الجمهورية العربية السورية

جامعة دمشق

كلية الهندسة المعلوماتية

قسم الذكاء الصنعي ومعالجة اللغات الطبيعية

نهج قائم على تعلم الآلة لتوقع انسحاب عملاء البنوك التجارية

مشروع مادة التعلم التلقائي

إعداد الطلاب

زاكروز نجم الدين اسماعيل حسن علي الزعبي

إشراف المهندسين:

زينة الدلال

علا الطبال

ملخص

في هذه الدراسة، نهدف إلى دراسة ظاهرة انسحاب عملاء البنوك التجارية وتطوير نماذج تعلم الآلة التي تتنبأ بالعملاء الذين من المرجح أن ينسحبوا. استخدمنا مجموعة بيانات شاملة تحتوي على سمات متنوعة للعملاء، مثل التفاصيل الديموغرافية، معلومات الحساب، وسجل المعاملات. تهدف الدراسة إلى فهم العوامل الأكثر تأثيرًا على انسحاب العملاء ومقارنة أداء عدة نماذج تعلم آلي.

تمت معالجة البيانات عن طريق التعامل مع القيم المفقودة، وترميز المتغيرات الفئوية باستخدام تقنيات مثل One Hot Encoding، وتقييس الميزات العددية. تم تدريب نماذج مثل الانحدار اللوجستي، الأشجار القرارية، الغابات العشوائية، LightGBM، SVM، XGBoost، و CatBoost وتقييمها باستخدام معايير مثل الدقة، F1-score، و ROC-AUC.

تشير نتائج البحث إلى أن نماذج Random Forest و XGBoost و CatBoost حققت أعلى دقة في التنبؤ بالانسحاب، حيث وصلت نسبة الدقة إلى 89%، متفوقة على جميع الأعمال المشابهة. كشف تحليل أهمية الميزات أن الشكاوى ورصيد الحساب كانت من بين أهم المؤشرات على انسحاب العملاء، وأن العملاء الأكبر سنًا كانوا أكثر عرضة للانسحاب. كما تسلط الدراسة الضوء على تأثير اختلال توازن البيانات وفعالية تقنيات مثل زيادة العينة لتحسين أداء النموذج.

في الختام، تُظهر نتائجنا أن نماذج تعلم الآلة يمكنها التنبؤ بانسحاب العملاء في القطاع المصرفي بفعالية، مما يمكّن البنوك من اتخاذ تدابير استباقية للاحتفاظ بالعملاء ذوي القيمة العالية. يمكن للأعمال المستقبلية أن تستكشف دمج بيانات سلوكية إضافية وقدرات التنبؤ الفوري لتحسين دقة التنبؤ بالانسحاب.

مقدمة

في صناعة البنوك التنافسية بشدة، أصبح الاحتفاظ بالعملاء عاملاً حاسماً للنجاح المستدام والربحية. ظاهرة فقدان العملاء، التي تشير إلى انتقال العملاء من بنك إلى منافس، لها تأثيرات مالية كبيرة، مما يجعل التنبؤ والوقاية منها مجالاً حاسماً للتركيز للبنوك. يتم معالجة تنبؤ الانسحاب على نطاق واسع في قطاعات مختلفة بمجموعة متنوعة من النهج التحليلية والمدفوعة بالبيانات، مما يعكس تعقيدها والعوامل المتنوعة التي تؤثر عليها، مثل جودة الخدمة، والتسعير، ورضا العملاء [1].

توفر التطورات الحديثة في تكنولوجيا البيانات الكبيرة والتعلم الآلي فرصًا جديدة لمعالجة الانسحاب بشكل أكثر فعالية. تتيح هذه التكنولوجيات تحليل كميات كبيرة من بيانات العملاء لتحديد العملاء المعرضين للخطر وفهم العوامل المؤدية للانسحاب. يستفيد هذا البحث من البيانات التاريخية للمعاملات ويطبق نماذج التعلم الآلي للتنبؤ بانسحاب عملاء البنوك التجارية، بهدف تقديم رؤى عملية يمكن أن تعزز استراتيجيات الاحتفاظ بالعملاء [1][3].

من خلال دمج هذه المنهجيات، يساهم هذا البحث في الإدارة الاستراتيجية لعلاقات العملاء والاحتفاظ بهم، ويقدم مقارنة مفصلة لأداء النماذج ويحدد أهم مؤشرات الانسحاب. الهدف النهائي هو تمكين البنوك من تخصيص تفاعلاتها مع العملاء ونهج الاحتفاظ بهم لتقليل الانسحاب وتحسين الولاء العام للعملاء.

الدراسة المرجعية

عند مناقشة المشاريع المماثلة في مجال توقع تحول عملاء البنوك، يمكننا الإشارة إلى الدراسة المحددة التي أجريت من قبل جامعة درسكل سينا شرندابي [2]. في هذه الدراسة، تم مقارنة أداء تقنيات التصنيف المختلفة لاقتراح نموذج فعال للتنبؤ بتحول عملاء البنوك، باستخدام 10 سمات ديموغرافية وشخصية من 10000 عميل للبنوك الأوروبية. من بين النماذج الستة المختلفة التي تم تطويرها، تبين أن الغابات العشوائية والشبكات العصبية (ANN) هما الأفضل من حيث الأداء

العام. ومع ذلك، تبين أن الغابات العشوائية عرضة بشكل ملحوظ لحالات الإفراط Overfiting في التكيف وهو مشكلة لم يتم حلها بشكل مرضٍ بعد اختيار السمات، ولكن تم تقليصها بعد القيام بموازنة الأصناف. من ناحية أخرى، لم تظهر الشبكات العصبية أي مشكلة خطيرة فيما يتعلق بحالات الإفراط في التكيف وقد تحسن الأداء بشكل كبير على البيانات المتوازنة. كما تبين أن الشبكات العصبية قوية في التعامل مع القيم المتطرفة في حين أن تداخل مثل هذه الأخطاء يمكن أن يؤثر سلبًا على أداء الغابات العشوائية بشكل أكبر.

إحدى ميزات الغابات العشوائية هي أنها على عكس الشبكات العصبية تسمح بترتيب المتغيرات وفقًا لمساهمتها في إجراء التصنيف. وبناءً على مجال التطبيق، قد تهتم البنوك بهذه المعلومة، مما يجعل الغابات العشوائية مرشحًا أفضل للتنبؤ. تشير الدراسات الأخرى إلى أن تحسين معدل الاحتفاظ بالعملاء بنسبة تصل إلى 5٪ يمكن أن يزيد أرباح البنك بما يصل إلى 85٪. هذه النتيجة تستند إلى دراسات عدة تؤكد على أهمية الاحتفاظ بالعملاء مقارنة بتكلفة جذب عملاء جدد. وفقًا لأبحاث متعددة[1][3]، تلعب البيانات الديموغرافية والشخصية دورًا حاسمًا في التنبؤ بانسحاب العملاء.

مجموعة البيانات

يتضمن مجموعة بيانات ["4]"Customer Churn تصنيف الأشخاص حسب الانسحاب أو عدم الانسحاب، حيث يتم توقع العملاء الذين قد ينسحبون من البنك. يتألف قاعدة البيانات من 10000 قيمة و 18عمودًا. تحتوي الميزات على عدة عوامل مثل رقم الصف، ورقم العميل، واللقب، والنقطة الائتمانية، والموقع الجغرافي، والجنس، والعمر، وفترة البقاء، والرصيد، وعدد المنتجات، وحملة البطاقة الائتمانية، والعضو النشط، والراتب المقدر، والشكوى، ودرجة الرضا. كما يتضمن العمود الأخير معلومات حول ما إذا كان العميل قد غادر البنك أم لا، ونوع البطاقة التي يحملها العميل، والنقاط التي يحصل عليها العميل لاستخدام بطاقة الائتمان.

تحليل واستكشاف البيانات

تتضمن مجموعة البيانات السمات التالية: رقم السطر، ورقم العميل، واللقب، والنقطة الائتمانية، وبلد الإقامة، والجنس، والعمر، وعدد سنوات الاشتراك، والرصيد، وعدد المنتجات، وحالة بطاقة الائتمان، ونشاط العميل، والراتب المقدر، وما إذا كان لديه شكوى أم لا، ودرجة الرضا التي يقدمها العميل لحل شكواه، ونوع البطاقة التي يحملها العميل، والنقاط التي يكسبها العميل عند استخدام بطاقة، والهدف هو معرفة ما إذا كان العميل قد ترك البنك أو لا.

	CreditScore	Age	Tenure	Balance	NumOfProducts	EstimatedSalary	Satisfaction Score	Point Earned
count	10,000.00	10,000.00	10,000.00	10,000.00	10,000.00	10,000.00	10,000.00	10,000.00
mean	650.53	38.92	5.01	76,514.62	1.53	100,136.47	3.01	606.52
std	96.65	10.46	2.89	62,108.99	0.58	57,353.54	1.41	225.92
min	350.00	18.00	0.00	0.00	1.00	11.58	1.00	119.00
25%	584.00	32.00	3.00	0.00	1.00	51,385.55	2.00	410.00
50%	652.00	37.00	5.00	96,653.79	1.00	100,136.47	3.00	605.00
75%	718.00	44.00	7.00	127,414.88	2.00	149,174.23	4.00	801.00
max	850.00	92.00	10.00	250,898.09	4.00	199,992.48	5.00	1,000.00

الشكل 1 وصف السمات الرقمية

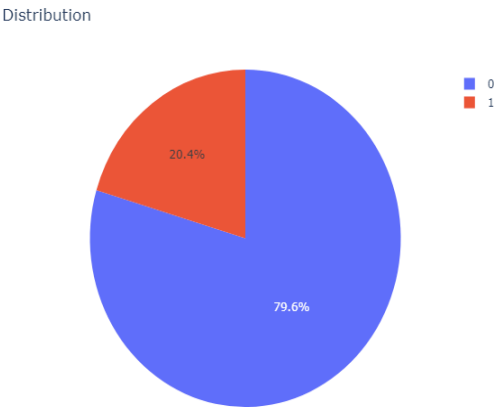
	Geography	Gender	HasCrCard	IsActiveMember	Complain	Card Type
count	10000	10000	10000.0	10000	10000	10000
unique	3	2	2.0	2	2	4
top	France	Male	1.0	1	0	DIAMOND
freq	5014	5457	7075.0	5151	7956	2507

الشكل 1 وصف السمات المحددة

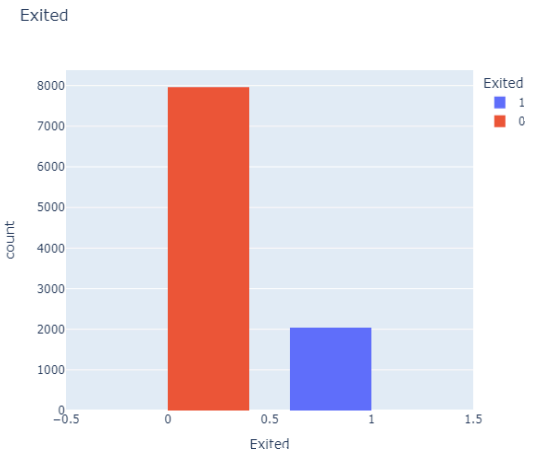
من خلال وصف السمات السابق، يتبين وضوحا أن البيانات نظيفة، فالسمات الرقمية تتبع توزيعا طبيعيا دون تجمع او تفرق ضمن أحد أرباع التوزع، والداتا المحدد متوزعة بشكل شبه متساوي بين قيم السمة دون طغيان لسمة على أخرى.

مخطط نسبة توزع العملاء الذين انسحبوا، من الملاحظ من الرسوم البيانية التالية أن نسبة

العملاء الموالين أكبر بما يقارب أربعة أضعاف العملاء المنسحبين، وعلى اعتبار أن هذه السمة هي الهدف المراد معرفته تكون مجموعة البيانات غير متوازنة

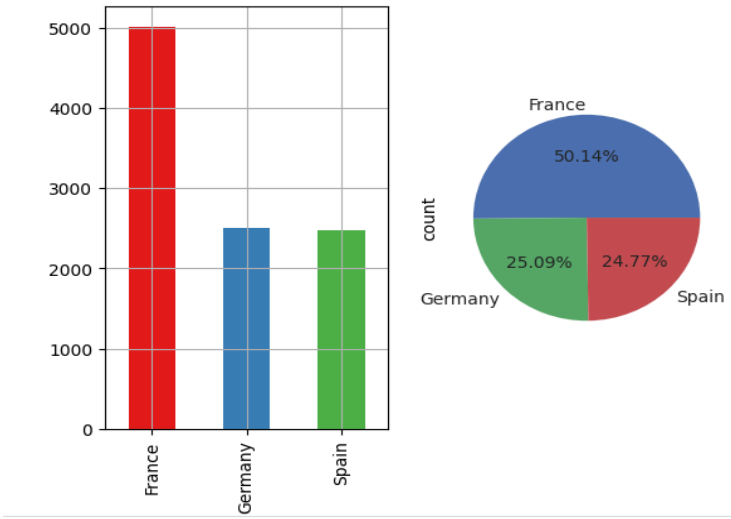


الشكل 4 نسبة العملاء الموالين من المنسحبين



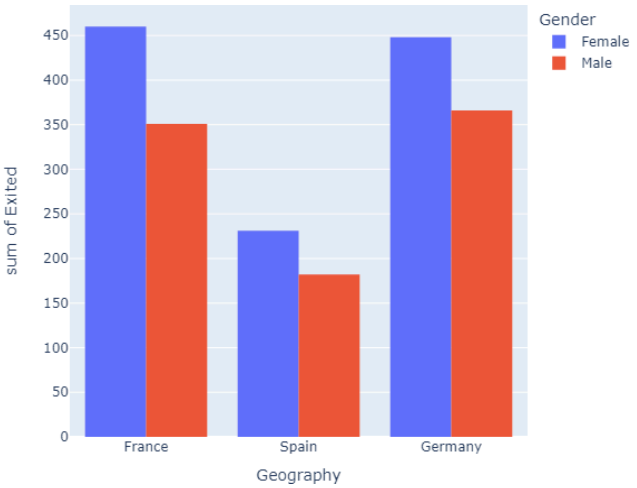
الشكل 3 عدد العملاء الموالين والمنسحبين

معرفة توزع العملاء على البلدان، من الملاحظ أن العملاء الفرنسيين هم نصف العملاء تقريبا.

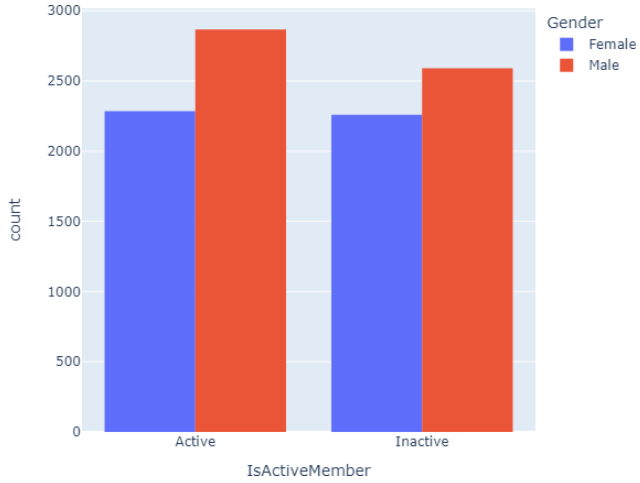


الشكل 5 عدد العملاء المنسحبين من كل بلدان

توزع العملاء الذين انسحبوا من البنك حسب الدولة والجنس، من الملاحظ أن العملاء الذكور هم أكثر انسحابا من الإناث، وعلى النقيض نجد أن الإناث هم أكثر فعالية من الذكور في البنك

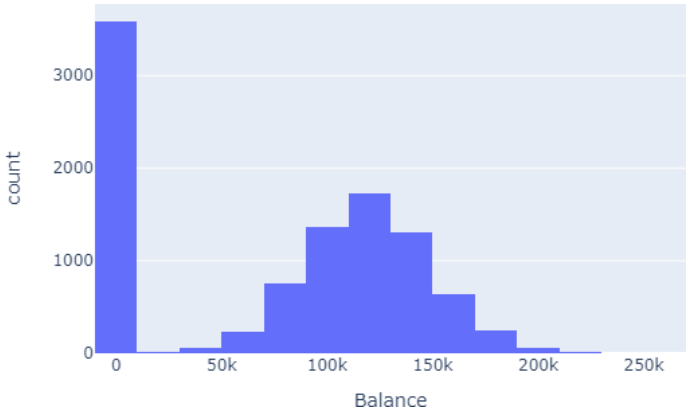


الشكل 6 توزع العملاء الذين انسحبوا من البنك حسب الدولة والجنس



الشكل 7 توزع العملاء الذين غادروا وبقيوا من البنك حسب الجنس

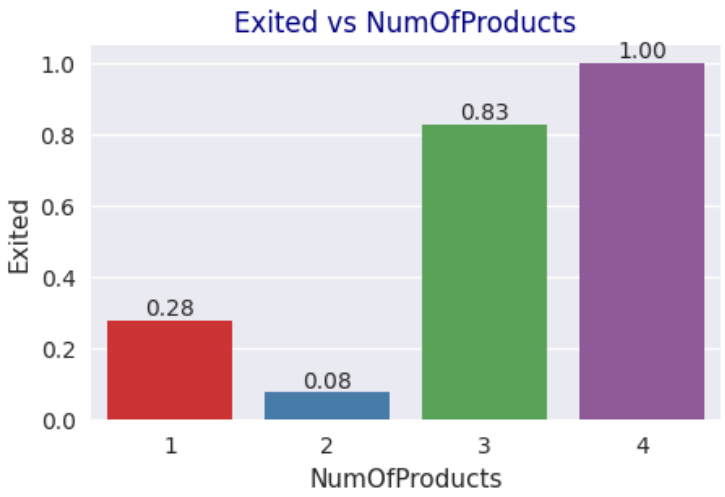
مخطط أرصدة العملاء، حيث يتبين أن عدد كبير من العملاء ليس لديهم أرصدة في البنك



نسبة العملاء الذين خرجوا (تم إرجاعهم) من البنك بناءً على عدد المنتجات التي كانوا يستخدمونها.

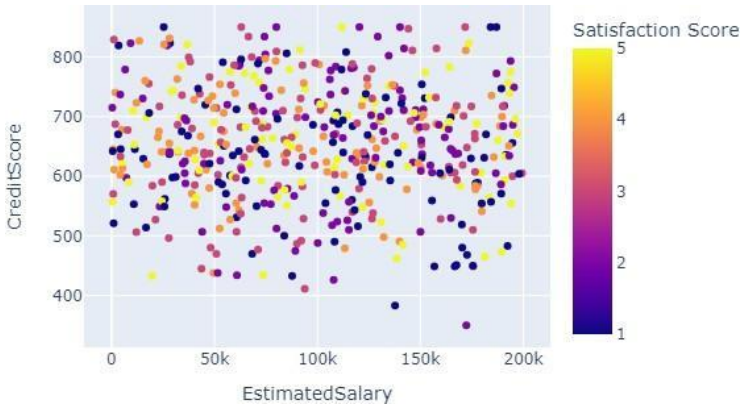
- منتج واحد: تبلغ نسبة الاحتفاظ به حوالي 28%، مما يشير إلى أن العديد من العملاء الذين لديهم منتج واحد فقط لا يميلون إلى مغادرة البنك.

- منتجان: معدل احتفاظ منخفض للغاية يبلغ 0.08%، مما يشير إلى أن العملاء الذين لديهم منتجين هم أقل عرضة للمغادرة بشكل كبير.
- ثلاثة منتجات: معدل احتفاظ اعلى م، حيث اختار 0.83% البقاء، مما قد يشير إلى أن زيادة عدد منتجات تنشأ مغادرة.
- أربعة منتجات: زيادة نسبة الاحتفاظ بالمنتجات إلى 1 %؛ لقد قرر جميع العملاء الذين لديهم أربعة منتجات بالمغادرة، مما يشير إلى مغادرة



الشكل 10 العلاقة بين مستوى رحيل العميل والعدد المنتجات التي لديه

من خلال الرسم التالي نلاحظ أنه لا يوجد أي علاقة بين مستوى رضا العميل وراتبه الشهري أو الرصيد الائتماني فهاك العملاء البسطاء ماديا لكن غير راضين عن الخدمات، وهناك العملاء الأغنياء الراضين عن الخدمات والعكس بالعكس.

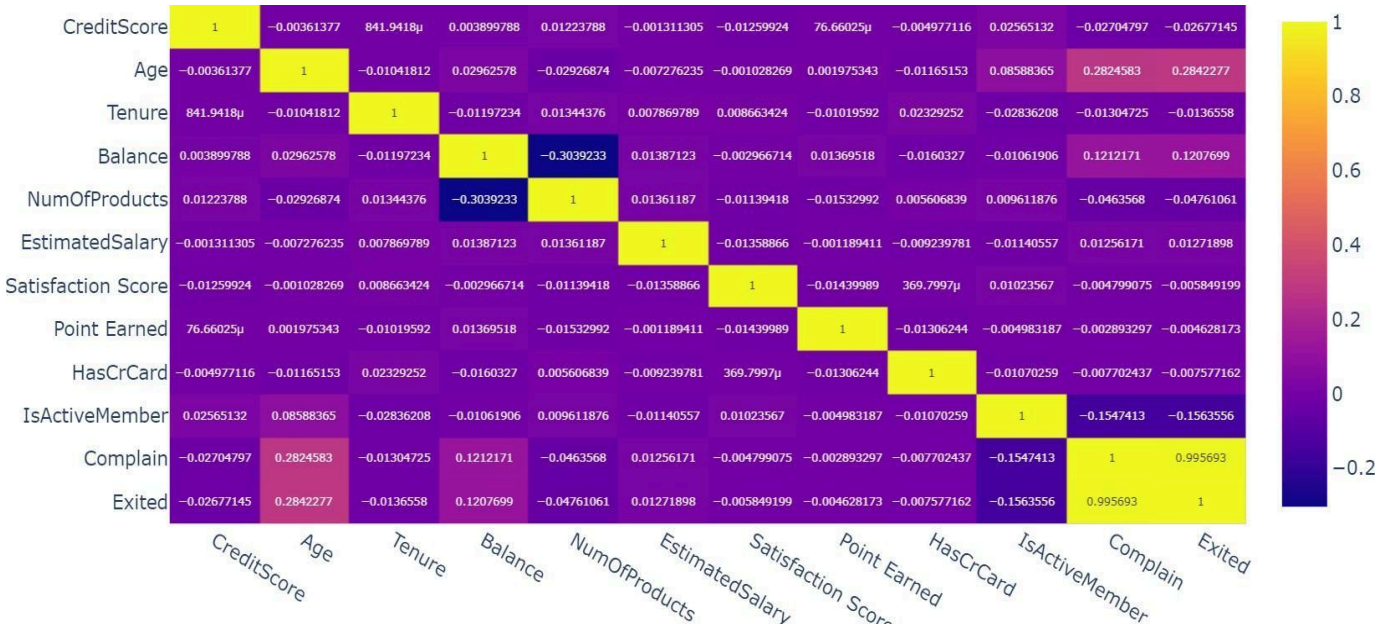


الشكل 10 العلاقة بين مستوى رضا العميل و الرصيد الائتماني

اكتشاف الترابط بين السمات

تشير البيانات إلى أن أعمدة العمر، الرصيد ، والجنس مرتبطة بشكل إيجابي بمعدل انسحاب العملاء وهناك علاقة سلبية بين كون العميل عضواً فعالاَ Is Active Member بمعدل انسحاب العملاء. وبمقارنة أعمدة شكوى والانسحاب (Exited) ، فسنلاحظ وجود ارتباط إيجابي قوي للغاية، حيث تم إنشاء الأخيرة بناًء على الأولى

في جميع الأوراق بحثية [1][2][3] مشابهة حيث تم حذف العمود Complain لأن لديه ارتباط ايجابي قوي للغاية, وهذا ليس بالأمر جيد للمودل لان سيقوم من خلاله فقط بالتنبؤ بال Target ويهمل الفيتشرز الأخرى, لذلك تم حذفه



الشكل 11 الترابط بين السمات

هندسة السمات

من خلال الاطلاع على السمات، تبين أنه يمكن استخلاص السمات التالية:

- **معدل الراتب إلى الرصيد (Balanced Salary rate) :** هو مقياس يستخدم لتقييم استقرار المال للعميل. يتم حساب هذا المقياس عن طريق مقارنة رصيد حساب العميل وراتبه المقدر، وحساب النسبة بينهما، حيث يمكن أن يساعد في تحديد العملاء الذين قد يواجهون صعوبات مالية أو يكونون عرضة لخطر مغادرة البنك.
- **معدل استخدام المنتج حسب السنة (Product usage rate by year) :** هو مقياس يقيس استخدام العميل لمنتجات البنك عبر الزمن. يتم حساب هذا المقياس عن طريق مقارنة عدد المنتجات التي يستخدمها العميل مع عدد السنوات التي قام فيها العميل بالتعامل مع البنك.
- **معدل الولاء حسب العمر (Tenure rate by age) :** وهو مقياس يقيس ولاء العميل للبنك على مدار حياته. يتم حساب هذا المقياس عن طريق مقارنة عدد السنوات التي قضاها العميل كعميل للبنك مقابل عمره بعد استبعاد سنوات المراهقة (بطرح 17عامًا من عمره).
- **معدل درجة الائتمان حسب الراتب (Credit score rate by age) :** هو مقياس يقيس جدارة الائتمان للعميل بناءً على دخله. يتم حساب هذا المقياس عن طريق مقارنة درجة الائتمان للعميل مع الراتب المقدر.

المنهجية المتبعة

بداية قمنا بتحويل المتغيرات النصية إلى متغيرات رقمية من خالل تحولها الى One Hot Encoding ومن ثم تقسيم البيانات إلى مجموعتي تدريب واختبار بمعدل 75/25 مع توزيع متوازن لأصناف في كل مجموعة من خالل إضافة stratify لعملية التقسيم بناء على عمود الهدف.

بما أن المسألة هي مسألة تصنيف ثنائي تم استخدام العديد من النماذج المخصصة في عمليات التصنيف وهي :

- Logistic Regression
- Decision Tree
- RandomForest
- KNeighbors
- XGBClassifier
- SVC
- GaussianNB
- CatBoostClassifier
- GradientBoostingClassifier
- LGBMClassifier

ايضا تم تحويل مسالة ل Unsupervised وتجريب عدد من موديلات الخاصة بال Clustering

- K-mean
- DBSCAN
- Hierarchical

قمنا ببناء خطوط تنفيذ (Pipeline) يحتوي كل واحد منها على نموذج لحصر القيم ضمن مجال معين باستخدام Standard Scaler ومن ثم النموذج بقيمة الافتراضية، وقمنا بتدريب هذه الخطوط على بيانات التدريب واختبارها على بيانات الاختبار، وبما أن البيانات غير متوازنة تم استخدام كل من العاملين "Accuracy" و "F1 Score" كمقاييس لدقة النموذج

1- النتائج الأولية للنماذج Classification السابقة موضحة بالجدول التالي:

F1 score	Accuracy	اسم النموذج
0.08	0.78	**Logistic Regression**
0.46	0.77	**Decision Tree**
0.52	0.84	**Random Forest**
0.45	0.82	**KNN**
0.53	0.84	**XGBoost**
0.31	082	**SVM**
0.08	0.79	**Gaussian NB**
0.54	0.85	**CatBoostClassifier**
0.56	0.85	**GradientBoostingClassifier**
0.51	0.84	**LGBM classifier**

جدول 1 النتائج البدائية للنماذج

وضوحا نجد أن قيم المعيار F 1 جدا سيئة وذلك بسبب أن البيانات غير متوازنة في الأصناف ومن أجل التأكد من ذلك قمنا بتجربة جميع مجالات المعاملات من أجل كل نموذج باستخدام Grid Search ولم يكن هناك تحسن ملحوظ بالنتائج فلذلك قمنا بموازنة عدد الأصناف ضمن مجموعة البيانات.

لعمل موازنة لمجموعة البيانات هناك طريقتان:

1. تقليل حجم البيانات إلى عدد عناصر الصنف الأقل .Under sampling

2. زيادة حجم البيانات إلى عدد عناصر الصنف الأكثر .Over sampling

من المتوقع أن تكون الطريقة الأولى (Under Sampling) هي الأفضل نظرا لأنها لا تولد بيانات

اصطناعية كما في الطريقة الثانية وتعتمد فقط على بيانات حقيقية غير أن نسبة العملاء الذين غادروا البنك قليلة وبالتالي سنخسر الكثير من البيانات.

تمت تجربة الطريقة الأولى على النماذج السابقة بمعاملاتها الافتراضية والنتائج موضحة بالجدول التالي

اسم النموذج	Accuray	F1 score
Logistic Regression	0.7	0.7
Decision Tree	0.68	0.68
Random Forest	0.75	0.75
KNN	0.70	0.70
XGBoost	0.73	0.73
SVM	0.75	0.74
Gaussian NB	0.53	0.17
CatBoostClassifier	0.76	0.75
GradientBoostingClassifier	0.76	0.75
LGBM classifier	0.75	0.75

جدول 2 النتائج مع تقليل عدد العناصر

نلاحظ أن قيم المعيار F1 تحسنت بشكل ملحوظ عن ذي قبل في حين أن الدقة الاجمالية انخفضت أيضا وذلك بسبب أن حجم البيانات انخفض بشكل كبير حيث أصبح عدد الأسطر في مجموعة البيانات 4076 سطر فقط.

ومن اجل التأكد من هذه النتيجة تمت تجربة جميع مجالات المعاملات من اجل كل نموذج باستخدام Grid search ولم يكن هناك تحسن ملحوظ في النتائج عما سبق.

من اجل ذلك قمنا بتجربة الطريقة الثانية(Over Sampling) على النماذج السابقة بمعاملاتها الافتراضية والنتائج

موضحة بالجدول التالي:

اسم النموذج	Accuracy	F1 score
Logistic Regression	0.71	0.72
Decision Tree	0.8	0.79
Random Forest	0.83	0.82
KNN	0.75	0.75
XGBoost	0.88	0.88
SVM	0.83	0.83
Gaussian NB	0.58	0.3
CatBoostClassifier	0.89	0.88
GradientBoostingClassifier	0.86	0.86
LGBM classifier	0.88	0.88

جدول 3 النتائج مع زيادة عدد العناصر

تم تحسن النتائج بشكل واضح عن ذي قبل إذ أصبح هناك تحسن ملحوظ في المعيار F1

لتحسين النتائج بشكل أفضل قمنا بالبحث عن أفضل المعاملات لكل نموذج على حدة وينتج لدينا الجدول التالي:

F1 score	Accuracy	اسم النموذج
0.72	0.72	**Logistic Regression**
0.8	0.81	**Decision Tree**
0.82	0.83	**Random Forest**
0.74	0.74	**KNN**
0.88	0.88	**XGBoost**
0.84	0.85	**SVM**
0.74	0.71	**Gaussian NB**

جدول 4 نتائج أفضل المعاملات بعد زيادة حجم البيانات

لاحظ أن كًلا من النموذجين Random Forestو KNN لم يكن هناك أي فرق في النتائج أما في بقية النماذج فكان هناك تحسن بسيط في النتائج لذلك سيتم الاعتماد فيما بعد على هذه النماذج.

من اجل الحصول على نتائج أفضل نتوقع أنه باستخدام طرق التعلم بالإجماع او بالتصويت أن تعطي نتيجة أفضل مما سبق, حيث أن كل نموذج قادر على التنبؤ بشكل صحيح بجزء من البيانات.

قمنا بتجربة كلا من الطريقتين Hard voting و Soft voting وقمنا باختيار افضل نموذجين من النماذج السابقة للقيام بعملية التصويت وهما CatBoostو XGBoostوكانت النتائج كما هو موضح

بالجدول التالي:

F1 score	Accuracy	
0.88	0.88	**Hard voting**
0.88	0.89	**Soft voting**

جدول 5 نتائج التصنيف من خلال التصويت

نلاحظ أن هناك تحسن ملحوظ باستخدام (Soft Voting) إذ تقوم هذه الطريقة بإعطاء أوزان أعلى للنماذج التي تقوم بإعطاء نتائج دقيقة أكثر. كمحاولة لإيجاد نتائج أفضل تم استخدام النموذج SVM و تدريبه على مجموعات مختلفة من البيانات من خلال الطريقة (bagging and pasting) وكانت النتائج على النحو التالي:

F1 score	Accuracy	
0.75	0.75	**Baging**
0.75	0.75	**Pasting**

نتائج 6 جدول Bagging and Pasting

من الواضح أن هذه العملية لم تعطي نتائج مفيدة.

أخيرا تم التوصل على أن النموذجان CatBoost , XGBoost أعطوا نتائج جيدة جدا وخاصة عند استخدامهما معا من خلال التصويت الموزن بدقو الاجابة (Soft Voting) مقارنة بالاعمال السابقة المنفذة على مجموعة البيانات Dataset Churn Customer وكانت النتيجة النهائية هي ACC=0.89 , F1=0.88 .

1- النتائج للنماذج **Clustering**:

- تم اختيار ثلاث خوارزمية k-mean , Dbscan , Hirecul
- تم اختيار افضل فيتشرز لأفضل المودل الذي قام بإعطاء دقة العالية
- Best Model : CatBoostClassifier

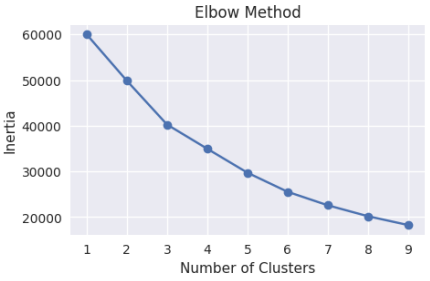
- Top 5 Feature in Best Features: ["Age", "NumOfProducts","Point Earned" "credit_score_rate_by_salary", "Balance", "balance_salary_rate"]

- الخطوات :
  - تطبيق Standard Scalar للفيتشرز
  - تطبيق PCA على نتائج للتقليل الابعاد وعملية العرض تكون سهلة وواضحة
  - تطبيق T-Sne على نتائج لتقليل الأبعاد عندما تكون الفيتشرز غير خطية

- النتائج:

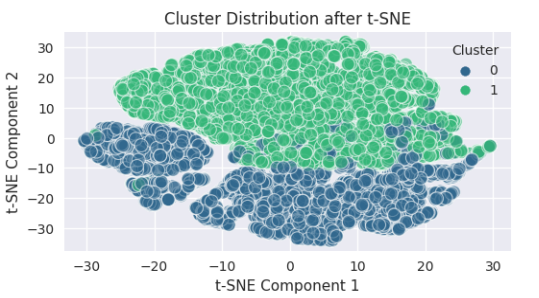
1- K-means

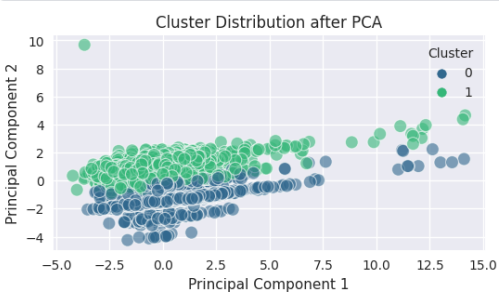
- تم اعتماد على الرسمة Elbow Method لاختيار أفضل قيمة K وكانت أفضل قيمة k=4
- تم اعتماد على قيمة K=2 لانه اقرب للعد كلاسات ضمن عمود exitied



نتائج 1 توزيع كلاسات K-mean

- نلاحظ أن عدد كلاسات 2 وتم تصنيف داتا اكثر شي للكلاس 1
- العرض نتائج باستخدام Pca , T-sne



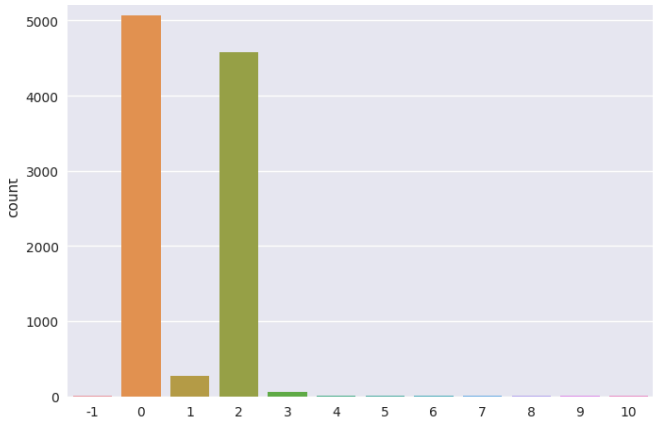


نتائج2 t-sne + K-mean using pca

2- DBSCAN

- تم تطبيق Grid Search للحصول على أفضل قيم ل eps, min_sample وكانت نتيجة نهائية للتدريب هي 0.95 لكن تم تصنيف جميع sample كالكلاس واحد وهذا امر ليس جيدا,
- تم اختيار قيم ل eps , min_sample حيث كانت دقة التدريب أقل 0.50
- تم اعتماد على القيم التالية eps=1.7, min_samples=2
- النتائج :

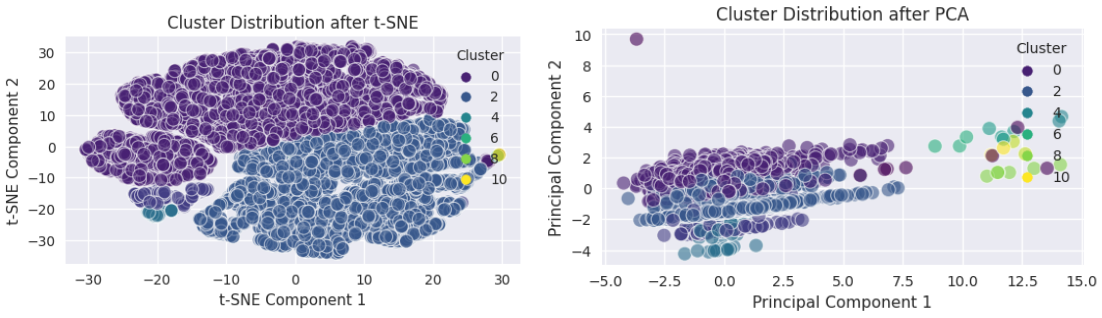
1 - تصنيف بيانات



- نلاحظ من الرسمة ان عدد كلاسات 10 ولكن اكثر قيم مصنفة للكلاس 2 و 3

نتائج 3 توزيع كلاسات Dbscan

- عرض بيانات باستخدام pca, T-sne

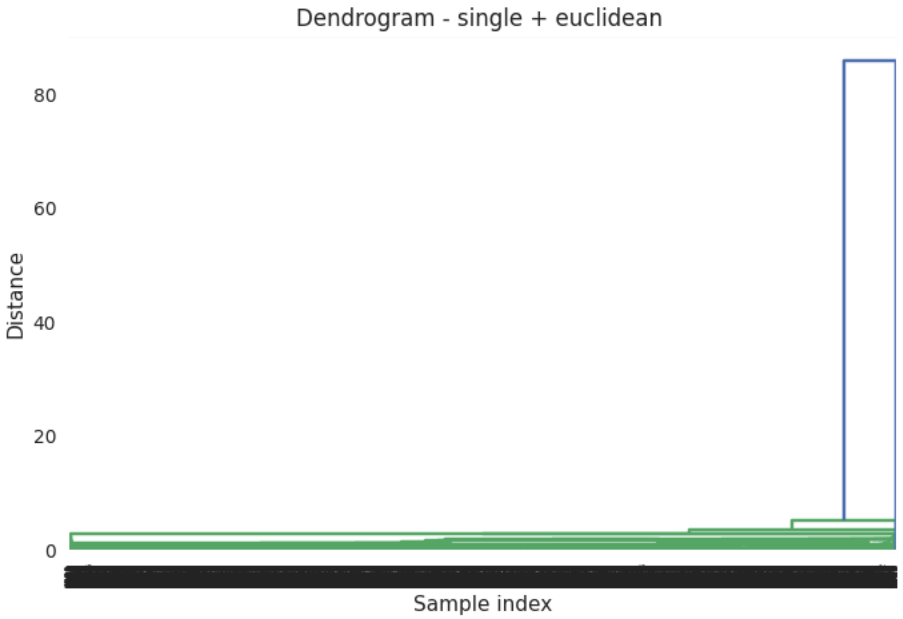


نتائج 3 Dbscan using pca + t-sne

3- Hierarchical Clustering

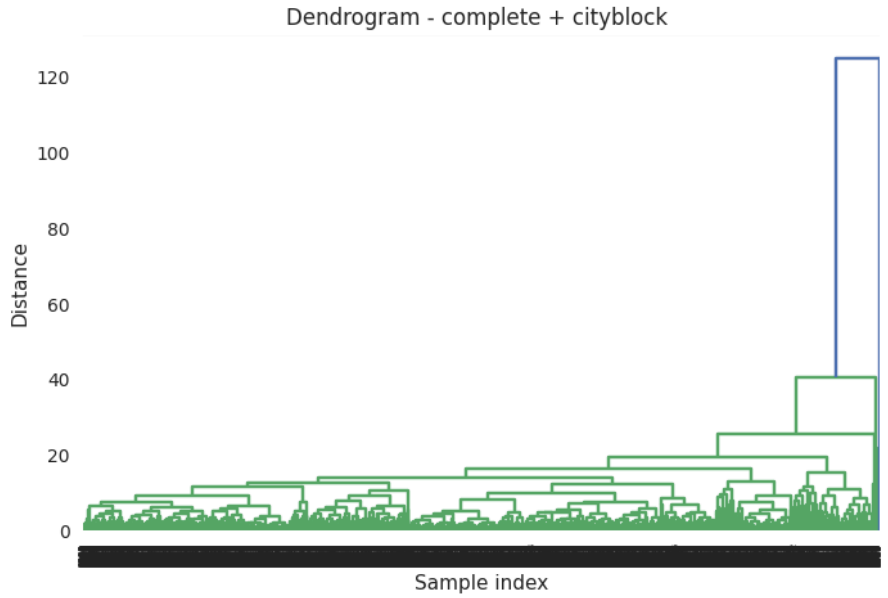
- استخدمنا جميع أنواع مقاييس المسافة ومعايير الارتباط لمعرفة أفضل نتيجة

- single + euclidean



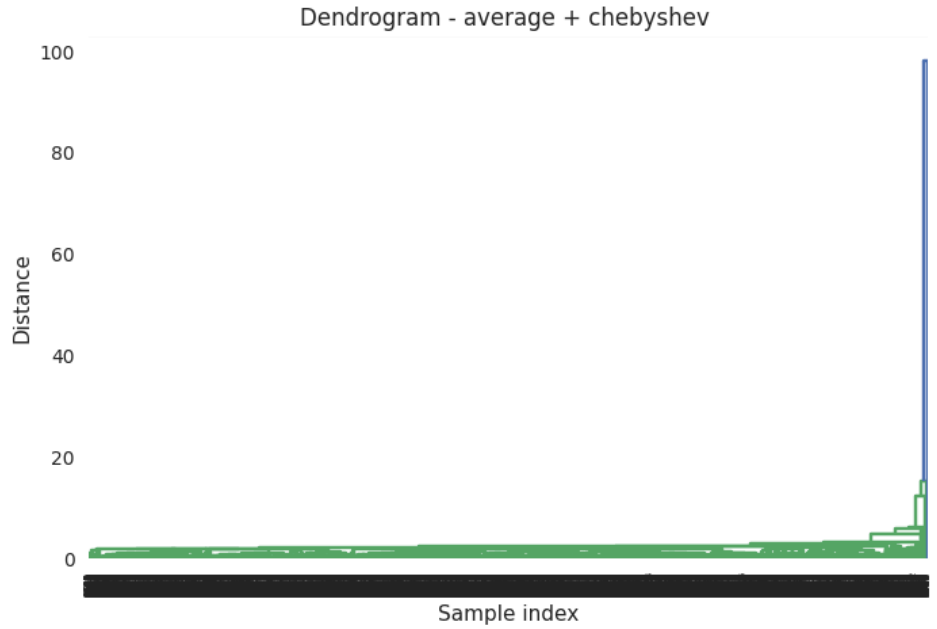
نتائج single + euclidean

- complete + cityblock



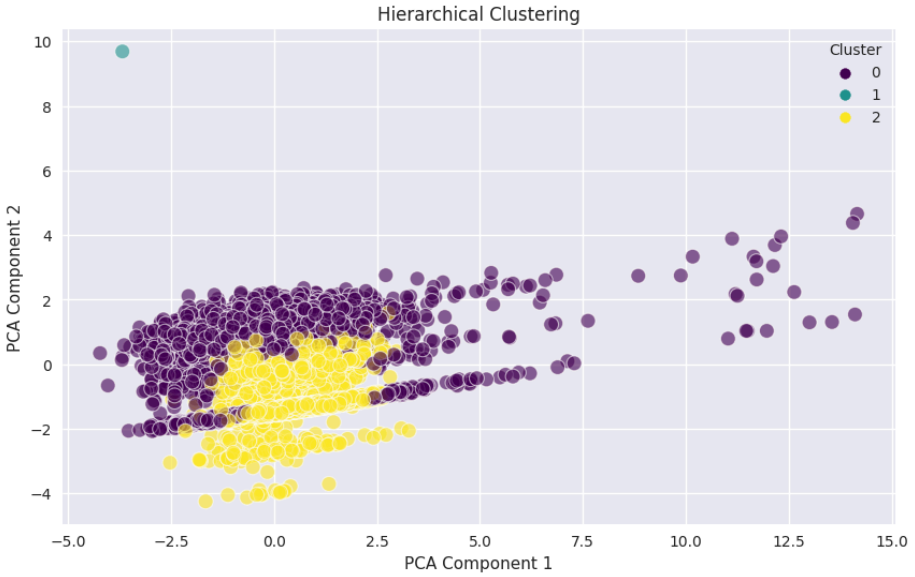
نتائج complete + cityblock

- average + chebyshe



نتائج average + chebyshe

- تشير جميع الطرق الثلاثة إلى أن مجموعتين قد تكونان الخيار الأمثل.



الخاتمة

نظرً ا للقيام بتحليل قاعدة البيانات الخاصة بنا، يمكننا ملاحظة أنه على الرغم من وجود العديد من الأعمدة والمعلومات في قاعدة البيانات، إلا أن عدد العملاء ليس كبيرً ا جدًا، وفيما يتعلق بتحليل بياناتنا، يمكننا أن نرى أن هناك تراب ًطا مطل ًقا بين المتغيرات المتعلقة بالشكاوى والمتغير المستهدف، وبنا ًء على ذلك نستبعد المتغير المتعلق بالشكاوى لكي نتمكن من تشغيل نماذجنا. وعند النظر في التحليل الاستكشافي، يمكننا التحقق من أن جز ًءا كبيرا من بياناتنا موزعة بشكل جيد، ويمكننا ملاحظة بعض الخصائص مثل أن غالبية العملاء من فرنسا، وعادة ما يكون لديهم بين 1و 2 منتج، وجزء كبير من عملائنا ليس لديهم أموال في حساباتهم، ويمكننا أن نرى أن عمر عملائنا يتبع توزيعً ا طبيعيًا. وشيء مهم جدًا رأيناه هو أن المتغير المستهدف لدينا غير متوازن، و أن كبار السن أكثر عرضة للانسحاب. بالنسبة لجزء تعلم الآلة، قمنا بإزالة بعض المتغيرات التي لا تلائم نماذجنا، وقمنا بتحويل المتغيرات التصنيفية إلى متغيرات مستمرة باستخدام ترميز .OneHot Label Encoder بعد معايرة فئة المتغير المستهدف وتشغيل نماذج تعلم الآلة، معظم النماذج حققت نتائج مرضية من حيث الدقة، وكانت أفضل النماذج هي XGBoost , SVMبنسبة دقة ،88% وعندما ننظر إلى المتغيرات الأكثر أهمية، لدينا العمر وعدد المنتجات والرصيد، والعمر هو الأهم منها، وهو ما يؤكد ما رأيناه في تحليلنا الاستكشافي

المراجع

[1] " Ramırez Mart´ıez, Erick David. "Churn detection on bank customers .," [Online]

[2] " Charandabi, Sina E. "Prediction of Customer Churn in Banking Industry.," [Online].

[3] "Hend Sayed, Manal A. Abdel-Fattah, Sherif Kholief " Predicting Potential Banking Customer Churn using Apache Spark ML and MLlib Packages: A Comparative Study," [Online]:

[4][Online]. Available: https://www.kaggle.com/datasets/radheshyamkollipara/bank customer-churn.