الفصل الأول

المقدمة

المقدمة Introduction

التخصص الجامعي يعني تحديد مسار لموضوع محدد في اي علم من علوم الحياة، حيث يمكن للطلاب التخصص فيه او اللجوء اليه أثناء التطلع الى الحصول على درجة جامعيه. عادة ما تكون المدارك العلمية التي اكتسبها الطالب في الفترة التي تسبق الدراسة الجامعية هي الاساس التي يمكن ان تساعد في الاختيار السليم للتخصص، ومن خلال اكمال تخصصك، فان ذلك ينمي الرغبة المستوحاة والتميز في المجال الذي اختاره الطالب.

أحيانا يعتبر اختيار مجال التخصص الدراسي قرارا صعبا، في الواقع، حيث أغلب خريجي الجامعات سيقومون بتغيير تخصصهم اذا اتاحت لهم الفرصة للعودة الى مرحلة اختيار التخصص الجامعي. العديد من الجامعات تقدم المئات من الخيارات، وهو تحديا لاختيار خيار واحد عندما تشعر وكأن بقية حياتك المهنية سيعتمد على هذا الخيار، وانه التزام كبير لمن لا يريد تضييع اربع سنين من عمره في دراسة تخصص لا يرغب به.

بسبب توفر العديد من التخصصات والمجالات اصبح من الصعب بل اصبح كالتحدي للطلبة اختيار التخصص المناسب لهم، لذلك جاءت فكرة الحاجة الى بناء نظام معين يقوم بالتنبؤ بالتخصص المناسب، على ضوء ذلك الكثير من الجامعات او مؤسسات التعليم العالي قامت بتحديث معايير قبولها للطلبة والبدء بإدخال تقنيات التنبؤ عن مجال التخصص.

نظرا لهذه التحديات او المشاكل التي يواجها الطالب ادت الحاجة الى استخدام تقنيات أو أنظمة اي بمعنى شامل نظريات مختلفة لتوفير فهم شامل لاختيار الطالب التخصص المناسب له، ستعتمد هذه الدراسة على استخدام خوارزمية شجرة القرارDecision Tree Algorithm (DTA)  لتوجيه الطلاب للمجال الجامعي المناسب والمقارنة بينهما والاعتماد على الافضل.

**مشكلة البحث Research problem**

بسبب التطور الهائل في التعليم الجامعي وازدياد رغبة العديد من الطلبة لدخول الجامعات، الا ان بعض طلبة الجامعات تواجههم العديد من المشاكل منها:

1. الانجرار وراء الاهل والاصدقاء عند خطوة اختيار التخصص الجامعي.
2. عدم وضع المؤهلات وسوق العمل في عين الاعتبار عند اختيار التخصص الجامعي.

**1أهداف البحث Research Objective**

من خلال تساؤلات البحث تهدف الدراسة المقترحة لتحقيق الاتي:

1. بناء نموذج ذكي لتنبؤ الطلبة باختيار التخصص الجامعي المناسب.
2. تطبيق خوارزمية شجرة القرار في التنبؤ.

**حدود البحث Proposed scope**

قمنا في هذه الدراسة بتجميع البيانات وتطبيق تقنيات التنقيب وخوارزميات الذكاء الاصطناعي على قاعدة البيانات الخاصة بالطلبة المسجلين في جامعة سبها.

**استراتيجية البحث**:

منهجية البحث العلمي خطوة من الخطوات المهمة التي يبنى عليها البحث العلمي، ويرجع السبب في هذا الامر الى انه من اهم ضوابط البحث العلمي ان يكون منظما ودقيق، بحيث يستفيد منه كل من يقرأه، أي ان تعتبر الوسيلة او الاسلوب العلمي الذي يتبعه الباحث والذي من خلاله سيجد الحل الجدير لمشكلة بحثه.

تستخدم الدراسة المقترحة المنهجية الاساسية للتنقيب عن البيانات(Data mining) ، بحيث تهدف المنهجية الى استخراج علاقات جديدة غير متوقعة من مجموعة بيانات، بعض الباحثون استخدموا هذه المنهجية في دراستهم، ولكن في تلك الدراسات معظمها استخدمت الخطوات بشكل كامل والاخرين استخدموا بعض تلك الخطوات، وتندرج خطوات هذه المنهجية:

1

3

4

5

2

**تجميع البيانات**

**معالجة البيانات**

**التصميم**

**تنفيذ خوارزمية شجرة القرار**

**الاختبار**

شكل (1) يوضح منهجية data mining

كما في الشكل (1) فان خطوات منهجية Data mining تتكون من الأتي:

1. **تجميع البيانات**: يتم تجميع البيانات الغير مهيكلة من مصادر عديدة وجمعها في مخزن بيانات واحد باستخدام تقنيات التنقيب عن البيانات، في هذه الدراسة سنقوم بأخذ البيانات من قاعدة البيانات الخاصة بطلبة كلية تقنية المعلومات.
2. **معالجة البيانات**: يتم في هذه المرحلة تحديد واسترجاع البيانات الملائمة لعملية التنقيب من مجموعة البيانات المجمعة، حيث يتم عزل البيانات التي تحتوي على شوائب او تشويش او بيانات مفقودة لمعالجتها، سنقوم في الدراسة المقترحة بحذف الأعمدة الغير ضرورية لبناء النموذج التنبئي مثل البيانات الشخصية للطلبة، ثم يتم تغيير البيانات الى شكل ملائم لتطبيق تقنيات التنقيب.
3. **التصميم:** بعد الانتهاء من مرحلة تجميع البيانات ومعالجتها، يتم تصميم الجداول و قواعد البيانات الخاصة ببيانات الطلبة، ايضا تصميم الخوارزمية المستخدمة في الدراسة.
4. **تنفيذ خوارزمية شجرة القرار**: تطبيق التقنيات الوصفية و التنبئية لاستخراج المعارف والأنماط المفيدة والغير مكتشفة سابقا، وسنقوم في الدراسة المقترحة بتطبيق خوارزمية شجرة القرار و خوارزمية التصنيف لملائمتها مع أهداف الدراسة المقترحة.
5. **الاختبار**: وهي المرحلة التي يراها المستفيد، سيتم فيها بناء موقع ويب ونشر النموذج التي تم بنائه على شبكة الانترنت، ليتم استخدامه مع البيانات والحالات المستقبلية.

فوائد البحث

تحتاج الجامعات الى انظمة جيدة ذات كفاءة عالية لتقديم خدماتها للطلبة المستجدين مع امكانية تحقيق كفاءة التنبؤ بالتخصص المناسب وتقليل الارتباك والحيرة عند عملية الاختيار، ومن الممكن تحقيق ذلك باستخدام ما يعرف بتقنيات الذكاء الاصطناعي والتصنيف، حيث تساعد هذه التقنيات في تحسين عملية اتخاد القرار من خلال زيادة نسبة التنبؤ، وتؤدي هذه الزيادة باستخدام تلك التقنيات في زيادة كفاءة القرار.

**الفصل الثاني**

**الإطار النظري**

**المقدمة**

يتطرق هذا الفصل للتعريف بخلفية البحث المتمثلة في مجال الذكاء الاصطناعي والتنقيب عن البيانات بشكل عام والتنقيب عن البيانات في المجال التعليمي وخوارزمياته، وبالأخص خوارزمية شجرة القرار وتقنيات التصنيف وفوائده وبالإضافة الى آلية عمله ومجالات استخدامه، كما سيسلط الضوء على الدراسات السابقة ذات العلاقة بموضوع البحث.

**الذكاء الاصطناعي**

ظهر الذكاء الاصطناعي في ستينيات القرن العشرين وبدا يغزوا مجالات واسعة من حياة الإنسان، امتد استخدامه من تخزين المعلومات وإرسالها، إلى الاعتماد عليه في مجالات أدق وأخص كالصناعة والتجارة وغيرها**.**

والذكاء الاصطناعي جزء من علوم الحاسوب يركز على تطوير أجهزة الحاسوب والتطبيقات المتقدمة بهدف محاكاة الذكاء البشري، في حين تم تطوير الذكاء الاصطناعي لأول مرة في الخمسينيات، نما التطبيق العام للعمليات والقدرات المتعلقة به بشكل كبير في العقود التي تلت ذلك، حيث تعد خوارزميات الذكاء الاصطناعي انها مبنية على مبادئ ومفاهيم الذكاء الاصطناعي، فتتميز هذه الخوارزميات بقدرتها على ابتداع طرق ديناميكية ملائمة لطبيعة المشكلة المراد حلها وتحديد طريقة علمية لإيجاد حل مناسب من بين مجموعة حلول ممكنة للمشكلة كما تتميز ايضا بأمثليه قيمة الحل وفقا للقيود والمتغيرات بأقل كلفة و أكثر استخداما و فعالية، فهذه الخوارزميات تعتبر من الأكثر استخداما في ايجاد الحلول المثالية لمشكلات متعددة في مجالات متنوعة(د. حسن العزيز 2016).

ولخوارزميات الذكاء الاصطناعي مجموعة من الخصائص نوجزها في النقاط التالية (سليمان صالح 2010):

1. استخدام الذكاء في حل المشاكل المعروضة مع غياب المعلومة الكاملة.
2. القدرة على التفكير والادراك.
3. القدرة على اكتساب المعرفة وتطبيقها
4. القدرة على التعلم من التجارب والخبرات السابقة.
5. القدرة على استخدام الخبرات القديمة وتوظيفها في مواقف جديدة.

ويضاف الى تلك الخصائص أن خوارزميات الذكاء الاصطناعي يخلق ألية لحل المشكلات داخل المنضمات تعتمد على الحكم الموضوعي والتقدير الدقيق للحلول، ورفع المستوى المعرفي لمسؤولي المنظمة من خلال تقديمه حلول للعديد من المشاكل التي يصعب تحليلها بواسطة العنصر البشري خلال فترة قصيرة، وبالتالي فان اهم ما يميزه ثباته النسبي، حيث لا يتعرض لما يتعرض له العنصر البشري من عوامل مؤثرة على قدراته كالنسيان.

وهناك مجموعة من الخصائص يتسم بها أي برنامج تعلم وهي: امكانية تمثيل المعرفة، استخدام الاسلوب التجريبي المتفائل، قابلية التعامل مع المعلومات الناقصة، القابلية للاستدلال والاستنتاج، التمثيل الرمزي معالجة اللغة الطبيعية، التعامل مع البيانات المتضاربة والمعلومات الناقصة.

ولقد ادى التطور في توليد وجمع البيانات الى وجود مجموعات بيانات ذات احجام هائلة في مجالات عدة وفي كافة فروع المعرفة العلمية، حيث وجدت المؤسـسـات نفسها غير قادرة على ترجمة وفهم الكم الهائل من البيانات الموجودة في مخازن البيانات ولم تعد وسائل التحليل التقليدية الإحصائية قادرة على التعامل معها، لذا ظهرت العديد من الدراسات منذ أواخر الثمانيات في محاولة لحل تلك المشكلات مع البحث عن حلول تجمع بين عدة تخصصـات سـواء كانت قواعد البيانات والذكاء الاصطناعي والحوسبة التناظرية، وقد تم التوصـل الى التنقيب في البيانات Data mining واكتشاف المعرفة اللذين أثبتا وجوداً كأحد الحلول الناجحة لتحليل كميات ضخمة من البيانات وذلك بتحويلها من بيانات متراكمة وغير مفهومة إلى معلومات قيمة يمكن استغلالها والاستفادة منها بعد ذلك لتصبح معرفة.

لذا يعد التنقيب في البيانات من أسرع المجالات نمواً في تخصـصـات علم الحاسب الآلي، ولقد جاءت شهرته وانتشـاره من الحاجة المتزايدة لأدوات تساعد في تحليل الكميات الهائلة من البيانات وفهمها، وتنتج هذه البيانات يومياً بواسطة المؤسسات المختلفة مثل البنوك، شركات التأمين، المصالح الحكومية، ومستودعات البيع، وعلى شبكة الانترنت، والمؤسسات التعليمية، فهو يعتبر إحدى فروع مجال الذكاء الاصطناعي ويستخدم أيضاً بعض الأساليب الإحصائية لاستخراج المعرفة من البيانات الضخمة.

**مفهوم التنقيب في البيانات:**

هو خطوة أساسية في عملية اكتشاف المعرفة، ولقد ظهر التنقيب في البيانات في أواخر الثمانينات والذي كان مجال حديثا مع قيمة بحثية مرتفعة في دراسة قواعد البيانات والذكاء الاصطناعي وتكنولوجيا قواعد البيانات وتعلم الآلي والاحصائيات وعرض البيانات وغيرها من المجالات النظرية والتكنولوجية.

ويعرف الباحثون تقنيات التنقيب في البيانات بعدة طرق:

* التنقيب في البيانات هو عملية اكتشاف الارتباطات والأنماط والاتجاهات الجديدة المفيدة من خلال التدقيق في كميات البيانات الضخمة باستخدام تقنيات تمييز النماذج بالإضافة الى التقنيات الرياضية والاحصائية (شج).
* التنقيب في البيانات هو استخدام تقنيات متنوعة لتحديد المعلومات أو المعرفة في قاعدة البيانات واستخراجها والاستفادة منها في دعم القرار والتنبؤ والتقدير ().
* ويعرف أيضاً بأنه هو تطبيق أساليب علمية على البيانات في حل المشاكل باستخدام تقنيات وخوارزميات معقدة وذلك للحصول معلومات عالية المستوى.

مما سبق يمكن تعريف التنقيب في البيانات كما يلي:

هو عملية معالجة وتحليل لمجموعة من البيانات غالباً تكون كبيرة، باستخدام وسائل آلية لإكتشاف

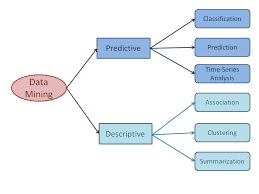
أنماط وسلوكيات لهذه البيانات على شكل نماذج وقواعد تفيد في حل المشاكل ودعم القرارات.

**مهام التنقيب في البيانات (Data Mining Tasks):**

المهام الأساسية للتنقيب في البيانات تنقسم كالتالي (س):

1. **التحليل الوصفي أو الاستكشافي:** يعمل على وصف واستكشاف وتحليل وفهم الخصائص العامة لمجموعة البيانات للوصول لنماذج مفيدة.
2. **التحليل التنبؤي:** يقوم بالاستفادة من البيانات الانية كملاحظات لتكوين نماذج قادرة على التنبؤ.

يوجد نوع ثالث من مهام التنقيب وهو التحليل المدمج والذي يعبر عن مجموعة مدمجة من التحليل الوصفي والتحليل التنبؤي، يركز الهدف التنبؤي على الدقة في المقدرة على التنبؤ، ويركز الهدف الوصفي على فهم البيانات التي تولد المعالجة التنبؤية.



**أنواع التنقيب في البيانات (Data mining method):**

يمكن للعديد من المشاكل المتعلقة بعملية اتخاذ القرارات في أغلب المجالات أن يتم نمذجتها وفق عدة

أساليب رئيسية وهي:

1. **التوصيف Description:** تركز هذه الطريقة على وصف وتشخيص هيكلية وسلوك البيانات فيما بينها بأسلوب مفهوم وواضح، وكمثال على ذلك الفجوة بين الجنسين في السياسة الأميركية حيث نشاهد كيف النساء يدعمون الديمقراطيين بأعداد أكبر من الرجال حيث يمكن هذا أن يثير قدراً كبيراً من الاهتمام ودراسة إضافية من جانب الصحفيين وعلماء الاجتماع والاقتصاد وعلماء السياسة ناهيك عن المرشحين للمناصب.
2. **التقدير Estimation:** تعتبر هذه الطريقة قريبة جداً من طريقة التصنيف لكنها لا تعطى في دالة الهدف كفئات بشكل مطلق بل تتعامل معها كمتغيرات مستمرة ضمن مجالات معينة، وتقوم على تحديد قيمة دالة الهدف الرقمية بشكل تقريبي باستخدام مجموعة من المتغيرات التنبؤية كما في حالة تقدير ضغط الدم للمريض بناء على قراءة بيانات تشخيص حالة المريض.
3. **العنقدة Clustering:** تقوم هذه الطريقة على تجميع البيانات ضمن مجموعات مختلفة تمتلك خصائص متشابهة تسمى عناقيد، وتختلف هذه الطريقة عن التصنيف بعدم وجود متغير هدف أي عدم معرفة هذه المجموعات بشكل مسبق، كما في حالة تجميع العملاء ضمن مجموعات.
4. **قواعد الاقتران:** تهدف هذه النماذج الى اكتشاف العلاقات والترابط بين عناصر مكونة بذلك علاقات وقواعد بين المتغيرات مثل: تحليل سلة السوق واكتشاف الترابط والعلاقات بين انماط طبيعية شراء الزبائن.
5. **التصنيف Classification:** تعتبر هذه الطريقة من أهم الطرق التنقيب في البيانات والتي تقوم على إسناد الكيانات إلى الفئات في دالة الهدف بناء على المتغيرات المستقلة وبالتالي تقوم على توصيف البيانات تبعاً لفئة دالة الهدف، كما يمكن تكييف هذه المنهجية لأداء مهمة التنبؤ بحيث تكون الفئة في دالة الهدف معروفة بشكل مسبق بالنسـبة للبيانات المستقلة كبيانات تنبؤية، كما في حالة تصنيف مخاطر إعطاء القروض للأشخاص هل هو منخفض أو متوسط أو عالي الخطورة. 3

ويعرف التصنيف أيضاً بأنه عملية التعرف على الأفكار والأشياء وفهمها وتجميعها في فئات محددة مسبقًا أو مجموعات فرعية باستخدام مجموعات بيانات التدريب المصنفة مسبقًا تستخدم برامج التعلم الآلي مجموعة متنوعة من الخوارزميات لتصنيف مجموعات البيانات المستقبلية إلى فئات، حيث تستخدم خوارزميات التصنيف في التعلم الآلي لبيانات تدريب الإدخال للتنبؤ باحتمالية أن البيانات اللاحقة ستندرج في إحدى الفئات المحددة مسبقًا، هذه المنهجية التي سنستخدمها في هذه الدراسة لتحقيق أهدافها، ونظراً لملائمتها مع طبيعية بيانات الدراسة بالمقارنة مع بقية منهجيات التنقيب في البيانات، وسنوضح بشكل أكثر استفاضة بعض المفاهيم والخوارزميات المستخدمة في المنهجية في الأقسام القادمة.

**التنقيب في البيانات في المؤسسات التعليمية Educational Data Mining**

ازدهر التنقيب في البيانات التعليمية (EDM) في النظم التعليمية مؤخرا يتيح تحليل أداء الطالب والتنبؤ به حتى يمكن اتخاذ التدابير مقدماً ()

دخلت تكنولوجيا تنقيب البيانات في الكثير من المجالات مثل: التسويق، كشف الاحتيال، الطب، الانتخابات الخدمات المصرفية، المراقبة والاكتشاف العلمي، وغيرة من المجالات.

فالتعليم واحد من أهم المجالات التي يمكن تنقيب البيانات في الكثير من الموضوعات مثل: استطاع رأي الطلاب في مادة معينة، تقيم المحتوى التعليمي، المساعدة في اتخاذ القرار، توقع مستوى الطلاب، مساعدة أو توجيه الطلاب في اختيارهم التخصص الدراسي وهذا النوع هو أنواع التنقيب في البيانات التعليمية الذي تتمحور عليه دراستنا الحالية.

ويعرف التنقيب في البيانات التعليمية بأنه هو العلم الذي يهتم بتطوير الطرائق والمنهجيات التي تساعد في اكتشاف المعرفة من البيانات الخاصة بالنظم التعليمية وذلك بهدف فهم الطلاب بشكل أكبر وفهم الشروط والظروف التي يتعلمون ضمنها ().

وقد تم تطبيق في الاوسط التعليمية للأهداف البحثية مثل تحسين عملية التعلم وتوجيه الطلاب التعلم أو اكتسابهم أعمق للظواهر التعليمية [6]. مثال لهذه التطبيقات في مجالات مثل الذكاء الاصطناعي، التعلم الآلي تحليل السوق، الإحصاء وقاعدة البيانات [15].

الهدف الرئيسي لمؤسسات التعليم العالي هو توفير التعليم الجيد لطلابها والطريق الوحيد الى تحقيق أعلى مستوى من الجودة هو تحديد العوامل التي تؤثر على الأداء الأكاديمي ثم محاولة حلها [18].

توجيه الطالب لاختيار التخصص الأكاديمي المناسب هو أحد تطبيقات التعليم يتطلب العديد من المعايير والتنبؤ الفعال بالتخصص الأكاديمي للطالب، وذلك يتطلب نموذج يشمل جميع الأحوال والبيانات العامة والخاصة بالطلاب مثل: التخصص الثانوي، الجنس، الحالة الاجتماعية، وغيرها، بناء على هذه البيانات وباستخدام تقنيات تنقيب في البيانات بالإمكان مساعدة وتوجيه الطالب لأقرب تخصصه بناء على بياناته.

تحتاج المؤسسات التعليمية إلى معرفة مسبقة ببيانات الطلاب الراغبين في الإلتحاق والتسجيل بها، وذلك للتنبؤ لهم بأفضل تخصص يمكن أن يلتحقون به بناء على بياناتهم ومعلوماتهم السابقة، مما يتيح للمؤسسة التعليمية على توجيه الطلاب بأفضل تخصص لهم.

يتم تخزين حجم متزايد من البيانات حاليا في قواعد البيانات التعليمية التي تحتوي على معلومات مخفية مختلفة يمكن أن تساعد في ارشاد الطلاب وبالتالي يتم استخدام تنقيب في البيانات التعليمية لدراسة البيانات واستخراج المعلومات المخفية للعمليات المستقبلية.

ويمكن استخدام البيانات التعليمية في العديد من العمليات مثل تقدير معدل تسرب الطلاب، الكشف عن القيم غير الطبيعية في نتائج الطلاب، والتنبؤ بأداء الطالب في التمارين والواجبات.

ويمكن استخدام تقنيات التنبؤ لمساعدة المدارس للتنبؤ لدرجات الطلاب في مادة معينة أو متوسط مجموعة المواد، إذا يشير ناتج التنبؤ هذا إلى أن يمكن استخدام تقنيات التنبؤ لمساعدة المدارس للتنبؤ لدرجات الطلاب في مادة معينة أو متوسط مجموعة المواد، إذا يشير ناتج التنبؤ هذا إلى أن الطالب سيحصل على تقدير نجاح أو رسوب في المادة، ثم بذل جهود إضافية يمكن إجراؤه لتحسين الأداء الأكاديمي للطالب في حالة التنبؤ الرسوب.

**للـتنقيب في البيانات التعليمية مميزات وإيجابيات تجعله مفيد أكثر من أساليب البحث التقليدية نذكر منها:**

1. أصبحت البيانات سهلة الوصول من خلال بعض المستودعات العامة للبيانات التعليمية التي تحوي على بيانات لطلاب حقيقيين وليس افتراضيين ومؤسسات تعليمية حقيقيـة والبيانات المتعلقة بـالأداء والظـروف المحيطة بالعملية التعليمية، بالتالي تكسب الباحثين مصداقية البحث والسهولة في إجراء الأبحاث نظراً لأنها بيانات متاحة للعموم.
2. نظراً لوجود بيانات ضخمة لآلاف الطلاب فإن EDM يوفر إمكانية دراسة تأثير العوامل المختلفة في أداء الطلاب والأساتذة وهذه الميزة لم تكن محققة بنفس المقدار من السهولة والسرعة في طرق البحث التقليدية.

يتطلب تنقيب في البيانات التعليمية المعرفة السابقة من مجموعات البيانات المتاحة، يمكن استخدامها لتمثيل وإيصال المعلومات إلى مديري هذه المؤسسات لمراقبة الظروف واتخاذ إجراءات لحل المشاكل.

يستخدم التنقيب عن البيانات في مجال التعليم لحل المشاكل التي يمكن عن تحدث في اداء الطالب الذي يمثل مصدر قلق كبير في المدارس والمعاهد التعليمية قد تؤثر عدة عوامل على الأداء.

**تطبيقات التنقيب في البيانات في المجال التعليمي:**

1. التواصل مع (الأشخاص المعنيين): الهدف هو مساعدة القائمين على المناهج والمعلمين على تحليل نشاطات الطلاب وكيفية استخدامهم للنظام التعليمي أما التقنيات المستخدمة

لتحقيق تلك الغاية فهي التحليل الإحصائي والتقارير.

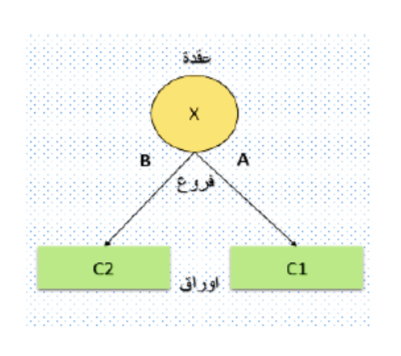
1. تحسين المناهج: الهدف هو مساعدة المعلمين على تحسين المناهج التعليمية من تطوير أو تغيير لمفردات المنهاج تبعاً لمستويات الطلاب وكيفية استخدامهم للنظام التعليمي. والتقنيات المستخدمة هي التصنيف classification والتجميع clustering والتحليل العلائقي association rule.
2. توقع درجات الطلاب: الهدف هو توقع درجات الطلاب في الامتحانات المقبلة ومدى قدرتهم على متابعة التعلم في المستقبل. نستخدم التصنيف والتجميع والتحليل العلائقي كتقنيات لتحقيق هذا الهدف.
3. نمذجة الطلاب: لنمذجة المستخدم في النظم التعليمية عدة فوائد منها اكتشاف حالة الطلاب الحالية... هل هم راضون عن مستواهم أو النظام التعليمي؟، هل هم في حالة من التحفيز؟، هل هناك مشاكل تؤثر على تحصيلهم العلمي؟ (ارتكاب أخطاء كثيرة، عدم الاستفادة من المساعدات المقدمة، محاولة الغش)، كل هذا يؤثر على عملية التعليم واختيارات الطلاب والمعلمين. تُستخدم تقنيات التجميع، التصنيف، التحليل العلائقي، التحليل الإحصائي، شبكات Bayes والنماذج النفسية لتحقيق الهدف المنشود.

**خوارزميات التنقيب في البيانات:**

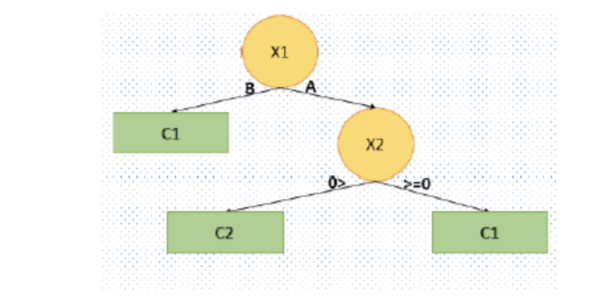
نظرا لأن تجميع البيانات المخزنة في قواعد البيانات التعليمية يتزايد بسرعة، ومن أجل الحصول على الاستفادة المطلوبة من هذه البيانات وإيجاد علاقات خفية بين المتغيرات، نلجأ إلى استخدام طرق التنقيب في البيانات المختلفة، وفي هذا البحث سوف نستخدم خوارزمية شجرة القرار لإنها من أكثر طرق التصنيف استخداماً وذلك بفضل قدرتها على إنشاء نماذج قادرة على شرح وتحليل عملية دعم القرارات المعقدة على شكل مجموعة من القرارات البسيطة قابلة للتفسير بشكل سهل، بالإضافة إلى أنها من الطرق التي يسهل تنفيذها واستخدامها لأنها لا تحتاج إلى تدريبات معقدة للحصول على النتائج لاعتمادها في عملية تقسيم الشجرة على أفضل المتغيرات، ومن ثم سنقوم بمقارنتها بخوارزميات اخرى نايف بايزن، الجار الأقرب وهي على النحو التالي:

**خوارزمية شجرة القرار**

هي إحدى أساليب التصنيف التي تقوم على بناء هيكل شجري لتمثيل القواعد المستخرجة من عملية التصنيف، وتتكون من مجموعة من العقد تمثل نقط اختبار، وتنتج عنها فروع تمثل مسار ناتج الاختبار، تؤدي في نهايتها إلى أوراق تمثل القرار، ويوضح الشكل التالي مكونات شجرة القرار:



وقد لا ترتبط عقد الاختبار مباشـرة بالأوراق عبر الفروع بل ترتبط الفروع الناتجة بعقد اختبار أخرى وفي النهاية تنتهي بأوراق، ويعبر عن العقدة العليا كعقدة أب للعقد الأدنى منها، أما العقد الدنيا فهي عقد أبناء للعقدة الأعلى منها، ويتم بناء شجرة القرار من جدول البيانات الذي يحوي عدة متغيرات مستقلة Xi ومتغير واحد هو دالة الهدف 7، وبذلك قد يكون لكل متغير مستقل عقدة اختبار، كما هو موضح بالشكل:



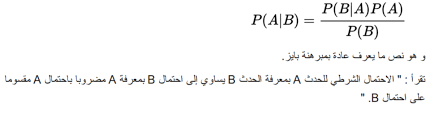
وهذا يعني أن عقد الاختبار في شجرة القرار تمثل الحقول الموجودة في قاعدة البيانات أي المتغيرات المستقلة، أما الفروع التي تخرج من هذه العقد فهي عبارة عن القيمة التي تحويها هذه الحقول، وتكون هذه القيمة هي اختبار العقدة بحيث تقســــم مجموعة البيانات تبعاً لإحدى الفئات في دالة الهدف إلى عدة مجموعات، وفي نهاية الشجرة عند حواف الفروع توجد الأوراق أي الفئات في دالة الهدف، وتقوم الشجرة بالتصنيف عن طريق الترتيب من الأعلى إلى الأسفل وتبدأ من عقدة وحيدة تسمى عقدة الجدر تشكل مسـارات إلى عقد أخرى عبر تقسـيم

البيانات إلى أجزاء بشكل تكراري حتى الوصول إلى الأوراق، مكونة بذلك عدة مسارات ويفسر كل مسار على أنه قاعدة للوصول إلى فئة معينة في دالة الهدف، وبالتالي الوصـول إلى القرار المطلوب وهذا يعني شجرة القرار

تتشكل مما يلي:

1. **عقدة الجذر:** هي عقدة وحيدة تمثل عقدة أب لكل العقد الموجودة في شجرة القرار كاملة، وليس لها أي فروع داخلة وتخرج منها فروع ترتبط إما بعقد دخل لتشـكل عقد أبناء لعقدة الجذر أو ترتبط مباشــرة بورقة، وتعبر عن أهم المتغيرات المستقلة التي تقسم مجموعة البيانات بحسب الفئات.
2. **عقد دخل:** هي مجموعة من العقد الأخرى المتشكلة في الشجرة، وكل عقدة لديها فرع داخل إما من عقد الجذر أو من عقد دخل أخرى تشكل لها عقد أب وفروع خارجة ترتبط بعقد دخل أخرى تشكل عقد أبناء لهذه العقدة أو ترتبط بورقة، وعند أي عقدة من هذه العقد تتشكل شجرة فرعية مع العقد الأبناء والأوراق في المستويات الأدنى المتفرعة عنها، وتعبر هذه العقدة عن المتغيرات المستقلة في مجموعة البيانات.
3. **الأوراق:** تعبر هذه الأوراق عن نهاية المسـار في الشجرة ولديها فرع داخل من عقدة دخل أو من عقدة الجذر مباشرة، ولا تخرج منها أي فروع، وتعبر عن الفئات في دالة الهدف وبالتالي القرار النهائي.

* **خوارزمية ناييف بايزن Bayesian Classifier:** هو طريقة تتطلب كمية أقل من إعداد البيانات لحساب المتغيرات، وهو منظم جيدا للتعامل مع كلا من البيانات الواقعية والمميزة، والاستقلالية الشرطية الطبقية واضحة بين المجموعات الفرعية من المتغيرات في مصنف بايزن، ويوفر نموذج بياني للعلاقة السببية لأداء عملية التعلم، ويعتمد مصنف بايز على قاعدة الاحتمال الشرطي لبايز، وهي تقنية لتقدير احتمال وجود خاصية في ضوء مجموعة بيانات كدليل أو مدخل، ويسمى قاعدة أو نظرية أو مبرهنة بايز.



على سبيل المثال، عندما يتم تصنيف الطلاب في كلية ما في شعبا مختلفة تنسيقا داخليا، شرط الحصول على درجات معينة في بعض المقررات.

* **مصنف الجار الأقرب k-nearest neighbor:**

هي اجدى خوارزميات التعلم الالي الموجه (الخاضع للاشراف)، وتعتبر خوارزمية الجار الاقرب من خوارزميات التصنيف التنبؤية والوصفية وتصنف هذه الخوارزمية البيانات لمعرفة الجار الاقرب بسهولة وفعالة، وتعتمد الخوارزمية في عملها على قياس المسافة الاقليدية بين كل نقطة والنقطة الاقرب اليها، وعندما تكون البيانات قريبة من بعضها تكون المسافة الاقليدية قليلة جدا بين كل نقطة والنقطة المجاورة لها ولكن كلما تباعدت قيم البيانات اصبحت المسافة بين النقاط كبيرة جدا ومن هنا جاء عنوان الخوارزمية، اذ يشير الحرف k الى الحالات التي سيتم تصنيفها بناء على المسافات بينها اي بين جيرانها، وتحسب المسافة بالمعادلة الأتية:

حيث ان:

D: تمثل المسافة بين نقطتين.

Xi: يمثلن نقاط البيانات في فضاء المتجه.

**طرق تقييم الخوارزميات**

من أجل تقييم فعالية نموذج التنبؤ والخوارزميات، يجب أن تحسب القيم المتوقعة مقارنة مع القيم الفعلية في حالة تساوى القيم بنسبة كبيرة يعني عن نموذج التنبؤ جيد، هناك معايير متعددة لفعالية التنبؤ.

يوضح الجدول النتائج المحتملة للتنبؤ بالقيم الثنائية.

|  |  |  |
| --- | --- | --- |
| Predicted as False | Predicted as True |  |
| False Negative | True Positive | Actually True |
| ‏True Negative | False Positive | Actually False |

جدول تقيم فعالية التنبؤ

تسمى المصفوفة التي تعرض نتائج التنبؤ المحتملة بمصفوفة الارتباك () هناك معايير تقييم مختلفة يمكن الحصول عليها من هذه القيم واحد هو الدقة ().

قانون الدقة:

Accuracy

الدقة هي في الأساس نسبة التوقعات الصحيحة سواء عن كانت ايجابية او سلبية ومع ذلك، فإن الدقة لها حدود في تقييم أداء التنبؤ.

على سبيل المثال، مجموعة البيانات التي تحتوي على 100 طالب، 90 منهم اجتازوا الاختبار، ربما يعتبر تنبؤ خام (يعرف باسم قاعدة الأغلبية) لا يستخدم في طرق التعلم الآلي، ولكن بدلاً من ذلك تتوقع أن يمر كل طالب بالامتحان بدقة 90 ٪ يجب أن يكون الأداء أفضل من مجرد تخمين.

في هذه الدراسة، يتم استخدام ثلاثة معايير أخرى، اثنان منهم الدقة والتذكر.

‏ Precision =

Recall =

يتم استخدام الدقة والاستدعاء معاً لإجراء تقييم أفضل.

الفكرة الرئيسية في ذلك هي التنبؤ بدقة النتائج الايجابية ليست كافية نموذج التنبؤ الجيد يجب لديه مزيج من التوقعات الإيجابية الناجحة والسلبية الناجحة.

المعيار الثالث الذي تستخدمه هذه الأطروحة يسمى F-measure ، وهو كذلك تم تعريفه على أنه[49]

‏ F1- measure =

‏ F-measureهو وسيلة للحصول على قيمة واحدة تأخذ الدقة والتذكر فيها الحساب F-measure هو معايير التقييم النهائية للمقارنات في هذه الرسالة.

**الدراسات السابقة:**

وراء ثورة الصحة الرقمية يواجه العديد من الطلاب في الوقت الحاضر مشاكل في تحديد مساراتهم المهنية، بعد تجاوز المرحلة الاْكاديمية ويحتاجون لإدراك لقدراتهم ولتحقق من المجالات التي تلبي اهتمامهم، هناك أيضًا تطورات منهجية باستخدام تقنيات الذكاء الاصطناعي والتعلم الآلي، هو النظام العلمي الذي يستخدم خوارزميات الكمبيوتر للتعلم من البيانات، و للمساعدةً في تحديد الأنماط في البيانات، والتنبؤ تتمثل إحدى الميزات الرئيسية التي تقوم عليها الإثارة وراء الذكاء الاصطناعي والتعلم الآلي في قدرتها على تحليل هياكل البيانات الكبيرة والمعقدة لإنشاء نماذج تنبؤ تقوم بتخصيص وتحسين التشخيص والمراقبة وإدارة العلاجات، بهدف تحسين النتائج الصحية الفردية. توجد نماذج التنبؤ لدعم اتخاذ القرار السريري منذ عقود، وواضعو السياسات ومطورو الإرشادات والمرضى وعامة الناس جميعًا مستخدمون محتملون لنماذج التنبؤ. يتزايد عدد دراسات نماذج التنبؤ بسرعة، مع تطوير مئات النماذج المختلفة لبعض نفس المجموعات السكانية والنتائج المستهدفة، لذلك تم في هذا القسم مناقشة عدة دراسات اجريت بخوارزميات الذكاء الاصطناعي ودورها في اتخاد القرار بغض النظر عمن أجراها او مكان اجرائها او المجال الذي تم الاجراء فيه وهي كالتالي:

في دراسة ل(W. W. T. Fok et al) تم البحث حول كيفية استخدام محرك Tensorflow للذكاء الاصطناعي لتصنيف أداء الطلاب والتنبؤ ببرنامج شهادات جامعاتهم المستقبلية. يعد التنبؤ المناسب والدقيق أمرًا مهمًا لتقديم المشورة الفورية للطالب بشأن اختيار البرنامج والجامعة. للحصول على دراسة أكثر شمولاً لجميع العوامل المستديرة ، حلل نموذج التعلم العميق ليس فقط الأداء الأكاديمي التقليدي بما في ذلك الرياضيات والصينية والإنجليزية والفيزياء والكيمياء والبيولوجيا والتاريخ ، ولكن أيضًا الأداء غير الأكاديمي مثل الخدمة والسلوك والرياضة والفن. يتم تعديل ومقارنة عدد قليل من المعلمات في محرك Tensorflow بما في ذلك عدد العقد الوسيطة وعدد طبقات التعلم العميق. مع مجموعة بيانات مكونة من ألفي طالب ، يتم استخدام 75٪ من هذه البيانات كبيانات تدريبية و 25٪ كبيانات اختبار ، وتراوحت الدقة من 80٪ إلى 91٪. يتم تحديد التكوين الأمثل لنموذج التعلم العميق Tensorflow الذي يحقق أعلى دقة في التنبؤ. حددت هذه الدراسة العوامل المؤثرة في دقة نموذج التنبؤ.

في دراسة ل( R. Reeta, et,al 2018 ) تقترح طريقة لتصنيف مجموعة الجينات المريضة الموجودة في الفرد المصاب بالتوحد. من مجموعة بيانات التدريب، يتنبأ النظام بسلوك التوحد للفرد من خلال مقارنة أوجه التشابه بين جينات الفرد والجينات المصاب في مجموعة التدريب. يتم تنفيذ ذلك باستخدام طريقة تصنيف NAIVE BAYESIAN. على سبيل المثال، إذا تم اختبار الحمض النووي للفرد وإذا كان الحمض النووي يحتوي على الجينات المريضة الموجودة في مجموعة التدريب، يتم التنبؤ بالتوحد. هذا النهج يجعل تشخيص التوحد أسهل وفي مرحلة مبكرة.

ايضا في ورقة (M. C. S. Geetha,et, al 2017) تم تطبيق وتحليل خوارزميات التصنيف الشائعة الاستخدام على مجموعة البيانات الطبية التي تساعد على التنبؤ بأمراض القلب التي تعتبر السبب الرئيسي للوفاة في جميع أنحاء العالم. من المعقد بالنسبة للممارسين الطبيين تصور النوبة القلبية لأنها تتطلب الخبرة والمعرفة. يحتوي قطاع الصحة اليوم على معلومات مخفية ولكنها مهمة لاتخاذ القرارات. كشفت التجارب التي تم إجراؤها أن الخوارزمية مثل J48 و SIMPLE CART و REPTREE توفر دقة تنبؤية أكثر من الخوارزميات الأخرى.

بناء على ذلك دراسة (R. Patil,et, al 2018)حاولت تطوير نظام من شأنه أن يساعد مؤسسة تعليمية على التكوين المسبق لأداء الطلاب من وظائفهم السابقة. من أجل تحقيق ذلك ، تم استخد م مفاهيم تقنيات استخراج البيانات تحت التصنيف. أيضًا من أجل تطوير الحل ، تم إعداد مجموعة بيانات تحتوي على معلومات حول الطلاب من حيث جنسهم وعلاماتهم ورتبهم في امتحانات القبول والنتائج في السنة الثالثة من الدفعة السابقة من الطلاب. تم تحليل مجموعات البيانات هذه لإعداد الحل النهائي. حيث يتم تعيين البيانات في مجموعات أو فئات محددة مسبقًا في تقنية التنقيب في بيانات التصنيف. إنها طريقة تعلم خاضعة للإشراف حيث يلزم إنشاء قواعد لتصنيف بيانات الاختبار إلى مجموعات محددة مسبقًا أو فصولاً تحمل بيانات التدريب. تنقسم العملية إلى مرحلتين. المرحلة الأولى وهي مرحلة التعلم، يتم فحص بيانات التدريب وبدء قواعد التصنيف في هذه المرحلة. المرحلة الثانية هي التصنيف الذي يتم فيه تصنيف بيانات الاختبار إلى فئات وفقًا لمجموعة بيانات التدريب. يتم تحديد الأداء العام والفردي لطلاب السنة الثالثة في الاختبارات المستقبلية باستخدام ID3 ،C 4.5 ، خوارزميات تصنيف ID3 المعدلة الموزونة المحسنة على هذه البيانات.

كما تناولت دراسة ((andrih,et,al,2018 بناء نموذج تنبؤي من مرحلتين باستخدام تقنيات تنقيب البيانات وتم تجميع بيانات الدراسة الخاصة بطلاب الذين اكملوا السنة الاولي من المهنة الاْكاديمية لتنبؤ بأدائهم الاكاديمي الكلي، وبلغ عدد الطلاب الذين جمعت بياناتهم 2459 طالب، و جاءت نسبة دقة النموذج اعلي بنسبة 95% بخوارزمية الغابات العشوائية

واقترحت دراسة اخرى اجراها الباحث (Mohamed Hegazy, Hoda Waguih 2018) بناء نموذج مرشد أكاديمي للطلاب لتعليمهم العالي والذي يستخدم التنقيب عن بيانات التصنيف للتوصية بالتخصص الأكاديمي المناسب. الطريقة المطبقة في البحث هي تقنيات تصنيف استخراج البيانات من خلال طريقة شجرة القرار لتوجيه الطلاب لاختيار التخصص المناسب والمساعدة في تخصيصهم للمسار الصحيح. يصنف النموذج المقترح الطلاب ويطابقهم مع مسارات الدراسة المناسبة وفقًا لخصائصهم.

بما ذلك دراسة اجراها () قامت بتصميم تطبيق ويب لمساعدة الطلاب الذين يواجهون مشكلة على اختيار التخصص المناسب لحياتهم المهنية، للعثور على أداء الطالب بمساعدة التحليلات، حيث سيساعد تطبيق الويب هذا الطالب على اختيار الخيار المناسب بناء على سمة شخصيته واهتمامه وقدرته على الالتحاق بالتخصص وكذلك توفير أفضل الكليات حسب منطقته وهيكل الرسوم.

ايضا تقترح ورقة () حلا لدعم اتخاذ القرار يوفر تحليلا دقيقا ودعما أفضل للقرار وقدرة على إعداد التقارير والتخطيط لمساعدة صانعي القرار من أجل تحسين جودة قراراتهم. لتحقيق هذا الهدف، يتم استخدام مجموعة من التعلم الآلي. يتم إجراء تجارب على بيانات حقيقية تصف كلية علوم وهندسة الكمبيوتر (CCSE) في جامعة طيبة في المملكة العربية السعودية. تظهر النتائج أنه يمكننا التنبؤ بمعدلات التخرج في دراسة حالة حقيقية لدعم عملية صنع القرار. بالإضافة إلى ذلك، تم اجراء مقارنة بين أربع تقنيات للتعلم الآلي وهي آلة ناقلات الدعم و Naïve Bayes و Decision Tree و Random Forest باستخدام الدقة والاستدعاء والقياس.

وفي دراسة ل(Hanan abdullah,2020) ركزت على طرق دعم الجامعات في أتخاد قرارات القبول باستخدام تقنيات التنقيب عن البيانات لتنبؤ بالإداء الاكاديمي للمتقدمين في الجامعة وتم استخدام مجموعة بيانات مكونة من 2039 طالب، واظهرت النتائج انه يمكن توقع الاداء الجامعي المبكر للمتقدمين بناء على معايير معينة مثل(sad)متوسط درجة الدراسة الثانوية ،ودرجة اختبار القبول لإنجاز الدراسي، ودرجة اختبار القدرات العامة)، وتظهر النتائج ان أفضل معيار لتنبؤ بأداء الطلاب في المستقبل هو درجة اختبار القبول لإنجاز الدراسي، واستخدمت الباحثة عدة خوارزميات لتنفيذ النموذج منها تقنية الشبكة العصبية ،شجرة القرار، نايف بايز، الة متجهة الدعم، وكانت تقنية الشبكة العصبية الاصطناعية أعلى معدل بنسبة دقة 79%.

ايضا دراسة أجراها(Karitey,et,al,2020)، تم بناء نظام ذكي يساعد الطلاب على تحديد المسار الوظيفي الأفضل لمستقبلهم، ويساعد ايضا في تحسين أداء الطالب وتحفيز اهتمامه بحيث يجعله يركز على حياته المهنية المستهدفة، ويعتمد هذا النظام على اختبار يجب على الطالب اجراىْه ويعتمد ايضا على الاْجابات التي يقدمها الطالب، واستعمل الباحث خوارزميتان لبناء النموذج لتنبؤ بإداء الطالب وهي خوارزمية شجرة القرار، وألة متجه الدعم.

**المناقشة:**

من خلال مراجعة العديد من الدراسات السابقة التي تصدت لإستخدام خوارزميات الذكاء الاصطناعي لاتخاذ القرارات ومنها دراسة (Karitey,et,al,2020، Mohamed Hegazy, Hoda Waguih 2018، Hanan abdullah,2020)، مع مراجعة العديد من الجامعات، اتضح ان العديد من تلك الجامعات تواجه هذه التحديات من ناحية عدم توفير الأساليب المناسبة لتوجيه الطالب، رغم ذلك العديد من الدراسات السابقة وعلى وجه التحديد التي تم الإشارة لها سابقًا تقترح حلولًا مبتكرة للتعامل مع هذه المشاكل مثل استخدام انظمة ذكية تقوم بالتنبؤ بالتخصص، او تطوير استطلاع لتحديد ما هي مهارات و شخصية الطالب.

فئة قليلة من الطلاب يعلمون الوظيفة المناسبة لهم منذ الصغر ولكن الأغلب لم يعطوا للأمر أهمية في تلك الفترة، على ضوء ذلك تناقش الدراسة المقترحة كيفية توجيه الطلاب للقرار المناسب باستخدام نظام مصمم بخوارزميات الذكاء الاصطناعي، حيث يقوم هذا النظام بجمع البيانات وتحليلها وتصنيفها لبيانات تاريخية وعرض النتائج التي تستخدم لاتخاذ القرارات الجوهرية المناسبة.

**الخاتمة:**

قدمنا في هذا الفصل دراسة نظرية حول مفهوم الذكاء الاصطناعي بشكل عام وآلية عملها ودورها في عملية اتخاذ القرار، وقدمنا دراسة مجال تنقيب البيانات والتنقيب في البيانات التعليمية، وقمنا بتوضيح أبرز أنواعها المستخدمة في عملية التعرف على الأفكار والأشياء وفهمها وتجميعها في فئات محددة مسبقًا، وركزنا على الأسلوب المتبع في هذه الدراسة وشرح خوارزمية شجرة القرار ومما تتكون وطريقة عملها وهيكليتها، وفي نهاية هذا الفصل قمنا بتلخيص وعرض أبرز الدراسات ذات العلاقة حول خوارزميات الذكاء الاصطناعي ودورها في عملية اتخاذ القرار في مجال التنقيب عن البيانات في عدة مجالات ، ومناقشتها.

**الفصل الثالث**

**المنهجية**

**المقدمة:**

منهجية البحث العلمي خطوة من الخطوات المهمة التي يبنى عليها البحث العلمي، ويرجع السبب في هذا الأمر الى انه من أهم ضوابط البحث العلمي أن يكون منظماً ودقيق، أي أن تعتبر الوسيلة او الأسلوب العلمي الذي يتبعه الباحث والذي من خلاله سيجد الحل لمشكلة بحثه.

هذا الفصل يتطرق الى شرح مفصل عن منهجية البحث بشكل مفصل، تتبع هذه الدراسة المنهجية الأساسية للتنقيب عن البيانات (Data mining)، وتوضيح ماذا تم في كل خطوة من خطوات المنهجية.

**منهجية التنقيب في البيانات:**

التنقيب في البيانات لا يقتصر على التحليل الآلي وتطبيق الخوارزميات بل يشمل مجموعة واسعة من العمليات على شكل خطوات متتالية، حيث تعتمد كل خطوة على الخطوة السابقة لها، كما أنه يمكن الرجوع عند كل خطوة إلى الخطوة السابقة، وبشكل عام توجد عدة منهجيات في مجال التنقيب في البيانات.

1. **تجميع البيانات:**

عملية جمع البيانات واحدة من أهم مهام مشاريع تنقيب البيانات يتم في هذه المرحلة تجميع البيانات من مصادر مختلفة أو مصدر بيانات، يمكن جمع البيانات بعدة طرق، تم في هذه الدراسة الاعتماد على بيانات جاهزة مخزنة في قاعدة البيانات مركز التطوير المعلوماتي الخاصة بالطلاب الذين التحقوا بجامعة سبها بداية من سنة 2017 الى سنة 2021، بمختلف كليات الجامعة، بالإضافة إلى أنه تمت الاستعانة بـ استبانة خاصة بالطلبة والغرض الذي اجريت بسببه معرفة الأسباب التي أدت بالطالب باختيار هذا التخصص، هذه الاستبانة تم اجراؤه من قبل (Amelalbohali, Almahdi Alshareef, Hassan Al Gaddafi, 2021)، اخترنا منها بعض الأسئلة لتقوية خصائص قواعد البيانات.

تم استخدام سبعة مجموعات مختلفة من البيانات تخص كليات جامعة سبها في هذه الدراسة، المجموعة الأولى خاصة بطلاب كلية الآداب، والمجموعة الثانية تحوي بيانات كلية العلوم، والثالثة خاصة ببيانات طلبة كلية تقنية المعلومات، والرابعة خاصة بطلبة كلية التجارة والعلوم السياسية، والخامسة تضم طلبة الكليات الطبية المختلفة (الطب البشري، طب الأسنان، الصيدلة، التمريض)، والسادسة نحوي بيانات كليات التربية المختلفة التابعة لجامعة سبها (كلية التربية أوباري، كلية التربية الشاطئ، كلية التربية غات)، والسابعة تحتوي قاعدة بيانات الخاصة بطلبة الهندسة (كلية الهندسة سبها، كلية الهندسة الشاطئ) تحتوي قاعدة البيانات على معلومات 2693 طالب، وتحتوي على 22 خاصية تم اختيارهن من قبل الباحثان كما في الجدول التالي ()

الجدول () يوضح وصف الخصائص لقاعدة البيانات:

|  |  |  |
| --- | --- | --- |
| الخاصية (العمود) | الوصف | نوع البيانات |
| الرقم الدراسي | تحتوي على الرقم الدراسي للطالب. | رقمي |
| مكان الميلاد | يحتوي على مكان الميلاد، اذا كان في (مدينة سبها او غير ذلك). | حرفي |
| الجنس | يحتوي على جنس الطالب، اذا كان (ذكر او انثى). | حرفي |
| التخصص الثانوي | يحتوي على التخصص الثانوي للطالب، اذا كان (علمي او ادبي). | حرفي |
| سنة الالتحاق | يحتوي على سنة الالتحاق بالجامعة. | رقمي |
| عنوان السكن | يحتوي على عنوان السكن، اذا كان في مدينة سبها او غير ذلك. | حرفي |
| الديانة | يحتوي على ديانة الطالب. | حرفي |
| الحالة الاجتماعية | يحتوي على الحالة الاجتماعية للطالب، اذا كان (أعزب او خاطب أو متجوز أو مطلق). | حرفي |
| المعدل الثانوي | يحتوي على المعدل الثانوي للطالب. | رقمي |
| المدرسة الثانوية | تحتوي على اسم المدرسة المتحصل منها الشهادة الثانوية. | حرفي |
| السؤال الاول | التسجيل في التخصص الحالي بناء على؟   1. رغبتك انت، ب. الكلية، ج. صديق، د. العائلة) | حرفي |
| السؤال الثاني | هل كان اختيارك للتخصص الجامعي بناء على أقرب جامعة من اقامتك؟  (أ. نعم، ب. لا) | حرفي |
| السؤال الثالث | هل كانت رغبتك العمل في الشركات الخاصة الدافع لاختيار التخصص؟   1. نعم، ب. لا) | حرفي |
| السؤال الرابع | هل كانت رغبتك العمل في القطاعات العامة الدافع لاختيار التخصص؟   1. نعم، ب. لا) | حرفي |
| السؤال الخامس | درجة الاب العلمية؟   1. دون مستوى، ب. ابتدائي، ج. اعدادي، د. ثانوي، ه. خريج جامعة، و. عليا) | حرفي |
| السؤال السادس | درجة الام العلمية؟   1. دون مستوى، ب. ابتدائي، ج. اعدادي، د. ثانوي، ه. خريج جامعة، و. عليا) | حرفي |
| السؤال السابع | هل لديك اخوة أكبر منك؟   1. نعم، ب. لا) | حرفي |
| السؤال الثامن | ماهي درجتهم العلمية(الاغلبية)؟  (أ. لا شيء، ب. ابتدائي، ج. اعدادي، د. ثانوي، ه. خريج جامعة) | حرفي |
| السؤال التاسع | مهنة الاب؟   1. موظف، ب. اعمال تجارة، ج. اعمال حرة، د. متقاعد، ه. دون عمل) | حرفي |
| السؤال العاشر | مهنة الأم؟   1. موظفة، ب. متقاعدة، ج. ربة منزل) | حرفي |
| السؤال الحادي عشر | هل حالة عائلتك المادية تسمح لك بأخذ دورات متخصصة تساعدك في الدراسة الجامعية؟   1. نعم، ب. لا) | حرفي |

1. **تهيئة البيانات:**

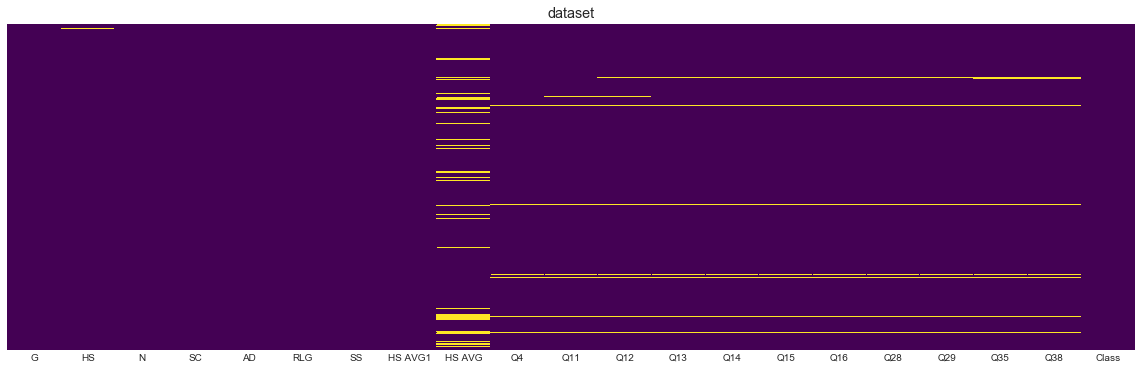
في هذه المرحلة تمت معالجة البيانات مسبقًا وتنظيف جميع الحالات المشوشة من البيانات، بحيث تركز هذه الخطوة المهمة في الغالب على الاحتفاظ بالبيانات والسجلات الأكاديمية لضمان التنبؤ الفعال، حيث تتمثل هذه المهمة في جعل البيانات مفيدة، نظرا لأننا جمعنا البيانات من مصادر مختلفة مما يعني أن البيانات ستكون بتنسيق غير منظم ويمكن العثور على الكثير من القيم الفارغة والقيم الغير صالحة والبيانات الغير مرغوب فيها، لذلك يجب تنظيف هذه البيانات بالبيانات المناسبة وازالة البيانات الفارغة والمفقودة، في بعض الاحيان قد تحتوي البيانات على قيم مشوشة وليس بتنسيق مناسب، و البيانات ذات الجودة المنخفضة تؤدي الى نتائج منخفضة عند ادخالها على خوارزميات تنقيب البيانات، ولذلك ينبغي رفع جودة البيانات قبل البدء بتحليلها.

الخطوات التي قمنا بها لتهيئة البيانات المجمعة هي كالتالي:

1. **معالجة القيم المفقودة:**

مشكلة البيانات المفقودة من المشكلات البحثية الشائعة والمتكررة عند جمع البيانات، وهي شائعة جداً في العديد من البيانات الواقعية، يمكن أن تؤذي القيم المفقودة الى تحيز نتائج التعلم الآلي أو التقليل من دقة النموذج، ولأن النموذج الذي سيتم استخدامه لتحليل البيانات يفترض دائما وجود معلومات كاملة عن جميع المتغيرات المستخدمة في التحليل، وتعد فقد البيانات مشكلة كبيرة للباحث، وعدم معالجتها بالشكل المناسب قد يسبب للباحث بعض المشكلات مثل قليل حجم العينة الى حجم غير ملائم، او الحصول على نتائج غير جيدة.

الشكل ادناه يبين عينة البيانات المفقودة من مجموعة البيانات الخاصة بالدراسة.



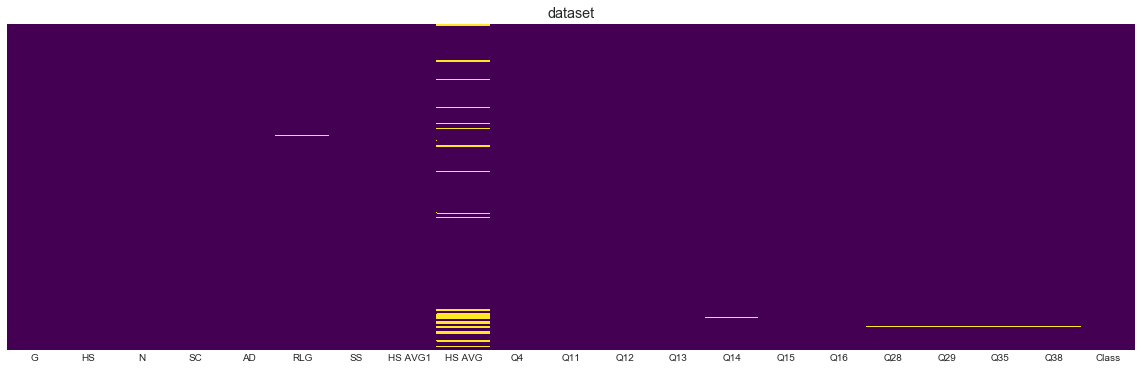
نلاحظ من خلال الشكل أن البيانات المفقودة هي الممثلة باللون الأصفر من مجموع البيانات الكلي.

**طريقة معالجة البيانات المفقودة:**

توجد عدة طرق تستخدم في معالجة القيم المفقودة، استخدمنا الطرق التي تتماشى مع انواع البيانات في قاعدة البيانات الخاصة بالدراسة وهي كالتالي:

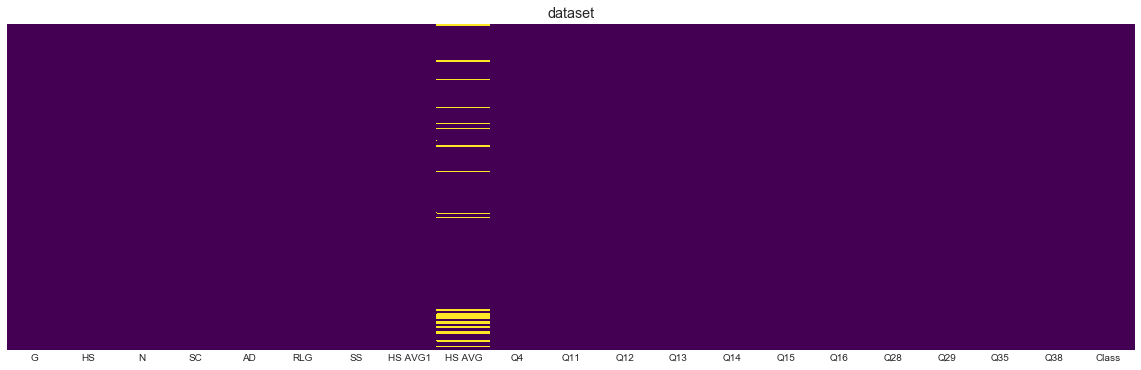
**1. حذف الصفوف ذات القيم المفقودة:**

في هذه الطريقة يتم حذف الصفوف أو الأعمدة التي تحتوي على قيم خالية، إذا كان الأعمدة تحتوي على أكثر من نصف الصفوف قيم فارغة، في هذه الحالة يمكننا حذف الصفوف التي تحتوي على قيم أعمدة فارغة، نلاحظ من الشكل () انه هناك صفوف قليلة اغلب قيمها فارغة سيتم حذفها لانها ليست قيم كثير ولا تؤثر في بالنسبة للمجموعة الكلي للبيانات، والشكل التالي يوضح بعد حذف السجلات التي أغلب قيمها مفقودة.



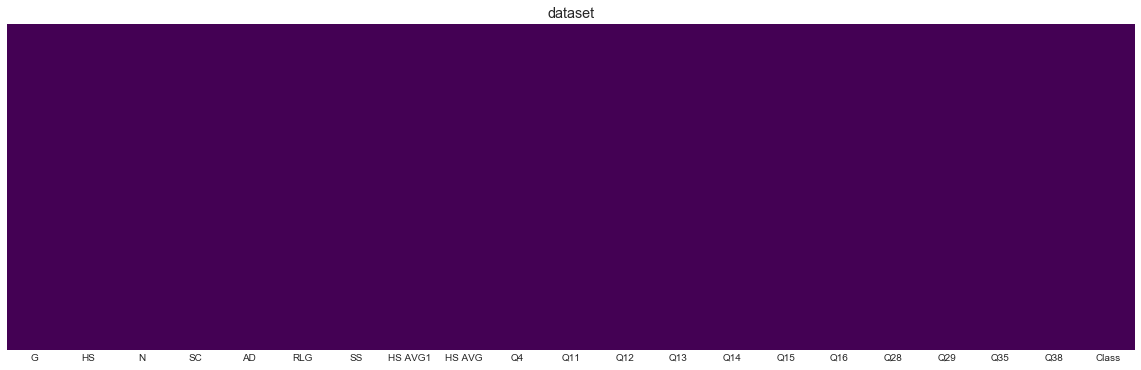
1. **طريقة الاستبدال بالقيم الأكثر تكرار:**

تستخدم هذه الطريقة عندما تكون القيم المفقودة في أعمدة بيانات فئوية سواء كانت (سلسلة أو رقمية)، ويقصد بالفئوية أن البيانات تكون على صورة فئات محددة مثل الجنس (ذكر، انثى)، وتتم المعالجة بعد النظر لتوزيع نسب الفئات في العمود، إذا كان 80% ذكر، و20% انثى فتتم ملئ القيم المفقودة من هذا العمود بقيمة ذكر، وبما أن جميع الأعمدة المتكونة منها قاعدة البيانات الخاصة بالدراسة أعمدة فئوية ماعدا عمود تقدير الطالب، سيتم استخدام هذه الطريقة لمعالجة هذه القيم المفقودة، الشكل التالي يوضح البيانات بعد معالجة الأعمدة ذات البيانات الفئوية.



1. ملئ القيم المفقودة بالمتوسط الحسابي:

في حالة البیانات التي يحويها العمود رقمیة حقیقیة، فمن الصعب تحدید الاكثر تكرارا ففي ھذه الحالة بالإمكان استبدال البیانات المفقودة بالمتوسط، سيتم معالجة القيم المفقودة التي بعمود تقدير الطالب بالمتوسط الحسابي للعمود، الشكل التالي يبين مجموعة البيانات بعد كل خطوات معالجة القيم المفقودة.



**دمج البيانات:**

هذه الخطوة احدى خطوات معالجة البيانات ضمن مشاريع التنقيب في البيانات، بحيث تتطلب عمليات تحليل وتنقيب البيانات في كثير من الأحيان إجراء عمليات دمج البيانات من عدة مصادر والدمج الجيد يساعد في اختصار المعلومات وتجنب التكرارات الزائدة او التناقضات، وهو ما يساهم في زيادة كفاءة وسرعة ودقة إجراءات التحليل او التنقيب، وعدم التجانس في بنية البيانات يشكل تحديا كبيرا لعمليات دمج البيانات، حيث انها تتطلب الدمج بطريقة صحيحة مهما تعددت المصادر والبيئات الحاضنة لها بحيث يتم الحصول في النهاية على قاعدة بيانات كاملة ذات هيكلية موحدة وبيانات منظمة موزعة بدقة في أماكنها المناسبة، بناء علي ذلك تمت عملية دمج بيانات الطلاب لكل من كلية الطب وكلية الآداب وكلية التربية بجميع تخصصاتها، بحيث تكون البيانات في كلاس مخصص لكلية الطب وكلاس واحد مخصص لكلية الآداب وكلاس واحد مخصص لكلية التربية ، بمعنى يحتوي الكلاس علي بيانات الطلاب في هده الكليات، أي كلية التمريض وكلية الصيدلة وكلية طب الاسنان وكلية طب البشري تندرج في كلاس مخصص لكلية الطب او كلية العلوم الطبية بجميع فروعها وتخصصاتها، وكلية الآداب بجميع تخصصاتها التي تشمل ايضا كلية القانون تم دمج سجلات طلاب تلك الكليات في كلاس واحد مخصص بكلية الآداب، وكلية التربية بجميع فروعها الموجدة في جامعة سبها، تم دمج سجلاتها في كلاس واحد مخصص لكلية التربية.

**اختيار السمات:**

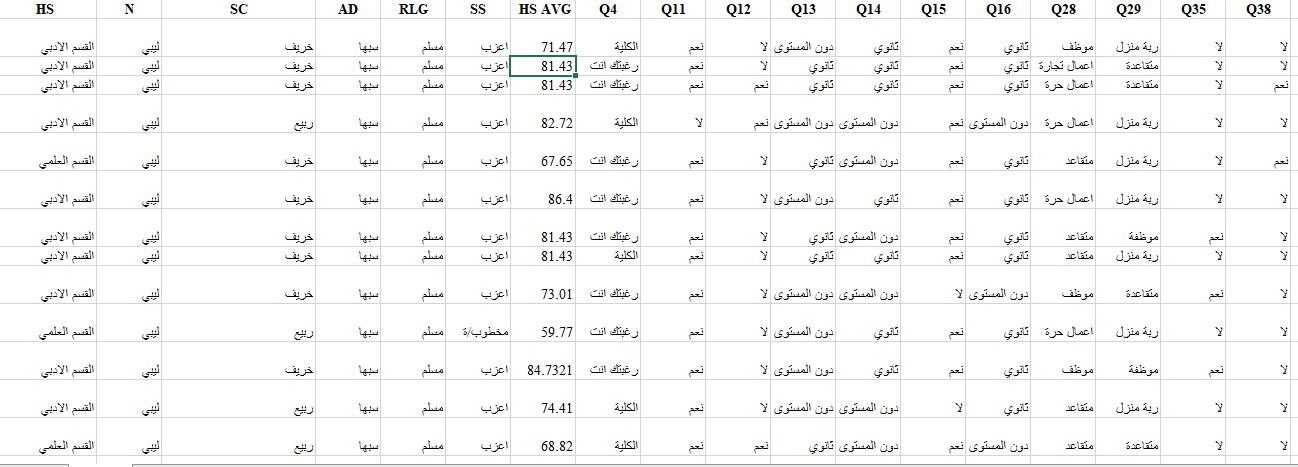
في هذه الخطوة قام الباحثان باختيار السمات ذات الفائدة والتأثير والتي تؤثر على المتغير التابع وهي

الجنس، التخصص الثانوي، سنة الالتحاق، عنوان السكن، الحالة الاجتماعية، المدرسة الثانوية، وتم اختيار احدى عشر سؤال من مجموع أسئلة الاستبانة والتي رآها الباحثان ذات التأثير أكبر من مجموع الأسئلة، وتم استثناء وازالة بعض السمات التي ليس لها تأثير على اختيار تخصص الطالب، مثل: مكان الميلاد، الرقم الدراسي، الديانة، وعدة أسئلة من الاستبانة المستعانة بيها لإعداد قاعدة البيانات والتي ليس لها تأثير كبير على اختيار التخصص.

1. **تغير البيانات:**

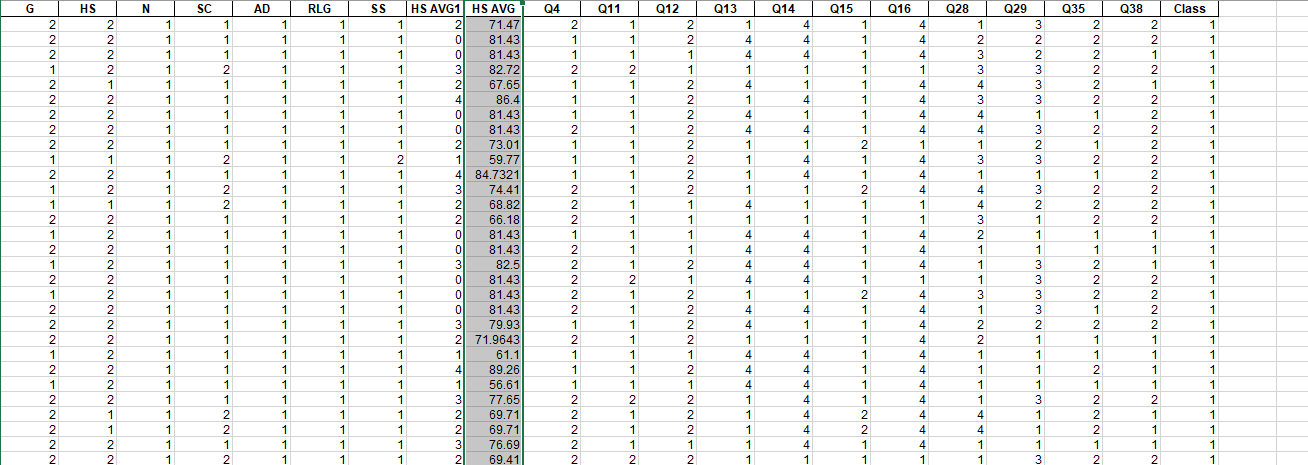
عادة ما تتضمن أي مجموعة بيانات واقعية أعمدة متعددة هذه الأعمدة يمكن ان تحتوي على بيانات فئوية نصية، في هذه العملية يتم تحويل وتغيير البيانات من المراحل المهمة في تحضير البيانات للتحليل والتنقيب، وذلك لإن الخوارزميات تفهم الأرقام وليس النصوص، نحتاج تحويل كل فئة نصية إلى أرقام حتى تتمكن الخوارزميات من معالجتها باستخدام المعادلات الرياضية، هذا هو السبب الرئيسي وراء الحاجة الى تحويل الأعمدة الفئوية إلى أعمدة رقمية حتى تتمكن الخوارزميات من فهمها والتعامل معها.

الشكل التالي أعمدة البيانات الفئوية التي تحتويها قاعدة البيانات.



توجد عدة طرق للتغير البيانات الفئوية، في هذه الدراسة سنطبق طريقة ترميز التسمية Label Encoding

في هذه الطريقة يتم تعيين عدد صحيح فريد لكل فئة في العمود مثل: عمود الجنس يتم استبدال قيمة ذكر برقم (1)، والأنثى برقم (2)، وهكذا يتم التعامل مع بقية الأعمدة التي تحتويها قاعدة البيانات، الشكل التالي يوضح قاعدة البيانات بعد تغير البيانات الفئوية إلى رقمية.



الجداول التالية بين كل عمود فئوي وكيف تم تغير قيمه لقيم رقمية.

الجدول () يوضح كيف تم تغير قيم عمود الجنس

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| Male(1), female(2) | Gender |

هذا العمود في قاعدة البيانات يبين جنس الطالب، والمرموز بالقيم: ذكر(1) او انثى(2).

الجدول () يوضح كيف تم تغير قيم حقل التخصص الثانوي

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| Scientific(1), Litreture(2) | High school |

هذا العمود في قاعدة البيانات يبين التخصص الثانوي للطالب، المقدر بقيمتين: علمي (1) و أدبي (2).

الجدول () يوضح كيف تم تغير قيم عمود التخصص الثانوي

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| Libyan(1), other(2) | Nationality |

يبين هذا الحقل في قاعدة البيانات جنسية الطالب، وتحتوي على القيم: ليبي (1) و غير ذلك(2).

الجدول () يوضح كيف تم تغير قيم عمود الالتحاق بالكلية

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| fall(1), spring(2) | Secession class |

يشير هذا العمود من أعمدة قاعدة بيانات الدراسة، الفصل الجامعي الملتحق فيه الطالب، اما ان يكون: فصل الخريف (1) او فصل الربيع (2).

الجدول () يوضح كيف تم تغير قيم عمود عنوان سكن الطالب

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| sebha(1), other(2) | Address |

يوضح هذا الحقل من حقول قاعدة بيانات الدراسة، على مكان السكن الحالي للطالب والمحدد بقيمتين: سبها (1) او غير ذلك (2).

الجدول () يوضح كيف تم تغير قيم عمود ديانة الطالب

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| muslim(1), other(2) | Religion |

يوضح هذا الحقل من حقول قاعدة بيانات الدراسة، على ديانة الطالب وتحتوي على القيمتين: مسلم (1) او غير ذلك (2).

الجدول () يوضح كيف تم تغير قيم عمود الحالة الاجتماعية لطالب

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| single(1), engaged(2), married(3), divorced(4) | Social status |

هذا الحقل من حقول قاعدة بيانات الدراسة يوضح الحالة الاجتماعية لدى الطالب بالقيم: أعزب (1) او خاطب/مخطوبة (2) او متجوز/ة(3) او مطلق/ة(4).

الجدول () يوضح كيف تم تغير قيم عمود معدل الطالب في الثانوي.

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| 50-64%(1), 65-74%(2), 75-84%(3), 85-100%(4) | High school average |

يوضح هذا الحقل المعدل الثانوي لدى الطالب، وبما ان قيم المعدل كثيرة قمنا بتقليص القيم الى اربع قيم، فالقيم من 50 الى 64% والذي يعتبر المقبول رمزنا له بالقيمة (1)، والقيم من 65 الى 74% والذي يعتبر المعدل الجيد رمزنا له بالقيمة (2)، والقيم من 75 الى 84% والذي يعتبر المعدل جيد جدا رمزنا له بالقيمة (3)، والقيم من 85 الى 100% والذي يعتبر المعدل الممتاز رمزنا له بالقيمة (4).

الجدول () يوضح كيف تم تغير قيم عمود السؤال الأول

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| Your desire(1), faculty(2), friend(3), family(4) | Q1: Have you enrolled in this specialty based on? |

قمنا بترميز قيم السؤال الاول في قاعدة بيانات الدراسة كالتالي: رغبتك أنت(1)، الكلية(2)، صديق(3)، العائلة(4).

الجدول () يوضح كيف تم تغير قيم عمود السؤال الثاني

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| yes(1), no(2) | Q2: is the desire to work in private companies the main reason for choosing your specialty? |

قمنا بترميز قيم السؤال الثاني في قاعدة بيانات الدراسة كالتالي: نعم (1)، لا(2).

الجدول () يوضح كيف تم تغير قيم عمود السؤال الثالث

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| yes(1), no(2) | Q3: is the desire to work in the public sectors the main reason for choosing your specialty? |

قمنا بترميز قيم السؤال الثالث في قاعدة بيانات الدراسة كالتالي: نعم (1)، لا(2).

الجدول () يوضح كيف تم تغير قيم عمود السؤال الرابع

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| Below level(1), primary(2), prepatory(3), secondary(4), university graduate(5), high(6) | Q4: father's education level? |

قمنا بترميز قيم السؤال الرابع في قاعدة بيانات الدراسة كالتالي: دون مستوى (1)، ابتدائي (2)، اعدادي(3)، ثانوي(4)، جامعي(5)، عليا(6).

الجدول () يوضح كيف تم تغير قيم عمود السؤال الخامس

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| Below level(1), primary(2), prepatory(3), secondary(4), university graduate(5), high(6) | Q5: mother's education level? |

قمنا بترميز قيم السؤال الخامس في قاعدة بيانات الدراسة كالتالي: دون مستوى (1)، ابتدائي(2)، اعدادي(3)، ثانوي(4)، جامعي(5)، عليا(6).

الجدول () يوضح كيف تم تغير قيم عمود السؤال السادس

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| yes(1), no(2) | Q6: do you have elder brothers? |

قمنا بترميز قيم السؤال السادس في قاعدة بيانات الدراسة كالتالي: نعم (1)، لا (2).

الجدول () يوضح كيف تم تغير قيم عمود السؤال السابع

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| none(1), primary(2), prepatory(3), secondary(4), university graduate(5 | Q5: what is their education level (majority)? |

قمنا بترميز قيم السؤال السابع في قاعدة بيانات الدراسة كالتالي: لا شي(1)، ابتدائي(2)، اعدادي(3)، ثانوي(4)، خريج جامعي(5).

الجدول () يوضح كيف تم تغير قيم عمود السؤال الثامن

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| employee(1), business(2), freelancer(3), retired(4), no work(5 | Q5: father's profession? |

قمنا بترميز قيم السؤال الثامن في قاعدة بيانات الدراسة كالتالي: موظف (1)، أعمال تجارة(2)، أعمال حرة(3)، متقاعد(4)، بلا عمل(5).

الجدول () يوضح كيف تم تغير قيم عمود السؤال التاسع

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| employee(1), retired(2), housewife(3) | Q5: father's profession? |

قمنا بترميز قيم السؤال التاسع في قاعدة بيانات الدراسة كالتالي: موظفة(1)، متقاعدة(2)، ربة منزل(3).

الجدول () يوضح كيف تم تغير قيم عمود السؤال العاشر

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| yes(1), no(2) | Q6: does your family's financial means allow you to enroll in training or specialized courses that help you excel in your university studies? |

قمنا بترميز قيم السؤال العاشر في قاعدة بيانات الدراسة كالتالي: نعم(1)، لا(2).

الجدول () يوضح كيف تم تغير قيم عمود الحادي عشر

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| yes(1), no(2) | Q6: is your choice of university major based on the university's proximity to the place of residence? |

قمنا بترميز قيم السؤال الحادي عشر في قاعدة بيانات الدراسة كالتالي: نعم(1)، لا(2).

الجدول () يوضح كيف تم تغير قيم عمود الهدف أو الكلاس.

|  |  |
| --- | --- |
| Possible value with numerical equilevants | Description |
| Faculty of arts(1)  Faculty of science(2)  Faculty of IT(3)  Faculty of economics and political science(4)  Faculty of medicine(5)  Faculty of education(6)  Faculty of engineering(7) | Current university (the class attribute) |

أخيرا قمنا بترميز هذا الحقل والذي يحتوي على قيم التخصص الجامعي والذي ايضا يعتبر حقل الclass بالقيم التالية: كلية الآداب(1)، كلية العلوم(2)، كلية تقنية المعلومات(3)، كلية التجارة والعلوم السياسية(4)، كلية الطب(5)، كلية التربية(6)، كلية الهندسة(7).

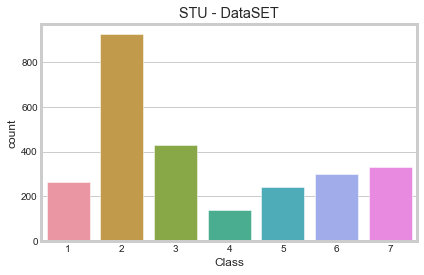
بعد الإنتهاء من هذه الخطوة تصبح البيانات جاهزة لإدخالها لخوارزميات التصنيف لتدرب عليها.

**المقدمة**

نقدم في هذا الفصل دراسة البيانات المعطاة لتطبيق الخوارزميات وبيئة العمل المستخدمة والتجارب التي اجريناها للحصول على نتائج أفضل.

**بيانات التدريب والاختبار:**

بعد الإنتهاء من خطوات معالجة البيانات، أصبحت البيانات جاهزة لإدخالها لخوارزميات التصنيف، هذه البيانات تتكون من 2632 سجل موزعة بمختلف الكليات السبعة الشكل () يبين عدد الطلاب بكل كلية، وفيما يتعلق ببناء النموذج فقد تم اتباع أسلوب التقسيم المئوي (percentage split)، في عملية التدريب والاختبار، والذي يقوم بتقسيم البيانات وفق نسبة يحددها الباحث أثناء عملية البحث حيث يتم تدريبها ومن ثم اختبارها وإعطاء النموذج النهائي، وقد قمنا بتقسيم البيانات الى 80% بيانات تدريب و20% المتبقية من البيانات لاختبار أداء النموذج.



**الآداب**

**العلوم**

**تقنية المعلومات**

**الطب**

**CLass**

**التربية**

**الهندسة**

**الاقتصاد**

نلاحظ من الشكل ان هناك تباين كبير في عدد الطلاب في كل كلية، حيث يبلغ عدد الطلاب كالتالي: (العلوم 923 طالب، تقنية المعلومات طالب 429، الهندسة 333 طالب، التربية 299 طالب، الآداب 266 طالب، الطب 244 طالب، الاقتصاد (التجارة والعلوم السياسية) 138 طالب)، هذا التباين قد يسبب خلل في تدرب نموذج ودقة نتائجه، حيث معالجته لخروج بنتائج جيدة ودقيقة.

**النتائج والمناقشة:**

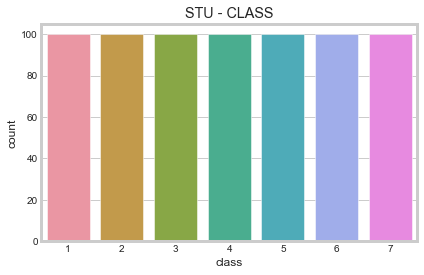
اعتمد الباحثان على استخدام شجرة القرار في التنبؤ باختيار التخصص الأقرب لطالب بناء على درجاته وبعض المعلومات الشخصية الخاصة به، ومقارنة هذه الخوارزمية ببعض خوارزميات التصنيف كخوارزمية ناييف باييز، وخوارزمية الجار الأقرب، وذلك للخروج بأفضل النتائج وإثبات صحة بعض الدراسات السابقة ()،()،()، التي تفوقت فيها خوارزمية شجرة القرار على باقي خوارزميات التصنيف، تم انشاء نماذج التنبؤ باستخدام أداة التعلم الأكثر شيوعاً Anaconda وبالاعتماد على الواجهة الطرفية jupyter notebook الموجودة بهذه الأداة، وذلك لإنها سهلة الاستخدام وتوفر واجهات رسومية يسهل التعامل معها.

نظراً لأن طبيعية البيانات غير متوازنة كما أشارنا في الفقرة السابقة ()، فقمنا بإجراء تجربتين على البيانات، التجربة الأولى قمنا فيها بتطبيق الخوارزميات السابقة على البيانات بتوزيعها الأولي، وفي التجربة الثانية قمنا بموازنة البيانات وجعلها متساوية في كل (Class) الكليات، من خلال طريقة under sampling تنص هذه الطريقة على مساواة بين كل الفئات وذلك بالتقليل من الفئات الكبيرة، ومساواتها بالأقلية بحيث تصبح متوازنة في كل الفئات، فقمنا بمساواتها وأخذ من كل فئة عينة عشوائية عددها 100 سجل من كل فئة من الفئات، وذلك من أجل تعزيز دقة النموذج.

جدول () يوضح نتائج الخوارزميات الثلاثة في التجربة الأولى:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **الخوارزمية** | **الدقة** | **الاستجابة** | **الاستدعاء** | **f-measure** |
| **شجرة القرار** | %61.85 | 65% | 66% | 65% |
| **ناييف باييز** | 10.05% | 15% | 16% | 3% |
| **الجار الأقرب** | 55.97% | 57% | 60% | 58% |

نلاحظ من الجدول السابق أن خوارزمية شجرة القرار أفضل خوارزمية مقارنةً ببقية الخوارزميات، بكل المقاييس المستخدمة، نسبة دقة خوارزمية شجرة القرار 61.85% وخوارزمية الجار الأقرب بنسبة 55.97%، وخوارزمية ناييف باييز بينت نتيجة ضعيفة مقارنة بالخوارزميتين، قد يرجع هذا التدني في هذه الخوارزمية نتيجة لطبيعية البيانات، أو لعددها وأبعادها، أو قد يرجع إلى أن البيانات كانت غير متوازنة، واتفقت نتيجة دراسة الحالية مع عدة دراسات ()، من حيث تفوق خوارزمية شجرة القرار على باقي الخوارزميات الأخرى، هذه نتيجة لا يمكننا الاعتماد عليها من حيث مقياس الدقة، وذلك نظراً لطبيعة البيانات الغير متوازنة، نتيجة لذلك قمنا بتحسين أداء النتائج بمساواة بين جميع الفئات، لاستخلاص نتيجة دقة يمكن الاعتماد عليها، حيث قمنا بالمساواة بين الخصائص عن طريق Under Sampling، تهدف هذه التقنية بالتقليل من الفئات الكبيرة ومساواتها بالقليلة، وذلك بأخذ عينة عشوائية من كل الفئات بشكل متساوي، وفي هذه الدراسة قمنا بالمساواة بأخذ 100 سجل طالب لكل من جميع الفئات بشكل عشوائي، الشكل () يوضح الفئات بعد عملية الموازنة.



**الاقتصاد**

**الآداب**

**العلوم**

**تقنية المعلومات**

**الطب**

**CLass**

**التربية**

**الهندسة**

الجدول التالي يبين نتائج الخوارزميات بعد عملية موازنة الفئات.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **الخوارزمية** | **الدقة** | **الاستجابة** | **الاستدعاء** | **f-measure** |
| **شجرة القرار** | 64.28 | 63% | 66% | 63% |
| **ناييف باييز** | 60.71% | 58% | 67% | 53% |
| **الجار الأقرب** | 49.28% | 49% | 47% | 46% |

يبين الجدول السابق نتائج الخوارزميات بعد عملية موازنة الفئات، إذ تبين حتى في هذه التجربة حتى بعد عملية موازنة الفئات، أن أفضل خوارزمية هي خوارزمية شجرة القرار بنسبة دقة 64.28%، ارتفعت بثلاث نسب مقارنة بالتجربة السابقة التي كانت فيها نسبة الدقة في خوارزمية شجرة القرار 61.85%، نلاحظ أنه ليس هناك فارق كبير بين التجربتين في هذه الخوارزمية، ولكن النتيجة التي أظهرتها هذه التجربة من حيث مقياس الدقة بالإمكان الاعتماد عليها، وذلك لإن الفئات كانت فيها متساوية على عكس التجربة الأولى، ونلاحظ أيضاً الارتفاع بنسبة كبيرة جداً في خوارزمية ناييف باييز من حيث مقياس إذ كانت في التجربة الأولى بنسبة 10.05%، وفي هذه التجربة بلغت نسبة الدقة 60.71%، فارق كبير بين التجربتين وهذا يبين أن خوارزمية ناييف باييز عندما تكون الفئات متعددة وغير متوازنة لا تعمل بشكل صحيح.

**مصفوفة الارتباك**

يبين الجدول () مصفوفة الارتباك لنموذج التنبؤ بخوارزمية شجرة القرار.

من خلال الشكل السابق نلاحظ أن الإناث أكثر من الذكور في قاعدة البيانات الخاصة بالدراسة.