الجمهورية العربية السورية
وزارة التعليم العالي والبحث العلمي
المعهد العالي لإدارة الأعمال
السنة الدراسية الخامسة
قسم إدارة العمليات و المعلومات

استخدام تقنيات التنقيب في البيانات بهدف الزيادة من الربحية

توقع ايداعات العملاء في بنك "Thera"

اعداد الطالب
عبادة شيخة

اشراف
د. ياسر رحال

العام الدراسي : 2022/2021

الفهرس

ملخص البحث

يتمثل هدف البحث في اقتراح نموذج لزيادة عدد عملاء القروض الشخصية في بنك بهدف زيادة الأرباح ،حيث تريد الإدارة استكشاف طرق لتحويل عملاء المسؤولية إلى عملاء قروض شخصية (مع الاحتفاظ بهم كمودعين). أظهرت الحملة التي أطلقها البنك العام الماضي لعملاء المسؤولية معدل تحويل تجاوز 9٪. وقد شجع هذا قسم تسويق التجزئة على ابتكار حملات لتسويق أفضل المستهدف لزيادة نسبة النجاح بأقل ميزانية.

لقد تم اجراء تحليلًا بسيطًا خطوة بخطوة لخصائص العميل لتحديد أنماط الاختيار الفعال للمجموعة الفرعية من العملاء الذين لديهم احتمالية أكبر لشراء منتج جديد "قرض شخصي" من البنك.

حيث تم تنفيذ الخطوات التالية على مجموعة بيانات مكونة من اثني عشرة مواصفة :

- التحقق من ارتباط كل المواصفات بحقيقة بيع المنتج أم لا.
- إيجاد المواصفات التي لها قوة ارتباط أعلى من المتوسط بالمنتج.
- تحليل الخصائص الرئيسية والحصول على شرائح في كل منها ذات قوة ارتباط مختلفة بالمنتج.
- محاولة تكوين مجموعة فرعية من العملاء بخصائص مثالية لديهم أعلى احتمالية لشراء المنتج. لسوء الحظ ، كان العدد قليل جداً , لذا انتقلنا للمرحلة التالية و هي بناء المصنفات .
- بناء نماذج بالاعتماد على أربعة خوارزميات تصنيف : 1- مصنف شجرة القرار . 2- مصنف الغابات العشوائية . 3- مصنف الجار الأقرب . 4- مصنف بايز الساذج .
- هدف البنك هو تحويل عملاء المسؤولية إلى عملاء قروض. يريدون إنشاء حملة تسويقية جديدة ومن ثم ، فهم بحاجة إلى معلومات حول العلاقة بين المتغيرات الواردة في البيانات. تم استخدام أربع خوارزميات تصنيف في هذه الدراسة .
- بعد التنفيذ كانت خوارزمية شجرة القرار تتمتع بأعلى دقة ويمكننا اختيار ذلك كنموذج نهائي لدينا .

Abstract

This case is about a bank (Thera Bank) which has a growing customer base. Majority of these customers are liability customers (depositors) with varying size of deposits. The number of customers who are also borrowers (asset customers) is quite small, and the bank is interested in expanding this base rapidly to bring in more loan business and in the process, earn more through the interest on loans. In particular, the management wants to explore ways of converting its liability customers to personal loan customers (while retaining them as depositors). A campaign that the bank ran last year for liability customers showed a healthy conversion rate of over 9% success. This has encouraged the retail marketing department to devise campaigns to better target marketing to increase the success ratio with a minimal budget.

The department wants to build a model that will help them identify the potential customers who have a higher probability of purchasing the loan. This will increase the success ratio while at the same time reduce the cost of the campaign. We conducted a simple step-by-step analysis of customer characteristics to identify patterns of effective selection of the subset of customers who are more likely to purchase a new "personal loan" product from the bank. We performed the following steps:

- We checked all twelve characteristics, whether or not each of them was related to the fact that the product was sold.

- We found five key characteristics that have an above-average strength of association with the product.

- We analyzed the main characteristics and obtained segments in each of them with different strengths of association with the product.

- We have tried to form a subset of customers with optimal characteristics who have the highest probability of purchasing the product. Unfortunately, there were too few, so we moved on to the next stage, building the classifiers.

- We built four classification algorithms: 1- Decision Tree Classifier. 2- Random forest classifier. 3- The nearest neighbor classifier. 4- Naive Bayes classifier.

The objective of the bank is to convert liability customers into loan customers. They want to create a new marketing campaign and hence, they need information about the relationship between the variables contained in the data. Four classification algorithms were used in this study.

The decision tree algorithm has the highest accuracy and we can choose that as our final model.

الفصل الاول

الاطار العام للبحث

المقدمة :

ان التطور التكنولوجي الحاصل في مختلف مجالات الحياة (الاعمال, العلوم, الطب, الهندسة, ...) أدى الى توليد كميات هائلة من البيانات يومياً . لمعالجة هذه الكمية من البيانات نحن بحاجة لأدوات تساعدنا على التحليل بشكل اسرع و ربما اكثر دقة, مما أدى الى نشوء عالم جديد هو التنقيب في البيانات او التنقيب عن المعلومات (Data Science) . انه علم يهتم باستخراج الأنماط و النماذج غير المعروفة بشكل مسبق او بديهي , و المخبأة في قواعد البيانات الضخمة بهدف فهم اعمق لما نملكه بين يدينا من بيانات بغية اتخاذ قراراتنا بشكل افضل .

يتعامل القطاع المصرفي مع شريحة كبيرة من العملاء المستفيدين من الخدمات المالية و النقدية التي يوفرها هذا القطاع , و على النحو الذي يتطلب الاحتفاظ بكميات كبيرة من البيانات عن هؤلاء العملاء و تعاملاتهم المالية , مما يستلزم توافر قاعدة بيانات خاصة لجمع وتخزين هذا الكم الهائل من البيانات، تمهيدا لعملية تنقيبها واكتشاف المعرفة منها، ومن هنا برزت أهمية تسخير التقنيات الحديثة وتوظيفها لترتيب وتجميع هذا الكم الهائل من البيانات لارتقاء اداء هذا القطاع والنهوض به .

مع هذه الكميات الضخمة من البيانات فإن الطرائق التقليدية لتحليل البيانات والتي هي مزيج من الطرائق الاحصائية وبعض النظم الحاسوبية المصممة لإدارة قواعد البيانات باتت تعاني الكثير من المشكلات في التعامل مع هذا النوع من البيانات.

من بين العلوم التطبيقية الحديثة في هذا المجال يأتي علم اكتشاف المعرفة في قواعد البيانات Knowledge Discovery in Databases وعلم التنقيب في البيانات Data Mining على رأس هذه العلوم .

ومما تقدم فان هدف البحث تتمثل في هذه الحالة ببنك لديه قاعدة عملاء متنامية . غالبية هؤلاء العملاء هم عملاء ذوو مسؤولية (مودعون) مع حجم ودائع متفاوتة . عدد العملاء الذين هم أيضًا مقترضون (عملاء الأصول) صغير جدًا ، ويهتم البنك بتوسيع هذه القاعدة بسرعة لجلب المزيد من أعمال القروض وفي هذه العملية كسب المزيد من خلال الفوائد على القروض. على وجه الخصوص ، تريد الإدارة استكشاف طرق لتحويل عملاء المسؤولية إلى عملاء قروض شخصية (مع الاحتفاظ بهم كمودعين). أظهرت الحملة التي أطلقها البنك العام الماضي لعملاء المسؤولية معدل تحويل تجاوز 9٪. وقد شجع هذا قسم تسويق التجزئة على ابتكار حملات لتسويق أفضل المستهدف لزيادة نسبة النجاح بأقل ميزانية.

مشكلة البحث :

يريد القسم بناء نموذج يساعدهم في تحديد العملاء المحتملين الذين لديهم احتمالية أكبر لشراء القرض. سيؤدي ذلك إلى زيادة نسبة النجاح وفي نفس الوقت تقليل تكلفة الحملة. و المطلوب :

- دراسة توزيع البيانات في كل Attribute .
- استخدام نموذج تصنيف للتنبؤ باحتمالية وجود عملاء مودعون سيقومون بشراء القروض الشخصية .

و بالتالي يمكن تلخيص مشكلة البحث بالأسئلة التالي :

هل توجد بعض الارتباطات بين الخصائص الشخصية وحقيقة أن العميل قد اشترى المنتج ؟ لو ذلك:

ما هي تلك الخصائص الرئيسية التي لها ارتباط بالمنتج وما هي قوة الارتباط ؟

ما هي أجزاء الخصائص الرئيسية التي لها ارتباط أقوى بالمنتج ؟

ما هي خوارزمية التصنيف التي يمكننا اعتمادها ؟

اهداف البحث :

- اقتراح نموذج يهدف الى زيادة مبيعات القروض الشخصية لعملاء البنك.
- ابتكار حملات تسويقية مستهدفة بشكل أفضل لزيادة نسبة النجاح بأقل ميزانية.
- تحديد العملاء المحتملين الذين لديهم احتمالية أكبر لشراء القرض.

منهجية البحث :

سيتم أجراء تحليلًا بسيطًا لخصائص العميل لتحديد أنماط الاختيار الفعال للمجموعة الفرعية من العملاء الذين لديهم احتمالية أكبر لشراء "قرض شخصي" من البنك .

سيتم تنفيذ الخطوات التالية:

- التحقق من جميع الخصائص الاثنتي عشرة سواء كان لكل منها ارتباط بحقيقة بيع المنتج أم لا.

- نجد الخصائص الرئيسية لها قوة ارتباط أعلى من المتوسط بالمنتج.
- نقوم بتحليل الخصائص الرئيسية والحصول على شرائح في كل منها ذات قوة ارتباط مختلفة بالمنتج.
- سنقوم بتكوين مجموعة فرعية من العملاء بخصائص مثالية لديهم أعلى احتمالية لشراء المنتج .
- بناء خوارزميات التصنيف و تحديد الخوارزمية التي لها نسبة دقة اعلى .

حدود البحث :

تتلخص حدود البحث وفق الاتي :

- حدود زمانية : تم اعداد البحث خلال المدة الزمنية الممتدة ما بين تاريخ 2022/4/1 و 2022/7/1 .
- الحدود المكانية : Thera Bank .

معوقات البحث :

- صعوبة الحصول على دراسات سابقة مفصلة مشابهة لمنهجية البحث المستخدمة كون المشاريع المشابهة عادة ما تتسم بالحماية الفكرية وبالتالي المعلومات التي تحويها تبقى محدودة.
- صعوبة الحصول على المعطيات المستخدمة في البحث .

الدراسات السابقة :

الدراسة الأولى :

الطالب :

محمد ناظم السبيعي , 2020 , استخدام تقنيات التنقيب في المعطيات للتنبؤ بنتيجة الحملات التسويقية المباشرة .

هدف الدراسة :

يهدف البحث إلى التنبؤ فيما إذا كان العميل سيشترك في وديعة استثمارية لأجل من خلال الحملات التسويقية المباشرة (الهاتفية) وذلك من خلال الإجابة عن تساؤلات البحث المذكورة في مشكلة البحث، ويمكن تلخيص الاهداف وفق الاتي:

1- إدارة الحملات التسويقية بطريقة فعالة.
2- تخفيض تكلفة الحملات التسويقية من خلال استهداف الشرائح الفعالة.
3- زيادة الإيرادات البنك من خلال الحملات التسويقية.
4- اكتساب عملاء جدد وضمهم للشرائح المستهدفة.
5- تصنيف العملاء الموجودين حاليا للتنبؤ بحاجات العملاء ذاتهم والعملاء المرتقبين.
6- زيادة ايداعات العملاء من خلال الحملات التسويق الهاتفية.
7- التأكد من رضا العميل من خلال تلبية احتياجاته.

نتائج الدراسة :

من خلال الدراسة العملية والتحليلية لبيانات هذا البحث كانت النتائج كالتالي:

1- يعرف بأن محور الحملة التسويقية المباشرة (الهاتفية) هي المكالمة لذلك يمكن من خلال الالتزام بمدة المكالمة التي حددت وهي بين (9) و (14) دقيقة فيمكن الحصول على إيداعات من العملاء بنسبة أكبر وبالتالي تخفيض تكلفة الحملات من خلال المعرفة التقريبية لنسبة مدة الاتصال.
2- يجب التركيز على الفئات العمرية التي بين (30) و (45) لأنها تملك أكبر شريحة عمرية من العملاء بشكل عام، وبشكل خا هم أكثر شريحة تشت رك بودائع بين العملاء المستهدفين، الاستفادة من الفئة العمرية ما فوق (60) بسبب حركة تجاوبهم مع الحملة التسويقية لكن بنسب صغيرة.
3- يمكن للمؤشرين سعر يوربيور (يوريبور 3) وثقة المستهلك (cons.conf.idx) عند معدلات معينة في أيام وأشهر إمكانية الاعتماد عليهم بالنسبة لتكثيف الحملة مرتبطا بمدة المكالمة الناتجة للحصول على حركة إيداع أكبر من قبل العملاء، وتخفيض الحملة نسبيا بالمعدلات التي فيها نسب كثيرة لعدم الإيداع ، حيث المؤشرين يؤثران بشكل ملحوظ بعملية نتائج الحملة وهذا منطقي من خلال المنطق الاقتصادي بحيث هناك ارتباط بين قدرة العميل على الشراء ورأيه وأمانه على نسبة إيداعه، وكذلك سعر اليوربيور بحث معرفته بسعر الفائدة على معاملات اليورو ما بين المصارف يأثر علة احتمالية إيداعه.
4- تم بناء (5) مصنفات للبيانات المدروسة واختيار أمثل مصنف واعتماده لتوقع إيداعات العملاء الاستثمارية، وقد كان مصنف الغابات العشوائية هو الأفضل من بين المصنفات المستخدمة متفوقا عليهم بجميع المعايير بنسبة دقة (% 95.3)

5- يجب الاخذ بعين الاعتبار جميع النتائج بالإضافة لمدة المكالمة مع للجميع فهي أساس نجاح الحملة إن أتمت على أكمل وجه مع النتائج السابقة.

الدراسة الثانية :

الطالب :

البدوي سعد البدوي المبارك , 2017 , استخدام تقنيات تنقيب البيانات لاستكشاف أنماط مؤثرات التحصيل الاكاديمي لطلاب المرحلة الثانوية .

هدف الدراسة :

أ- التعرف علي العوامل المؤدية الي تدني التحصيل في مادتي الرياضيات والفيزياء لطلاب مرحلة الشهادة السودانية والتعرف الي الاختلاف في العوامل المؤدية الي تدني التحصيل في مادتي الرياضيات والفيزياء لدى طلاب مرحلة الشهادة السودانية باختلاف الجنس وخبرة المعلم.

ب- بناء مستودع بيانات منطقي لقاعدة بيانات الوزارة يساهم في اتخاذ القرار وتبصير القائمون علي الوزارة بالاخطاء وطرق معالجتها وتقييم الاعمال وتكون مرجعية.

ت-التوصل الى أسباب تدني التحصيل الاكاديمي و ذلك عبر تصميم خوارزمية (K-means)للخروج بنتائج تساعد في اتخاذ القرار في الجانب الاكاديمي في وزارة التربية و التعليم .

ث-التوصل الى أسباب تدني التحصيل الاكاديمي و ذلك عبر تصميم خوارزمية شجرة القرار Decision Tree لاكتشاف الأنماط السائدة في بيانات الطلبة الاكاديمية في وزارة التربية و التعليم بالسودان .

نتائج الدراسة :

في هذا البحث:

أ- تم بناء مستودع بيانات منطقي وتطبيق بعض خوارزميات التنقيب في البيانات على قاعدة البيانات بوزارة التربية والتعليم الاتحادية للمساهمة في توفير قاعدة معرفية لصناع القرار في الوزارة .

ب- تم التوصل من خلال هذا البحث إلى استنتاجات هامة، أهمها هو الحاجة الماسة إلى بناء مستودع بيانات مترابط ومتكامل ونقي وخالي من الأخطاء للوزارة .

ت- إضافة إلى نتائج أخرى هامة متعلقة بسجلات الطلاب مثل علاقة معدلات مرحلة الاساس وفترة الانقطاع بعد مرحلة الاساس باتجاهات الطلاب الأكاديمية ومستوى أدائهم خلال دراستهم الثانوية

ث- أضف إلى ذلك علاقة كثير من المواد الدراسية بتسرب الطلاب وانقطاعهم عن الدراسة، إما لصعوبة المفردات أو لخلل في الخطط الدراسية والمناهج .

ج- وهناك عدة جوانب لم يتم التطرق إليها في هذا البحث إليها نظراً لعدم توفر المصادر المطلوبة أو

لعدم وجودها ضمن حدود البحث مبدئيا ، وهذه يمكن تطويرها كأعمال مستقبلية لهذا البحث مثل تحديث بيانات الطلاب، واستكمال البيانات الناقصة للطلاب في قاعدة بيانات الوزارة والتي قد تكون فعالة في استنتاج مزيد من الأنماط والعلاقات، مثل مكان الميلاد، والحالة الصحية، والحالة الاجتماعية، دراسة بيانات الكادر التدريسي ومدى تأثيرهم على مستوى الطلاب، واستخدام بعض الطرق الأحدث في التنقيب في البيانات، مثل استخدام تقنيات Genetic Algorithms و تقنيات Neural Networks .

الدراسة الثالثة :

الطالب :

عماد حجار , 2016 , التنقيب في بيانات المؤسسات التعليمية لتحليل مستوى الطلاب .

نتائج الدراسة :

أظهر التطبيق العملي كيف يمكن الاستفادة من تقنيات التنقيب في تحليل وفهم بيانات الطلاب بشكل أكثر عمقاً وربما إظهار نماذج ما كانت لتظهر بنفس السرعة والسهولة بدون تطبيقها.

- من خلال تطبيق تقنية قواعد الارتباط، تم الكشف عن أكثر العوامل التي تسبب رسوب الطالب وهي المشروع العملي Project Nominal
- من خلال تطبيق تقنية التصنيف وخوارزمية شجرة القرار، تم الحصول على شجرة سهلة الفهم وتساعد المدرس على توقع النتائج المستقبلية فيستطيع من خلالها اتخاذ الإجراءت الاحتياطية لمنع حدوثها وكان الامتحان العملي Exam Amali Nominal العامل الأول والأهم (جذر الشجرة) في عملية التوقع.
- من خلال تطبيق تقنية التجميع، تم فرز الطلاب إلى مجموعتين (ناجحين وراسبين) ورؤية ما يميز كل مجموعة مما يساعد ال مدرس على قيادة وتوجيه كل مجموعة على حدى. كما تم الاستفادة من تقنية التجميع في كشف الحالات الشاذة بين الطلاب والتي تعتبر مهمة جداً بالنسبة للمدرس.
- من خلال موقع الانترنت، تم تقديم فكرة جديدة و هي توفير أدوات تحل يل البيانات لجميع المدرسين كواجهة بسيطة قدر الإمكان وإمكانية دمج صفحاته ضمن صفحات بيلة التعليم الالكتروني للجامعة بدلاً من تنصيب أدوات التحليل على كل حاسب من حواسيب الأشخاص المهتمين بالقيام بعملية التحليل.

الدراسة الرابعة :

الطالب :

نهى حمود , تصميم عروض حسب الشرائح الزمنية باستخدام التنقيب في المعطيات, 2021

اهداف الدراسة :

يهدف البحث إلى تصميم عروض لزبائن شركات الاتصال بالاعتماد على الشرائح الزمنية وذلك من خلال الإجابة عن تساؤل البحث المذكور في مشكلة البحث، ويمكن تلخيص الأهداف وفق الآتي:

1- تصميم عروض باستخدام إحدى تقنيات التنقيب في المعطيات.

2- تحديد ميزات المعطيات لتسهيل تصميم العروض حسب الشرائح الزمنية.

نتائج الدراسة :

من خلال الدراسة التي أجريت ومن خلال ما تم توضيحه في الشقين النظري والعملي من البحث يمكن تلخيص النتائج التي توصلت إليها الباحثة من خلال بحثها فيما يلي:

1- يمكن استخدام تقنيات التنقيب في المعطيات لتصميم عروض لزبائن شركات الاتصال.

2- من خلال القسم النظري تم الحصول على الفئات التي يمكن تصميم عروض خاصة لها حيث يمكن تصميم عروض خاصة للفئة المنتمية للعنقود ذو الرقم صفر لأنه كان يحمل السمة (TeleV)بشكل أكبر، ويمكن تصميم عروض خاصة للعنقود ذو الرقم واحد لجذب الزبائن بشكل أكبر لأنه كان يحمل السمة (TeleV) بشكل أقل من العنقود السابق.

الفصل الثاني

القسم النظري

مقدمة :

ان التقدم العلمي و انتشار استخدام التكنولوجيا في شتى مناحي الحياة اليومية ادلا الى رفع القدرة في توليد و جمع البيانات بشكل سريع في هذا العصر , و قد ساهم ذلك في حوسبة معظم الاعمال و العلوم و الخدمات التي يتم تقديمها يومياً في كل مكان حول العالم , حتى ان معظم المنتجات بكافة أنواعها اصبح لها شيفرة رقمية (رمز استجابة سريع) تميزها عن بعضها البعض Bar Code , كما ان التقدم التكنولوجي تسبب في ظهور أنواع جديدة من البيانات كالنصوص و الصور و الفيديو و أنظمة متعددة المهام بالإضافة الى شبكة الانترنت التي تحتوي على كميات هائلة من البيانات بكافة اشكالها . كل ذلك أدى الى تضخم غير مسبوق في كميات البيانات التي يتم تخزينها يومياً , الامر الذي اظهر حاجات ملحة لتقنيات جديدة و أدوات ذكية يمكن ان تساعد في تحويل هذا الكم الهائل من البيانات الى معلومات او معرفة مفيدة , و التي تمثلت في أدوات تنقيب البيانات باعتبارها أدوات جديدة في منهجية البحث العلمي العصري الذي يختلف عن البحث العلمي الكلاسيكي .

تعريفات التنقيب في البيانات :

1- التنقيب في مجموعه ضخمه من مجلدات البيانات فضلا عن اكتشاف العلاقة بينهما او الاجابة عن الاسئلة المتخصصة التي تكون واسعه جدا عند استخدام ادوات الاستعلام التقليدية .
2- عباره عن تحليلات لكميه كبيره من البيانات بغرض ايجاد قواعد و امثلة ونماذج يمكن ان تستخدم وتقود وتدل اصحاب القرار وتتنبأ بالسلوك المستقبلي.
3- عباره عن تطبيق لاستخراج او اكتشاف معرفه مفيدة وقابله للاستغلال من خلال مجموعه كبيره من البيانات حيث يساعد في اكتشاف المعرفة المخفية والنماذج غير المتوقعة بالإضافة الي استكشاف قواعد جديده موجوده في قواعد بيانات كبيرة .

اهداف التنقيب في البيانات :

ان التنقيب في البيانات (او تعدين البيانات)يهدف الى استخلاص المعلومات المخبأة في كتل البيانات الكبيرة . و تنقيب البيانات تكنولوجيا حديثة فرضت نفسها بقوة في عصر المعلوماتية , و استخدامها يوفر للشركات و المؤسسات في جميع المجالات القدرة على استكشاف و التركيز على اهم المعلومات في كتل

البيانات الكبيرة . كما تركز تقنيات التنقيب على الاستشعار و بناء التنبؤات المستقبلية و استكشاف الأنماط و الارتباطات و السلوك و الاتجاهات مما يسمح بتقدير القرارات الصحيحة و اتخاذها في الوقت المناسب و وضع الحلول المناسبة للمشكلات و التخطيط و التطوير و التحديث في جميع المجالات . و تجيب تقنيات التنقيب على العديد من الأسئلة و في وقت قياسي . بخاصة تلك النوعية من الاسئلة التي كان من الصعب الإجابة عليها , ان لم يكن مستحيلاً , باستخدام تقنيات التحليل الاحصائي الكلاسيكية . و التي كانت ان وجدت فأنها تستغرق وقتاً طويلاً و العديد من إجراءات التحليل .[1]

أهمية التنقيب في البيانات :

للتنقيب في البيانات أهمية كبيرة في الكثير من التطبيقات الحقيقة حول العالم وأهمها القدرة على التعامل مع المشاكل المعقدة، والقدرة على اكتشاف المعلومات المثيرة للاهتمام والغير متوقعة، واستخراج العديد من البيانات المخفية، والقدرة على التعلم الذاتي، وإمكانية استخدام التجارب والاخطاء من الماضي لتحسين نوعية النماذج تلقائيا، والتعرف على مسارات البيانات المخفية وهو ما يؤهله لتقديم العون والمساعدة في بناء التنبؤات المستقبلية، واستكشاف السلوك، والاتجاهات مما يسمح لاتخاذ القرارات الصحيحة وفي الوقت المناسب[2].

تنقيب البيانات و ذكاء الاعمال :

ان علم تنقيب البيانات يعتبر من العلوم الحديثة نسبياً , و هو يشكل امتداداً لعلم التحليل الاحصائي و العصب الرئيسي لعلوم ذكاء الاعمال او استخبارات الاعمال بكافة اشكالها و المستخدمة بشكل أساسي في مجال الاعمال . و قد نشأ علم التنقيب كنتيجة طبيعية للتطور الكبير الذي حدث في مجال نظم المعلومات و التضخم الكبير في المعلومات التي تنمو بشكل اسي , و خاصة بعد الانتشار الواسع لاستخدام أنظمة المعلومات و تراكم الكم الهائل من البيانات التي أصبحت متداولة يومياً في العديد من المجالات , الامر الذي أدى الى الحاجة الملحة للإجابة على العديد من الأسئلة و استكشاف المعرفة و التقديرات و التنبؤات المستقبلية . و تعتبر عمليات التحليل و التنقيب من أولويات عمل دوائر التخطيط في الشركات و المؤسسات في العالم في الوقت الراهن و الأداة المثلى للمستويات الإدارية العليا التي

[1] (البدوي2017)
[2] (han, kamber, & pei, 2000)

تطمح للنجاح و تضمن استمراره بشكل استراتيجي , نظراً لما توفره من إمكانية لإنتاج المعرفة الحقيقية المخبأة في كتل البيانات الكبيرة الخاصة بأنشطة كل شركة او مؤسسة و التي يتم تخزينها و تراكمها يومياً .

أنواع التنقيب في البيانات :

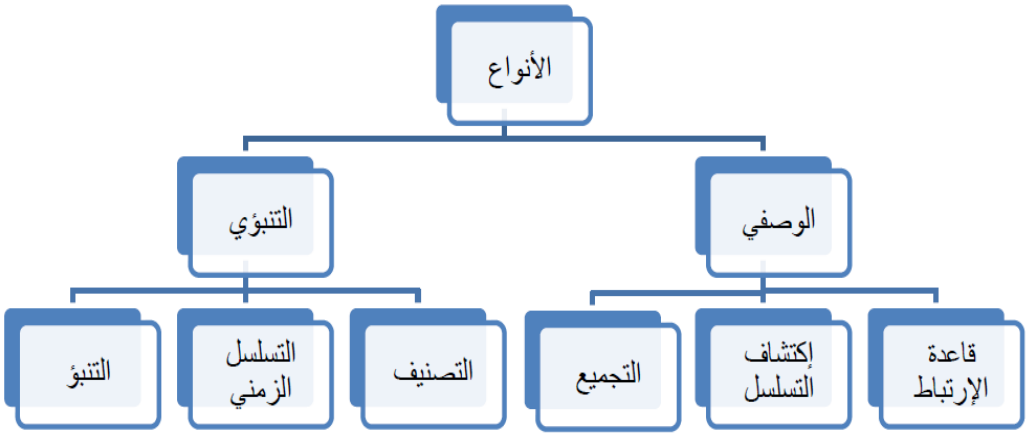
توجد عدة تقنيات تستخدم في تنقيب البيانات و اختيار التقنية المناسبة يعتمد على طبيعة البيانات التي تتم دراستها فضلا عن حجم البيانات وهناك نموذجان لتنقيب البيانات :

التنقيب التنبؤي (Prediction) و ضيم عدة تقنيات , منها :

- التصنيف Classification .
- الانحدار Regression .
- تحليل السلاسل الزمنية Time Series Analysis .

التنقيب الوصفي (Describe) و يحتوي على عدة تقنيات , منها :

- قواعد الارتباط Association Rules .
- العنقدة Clustering .
- اكتشاف التسلسل Discovery .
- المرئية Visualization .
- التلخيص Summarization .[3]



اولاً – التنقيب التنبؤي Predictive Model

وهو النموذج الذي يستخدم النتائج المعروفة المستنبطة من البيانات المختلفة لأجل التنبؤ بقيم لاحقة للبيانات ويحاول إيجاد أفضل التنبؤات اعتماداً على المعطيات , كمعرفة المنتج الأفضل لزبون معين . باختصار يعتمد هذا التنقيب على استخدام معلومات قديمة لتوقع ما سيحدث في المستقبل وتكون لدى مثلهذه البيانات هدف . ويتضمن هذا النموذج أشهر تقنيات أدوات تنقيب البيانات.

1- التصنيف Classification

يستخدم التصنيف بشكل واسع في حل الكثير من المشكلات خاصة تلك التي تتعلق بالأعمال من خلال تحليل مجموعة من البيانات ووضعها على شكل أصناف أو أقسام يمكن استخدامها فيما بعد لتصنيف البيانات مستقبلا ، وهنا يكمن الفرق بين التصنيف وانشاء العناقيد، فالتصنيف يقصد به تقسيم البيانات إلى مجاميع يتم تحديدها مسبقا . اما انشاء العناقيد او ما يسمى بال Clustering فهو يعني تقسيم البيانات الى مجاميع ليست معروفة مسبقاً . هناك العديد من الطرق التي يمكن استخدامها في تصنيف البيانات باستخدام خوارزميات مختلفة مثل الخوارزميات الإحصائية Statistical Algorithms و الشبكات العصبية Neural Network و الخوارزميات الجينية Genetic Algroithms و طريقة الجار الأقرب Nearest Neighbor Method . و طريقة ال Decision Trees و هي هيكل شجري يقدم مجموعة من القرارات التي تولد قواعد لمجموعة البيانات المصنفة (Classification Data) تشترط هذه التقنية وجود حقل قرار يتم تصنيف البيانات بناءً عليه .

2- الانحدار Regression

هو تقنية تسمح بتحليل البيانات لوصف العلاقة بين متغيرين أو أكثر ، ويستخدم في تقدير قيمة أحد المتغيرين إذا عرف المتغير الاخر. و يفترض أن توضع البيانات بنوع معروف من الدوال ومن ثم يتم تحديد أفضل دالة للبيانات المعطاة.

3- تحليل السلاسل الزمنية Time Series Analysis

إن مشاهدة البيانات عبر الزمن تنتج تحليلا مفيدا لأنه تتم مشاهدة سلوك البيانات عبر الزمن بشكل أساسي . وهذا يعني أن قيم الصفة المميزة للبيانات التي يتم فحصها تكون متغيرة عبر الزمن .

4- التنبؤ Prediction

يعد التنبؤ من الأدوات التي تجذب الانتباه لأنها تتمكن من إعطاء مغزى التوقع الناجح في سياق العمل لذا فإنه يمكن النظر إلى العديد من تطبيقات تنقيب بيانات العالم الحقيقي كأنها تنبؤ بحالة بيانات مستقبلية معتمدة على بيانات سابقة وحالية.

ثانياً – التنقيب الوصفي Descriptive Model

وهو النموذج الذي يُعَرف الأنماط والعلاقات في البيانات ، وعلى العكس من النموذج التنبؤي ، فإن النموذج الوصفي يستخدم كطريقة لاستكشاف خصائص البيانات التي تتم دراستها وليس للتنبؤ بخصائص جديدة ، ويعتمد علي إعادة تنظيم البيانات ، والتنقيب في أعماقها لاستخراج النماذج الموجودة فيها كتشابه الزبائن الذي يسمح لك بإنشاء وصف بسيط عن مجموعه زبائن متشابهين ولا يستوجب وجود هدف لمثل تلك التشابهات ، ويعد من أشهر تقنيات هذا النموذج من أدوات تنقيب البيانات .

1- العنقدة Clustering

الهدف منها هو تحديد الاتجاهات داخل البيانات ، ويحاول هذا الاسلوب العثور على مجموعات من العناصر التي توجد عادة معا . وهي عملية تقسيم البيانات إلى مجموعات من الأصناف اعتمادا على
اشتراكها بالخواص المتشابها وأن العنقدة هي تقسيم غير موجه للبيانات ، وهي عكس التصنيف ، كما أنها تساعد المستفيد على فهم التركيب الطبيعي لمجموعات من البيانات.

2- قواعد الارتباط Association Rules

قواعد الارتباط هي احدى التقنيات الواعدة في ال Data Mining كأداة من أدوات استكشاف المعرفة KDD و لديها القدرة على معالجة كميات هائلة من البيانات , و تسمح باستنتاج كل القوانين الممكنة التي تشرح بعض الصفات الموجودة اعتماداً على وجود الصفات الأخرى و بمعنى اخر هي قواعد ارتباط معينة بين عدة مجموعات من البيانات في قاعدة بيانات واحدة و تعتمد طريقة قواعد الارتباط على إيجاد Large Item Set

3- اكتشاف التسلسل Sequence Discovery

إن تحليل أو اكتشاف التسلسل يستخدم لتحديد أنماط متسلسلة في البيانات وهذه الأنماط معتمدة على تسلسل زمن التأثيرات . في هذه الطريقة يتم البحث لاكتشاف نماذج تحدث بالتسلسل إذ تكون المدخلات عبارة عن بيانات تشكل مجموعه متسلسله و كل سلسله من البيانات هي قائمه منظمه من العمليات أو المصطلحات و عندما تكون العملية عبارة عن مجموعات من المصطلحات لابد أن يحسب معها الوقت المصاحب لكل عملية ولكن مشكلة هذا النموذج تكمن في إيجاد كل النماذج المتسلسلة مع أقل دعم يخصصه المستخدم عندما يكون الدعم لهذا النموذج هو نسبة تسلسل البيانات التي يتضمنها النموذج.

4- المرئية Visualization

إن مرئية النتائج يكون مساعدا في تسهيل ملاحظات مخرجات خوارزميات تنقيب البيانات وفهمها .

5- التلخيص Summarization

و يسمى ايضاً بالمميزات او الخواص Characterization او العمومية Generalization و يمكن تعريفها على انها وصف الخصائص العامة للنماذج .[4]

مراحل استخراج المعرفة و التنقيب بها :

تظهر عملية استخراج المعرفة كتسلسل تكراري للمراحل التالية :

1- **تنظيف البيانات** (لإزالة الضوضاء و البيانات غير المتسقة) .
2- **دمج البيانات** (حيث يمكن دمج مصادر بيانات متعددة) .
3- **اختيار البيانات** (حيث يثم استرداد البيانات ذات الصلة بمهمة التحليل من قاعدة البيانات) .
4- **تحويل البيانات** (حيث يتم تحويل البيانات و دمجها في اشكال مناسبة للتعدين من خلال اجراء عمليات الملخص او التجميع) .[5]

[4] (البدوي 2017)
[5] (Palace, 1996)

5- **استخراج البيانات** (عملية أساسية حيث يتم تطبيق الأساليب الذكية لاستخراج أنماط و نماذج البيانات) .

6- **تقييم النماذج** (لتحديد النماذج التي تمثل المعرفة بناءً على المقاييس المهمة).

7- **عرض المعرفة** (حيث يتم استخدام تقنيات التصور و تمثيل المعرفة لتقديم المعرفة للمستخدمين).[6]

تطبيقات التنقيب في البيانات :

وسائل التنقيب في البيانات تستعمل و بنجاح في كثير من التطبيقات الحقيقية حول العالم:

- **التجارة**

1- الكشف عن حالات الاحتيال : تحديد المعاملات الاحتيالية.

2- الموافقة على القروض : تحديد الاهلية الائتمانية للعميل طالب القرض.

3- تحليل الاستثمار : التنبؤ بمحفظة العائد على الاستثمار.

4- التسويق و المبيعات و تحليل البيانات : تحديد العملاء المحتملين ، اقامة فعالية حملة المبيعات.

- **الطب**

1- تحليل تأثير المخدرات : من حالات المرضى لمعرفة اثار المخدرات.

2- مكافحة الامراض وتحليل العلاقة السببية.

3- تصنيف الامراض : تصنيف مرض السرطان.

- **السياسة**

السياسة الاجتماعية : الضرائب و الفوائد .

- **التصنيع**

1- الصناعة التحويلية في عملية التحليل : تحديد أسباب المشاكل و التصنيع .

2- التحليل لنتائج التجربة : ملخص لنتائج التجربة و خلق نماذج تنبؤ .

[6] (han, kamber, & pei, 2000)

تطبيقات التنقيب في البيانات بدأت تنمو بصورة كبيرة للأسباب الاتية :

1- كمية البيانات الموجودة في مخزن البيانات و سوق البيانات تنمو بصورة كبيرة , اذ ان وجود بيئة معلوماتية كبيرة تدفع المهتمين للاستفادة منها .

2- ظهور الكثير من أدوات التنقيب الفعالة , شجع على ازدياد عملية التنقيب في البيانات بكثرة .

3- المنافسة الشديدة الموجودة في السوق تدفع الشركات الى البحث عن طرق تساعدها على الإنتاج باقل التكاليف .[7]

فوائد التنقيب في البيانات :

بشكل عام ، تأتي الفوائد التجارية لاستخراج البيانات من زيادة القدرة على الكشف عن الأنماط المخفية والاتجاهات والارتباطات والشذوذ في مجموعات البيانات. يمكن استخدام هذه المعلومات لتحسين عملية اتخاذ القرارات التجارية والتخطيط الاستراتيجي من خلال مزيج من تحليل البيانات التقليدية والتحليلات التنبؤية.

تشمل مزايا التنقيب عن البيانات المحددة ما يلي:

- تسويق ومبيعات أكثر فعالية. يساعد التنقيب عن البيانات المسوقين على فهم سلوك العملاء وتفضيلاتهم بشكل أفضل ، مما يمكنهم من إنشاء حملات تسويقية وإعلانية مستهدفة. وبالمثل ، يمكن لفرق المبيعات استخدام نتائج استخراج البيانات لتحسين معدلات تحويل العملاء المحتملين وبيع منتجات وخدمات إضافية للعملاء الحاليين.
- خدمة عملاء أفضل. بفضل التنقيب عن البيانات ، يمكن للشركات تحديد مشكلات خدمة العملاء المحتملة بشكل أسرع وإعطاء وكلاء مركز الاتصال معلومات محدثة لاستخدامها في المكالمات والمحادثات عبر الإنترنت مع العملاء.
- تحسين إدارة سلسلة التوريد. يمكن للمؤسسات تحديد اتجاهات السوق والتنبؤ بالطلب على المنتجات بشكل أكثر دقة ، مما يمكنها من إدارة مخزونات السلع والإمدادات بشكل أفضل. يمكن لمديري سلسلة التوريد أيضًا استخدام المعلومات من استخراج البيانات لتحسين التخزين والتوزيع والعمليات اللوجستية الأخرى.
- زيادة وقت تشغيل الإنتاج. يدعم استخراج البيانات التشغيلية من أجهزة الاستشعار الموجودة على آلات التصنيع وغيرها من المعدات الصناعية تطبيقات الصيانة التنبؤية لتحديد المشكلات المحتملة قبل حدوثها ، مما يساعد على تجنب فترات التوقف غير المجدولة.

[7] (البدوي , 2017)

- إدارة أقوى للمخاطر. يمكن لمديري المخاطر والمديرين التنفيذيين للأعمال تقييم المخاطر المالية والقانونية والأمن السيبراني والمخاطر الأخرى التي تتعرض لها الشركة بشكل أفضل ووضع خطط لإدارتها.
- انخفاض التكاليف. يساعد التنقيب عن البيانات في تحقيق وفورات في التكاليف من خلال الكفاءات التشغيلية في العمليات التجارية وتقليل التكرار والهدر في إنفاق الشركات.

في النهاية ، يمكن أن تؤدي مبادرات التنقيب عن البيانات إلى زيادة الإيرادات والأرباح ، فضلاً عن المزايا التنافسية التي تميز الشركات عن منافسيها التجاريين.

التصنيف Classification :

التصنيف هو تقنية للتنبؤ بنتيجة معينة بناءً على مدخلات معينة. تستخرج نماذج تصف بشكل دقيق فئات وتصنيفات البيانات الهامة، وتتنبأ مثل هذه النماذج، بالتصنيفات الفئوية (المنفصلة وغير المرتبة) .

تحلل الخوارزمية مجموعة تدريب تحتوي على مجموعة من السمات جنبًا إلى جنب مع نتائج كل منها (هدف أو سمة تنبؤ)، ويتم استخدام هذا بعد ذلك للتنبؤ بنتائج بيانات غير معروفة تسمى مجموعة الاختبار. أولاً، تحاول الخوارزمية اكتشاف العلاقات بين سمات مجموعة التدريب التي تجعل من الممكن التنبؤ بالنتيجة، وبعد ذلك يتم إعطاء الخوارزمية مجموعة بيانات لم يتم رؤيتها من قبل، تسمى مجموعة التنبؤ أو مجموعة الاختبار، والتي تحتوي على نفس مجموعة السمات، باستثناء سمة التنبؤ تكون (غير المعروفة) حتى الان .

تقوم الخوارزمية بتحليل المدخلات وإصدار التنبؤ. تحدد دقة التنبؤ مدى جودة "الخوارزمية". على سبيل المثال، يمكن استخدام التصنيف في قاعدة بيانات طبية للتنبؤ بما إذا كان المريض يعاني من أمراض القلب بناءً على المعلومات المسجلة سابقا والتي تحمل سمة الإخراج (يعاني/ لا يعاني)، مثال أخر تحتاج قروض البنك التحليل في بياناتها لمعرفة أي من المتقدمين للحصول على القروض ("آمنون" / "محفوف بالمخاطر") بالنسبة للبنك، أو توقع ما إذا كانت ستمطر في يوم معين أم لا.

يتكون النهج العام للتصنيف بانه عملية من خطوتين :

1- الخطوة الأولى :

نقوم ببناء نموذج تصنيف بناءً على البيانات السابقة و تسمى خطوة التعلم او مرحلة التعلم , و في هذه الخطوة تقوم خوارزميات التصنيف ببناء المصنف . المصنف مبني على مجموعة التدريب (البيانات السابقة) المكونة من قاعدة بيانات Tuples و تصنيفات الفصل المرتبطة (سمة التنبؤ) .

2- الخطوة الثانية :

نحدد ما إذا كانت دقة النموذج مقبولة، وإذا كان الأمر كذلك، فإننا نستخدم النموذج لتصنيف البيانات الجديدة، في هذه الخطوة يتم استخدام المصنف للتنبؤ ببيانات غير معروفة، هنا يتم استخدام بيانات الاختبار لتقدير الدقة قواعد التصنيف. يمكن تطبيق قواعد التصنيف على البيانات الجديدة، تتم مقارنة هذا المصنف مع المصنف الفعلي[8].

العوامل المؤثرة على أداء دقة التصنيف :

1) تحويل البيانات : Data transformation

تحتوي عملية تحويل البيانات على عمليتين , و ليس بالضرورة استخدامهم معا أي ان طبيعة البيانات هي التي تحدد .

1- التطبيع : Normalization

يتم تحويل البيانات باستخدام التطبيع , و يشمل التطبيع قياس جميع القيم لسمة معينة لجعلها تقع ضمن نطاق صغير محدد . يستخدم التطبيع عندما يتم استخدام الشبكات العصبية او الأساليب التي تنطوي على قياسات في خطوة التعلم .

2- التعميم Generalization

يمكن ايضاً تحويل البيانات من خلال تعميمها على المفهوم الأعلى لهذا الغرض يمكننا استخدام التسلسل الهرمي للمفاهيم .

[8] (السبيعي,2020)

2) وجود قيم المتطرفة : Presence of outliers

القيم المتطرفة هي نقاط بيانات لا تتوافق مع غالبية البيانات. القيم المتطرفة هي أيضًا نقاط يصعب تصنيفها بسبب يمكننا أن نقول انهم لا يملكون خصائص ناقلات ميزة مماثلة مثل غالبية البيانات. لذا تصنيف القيم المتطرفة مهمة محفوفة بالمخاطر ويمكن أن تؤثر على الدقة بشكل سلبي. طريقة جيدة يمكن استخدامها هنا مقدما هي ازالة القيم المتطرفة. هناك عدد قليل من الطرق الجيدة للقيام بذلك التجميع وتركيب المنحنى.

3) إزالة الضوضاء : Noise removal

يمكن تعري الضوضاء عادة على أنها خطأ عشوائي أو التباين في متغير مقاس يكون النوعان النموذجيان فيه غير متناسقين قيم الميزات أو الفئات. الضوضاء عادة ما تكون أقلية في مجموعة البيانات. ويمكن إزالته باستخدام خوارزميات العنقدة.

4) تحليل مدى الارتباط : Relevance analysis

قد تحتوي البيانات في بعض الأحيان على مجموعات قليلة من الصفات التي لا توفر الكثير من المعلومات للمصنف. هذا يعني انه لا تشكل هذه السمات إدخالات مهمة في متجه الميزة. ازالة هذه السمات يمكن لهذه الصفات تسريع أداء المصنف. يمكن أيضا القيام بذلك باستخدام (Chi-square) ، (LDA (Linear Discriminant Analysis تساعد هذه الطرق في تقليل مساحة الميزة بشكل فعال.

5) طرق التقدير الخاطئة : Wrong estimation methods

لا ينبغي ان تكون دقة التصنيف مثالية تقاس في تجربة واحدة . التحقق المتقاطع هو طريقة جيدة جداً لقياس الدقة التي تستخدم نوعاً من الاجراء و تترك الدقة لجميع التكرارات . لكن الدقة لا تزال تدبير شخصي ولا يمكن تزويدنا بمعلومات كاملة عن أداء المصنف .[9]

[9] (السبيعي,2020)

خوارزميات التصنيف :

خوارزمية شجرة القرار Decision Tree Classifier

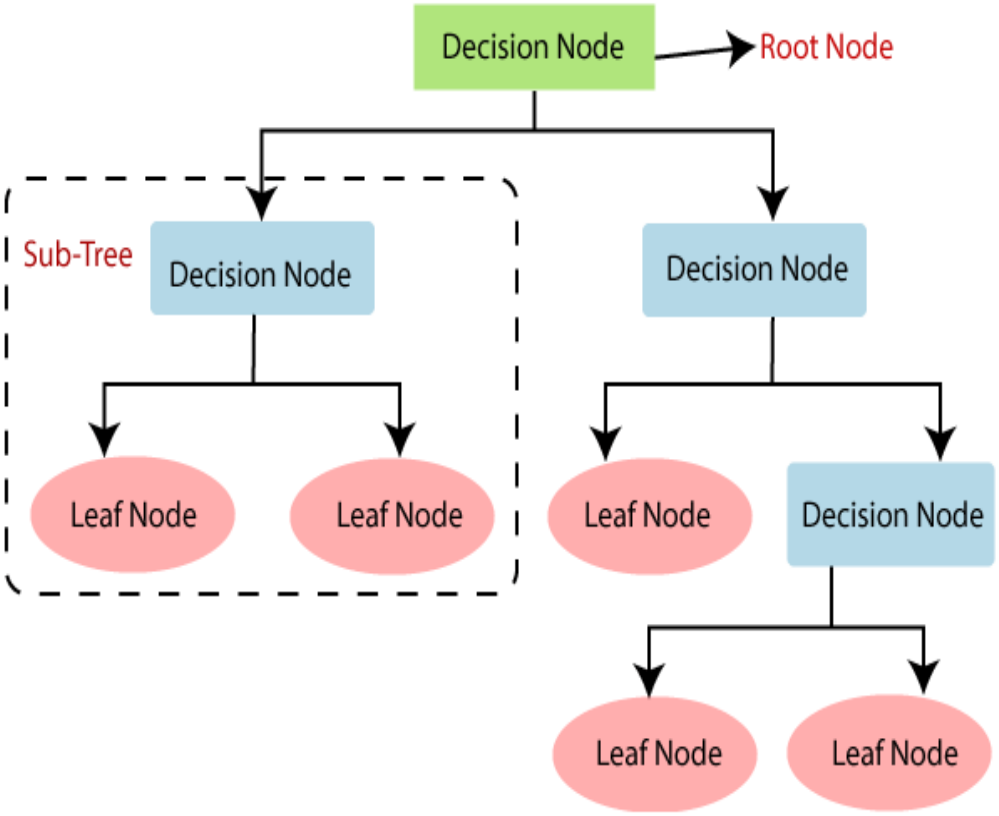
شجرة القرار هي تقنية تعلم خاضعة للإشراف يمكن استخدامها لكل من مشاكل التصنيف والانحدار ، ولكنها في الغالب مفضلة لحل مشاكل التصنيف. إنه مصنف منظم على شكل شجرة ، حيث تمثل العقد الداخلية ميزات مجموعة البيانات ، وتمثل الفروع قواعد القرار وتمثل كل عقدة ورقية النتيجة.

في شجرة القرار ، توجد عقدتان ، وهما عقدة القرار Decision Node والعقدة الورقية Leaf Node . تُستخدم عقد القرار لاتخاذ أي قرار ولها فروع متعددة ، في حين أن العقد الورقية هي نتاج تلك القرارات ولا تحتوي على أي فروع أخرى.

إنه تمثيل رسومي للحصول على جميع الحلول الممكنة لمشكلة / قرار بناءً على شروط معينة.

يطلق عليها شجرة القرار لأنها ، على غرار الشجرة ، تبدأ بالعقدة الجذرية ، التي تتوسع في المزيد من الفروع وتبني بنية شبيهة بالشجرة. تطرح شجرة القرار سؤالاً ببساطة ، وبناءً على الإجابة (نعم / لا) ، فإنها تقسم الشجرة إلى أشجار فرعية.

يوضح الرسم البياني أدناه الهيكل العام لشجرة القرار:



لماذا نستخدم شجرة القرار :

هناك العديد من الخوارزميات في التعلم الآلي ، لذا فإن اختيار أفضل خوارزمية لمجموعة البيانات والمشكلة المعينة هو النقطة الرئيسية التي يجب تذكرها أثناء إنشاء نموذج التعلم الآلي. فيما يلي سببان لاستخدام شجرة القرار:

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

طريقة عمل الخوارزمية :

في شجرة القرار ، للتنبؤ بفئة مجموعة البيانات المحددة ، تبدأ الخوارزمية من العقدة الجذرية للشجرة. تقارن هذه الخوارزمية قيم سمة الجذر مع سمة 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 ، لنقول S ، والتي تحتوي على مجموعة البيانات الكاملة.

الخطوة 2: ابحث عن أفضل سمة في مجموعة البيانات باستخدام مقياس تحديد السمة (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.

مزايا شجرة القرار :

- من السهل فهمها لأنها تتبع نفس العملية التي يتبعها الإنسان أثناء اتخاذ أي قرار في الحياة الواقعية.
- يمكن أن تكون مفيدة جدًا في حل المشكلات المتعلقة بالقرار.
- تساعد على التفكير في جميع النتائج المحتملة لمشكلة ما.
- هناك متطلبات أقل لتنظيف البيانات مقارنة بالخوارزميات الأخرى.

عيوب شجرة القرار :

- تحتوي شجرة القرار على الكثير من الطبقات ، مما يجعلها معقدة.
- قد يكون هناك مشكلة في التركيب ، والتي يمكن حلها باستخدام خوارزمية Random Forest.
- لمزيد من تسميات الفئات ، قد يزداد التعقيد الحسابي لشجرة القرار .

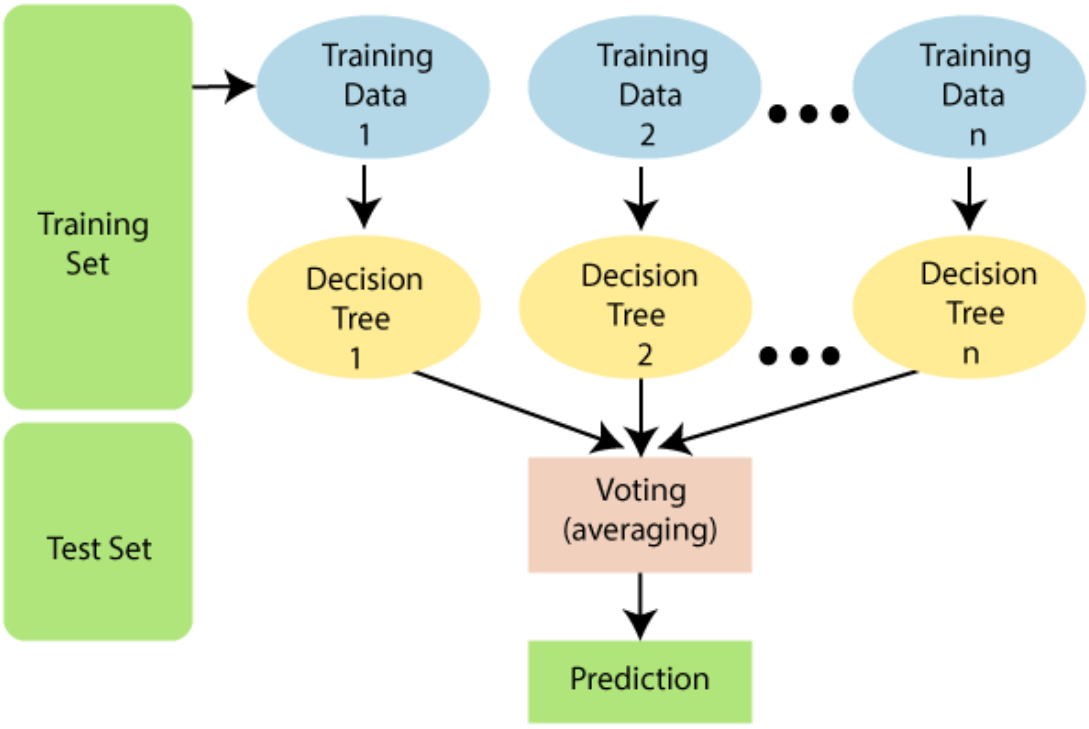
خوارزمية الغابات العشوائية Random Forest Algorithm

الغابات العشوائية خوارزمية شائعة للتعلم الآلي تنتمي إلى تقنية التعلم الخاضع للإشراف. يمكن استخدامه لكل من مشاكل التصنيف والانحدار في التعلم الالي . يعتمد على مفهوم التعلم الجماعي ، وهو عملية الجمع بين عدة مصنفات لحل مشكلة معقدة وتحسين أداء النموذج.

كما يوحي الاسم "الغابات العشوائية" هو مصنف يحتوي على عدد من أشجار القرار في مجموعات فرعية مختلفة من مجموعة البيانات المحددة ويأخذ المتوسط لتحسين الدقة التنبؤية لمجموعة البيانات هذه. بدلاً من الاعتماد على شجرة قرار واحدة ، تأخذ الغابة العشوائية التنبؤ من كل شجرة وتعتمد على أغلبية الأصوات للتنبؤات ، وتتوقع الناتج النهائي.

يؤدي العدد الأكبر من الأشجار في الغابة إلى دقة أعلى ويمنع مشكلة التجهيز الزائد.

يوضح الرسم البياني أدناه عمل خوارزمية Random Forest:



لماذا تستخدم الغابات العشوائية ؟

فيما يلي بعض النقاط التي تشرح سبب استخدامنا لخوارزمية Random Forest:

- يستغرق وقت تدريب أقل مقارنة بالخوارزميات الأخرى.
- يتنبأ بالإخراج بدقة عالية ، حتى بالنسبة لمجموعة البيانات الكبيرة التي يتم تشغيلها بكفاءة.
- يمكنه أيضًا الحفاظ على الدقة عند فقدان نسبة كبيرة من البيانات.

كيف تعمل خوارزمية Random Forest؟

تعمل Random Forest على مرحلتين الأولى هي إنشاء الغابة العشوائية من خلال الجمع بين شجرة القرار N ، والثاني هو عمل تنبؤات لكل شجرة تم إنشاؤها في المرحلة الأولى.

يمكن شرح عملية العمل في الخطوات أدناه:

الخطوة 1: حدد نقاط بيانات K عشوائية من مجموعة التدريب.

الخطوة 2: قم ببناء أشجار القرار المرتبطة بنقاط البيانات المحددة (المجموعات الفرعية).

الخطوة 3: اختر الرقم N لأشجار القرار التي تريد بناءها.

الخطوة 4: كرر الخطوة 1 و 2.

مزايا الغابة العشوائية :

- Random Forest قادرة على أداء مهام التصنيف والانحدار.
- إنها قادرة على التعامل مع مجموعات البيانات الكبيرة ذات الأبعاد العالية.
- يعزز دقة النموذج ويمنع مشكلة التجاوز.

عيوب الغابات العشوائية :

- على الرغم من أنه يمكن استخدام مجموعة التفرعات العشوائية لكل من مهام التصنيف والانحدار ، إلا أنها ليست أكثر ملاءمة لمهام الانحدار.

خوارزمية مصنف بايز Naïve Bayes Classifier Algorithm

خوارزمية Naïve Bayes هي خوارزمية تعلم خاضعة للإشراف ، والتي تستند إلى نظرية بايز وتستخدم لحل مشاكل التصنيف.

يتم استخدامه بشكل أساسي في تصنيف النص الذي يتضمن مجموعة بيانات تدريبية عالية الأبعاد.

يعد تصنيف Naïve Bayes أحد خوارزميات التصنيف البسيطة والأكثر فاعلية والتي تساعد في بناء نماذج التعلم الآلي السريعة التي يمكنها إجراء تنبؤات سريعة.
إنه مصنف احتمالي ، مما يعني أنه يتنبأ على أساس احتمال وجود كائن.
بعض الأمثلة الشائعة لخوارزمية Naïve Bayes هي ترشيح البريد العشوائي والتحليل العاطفي وتصنيف المقالات.

نظرية بايز Bayes' theorem :

تُعرف نظرية بايز أيضًا باسم قانون بايز ، والذي يستخدم لتحديد احتمالية الفرضية بمعرفة مسبقة. يعتمد ذلك على الاحتمال الشرطي.
تُعطى صيغة نظرية بايز على النحو التالي:

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$$

حيث :
P (A | B) هو الاحتمال اللاحق: احتمال الفرضية A على الحدث المرصود B.
P (B | A) هو احتمال الاحتمال: احتمالية الدليل بالنظر إلى أن احتمال الفرضية صحيح.
P (A) هو احتمال مسبق: احتمال فرضية قبل ملاحظة الدليل.
P (B) هو احتمال هامشي: احتمال وجود دليل.

مزايا مصنف بايز :

- Naïve Bayes هي واحدة من خوارزميات التعلم الالي السريعة والسهلة للتنبؤ بفئة من مجموعات البيانات.
- يمكن استخدامه لتصنيفات ثنائية بالإضافة إلى تصنيفات متعددة الفئات.
- يعمل بشكل جيد في التنبؤات متعددة الفئات مقارنة بالخوارزميات الأخرى.
- إنه الخيار الأكثر شيوعًا لمشاكل تصنيف النص.

مساوئ مصنف بايز :

- يفترض Naive Bayes أن جميع الميزات مستقلة أو غير مرتبطة ، لذلك لا يمكنها معرفة العلاقة بين الميزات.

مصنف K الجار الأقرب K-Nearest Neighbor(KNN)

K–Nearest Neighbor هي واحدة من أبسط خوارزميات التعلم الآلي القائمة على تقنية التعلم الخاضع للإشراف.

تفترض خوارزمية K–NN التشابه بين الحالة / البيانات الجديدة والحالات المتاحة وتضع الحالة الجديدة في الفئة الأكثر تشابهًا مع الفئات المتاحة.

تقوم خوارزمية K–NN بتخزين جميع البيانات المتاحة وتصنيف نقطة بيانات جديدة بناءً على التشابه.

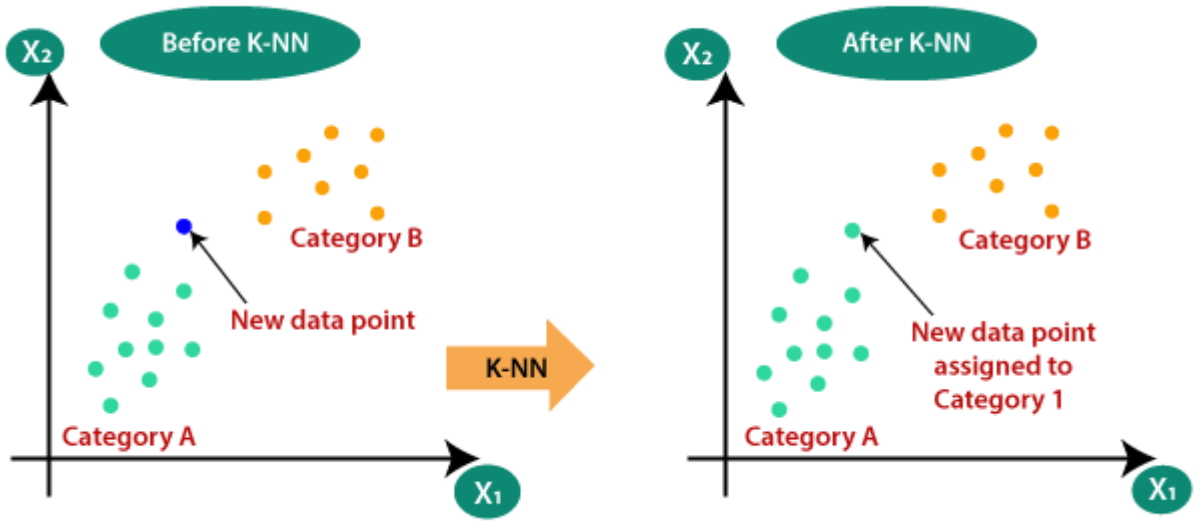
هذا يعني أنه عند ظهور بيانات جديدة ، يمكن تصنيفها بسهولة إلى فئة مجموعة جيدة باستخدام خوارزمية K– NN.

يمكن استخدام خوارزمية K–NN للانحدار وكذلك للتصنيف ولكن في الغالب يتم استخدامه لمشاكل التصنيف.

K–NN هي خوارزمية غير معلمية ، مما يعني أنها لا تقوم بأي افتراض على البيانات الأساسية.

يُطلق عليها أيضًا خوارزمية المتعلم الكسول لأنها لا تتعلم من مجموعة التدريب على الفور بدلاً من ذلك تقوم بتخزين مجموعة البيانات وفي وقت التصنيف ، تقوم بتنفيذ إجراء على مجموعة البيانات.

تقوم خوارزمية KNN في مرحلة التدريب فقط بتخزين مجموعة البيانات وعندما تحصل على بيانات جديدة ، فإنها تصنف تلك البيانات في فئة تشبه إلى حد كبير البيانات الجديدة.



طريقة عمل الخوارزمية :

يمكن شرح عمل K–NN على أساس الخوارزمية التالية:

الخطوة 1: حدد عدد K من الجيران

الخطوة 2: احسب المسافة الإقليدية لعدد K من الجيران

الخطوة 3: خذ أقرب جيران K وفقًا للمسافة الإقليدية المحسوبة.
الخطوة 4: من بين هؤلاء الجيران k ، احسب عدد نقاط البيانات في كل فئة.
الخطوة 5: قم بتعيين نقاط البيانات الجديدة إلى تلك الفئة التي يكون عدد الجار فيها الحد الأقصى.
الخطوة 6: نموذجنا جاهز.

مزايا خوارزمية الجار الأقرب :

- إنها سهلة التنفيذ.
- إنها قوية لبيانات التدريب المشتتة .
- يمكن أن تكون أكثر فعالية إذا كانت بيانات التدريب كبيرة.

مساوئ خوارزمية الجار الأقرب :

- تحتاج دائمًا إلى تحديد قيمة K والتي قد تكون معقدة في بعض الاوقات.
- تكلفة الحساب عالية بسبب حساب المسافة بين نقاط البيانات لجميع عينات التدريب.

الفصل الثالث

القسم العملي

الحالة هي أن لدى البنك بيانات العملاء ذات الخصائص المختلفة. قامت الإدارة ببناء منتج جديد – القرض الشخصي ، وقامت بحملة صغيرة لبيع المنتج الجديد لعملائها. بعد مرور بعض الوقت ، حصل 9٪ من العملاء على قرض شخصي من البنك .

سنقوم بتحليل البيانات لتحديد العملاء المحتملين الذين لديهم احتمالية أكبر لشراء القرض .

تتعلق هذه الحالة ببنك لديه قاعدة عملاء متنامية . غالبية هؤلاء العملاء هم عملاء ذوو مسؤولية (مودعون) مع حجم ودائع متفاوتة . عدد العملاء الذين هم أيضًا مقترضون (عملاء الأصول) صغير جدًا ، ويهتم البنك بتوسيع هذه القاعدة بسرعة لجلب المزيد من أعمال القروض وفي هذه العملية كسب المزيد من خلال الفوائد على القروض. على وجه الخصوص ، تريد الإدارة استكشاف طرق لتحويل عملاء المسؤولية إلى عملاء قروض شخصية (مع الاحتفاظ بهم كمودعين). أظهرت الحملة التي أطلقها البنك لعملاء المسؤولية معدل تحويل تجاوز 9٪. وقد شجع هذا قسم تسويق التجزئة على ابتكار حملات لتسويق أفضل المستهدف لزيادة نسبة النجاح بأقل ميزانية.

سنقوم ببناء نموذج يساعدهم في تحديد العملاء المحتملين الذين لديهم احتمالية أكبر لشراء القرض. سيؤدي ذلك إلى زيادة نسبة النجاح وفي نفس الوقت تقليل تكلفة الحملة.

توصيف قاعدة البيانات :

تتضمن مجموعة البيانات 5000 مشاهدة بأربعة عشر متغيرًا مقسمة إلى أربع فئات قياس مختلفة.

تحتوي فئة الفاصل الزمني على خمسة متغيرات: العمر والخبرة والدخل ومتوسط الانفاق على بطاقات الائتمان شهرياً والرهن العقاري . تحتوي الفئة الثنائية على خمسة متغيرات ، بما في ذلك القرض الشخصي المتغير المستهدف ، وكذلك حساب الأوراق المالية ، وحساب شهادة الايداع ، والخدمات المصرفية عبر الإنترنت ، وبطاقات الائتمان. تشمل الفئة الترتيبية المتغيرات الأسرة والتعليم. الفئة الأخيرة اسمية بالID والرمز البريدي. لا يضيف متغير الID أي معلومات مثيرة للاهتمام على سبيل المثال لا

يقدم الارتباط الفردي بين الشخص (المشار إليه بالرقم التعريفي) والقرض أي استنتاج عام لعملاء القرض المحتملين في المستقبل. لذلك ، سيتم إهماله في الفحص .

ID	Age	Exp	Inc	ZIP	Family	CCAvg	Edu	Mortgage	Personal Loan	Securities Account	CD Account	Online	CreditCard
1	25	1	49	91107	4	1.6	1	0	0	1	0	0	0
2	45	19	34	90089	3	1.5	1	0	0	1	0	0	0
3	39	15	11	94720	1	1.0	1	0	0	0	0	0	0
4	35	9	100	94112	1	2.7	2	0	0	0	0	0	0
5	35	8	45	91330	4	1.0	2	0	0	0	0	0	1

معنى المتغيرات :

العمود الأول (ID) : ال ID الخاص بكل عميل .

العمود الثاني (Age) : عمر كل عميل .

العمود الثالث (Experience) : عدد سنين الخبرة في العمل .

العمود الرابع (Income) : دخل كل عميل .

العمود الخامس (ZIP Code) : رمز عنوان العملاء .

العمود السادس (Family) : عدد افراد الاسرة .

العمود السابع (CCAVG) : متوسط الصرف الشهري من بطاقة الائتمان (1000 دولار).

العمود الثامن (Education) : درجة التعليم (1 = غير متخرج , 2 = متخرج, 3 = مختص).

العمود التاسع (Mortgage) : قيمة رهن المنزل (ان وجد) .

العمود العاشر (Personal Loan) : هل قبل العميل القرض الذي عرض عليه في الحملة الماضية ؟

العمود الحادي عشر (Securities Account) : هل لدى العميل حساب للأوراق المالية مع البنك ؟

العمود الثاني عشر (CD Account) : هل لدى العميل حساب إيداع مع البنك ؟

العمود الثالث عشر (Online) : هل يستخدم العميل الخدمات المصرفية عبر الانترنت ؟

العمود الرابع عشر (Credit Card) : هل يستخدم العميل بطاقة ائتمان من البنك ؟

الأدوات المستخدمة في التنقيب عن المعطيات :

أداة Jupyter Notebook أحد أشهر وأهم الأدوات المُستخدمة أثناء تحليل البيانات. بالإضافة إلى ذلك، هي شيء أساسي لا يستغني عنها عالم البيانات، وذلك لما تُقدمه من مميزات وخصائص تُسهل من التعامل مع البيانات والشيفرة البرمجية. تُعتبر Jupyter أداة قوية يُمكن استخدامها تفاعليًا في مشاريع علم البيانات وتعليم الآلة.

و هو يدعم اكثر من 40 لغة برمجة بما في ذلك Python , R , Julia, Scala و اكثر من ذلك (مما يعني انه يمكنه تشغيل كود مكتوب بجميع هذه اللغات) .

و سيتم استخدام لغة Python 3 في حالتنا هذه .

المكتبات الخاصة بعملية التنقيب في البيانات :

Pandas

Numpy

Seaborn

Matplotlib

التنقيب في المعطيات :

سنقوم هنا بالبدء في عملية التنقيب في المعطيات عن طريق أداة Jupyter Notebook و استخدام لغة البرمجة Python 3 و و تطبيق Association Rules , Decision tree , Classification, لمعرفة الزبائن الأنسب لتقديم العروض لهم .

أولاً :

تحضير البيانات :

سنقوم بتهيئة التطبيق و استيراد المكتبات اللازمة :

```
import os , sys
import pandas as pd
import numpy as np

#get rid of future warnings with seaborn
import warnings
warnings.simplefilter(action='ignore', category=FutureWarning)
```

تم استيراد مكتبات Pandas و Numpy و Warning .

Numpy : للتعامل مع المصفوفات العددية .

Pandas : لقراءة ملفات ال csv و xlsx الخاصة ب اكسل .

Warning : للتعامل مع التحذيرات و الاستثناءات .

و الان سنقوم بتغيير مكان عمود (القرض الشخصي) و وضعه في الأخير لأنه العمود المستهدف و لراحة اكبر .

```
a = master['Personal Loan']
master.drop('Personal Loan', axis = 1, inplace = True)
master['Personal Loan'] = a
```

سوف نعرض الجدول بأول نتيجة للتأكد من صحة تطبيق الأمر .

```
master.head(1)
```

و كانت المعطيات كما يلي :

ID	Age	Exp	Inc	ZIP	Family	CCAvg	Edu	Mortgage	Securities Account	CD Account	Online	CreditCard	Personal Loan
1	25	1	49	91107	4	1.6	1	0	1	0	0	0	0

ثانياً :

الوصول للبيانات و تنظيفها :

سنقوم بالكشف عن أنواع البيانات و رؤية ما اذا كان يوجد قيم فارغة :

```
df = master.copy()
df.info()
```

و كانت المعطيات كما يلي :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 14 columns):
 #   Column              Non-Null Count  Dtype
---  ------              --------------  -----
 0   ID                  5000 non-null   int64
 1   Age                 5000 non-null   int64
 2   Experience          5000 non-null   int64
 3   Income              5000 non-null   int64
 4   ZIP Code            5000 non-null   int64
 5   Family              5000 non-null   int64
 6   CCAvg               5000 non-null   float64
 7   Education           5000 non-null   int64
 8   Mortgage            5000 non-null   int64
 9   Securities Account  5000 non-null   int64
 10  CD Account          5000 non-null   int64
 11  Online              5000 non-null   int64
```

```
 12  CreditCard          5000 non-null   int64
 13  Personal Loan       5000 non-null   int64
dtypes: float64(1), int64(13)
memory usage: 547.0 KB
```

- لا يوجد قيم فارغة او قيم مفقودة كما رأينا في الجدول .

و الان سنتأكد من ان البيانات نظيفة :

```
df.describe().transpose()
```

و كانت المعطيات كما يلي :

	count	mean	std	min	25%	50%	75%	max
ID	5000.0	2500.500000	1443.520003	1.0	1250.75	2500.5	3750.25	5000.0
Age	5000.0	45.338400	11.463166	23.0	35.00	45.0	55.00	67.0
Experience	5000.0	20.104600	11.467954	-3.0	10.00	20.0	30.00	43.0
Income	5000.0	73.774200	46.033729	8.0	39.00	64.0	98.00	224.0
ZIP Code	5000.0	93152.503000	2121.852197	9307.0	91911.00	93437.0	94608.00	96651.0
Family	5000.0	2.396400	1.147663	1.0	1.00	2.0	3.00	4.0
CCAvg	5000.0	1.937913	1.747666	0.0	0.70	1.5	2.50	10.0
Education	5000.0	1.881000	0.839869	1.0	1.00	2.0	3.00	3.0
Mortgage	5000.0	56.498800	101.713802	0.0	0.00	0.0	101.00	635.0

	count	mean	std	min	25%	50%	75%	max
Securities Account	5000.0	0.104400	0.305809	0.0	0.00	0.0	0.00	1.0
CD Account	5000.0	0.060400	0.238250	0.0	0.00	0.0	0.00	1.0
Online	5000.0	0.596800	0.490589	0.0	0.00	1.0	1.00	1.0
CreditCard	5000.0	0.294000	0.455637	0.0	0.00	0.0	1.00	1.0
Personal Loan	5000.0	0.096000	0.294621	0.0	0.00	0.0	0.00	1.0

- المتغير Experience يوجد فيه بيانات سالبة كما رأينا في العمود min و تحتاج للتصليح .
- المتحولات الأخرى كلها نظيفة و جاهزة للتحليل .

الان سوف نرى كمية البيانات السلبية في المتغير Experience .

```
df[df['Experience'] < 0]['Experience'].value_counts()
```

و كانت المعطيات كما يلي :

```
-1    33
-2    15
-3     4
Name: Experience, dtype: int64
```

لدينا 33 قيمة -1 و 15 قيمة -2 و 4 قيم -3 .

- سوف نرى ما اذا كان هناك متغير ارتباطه قوي بالمتغير Experience نستطيع ان نملأ القيم السلبية بمساعدته .
- اولاً سنقوم باستيراد المكاتب التالية : Matplotlib و Seaborn .
- Matplotlib : للتعامل مع الرسوم البيانية .

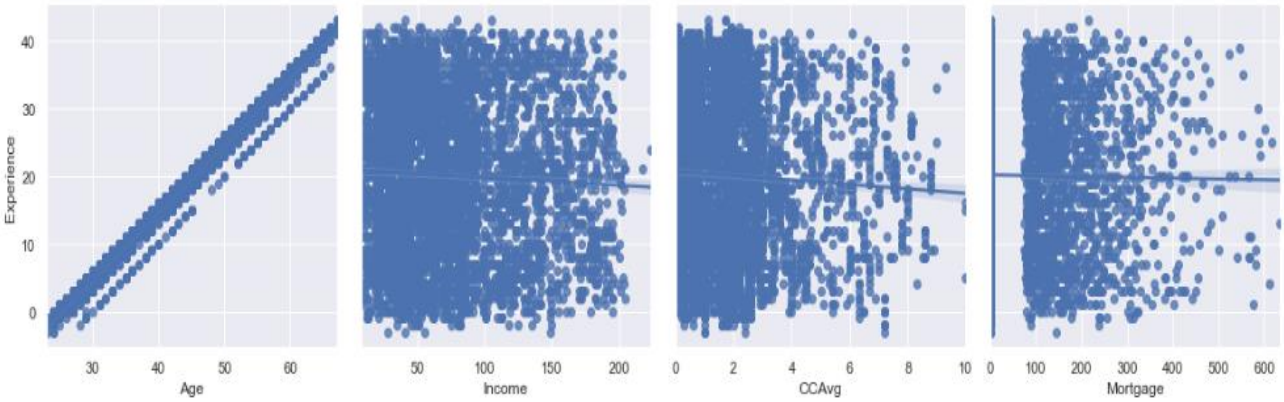
- Seaborn : للتعامل مع الرسوم الاحصائية .

```
import matplotlib.pyplot as plt
import seaborn as sns
sns.set(color_codes = True)
```

و الان سوف نجد المتغيرات الكمية التي لها ارتباط قوي مع المتغير Experience .

```
ncol = ['Age', 'Income','CCAvg', 'Mortgage']
grid = sns.PairGrid(df, y_vars = 'Experience', x_vars = ncol, height = 4)
grid.map(sns.regplot);
```

و كانت المعطيات كما يلي :



كما نرى انه يوجد متغير وحيد له ارتباط قوي مع المتغير Experience و هو المتغير Age .

- و الان لنحل مشكلة القيم السلبية يمكننا ان نستبدل كل قيمة سلبية بمتوسط القيمة الإيجابية المرتبطة بمتغير العمر Age .

اولاً سنقوم بإيجاد كل الاعمار التي كانت فيها قيم سلبية .

```
ages = df[df['Experience'] < 0]['Age'].unique().tolist()
ages
```

و كانت المعطيات كما يلي :

[25, 24, 28, 23, 29, 26]

بعدها سنحصل على الفهارس (Indexes) ذات القيم السالبة :

```
indexes = df[df['Experience'] < 0].index.tolist()
```

و الان سنقوم بتبديل القيم السالبة في متغير الExperience بمتوسط الخبرة في كل عمر .

```
for i in indexes:
    for x in ages:
        df.loc[i,'Experience'] = df[(df.Age == x) & (df.Experience > 0)].Experience.mean()
```

و الان سنقوم بالاختبار :

```
df.Experience.describe()
```

و كانت المعطيات كما يلي :

```
count    5000.000000
mean       20.135743
std        11.413140
min         0.000000
25%        10.000000
50%        20.000000
75%        30.000000
max        43.000000
Name: Experience, dtype: float64
```

لا يوجد ولا قيمة سلبية .

- الان كل البيانات إيجابية و نظيفة و نستطيع ان نبدأ بالتحليل .

ثالثاً :

تحليل البيانات :

سنحاول البحث عن أجوبة بحث :

اولاً- هل هناك ارتباط ما بين السمات الشخصية وحقيقة حصول ذلك الشخص على قرض شخصي ؟

دعنا نتحقق من القيم أو مجموعة القيم لكل متغير التي لديها ارتباط بال "قرض شخصي" .

نظرًا لأننا وجدنا ارتباطًا قويًا بين "العمر" و "الخبرة" ، فقد قررنا استبعاد "الخبرة" من خطوات التحليل لتجنب العلاقات الخطية المتعددة (Multicollinearity).

المتغيرات الكمية (Quantitative Variables) :

['Age', 'Income', 'CCAvg', 'Mortgage']

سنقوم بوضعها بمتحول واحد :

```
quant_df = df[['Personal Loan', 'Age', 'Income', 'CCAvg', 'Mortgage']].copy()
```

و سنعمل جدول ارتباط :

```
quant_df.corr()
```

و كانت المخرجات على الشكل التالي :

	Personal Loan	Age	Income	CCAvg	Mortgage
Personal Loan	1.000000	−0.007726	0.502462	0.366891	0.142095
Age	−0.007726	1.000000	−0.055269	−0.052030	−0.012539
Income	0.502462	−0.055269	1.000000	0.645993	0.206806
CCAvg	0.366891	−0.052030	0.645993	1.000000	0.109909
Mortgage	0.142095	−0.012539	0.206806	0.109909	1.000000

الان سنقوم بالحصول على معاملات الارتباط لمتغير الPersonal Loan و استبعاده من السلسلة :

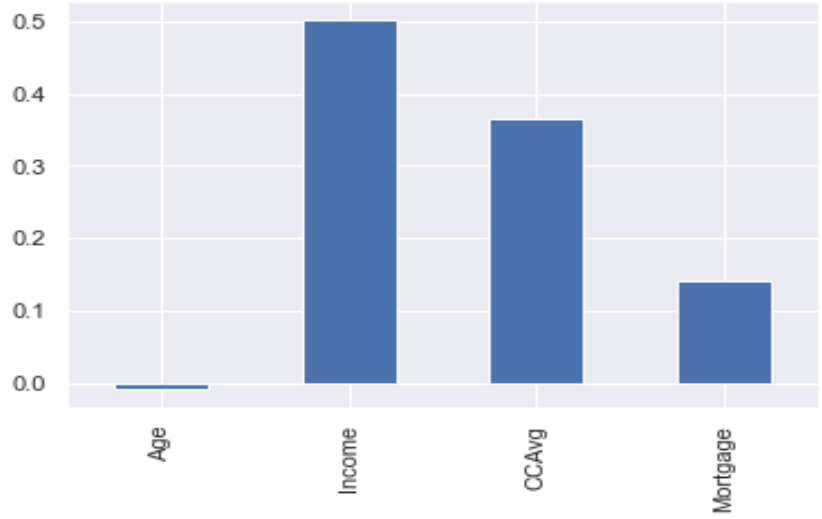
```
quant_df.corr()['Personal Loan'][1:]
```

و كانت المعطيات على الشكل التالي :

```
Age        -0.007726
Income      0.502462
CCAvg       0.366891
Mortgage    0.142095
Name: Personal Loan, dtype: float64
```

سنقوم بالنمذجة بيانياً الان :

```
quant_df.corr()['Personal Loan'][1:].plot.bar();
```



- نلاحظ ان المتغيرين Age و Mortgage لديهم ارتباط ضعيف بمتغير ال Personal Loan .
- و المتغيرين Income و CCAvg لديهم ارتباط بالمتغير Personal Loan .

سنقوم بالتحقق من ثقة Confidence الجملتين السابقتين باستخدام نموذج انحدار :

```
quant_df['intercept'] = 1
log_mod = sm.Logit(quant_df['Personal Loan'], quant_df[['intercept', 'Age', 'Income', 'CCAvg', 'Mortgage']]).fit()
```

```
log_mod.summary()
```

و كانت المخرجات على الشكل التالي :

Logit Regression Results

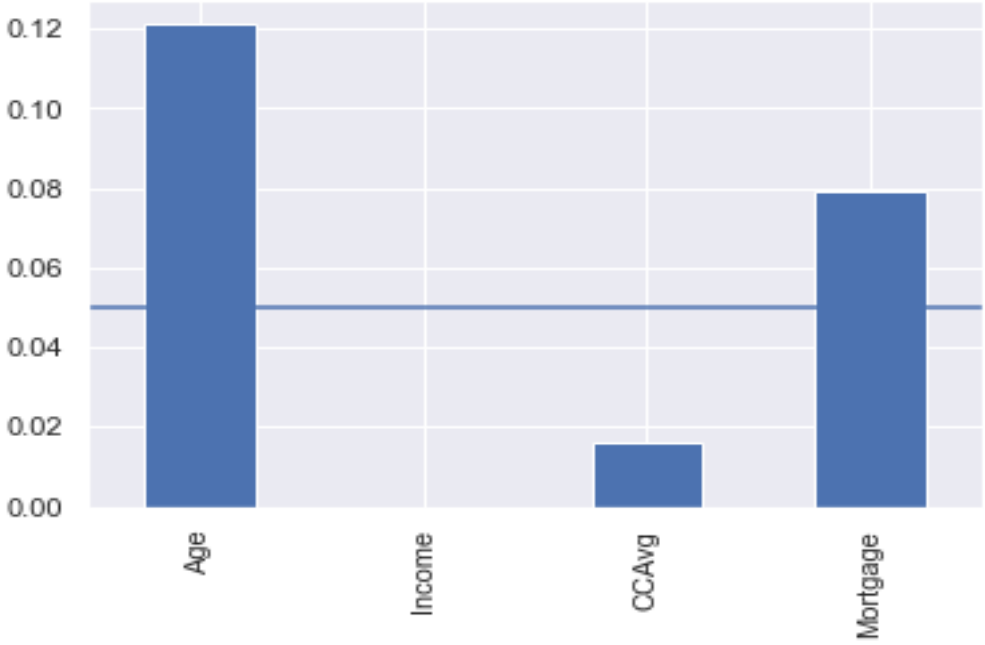
Dep. Variable:	Personal Loan	**No. Observations:**	5000
Model:	Logit	**Df Residuals:**	4995
Method:	MLE	**Df Model:**	4
Date:	Sun, 15 May 2022	**Pseudo R-squ.:**	0.3657
Time:	18:10:15	**Log-Likelihood:**	-1002.9
converged:	True	**LL-Null:**	-1581.0
Covariance Type:	nonrobust	**LLR p-value:**	4.743e-249

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-6.5144	0.308	-21.155	0.000	-7.118	-5.911
Age	0.0080	0.005	1.550	0.121	-0.002	0.018
Income	0.0351	0.002	22.313	0.000	0.032	0.038
CCAvg	0.0688	0.029	2.409	0.016	0.013	0.125
Mortgage	0.0007	0.000	1.757	0.079	-8.49e-05	0.002

سنقوم بعمل مخطط لل P-Values لتبسيط القراءة :

```
log_mod.pvalues[1:5].plot.bar()
plt.axhline(y = 0.05);
```

و كانت المعطيات على الشكل التالي :



يمكننا أن نقول بثقة أن كلا من "الدخل" و "CCAvg" لهما علاقة ذات دلالة إحصائية مع "القرض الشخصي" ، نظرًا لأن قيمتهما الاحتمالية في الانحدار اقل من 0.05 .

الان سنقوم بحفظ المتغيرات و القيم التي قيمة انحدارها اقل من 0.05 في Dictionary .

```
quant_df_main = {}
for i in log_mod.params[1:5].to_dict().keys():
    if log_mod.pvalues[i] < 0.05:
        quant_df_main[i] = log_mod.params[i]
    else:
        continue
```

سنقوم الان بقياس الاحتمالات لكل من المتغيرين Inome و CCAvg :

```
quant_df_main_odds = {k : np.exp(v) for k, v in quant_df_main.items()}
```

```
quant_df_main_odds
```

و قد حصلنا على المخرجات التالية :

{'Income': 1.0357095990302452, 'CCAvg': 1.0712155752174697}

نستخلص :

القرض الشخصي له علاقة ذات دلالة إحصائية بـ :

- Income : coef = 0.03508
- CCAvg : coef = 0.06879

كما رأينا في نموذج الانحدار .

كلا المتغيرين مرتبطان بشكل إيجابي بمتغير ال Personal Loan. بمجرد أن يحتوي كلاهما على وحدة واحدة بقيمة 1000 دولار ، يمكننا أن نقول ما يلي :

- كل زيادة بمقدار 1000$ في متغير الIncome ينتج عنه زيادة بمقدار 3.58% في احتمالية بيع القرض الشخصي , مع ثبات العوامل الأخرى .
- كل زيادة بمقدار 1000$ في متغير الCCAvg ينتج عنه زيادة بمقدار 7.12% في احتمالية بيع القرض الشخصي , مع ثبات العوامل الأخرى .

المتغيرات الصنفية (Categorical Variables):

ZIP Code, Family, Education

المتغيرين Family و Education متغيرين ترتيبيين فئويين ، لذا يمكننا تطبيق الانحدار . مباشرة عليهم. Zip Code هو متغير اسمي (Nominal) ، و لا نعتقد ان يكون له علاقة ببيع ال "قرض الشخصي" لذلك سيتم استعاده من الاختبار .

- اولاً سنقوم بوضع المتغيرات ضمن متغير واحد .

```
cat_df = df[[ 'Family', 'Education', 'Personal Loan']].copy()
```

و الان سنطبق اختبار الارتباط على المتغيرين Family و Education مع ال Personal Loan.

```
cat_df.corr()
```

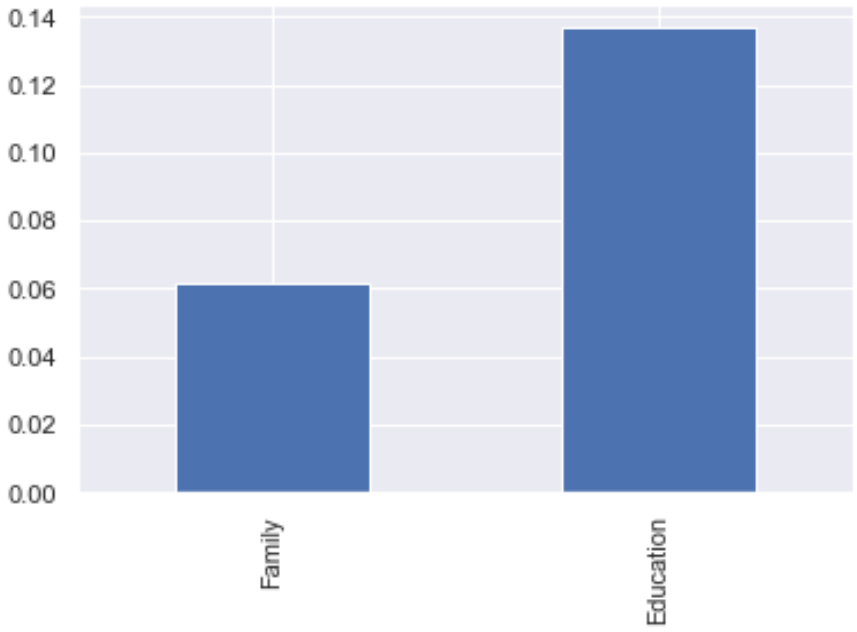
و كانت المخرجات على الشكل التالي :

	Family	Education	Personal Loan
Family	1.000000	0.064929	0.061367
Education	0.064929	1.000000	0.136722
Personal Loan	0.061367	0.136722	1.000000

سنقوم بنمذجتها بيانياً لسهولة القراءة :

```
cat_df.corr()['Personal Loan'][0:2].plot.bar();
```

و كان البيان على الشكل التالي :



يرتبط Family و Education ارتباطًا ضعيفًا بـ Personal Loan ، سنقوم بالتحقق من ثقة Confidence الارتباط باستخدام نموذج انحدار :

```
log_mod.summary()
```

و كانت المخرجات على الشكل التالي :

Logit Regression Results

Dep. Variable:	Personal Loan	**No. Observations:**	5000
Model:	Logit	**Df Residuals:**	4997
Method:	MLE	**Df Model:**	2
Date:	Mon, 16 May 2022	**Pseudo R-squ.:**	0.03415
Time:	15:25:52	**Log-Likelihood:**	-1527.0
converged:	True	**LL-Null:**	-1581.0
Covariance Type:	nonrobust	**LLR p-value:**	3.575e-24

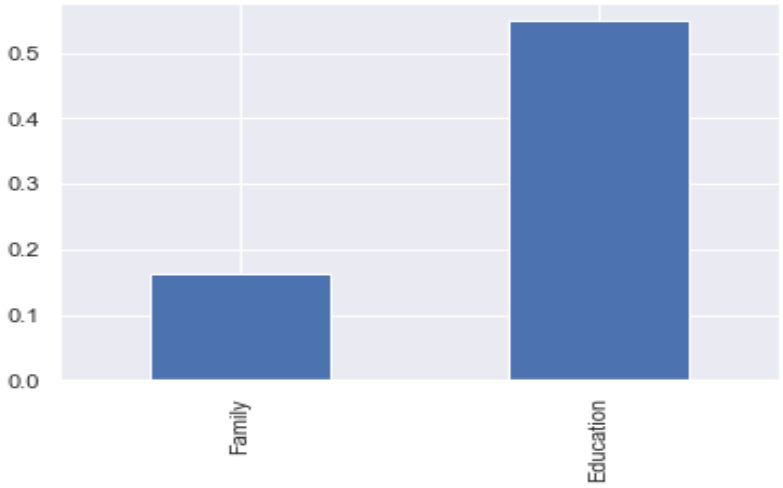
	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-3.7670	0.175	-21.574	0.000	-4.109	-3.425
Family	0.1623	0.042	3.863	0.000	0.080	0.245
Education	0.5487	0.059	9.260	0.000	0.433	0.665

يمكننا أن نقول أن كلا من Family و Education لهما علاقة إحصائية مهمة مع Personal Loan ، نظرًا لأن قيمتهما الاحتمالية في الانحدار اللوجستي اقل من 0.05 .

سنقوم بعمل مخطط لل P-Values لتبسيط القراءة :

```
log_mod.params[1:3].plot.bar();
```

و كان المخطط على الشكل التالي :



المتغير Education لديه ارتباط اقوى من المتغير Family مع المتغير Personal Loan .

- الان سنقوم بحفظ المتغيرات و القيم التي قيمة انحدارها اقل من 0.05 في Dictionary .

```
cat_df_main = {}
for i in log_mod.params[1:3].to_dict().keys():
    if log_mod.pvalues[i] < 0.05:
        cat_df_main[i] = log_mod.params[i]
    else:
        continue
```

سنقوم الان بقياس الاحتمالات لكل من المتغيرين Family و Education :

```
cat_df_odds = {k : np.exp(v) for k, v in cat_df_main.items()}
```

```
cat_df_odds
```

و كانت المخرجات على الشكل التالي :

{'Family': 1.1762340487826346, 'Education': 1.7310508695002493}

نستخلص :

القرض الشخصي له علاقة ذات دلالة إحصائية بـ :

- Family : coef = 0.16231
- Education : coef = 0.54873

كما رأينا في نموذج الانحدار .

كلا المتغيرين مرتبطان بشكل إيجابي بـ "القرض الشخصي". يمكننا ان نقول ما يلي:

- كل زيادة بمقدار وحدة واحدة في متغير الFamily ينتج عنه زيادة بمقدار 17.62% في احتمالية بيع القرض الشخصي , مع ثبات العوامل الأخرى .
- كل زيادة بمقدار وحدة واحدة في متغير الEducation ينتج عنه زيادة بمقدار 73.10% في احتمالية بيع القرض الشخصي , مع ثبات العوامل الأخرى .

المتغيرات الثنائية (Binary Variables) :

'Securities Account', 'CD Account', 'Online', 'Credit Card'

سنقوم بوضعها ضمن متحول واحد :

```
bin_df = df[['Personal Loan', 'Securities Account', 'CD Account', 'Online', 'CreditCard']].copy()
```

و سنعمل جدول ارتباط :

```
bin_df.corr()
```

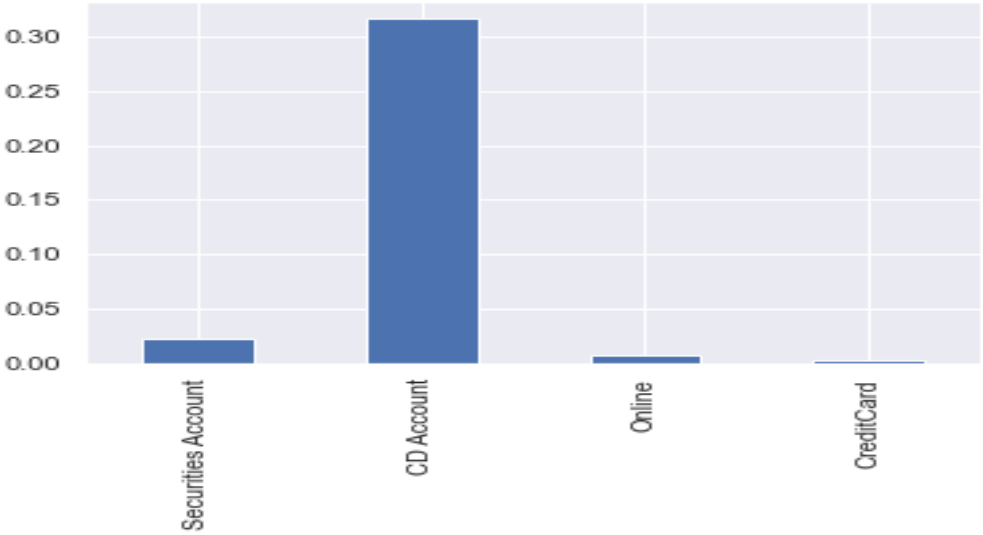
و كانت المخرجات على الشكل التالي :

	Personal Loan	Securities Account	CD Account	Online	CreditCard
Personal Loan	1.000000	0.021954	0.316355	0.006278	0.002802
Securities Account	0.021954	1.000000	0.317034	0.012627	−0.015028
CD Account	0.316355	0.317034	1.000000	0.175880	0.278644
Online	0.006278	0.012627	0.175880	1.000000	0.004210
CreditCard	0.002802	−0.015028	0.278644	0.004210	1.000000

سوف نقوم بالنمذجة بيانياً لتبسيط القراءة :

```
bin_df.corr()['Personal Loan'][1:].plot.bar();
```

و كانت المخرجات على الشكل التالي :



المتغير CD Account هو المتغير الوحيد الذي لديه ارتباط متوسط مع ال Personal Loan .

- و الان سنقوم بقياس الانحدار للمتغير CD Account مع المتغير Personal Loan :

```
log_mod = sm.Logit(bin_df['Personal Loan'], bin_df[['intercept', 'CD Account']]).fit()
```

```
log_mod.summary()
```

و كانت المعطيات على الشكل التالي :

Logit Regression Results

Dep. Variable:	Personal Loan	**No. Observations:**	5000
Model:	Logit	**Df Residuals:**	4998
Method:	MLE	**Df Model:**	1
Date:	Tue, 17 May 2022	**Pseudo R-squ.:**	0.09632
Time:	15:29:54	**Log-Likelihood:**	-1428.7
converged:	True	**LL-Null:**	-1581.0
Covariance Type:	nonrobust	**LLR p-value:**	3.340e-68

	coef	std err	z	P>|z|	[0.025	0.975]
intercept	-2.5508	0.056	-45.301	0.000	-2.661	-2.440
CD Account	2.4049	0.128	18.730	0.000	2.153	2.657

يمكننا أن نقول أن CD Account لها علاقة إحصائية مهمة مع Personal Loan ، نظرًا لأن قيمتهما الاحتمالية في الانحدار اللوجستي اقل من 0.05 .

سنقوم بقياس الاحتمالات الان :

```
bin_odds = {'CD Account' : np.exp(log_mod.params[1])}
```

```
bin_odds
```

و كانت المخرجات على الشكل التالي :

{'CD Account': 11.076978939724048}

نستخلص :

القرض الشخصي له علاقة ذات دلالة إحصائية بـ :

- CD Account : coef = 2.40

كما رأينا في نموذج الانحدار .

المتغير مرتبط بشكل إيجابي بـ "القرض الشخصي". يمكننا ان نقول ما يلي:

- نظرًا لأن العميل كان لديه حساب CD لدى البنك ، نتوقع أن تزداد احتمالات بيع القرض الشخصي 10 مرات ، مع ثبات العوامل الأخرى .

الخلاصة :

القرض الشخصي له علاقة ذات دلالة إحصائية بـ :

- Income : coef = 0.03508
- CCAvg : coef = 0.06879
- Family : coef = 0.16231
- Education : coef = 0.54873
- CD Account : coef = 2.40

كل المتغيرات السابقة مرتبطة بشكل إيجابي بـ "القرض الشخصي". يمكننا ان نقول ما يلي:

- كل زيادة بمقدار $1000 في متغير الIncome ينتج عنه زيادة بمقدار 3.58% في احتمالية بيع القرض الشخصي , مع ثبات العوامل الأخرى .
- كل زيادة بمقدار $1000 في متغير الCCAvg ينتج عنه زيادة بمقدار 7.12% في احتمالية بيع القرض الشخصي , مع ثبات العوامل الأخرى .
- كل زيادة بمقدار وحدة واحدة في متغير الFamily ينتج عنه زيادة بمقدار 17.62% في احتمالية بيع القرض الشخصي , مع ثبات العوامل الأخرى .
- كل زيادة بمقدار وحدة واحدة في متغير الEducation ينتج عنه زيادة بمقدار 73.10% في احتمالية بيع القرض الشخصي , مع ثبات العوامل الأخرى .
- نظرًا لأن العميل كان لديه حساب CD لدى البنك ، نتوقع أن تزداد احتمالات بيع القرض الشخصي 10 مرات ، مع ثبات العوامل الأخرى .

بمجرد أن وجدنا أن "القرض الشخصي" يعتمد على خمس خصائص رئيسية ، فسنقوم بتقسيم إطار البيانات الخاص بنا وإلقاء نظرة فاحصة على البيانات.

فيما يلي مجموعة فرعية من إطار البيانات الأولية ذات الخصائص فقط التي لها ارتباط إيجابي بـ "القرض الشخصي" وحجم الارتباط أعلى من المتوسط .

```
exp_df = df[['Income', 'CCAvg', 'Family', 'Education', 'CD Account', 'Personal Loan']].copy()
```

تم تشكيل المجموعة و الان سنقوم باستعراض اول نتيجتين للتأكد :

```
exp_df.head(2)
```

و كانت المخرجات على الشكل التالي :

	Income	CCAvg	Family	Education	CD Account	Personal Loan
0	49	1.6	4	1	0	0
1	34	1.5	3	1	0	0

و الان سنطبق تحليل الانحدار على المجموعة الفرعية :

```
log_mod = sm.Logit(exp_df['Personal Loan'], exp_df[['intercept','Income', 'CCAvg', 'Family', 'Education', 'CD Account']]).fit()
```

```
log_mod.pvalues[1:]
```

و كانت المخرجات على الشكل التالي :

```
Income          1.568180e-104
CCAvg            4.872101e-03
Family          2.077119e-21
Education        2.509902e-53
CD Account       1.430875e-28
dtype: float64
```

و الان سنقوم بحساب الاحتمالات :

```
odds = np.exp(log_mod.params)
```

```
odds
```

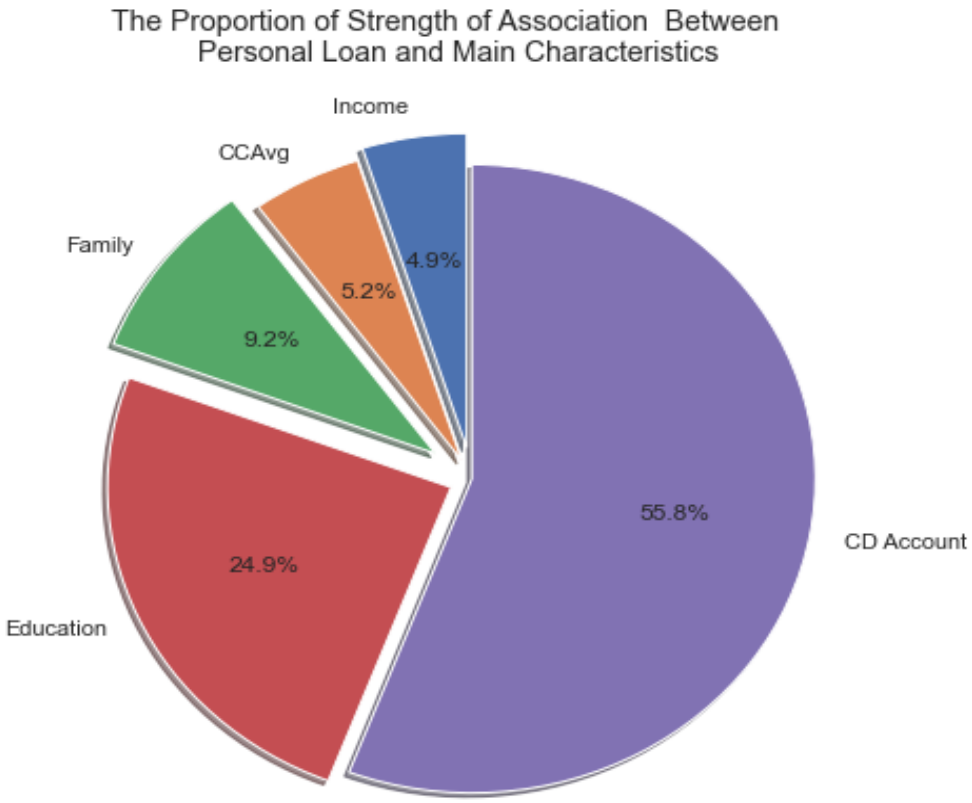
و كانت المخرجات على الشكل التالي :

```
intercept       0.000001
Income           1.056149
CCAvg            1.114028
Family          1.991695
Education        5.352795
CD Account      12.021627
dtype: float64
```

و الان سنقوم بعرض مخططاً يوضح نسبة قوة الارتباط بين ال Personal Loan و الخصائص الرئيسية الأخرى .

```
sizes = odds_df.Odds.tolist()# list of sizes of slices
labels = odds_df.index.tolist() # list of labels
explode = (0.15, 0.1, 0.2, 0.1, 0)  # "explode" the 2nd and 3rd slices
fig = plt.figure(figsize=(10, 5))
plt.suptitle('The Proportion of Strength of Association  Between  \n Personal Loan and Main Characteristics', \
          fontsize = 14, y = 1.18)
plt.axis('equal'); # set aspect ration as equal to make sure the pie is drawn as a circle
plt.pie(sizes, labels = labels, explode = explode, radius = 1.5, \
        shadow = True, startangle = 90,autopct= '%1.1f%%')

plt.savefig('proportion_of_stregth_of_association.png', bbox_inches = 'tight');
```

و كانت المعطيات على الشكل التالي :



مما يعني ان اكثر ما يؤثر على عملية شراء القرض Personal Loan هو بالترتيب :

1- CD Account بنسبة 55.8% .

2- Education بنسبة 24.9% .

3- Family بنسبة 9.2% .

4- CCAvg بنسبة 5.2% .

5- Income بنسبة 4.9 % .

الان سنقوم بالبحث عن جواب للسؤال الثاني من أسئلة البحث :

ثانياً – ما هي الشرائح من الخصائص الرئيسية التي لها ارتباط اقوى بال Personal Loan.

دعنا نلقي نظرة فاحصة على كل من الخصائص الرئيسية :

CD Account :

فيما يلي توزيع قيم Personal Loan بين مجموعات قيم CD Account :

```
series_cd = exp_df[exp_df['Personal Loan'] == 1]['CD Account'].value_counts()
series_cd
```

و كانت المخرجات على الشكل التالي :

```
0    340
1    140
Name: CD Account, dtype: int64
```

أي 140 من اصل 480 الذين اشتروا القرض كان لديهم CD Account .

الان سنعرض توزيع قيم الذين لم يشتروا القرض بين مجموعات قيم CD Account :

```
series_cdd = exp_df[exp_df['Personal Loan'] == 0]['CD Account'].value_counts()
series_cdd
```

و كانت المخرجات على الشكل التالي :

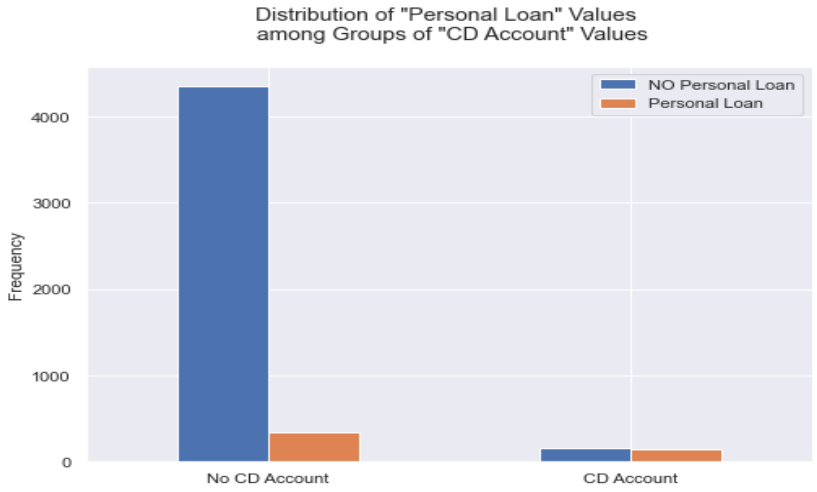
```
0    4358
1     162
Name: CD Account, dtype: int64
```

أي 162 عميل كان لديه CD Account و لم يقم بشراء القرض الشخصي .

و الان سنقوم بعرض النتائج بيانياً لتسهيل القراءة :

```
pd.DataFrame(dict( NO_PL= series_cdd, PL= series_cd,)).plot.bar(figsize = (8,6))
plt.ylabel('Frequency')
plt.xticks(np.arange(2),('No CD Account','CD Account'), rotation = 'horizontal')
plt.legend(('NO Personal Loan', 'Personal Loan'));
plt.title('Distribution of "Personal Loan" Values \n among Groups of "CD Account" Values', fontsize = 14, y = 1.05);
plt.savefig('distribution_of_PL_among_CDacc.png', bbox_inches = 'tight')
```

و كانت المخرجات على الشكل التالي :



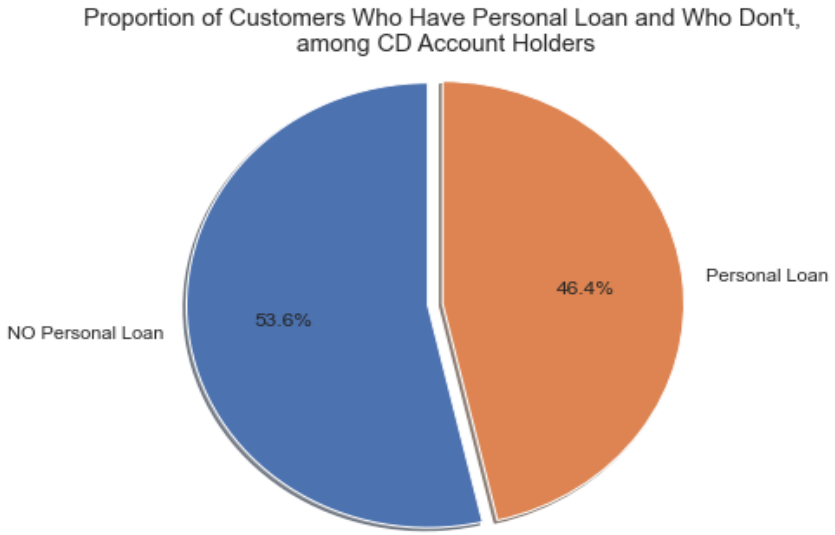
يمكننا ان نقول أن نسبة الأشخاص الذين لديهم قرض شخصي من بينهم ممن لديهم حساب CD مع البنك مرتفعة.

دعونا نرى العدد الدقيق لنسبة المقترضين بين المودعين .

```
series = exp_df[exp_df['CD Account'] == 1]['Personal Loan'].value_counts()
```

```
plt.axis('equal')
plt.title('Proportion of Customers Who Have Personal Loan and Who Don\'t,\n among CD Account Holders', \
          fontsize = 14, y = 1.2)
labels = ['NO Personal Loan','Personal Loan']
plt.pie(series, labels = labels,autopct= '%1.1f%%', shadow = True,explode = (0.1, 0), radius = 1.6, startangle = 90)
plt.savefig('Proportion_of_loanees_among_depositees.png', bbox_inches = 'tight');
```

و كانت المخرجات على الشكل التالي :



- 46.4% من حاملي CD Account قد قاموا بشراء القرض الشخصي .
- بما ان النسبتين متساويتين تقريباً سنقوم بأخذ القيمتين معاً .

: Education

فيما يلي توزيع قيم Personal Loan بين مجموعات قيم Education :

```
series_ed = exp_df[exp_df['Personal Loan'] == 1]['Education'].value_counts()
series_ed
```

و كانت المخرجات على الشكل التالي :

```
3    205
2    182
1     93
Name: Education, dtype: int64
```

أي 205 من العملاء قاموا بشراء القرض و كانت درجة تعليمهم 3 .

و 182 من العملاء قاموا بشراء القرض و كانت درجة تعليمهم 2 .

و 93 من العملاء قاموا بشراء القرض و كانت درجة تعليمهم 1 .

الان سنعرض توزيع قيم الذين لم يشتروا القرض بين مجموعات قيم Education :

```
series_edd = exp_df[exp_df['Personal Loan'] == 0]['Education'].value_counts()
series_edd
```

و كانت المخرجات على الشكل التالي :

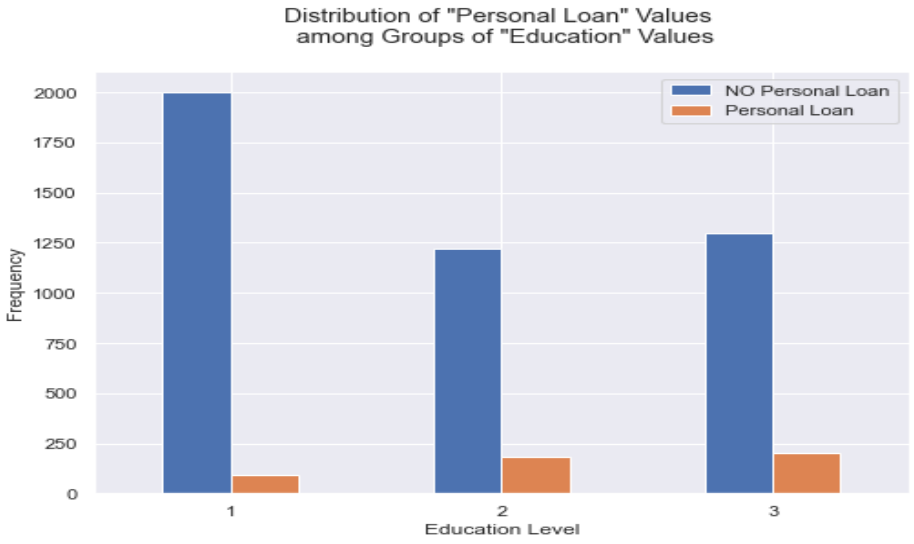
```
1    2003
3    1296
2    1221
Name: Education, dtype: int64
```

ان النسبة الأكبر التي لم تشتري القرض كانت درجة تعليمها 1 ثم 3 ثم 2 .

الان سنقوم بعرض النتائج بيانياً لتسهيل القراءة :

```
pd.DataFrame(dict(NO_PL= series_edd, PL= series_ed)).plot.bar(figsize = (8,6))
plt.ylabel('Frequency')
plt.xlabel('Education Level')
plt.xticks(np.arange(3),('1','2','3'), rotation = 'horizontal')
plt.legend(('NO Personal Loan', 'Personal Loan'))
plt.title('Distribution of "Personal Loan" Values \n among Groups of "Education" Values', fontsize = 14, y = 1.05);
plt.savefig('distribution_PL_among_Education.png', bbox_inches = 'tight')
```

و كانت المخرجات على الشكل التالي :



يمكننا ان نقول أن نسبة الأشخاص الذين لديهم قرض شخصي بينهم ممن لديهم المستوى الثالث والثاني من التعليم أعلى من النسبة بين الأشخاص الذين لديهم المستوى الأول من التعليم .

الان سنعرض الأرقام الصحيحة للنسب :

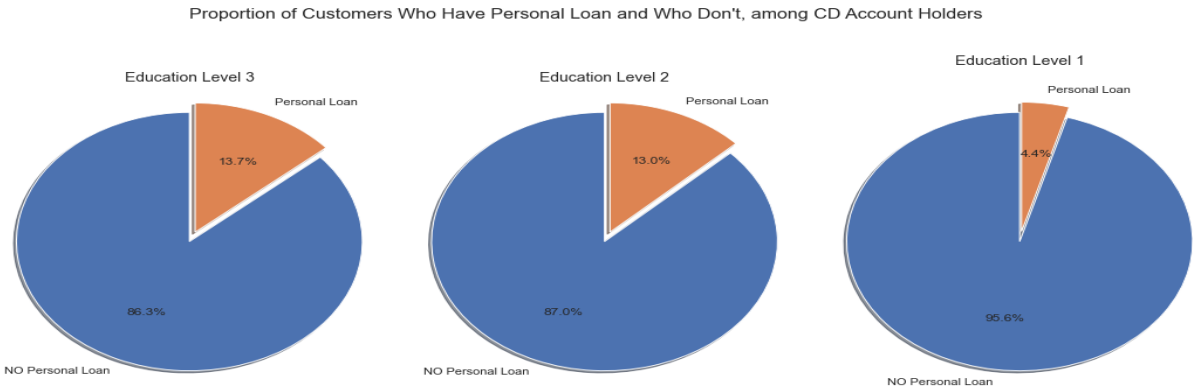
```
series_edu_3 = exp_df[exp_df['Education'] == 3]['Personal Loan'].value_counts()
```

```
series_edu_2 = exp_df[exp_df['Education'] == 2]['Personal Loan'].value_counts()
```

```
series_edu_1 = exp_df[exp_df['Education'] == 1]['Personal Loan'].value_counts()
```

```
labels = ['NO Personal Loan','Personal Loan']
fig, (ax1, ax2, ax3) = plt.subplots(1,3, figsize = (18,6),subplot_kw=dict(aspect="equal"))
plt.axis('equal')
ax1.pie(series_edu_3, labels = labels, autopct= '%1.1f%%', shadow = True,explode = (0, 0.1), radius = 1.25, startangle = 90)
ax1.set_title('Education Level 3',fontsize = 14, y = 1.1)

ax2.pie(series_edu_2, labels = labels, autopct= '%1.1f%%', shadow = True,explode = (0, 0.1), radius = 1.25, startangle = 90)
ax2.set_title('Education Level 2', fontsize = 14, y = 1.1)

ax3.pie(series_edu_1, labels = labels, autopct= '%1.1f%%', shadow = True,explode = (0, 0.1), radius = 1.25, startangle = 90);
ax3.set_title('Education Level 1',fontsize = 14, y = 1.1)

plt.suptitle('Proportion of Customers Who Have Personal Loan and Who Don\'t, among CD Account Holders', \
             fontsize = 16, y = 1.12);

plt.savefig('Proportion_of_PL_among edu_levels.png', bbox_inches = 'tight');
```

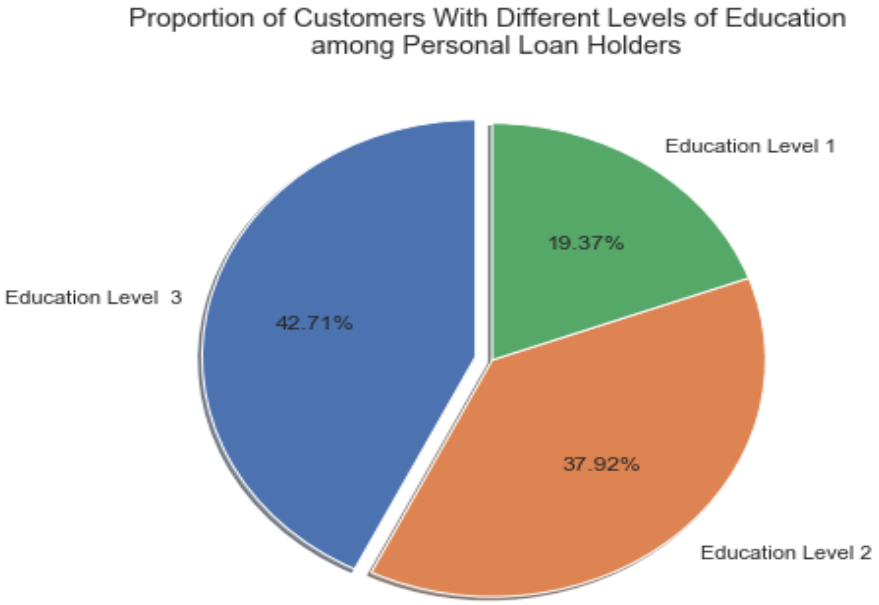
و كانت المعطيات على الشكل التالي :



الان سنعرض نسب الذين قاموا بشراء القرض فقط و عرضهم ضمن متغير واحد :

```
plt.axis('equal')
plt.title('Proportion of Customers With Different Levels of Education \n among Personal Loan Hol
ders', \
          fontsize = 14, y = 1.3)
labels = ['Education Level  3',' Education Level 2','Education Level 1']
plt.pie(series_edu_4, labels = labels, autopct= '%1.2f%%', shadow = True,explode = (0.1, 0, 0),
radius = 1.6, startangle = 90);
plt.savefig('Proportion_edu_levels_among_PL.png', bbox_inches = 'tight');
```

و كانت المعطيات على الشكل التالي :



- 42.7٪ و 37.9٪ من الأشخاص الذين لديهم قرض شخصي ، لديهم مستوى تعليمي 3 ومستوى 2 على التوالي.
- بالنسبة لمتغير ال Education – الشرائح الرئيسية لبيع القرض الشخصي هي الأشخاص الذين لديهم المستويان الثاني والثالث من التعليم .
- القيم المستهدفة لمتغير Education هي 3 و 2 بترتيب تنازلي .

Family :

فيما يلي توزيع قيم Personal Loan بين مجموعات قيم Family :

```
series_fam = exp_df[exp_df['Personal Loan'] == 1]['Family'].value_counts()
series_fam
```

و كانت المخرجات على الشكل التالي :

4 134
3 133
1 107
2 106
Name: Family, dtype: int64

ان النسبة الاكبر كانت للعملاء الذي يتكون اعداد افرادهم من 4 و كان عددهم 134 عميل , ثم للذين كان عدد افرادهم 3 و كان عددهم 133 , ثم الذين كان عدد افرادهم 1 و 2 و كانت اعدادهم 107 و 106 على التوالي .

الان سنعرض توزيع قيم الذين لم يشتروا القرض بين مجموعات قيم Education :

```
series_famm = exp_df[exp_df['Personal Loan'] == 0]['Family'].value_counts()
series_famm
```

و كانت المخرجات على الشكل التالي :

```
1    1365
2    1190
4    1088
3     877
Name: Family, dtype: int64
```

ان النسبة الأكبر التي لم تشتري القرض كان عدد افراد عائلاتهم 1 ثم 2 ثم 4 ثم 3 .

الان سنقوم بعرض النتائج بيانياً لتسهيل القراءة :

و كانت المعطيات على الشكل التالي :

قد نقول أن نسبة الأشخاص الذين لديهم قرض شخصي بينهم من حجم افراد 2 و 3 هي أعلى نسبة.

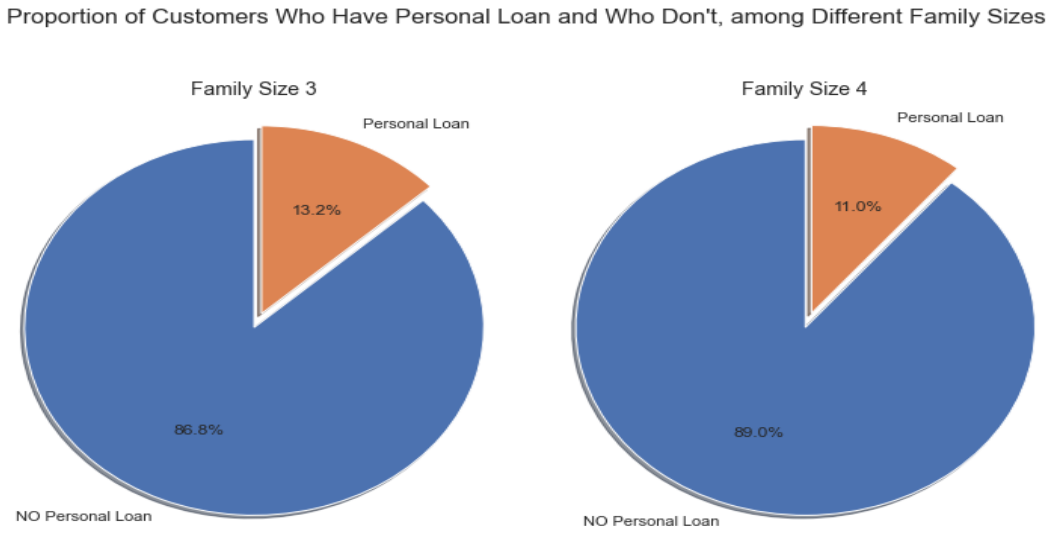
دعونا نرى العدد الدقيق لتلك النسب من المقترضين بين المودعين .

```
series_fam_3 = exp_df[exp_df['Family'] == 3]['Personal Loan'].value_counts()
```

```
series_fam_4 = exp_df[exp_df['Family'] == 4]['Personal Loan'].value_counts()
```

```
labels = ['NO Personal Loan','Personal Loan']

fig, (ax1, ax2) = plt.subplots(1,2, figsize = (12,6),subplot_kw=dict(aspect="equal"))
fig.suptitle('Proportion of Customers Who Have Personal Loan and Who Don\'t, \
among Different Family Sizes', fontsize = 16, y = 1.1, x = 0.51);

ax1.pie(series_fam_3, labels = labels, autopct= '%1.1f%%', shadow = True,explode = (0, 0.1), radius = 1.25, startangle = 90)
ax1.set_title('Family Size 3',fontsize = 14, y = 1.1)

ax2.pie(series_fam_4, labels = labels, autopct= '%1.1f%%', shadow = True,explode = (0, 0.1), radius = 1.25, startangle = 90)
ax2.set_title('Family Size 4', fontsize = 14, y = 1.1);

plt.savefig('Proportion_of_PL_among_family_levels.png', bbox_inches = 'tight');
```

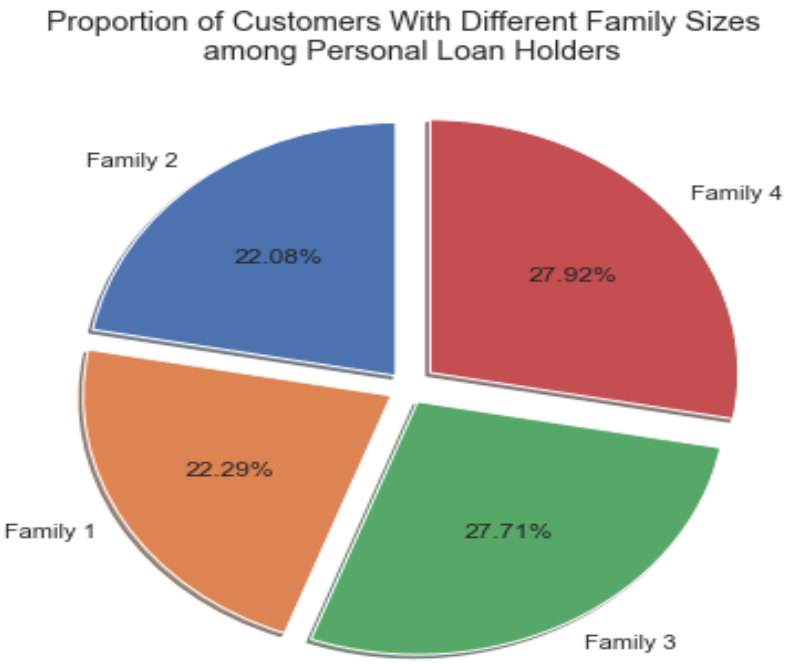
و كانت المخرجات على الشكل التالي :



الان سنعرض نسب الذين قاموا بشراء القرض فقط و عرضهم ضمن متغير واحد :

```
plt.axis('equal')
plt.title('Proportion of Customers With Different Family Sizes \n among Personal Loan Holders', \
          fontsize = 14, y = 1.3)
labels = ['Family 2',' Family 1','Family 3','Family 4']
plt.pie(series_fam.sort_values(ascending = True), labels = labels, \
        autopct= '%1.2f%%', shadow = True, explode = (0.1, 0.1, 0.1,0.15), radius = 1.6, startangle = 90);
plt.savefig('Proportion_family_size_among_PL.png', bbox_inches = 'tight');
```

و كانت المخرجات على الشكل التالي :



- 27.9٪ و 27.7٪ من الأشخاص الذين لديهم قرض شخصي لديهم حجم عائلي 4 و 3 على التوالي.
- بالنسبة للمتغير Education – الشرائح الرئيسية لبيع القرض الشخصي هي الأشخاص الذين لديهم حجم عائلي 3 و 4.
- القيم المستهدفة لمتغير Education هي 3 و 4 ، حيث أن نسبة الأشخاص الذين لديهم قرض شخصي هي الأعلى مع حجم الأسرة 3 – 13.2٪.

CCAvg :

فيما يلي توزيع قيم Personal Loan بين مجموعات قيم CCAvg :

```
series_cca = exp_df[exp_df['Personal Loan'] == 1]['CCAvg'].value_counts()
series_cca.describe()
```

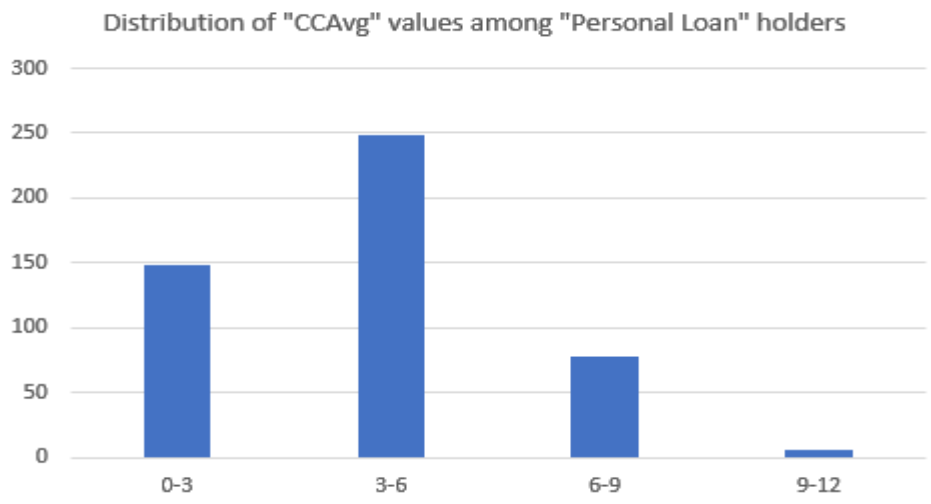
و كانت المخرجات على الشكل التالي :

```
count    95.000000
mean      5.052632
std       3.349717
min       1.000000
25%       2.000000
50%       5.000000
75%       7.000000
max      19.000000
Name: CCAvg, dtype: float64
```

سنقوم بعرض النتائج بيانياً لتسهيل القراءة :

```
width = 1.5
series_cca.plot.hist(bins = np.arange(series_cca.min(), series_cca.max() + width, width ), figsize = (8,6))
plt.xlabel('CCAvg')
plt.title('Distribution of "CCAvg" values among "Personal Loan" holders', fontsize = 14, y = 1.05);
plt.savefig('Distrib_ccavg_among_PL.png', bbox_inches = 'tight')
```

و كانت المعطيات على الشكل التالي :



يمكننا ان نقول أنه يمكن تقسيم قيم خصائص CCAvg إلى ثلاث مجموعات ، ضع في اعتبارك مدى تكرارها بين صاحب القرض الشخصي:

- المجموعة الأولى : $0 < CCAvg < 3$.
- المجموعة الثانية : $3 < CCAvg < 6$.
- المجموعة الثالثة : $6 < CCAvg < 9$.

إن اول مجموعتين من CCAvg هي مجموعات مستهدفة و ذات اولوية

- المجموعات المستهدفة من المتغير CCAvg حسب الأولوية:
- المجموعة الأولى : $6 > CCAvg > 3$.
- المجموعة الثانية : $3 > CCAvg > 0$.

: Income

فيما يلي توزيع قيم Personal Loan بين مجموعات قيم Income :

```
series_inc = exp_df[exp_df['Personal Loan'] == 1]['Income'].value_counts()
series_inc.describe()
```

و كانت المعطيات على الشكل التالي :

count	102.000000
mean	4.705882
std	2.723525
min	1.000000
25%	2.000000
50%	4.000000
75%	7.000000
max	11.000000

Name: Income, dtype: float64

سنقوم بعرض النتائج بيانياً لتسهيل القراءة :

```
width = 1.5
series_inc.plot.hist(bins = np.arange(series_inc.min(), series_inc.max() + width, width ), figsize = (8,6))
plt.xlabel('Income')
plt.title('Distribution of "Income" values among "Personal Loan" holders', fontsize = 14, y = 1.05);
plt.savefig('Distrib_income_among_PL.png', bbox_inches = 'tight')
```

و كانت المخرجات على الشكل التالي :

يمكننا ان نقول أنه يمكن تقسيم القيم المميزة لـ Education إلى ثلاث مجموعات مع مراعاة تواترها بين صاحب القرض الشخصي:

- المجموعة الأولى :100 > Income > 51 .
- المجموعة الثانية : 150 > Income > 101 .
- المجموعة الثالثة : 200 > Income > 151 .

أن المجموعة الثانية و الثالثة من Education هي مجموعات ذات أولوية لبيع القرض الشخصي .

المجموعات المستهدفة من المتغير Education مرتبة ترتيبًا تنازليًا حسب الأولوية:

- المجموعة الأولى : 150 > Income > 101 .
- المجموعة الثانية : 200 > Income > 151 .

الان سنقوم بالبحث عن جواب السؤال الثالث من البحث :

ما هي عينة البيانات مع العملاء من الشرائح الرئيسية ؟

كما وجدنا سابقاً ، هناك بعض الأجزاء ذات الخصائص الرئيسية التي تتمتع بقوة ارتباط أكبر بالقرض الشخصي.

فيما يلي المجموعة الفرعية ذات الارتباط الأعلى بالقرض الشخصي ، أي تعني أعلى احتمالية لبيع المنتج :

لنلقِ نظرة على نماذج قاعدة البيانات التي تحتوي على تلك الأجزاء ...

CD Account = 0,1

Education = 2,3

Family = 3,4

CCAvg = 0 to 6

Income = 101 to 200

```
Perfect = df[(df['Personal Loan'] == 0) &\
   (df['CD Account'] >= 0) &\
   (df['Education'] >= 2) &\
   (df['Family'] >= 3)&\
   (df['CCAvg'] >= 0)&\
   (df['CCAvg'] <= 6)&\
   (df['Income'] >=100)&\
   (df['Income'] <=200)                    ]
```

```
Perfect.shape[0]
```

و كانت المعطيات على الشكل التالي :

58

أي لدينا 58 عميل لديهم النسبة الأعلى لشراء القروض .

التصنيف (Classification) :

سيتم تطبيق 4 خوارزميات للتصنيف على البيانات الموجودة و المقارنة بين المصنفات المخرجة للوصول الى افضل مصنف ليتم اعتماده كمصنف للحملات التسويقية في البنك و سيتم تطبيق خوارزميات :

- شجرة القرار Decision Tree .
- نايف بايز Naïve Bayes .

- الغابات العشوائية Random Forest .
- الجار الأقرب (K – Nearest Neighbor) KNN .

شجرة القرار Decision Tree :

اولاً سنقوم بفصل العينة الى عينتين , عينة تشمل المتغيرات الخمس التي تؤثر على عملية شراء القرض و التي قمنا بمعرفتها سابقاً , و عينة سنقوم بوضع المتغير الرئيسي داخلها :

```
X = my_data[['Family', 'Education', 'CCAvg', 'CDAccount', 'Income']].values
```

```
y = my_data["PersonalLoan"]
```

و الان سنقسم العينتين الى أربعة اقسام :

قسم من x و قسم من y لتقوم الخوارزمية بالتدريب عليها و ستكون نسبتها 70% من اجمالي العينة .

و قسم من x و قسم من y لتقوم الخوارزمية بإجراء اختبار عليها و معرفة نسبة الدقة و تشكل ال30% الباقية من العينة الاجمالية .

```
X_trainset, X_testset, y_trainset, y_testset = train_test_split(X, y, test_size=0.3, random_state=3)
```

و الان سنقوم بتطبيق الخوارزمية :

```
drugTree = DecisionTreeClassifier(criterion="entropy", max_depth = 4)
drugTree # it shows the default parameters
```

```
from sklearn import metrics
import matplotlib.pyplot as plt
print("DecisionTrees's Accuracy: ", metrics.accuracy_score(y_testset, predTree))
```

و كانت المخرجات على الشكل التالي :

```
DecisionTrees's Accuracy:  0.978
```

أي ان نسبة ال Accuracy لدى المصنف Decision Tree هي 97.8% .

نايف بايز Naïve bayes :

سنقوم بتطبيق الخوارزمية :

```
naive_model = GaussianNB()
naive_model.fit(X_trainset, y_trainset)

prediction = naive_model.predict(X_testset)
print("Naive Bayes's Accuracy: " , naive_model.score(X_testset, y_testset))
```

و كانت المخرجات على الشكل التالي :

```
Naive Bayes's Accuracy:  0.8953333333333333
```

أي ان نسبة ال Accuracy لدى المصنف Naïve bayes هي 89.5% .

الغابات العشوائية Random Forest Classifier :

سنقوم بتطبيق الخوارزمية :

```
randomforest_model = RandomForestClassifier(max_depth=2, random_state=0)
randomforest_model.fit(train_set, train_labels)
```

```
predicted_random=randomforest_model.predict(test_set)
print("Random Forest's Accuracy: " ,randomforest_model.score(test_set,test_labels))
```

و كانت المخرجات على الشكل التالي :

```
Random Forest's Accuracy:  0.904
```

أي ان نسبة ال Accuracy لدى المصنف Random Forest هي 90.4% .

الجار الأقرب KNN (K – Nearest Neighbour) :

```
knn = KNeighborsClassifier(n_neighbors= 21 , weights = 'uniform', metric='euclidean')
knn.fit(X_Train, Y_Train)
predicted = knn.predict(X_Test)
from sklearn.metrics import accuracy_score
print("KNN's Accuracy: ",accuracy_score(Y_Test, predicted))
```

و كانت المخرجات على الشكل التالي :

```
KNN's Accuracy:  0.9159439626417611
```

أي ان نسبة ال Accuracy لدى المصنف KNN K–Nearest Neighbor هي 91.5% .

و من خلال المقارنة بين نتائج المصنفات الأربعة المستخدمة اتضح ان مصنف شجرة القرار (Decision Tree) هو افضل من المصنفات الأخرى , و بالتالي سيتم الاعتماد عليه لاختيار العملاء الأكثر قابلية لشراء القرض .

النتائج و التوصيات :

النتائج :

لقد أجرينا تحليلًا بسيطًا خطوة بخطوة لخصائص العميل لتحديد أنماط الاختيار الفعال للمجموعة الفرعية من العملاء الذين لديهم احتمالية أكبر لشراء منتج جديد "قرض شخصي" من البنك. نفذنا الخطوات التالية:

- تحققنا من جميع الخصائص الاثنتي عشرة سواء كان لكل منها ارتباط بحقيقة بيع المنتج أم لا.
- وجدنا خمس خصائص رئيسية لها قوة ارتباط أعلى من المتوسط بالمنتج.
- قمنا بتحليل الخصائص الرئيسية والحصول على شرائح في كل منها ذات قوة ارتباط مختلفة بالمنتج.
- لقد حاولنا تكوين مجموعة فرعية من العملاء بخصائص مثالية لديهم أعلى احتمالية لشراء المنتج. لسوء الحظ ، كان العدد قليل جداً , لذا انتقلنا للمرحلة التالية و هي بناء المصنفات .
- قمنا ببناء أربعة خوارزميات تصنيف : 1- مصنف شجرة القرار . 2- مصنف الغابات العشوائية . 3- مصنف الجار الأقرب . 4- مصنف بايز الساذج .
- هدف البنك هو تحويل عملاء المسؤولية إلى عملاء قروض. يريدون إنشاء حملة تسويقية جديدة ومن ثم ، فهم بحاجة إلى معلومات حول العلاقة بين المتغيرات الواردة في البيانات. تم استخدام أربع خوارزميات تصنيف في هذه الدراسة .
- أن خوارزمية شجرة القرار تتمتع بأعلى دقة ويمكننا اختيار ذلك كنموذج نهائي لدينا .

التوصيات :

- اقتصرت الدراسة على بيانات قليلة نسبياً خاصة بالعميل فيجب على إدارة الحملات التسويقية للبنوك الوصول لمعلومات عن العميل تفيد بشكل مناسب .
- التنبؤ المسبق لحاجات العميل و رغباته من اجل اشباعها يكمن باستخدام تقنيات التنقيب في البيانات .
- استخدام تقنيات التنقيب في البيانات على جميع الأصعدة تزيد من إيرادات البنوك و تسهل عملياتها الداخلية و الخارجية و الاحتكاك مع العميل بكل وقت .
- تطبيق تقنيات التنقيب في البيانات على جميع اقسام البنك و عدم اقتصار قسم معين (تسويق) لما له من ميزات على المدى الطويل و الارتباط القوي داخل البنك و تعامله مع العميل .
- عندما يتم إقرار البيانات التي يجب الحصول عليها من العملاء يجب ان تكون مدروسة بشكل صحيح لتعطي معلومات مفيدة و التخلص من البيانات التي لا قيمة لها لكي يتم الاستخدام الأمثل لها من خلال تقنيات التنقيب في البيانات .
- عدد العملاء كان قليل نسبياً لذلك ينصح بإعادة الدراسة كل فترة مع زيادة اعداد العملاء , لأن التنقيب في البيانات يحتاج لعدد كبير من البيانات .
- اعتماد نموذج شجرة القرار للتنبؤ بالعملاء المحتملين مستقبلاً .

المراجع :

- محمد ناظم السبيعي , 2020 , استخدام تقنيات التنقيب في المعطيات للتنبؤ بنتيجة الحملات التسويقية المباشرة .
- البدوي سعد البدوي المبارك , 2017 , استخدام تقنيات تنقيب البيانات لاستكشاف أنماط مؤثرات التحصيل الاكاديمي لطلاب المرحلة الثانوية .
- عماد حجار , 2016 , التنقيب في بيانات المؤسسات التعليمية لتحليل مستوى الطلاب .
- نهى حمود , تصميم عروض حسب الشرائح الزمنية باستخدام التنقيب في المعطيات, 2021
- han, j., kamber, n., & pei, j. (2000). *Data Mining concepts and techniques.*
- Palace, B. (1996). *Data Mining, Technology note prepared for Management.* UCLA.
- https://www.javatpoint.com/machine-learning
- https://www.sciencedirect.com/topics/computer-science/decision-tree-classifier
- https://www.baianat.com/ar/articles/mine-gold-in-your-data
- https://techzonee.com/
- https://economistsarab.com/