إستخدام تقنية SMOTE لمعالجة مشكلة عدم توازن البيانات في استبيان الطلاب لتقييم العملية التدريسية في الجامعة

Using the SMOTE technology to fit the problem of Data Imbalance in the Student Survey to evaluate the teaching process at the University

رانية عبد العال عثمان سليمان

أستاذ محاضر

Nucsit016@gmil.com

كلية علوم الحاسوب و تقانة المعلومات – جامعة النيلين

الخرطوم ، السودان

المستخلص:

في الآونة الأخيرة حرصت أغلب المؤسسات الأكاديمية على إيجاد العوامل التي تزيد من نتائج التعليم وأصبحت جودة التعليم هدف مهم تسعى جميع المؤسسات التعليميه اليه خاصة في ظل توفر قدر ضخم من البيانات التي يمكن الاستفادة منها للحصول على نتائج ممتازة. و كثر استخدام خوارزميات تنقيب البيانات لتحليل هذه العوامل ، في هذه الدراسة تم بناء نموذج لتقييم الأداء الأكاديمي للأستاذ الجامعي وذلك عبر تصنيف اجابات الطلاب المتحصل عليها بواسطة استبانة الطلاب والتي تم بناءها باستخدام نموذج جوجل Google form وهي مكونة من 36 سؤال بالإضافة لبعض البيانات الأخرى الخاصة بالأستاذ مثل معدله التراكمي الذي تخرج به ، الدرجة العلمية ، و عدد سنوات الخبرة. وقد تم التصنيف باستخدام

ثلاث خوارزميات تصنيف (classification) وهي (Random Forest, Decision Tree, Naïve Bayes) وذلك لأنها تعمل مع الـــ Multi Class classification وعند المقارنة بين اداء الخوارزميات اعطت خوارزمية الغابة العشوائية معدل أعلى من خوارزميتي Naïve Bayes وشجرة القرار حيث كان معدل الدقة 0.68 في خوارزمية الغابة العشوائية و هو منخفض نوعا ما بالإضافة الى أن نتائج مقاييس مصفوفة الارتباك recall و precision مع الفصائل 0 و 1 كانت منخفضة بينما النتائج مرتفعة مع الفصيل 2 أي أن النموذج منحاز للأغلبية و هذا نسبة لعدم التوازن في توزيع البيانات ، لذلك استخدم الباحث طريقة SMOTE لمعالجة هذه المشكلة. استخدمت لغة بايثون لبناء النموذج ، يمكن تقسيم عملية تصميم النموذج النهائي الى أربع مراحل ، المرحلة الأولى هي المعالجة المسبقة (preprocessing) ، و الثانية هي تطبيق الثلاث خوارزميات المقترحة ، في المرحلة الثالثة يتم اختبار النموذج باستخدام أدوات مصفوفة الارتباك (precision, recall , f1-score) و المرحلة الأخيرة تم استخدام طريقة SMOTE للتعامل مع عدم توازن البيانات. خلصت الدراسة إلى أن أداء النموذج تحسن أي أعطى معدل دقة أعلى عند تطبيق SMOTE حيث تحسنت الدقة العامة وايضا قدرة وأداء النموذج على التنبؤ و كانت خوارزمية Random Forest هي الأفضل من حيث التنبؤ بالعوامل الأكثر تأثيرا على معدلات تقييم الطالب للأستاذ ، حيث ارتفع معدل الدقة الى 0.82 بدلا عن 0.68

الكلمات المفتاحية : تقييم الأداء، SMOTE ، تنقيب البيانات التعليمية، بيانات غير متوازنة، خوارزميات التصنيف.

Abstract:

Recently, most academic institutions have been keen to find factors that increase educational results, and the quality of education has become an important goal that all educational institutions seek, especially in light of the availability of a huge amount of data that can be used to obtain excellent results. And data mining algorithms were frequently used to analyze these factors. In this study, a model was built to evaluate the academic performance of the university professor, by classifying the students' answers obtained by the students' questionnaire, which was built using a Google form, which is made up of 36 questions in addition to some other data for the professor. Such as his cumulative grade point average, academic degree, and number of years of experience. The classification was done using three classification algorithms (classification) which are (Random Forest, Decision Tree, Naïve Bayes) because it works with the Multi Class Classification When comparing the performance of the algorithms, the random forest algorithm

gave a higher rate than the two Naïve Bayes algorithms and the decision tree, where the accuracy rate was 0.68 in the random forest algorithm, which is rather low. In addition, the results of the measures of confusion matrix recall and precision with classes 0 and 1 were low, while the results are high with Faction 2 means that the model is biased to the majority and this is due to the imbalance in the data distribution, so the researcher used the SMOTE method to address this problem. Python language was used to build the model. The process of designing the final model can be divided into four stages. The first stage is preprocessing. The second stage is the application of the three proposed algorithms. In the third stage, the model is tested using the tools of the precision, recall, and f1-score confusion matrix.)) And in the last stage, the SMOTE method was used to deal with the data imbalance. The study concluded that the performance of the model improved, that is, it gave a higher accuracy rate when applying SMOTE, where the general accuracy and the model's predictive ability and performance improved. The Random Forest algorithm was the best in terms of predicting the factors most affecting the student's assessment rates for the professor, where the accuracy rate increased to 0.82 instead of 0.68

1. المقدمة :

يمثل التعليم الجامعي في الدول المتقدمة والنامية مصدر إشعاع علمي، وثقافي، وحضاري؛ لأنَّه المسئول عن إعداد الكفاءات المتخصصة اللازمة للنهوض بأعباء التنمية في مختلف المجالات.

وما من شك أنَّه لن يستطيع أن يقوم بتلك الوظيفة إلاَّ إذا توافرت له الإمكانات التي تعينه، وفي مقدمتها أستاذ الجامعة الذي يستطيع بإمكاناته العلمية والخُلقية والنفسية أن يساهم مساهمةً فعالةً في تحقيق الأهداف المنشودة من الجامعة . [1]

وقد درج مؤخرا استخدام تنقيب البيانات في تقييم الأداء الأكاديمي للأستاذ الجامعي وذلك عبر تقديم نماذج مختلفة لتقييم الأداء و ذلك باستخدام خوارزميات التعلم الآلي المختلفة .[2] ومن هنا جاءت الدراسة الحالية لتقييم الأداء الأكاديمي لأستاذ الجامعة وما ينبغي أن يكون عليه في جامعة النيلين – كلية علوم الحاسوب و تقانة المعلومات– من وجهة نظر الطلاب؛ ممَّا يتيح الفرصة لأستاذ الجامعة وإدارتها للتعرف على نواحي القوة وتشجيعها، والوقوف على نواحي الضعف لعلاجها.

وجاءت فكرة الدراسة بسبب قلـــة أو انعدام الدراسات التي أجريت في مجال تقييم الأداء الأكاديمي

لأستاذ الجامعة وما ينبغي أن يكون عليه في السودان عامة، وجامعة النيلين بشكل خاص.
وتميزت الدراسة بقدرتها على معالجة عدم توازن البيانات و تقديم نموذج قادر على أعطاء نتائج دقيقة لتقييم الأداء الأكاديمي للأستاذ الجامعي حتى في ظل عدم توفر مجموعة بيانات كبيرة، وبالتالي يمكن اعتماده كإطار عام و قابل للتطبيق و الاستخدام في جامعات أخرى

2. تنقيب البيانات التعليمية Educational Data Mining: [9]

مع إنشاء المؤتمر الدولي السنوي لاستخراج البيانات التعليمية ومجلة تعدين البيانات التعليمية في عام 2008، برزت EDM كمجال بحثي موثوق به (.Baker et al، 2010). تقدم الجمعية الدولية لتعدين البيانات التعليمية، التي تستضيف المؤتمر الدولي لتعدين البيانات التعليمية وتنشر مجلة تعدين البيانات التعليمية ، هذا التعريف لـــ EDM:

"التنقيب في البيانات التعليمية هو تخصص ناشئ، يهتم بتطوير طرق لاستكشاف البيانات الفريدة والواسعة النطاق المتزايدة التي يتم الحصول عليها من الإعدادات التعليمية، ويستخدم هذه الأساليب لفهم الطلاب بشكل أفضل والإعدادات التي يتعلمون فيها"

وفقًا لجمعية التنقيب عن البيانات التعليمية الدولية (2011) ، غالبًا ما تتكون المعلومات في أي سياق تعليمي من مستويات هرمية متعددة ، والتي لا يمكن تحديدها مسبقًا ولكن يجب التحقق منها من خلال الخصائص الموجودة في البيانات. تعتبر العوامل ، مثل الوقت والتسلسل والسياق مهمة أيضًا في الاعتبار عند دراسة البيانات التعليمية. على سبيل المثال ، يمكن تحليل السلوكيات التعليمية للطلاب (مشاركة الطلاب، وتكرار تسجيل الدخول، وعدد رسائل الدردشة، ونوع الأسئلة المرسلة إلى المعلم) جنبًا إلى جنب مع درجاتهم النهائية.

3. البيانات الغير متوازنة Imbalanced Data :

أحد التحديات المهمة في التنقيب عن البيانات هو التعامل مع البيانات غير المتوازنة في التصنيف. ، يظهر عدم التوازن عندما يتم توزيع البيانات بشكل غير متساوٍ إلى فصائل ؛ قد تحتوي بعض الفصائل على كمية كبيرة من البيانات تسمى فصائل الأغلبية Majority class وبعضها قد يحتوي على حالات قليلة فقط من البيانات تسمى فئات الأقليات Minority class . يتسبب هذا التوزيع غير المتكافئ في أداء متحيز للمصنفات التقليدية لأنها تأخذ في الاعتبار معدل الخطأ وليس توزيع البيانات ، وبسبب قلة عدد حالات البيانات ، يتم تجاهل فئات الأقليات في نتيجة التصنيف الإجمالية. تظهر هذه المشكلة في العديد من تطبيقات العالم الحقيقي ، مثل قطاع الرعاية الصحية ، و اكتشاف الانسكاب النفطي ، و اكتشاف

الاحتيال في استخدام بطاقات الائتمان ، و نمذجة للثقافات ، و اكتشاف التسلل في الشبكات ، وتصنيف النصوص ، وما إلى ذلك. [17]

4. تقنية Stratified K-fold Cross-Validation : [24]

في التعلم الآلي ، عندما نريد تدريب النموذج ، نقوم بتقسيم مجموعة البيانات إلى مجموعتان واحدة للتدريب و الاخرى للاختبار ، ثم ندرب النموذج على مجموعة التدريب ونختبره على مجموعة الاختبار،المشاكل التي تحدث بهذه الطريقة هي أنه كلما تم تغيير معلمة random_state الموجودة في (train_test_split) ، نحصل على دقة مختلفة لحالة عشوائية مختلفة ، وبالتالي لا يمكننا تحديد دقة النموذج بالضبط.

لحل مشكلة العينة العشوائية و التي قد تؤدي الى بيانات غير متوازنة أو توزيع غير متوازن بين الفصائل تم استخدام Stratified K-fold Cross-Validation ، و على غرار أداة التحقق المتقاطعة K_Fold ، تقوم StratifiedKfold بإرجاع الطيات الطبقية (folds) ، أي أثناء عمل الطيات ، فإنها تحافظ على النسبة المئوية للعينات لكل فصيل في كل طية. بحيث يحصل هذا النموذج على بيانات موزعة بالتساوي من أجل طيات التدريب / الاختبار.

5. **ما هي** SMOTE (Synthetic Minority Oversampling Technique) ؟ [10]

كما يوحي الاسم ، SMOTE هي تقنية لأخذ عينات مفرطة. بعبارة أخرى ، ستنشئ SMOTE نقاط بيانات تركيبية لفصيل الأقلية. أي أنها ستخلق حالات جديدة بين نقاط طبقة الأقلية.

يعمل SMOTE من خلال استخدام خوارزمية k–الأقرب لإنشاء بيانات تركيبية. تبدأ SMOTE أولاً باختيار بيانات عشوائية من فئة الأقلية ، ثم يتم تعيين k– أقرب جيران من البيانات. سيتم بعد ذلك إجراء البيانات الاصطناعية بين البيانات العشوائية والجار الأقرب k المختار عشوائيًا.

يتم تكرار الإجراء مرات كافية حتى تحصل فئة الأقلية على نفس نسبة فئة الأغلبية.

6-1. خطوات خوارزمية SMOTE : [16]

- الخطوة 1: تعيين فئة فئة الأقلية A ، لكل x € A ، يتم الحصول على أقرب جيران k لـــ x عن طريق حساب المسافة الإقليدية بين x وكل عينة أخرى في المجموعة A.
- الخطوة 2: يتم ضبط معدل أخذ العينات N وفقًا للنسبة غير المتوازنة. لكل x € A ، يتم اختيار أمثلة N (أي x1، x2، xn ...) عشوائيًا من جيرانها الأقرب لـــ k ، وهذه تشكل المجموعة A1.

- الخطوة 3: لكل مثال (k = 1 xk € A1، 2، N ... 3) ، يتم استخدام الصيغة التالية لإنشاء مثال جديد: (0 x' = x + rand، x_k - x | * (1 | حيث يمثل (0 rand، 1) الرقم العشوائي بين 0 و 1.

6. خوارزمية Random Forest Classifier : [13]

كيف تعمل خوارزمية الغابة العشوائية ؟

فيما يلي الخطوات الأساسية المتضمنة في تنفيذ خوارزمية الغابة العشوائية:

1. اختر N سجلات عشوائية من مجموعة البيانات.
2. قم ببناء شجرة قرار بناءً على هذه السجلات N.
3. اختر عدد الأشجار التي تريدها في الخوارزمية وكرر الخطوتين 1 و 2.

4.في حالة وجود مشكلة انحدار ، بالنسبة لسجل جديد ، تتنبأ كل شجرة في الغابة بقيمة Y (الإخراج). يمكن حساب القيمة النهائية بأخذ متوسط جميع القيم التي تنبأت بها جميع الأشجار في الغابة. أو ، في حالة وجود مشكلة تصنيف ، تتنبأ كل شجرة في الغابة بالفئة التي ينتمي إليها السجل الجديد

7. خوارزمية Decision Tree [14]

شجرة القرار هي تقنية تعليمية خاضعة للإشراف يمكن استخدامها لكل من مشاكل التصنيف والانحدار ، ولكنها في الغالب مفضلة لحل مشاكل التصنيف.

تمثل العقد الداخلية سمات مجموعة البيانات ، وتمثل الفروع قواعد القرار وتمثل كل عقدة ورقية النتيجة. في شجرة القرار ، توجد عقدتان ، وهما عقدة القرار والعقدة الورقية. تُستخدم عقد القرار لاتخاذ أي قرار ولها فروع متعددة ، في حين أن العقد الورقية هي نتاج تلك القرارات ولا تحتوي على أي فروع أخرى. يتم تنفيذ القرارات أو الاختبار على أساس ميزات مجموعة البيانات المحددة.

لماذا نستخدم أشجار القرار؟

فيما يلي سببان لاستخدام شجرة القرار:

1. عادةً ما تحاكي أشجار القرار قدرة التفكير البشري أثناء اتخاذ القرار ، لذلك من السهل فهمها.
2. يمكن فهم المنطق وراء شجرة القرار بسهولة لأنها تظهر بنية شبيهة بالشجرة.

مصطلحات شجرة القرار :

عقدة الجذر Root Node**:** عقدة الجذر هي من حيث تبدأ شجرة القرار. يمثل مجموعة البيانات بأكملها ، والتي يتم تقسيمها إلى مجموعتين أو أكثر من المجموعات المتجانسة.

العقدة الورقية Leaf Node**:** العقد الورقية هي عقدة الإخراج النهائية ، ولا يمكن فصل الشجرة أكثر بعد الحصول على عقدة ورقية.

التقسيم Splitting**:** التقسيم هو عملية تقسيم عقدة القرار / العقدة الجذرية إلى عقد فرعية وفقًا للشروط المحددة.

الفرع / الشجرة الفرعية Branch/Sub Tree**:** الشجرة التي تكونت بتقسيم الشجرة.

التقليم Pruning**:** التقليم هو عملية إزالة الفروع غير المرغوب فيها من الشجرة.

العقدة الأصلية / الفرعية Parent/Child node**:** تسمى العقدة الجذرية للشجرة العقدة الأصلية ، وتسمى العقد الأخرى بالعقد الفرعية.

كيف تعمل خوارزمية شجرة القرار؟

في شجرة القرار ، للتنبؤ بفئة مجموعة البيانات المحددة ، تبدأ الخوارزمية من العقدة الجذرية للشجرة. تقارن هذه الخوارزمية قيم سمة الجذر بسمة السجل (مجموعة البيانات الحقيقية) ، وبناءً على المقارنة ، تتبع الفرع وتنتقل إلى العقدة التالية.

بالنسبة للعقدة التالية ، تقارن الخوارزمية مرة أخرى قيمة السمة مع العقد الفرعية الأخرى وتتحرك أكثر. تستمر العملية حتى تصل إلى العقدة الورقية للشجرة. يمكن فهم العملية الكاملة بشكل أفضل باستخدام الخوارزمية أدناه:

الخطوة 1: ابدأ الشجرة بالعقدة الجذرية ، لنرمز لها بالرمز S، والتي تحتوي على مجموعة البيانات الكاملة.

الخطوة 2: ابحث عن أفضل سمة في مجموعة البيانات باستخدام مقياس تحديد السمة Attribute Selection Measures(ASM)

الخطوة 3: قسّم S إلى مجموعات فرعية تحتوي على القيم الممكنة لأفضل السمات.

الخطوة 4: إنشاء عقدة شجرة القرار ، والتي تحتوي على أفضل سمة.

الخطوة 5: إنشاء أشجار قرارات جديدة بشكل متكرر باستخدام مجموعات فرعية من مجموعة البيانات التي تم إنشاؤها في الخطوة 3. استمر في هذه العملية حتى يتم الوصول إلى مرحلة حيث لا يمكنك تصنيف العقد بشكل أكبر وتسمى العقدة النهائية كعقدة طرفية.

8. **خوارزمية** Naive Bayes

Naive Bayes هي تقنية تصنيف إحصائي تعتمد على نظرية بايز. إنها واحدة من أبسط خوارزميات التعلم الخاضع للإشراف. مصنف Naive Bayes هو خوارزمية سريعة ودقيقة وموثوقة. تتميز مصنفات Naive Bayes بدقة عالية وسرعة في مجموعات البيانات الكبيرة. [15]

يعمل مصنف Naive Bayes على الاحتمال الشرطي. الاحتمال الشرطي هو احتمال حدوث شيء ما ، بالنظر إلى أن شيئًا آخر قد حدث بالفعل. باستخدام الاحتمال الشرطي ، يمكننا حساب احتمال حدث باستخدام معرفته السابقة. [18]

يوجد أدناه معادلة حساب الاحتمال الشرطي.

$$P(H|E) = \frac{P(E|H)*P(H)}{P(E)}$$

معادلة حساب الاحتمال الشرطي لخوارزمية Naive Bayes [18]

حيث أن :

P (H) هو احتمال أن تكون الفرضية H صحيحة. يُعرف هذا بالاحتمال السابق.

P (E) هو احتمال الدليل (بغض النظر عن الفرضية).

P (E | H) هو احتمال وجود الدليل بالنظر إلى صحة الفرضية.

P (H | E) هو احتمال الفرضية بالنظر إلى وجود الدليل.

9. مصفوفة الارتباك Confusion Matrix [19]

مصفوفة الارتباك هي مصفوفة N x N تُستخدم لتقييم أداء نموذج التصنيف ، حيث N هو عدد الفئات المستهدفة. تقارن المصفوفة القيم المستهدفة الفعلية بتلك التي تنبأ بها نموذج التعلم الآلي. يمنحنا هذا نظرة شاملة عن مدى جودة أداء نموذج التصنيف الخاص بنا وأنواع الأخطاء التي يرتكبها.

يتم حساب دقة النموذج بناء على المعادلات الآتية : [20]

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{Correct\ Predictions}{Total\ Predictions}$$

معادلة حساب دقة النموذج [20]

ما هو TP و FP و FN و TN هنا؟ هذا هو الجزء الأساسي من مصفوفة الارتباك [21]

إيجابي حقيقي (TP)

القيمة المتوقعة تطابق القيمة الفعلية

كانت القيمة الفعلية موجبة وتنبأ النموذج بقيمة موجبة

سلبي حقيقي (TN)

القيمة المتوقعة تطابق القيمة الفعلية

كانت القيمة الفعلية سالبة وتنبأ النموذج بقيمة سالبة

موجب كاذب (FP) – **خطأ من النوع الأول**

تم التنبؤ بالقيمة المتوقعة بشكل خاطئ كانت القيمة الفعلية سالبة لكن النموذج تنبأ بقيمة موجبة

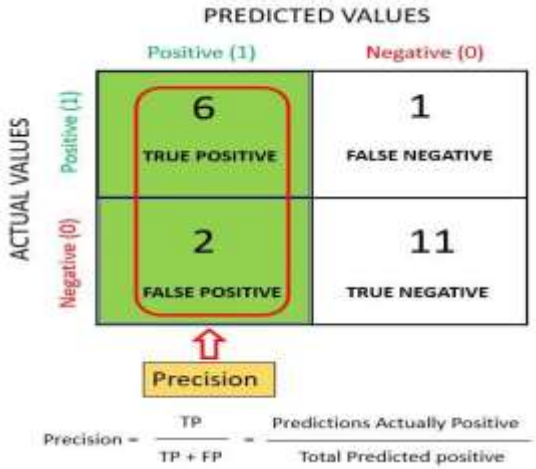
سلبي كاذب (FN) – **خطأ من النوع 2**

تم التنبؤ بالقيمة المتوقعة بشكل خاطئ

كانت القيمة الفعلية موجبة لكن النموذج تنبأ بقيمة سالبة

تخبرنا precision عن عدد الحالات الايجابية التي تم التنبؤ بها بشكل صحيح . أي أنها النسبة بين الإيجابيات الحقيقية التي تم التنبؤ بها و بين كل الإيجابيات التي تم التنبؤ بها. [22] أي أنها تجاوب عن السؤال التالي : كم من عدد الأساتذة الذين تم تصنيفهم من قبل الطلاب على أن مستوى أداءهم جيد أو وسط أو منخفض هم بالفعل كذلك ؟ ، "الدقة هي مقياس مفيد في الحالات التي يكون فيها الإيجابي الكاذب مصدر قلق أكبر من السلبيات الكاذبة" . [23]

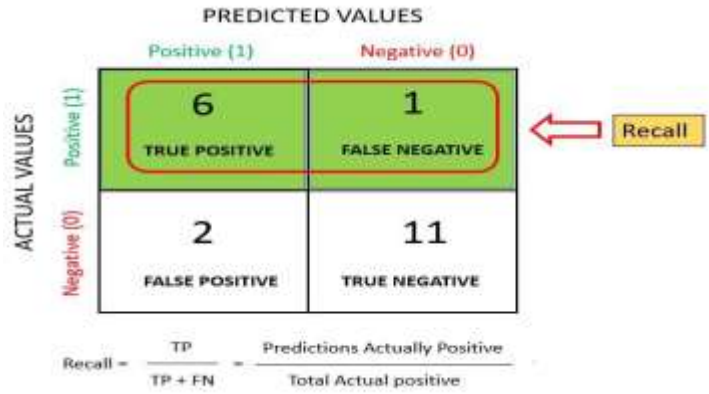
والشكل (1) يوضح طريقة حساب الـــ Precision :



الشكل (1) : طريقة حساب الـــ Precision [23]

تخبرنا Recall عن عدد الحالات الإيجابية الفعلية التي تمكن النموذج من التنبؤ بها بشكل صحيح. أي أنه هو مقياس النموذج الذي يحدد الإيجابيات الحقيقية بشكل صحيح ، و نلجأ اليه عندما يكون أهمية السلبية الكاذبة عالية أي أن يخبرنا النموذج أن أداء الأستاذ منخفض و هو ليس كذلك .

الشكل (2) يوضح طريقة حساب Recall :



الشكل (2) : طريقة حساب Recall [23]

F1 score :

F1-Score هي الوسيلة التوافقية للدقة والاستدعاء ، ولذا فهي تعطي فكرة مجمعة عن هذين المقياسين. وتعطي نفس الوزن لكل من الـــ Recall و precision .

$$\boldsymbol{F1\ score} = \frac{\mathbf{2}}{\frac{\mathbf{1}}{\boldsymbol{Precision}} + \frac{\mathbf{1}}{\boldsymbol{Recall}}} = \frac{\mathbf{2} * (\boldsymbol{Precision} * \boldsymbol{Recall})}{(\boldsymbol{Precision} + \boldsymbol{Recall})}$$

معادلة حساب f1 score [23]

Macro average

هو متوسط f1 score / recall / precision

$$\boldsymbol{Macro\ avg\ Precion} = \frac{\boldsymbol{P1} + \boldsymbol{P2}}{\mathbf{2}}$$

معادلة حساب متوسط f1 score / recall / precision [23]

حيث أن P1 و P2 هي التنبؤ الخاص بكلا الفصيلين

10. الدراسات السابقة :

1. في هذه الدراسة [2] تم استخدام نموذج التصنيف classification model للتنبؤ بأداء الأساتذة وذلك بالاعتماد على بيانات من الطلاب و الزملاء ، كما تم تعريف العوامل التي تؤثر على أداء الأستاذ. استخدم الباحث خوارزمية : J48 decision tree(random tree) and naïve bayes . بإستخدام أداة ويكا weka . وتم التوصل الى أن رأي الطالب وحده غير كافي لتقييم الأستاذ لذلك لا بد من استخدام بيانات أخرى مثل بيانات الزملاء ورؤساء الأقسام .
2. تهدف هذه الدراسة [3] للتنبؤ بجودة أداء الاستاذ الجامعي و تحليل العوامل التي تؤثر على إستفادة الطلاب لتحسين جودة النظام التعليمي . تم العمل علي بيانات طلاب في جامعة تركية واستخدم الباحث خوارزميات التصنيف : J48 Decision tree, Multilayer Perception, Naïve Bayes, and Sequential Minimal Optimization . وبعد مقارنة الخوارزميات الأربعة اظهرت الدراسة مدى أهمية استخدام تنقيب البيانات في مجال التعليم ، ايضا توصلت الدراسة الى أن استخدام بيانات الطالب لتقييم المقررات مفيد في التنبؤ بالعوامل التي تؤثر على تحصيله وعلى التنبؤ بجودة أداء الأستاذ .
3. في هذه الدراسة [4] تم انشاء نظام تصنيف يعتمد على طبقتين: الشبكة العصبية Artificial Neural Network (ANN) و شجرة القرار Decision Tree . تم اختبار النظام بنجاح باستخدام بيانات دراسة حالة من الجامعة النيجيرية . البيانات هي: المؤهلات الاكاديمية لأهضاء هيئة التدريس وكذلك الخبرة درجات الطلاب في المواد التي يدرسونها . اظهرت النتائج انه من بين الست سمات المستخدمة تم تصنيف الخبرة و المؤهل كأفضل سمتين ساهمتا في اداء أعضاء هيئة التدريس . ايضا بالاخد في الاعتبار الزمن المستغرق في بناء النماذج و مستوى دقة اداء شجرة القرار C4.5 تفوقت على الخوارزميتين MLP و ID3 .
4. تبحث هذه الدراسة [5] في العوامل المرتبطة بتقييم أداء الأساتذة باستخدام تقنيتين : الانحدار التدريجي stepwise regression و أشجار القرار decision trees . تم جمع البيانات من تقييمات الطلاب بالإضافة إلى ذلك تم تضمين بعض المتغيرات الأخرى . أشارت النتائج إلى أن حالة توظيف الأستاذ ، و عبء العمل ، وحضور الطلاب و مستوى الطلاب الأكاديمي هي أبعاد مهمة في تقييم الأستاذ .
5. حاولت هذه الورقة [6] الإجابة على أسئلة مثل : هل لدى الطلاب النضج و المعرفة الكافية لتقديم ملاحظات مفيدة يمكن الاعتماد عليها لتحسين القدرات التدريسية للأساتذة ؟ ، هل صعوبة المقرر

لديه علاقة قوية مع التصنيف الذي يعطيه الطالب للأستاذ؟ . تم استخدام بعض تقنيات تنقيب البيانات الإحصائية الحديثة مثل :

Support vector machines, classification and regression trees, boosting, random forest, factor analysis, k-Means clustering. Hierarchical clustering.

تم استخدام جوانب مختلفة من البيانات من منظورين مختلفين ، التعلم الخاضع للإشراف و الغير خاضع للإشراف supervised and unsupervised learning. البيانات التي تم تحليلها في هذه الورقة تم جمعها من جامعة تركية ، تم اكتشاف بعض الانماط مثل الارتباط القوي بين جدية الطلاب (تقاس بالحضور) و بين نوع الدرجات التي يمنحونها لاساتذتهم.

6. في هذه الورقة [7] تم استخدام 4 خوارزميات تصنيف : decision tree algorithms, support vector machines, artificial neural networks, and discriminant analysis.

تم مقارنة اداءها بالتطبيق على بيانات تم الحصول عليها بواسطة استبيان الطلاب لتقييم المقرر الدراسي . أظهرت خوارزمية C5.0 نتائج أكثر دقة مقارنة بالخوارزميات الأخرى ، و بتحليل متغيرات كل نموذج تصنيفي تم التوصل الى انه العديد من اسئلة الاستبيان ليست ذات صلة بمشكلة الدراسة ، علاوة على ذلك اظهرت النتائج ان نجاح الاساتذة يعتمد على اهتمام الطلاب باالمقرر وعليه يمكن استخدام النتائج لتحسين ادوات القياس.

باستقراء الدراسات السابقة يلاحَظ أنَّ العديد من الدراسات تناولت موضوع أستاذ الجامعة من جوانب متعددة ومختلفة؛ لأهمية أستاذ الجامعة، ودوره في خدمة الجامعة والمجتمع. وقد اختلفت هذه الدراسات وفقًا لطبيعتها، وأهدافها، والعينة المختارة، والنظام التعليمي، والمجتمعات التي أجريت فيها. والدراسة الحالية تتفق مع الدراسات السابقة من حيث موضوعها –أستاذ الجامعة –، ومع ذلك فإنَّ الدراسة الحالية تكتسب وضعًا مختلفًا في معالجتها لتقويم الأداء الأكاديمي لأستاذ الجامعة وفقا لعدد من المتغيرات المتمثلة في فقرات الاستبيان الذي يمثل إحدى ادوات جمع بيانات الدراسة ، كما أن الدراسة الحالية تميزت بقدرتها على ايجاد حلول للتعامل مع البيانات الغير متوازنة وكان ذلك واضحا في تحسن أداء النموذج .

11. المنهجية :

12-1. مرحلة جمع البيانات:

تضمنت هذه المرحلة جمع البيانات بواسطة الاستبيان وبعض البيانات الأخرى الخاصة بالاستاذ مثل المؤهل العلمي، سنوات الخبرة و المعدل التراكمي و التي تم الحصول عليها من ملفات شئون العاملين و الشئون العلمية بالجامعة .

كان حجم مجموعة البيانات 709 صف . جمعت اجابات الطلاب من 6 اقسام (علوم الحاسوب، تقنية المعلومات ، هندسة البرمجيات، نظم المعلومات الادارية، نظم المعلومات المحاسبية، نظم معلومات المكتبات) في كلية علوم الحاسوب وتقنية المعلومات بجامعة النيلين.

12-2. مرحلة تنظيف البيانات : Data Cleaning

بعد أن قام الباحث بدراسة مجموعة البيانات و اجراء بعض المقارنات الاحصائية بين السمات قرر استبعاد السمات : الجامعة ، الكلية ، القسم ، الرقم الجامعي ، اسم الطالب ، اسم المقرر ، اسم أستاذ المقرر و المعدل التراكمي وذلك للاسباب التالية :

1. الجامعة، الكلية، القسم واسم المقرر واسم الاستاذ سمات غير مؤثرة في النتائج.
2. المعدل التراكمي انحصرت أغلب القيم بين 72 و 74 أي أن المدى بسيط نوعا ما .
3. الرقم الجامعي واسم الطالب تحتوي على قيم شاغرة كثيرة.

12-3. طريقة تقسيم البيانات :

بعد مرحلة تنظيف البيانات أصبح حجم البيانات هو 28 سمة و 709 صف ، تم تقسيم البيانات الي قسمين : للتدريب و الاختبار بنسبة 80% و 20% على التوالي فاصبح الجزء المخصص للتدريب هو عبارة عن 567 صف و الباقي للاختبار و هو 142 صف .

تم اختيار السمة Std_satisfaction و التي تعبر عن رضاء الطالب العام عن أداء الاستاذ، تم اختيارها كمتغير هدف في عملية التصنيف ، عدد الفصائل/خيارات هذه السمة هي 0،1،2 حيث أن 0 تعني أن درجة الرضاء منخفضة و 1 تعني أنها متوسطة و 2 تعني أنها جيدة .

الشكل (5) يوضح توزيع البيانات على الفصائل و كما نلاحظ أن انحياز أغلب الاجابات الى الفصيل (2) و ذلك بقيمة 335 من أصل 709 ، يليها الاداء الوسط (الفصيل (1)) ثم اخيرا الاداء الضعيف (الفصيل (0)) . هذا الانحياز يوضح أن البيانات غير متوازنة imbalanced مما سيؤثر لاحقا في نتائج النموذج . لذلك ولمزيد من التأكد من أداء النموذج و لتحسين دقة خوارزميات التصنيف الثلاث، استخدمت تقنية SMOTE .

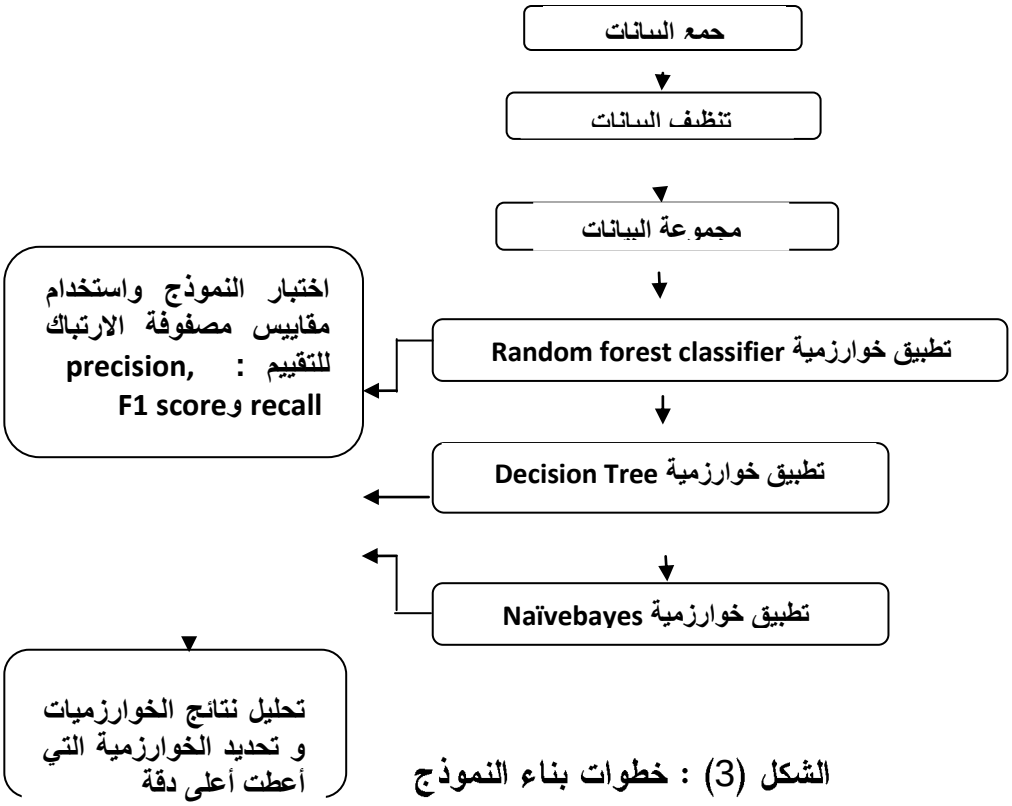
```
1  y_train.value_counts()
```

```
2    335
1    149
0     83
Name: Std_satisfaction, dtype: int64
```

الشكل (5) : توزيع البيانات على الفصائل

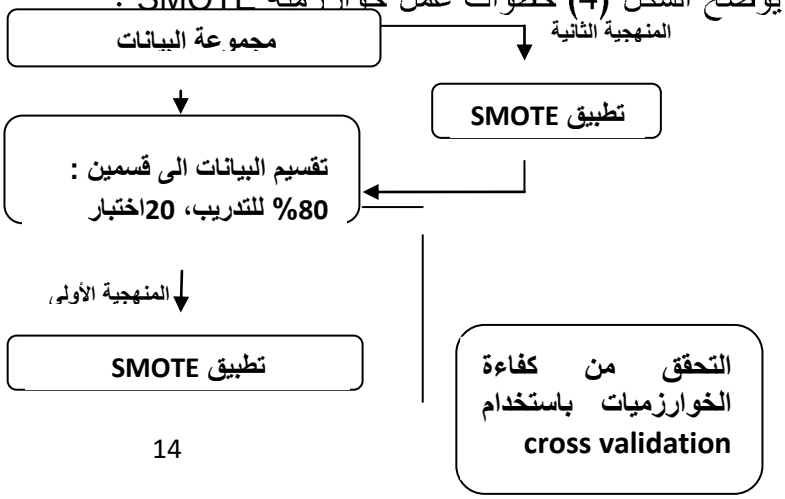
12. النموذج :

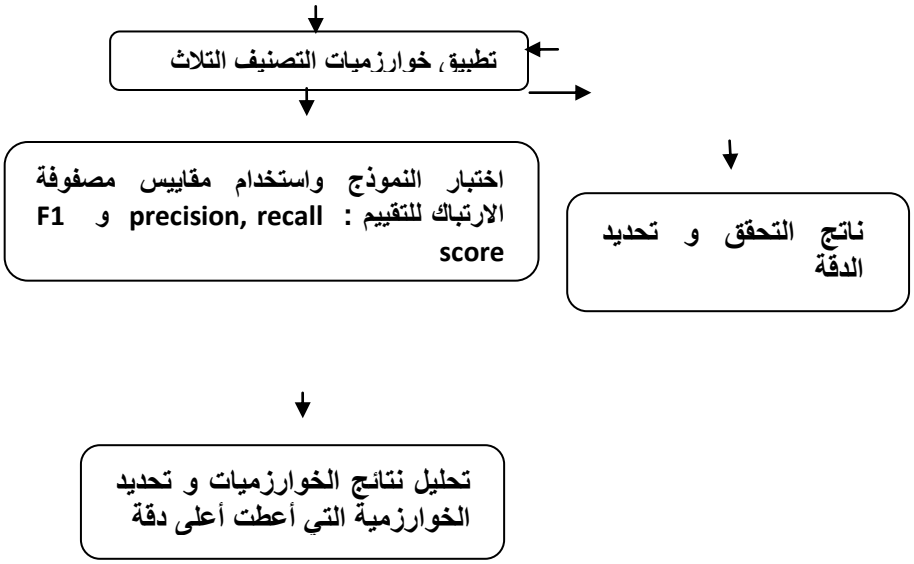
كما هو ملاحظ في الشكل (3) فان خطوات النموذج موضحة بداية بجمع البيانات وانتهاء بنتيجة التقييم



الشكل (3) : خطوات بناء النموذج

يوضح الشكل (4) خطوات عمل خوارزمية SMOTE :





الشكل (4) خطوات خوارزمية SMOTE

تطبيق خوارزميات التصنيف الثلاث :

Random Forest Classifier:

تم استخدام خوارزمية الغابة العشوائية لتدريب النموذج على بيانات التدريب و كانت دقة التعلم تساوي 0.97 .

و عند اختبار النموذج و من ثم استخدام مصفوفة الارتباك لتقييم النموذج كانت الدقة تساوي 0.68 ، كما موضح في الشكل (6)

```
print (metrics.classification_report(y_test,y_pred_rd))
```

	precision	recall	f1-score	support
0	0.45	0.60	0.51	15
1	0.61	0.49	0.54	51
2	0.78	0.83	0.80	76
accuracy			0.68	142
macro avg	0.61	0.64	0.62	142
weighted avg	0.68	0.68	0.68	142

الشكل (6) : تقييم خوارزمية Random Forest **بعد اختيار السمات**

خوارزمية Decision Tree:

تعلم هذه النموذج بدقة تساوي 0.1 و عند اختبار النموذج و من ثم استخدام مصفوفة الارتباك لتقييم النموذج كانت الدقة تساوي 0.62 ، كما موضح في الشكل (7)

```
1 print (metrics.classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
0	0.45	0.67	0.54	15
1	0.48	0.41	0.44	51
2	0.75	0.75	0.75	76
accuracy			0.62	142
macro avg	0.56	0.61	0.58	142
weighted avg	0.62	0.62	0.62	142

الشكل (7) : تقييم خوارزمية Decision Tree بعد اختيار السمات

خوارزمية Naive Bayes :

تعلم هذه النموذج بدقة تساوي 0.78 و عند اختبار النموذج و من ثم استخدام مصفوفة الارتباك لتقييم النموذج كانت الدقة تساوي 0.65 ، كما موضح في الشكل (8)

```
1 print (metrics.classification_report(y_test,y_pred))
```

	precision	recall	f1-score	support
0	0.55	0.73	0.63	15
1	0.58	0.22	0.31	51
2	0.68	0.92	0.78	76
accuracy			0.65	142
macro avg	0.60	0.62	0.57	142
weighted avg	0.63	0.65	0.60	142

الشكل (8) : تقييم خوارزمية Naive Bayes بعد اختيار السمات

تقييم النتائج :

كانت نتائج الخوارزميات الثلاث كما موضح في الجدول التالي:

الجدول (1) : ملخص نتائج الخوارزميات

Algorithm	classes	Micro avg/ accuracy	precision	recall	
Random Forest	0	0.68	0.45	0.60	0.51
	1		0.61	0.49	0.54
	2		0.78	0.83	0.80

		Macro avg	0.61	0.64	0.62
Decision Tree	0	0.62	0.45	0.67	0.45
	1		0.48	0.41	0.44
	2		0.75	0.75	0.75
		Macro avg	0.56	0.61	0.58
Naïve Bayes	0	0.65	0.55	0.73	0.63
	1		0.58	0.22	0.31
	2		0.68	0.92	0.78
		Macro avg	0.60	0.62	0.57

نلاحظ من الجدول (1) أن خوارزمية الغابة العشوائية أعطت معدل دقة أعلى من خوارزميتي Naïve Bayes وشجرة القرار

استخدام SMOTE لتحسين أداء النموذج :

تم تطبيق خوارزمية SMOTE بمنهجيتين :

المنهجية الأولى :

استخدمت على مجموعة البيانات كاملة ، تم تقسيم البيانات الى جزئين للتدريب و الاختبار ومن ثم تم تطبيق SMOTE على بيانات التدريب و الاختبار ، كل على حدى وذلك للتأكد أن البيانات التي تم توليدها في جزء الاختبار لا توجد نسخة منها في جزء التدريب وكانت النتائج كما يلي :

الشكل (9) و الشكل (10) يوضحان توزيع البيانات على الفصائل في كل من مجموعة بيانات التدريب و الاختبار و الرسم البياني المصاحب لهما

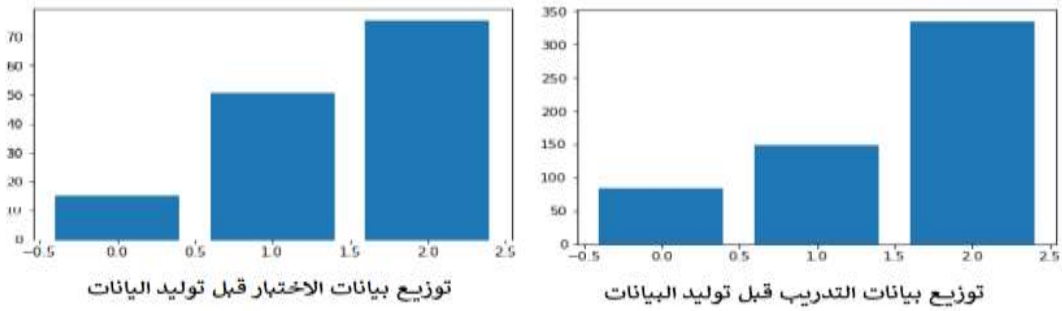
```
1  y_train.value_counts()
```

```
2    335
1    149
0     83
Name: Std_satisfaction, dtype: int64
```

```
1  y_test.value_counts()
```

```
2    76
1    51
0    15
Name: Std_satisfaction, dtype: int64
```

الشكل (9) : توزيع البيانات على الفصائل في كل من مجموعة بيانات التدريب و الاختبار

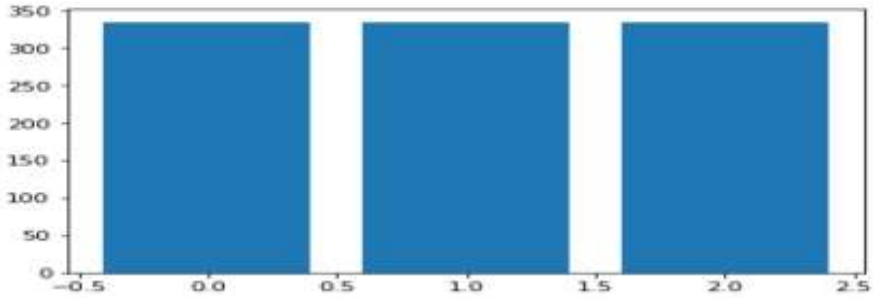


الشكل (10) : المخطط البياني لتوزيع البيانات على الفصائل في كل من مجموعة بيانات التدريب و الاختبار

قامت خوارزمية SMOTE بزيادة بيانات مجموعة التدريب في كل فصائل الاقلية لمستوى فصيل الاغلبية كما في الشكل (11)

```
Before OverSampling, counts of label '1': 149
Before OverSampling, counts of label '0': 83
Before OverSampling, counts of label '2': 335

After OverSampling, counts of label '1': 335
After OverSampling, counts of label '0': 335
After OverSampling, counts of label '2': 335
```

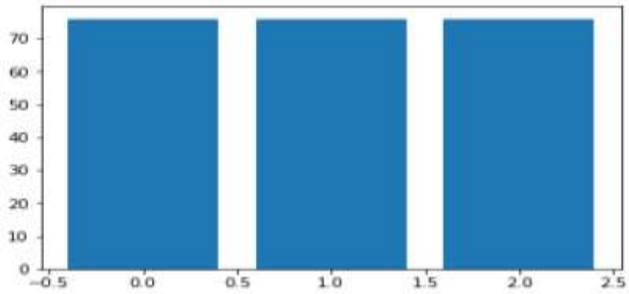


الشكل (11) : توزيع بيانات مجموعة التدريب على الفصائل بعد توليد البيانات

بعد ذلك قامت خوارزمية SMOTE بزيادة بيانات مجموعة الاختبار في كل فصائل الاقلية لمستوى فصيل الاغلبية كما في الشكل (12)

```
Before OverSampling, counts of label '1': 51
Before OverSampling, counts of label '0': 15
Before OverSampling, counts of label '2': 76

After OverSampling, counts of label '1': 76
After OverSampling, counts of label '0': 76
After OverSampling, counts of label '2': 76
```



الشكل (12) : توزيع بيانات مجموعة الاختبار على الفصائل بعد توليد البيانات

خوارزمية شجرة القرار :

الشكل (13) يوضح ناتج خوارزمية شجرة القرار بتطبيق SMOTE بعد تقسيم البيانات

```
1 print (metrics.classification_report(y_test_res,y_pred_test)
```

	precision	recall	f1-score	support
0	0.74	0.80	0.77	76
1	0.54	0.54	0.54	76
2	0.74	0.68	0.71	76
accuracy			0.68	228
macro avg	0.68	0.68	0.67	228
weighted avg	0.68	0.68	0.67	228

الشكل (13) : ناتج خوارزمية شجرة القرار بتطبيق SMOTE بعد تقسيم البيانات

خوارزمية الغابة العشوائية :

الشكل (14) يوضح ناتج خوارزمية الغابة العشوائية بتطبيق SMOTE بعد تقسيم البيانات

```
print (metrics.classification_report(y_test_res,y_pred_test))
```

	precision	recall	f1-score	support
0	0.90	0.84	0.87	76
1	0.73	0.75	0.74	76
2	0.82	0.86	0.84	76
accuracy			0.82	228
macro avg	0.82	0.82	0.82	228
weighted avg	0.82	0.82	0.82	228

الشكل (14) : ناتج خوارزمية الغابة العشوائية بتطبيق SMOTE بعد تقسيم البيانات

خوارزمية Naïve Bayes :

الشكل (15) يوضح ناتج خوارزمية Naïve Bayes بتطبيق SMOTE بعد تقسيم البيانات وقبل اختيار السمات

```
print (metrics.classification_report(y_test_res,y_pred_test))
```

	precision	recall	f1-score	support
0	0.90	0.87	0.89	76
1	0.69	0.45	0.54	76
2	0.65	0.91	0.76	76
accuracy			0.74	228
macro avg	0.75	0.74	0.73	228
weighted avg	0.75	0.74	0.73	228

الشكل (15) : ناتج خوارزمية Naïve Bayes بتطبيق SMOTE بعد تقسيم البيانات

المنهجية الثانية :

أستخدمت على مجموعة البيانات كاملة أي قبل تقسيم البيانات الى جزئين للتدريب و الاختبار حيث تم تطبيق SMOTE على كل بيانات مجموعة البيانات ثم تقسيم مجموعة البيانات ومن ثم تطبيق خوارزميات التصنيف واجراء اختبار عن طريق Stratified K-fold Cross-Validation لكن على مجموعة البيانات كاملة قبل توليد البيانات وكانت النتائج كما يلي :

خوارزمية شجرة القرار :

الشكل (16) يوضح نتائج خوارزمية شجرة القرار بعد استخدام SMOTE على مجموعة البيانات بدون تقسيمها الى مجموعة تدريب و اختبار

```
1 print (metrics.classification_report(y_test_res,y_pred_test))
```

	precision	recall	f1-score	support
0	0.89	0.88	0.88	81
1	0.76	0.79	0.78	81
2	0.84	0.82	0.83	85
accuracy			0.83	247
macro avg	0.83	0.83	0.83	247
weighted avg	0.83	0.83	0.83	247

الشكل (16) : ناتج خوارزمية شجرة القرار بعد استخدام SMOTE على مجموعة البيانات بدون تقسيمها الى مجموعة تدريب و اختبار

وعند تطبيق الـــ Stratified K-fold Cross-Validation كانت الدقة تساوي 66.72% كما في الشكل (17)

```
print("Accuracy: %.2f%%" % (results_skfold.mean()*100.0))
Accuracy: 66.72%
```

الشكل (17) : ناتج Stratified K-fold Cross-Validation على خوارزمية شجرة القرار بعد توليد بيانات لكل مجموعة البيانات بدون تقسيم

خوارزمية الغابة العشوائية :

الشكل (18) يوضح نتائج خوارزمية الغابة العشوائية بعد استخدام SMOTE على مجموعة البيانات بدون تقسيمها الى مجموعة تدريب و اختبار

```
1 print (metrics.classification_report(y_test_res,y_pred_test))
```

	precision	recall	f1-score	support
0	0.96	0.91	0.94	81
1	0.88	0.85	0.87	81
2	0.86	0.93	0.89	85
accuracy			0.90	247
macro avg	0.90	0.90	0.90	247
weighted avg	0.90	0.90	0.90	247

الشكل (18) : ناتج خوارزمية الغابة العشوائية بعد استخدام SMOTE على مجموعة البيانات بدون تقسيمها الى مجموعة تدريب و اختبار

وعند تطبيق الـــ Stratified K-fold Cross-Validation كانت الدقة تساوي 78.70% كما في الشكل (19)

```
print("Accuracy: %.2f%%" % (results_skfold.mean()*100.0))
Accuracy: 78.70%
```

الشكل (19) : ناتج Stratified K-fold Cross-Validation على خوارزمية الغابة العشوائية بعد توليد بيانات لكل مجموعة البيانات بدون تقسيم

خوارزمية Naïve Bayes :

الشكل (20) يوضح نتائج خوارزمية Naïve Bayes بعد استخدام SMOTE على مجموعة البيانات بدون تقسيمها الى مجموعة تدريب و اختبار

```
print (metrics.classification_report(y_test_res,y_pred_test))
```

	precision	recall	f1-score	support
0	0.77	0.84	0.80	81
1	0.73	0.51	0.60	81
2	0.76	0.92	0.83	85
accuracy			0.76	247
macro avg	0.75	0.75	0.74	247
weighted avg	0.75	0.76	0.75	247

الشكل (20) : ناتج خوارزمية Naïve Bayes بعد استخدام SMOTE على مجموعة البيانات بدون تقسيمها الى مجموعة تدريب و اختبار

وعند تطبيق الـــ Stratified K-fold Cross-Validation كانت الدقة تساوي 72.78% كما في الشكل (21)

```
print("Accuracy: %.2f%%" % (results_skfold.mean()*100.0))
Accuracy: 72.78%
```

الشكل (21) : ناتج Stratified K-fold Cross-Validation على خوارزمية Naïve Bayes بعد توليد بيانات لكل مجموعة البيانات بدون تقسيم

ملخص نتائج الخوارزميات بعد استخدام SMOTE :

الجدول (2) يوضح ملخص نتائج تطبيق طريقة SMOTE بالمنهجيتين التي اتبعهما الباحث :

الجدول (2) يوضح ملخص نتائج تطبيق طريقة SMOTE بالثلاث طرق

The Way	Algorithm	classes	Micro avg/ accuracy	precision	recall	F1-score	Accuracy after validation
First way	Random Forest	0	0.82	0.90	0.84	0.87	No validation
		1		0.73	0.75	0.74	

		2		0.82	0.86	0.84	occure
	Macro avg			0.82	0.82	0.82	
	Decision Tree	0	0.68	0.74	0.80	0.77	No validation occure
		1		0.54	0.54	0.54	
		2		0.74	0.68	0.71	
	Macro avg			0.68	0.68	0.67	
	Naïve Bayes	0	0.74	0.90	0.87	0.89	No validation occure
		1		0.69	0.45	0.54	
		2		0.65	0.91	0.76	
	Macro avg			0.75	0.74	0.73	
The Way	Algorithm	classes	Micro avg/ accuracy	precision	recall	F1-score	Accuracy after validation
Second way	Random Forest	0	0.90	0.96	0.91	0.94	78.70%
		1		0.88	0.85	0.87	
		2		0.86	0.93	0.89	
	Macro avg			0.90	0.90	0.90	
	Decision Tree	0	0.83	0.89	0.88	0.88	66.72%
		1		0.76	0.79	0.78	
		2		0.84	0.82	0.83	
	Macro avg			0.83	0.83	0.83	

	Naïve Bayes	0	0.76	0.77	0.84	0.80	72.78%
		1		0.73	0.51	0.60	
		2		0.76	0.92	0.83	
	Macro avg			0.75	0.75	0.74	

النتائج و التوصيات :

نلاحظ من الجدول (1) أن قيم recall ، precision و f1 مرتفعة فقط مع الفصيل 2 و هو فصيل الأغلبية أي النموذج لا يتنبأ بصورة جيدة مع بقية الفصائل و ذلك بسبب عدم توازن البيانات ، أما في الجدول (2) و بعد استخدام SMOTE ارتفعت القيم في كل الفصائل و بالتالي تحسن أداء النموذج .
النموذج الذي تم بناءه تحسن أداءه بعد استخدام SMOTE حيث تحسنت دقة الخوارزميات كما يلي: الغابة العشوائية من 0.68 الى 0.82 ، شجرة القرار من 0.62 الي 0.68 و Naïve Bayes من 0.65 الى 0.74 ، أيضا كانت خوارزمية الغابة العشوائية هي الأفضل .

وكما هو واضح أن الفرق الكبير في دقة الخوارزميات بعد استخدام SMOTE وفقا للمنهجية الأولى، عليه يوصي الباحث باستخدام المنهجية الأولى ، كذلك نلاحظ أنه حتى بعد استخدام SMOTE ظلت قيم recall و precision هي الأقل مع الفصيل 1 ، وهذا يعني أن النموذج لم يتعلم بشكل جيد مع هذا الفصيل الذي يمثل الأداء المتوسط للأستاذ وقد يعزي ذلك الى أن بعض الطلاب لم يتعاملوا بجدية في ملئ الاستبيان واختاروا قيم وسطية لأغلب الأسئلة.
نتائج تطبيق SMOTE حسب المنهجية الثانية كانت مرتفعة لكن رأى الباحث أنه لا يعول عليها لأن استخدام SMOTE بهذه الطريقة قد يسمح للنموذج أن يقوم بعملية غش في مرحلة الاختبار و بالتالي تكون النتائج عالية ولكنها غير حقيقية ، و من الجدول (2) نلاحظ أن نتائج Stratified K-fold Cross-Validation تثبت ذلك حيث أنها أعطت قيم منخفضة مقارنة بالقيم المتحصل عليها بعد تطبيق المنهجية الثانية .
من هنا خلص الباحث الى أن افضل طريقة لتصميم و تنفيذ النموذج هو إستخدام SMOTE وفقا للمنهجية الأولى و من ثم استخدام خوارزمية الغابة العشوائية باعتبارها هي الافضل دائما حسب نتائج هذه الدراسة . و عليه يمكن أن يعول على السمات المستخدمة في مجموعة البيانات و اعتمادها كإطار عام و قابل للتطبيق و الاستخدام في جامعات أخرى و بيانات مختلفة .

كما يوصى الباحث باستخدام مجموعة بيانات أخرى من جامعة أخرى و اختبارها على النموذج قيد الدراسة ومقارنة النتائج للتأكد من دقة النموذج قيد الدراسة .

المراجع :

[1] كمال بـــــــورزق .عبد الحميد شـــــــــطة .محمد رحماني (2018) ، " نظرة الطلبة نحو أداء أستاذ البدنية والرياضية بين الواقع والمأمول" ، دراسة ميدانية في معهد علوم وتقنيات الأنشطة البدنية والرياضية ، https://www.asjp.cerist.dz/en/article/64575

[2] Debalkew Geremew & Rupali Gangrade, " Data Mining Techniques to Evaluate and Predict Teachers's Performance in Higher Education", International Journal of Innovations & Avancement in computer scince , IJIACS , (october 2017)

[3] Ahmed Mohamed Ahmeda. AhmetRizanerc. Ali HakanUlusoyc, " Using data mining to predict instructor performance" , International Conference on Application of Fuzzy Systems and Soft Computing, ICAFS, (29-30 August 2016)

[4] Sona MARD˙IKYAN, Bertan BADUR, "Analyzing Teaching Performance of Instructors Using Data Mining Techniques" ,Department of Management Information Systems, Bogazici University ,Istanbul, Turkey, (July 2011)

[5] Asanbe M.O. Osofisan A.O. William W.F, "Teachers' Performance Evaluation in Higher Educational Institution using Data Mining Technique" , University of Ibadan Ibadan, Nigeria , International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 10 – No.7 , (March 2016) , www.ijais.org

[6] Sona MARD˙IKYAN, Bertan BADUR , " Analyzing Teaching Performance of Instructors Using Data Mining Techniques" , Department of Management Information Systems, Bogazici University Istanbul, Turkey , (July 2011)

[7] Necla G¨und¨uza_ and Ernest Fokou´eb , " Pattern Discovery in Students' Evaluations of Professors A Statistical Data Mining Approach" : https://arxiv.org/abs/1501.02263 , 11/2/2020 . 3:00 PM , (JAN 2015)

[8] MUSTAFA AGAOGLU , " Predicting Instructor Performance Using Data Mining Techniques in Higher Education" , IEEE, Received May 6, 2016, accepted May 10, 2016, date of publication May 13, 2016, date of current version June 3, 2016. Digital Object Identifier 10.1109/ACCESS.2016.2568756 (June 3, 2016)

[9] “Why is Educational Data Mining important in the research?” [online], Available: https://towardsdatascience.com/why-is-educational-data-mining-important-in-the-research-e78ed1a17908 , Last Updated : 19 Jan, 2019 , Accessed: 18/3/2021 . 3:50 PM

[10] “ 5 SMOTE Techniques for Oversampling your Imbalance Data” [online], Available: https://towardsdatascience.com/5-smote-techniques-for-oversampling-your-imbalance-data-b8155bdbe2b5, Last Updated: 14 sep 2020 , Accessed: 9/4/2021 . 10:45 PM

[11] Costing, H. (1994). “Reading in Total Quality Management”, Copyright, by Har Court Brace & Company, Sandigo, New York , Global Journal of Management and Business Research: A Administration and Management , (2019)

[12] “What Are Feature Selection Techniques In Machine Learning?” [online] , Available : https://analyticsindiamag.com/what-are-feature-selection-techniques-in-machine-learning/ , Last Updated : 1 Apr, 2019, Accessed: 2021/3/5 . 12:35AM

[13] "Random Forest Algorithm with Python and Scikit-Learn" [online] , Available: https://stackabuse.com/random-forest-algorithm-with-python-and-scikit-learn/, Last Updated : 21 Mar, 2021 , Accessed: 7/3/2021. 11:53 PM

[14] "Decision Tree Classification Algorithm" [online] ,

Available: https://www.javatpoint.com/machine-larning-decision-tree-classification-algorithm , Last Updated : 2018 , Accessed: 18/3/2021 . 3:45 PM

[15] "Naive Bayes Classification using Scikit-learn" [online], Available:https://www.datacamp.com/community/tutorials/naive-bayes-scikit-learn , Last Updated : 4 Des, 2018 , Accessed: 18/3/2021. 4:45PM

[16] "Enhancing Imbalanced Dataset by Utilizing (K-NN Based SMOTE_3D Algorithm)" [online] ,

Available: https://www.peertechzpublications.com/articles/ARA-4-102.php,

Accessed: 9/4/2021 . 11:07 PM

[17] Harshita Patel. Dharmendra Singh Rajput. G Thippa Reddy1. Celestine Iwendi. Ali Kashif Bashir. Ohyun Jo ," A review on classification of imbalanced data for wireless sensor networks" ,International Journal of Distributed Sensor Networks , Received: January 06, 2020; Accepted: March 04, 2020 .

[18] https://dataaspirant.com/naive-bayes-classifier-machine-learning/, "HOW THE NAIVE BAYES CLASSIFIER WORKS IN MACHINE LEARNING" , February 6, 2017 , accessed : 28/4/2021 1:26 PM

[19] "Everything you Should Know about Confusion Matrix for Machine Learning" [online],

Available: https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machine-learning/, Last Updated : 17 Apr 2020 , Accessed: 4/3/2021 . 11:32 PM

[20] "Classification: Accuracy" [online] ,

Available: https://developers.google.com/machine-learning/crash-course/classification/accuracy , Accessed: 10/4/2021 . 2:10 AM

[21] " Understanding True Positive, True Negative, False Positive and False Negative in a Confusion Matrix " [online] ,

Available: https://www.analyticsvidhya.com/blog/2020/04/confusion-matrix-machin e-learning/ , Accessed: 10/4/2021 . 2:10 AM

[22] "Precision vs. Recall – An Intuitive Guide for Every Machine Learning Person" [online], Available : https://www.analyticsvidhya.com/blog/2020/09/precision-recall-machine-learning/ , Accessed: 10/4/2021 . 2:23 AM

[23] " What is a confusion matrix?" [online], Available : https://medium.com/analytics-vidhya/what-is-a-confusion-matrix-d1c0f8feda5, Accessed: 10/4/2021 . 3:04 AM

[24] "Why is Educational Data Mining important in the research?" [online],

Available: https://towardsdatascience.com/why-is-educational-data-mining-important-in-the-research-e78ed1a17908 , Last Updated : 19 Jan, 2019 , Accessed: 18/3/2021 . 3:50 PM