**تحديد حالة دخل الأفراد**

**(فوق أو تحت العتبة)**

وسام زودي يحيى الطرودي

نبيل غصن مصطفى الرحال

**الفهرس**

**الفهرس.......................................................................................................2**

**الملخص التجريدي.............................................................................................3**

**1.مقدمة..........................................................................................................4**

**2.أرضية العمل...................................................................................................5**

2.1 الدراسة المرجعية......................................................................................5

2.2 مجموعة المعطيات...................................................................................5

**3. التجارب.......................................................................................................7**

3.1 عمليات ما قبل المعالجة.............................................................................7

3.2 عمليات الاستكشاف..................................................................................8

3.2.1 دراسة السمات............................................................................8

3.3 الاختبار..................................................................................................9

**4. الخاتمة.......................................................................................................10**

**المراجع...........................................................................................................11**

**الملخص التجريدي**

إن معرفة عتبة دخل الأفراد ضمن الدولة وظيفة مهمة لدى المؤسسات المسؤولة. فتسطيع المؤسسات بناءً على ذلك اتخاذ قرارات صحيحة من أجل توجيه الدعم والمخصصات للأكثر حاجة من عامة الناس.

نتيجة الكم الهائل من البيانات أثناء عملية التحليل, يمكن الاستعانة بالتعلم التلقائي لحل هذه المسألة والمسائل المشابهة. وعليه, في هذا العمل قمنا باستخدام خوارزميات التعلم التلقائي على قاعدة البيانات الخاصة بدخل الأفراد ووصلنا إلى دقة أقصاها %86.

**1. مقدمة**

تعد دراسة وتحليل دخل الأفراد من أهم الوظائف التي تقوم بها مؤسسات الدولة, فهي تلعب دوراً محورياً في استنباط واستنتاج معلومات حول الوضع المادي للأشخاص ضمن الدولة, من أجل اتخاذ قرارات صحيحة تصب في مصلحة الأفراد وتقوم بإيصال الدعم المادي لمستحقيه من عامة الشعب.

نظراً للزيادة الكبيرة في أحجام البيانات, فإن التحليل والدراسة لهذه الكميات الهائلة بالطرق التقليدية بات أمر في غاية الصعوبة بل مستحيلاً في بعض الأحيان. وعليه, وجود أدوات مثل التعلم التلقائي ونظم استكشاف المعرفة قدَمت تسهيلات كبيرة بهذا الشأن. ومنه, سنقوم في هذه الورقة البحثية بالمساهمة في تسهيل هذه العملية عن طريق بناء نموذج تعلَم تلقائي للتنبؤ بدخل الأفراد في حال كان تحت أو فوق عتبة معينة (50 ألف) بناءً على عدة معايير.

**2. أرضية العمل**

2.1 الدراسة المرجعية

استخدم [1] الشبكات العصبونية مع الخوارزميات الجينية لاستخراج السمات ليصل إلى دقة %85, واستعمل [2] عدة خوارزميات منها k star حيث وصل إلى دقة %72.

ايضاً حاول [3] استعمال extra tree classifier لاستخراج أهم السمات اعتماداً على درجة أهمية كل سمة بناءً على الخرج الناتج من الخوارزمية السابقة. واستخدم خوارزمية GDC ليصل إلى دقة %88.

فيما قام [4] باستعمال عدد من خوارزميات التعلم التلقائي وأعلى دقة وصل إليها %88 باستخدام خوارزمية الغابات العشوائية.

2.2 مجموعة المعطيات

سنقوم بالعمل على مجموعة المعطيات Adult الموفرة من قبل موقع UCI1, حيث جمعت من قبل مكتب تعداد الولايات المتحدة USCB المسؤول عن جمع المعلومات الديموغرافية والاقتصادية للأفراد.

تحوي قاعدة المعطيات قرابة 49000 عينة وتتضمّن 14 سمة لتوصيف الأفراد سنذكرها في الجدول التالي:

|  |  |  |
| --- | --- | --- |
| النمط | اسم السمة | المعرف |
| رقمي | Age | 1 |
| فئوي | workclass | 2 |
| رقمي | fnlwgt | 3 |
| فئوي | education | 4 |
| رقمي | education\_num | 5 |
| فئوي | martial\_status | 6 |
| فئوي | occupation | 7 |
| فئوي | relationship | 8 |
| فئوي | race | 9 |
| فئوي | sex | 10 |
| رقمي | capital\_gain | 11 |
| رقمي | capital\_loss | 12 |
| رقمي | hours\_per\_week | 13 |
| فئوي | native\_country | 14 |

**3. التجارب**

3.1 عمليات ما قبل المعالجة

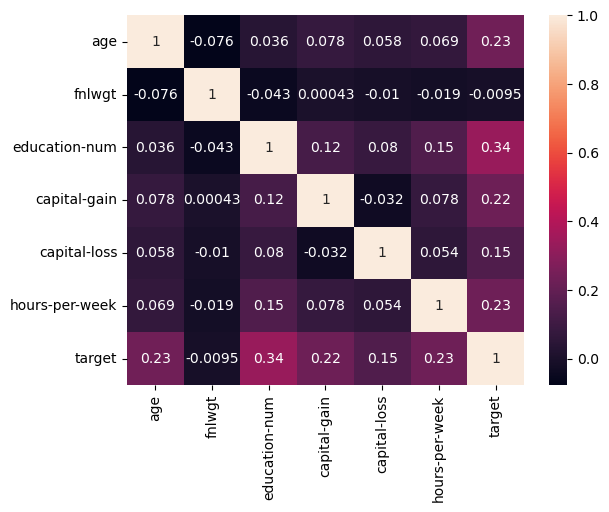
قمنا بدايةً بإزالة القيم المكررة من قاعدة المعطيات حيث تعدادها 24 قيمة. بعد ذلك قمنا بإحصاء القيم المفقودة كنسبة مئوية واستبدلناها بالقيمة الأكثر تكراراً (المنوال) للسمة الموافقة.

تقسم السمات بشكل رئيسي إلى ست سمات رقمية وثمان سمات فئوية، بالنسبة للسمات الرقمية حاولنا استكشاف ودراسة العلاقة الخطية فيما بينها، قمنا بعرض خريطة حرارية لمصفوفة الارتباط الخطي للمساعدة في معرفة أزواج السمات المترابطة مع بعضها من أجل التخلص من أحد الأزواج استناداً إلى [5].

كذلك قمنا بعملية التوحيد القياسي (تسوية درجة z) للسمات الرقمية نظراً لفائدتها للخوارزميات مثل خوارزمية الانحدار اللوجيستي (Logistic regression).

لقد تم تجريب عمل تقطيع للسمات الرقمية ولكنه تبين أنه غير مفيد لعدم توزع الداتا بشكل طبيعي ولأهمية تراتبيتها.

أما بالنسبة للسمات الفئوية تم تمثيلها كمتجهات بترميز واحد-ساخن وإزالة الترددات الضعيفة للفئات (التي تحوي أصفاراً بنسبة عالية أكثر من 99 في المئة) وحذف أول متجه ناتج عن هذا الترميز حيث يمثل في الفئات الأخرى.



الشكل-1

3.2 عمليات الاستكشاف

3.2.1 دراسة السمات

استناداً إلى الخريطة الحرارية في الشكل-1 قمنا بدراسة العلاقة بين أزواج السمات من أجل الإبقاء على السمات الهامة في عملية التدريب، ولاحظنا أن العلاقة بين الأزواج ضعيفة عموماً فلا يمكننا التخلص من أحد الأزواج والاكتفاء بالأخر. ايضاً من الشكل-1 نلاحظ وجود علاقة ارتباط ضعيفة بين الهدف وسمة fnlwgt, ما يرشح إزالتها من قاعدة المعطيات بعد التأكد من عدم تأثيرها على دقة عملية التدريب.

أما بالنسبة للسمات الفئوية فيما قمنا بإزالة عدة سمات مثل السمة education حيث يوجد سمة مكافئة لها education\_num نمطها رقمي, وسمات أخرى مثل native\_country, race و sex بناءً على عملية التجريب.

3.3 الاختبار

نقوم بالعمل على نسبة %80 من قاعدة المعطيات من أجل عملية الضبط الدقيق (fine tuning) ونسبة %20 من أجل عملية الاختبار النهائي والتحقق من دقة النموذج الأفضل المدرب.

فيما قمنا باختيار نموذج dummy classifier يقوم على أساس المنوال كخط الأساس لتحديد الحد الأدنى من الأداء المتوقع وحصلنا على دقة %75 وتفسر النتيجة بسبب عدم توازن المعطيات.

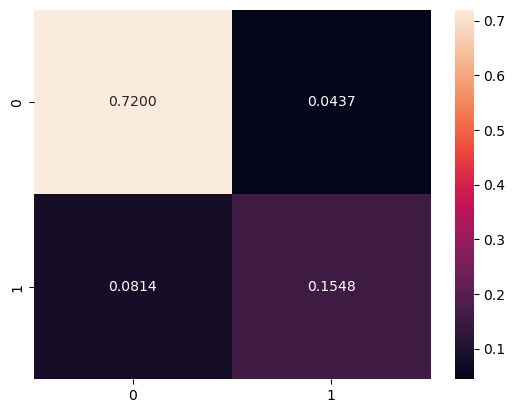
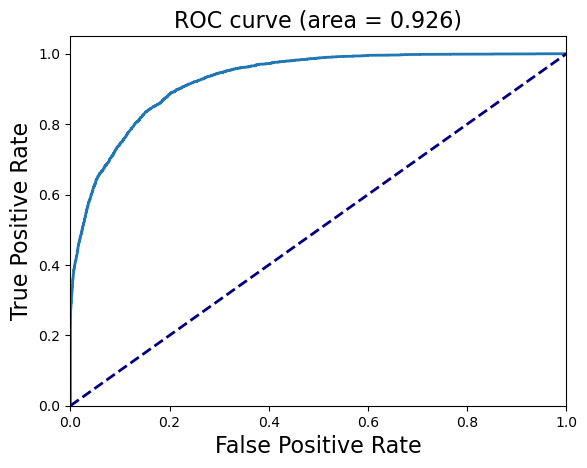
ولعدم توازن السمة الهدف نحتاج إلى معايير تعير أهمية لذلك فاعتمدنا المنطقة تحت منحني روك في عملية التحسين للوصول إلى أفضل دخل للنماذج حيث جربنا عدد من الخوارزميات موضحة في الجدول التالي دقة كل خوارزمية على معطيات التطوير. علماً أن باقي المعايير المستخدمة بالتفصيل في التقييم موثقة في بيئة التجريب.

|  |  |  |
| --- | --- | --- |
| **المنطقة الواقعة تحت منحني ROC** | **الدقة** | **اسم النموذج** |
| - | 75% | Dummy classifier |
| 0.911 | %85 | Logistic regression |
| 0.905 | %85 | K-nearest neighbor |
| 0.896 | %85 | Decision tree |
| 0.926 | %86.7 | AdaBoost |
| 0.929 | %86.7 | Gradient Boosting |

وبعد استنتاج أفضل نموذج محققاً أعلى دقة وأعلى قدر من التعميم وهو GBoost بعمق أعظمي مساو 4 وعدد estimators مساو 250 وتم اختبار معطيات التدريب عليه والحصول على النتائج الآتية:

|  |  |  |  |
| --- | --- | --- | --- |
| auc | precision | recall | accuracy |
| 0.9261942966661167 | 0.779771110423755 | 0.6554862194487779 | 0.8748848350838401 |

وهنا رسمان توضيحيان لمنحني روك ولمصفوفة ال confusion حيث تمثل أفقياً التوزع الحقيقي بين الصفر والواحد وشاقولياً ما تم تصنيفه من قبل النموذج وغماقة القطر الثانوي تدل على جودته



**4. الخاتمة**

في هذا البحث قمنا بتجريب عدد من نماذج التعلّم التلقائي ووثّقنا دقة كل خوارزمية بالجدول الذي في الأعلى. ختاماً نود أن نوضح أن قاعدة المعطيات خاصة بالافراد المقيمين في الولايات المتحدة الأميركية, فلو كانت قاعدة المعطيات أكثر شمولية جغرافياً لكان من الممكن الحصول على نتائج أدق نظراً لوجود عدة معايير وعوامل لها علاقة بدخل الأفراد باختلاف الجغرافيا.

**المراجع**

[1] Li, T. (2018). *Neural network application in income classification*. Research School Computer Science. Canberra, Australia.

[2] Deepajothi, S., & Selvarajan, S. (2012). A comparative study of classification techniques on adult data set. International journal of engineering & technology, ISSN: 2278-0181 Vol. 1 Issue8, October-2012.

[3] Chakrabarty, N., & Biswas, S. (2018, October). *A statistical approach to adult census income level prediction* [*Conference paper*]. International conference on advances in computing, communication control and networking (ICACCCN) 2018, Country name.

[4] Wang, J. (2022, June). Research on income forecasting based on machine learning methods and the importance of features [Conference session]. ICIDC 2022, Qingdao, China.

[5] https://www.diva-portal.org/smash/get/diva2:1632660/FULLTEXT01.pdf