تقييم كان للتدرج المعزز، بشكل العام تعتبر دقة النماذج جيدة إلى أقل من الجيد في التعلم من البيانات المقدمة، وبالتالي يمكن القول أن النماذج يمكنها توقع تخصص الطالب بدرجة معقولة، وربما يعزى ذلك إلى قلة البيانات المقدمة، حيث كانت البيانات في تخصص تقنية المعلومات قليلة مقارنة بغيرها، وهذا ما يجعل النماذج تنحاز أكثر لاختيار تخصص الحاسوب، أو أنها لا تتبع نمط واضح، وبتجربة النماذج المدربة

[1] F. Pérez García, J. Aldás-Manzano, and I. Zaera, *U-Ranking 2019: Indicadores sintéticos de las universidades españolas. 7ª edición*, 7th ed. ES: Fundación BBVA ; Ivie, 2019. Accessed: Jan. 01, 2024. [Online]. Available: https://doi.org/10.12842/RANKINGS_SP_ISSUE_2019

[2] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *WIREs Data Min. Knowl. Discov.*, vol. 10, no. 3, p. e1355, May 2020, doi: 10.1002/widm.1355.

[3] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an 'early warning system' for educators: A proof of concept," *Comput. Educ.*, vol. 54, no. 2, pp. 588–599, Feb. 2010, doi: 10.1016/j.compedu.2009.09.008.

[4] F. Marbouti, H. A. Diefes-Dux, and K. Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," *Comput. Educ.*, vol. 103, pp. 1–15, Dec. 2016, doi: 10.1016/j.compedu.2016.09.005.

[5] A. J. Fernandez-Garcia, J. C. Preciado, F. Melchor, R. Rodriguez-Echeverria, J. M. Conejero, and F. Sanchez-Figueroa, "A Real-Life Machine Learning Experience for Predicting University Dropout at Different Stages Using Academic Data," *IEEE Access*, vol. 9, pp. 133076–133090, 2021, doi: 10.1109/ACCESS.2021.3115851.

[6] J. Demsˇar *et al.*, "Orange: Data Mining Toolbox in Python".

[7] G. Aurélien, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed. O'Reilly Media, Inc, 2019. [Online]. Available: https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/

[8] I. Csiszar, "I-Divergence Geometry of Probability Distributions and Minimization Problems," *Ann. Probab.*, vol. 3, no. 1, pp. 146–158, Feb. 1975, doi: 10.1214/aop/1176996454.

[9] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[10] R. Dorfman, "A Formula for the Gini Coefficient," *Rev. Econ. Stat.*, vol. 61, no. 1, pp. 146–149, 1979, doi: 10.2307/1924845.

[11] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobotics*, vol. 7, 2013, doi: 10.3389/fnbot.2013.00021.

[12] Leif E. Peterson, "K-nearest neighbor," *scholarpedia*. 2009. [Online]. Available: http://scholarpedia.org/article/K-nearest_neighbor

[13] C. Schaffer, "Selecting a classification method by cross-validation," *Mach. Learn.*, vol. 13, no. 1, pp. 135–143, Oct. 1993, doi: 10.1007/BF00993106.

[14] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the Effect of Accuracy on Trust in Machine Learning Models," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–12. doi: 10.1145/3290605.3300509.

[15] D. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation".

[16] Y. Sasaki, "The truth of the F-measure".

[17] B. W. Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme," *Biochim. Biophys. Acta BBA - Protein Struct.*, vol. 405, no. 2, pp. 442–451, Oct. 1975, doi: 10.1016/0005-2795(75)90109-9.

[18] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.