الجمهورية العربية السورية

جامعة دمشق

كلية الهندسة المعلوماتية

قسم الذكاء الصنعي

نظام مساعد ذكي في تحليل التقارير السنوية للشركات و التوقع بالافلاس

إعداد الطلاب :

خالد أحمد رشواني

سوزانا حسان حمزة

رأفت زين الدين

ميار نوفل

كرم فروان

بإشراف الدكتورة :

د. ندى غنيم

الفهرس:

المقدمة: 4

المشاكل التي يحلها النظام : 4

ما هو نموذج Q-10 ؟ 5

أهداف المشروع: 6

منهجية العمل: 6

التحديات والاعتبارات: 7

مقارنة المشروع مع تطبيقات أخرى مشابهة : 7

مخطط الاستخدام : 9

Activity diagram : 10

استخدام نظم قواعد المعرفة KBS : 11

1) المكونات الرئيسية لتحليل الربحية: 12

2) المكونات الرئيسية لتحليل الملاءة المالية: 13

3) المكونات الرئيسية لتحليل السيولة : 14

4) نموذج Beneish M-Score: 16

المؤشرات المالية في نموذج M-Score: 16

أهمية M-Score: 18

5) نموذج Piotroski F-Score : 18

6) نموذج Springate Score : 21

استخدام معالجة اللغات الطبيعية NLP : 23

التحليل التنبؤي لتحديد حالة الإفلاس المحتملة للشركات باستخدام ML: 25

1. الفكرة (المشكلة): 25

الحل النظري: 25

الحل التقني : 25

SMOTE (Synthetic Minority Over-sampling Technique) 42

Logistic Regression 46

Gradient Boosting 47

Multilayer Perceptron (MLP) 47
Bagging 48
Decision Tree 49
تحليل مشاعر التقارير المالية: 55
مقدمة : 55
Loughran-McDonald Master Dictionary : 55
FinBert: 55
مكونات التحليل : 55
المكتبات والتقنيات المستخدمة : 57
النتائج : 57
التطويرات والتحديثات المستقبلية: 58
References : 59

المقدمة:

في ظل التطورات السريعة في مجال الذكاء الاصطناعي وتحليل البيانات، أصبح استخدام التقنيات الحديثة ضرورة لتمكين الشركات والمستثمرين من اتخاذ قرارات مالية مدروسة. يعتبر التنبؤ بصحة الشركات المالية وتجنب الإفلاس من أهم التحديات التي تواجهها المؤسسات المالية والمستثمرين. لذلك تم تطوير مشروع يعتمد على الذكاء الاصطناعي لتحليل تقارير Q-10 المالية الفصلية التي تقدمها الشركات المدرجة في الأسواق المالية.

يهدف هذا المشروع إلى استخدام تقنيات معالجة اللغة الطبيعية (NLP) لاستخراج البيانات الحيوية من تقارير Q-10 بدقة وسرعة. بعد ذلك، يتم حساب النسب الاقتصادية الرئيسية باستخدام نظام قائم على المعرفة (KBS)، والذي يحلل الربحية والسيولة والملاءة المالية، بالإضافة إلى اكتشاف التلاعب بالأرباح وقياس الصحة المالية العامة وحالة الشركة. وأخيرًا، يتم تمرير النسب المالية إلى نموذج تعلم آلي (ML) متقدم لتوقع احتمالية إفلاس الشركة في المستقبل القريب. ومن خلال استخدام تحليل المشاعر (Sentiment Analysis)، يتم تقييم اللهجة العامة للنص في التقارير لتقديم صورة شاملة عن النظرة المستقبلية للشركة. يسعى المشروع إلى توفير أداة قوية تساعد في تحليل التقارير المالية بشكل أسرع وأكثر دقة، مما يعزز من قدرة المستثمرين والمؤسسات المالية على اتخاذ قرارات استثمارية وتفادي المخاطر المحتملة.

المشاكل التي يحلها النظام :

في عالم الاستثمار، تعد السرعة والدقة عوامل حاسمة في اتخاذ قرارات البيع والشراء للأسهم. عندما تصدر الشركات تقاريرها المالية الفصلية Q-10، يحتاج المستثمرون عادة إلى وقت طويل لتحليل البيانات واتخاذ قرارات مدروسة، مما يؤدي إلى تغييرات في أسعار الأسهم بعد فترة قصيرة من نشر التقارير. هذا التأخير قد يؤثر سلباً على قدرة المستثمرين على الاستفادة من الفرص المتاحة أو تجنب المخاطر المحتملة بشكل فوري.

يهدف تطبيقنا إلى حل هذه المشكلة من خلال توفير أداة تعتمد على الذكاء الاصطناعي لتحليل نصوص التقارير المالية بسرعة وفعالية. باستخدام تقنيات معالجة اللغة الطبيعية (NLP) ونماذج التعلم الآلي (ML)، يمكن للتطبيق استخراج المعلومات الحيوية وحساب النسب الاقتصادية الهامة في وقت قصير جداً، مما يمنح المستثمرين القدرة على اتخاذ قرارات استثمارية أسرع وأكثر دقة.

هذا يقلل من الفجوة الزمنية بين إصدار التقرير واتخاذ القرار، مما يمكن المستثمرين من الاستجابة السريعة لتغيرات السوق والحفاظ على ميزة تنافسية.

إضافة إلى ذلك، يساعد التطبيق في معالجة الحجم الكبير والمعقد من البيانات الموجودة في تقارير Q-10، مما يوفر الوقت والجهد الذي يتطلبه التحليل اليدوي. كما يقلل من احتمالية وقوع الأخطاء البشرية التي قد تحدث عند تحليل هذه البيانات المعقدة.

يقوم نظام المعرفة (KBS) بتحليل مجموعة واسعة من النسب المالية، مثل الربحية والسيولة والملاءة المالية، ويساعد في اكتشاف أي تلاعب محتمل بالأرباح وتقييم الصحة المالية العامة للشركة. بالإضافة إلى ذلك، يساهم التطبيق في الكشف المبكر عن الإشارات المالية الخطرة، مما يساعد المستثمرين على اتخاذ إجراءات وقائية قبل تفاقم الوضع المالي للشركة.

و بالتالي يعزز التطبيق مستوى الشفافية والثقة من خلال تقديم تحليلات دقيقة قائمة على البيانات، مما يبني ثقة أكبر لدى المستثمرين في اتخاذ قراراتهم.

ما هو نموذج Q-10 ؟

نموذج Q-10 هو تقرير ربع سنوي يجب على الشركات المتداولة علنًا في الولايات المتحدة تقديمه إلى لجنة الأوراق المالية والبورصات (SEC). يوفر هذا التقرير نظرة شاملة على الأداء المالي للشركة ويحتوي على:

- القوائم المالية: تشمل الميزانية العمومية، وقائمة الدخل، وقائمة التدفقات النقدية للربع المالي.
- مناقشة وتحليل الإدارة (MD&A): تقدم تحليلاً لحالة الشركة المالية، بما في ذلك الاتجاهات والمخاطر والتوقعات المستقبلية.
- ملاحظات على القوائم المالية: تقدم معلومات تفصيلية عن السياسات المحاسبية، والالتزامات المحتملة، والبيانات المالية الأخرى المهمة.
- الإجراءات القانونية: تحديثات حول أي إجراءات قانونية مهمة تتعلق بالشركة.
- الإفصاحات: معلومات حول مخاطر السوق، والضوابط الداخلية، وأي أحداث هامة حدثت بعد نهاية الربع المالي.

أهداف المشروع:

يهدف المشروع إلى تحقيق عدة أهداف استراتيجية:

1. **تحسين كفاءة التحليل المالي**: باستخدام تقنيات الذكاء الاصطناعي، سيتمكن المستثمرون من تحليل التقارير المالية بسرعة ودقة أكبر، مما يعزز من قدرتهم على اتخاذ قرارات استثمارية مبنية على بيانات موثوقة.
2. **التنبؤ بالمخاطر المالية**: من خلال استخدام نماذج تعلم آلي متقدمة، يهدف المشروع إلى توفير تنبؤات دقيقة حول احتمالية إفلاس الشركات، مما يمكن المؤسسات المالية من اتخاذ إجراءات وقائية في وقت مبكر.
3. **تقليل التأخير في اتخاذ القرارات**: عبر أتمتة عملية التحليل، يمكن للمستثمرين تقليص الوقت اللازم لدراسة التقارير المالية، مما يتيح لهم الاستجابة بشكل أسرع للتغيرات في السوق.
4. **زيادة الشفافية والدقة**: يسهم النظام في تحسين دقة التحليل المالي من خلال تقديم تقييمات مبنية على المعرفة وبيانات دقيقة، مما يزيد من الثقة في القرارات المالية المبنية على هذه التحليلات.

منهجية العمل:

لتحقيق أهداف المشروع، تم تبني منهجية متعددة المراحل تشمل:

1. **معالجة اللغة الطبيعية (NLP):** تم استخدام تقنيات NLP لاستخراج المعلومات الحيوية من نصوص تقارير Q-10، مثل البيانات المالية، والتعليقات الإدارية، والإفصاحات القانونية.

2. **حساب النسب المالية باستخدام KBS:** بعد استخراج البيانات، يقوم النظام بحساب النسب المالية الرئيسية مثل نسب السيولة، ونسب الربحية، ونسب الدين، والتي تعد مؤشرات هامة على الأداء المالي للشركة.

3. **التنبؤ باستخدام نماذج التعلم الآلي (ML)**: يتم تمرير النسب المالية المحسوبة إلى نموذج تعلم آلي تم تدريبه على مجموعة كبيرة من بيانات الشركات السابقة لتوقع احتمالية الإفلاس.

4. **التقييم والتحسين:** يتم تقييم دقة النموذج بشكل دوري وتحسينه بناءً على نتائج التوقعات الفعلية، مما يضمن تطور الأداء بمرور الوقت.

التحديات والاعتبارات:

- **جودة البيانات** : تعتبر جودة البيانات المستخرجة من التقارير المالية عنصراً حاسماً في دقة التحليل والتنبؤ. لذلك، تم التركيز على تطوير تقنيات متقدمة لتحسين جودة البيانات المستخرجة.

- **تفسير النتائج**: يمثل تفسير نتائج النموذج تحدياً مهماً، حيث يجب أن تكون النتائج واضحة وسهلة الفهم للمستخدمين غير المتخصصين في مجال الذكاء الاصطناعي أو الاقتصاد .

- **التحديث المستمر**: مع تطور الأسواق المالية، يجب أن يتم تحديث النماذج المستخدمة بانتظام لضمان توافقها مع أحدث الاتجاهات الاقتصادية والمالية.

مقارنة المشروع مع تطبيقات أخرى مشابهة :

Sentieo:

الوظيفة: يوفر Sentieo منصة تحليلية شاملة تعتمد على الذكاء الاصطناعي لتحليل التقارير المالية بما في ذلك تقارير Q-10. يستخدم تقنيات البحث المتقدم، وتحليل المشاعر، واستخراج البيانات لتحليل المحتوى المالي.

المزايا: يتيح للمستخدمين التنقيب عن النصوص المالية بسرعة فائقة، ويقدم أدوات لتحليل البيانات بشكل مرئي.

العيوب: يعتمد بشكل أساسي على استخراج المعلومات الموجودة، دون دمج مباشر للنسب المالية مع نماذج تعلم آلي لتوقع الإفلاس.

AlphaSense:

الوظيفة: يستخدم AlphaSense تقنية الذكاء الاصطناعي للبحث والتحليل المالي، مع التركيز على تسهيل البحث عن المعلومات الحيوية في تقارير الشركات.

المزايا: يتميز بسهولة استخدامه وسرعة الوصول إلى المعلومات الحرجة. يوفر أيضًا تحليل مشاعر للنصوص المالية.

العيوب: لا يقدم نماذج متقدمة لتوقع الإفلاس، والتركيز أكثر على البحث من التحليل العميق للنسب المالية.

Bloomberg Terminal:

الوظيفة: منصة معلومات مالية متكاملة توفر بيانات فورية عن الأسواق المالية، بما في ذلك تقارير Q-10، مع أدوات تحليل متقدمة.

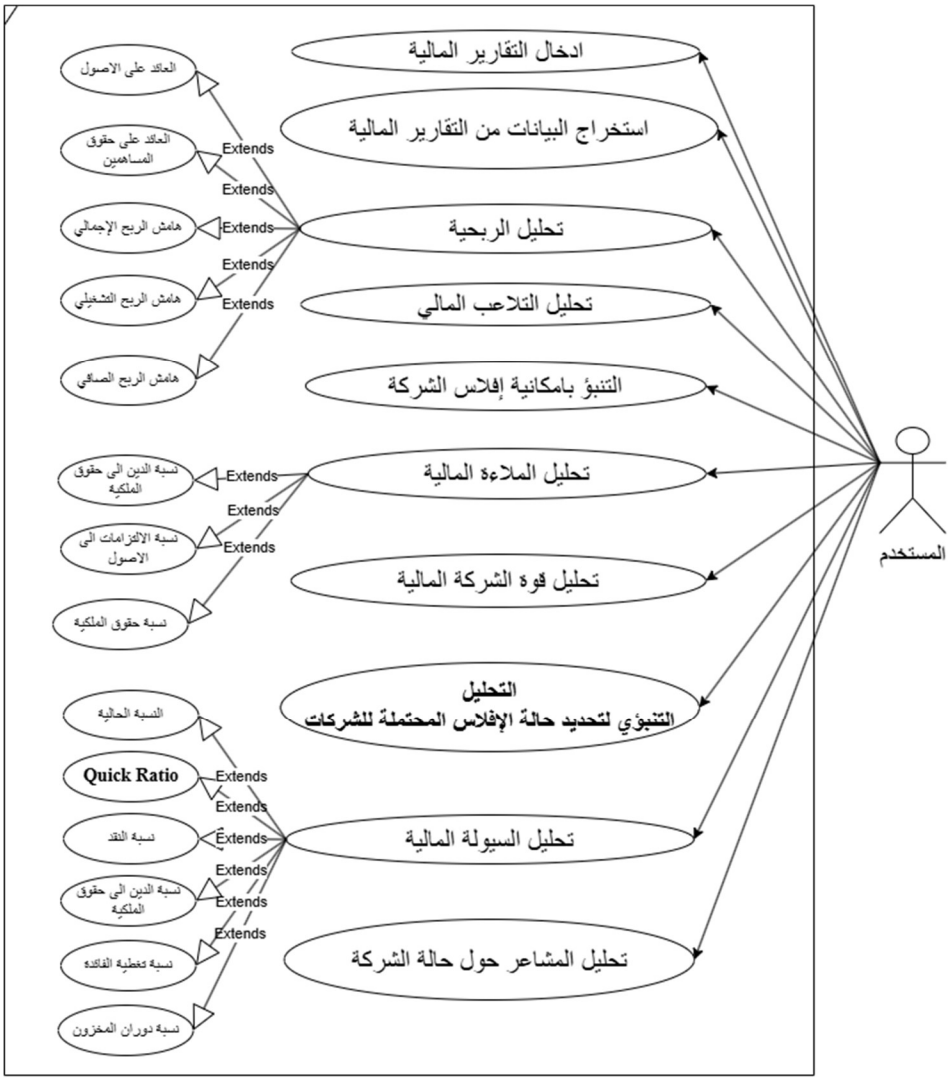
المزايا: يعد من أقوى الأدوات المالية التي توفر معلومات شاملة، مع إمكانيات تحليل متقدمة.

العيوب: تكلفة الاشتراك العالية، ولا يركز بشكل خاص على استخدام الذكاء الاصطناعي لتحليل النسب المالية أو توقع الإفلاس.

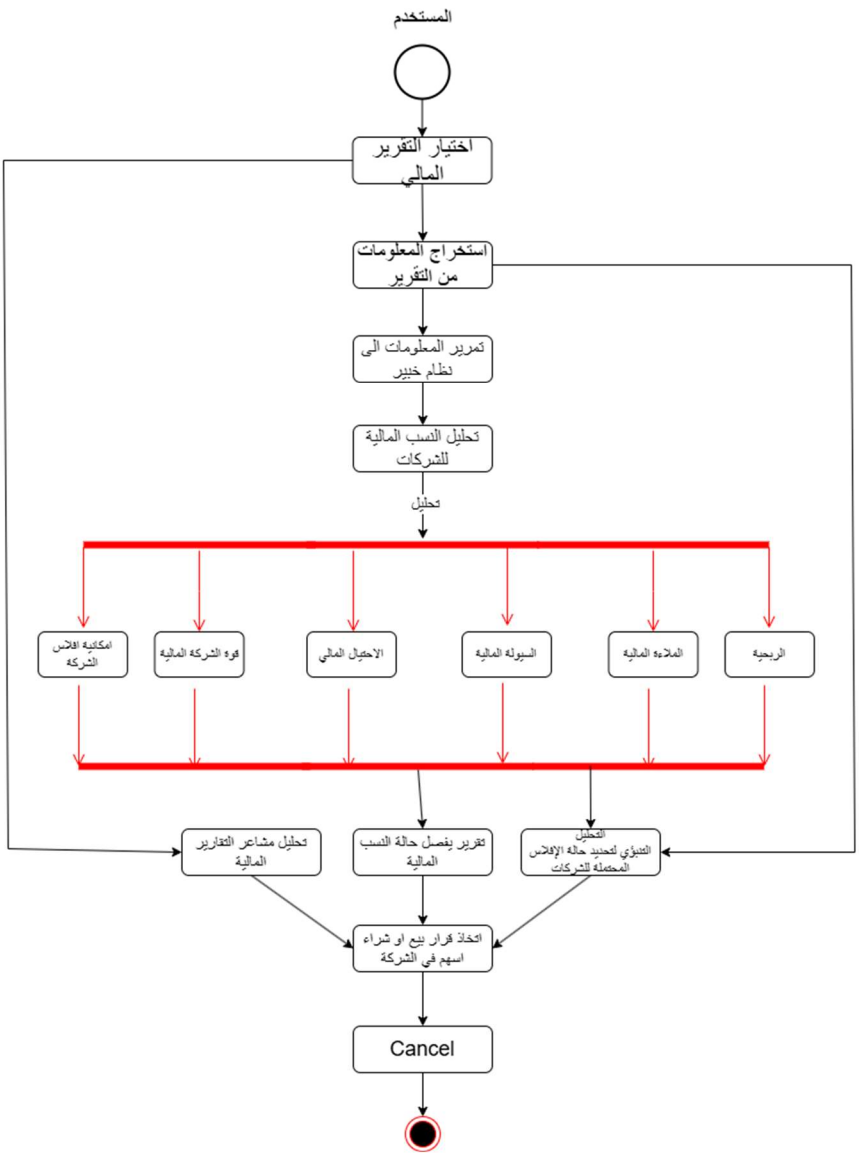
مزايا المشروع المقترح:

تكامل متقدم: يجمع المشروع بين استخراج البيانات، وتحليل النسب المالية، وتوقع الإفلاس باستخدام الذكاء الاصطناعي.

مخطط الاستخدام :



Activity diagram :



استخدام نظم قواعد المعرفة KBS :

كيفية عمل النظام :

النظام يستخدم مكتبة "Experta" في python لتنفيذ القواعد بناءً على الحقائق المالية المستخرجة من التقرير. يقوم النظام بحساب كل من المؤشرات المالية المذكورة في الاسفل باستخدام البيانات المقدمة للسنة الحالية والسنة السابقة، ثم يقوم بتحليل التغيرات في هذه المؤشرات :

يتم تحليل التقرير في نظم قواعد المعرفة وفقا للمحاور التالية :

1 . تحليل الربحية : تحليل الربحية هو عملية تقييم قدرة شركة على تحقيق أرباح من عملياتها. يهدف هذا التحليل إلى فهم مدى كفاءة الشركة في إدارة مواردها وتحويلها إلى أرباح. يتم ذلك من خلال حساب مجموعة من النسب المالية التي تقارن بين الأرباح والتكاليف والإيرادات والأصول. يوفر هذا التحليل فهماً عميقاً لكفاءة الشركة في تحقيق الأرباح.

2 . تحليل الملاءة المالية : حيث أن **الملاءة المالية** تشير إلى قدرة الشركة على الوفاء بجميع التزاماتها المالية طويلة الأجل عند استحقاقها، وضمان استمرارها في العمل دون الحاجة إلى التصفية أو الإفلاس. بعبارة أخرى، هي مقياس لقدرة الشركة على تغطية ديونها وأعبائها المالية باستخدام أصولها الإجمالية.

3 . تحليل السيولة : تحليل السيولة المالية هو عملية تقييم قدرة شركة على سداد التزاماتها قصيرة الأجل، أي ديونها المستحقة خلال العام. هذا التحليل مهم للغاية للمستثمرين والدائنين على حد سواء، حيث يساعدهم في تقييم مدى صحة الشركة المالية وقدرتها على الاستمرار في العمل.

4 . **نموذج** M-Score **:** يستخدم نموذج Beneish M-Score لتحديد ما إذا كانت الشركة قد تلاعبت في أرباحها المالية. يُعد هذا النموذج أداة فعالة للكشف عن الاحتيال المالي.

5 . **نموذج** Piotroski F-Score **:** يقيم هذا النموذج قوة الشركة المالية من خلال تحليل مجموعة من المؤشرات المالية.

6 . **نموذج** Springate Score **:** يستخدم هذا النموذج للتنبؤ بإمكانية إفلاس الشركة من خلال تحليل مجموعة من المؤشرات المالية.

1) المكونات الرئيسية لتحليل الربحية:

1. العائد على الأصول Return on Assets (ROA) :

- التعريف : يقيس العائد على الأصول كفاءة الشركة في استخدام أصولها لتوليد الأرباح.
- طريقة الحساب : يتم حساب ROA بقسمة صافي الدخل على إجمالي الأصول.
- القاعدة : إذا كان العائد على الأصول الحالي أعلى من العائد السابق، فهذا يشير إلى تحسن في كفاءة استخدام الأصول.

2. العائد على حقوق المساهمين Return on Equity (ROE):

- التعريف : يقيس العائد على حقوق المساهمين مقدار العائد الذي تحصل عليه الشركة من حقوق المساهمين.
- طريقة الحساب : يتم حساب ROE بقسمة صافي الدخل على إجمالي حقوق المساهمين.
- القاعدة : إذا كان العائد على حقوق المساهمين الحالي أعلى من العائد السابق، فهذا يشير إلى فعالية الشركة في توليد أرباح من استثمارات المساهمين.

3. هامش الربح الإجمالي Gross Margin:

- التعريف : يقيس هامش الربح الإجمالي مدى ربحية الشركة بعد خصم تكلفة المبيعات من الإيرادات.
- طريقة الحساب : يتم حساب هامش الربح الإجمالي بقسمة الربح الإجمالي على إجمالي الإيرادات.
- القاعدة : زيادة في هامش الربح الإجمالي تشير إلى تحسن في قدرة الشركة على التحكم في التكاليف المباشرة المتعلقة بالإنتاج.

4. **هامش الربح التشغيلي Operating Profit Margin :**

- التعريف : يقيس هامش الربح التشغيلي الربحية التشغيلية للشركة، أي مقدار الربح الناتج عن الأنشطة الأساسية بعد خصم المصاريف التشغيلية.
- طريقة الحساب : يتم حساب هامش الربح التشغيلي بقسمة الدخل التشغيلي على إجمالي الإيرادات.
- القاعدة : زيادة في هامش الربح التشغيلي تشير إلى قدرة الشركة على تحسين الكفاءة التشغيلية وخفض التكاليف غير المباشرة.

5. **هامش الربح الصافي Net Profit Margin** :

- التعريف : يقيس هامش الربح الصافي نسبة الربح المتبقي للشركة بعد خصم جميع التكاليف، بما في ذلك الضرائب والفوائد.
- طريقة الحساب : يتم حساب هامش الربح الصافي بقسمة صافي الدخل على إجمالي الإيرادات.
- القاعدة : زيادة في هامش الربح الصافي تشير إلى تحسن في الأداء المالي العام للشركة.

2) المكونات الرئيسية لتحليل الملاءة المالية:

1. **نسبة الدين إلى حقوق الملكية Debt-to-Equity Ratio :**

- التعريف : تقيس نسبة الدين إلى حقوق الملكية مقدار الدين الذي تستخدمه الشركة مقارنة برأس مال المساهمين.
- طريقة الحساب : يتم حساب هذه النسبة بقسمة إجمالي الالتزامات على إجمالي حقوق المساهمين.
- القاعدة : زيادة هذه النسبة تشير إلى ارتفاع في مستوى الدين بالنسبة لرأس المال المملوك، مما قد يزيد من المخاطر المالية.

2. **نسبة الالتزامات إلى الأصول Total Liabilities/Total Assets :**

- التعريف : تقيس نسبة الالتزامات إلى الأصول مدى اعتماد الشركة على الديون في تمويل أصولها.
- طريقة الحساب : يتم حساب هذه النسبة بقسمة إجمالي الالتزامات على إجمالي الأصول.
- القاعدة : زيادة هذه النسبة تشير إلى ارتفاع في مستوى الدين المستخدم لتمويل الأصول، مما قد يزيد من المخاطر المالية.

3. **نسبة حقوق الملكية Equity Ratio:**

- التعريف : تقيس نسبة حقوق الملكية الجزء الذي يمثله رأس المال المملوك من إجمالي الأصول.
- طريقة الحساب : يتم حساب هذه النسبة بقسمة إجمالي حقوق المساهمين على إجمالي الأصول.
- القاعدة : زيادة هذه النسبة تشير إلى انخفاض في مستوى الدين المستخدم وتمثل حالة مالية أكثر أمانًا.

4. **نسبة الدين Debt Ratio :**

- التعريف : تقيس نسبة الدين إجمالي الدين كمقدار من إجمالي الأصول.
- طريقة الحساب : يتم حساب هذه النسبة بقسمة إجمالي الالتزامات على إجمالي الأصول.
- القاعدة : ارتفاع نسبة الدين يشير إلى ارتفاع في مخاطر التمويل، حيث تعتمد الشركة بشكل أكبر على الديون.

3) المكونات الرئيسية لتحليل السيولة :

1. النسبة الحالية Current Ratio :

- يتم حسابها بقسمة إجمالي الأصول الحالية على إجمالي الخصوم الحالية. هذه النسبة تعكس قدرة الشركة على سداد التزاماتها قصيرة الأجل باستخدام أصولها الحالية.
- **التفسير:** إذا كانت النسبة الحالية للشركة عالية (أكبر من 1)، فإن ذلك يعني أن الشركة تتمتع بسيولة جيدة. إذا كانت منخفضة (أقل من 1)، فقد تواجه الشركة صعوبات في تلبية التزاماتها الفورية.

2. حساب Quick Ratio :

- يتم حسابها بقسمة إجمالي الأصول الحالية (باستثناء المخزون) على إجمالي الخصوم الحالية. تعكس هذه النسبة قدرة الشركة على سداد التزاماتها قصيرة الأجل باستخدام أصولها السائلة فقط، دون الاعتماد على بيع المخزون.
- **التفسير :** نسبة سريعة عالية تشير إلى أن الشركة يمكنها الوفاء بالتزاماتها بسرعة دون الحاجة إلى تصفية المخزون، مما يدل على قوة السيولة.

3. نسبة النقد Cash Ratio :

- يتم حسابها بقسمة النقد وما يعادله على إجمالي الخصوم الحالية. هذه النسبة تعتبر الأكثر تحفظًا حيث تقيس القدرة على الوفاء بالالتزامات الفورية باستخدام النقد فقط.
- **التفسير :** نسبة النقد الأعلى من 1 تعني أن الشركة يمكنها تغطية جميع التزاماتها الحالية بالنقد فقط، مما يدل على استقرار مالي كبير.

4. نسبة الدين إلى حقوق الملكية Debt to Equity Ratio :

- تقيس هذه النسبة مدى اعتماد الشركة على الديون لتمويل أصولها، مقارنة بحقوق المساهمين.
- **التفسير :** نسبة أعلى تعني أن الشركة تعتمد بشكل أكبر على الديون، مما يزيد من المخاطر المالية. نسبة منخفضة تشير إلى أن الشركة تعتمد بشكل أكبر على تمويل حقوق الملكية، مما يقلل من المخاطر.

5. نسبة تغطية الفائدة Interest Coverage Ratio :

- تقيس قدرة الشركة على دفع مصاريف الفوائد من أرباحها التشغيلية.
- **التفسير :** نسبة تغطية الفائدة العالية تشير إلى أن الشركة قادرة على الوفاء بتكاليف الفائدة بسهولة، مما يدل على وضع مالي صحي.

6. نسبة دوران المخزون Inventory Turnover Ratio :

- تقيس مدى كفاءة الشركة في إدارة مخزونها وبيعه.

- **التفسير :** نسبة دوران المخزون العالية تشير إلى أن الشركة تدير مخزونها بكفاءة ولديها مبيعات قوية، بينما تشير النسبة المنخفضة إلى مشاكل محتملة في الطلب أو فائض المخزون.

4) نموذج Beneish M-Score:

نموذج Beneish M-Score هو أداة تحليلية تستخدم لتحديد ما إذا كانت الشركة قد تلاعبت في أرباحها المالية من خلال تحليل مجموعة من النسب المالية. يعتمد هذا النموذج على مجموعة من المؤشرات المالية التي تقيس مدى احتمال وجود احتيال مالي في القوائم المالية للشركة. تم تطوير هذا النموذج من قبل الأستاذ Daniel Beneish .M في عام 1999، ويعتبر واحداً من الأدوات الهامة المستخدمة في مجال المحاسبة الجنائية وتحليل الاحتيال.

المؤشرات المالية في نموذج M-Score:

مؤشر :Days' Sales in Receivables Index (DSRI)

- يمثل هذا المؤشر زيادة كبيرة في أيام التحصيل على الديون المستحقة، وهو ما قد يشير إلى تسريع الاعتراف بالإيرادات لزيادة الأرباح. ارتفاع هذا المؤشر يمكن أن يدل على أن الشركة تحاول تضخيم أرباحها من خلال إيرادات لم يتم تحصيلها بعد.

مؤشر :Gross Margin Index (GMI)

- يشير هذا المؤشر إلى تدهور في هامش الربح الإجمالي، وهو ما يرسل إشارة سلبية حول مستقبل الشركة ويزيد من الحافز لتضخيم الأرباح. إذا كانت الشركة تعاني من تراجع في هوامش الربح، فقد تلجأ إلى التلاعب بالأرباح لتحسين صورتها المالية.

مؤشر :Asset Quality Index (AQI)

- يقيس هذا المؤشر زيادة في الأصول طويلة الأجل (مثل رسملة التكاليف) بالنسبة إلى إجمالي الأصول، مما قد يشير إلى زيادة في تأجيل التكاليف بهدف تضخيم

الأرباح. استخدام مثل هذه الاستراتيجيات يمكن أن يكون مؤشراً على محاولة الشركة إخفاء ضعف الأداء الفعلي.

مؤشر Sales Growth Index (SGI):

- يركز هذا المؤشر على نمو المبيعات، حيث أن الشركات ذات النمو المرتفع ليست بالضرورة متلاعبة، لكن الشركات ذات النمو المرتفع قد تكون أكثر عرضة لارتكاب الاحتيال المالي بسبب الضغوط المالية والاحتياجات الرأسمالية لتحقيق أهداف الأرباح.

مؤشر Depreciation Index (DEPI):

- يشير انخفاض مستوى الاستهلاك بالنسبة إلى الأصول الثابتة الصافية إلى احتمال أن تكون الشركة قد عدلت تقديرات العمر الافتراضي للأصول أو تبنت طريقة جديدة تزيد من الدخل. هذا التغيير قد يكون مؤشرًا على محاولة تضخيم الأرباح.

مؤشر Sales, General and Administrative Expenses Index (SGAI):

- يعتبر المحللون أن الزيادة غير المتناسبة في مصاريف البيع والإدارة العامة بالنسبة إلى المبيعات إشارة سلبية حول آفاق الشركة المستقبلية، مما قد يحفز الشركة على تضخيم الأرباح.

مؤشر Leverage Index (LVGI):

- يقيس هذا المؤشر الزيادة في الرافعة المالية من خلال إجمالي الديون بالنسبة إلى إجمالي الأصول. زيادة الرافعة المالية قد تدفع الشركة إلى التلاعب بالأرباح لتحقيق متطلبات الديون.

مؤشر Total Accruals to Total Assets (TATA):

- يعكس هذا المؤشر مستوى التجميعات الكلية بالنسبة إلى إجمالي الأصول، حيث يشير ارتفاع مستوى التجميعات إلى احتمالية أكبر للتلاعب بالأرباح من خلال الخيارات المحاسبية التقديرية.

تفسير :M-Score

- إذا كان **M-Score** أقل من -2.22، فهذا يشير إلى أن الشركة من المحتمل أن تكون لم تتلاعب بأرباحها.
- إذا كان **M-Score** أعلى من -2.22، فهذا يشير إلى أن الشركة قد تكون تلاعبت بأرباحها.

أهمية M-Score:

نموذج M-Score هو أداة قوية للكشف عن الاحتيال المالي، ويمكن استخدامه من قبل المحللين الماليين والمراجعين والمستثمرين لفحص صدقية القوائم المالية. يساعد النموذج في الكشف عن الشركات التي قد تلجأ إلى التضليل في الإفصاح عن أرباحها بهدف تحسين صورتها أمام المستثمرين والسوق. يعد هذا النموذج جزءاً مهماً من ترسانة الأدوات المستخدمة في تقييم الأداء المالي للشركات وتجنب الاستثمارات الخطرة.

5) نموذج Piotroski F-Score :

تمثيل نموذج Piotroski F-Score باستخدام نظام قائم على المعرفة (: (KBS

في إطار تعزيز قدرة النظام القائم على المعرفة (KBS) في تحليل الأداء المالي للشركات وتقييم جودة أرباحها، تم دمج نموذج **Piotroski F-Score** كجزء أساسي من منظومتنا التحليلية. يُعتبر Piotroski F-Score أداة مهمة لقياس جودة أرباح الشركات وتقييم مدى صحة أوضاعها المالية، وهو يعتمد على مجموعة من النسب المالية المشتقة من القوائم المالية.

كيف يعمل : Piotroski F-Score

طور المحلل المالي جوزيف بيتروسكي (Joseph Piotroski) نموذج F-Score لتقييم الشركات التي تتمتع بتقييم منخفض (قيمة دفترية أقل) ولكن ذات جودة أرباح عالية. يعتمد النموذج على تسع إشارات مالية (مؤشرات) تُصنّف ضمن ثلاث فئات رئيسية: الربحية، الكفاءة التشغيلية، وهيكل رأس المال.

الفئات الثلاثة لمؤشرات Piotroski F-Score

1) مؤشرات الربحية:

العائد على الأصول : يُقاس العائد على الأصول كمؤشر على قدرة الشركة على تحقيق الأرباح باستخدام أصولها.

التدفق النقدي من العمليات (CFO): يُشير إلى قدرة الشركة على توليد النقدية من عملياتها الأساسية.

تغيير العائد على الأصول: يُقارن العائد على الاصول بين السنة الحالية والسنة السابقة لتحديد الاتجاه.

جودة الأرباح: يُقارن بين التدفق النقدي من العمليات وصافي الربح لمعرفة مدى جودة الأرباح.

2) مؤشرات الكفاءة التشغيلية:

التغيير في هامش الربح الإجمالي: يقيس التحسن أو التدهور في كفاءة العمليات من خلال مقارنة هامش الربح الإجمالي بين العامين.

التغيير في دوران الأصول: يُقيس كفاءة استخدام الشركة لأصولها في توليد الإيرادات.

3) مؤشرات هيكل رأس المال:

التغيير في الرافعة المالية: يُحلل التغيرات في نسبة الديون إلى حقوق الملكية.

التغيير في السيولة: يقيس قدرة الشركة على الوفاء بالتزاماتها قصيرة الأجل من خلال مقارنة نسبة السيولة الحالية بين العامين.

إصدار الأسهم: ينظر في ما إذا كانت الشركة قد أصدرت أسهم جديدة لتمويل عملياتها، مما قد يُشير إلى مخاطر تخفيف حقوق المساهمين الحاليين.

كيف تم دمج Piotroski F-Score في النظام القائم على المعرفة (KBS)؟

استخراج البيانات المالية:

كما هو الحال مع M-Score، يتم أولاً استخدام تقنيات معالجة اللغة الطبيعية (NLP) لاستخراج البيانات المالية الأساسية اللازمة لحساب مؤشرات Piotroski F-Score من التقارير المالية، مثل العائد على الأصول، التدفق النقدي من العمليات، ونسبة الرافعة المالية.

حساب المؤشرات:

بعد استخراج البيانات، يقوم النظام بحساب المؤشرات التسعة بناءً على القيم المستخرجة. يتم تمثيل كل مؤشر بنقطة واحدة (1) إذا كان إيجابيًا وبصفر (0) إذا كان سلبيًا.

تطبيق نموذج :Piotroski F-Score

بعد حساب جميع المؤشرات، يتم تجميع النقاط للحصول على Piotroski F-Score الذي يتراوح بين 0 و9. كلما كانت القيمة أعلى، كانت جودة أرباح الشركة وأدائها المالي أفضل.

تفسير النتائج:

يقوم النظام بتفسير النتيجة النهائية وتقديم توصيات بناءً على Piotroski F-Score. على سبيل المثال، إذا حصلت الشركة على درجة عالية (7-9)، يُشير ذلك إلى أنها تتمتع بأداء مالي قوي وربحية عالية. إذا كانت النتيجة منخفضة (0-3)، فقد يُشير ذلك إلى مشاكل مالية محتملة.

الفوائد من دمج : Piotroski F-Score

تحليل شامل للأداء المالي:

يسمح النظام القائم على المعرفة بتحليل شامل لأداء الشركات باستخدام Piotroski F-Score، مما يوفر صورة واضحة عن جودة أرباح الشركة وكفاءة عملياتها المالية.

الكشف المبكر عن المخاطر:

يُمكن لهذا النظام أن يكون أداة قوية في الكشف المبكر عن المخاطر المالية وتحذير المستخدمين من الشركات التي قد تكون في وضع مالي ضعيف.

توفير الوقت والموارد:

يساعد النظام على توفير الوقت والموارد عن طريق إجراء تحليل مالي معقد بطريقة آلية ودقيقة، مما يُتيح للمستثمرين التركيز على اتخاذ القرارات الإستراتيجية.

تحسين دقة التحليل:

من خلال دمج Piotroski F-Score مع تقنيات KBS، يمكن تحقيق دقة أكبر في تحليل جودة أرباح الشركات وتقييم استقرارها المالي على المدى الطويل.

6) نموذج Springate Score :

كيف يعمل Springate Score :

تم تطوير Springate Score بواسطة المحلل جوردون سبرينغيت في عام 1978 كوسيلة لتقييم الاستقرار المالي للشركات وتحديد ما إذا كانت عرضة لخطر الإفلاس. يعتمد النموذج على مجموعة من النسب المالية المستخلصة من القوائم المالية، ويتم تجميعها للحصول على نتيجة نهائية تعكس الوضع المالي للشركة.

النسب المالية المستخدمة في :Springate Score

يتم حساب Springate Score باستخدام أربع نسب مالية رئيسية:

1) نسبة رأس المال العامل إلى إجمالي الأصول:

تقيس هذه النسبة كفاءة الشركة في استخدام رأس المال العامل لتغطية أصولها الإجمالية، وتعد مؤشرًا على قدرة الشركة على تلبية التزاماتها قصيرة الأجل.

2) نسبة الأرباح قبل الفوائد والضرائب إلى إجمالي الأصول:

تقيس هذه النسبة مدى كفاءة الشركة في توليد الأرباح باستخدام أصولها، وهي مؤشر على الربحية التشغيلية للشركة.

3) نسبة الأرباح قبل الفوائد والضرائب إلى الخصوم المتداولة:

تعكس هذه النسبة قدرة الشركة على توليد الأرباح الكافية لتغطية التزاماتها الحالية، مما يعكس مستوى المخاطر المالية.

4) نسبة المبيعات إلى إجمالي الأصول:

تقيس هذه النسبة قدرة الشركة على استخدام أصولها لتحقيق المبيعات، وهي مؤشر على كفاءة تشغيل الشركة.

كيفية دمج Springate Score في النظام القائم على المعرفة :

حساب النسب المالية:

يقوم النظام بحساب النسب المالية الأربعة باستخدام البيانات المستخرجة، ثم يتم دمجها في معادلة Springate Score لحساب النتيجة النهائية.

تفسير النتيجة:

بناءً على النتيجة النهائية لـ Springate Score، يقوم النظام بتقديم توصيات تتعلق بالاستقرار المالي للشركة. إذا كانت النتيجة **أكبر من 0.862**، فإن النظام يفسر ذلك بأن **الشركة في حالة مستقرة**، مما يعني أن الشركة لا تواجه مخاطر مالية كبيرة. أما إذا كانت النتيجة **أقل من أو تساوي 0.862**، فإن النظام يشير إلى أن **الشركة قد تكون تحت ضغوط مالية**، مما يتطلب متابعة دقيقة للمخاطر المالية المحتملة.

استخدام معالجة اللغات الطبيعية NLP :

يهدف إلى استخراج وتحليل البيانات المالية من ملف PDF يحتوي على التقارير المالية لشركة معينة. يتم ذلك من خلال عدة خطوات تشمل قراءة الملف، تنظيف النصوص، استخراج القيم المالية، وتحويلها إلى صيغة يتم استخدامها في نواحي المشروع الاخرى .

1. استيراد المكتبات اللازمة :

نقوم باستيراد عدة مكتبات أساسية نذكر منها :

re: مكتبة مدمجة في بايثون لمعالجة النصوص باستخدام التعبيرات النمطية (Regular Expressions).

PyPDF2: مكتبة تُستخدم لاستخراج النصوص من ملفات .PDF

2. فتح ملف PDF :

يتم فتح ملف PDF والذي يحتوي على التقرير المالي لشركة ما ويُستخدم لذلك وضعية القراءة الثنائية (.(binary read mode

3. تعريف متغيرات مالية وقيمها المحتملة :

يتم إنشاء قاموس يحتوي على المتغيرات المالية المهمة كـ "صافي الدخل"، "إجمالي الأصول"، إلخ. كل متغير يحتوي على قائمة من المفاتيح المحتملة (تسميات مختلفة قد تظهر في التقرير للإشارة إلى نفس المتغير).

4. تعريف الأنماط لتنظيف النص :

يتم إعداد قائمة من الأنماط لتنظيف النصوص المستخرجة من ملف PDF. تشمل هذه الأنماط:

إزالة فواصل الأسطر واستبدالها بمسافات.

إزالة الأحرف الخاصة مثل علامات الترقيم والرموز.

تقليص المسافات المتعددة إلى مسافة واحدة.

5. استخراج النص من ملف : PDF

يتم استخدام مكتبة PyPDF2 لقراءة جميع صفحات ملف PDF واستخراج النص منها. يتم تخزين النص المستخرج في قائمة.

6. تنظيف النص المستخرج :

كل صفحة من النص المستخرج يتم تنظيفها باستخدام الأنماط المحددة مسبقًا لضمان إزالة الأحرف غير الضرورية وجعل النص مناسبًا للمعالجة اللاحقة.

7. استخراج القيم المالية :

يتم استخدام تعبير نمطي (Regular Expression) للبحث عن القيم المالية المرتبطة بكل متغير. يتم البحث في النص عن كل مفتاح من المفاتيح المحتملة المرتبطة بالمتغيرات المالية وتعريف نمط لاستخراج الأرقام المرتبطة بكل متغير.

8. تحويل القيم المستخرجة :

القيم المالية المستخرجة عادة ما تكون في شكل نصي مع وجود فواصل الأرقام، لذلك يتم تحويلها إلى أرقام عشرية (float) بعد إزالة الفواصل.

9. تحويل القيم إلى الملايين :

يتم تحويل القيم المالية المستخرجة إلى وحدة الملايين نظرا لان الارقام في القوائم واحدتها المليون.

10. عرض النتائج :

في النهاية، يتم عرض القيم المالية المستخرجة من التقرير المالي بشكل منظم، حيث يتم عرض كل متغير مع قيمته الحالية والقيمة المقابلة له من السنة المالية السابقة.

التحليل التنبؤي لتحديد حالة الإفلاس المحتملة للشركات باستخدام ML:

1. الفكرة (المشكلة):

في ظل التحديات الاقتصادية المتزايدة وعدم الاستقرار المالي الذي تعاني منه العديد من الشركات حول العالم، أصبح من الضروري تطوير أدوات تساعد في التنبؤ بالحالة المالية المستقبلية للشركات. من أهم هذه التحديات هو التنبؤ بإمكانية إفلاس الشركات، وهو ما يتيح للإدارات اتخاذ التدابير الوقائية اللازمة لتجنب السقوط في أزمة مالية حادة

يمثل الإفلاس تهديدًا كبيرًا للشركات، حيث يؤدي إلى خسائر كبيرة على مستوى المساهمين والموظفين والعملاء وحتى الاقتصاد العام. لذلك، يعد التنبؤ المبكر بالإفلاس خطوة حاسمة تتيح للشركات اتخاذ إجراءات تصحيحية قبل فوات الأوان

الحل النظري:

يمكن استخدام البيانات المالية والتشغيلية للشركة للتنبؤ حول حالتها المستقبلية اذ أن التعلم الآلي يمكن أن يساعد في تحديد الأنماط والمؤشرات التي تسبق حدوث الإفلاس.

التنبؤ بالإفلاس يعتمد على تحليل البيانات المالية والتشغيلية للشركات. يتم استخدام تقنيات مختلفة من التحليل الإحصائي والذكاء الاصطناعي (مثل التعلم الآلي) لتحديد المتغيرات الرئيسية التي تساهم في تحديد إمكانية الإفلاس. من بين هذه المتغيرات: نسب الربحية، مستويات الديون، التدفقات النقدية، والقدرة على الوفاء بالالتزامات المالية.

الحل التقني :

لنتحدث بدايةً عن الداتا سيت المستخدمة :

تم جمع البيانات من مجلة تايوان الاقتصادية للسنوات من 1999 إلى 2009. وتم تعريف إفلاس الشركات استناداً إلى اللوائح التجارية لبورصة تايوان للأوراق المالية .

	Bankrupt?	ROA(C) before interest and depreciation before interest	ROA(A) before interest and % after tax	ROA(B) before interest and depreciation after tax	Operating Gross Margin	Realized Sales Gross Margin	Operating Profit Rate	Pre-tax net Interest Rate	After-tax net Interest Rate	Non-industry income and expenditure/revenue	...	Net Income to Total Assets	Total assets to GNP price	No-credit Interval
0	1	0.370594	0.424389	0.405750	0.601457	0.601457	0.998969	0.796887	0.808809	0.302646	...	0.716845	0.009219	0.622879
1	1	0.464291	0.538214	0.516730	0.610235	0.610235	0.998946	0.797380	0.809301	0.303556	...	0.795297	0.008323	0.623652
2	1	0.426071	0.499019	0.472295	0.601450	0.601364	0.998857	0.796403	0.808388	0.302035	...	0.774670	0.040003	0.623841
3	1	0.399844	0.451265	0.457733	0.583541	0.583541	0.998700	0.796967	0.808966	0.303350	...	0.739555	0.003252	0.622929
4	1	0.465022	0.538432	0.522298	0.598783	0.598783	0.998973	0.797366	0.809304	0.303475	...	0.795016	0.003878	0.623521
5	1	0.388680	0.415177	0.419134	0.590171	0.590251	0.998758	0.796903	0.808771	0.303116	...	0.710420	0.005278	0.622605
6	0	0.390923	0.445704	0.436158	0.619950	0.619950	0.998993	0.797012	0.808960	0.302814	...	0.736619	0.018372	0.623655
7	0	0.508361	0.570922	0.559077	0.601738	0.601717	0.999009	0.797449	0.809362	0.303545	...	0.815350	0.010005	0.623843
8	0	0.488519	0.545137	0.543284	0.603612	0.603612	0.998961	0.797414	0.809338	0.303584	...	0.803647	0.000824	0.623977
9	0	0.495686	0.550916	0.542963	0.599209	0.599209	0.999001	0.797404	0.809320	0.303483	...	0.804195	0.005798	0.623865

الداتا مؤلفة من 96 سمة و 6819 مثال

Bankrupt?	0
ROA(C) before interest and depreciation before interest	0
ROA(A) before interest and % after tax	0
ROA(B) before interest and depreciation after tax	0
Operating Gross Margin	0
..	
Liability to Equity	0
Degree of Financial Leverage (DFL)	0
Interest Coverage Ratio (Interest expense to EBIT)	0
Net Income Flag	0
Equity to Liability	0

تم التأكد بأن الداتا سيت ل اتحتوي على أي قيم مفقودة أي أنه لا توجد حاجة لاتخاذ أي إجراء إضافي لمعالجة القيم المفقودة مثل الاستبدال بالقيم المتوسطة أو الحذف لأن البيانات مكتملة في جميع الأعمدة.

كما نلاحظ وجود 96 سمة فإن التعامل مع هذا العدد الكبير من السمات قد يؤدي الى زيادة التعقيد الحسابي حيث أنه كلما زاد عدد السمات، زاد التعقيد الحسابي للنماذج. هذا يؤدي إلى زيادة الوقت اللازم لتدريب النموذج وزيادة استهلاك الموارد الحاسوبية (مثل الذاكرة والمعالجة)

وهناك احتمال كبير أن تكون بعض هذه السمات مرتبطة ببعضها البعض بقوة، مما يؤدي إلى التعدد الخطي. هذا يؤدي إلى نتائج نمذجة غير دقيقة وصعوبة في تفسير النموذج

وجود عدد كبير من السمات يزيد من احتمالية أن يكون النموذج حساسًا جدًا للبيانات التي يتم تدريبه عليها. هذا يؤدي إلى بناء نموذج جيد جدًا في تفسير البيانات التدريبية ولكنه يفشل في التعميم على البيانات الجديدة

هذا ما دفعنا الى القيام بتحليل الارتباط وتقليل السمات:

حيث تحليل الارتباط يساعد في تحديد السمات التي تحمل معلومات زائدة عن الحاجة بسبب علاقتها القوية مع سمات أخرى. من خلال تقليل عدد السمات، يصبح النموذج أبسط وأسرع في التدريب وأقل استهلاكًا للموارد.

و يساعد في اكتشاف السمات المرتبطة ببعضها بشكل كبير. من خلال حذف أو دمج السمات المرتبطة، يمكن تقليل التعدد الخطي وتحسين دقة النموذج.

كما أن بتقليل عدد السمات يقوم بالتركيز على السمات الأكثر أهمية، مما يسهل معالجة البيانات المفقودة أو التعامل مع البيانات بشكل أكثر فعالية.

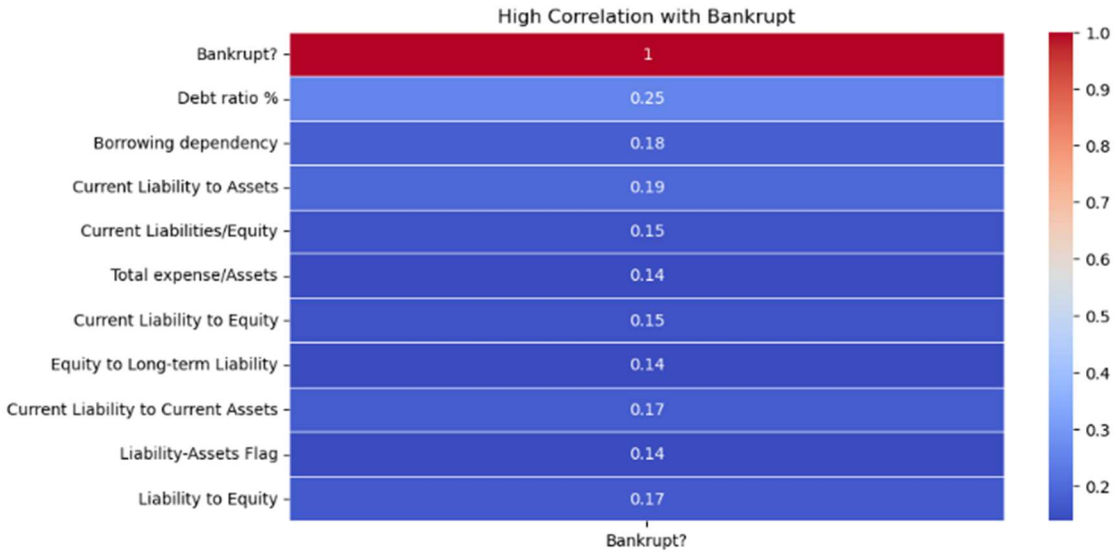
مما يؤدي الى بناء نموذج أكثر قدرة على التعميم، مما يقلل من احتمالية الإفراط في التكيّف على البيانات التدريبية.

فنحن قمنا بتحليل العلاقات بين السمة المستهدفة وبقية السمات في البيانات. حيث الهدف هو تحديد السمات التي لديها ارتباط قوي مع السمة المستهدفة والتي يمكن أن تكون مفيدة في التنبؤ بإفلاس الشركة.

وذلك من خلال حساب مصفوفة الارتباط التي تقيس قوة العلاقة بين السمة الهدف وباقي السمات . إن الناتج هو جدول ذو أبعاد ثنائية حيث يحتوي كل عنصر على معامل ارتباط بيرسون(مقياس إحصائي يحدد قوة واتجاه العلاقة الخطية بين متغيرين مستمرين)

و أيضا من خلال تحديد السمات ذات الارتباط العالي ,أكبر أو تساوي 0.1, (هذا يساعد في تحديد السمات التي تكون لها علاقة قوية نسبيًا مع احتمال الإفلاس)

```
Columns with correlation > 0.0 with Bankrupt?:
                                  Bankrupt?
Bankrupt?                          1.000000
 Debt ratio %                      0.250161
 Borrowing dependency              0.176543
 Current Liability to Assets       0.194494
 Current Liabilities/Equity        0.153828
 Total expense/Assets              0.139049
 Current Liability to Equity       0.153828
 Equity to Long-term Liability     0.139014
 Current Liability to Current Assets  0.171306
 Liability-Assets Flag             0.139212
 Liability to Equity               0.166812
```

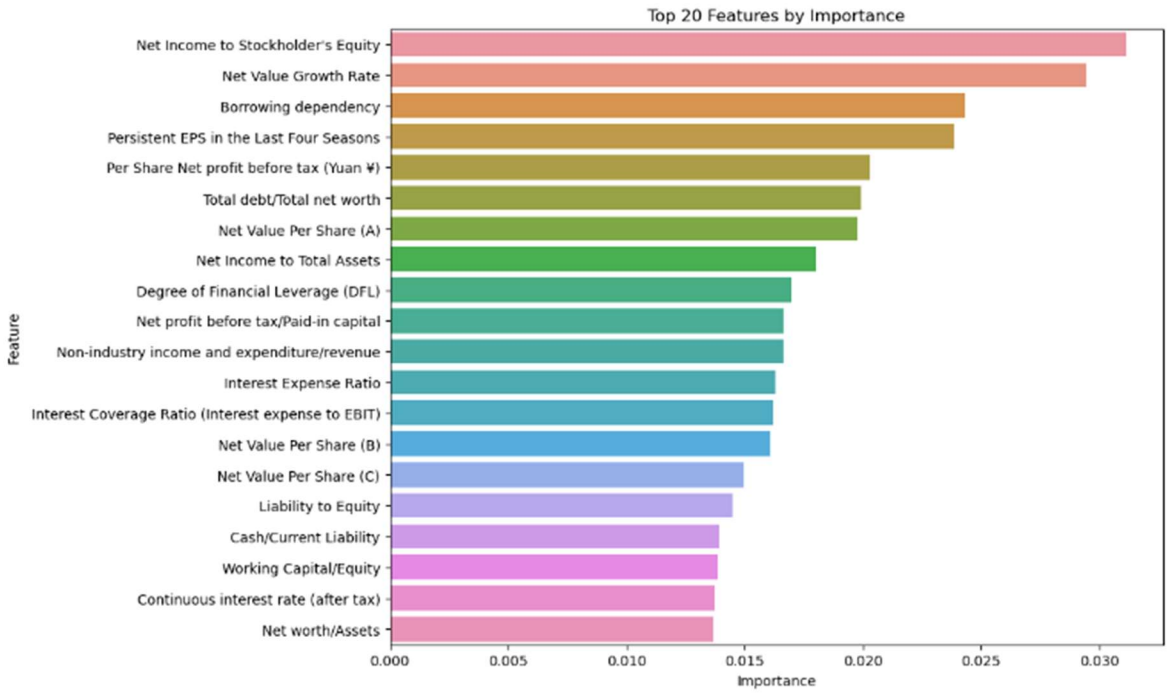


من المخطط السابق حصلنا على قائمة بالسمات التي لها ارتباط أكبر من 0.1 مع السمة الهدف على سبيل المثال، وجدنا أن نسبة الديون لها ارتباط قيمته حوالي 0.25 مع الإفلاس، مما يعني أن هناك علاقة إيجابية قوية نسبيًا بين هاتين السمتين: كلما زادت نسبة الديون، زادت احتمالية الإفلاس.

بالإضافة الى ماسبق قمنا ببناء نموذج Random Forest Classifier لتحديد أهمية (مدى تأثير كل سمة على قرار النموذج.)كل سمة من السمات في معرفة ما إذا كانت الشركة ستفلس أم لا. بعد تدريب النموذج، قمنا باستخراج الأهمية النسبية لكل سمة وقمنا بترتيبها بناءً على تأثيرها في التنبؤ بالإفلاس.

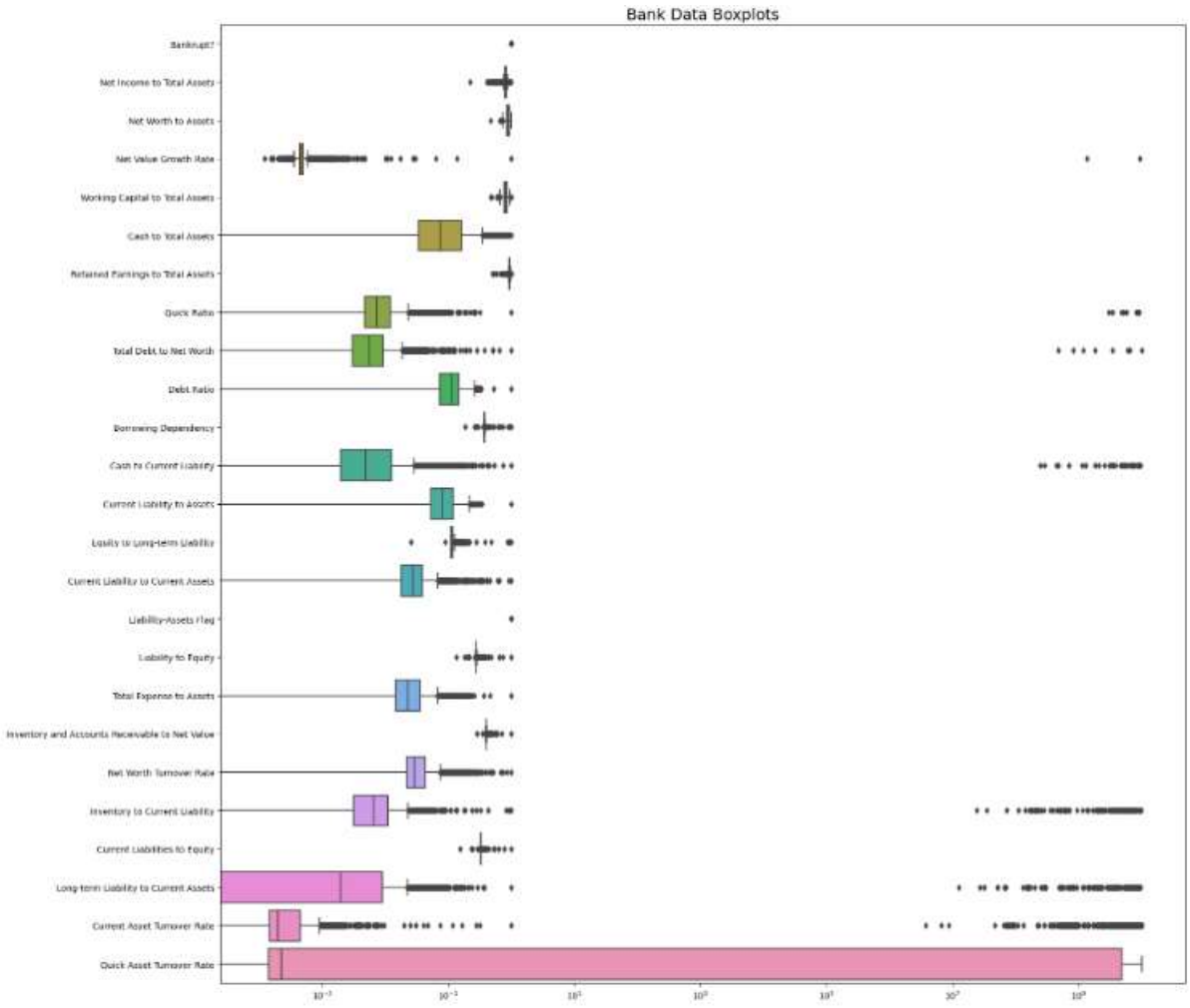
	Feature	Importance
89	Net Income to Stockholder's Equity	0.031123
29	Net Value Growth Rate	0.029454
39	Borrowing dependency	0.024325
18	Persistent EPS in the Last Four Seasons	0.023858
22	Per Share Net profit before tax (Yuan ¥)	0.020301
..	...	...
75	Fixed Assets to Assets	0.005915
31	Cash Reinvestment %	0.005626
14	Tax rate (A)	0.002257
84	Liability-Assets Flag	0.000242
93	Net Income Flag	0.000000

[95 rows x 2 columns]

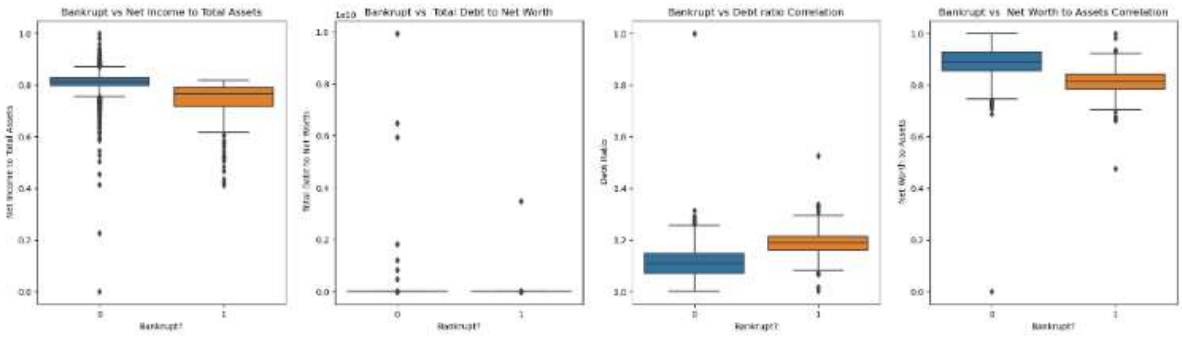


بإجراء هذا التحليل، يمكننا تحديد السمات الأكثر تأثيرًا في قرار النموذج، مما يساعد في تحسين دقة وكفاءة النماذج المستقبلية.

وبعد تطبيق ما سبق اختصرنا عدد السمات من 96 الى 24 سمة فعالة .



المخطط الذي تم عرضه هو Boxplotلعدد من السمات العددية فيBank Data. تمثل الـ Boxplotsتوزيع البيانات لكل سمة، مما يتيح لنا رؤية القيم المركزية والانتشار، وكذلك القيم المتطرفة(outliers)

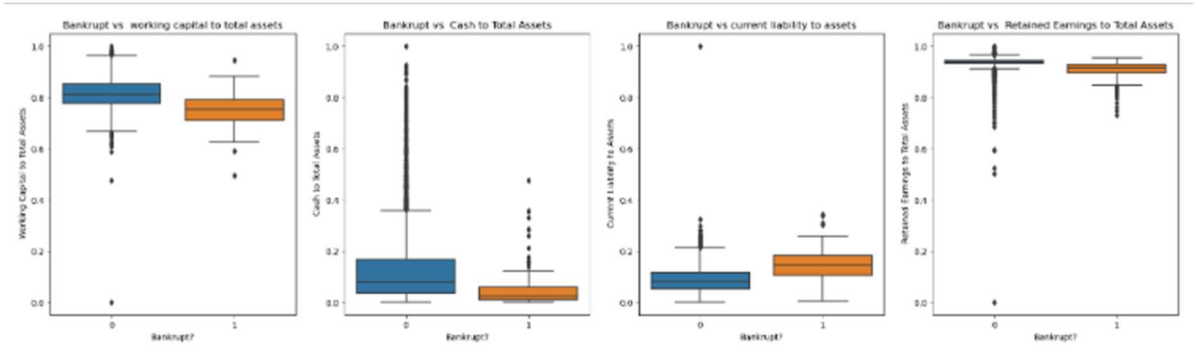


المخطط المعروض يتضمن عدة رسوم بيانية (Boxplots) لكل من السمات المالية الأساسية بالمقارنة مع حالة الإفلاس(Bankrupt?) ، حيث:

- **Bankrupt vs Net Income to Total Assets:**
  - هذا الرسم البياني يعرض العلاقة بين حالة الإفلاس والنسبة بين صافي الدخل إلى إجمالي الأصول.
  - يُظهر الرسم أن الشركات التي لم تفلس (0) تميل إلى أن تكون لديها نسبة أعلى من صافي الدخل إلى إجمالي الأصول مقارنةً بالشركات التي أفلست.(1)
  - ومع ذلك، هناك تداخل كبير بين الفئتين مع وجود بعض القيم المتطرفة.
- **Bankrupt vs Total Debt to Net Worth:**
  - هذا الرسم يظهر العلاقة بين حالة الإفلاس وإجمالي الدين إلى صافي القيمة.
  - يمكن ملاحظة أن الشركات المفلسة (1) غالباً ما تكون لديها نسبة دين أعلى إلى صافي القيمة مقارنةً بالشركات غير المفلسة .(0) ويظهر أيضًا وجود قيم متطرفة عالية في البيانات.
- **Bankrupt vs Debt Ratio Correlation:**
  - الرسم البياني يعرض العلاقة بين حالة الإفلاس ونسبة الدين.

- نلاحظ أن الشركات المفلسة (1) تميل إلى أن تكون لديها نسب ديون أعلى مقارنة بالشركات غير المفلسة.(0)

- **Bankrupt vs Net Worth to Assets Correlation:**
  - هذا الرسم يوضح العلاقة بين حالة الإفلاس وصافي القيمة إلى الأصول.
  - الشركات التي لم تفلس (0) لديها نسبة أعلى من صافي القيمة إلى الأصول مقارنةً بالشركات التي أفلست.(1)



المخطط يحتوي على أربعة رسوم بيانية (Boxplots) تعرض العلاقة بين حالة الإفلاس (Bankrupt?)وبعض المؤشرات المالية المختلفة. إليك شرح كل منها:

- **Bankrupt vs Working Capital to Total Assets:**
  - يوضح العلاقة بين رأس المال العامل إلى إجمالي الأصول وحالة الإفلاس.
  - يُلاحظ أن الشركات التي لم تفلس (0) تمتلك بشكل عام نسبة أعلى من رأس المال العامل إلى إجمالي الأصول مقارنة بالشركات التي أفلست .(1) ومع ذلك، هناك تداخل واضح وقيم متطرفة.

- **Bankrupt vs Cash to Total Assets:**
  - يظهر العلاقة بين نسبة النقد إلى إجمالي الأصول وحالة الإفلاس.

- الشركات التي لم تفلس (0) تظهر نطاقًا أوسع من نسب النقد إلى إجمالي الأصول، بينما الشركات المفلسة (1) لديها نسب أقل في معظمها. يظهر أيضًا العديد من القيم المتطرفة في هذا الرسم.

- **Bankrupt vs Current Liability to Assets:**
  - يوضح العلاقة بين الخصوم الحالية إلى الأصول وحالة الإفلاس.
  - الشركات المفلسة (1) تميل إلى أن تكون لديها نسبة أعلى من الخصوم الحالية إلى الأصول مقارنة بالشركات غير المفلسة .(0) هذا قد يشير إلى أن الشركات التي لديها التزامات قصيرة الأجل أكبر من الأصول لديها احتمالية أكبر للإفلاس.

- **Bankrupt vs Retained Earnings to Total Assets:**
  - يظهر العلاقة بين الأرباح المحتجزة إلى إجمالي الأصول وحالة الإفلاس.
  - يمكن ملاحظة أن الشركات التي لم تفلس (0) تميل إلى أن تكون لديها نسبة أعلى من الأرباح المحتجزة إلى إجمالي الأصول، في حين أن الشركات المفلسة (1) لديها نسب أقل.

نظراً للمخططات السابقة تبيا معنا بأنه يوجد قيم متطرفة (هي بيانات تختلف بشكل كبير عن بقية البيانات ويمكن أن تؤثر سلبًا على التحليل الإحصائي والنمذجة)حيث في النماذج التي تعتمد على المسافة، مثل الانحدار الخطي أو خوارزميات التعلم الآلي مثل-k) k-Nearest Neighbors (NN، يمكن أن تسبب القيم المتطرفة في تحريف النموذج بعيدًا عن الاتجاه العام للبيانات. قد يؤدي هذا إلى نتائج تنبؤية غير دقيقة أو نماذج غير صالحة.

كما أنه قد يميل النموذج إلى التكيف بشكل مفرط مع هذه القيم، مما يؤدي إلى overfitting، حيث يعمل النموذج بشكل جيد على البيانات التي تحتوي على القيم المتطرفة ولكنه يفشل في التعميم على بيانات جديدة.

القيم المتطرفة قد تخفي العلاقات الحقيقية بين المتغيرات أو تسبب في الكشف عن علاقات مزيفة (spurious correlations)، مما يصعب من تفسير نتائج التحليل.

لذلك عملنا على ازالة القيم المتطرفة عن طريق إيجاد Quartile 25 و Quartile 75: اللذان يمثلان الحدود الأدنى (25th percentile) والأعلى (75th percentile) لقيمة معينة. على سبيل المثال، في حالة"Net Income to Total Assets" ، نجد أن:

- Quartile 25: 0.7967
- Quartile 75: 0.8265

وحساب كل من(Interquartile Range) IQR و Cut Off :

IQR (Interquartile Range) هو الفارق بين الربع الأعلى والربع الأدنى. وهو مؤشر على مدى انتشار البيانات في المنتصف.

- مثلاً، في حالة"Net Income to Total Assets" ، الـ IQR هو 0.0297.

Cut Off يتم تحديده بضرب الـ IQR في 1.5. ويستخدم لتحديد الحدود التي بناءً عليها تعتبر القيمة متطرفة.

- مثلاً، في حالة"Net Income to Total Assets" ، الـ Cut Off هو 0.0446.

lower, upper = q25 - feat_cut_off, q75 + feat_cut_off

Lower و Upper Bounds هذه الحدود تحدد القيم التي تقع خارجها القيم المتطرفة.

- مثلاً، في حالة"Net Income to Total Assets" ، الحد الأدنى هو 0.7522 والحد الأعلى هو 0.8710.

ونعرض عدد القيم التي تعتبر متطرفة بالنسبة للمؤشر المدروس. على سبيل المثال:

- في حالة"Net Income to Total Assets" ، هناك 561 قيمة متطرفة.

الخرج لاحد السمات :

Quartile 25: 0.7967498491931705 | Quartile 75: 0.8264545295408715

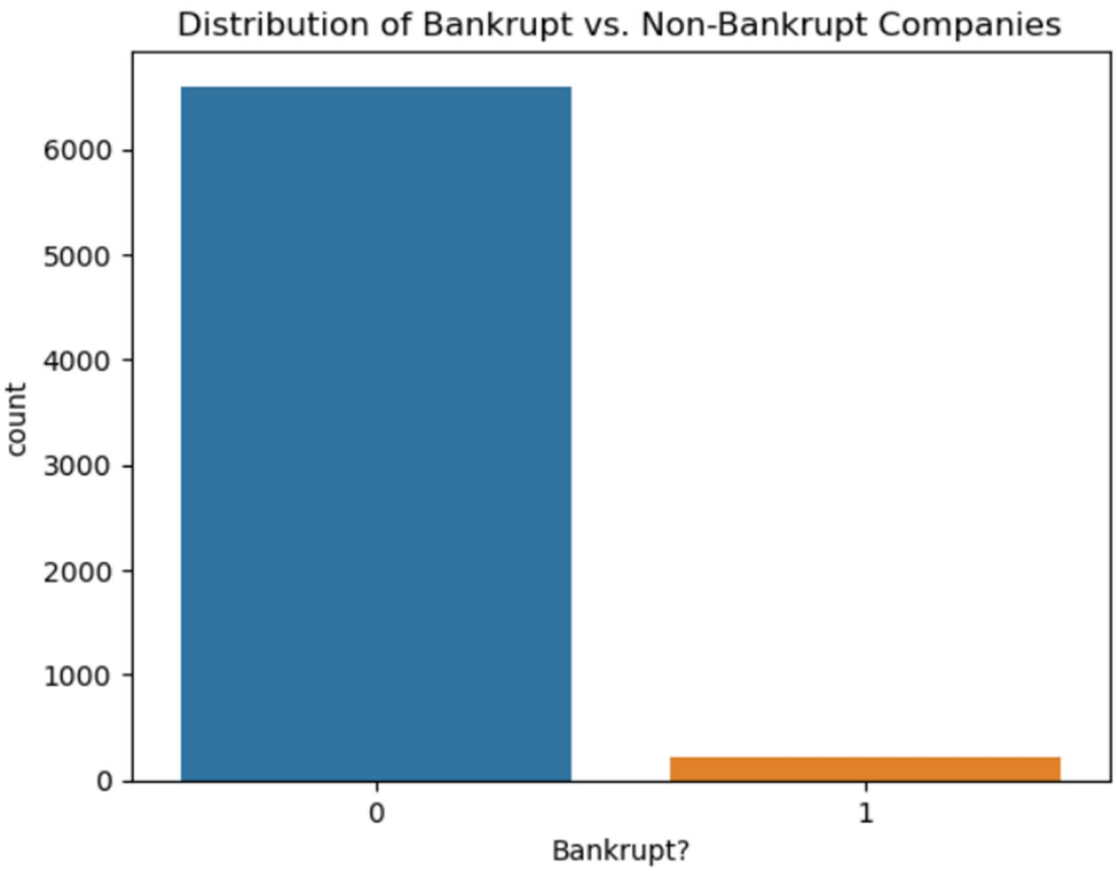
iqr: 0.029704680347701018

Cut Off: 0.04455702052155153

Net Income to Total Assets Lower: 0.752192828671619

Net Income to Total Assets Upper: 0.871011550062423

Net Income to Total Assets outliers for close to bankruptcy cases: 561



نلاحظ أن الداتا سيت غير متوازنة حيث أن نسبة الشركات الغير مفلسة 96.67737% أما نسبة الشركات المفلسة 3.2263%

قمنا بعمل scaling (عن طريق استخدام StandardScaler هو مقياس يتم استخدامه لتحجيم البيانات بحيث يتمركز المتوسط عند 0 والانحراف المعياري يكون 1 لكل ميزة في البيانات.) للبيانات الموجودة في مصفوفة الميزات وذلك لأن العديد من خوارزميات التعلم الآلي تعمل بشكل أفضل عندما تكون الميزات محجمة بشكل صحيح وخصوصا الخوارزميات التي تعتمد على المسافة .

كما أن بعض خوارزميات التحسين (optimization algorithms) تتقارب بشكل أسرع عندما تكون البيانات محجمة.

وقمنا بتقسيم البيانيات الى training and testing sets

الغابة العشوائية (Random Forest) هي خوارزمية تعلم آلي تقوم بدمج عدة أشجار قرار (Decision Trees) لتحسين دقة واستقرار التنبؤات بشكل عام.

كيفية عملها:

1. **(Bootstrapping)** تبدأ الخوارزمية بإنشاء عدة مجموعات فرعية من البيانات الأصلية باستخدام Bootstrapping مع الإعادة (أي يمكن تكرار العينات).
2. **بناء أشجار القرار :**لكل عينة استعيان، يتم بناء شجرة قرار باستخدام مجموعة فرعية عشوائية من الميزات.
3. **التصويت الجماعي :(Ensemble Voting)** عند إجراء التنبؤ، تقوم كل شجرة في الغابة بالإدلاء بصوتها بشأن النتيجة الأكثر احتمالية. في مشاكل التصنيف، يتم اختيار النتيجة التي تحصل على غالبية الأصوات.

المزايا:

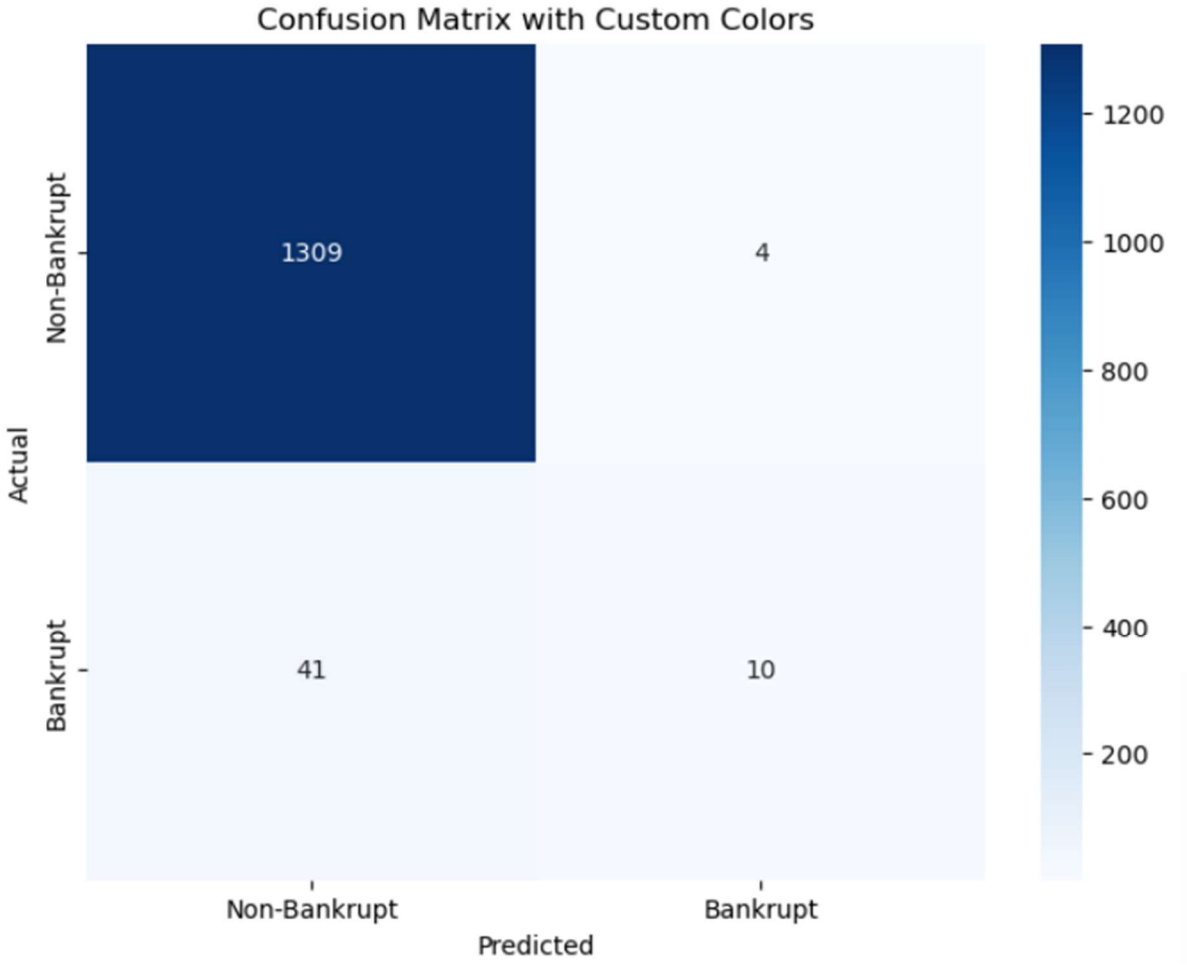
- **دقة عالية :**تقدم الغابة العشوائية دقة أفضل لأنها تقوم بتقليل أخطاء الأشجار الفردية.

- **مقاومة ل overfitting:**العشوائية في اختيار الميزات وبناء الأشجار تقلل من خطر الإفراط في التكيف مع البيانات.

قمنا بتطبيق randomForestClassifier بدون معالجة الداتا الغير متوازنة ولاحظنا النتائج التالية :

```
Accuracy: 0.967008797653959
Recall: 0.1960784313725490 2
F1-score: 0.3076923076923077
```

الدقة وحدها لا تكفي للبيانات غير المتوازنة. يجب أيضًا النظر في مصفوفة الارتباك وتقرير التصنيف مع خط الأساس Recall و F1-score للفئة 1. يجب أيضًا النظر في مصفوفة الارتباك وتقرير التصنيف مع خط الأساس Recall ودرجة F1 للفئة 1.



	precision	recall	f1-score	support
0	0.97	1.00	0.98	1313
1	0.71	0.20	0.31	51
accuracy			0.97	1364
macro avg	0.84	0.60	0.65	1364
weighted avg	0.96	0.97	0.96	1364

الموديل لديه دقة عالية جدًا للفئة 0، ولكنه يعاني من ضعف كبير في اكتشاف الفئة 1

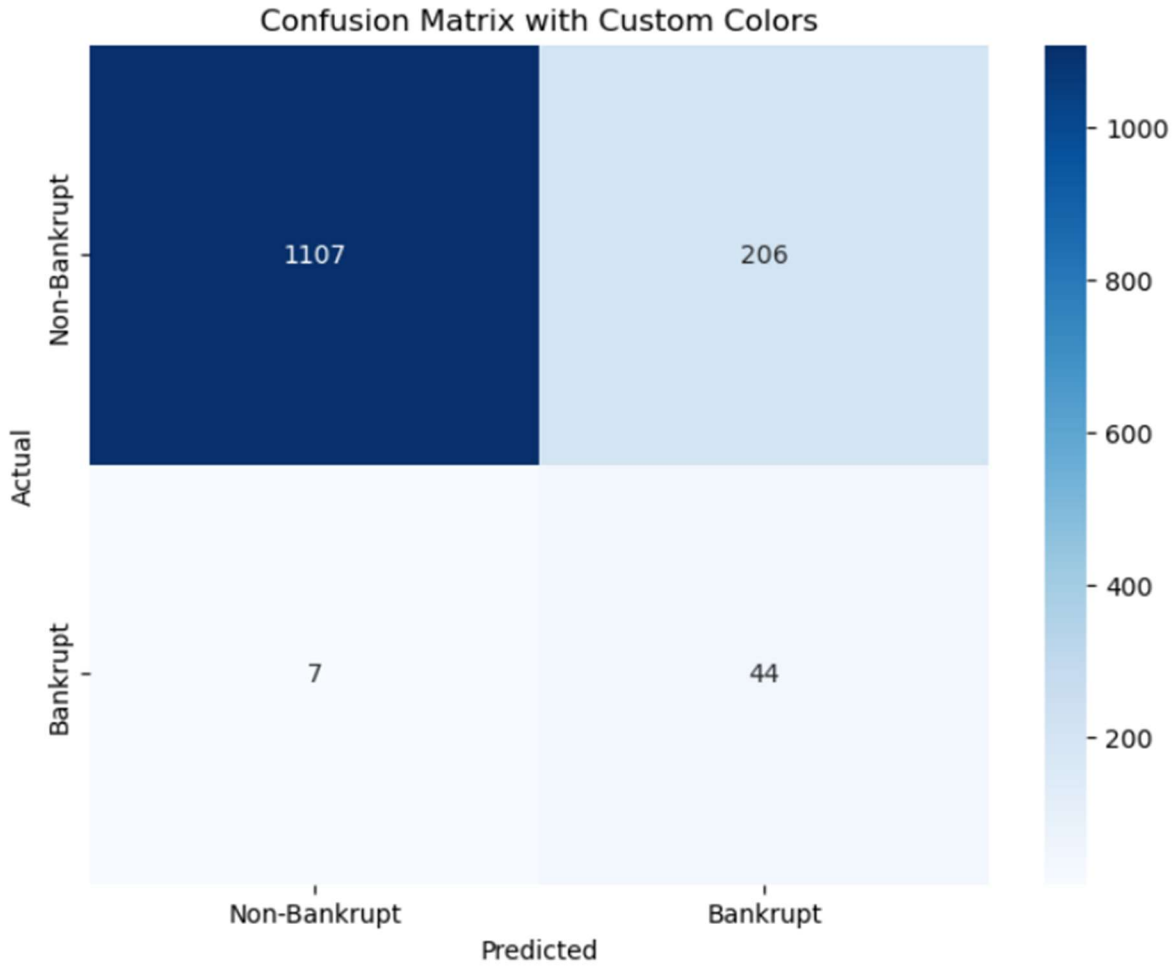
(recall للفئة 1 ضعيف جدًا)

لنقوم بتطبيق **Undersampling**

يتضمن تقليل عدد الحالات في فئة الأغلبية لموازنة توزيع الفئة. يتم ذلك عن طريق إزالة عينات عشوائيًا من فئة الأغلبية بحيث يكون حجمها أقرب إلى حجم فئة الأقلية.

بعد تطبيقها اصبح لدينا 338 مثال من أصل 6819 مثال

```
Undersampling Accuracy: 0.843841642228739
Undersampling Recall: 0.8627450980392157
Undersampling F1-score: 0.29235880398671094
```



	precision	recall	f1-score	support
0	0.99	0.84	0.91	1313
1	0.18	0.86	0.29	51
accuracy			0.84	1364
macro avg	0.58	0.85	0.60	1364
weighted avg	0.96	0.84	0.89	1364

في هذه التقنية، يحقق Recall معدل أعلى بكثير مع المفاضلة مع Precision ، مما يؤدي إلى درجة F1 أقل من النتائج السابقة للفئة 1. هذا يعني أن المستثمرين سيكونون آمنين للغاية إذا استندوا إلى هذه النسخة من التنبؤ، ولكن سيفقدون فرصة الاستثمار في العديد من الشركات التي تعمل بشكل جيد ولكن تم التنبؤ بها على أنها تتعامل مع الإفلاس. كما تنخفض الدقة الإجمالية بشكل كبير مقارنة بالنموذج الافتراضي.

تحسن كبير في قدرة الموديل على اكتشاف الفئة 1 (recall مرتفع جدا)

ولكن على حساب انخفاض accuracy والـ precision للفئة 1.

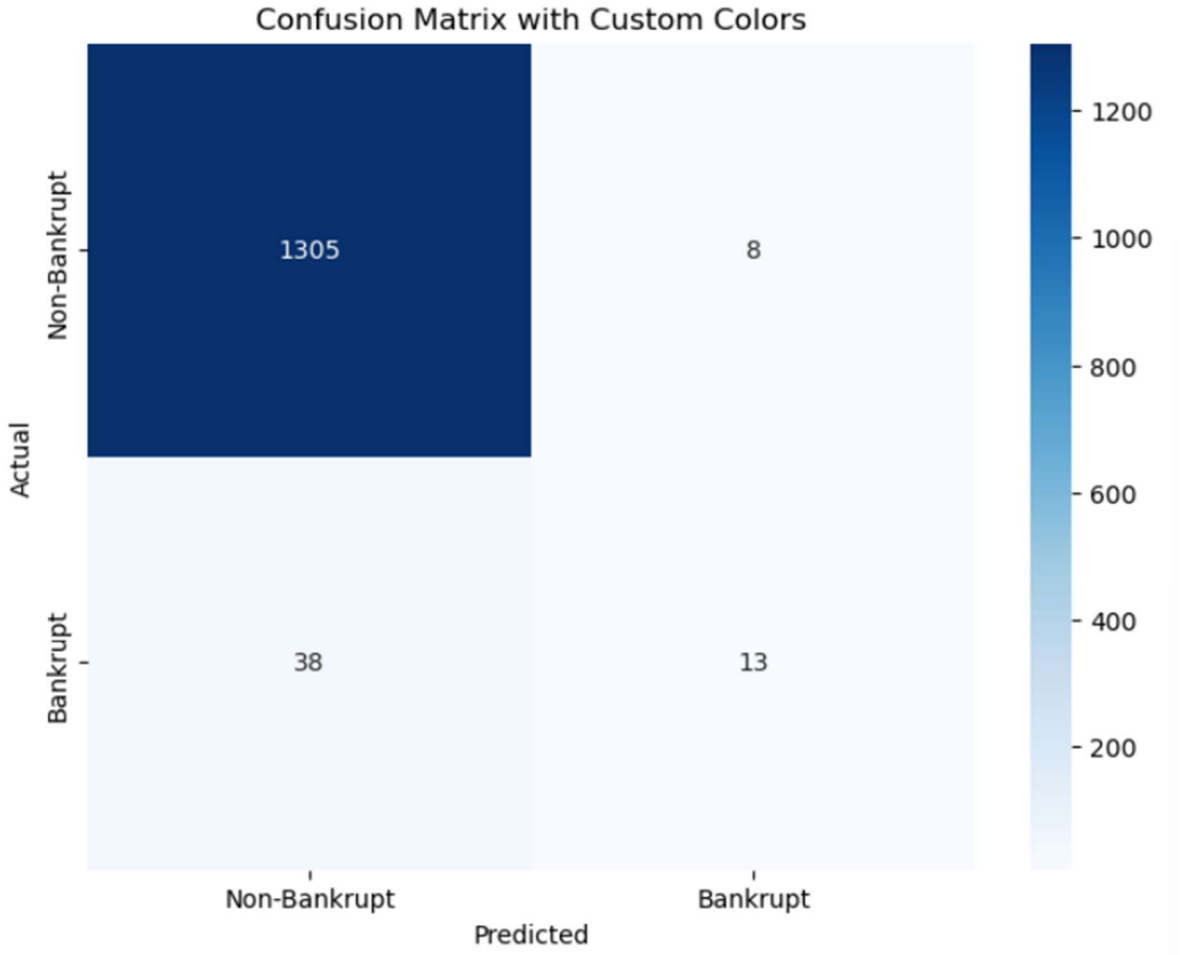
لا يُعد تطبيق **Undersampling** على مجموعة البيانات غير المتوازنة ونموذج الغابة العشوائية في هذه المهمة حلاً جيدًا.

لنقوم بتطبيق **Oversampling:**

يعد أخذ العينات الزائدة أسلوبًا آخر يستخدم لمعالجة اختلال توازن الفئة، ويتضمن زيادة عدد المثيلات في فئة الأقلية لموازنة توزيع الفئة.

بعد تطبيقها أصبح لدينا 10572 مثال أي أكثر ب 3,753 مثال من عدد الأمثلة في ال داتا سيت

```
Oversampling Accuracy: 0.966275659824047
Oversampling Recall: 0.2549019607843137
Oversampling F1-score: 0.3611111111111111
```



	precision	recall	f1-score	support
0	0.97	0.99	0.98	1313
1	0.62	0.25	0.36	51
accuracy			0.97	1364
macro avg	0.80	0.62	0.67	1364
weighted avg	0.96	0.97	0.96	1364

هذه التقنية على زيادات في كل من درجة الاستدعاء ودرجة F1 للفئة 1، وهو الهدف الذي كان يتم البحث عنه ,توازن أفضل بين precision والـ recall للفئة 1 مقارنة بالـundersampling ، مع الحفاظ على دقة عالية للفئة 0.

يعتبر أخذ العينات الزائدة(**Oversampling**) للغابة العشوائية هو الإصدار الأفضل مقارنة بالنموذج بدون معالجة البيانات والنموذج مع **Undersampling**

لنقوم بتطبيق **SMOTE**:

SMOTE (Synthetic Minority Over-sampling Technique)

هي تقنية تُستخدم لمعالجة اختلال التوازن الطبقي في مجموعات البيانات. فهي تُنشئ عينات اصطناعية لفئة الأقلية

كيف تعمل SMOTE:

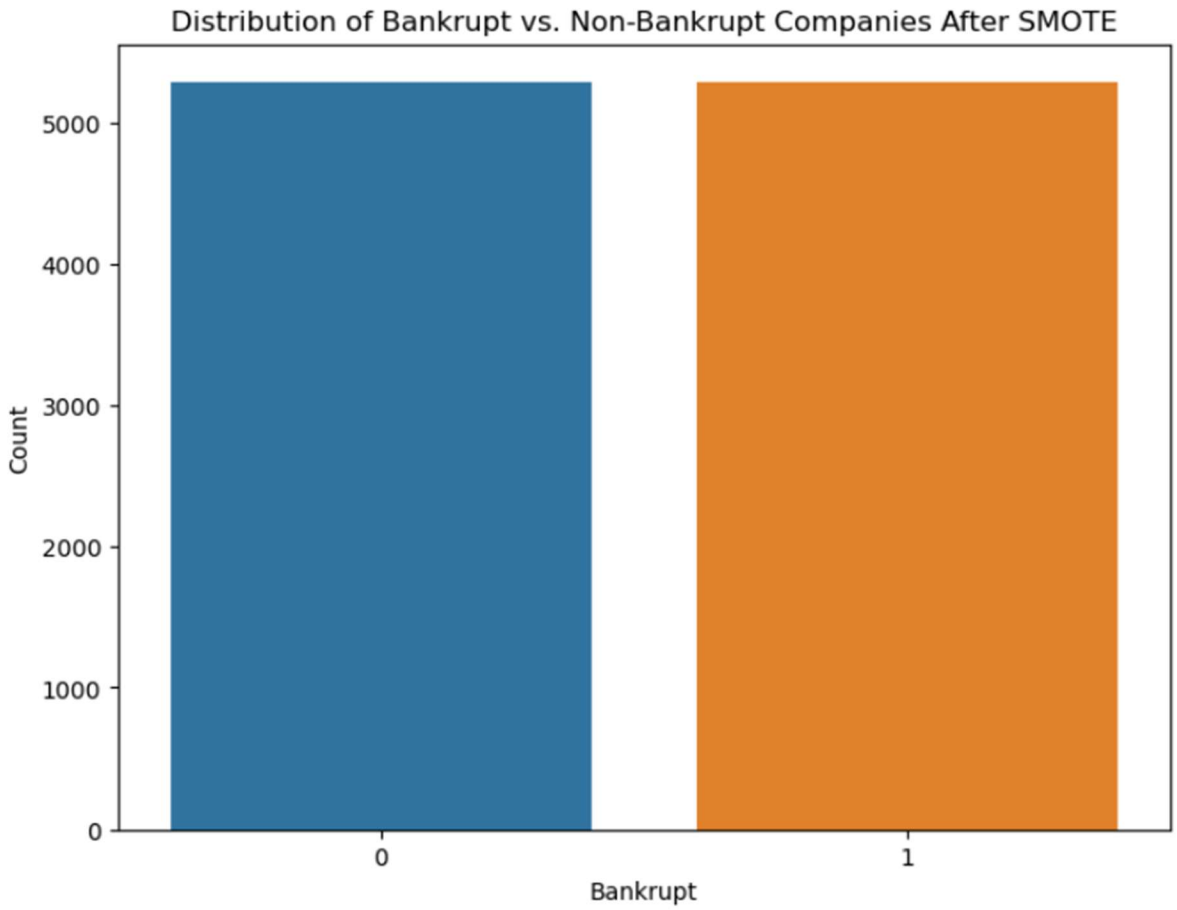
يحدد SMOTE أولاً العينات التي تنتمي إلى فئة الأقلية في مجموعة البيانات.

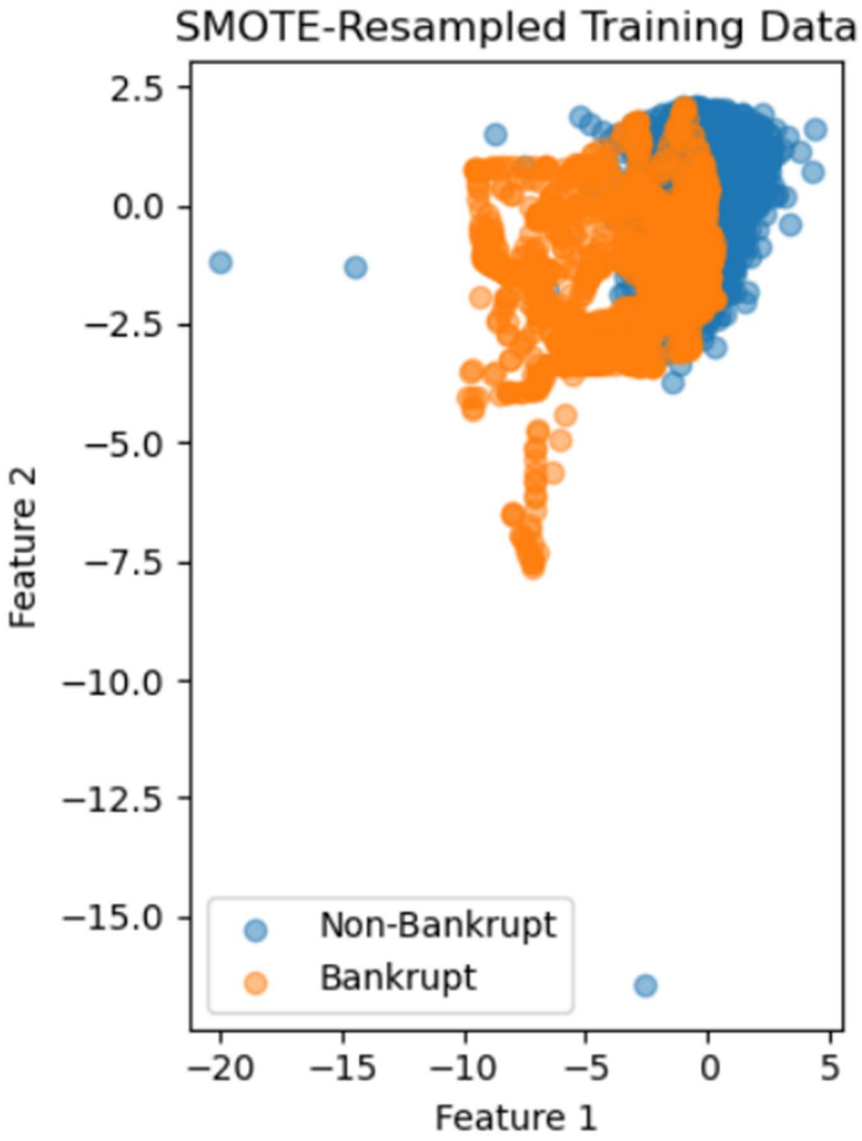
و بالنسبة لكل عينة من فئة الأقلية، تحدد SMOTE أقرب جيرانها من نفس فئة الأقلية. عادةً ما يحدد المستخدم قيمة k، وعادةً ما تكون 5.

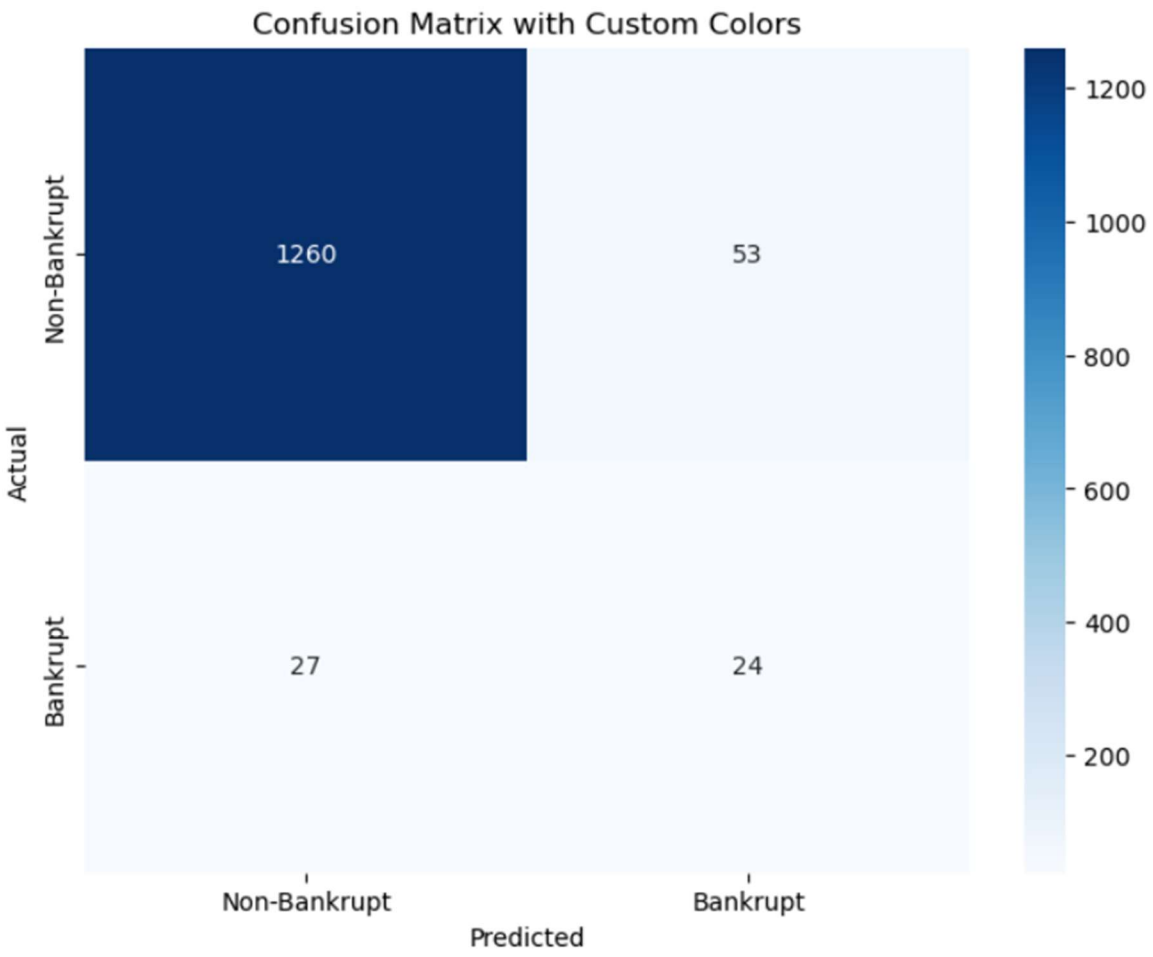
تختار SMOTE عشوائيًا أحد الجيران الأقرب k وتولّد عينة اصطناعية جديدة عن طريق الاستيفاء بين عينة فئة الأقلية وجارتها. يتم هذا الاستيفاء عن طريق أخذ متوسط مرجح بين العينتين، مما يؤدي فعليًا إلى إنشاء نقطة على طول القطعة المستقيمة التي تربط بين العينتين في فضاء الميزة.

وبعد ذلك تُضاف العينات الاصطناعية إلى مجموعة البيانات، مما يزيد من عدد مثيلات فئة الأقلية وبالتالي يقلل من اختلال التوازن بين الفئتين.

بعد تطبيقها أصبح لدينا 10572 مثال أي أكثر ب 3,753 مثال من عدد الأمثلة في ال داتا سيت


Distribution of Bankrupt vs. Non-Bankrupt Companies After SMOTE
Count
5000
4000
3000
2000
1000
0
0
1
Bankrupt


SMOTE-Resampled Training Data
Feature 2
Feature 1
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−12.5
−15.0
−20
−15
−10
−5
0
5
Non-Bankrupt
Bankrupt



	precision	recall	f1-score	support
0	0.98	0.96	0.97	1313
1	0.31	0.47	0.37	51
accuracy			0.94	1364
macro avg	0.65	0.72	0.67	1364
weighted avg	0.95	0.94	0.95	1364

: SMOTE

يوفر توازنًا معقولًا بين precision و recall للفئة 1، ويعتبر خيارًا جيدًا من أجل تحسين القدرة على اكتشاف الفئة 1 مع الحفاظ على مستوى معقول من الدقة.

تحصل هذه التقنية على زيادة في كل من درجةRecall ودرجة F1 للفئة 1،

وهو الهدف الذي كان يتم البحث عنه. وهذا يساعد على تقليل مخاطر الاستثمار الخاطئ.

نستنتج أن SMOTE للغابة العشوائية هو الإصدار الأفضل من بين النماذج المنفذة.

الأن سوف نطبق كل من هذه الخوارزميات **Gradient , Logistic Regression Boosting**

Random Forest , Decision Tree, Bagging,Multilayer Perceptron (MLP)

ونحصل على أفضل خوارزمية من خلال مقارنة النتائج :

Logistic Regression

(Logistic Regression) هي خوارزمية تصنيف تُستخدم للتنبؤ بالنتائج التي تأخذ قيمتين فقط (ثنائية)، مثل "نعم" أو "لا"، "صح" أو "خطأ". وتستند إلى استخدام دالة لوجيستية، التي تُعرف أيضًا باسم دالة سيجمويد، لتحويل التنبؤات الخطية إلى احتمالات.

كيفية العمل:

- تبدأ الخوارزمية بحساب التنبؤ الخطي بناءً على المتغيرات المستقلة.(features)
- ثم يتم تمرير هذا التنبؤ عبر دالة سيجمويد لتحويله إلى قيمة احتمالية بين 0 و 1.

- إذا كانت الاحتمالية أكبر من العتبة (عادة 0.5)، يتم تصنيف العينة في الفئة 1، وإلا في الفئة 0

Gradient Boosting

Gradient Boostingهي تقنية تجميع تهدف إلى تحسين أداء نموذج ضعيف عبر بناء سلسلة من النماذج، حيث يحاول كل نموذج جديد تقليل الأخطاء التي ارتكبها النموذج السابق. تُستخدم عادةً في مشاكل التصنيف والتنبؤ.

كيفية العمل:

- يتم تدريب نموذج ضعيف (عادةً شجرة قرار صغيرة) على البيانات.
- يتم حساب الأخطاء المتبقية (residuals) بين التوقعات والنتائج الفعلية.
- يتم تدريب نموذج جديد على هذه الأخطاء المتبقية بهدف تقليلها.
- جمع النماذج المتتابعة، مع التركيز على الأخطاء في كل مرحلة، للحصول على نموذج نهائي أكثر دقة وأداءً.

Multilayer Perceptron (MLP)

MLPهو نوع من الشبكات العصبية الاصطناعية يتكون من طبقات متعددة من الخلايا العصبية، تشمل طبقة إدخال، عدة طبقات مخفية، وطبقة إخراج. يُستخدم في مهام مثل التصنيف والتنبؤ، حيث يتيح تعدد الطبقات قدرة الشبكة على اكتشاف الأنماط المعقدة في البيانات.

كيفية العمل:

- يتم تقديم البيانات المدخلة إلى طبقة الإدخال.
- تمر البيانات عبر الطبقات المخفية، حيث تخضع للعمليات الرياضية مثل الضرب بالمصفوفات وتطبيق وظائف التنشيط (Activation Functions) لتحويل المدخلات إلى إشارات.

- تصل البيانات إلى طبقة الإخراج التي تنتج التنبؤ النهائي. يعتمد نوع وظيفة التنشيط في طبقة الإخراج على نوع المهمة (تصنيف أو انحدار).

Bagging

Bagging (Bootstrap Aggregating)هي طريقة تجميع تهدف إلى تقليل التذبذب وزيادة استقرار النماذج التنبؤية، من خلال إنشاء عدة نماذج موازية (غالبًا أشجار قرار) وتدريب كل منها على عينة مختلفة من البيانات,بعد ذلك، يتم تجميع نتائج هذه النماذج للحصول على تنبؤ أكثر دقة وأقل تذبذب

كيفية العمل:

- يتم إنشاء عدة عينات من البيانات عن طريق اختيار عينات عشوائية مع الاستبدال (bootstrap samples). على سبيل المثال، إذا كان لدينا 1000 نقطة بيانات في مجموعة البيانات الأصلية، فقد نختار 1000 نقطة بيانات جديدة بطريقة عشوائية، حيث يمكن أن تتكرر بعض النقاط في العينة النهائية، بينما قد يتم استبعاد نقاط أخرى
- يتم تدريب نموذج مستقل (مثل شجرة قرار) على كل عينة: عند إنشاء عينات متعددة من البيانات باستخدام البوتستراب، يتم تدريب نموذج مستقل (مثل شجرة قرار) على كل عينة من هذه العينات. وبما أن كل عينة مختلفة قليلًا عن الأخرى، فإن كل نموذج يتعلم شيئًا مختلفًا عن البيانات، مما يزيد من تنوع النماذج
- يتم تجميع النتائج النهائية عن طريق أخذ متوسط التنبؤات أو التصويت للأغلبية بين النماذج المختلفة، مما يحسن من الدقة ويقلل من احتمال الخطأ.

Decision Tree

شجرة القرار هي نموذج يعتمد على تقسيم البيانات بشكل متكرر إلى مجموعات فرعية بناءً على ميزات محددة، حتى يتم الوصول إلى التنبؤ النهائي في الأوراق .(Leaves) تستخدم شجرة القرار في كل من التصنيف والانحدار.

كيفية العمل:

- تبدأ الشجرة بعقدة جذر تحتوي على جميع البيانات.
- يتم اختيار الميزة الأكثر أهمية في البيانات وتقسيم البيانات إلى مجموعات فرعية استنادًا إلى هذه الميزة.
- يتم تكرار عملية التقسيم لكل مجموعة فرعية حتى تصل إلى عقد نهائية (أوراق) تكون فيها جميع البيانات في المجموعة متجانسة أو تصل إلى شرط توقف معين.
- تمثل الأوراق النهائية التنبؤات بناءً على شروط التقسيم التي تم الوصول إليها.

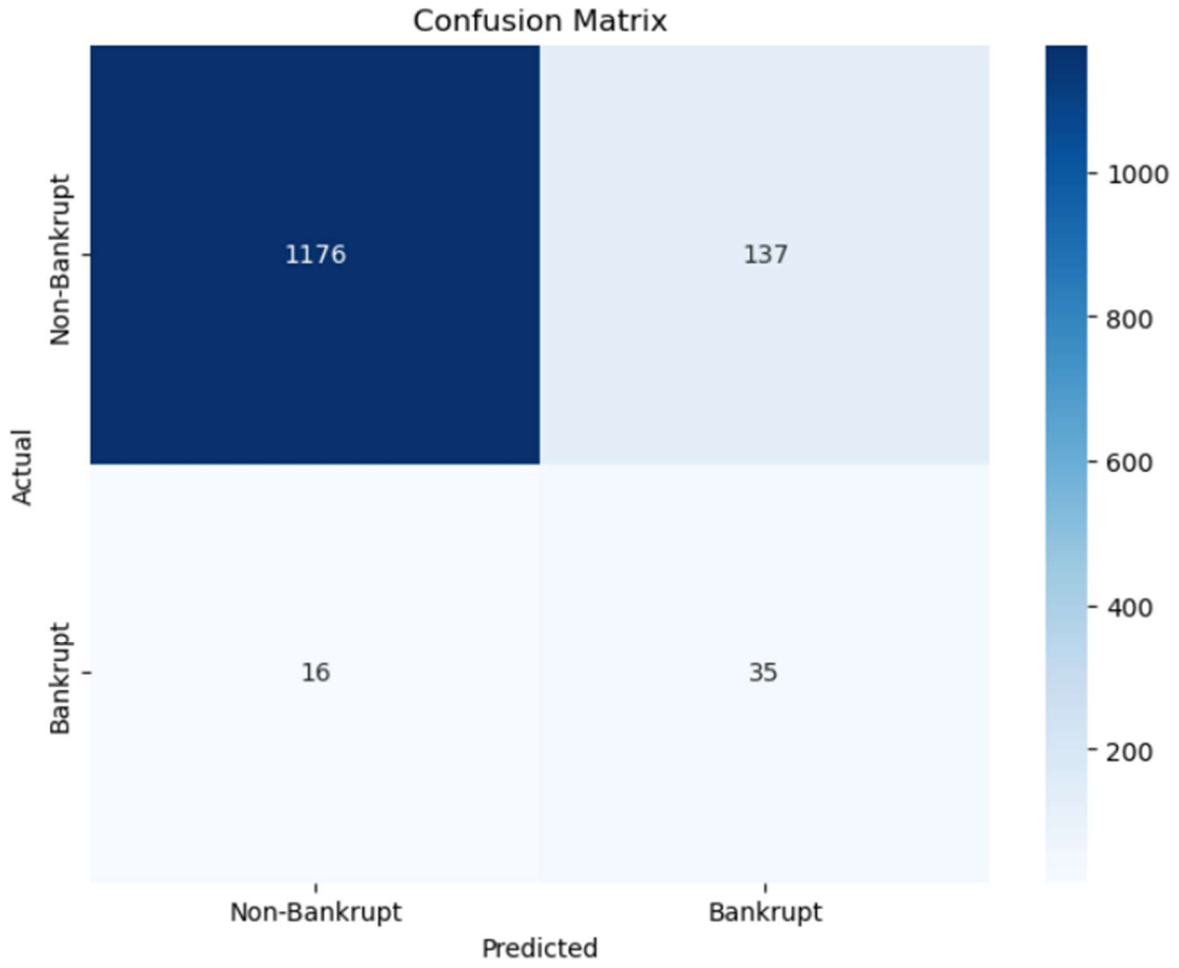
```
Logistic Regression Best Threshold: 0.8299999999999996
Logistic Regression Accuracy: 0.9451879010082493
Logistic Regression Precision: 0.3006134969325153
Logistic Regression Recall: 0.5798816568047337
Logistic Regression F1 Score: 0.39595959595959596
Random Forest Best Threshold: 0.46999999999999986
Random Forest Accuracy: 0.9424381301558203
Random Forest Precision: 0.29461756373937675
Random Forest Recall: 0.6153846153846154
Random Forest F1 Score: 0.3984674329501916
Gradient Boosting Best Threshold: 0.8899999999999996
Gradient Boosting Accuracy: 0.9615032080659945
Gradient Boosting Precision: 0.39267015706806285
Gradient Boosting Recall: 0.4437869822485207
Gradient Boosting F1 Score: 0.41666666666666663
MLP Best Threshold: 0.7199999999999996
MLP Accuracy: 0.9510540788267644
MLP Precision: 0.2850877192982456
MLP Recall: 0.38461538461538464
MLP F1 Score: 0.327455919395466
Bagging Best Threshold: 0.6099999999999998
Bagging Accuracy: 0.9514207149404217
Bagging Precision: 0.2966101694915254
Bagging Recall: 0.41420118343195267
Bagging F1 Score: 0.345679012345679
Decision Tree Best Threshold: 0.1
Decision Tree Accuracy: 0.9231897341888176
Decision Tree Precision: 0.18112244897959184
Decision Tree Recall: 0.42011834319526625
Decision Tree F1 Score: 0.2531194295900178
```

	Model	Best Threshold	Accuracy	Precision	Recall	F1 Score
0	Logistic Regression	0.83	0.945188	0.300613	0.579882	0.395960
1	Random Forest	0.47	0.942438	0.294618	0.615385	0.398467
2	Gradient Boosting	0.89	0.961503	0.392670	0.443787	0.416667
3	MLP	0.72	0.951054	0.285088	0.384615	0.327456
4	Bagging	0.61	0.951421	0.296610	0.414201	0.345679
5	Decision Tree	0.10	0.923190	0.181122	0.420118	0.253119

أفضل نموذج هو Gradient Boosting مع F1 Score: 0.41666666666666663 وعتبة القرار المثلى:
0.8899999999999996
تم حفظ النموذج الأفضل باسم best_model_with_threshold.pkl
تم حفظ نتائج المقارنة في model_comparison_results.csv

Accuracy: 0.89
ROC-AUC: 0.91
Classification Report:

	precision	recall	f1-score	support
0	0.99	0.90	0.94	1313
1	0.20	0.69	0.31	51
accuracy			0.89	1364
macro avg	0.60	0.79	0.63	1364
weighted avg	0.96	0.89	0.92	1364



لقد قمنا بتنفيذ التحقق المتبادل باستخدام StratifiedKFold من أجل:

- معالجة عدم توازن الفئة: تحافظ StratifiedKFold على توزيع الفئة في كل طية، مما يضمن أن تمثل جميع الطيات مجموعة البيانات الإجمالية، وهو أمر بالغ الأهمية للبيانات غير المتوازنة.

- تقليل التباين: من خلال حساب متوسط أداء النموذج عبر طيات متعددة، يوفر التحقق المتبادل تقديرًا أكثر موثوقية واستقرارًا مقارنة بتقسيم التدريب والاختبار الفردي.

- تخفيف الإفراط في التجهيز: يساعد التدريب والاختبار على مجموعات فرعية مختلفة من البيانات في تحديد النماذج التي تعمم بشكل جيد، مما يقلل من خطر الإفراط في التجهيز.

كما أنه طبقنا SMOTE في كل طية من طيات التحقق المتبادل من أجل:

- بيانات تدريب متوازنة: يتم تطبيق SMOTE في كل طية لضمان توزيع متوازن للفئة في مجموعة التدريب، وهو أمر مهم لتدريب النموذج المتسق.

- منع تسرب البيانات: يؤدي تطبيق SMOTE في كل طية إلى تجنب تسرب البيانات، مما يضمن عدم ظهور الأمثلة الاصطناعية في مجموعة الاختبار، والحفاظ على سلامة تقديرات الأداء.

- الاتساق عبر الطيات: يضمن هذا النهج تقييم كل نموذج على بيانات تمت معالجتها باستمرار، مما يؤدي إلى مقاييس أداء أكثر موثوقية وقابلية للمقارنة.

طبقنا أيضا مبدأ **threshold** لانه عادة عند بناء نموذج تصنيف ثنائي يوجد عتبة (0.5 افتراضيا)يتم من خلالها التصنيف أي أنه احتمال أكبر من أو يساوي 0.5 يتم تصنيفه على أنه الفئة الإيجابية، وأي شيء أقل من 0.5 يتم تصنيفه على أنه الفئة السلبية

في الحالات التي تكون فيها البيانات غير متوازنة، قد لا تكون هذه العتبة الافتراضية هي الأفضل. قد يكون من الأفضل تحريك العتبة لأعلى أو لأسفل لتحقيق توازن أفضل بين المقاييس المختلفة مثل Precision و Recall و F1 Score

العتبة لدينا تأخذ مجالاً من 0.1 إلى 0.9، حيث نقوم بتجريب قيم متعددة ضمن هذا النطاق لاختيار العتبة التي تعطي أفضل .**F1 Score**

وقمنا بحساب predicted_probs التي بدورها تسمح لنا بتجريب عتبات مختلفة لتحديد النقطة المثلى لتصنيف العينات كفئة إيجابية (1) أو سلبية.(0)

بدلاً من الاعتماد على العتبة الافتراضية (0.5) التي قد لا تكون مثلى في حالة البيانات غير المتوازنة،

فعند استخدام cross_val_predictمع predict_proba، فإن النموذج لا يقوم فقط بتصنيف العينات على أنها تنتمي إلى الفئة 0 أو 1. بل يقوم بإرجاع **احتمالية** انتماء كل عينة إلى كل فئة. باستخدامه يمكننا العثور على العتبة التي تحقق أفضل **F1 Score**، وهو مقياس يجمع بين كل من PrecisionوRecall ، وهو مهم جداً في الحالات غير المتوازنة.

تحليل مشاعر التقارير المالية:

مقدمة :

يهدف إلى تحليل المشاعر المالية للنصوص من خلال استخدام أساليب تحليل تعتمد على قاموس Loughran-McDonald و نموذج تعلم عميق مدرّب مسبقًا يسمى FinBERT. يتم استخدام هذه الأدوات لتقديم تحليل شامل ودقيق للمشاعر المالية بناءً على محتوى النصوص.

Loughran-McDonald Master Dictionary :

هو قاموس متخصص يستخدم بشكل رئيسي في مجال التمويل وتحليل البيانات المالية. تم تطويره من قبل جيمس لوغرن وبيتر مكدونالد، ويهدف إلى تقديم مجموعة دقيقة ومنظمة من المصطلحات المالية والإدارية.

يحتوي القاموس على مجموعة من المصطلحات التي تُستخدم في النصوص المالية والإفصاحات والتقارير السنوية، ويعزز من دقة وتحليل المعلومات المالية من خلال تحديد وتحليل استخدام المصطلحات في النصوص المالية.

FinBert:

نموذج تعلم عميق مدرب مسبقًا على مجموعة كبيرة من البيانات المالية، مصمم لتحليل المشاعر في النصوص المالية. يتم تحميل النموذج مع المعالج النصي الخاص به لتحليل النصوص باستخدام شبكة عصبية عميقة

مكونات التحليل :

1. تحميل البيانات والمعاجم

قائمة الكلمات المستبعدة (Stopwords): يتم تحميل قائمة بالكلمات الشائعة غير المهمة مثل "و"، "إلى"، "في"، والتي لا تساهم بشكل فعّال في تحليل المشاعر. يتم استبعاد هذه الكلمات من النص المدخل لتحسين دقة التحليل.

قاموس Loughran-McDonald: يتم تحميل هذا القاموس، وهو معجم متخصص يحتوي على كلمات ذات طابع مالي، مصنفة إلى إيجابية وسلبية. يتم استخدام هذا القاموس لقياس المشاعر المالية للنصوص بناءً على نسبة الكلمات الإيجابية والسلبية.

2. تحميل نموذج FinBERT

FinBERT: نموذج تعلم عميق مدرب مسبقًا على مجموعة كبيرة من البيانات المالية، مصمم لتحليل المشاعر في النصوص المالية. يتم تحميل النموذج مع المعالج النصي الخاص به لتحليل النصوص باستخدام شبكة عصبية عميقة.

معالجة النصوص

1. التحضير المسبق للنصوص

يتم تحويل النصوص إلى حروف صغيرة لزيادة التجانس بين الكلمات.

يتم إزالة الكلمات المستبعدة والرموز والنقاط والأرقام لتحسين جودة النصوص المدخلة.

تحليل المشاعر

1. تحليل المشاعر باستخدام قاموس Loughran-McDonald

بعد تحضير النص، يتم حساب عدد الكلمات الإيجابية والسلبية الموجودة في النص وفقًا لقاموس Loughran-McDonald.

يتم حساب نسبتين:

الأولى تعتمد على الفرق بين الكلمات الإيجابية والسلبية مقسومًا على إجمالي عدد الكلمات.

الثانية تعتمد على الفرق بين الكلمات الإيجابية والسلبية مقسومًا على مجموع الكلمات الإيجابية والسلبية فقط.

بناءً على هذه النسب، يتم تصنيف المشاعر إلى إيجابية، سلبية أو محايدة.

2. تحليل المشاعر باستخدام FinBERT

يتم استخدام النموذج لتحليل النصوص المدخلة وتوليد احتمالات للمشاعر (إيجابية، سلبية، محايدة). بناءً على هذه الاحتمالات، يتم تصنيف النص تحت الفئة ذات الاحتمالية الأكبر.

يوفر هذا النظام تحليلًا شاملاً للمشاعر المالية من خلال استخدام أساليب تعتمد على القواميس المتخصصة مع تقنيات التعلم العميق، مما يتيح الحصول على نتائج دقيقة وموثوقة يمكن استخدامها في اتخاذ قرارات مالية أو استثمارية.

المكتبات والتقنيات المستخدمة :

يعتمد على عدد من المكتبات المتخصصة التي تساهم في تحليل النصوص واستخراج المعلومات المالية:

Transformers : تُستخدم مكتبة Transformers لتحميل واستخدام نماذج التعلم العميق مثل FinBERT. هذه المكتبة تسهل التعامل مع النماذج المدربة مسبقًا على مهام محددة مثل تصنيف النصوص أو استخراج المعلومات.

Scipy وPyTorch: تُستخدم هاتان المكتبتان لمعالجة مخرجات النماذج العميقة، حيث يتم تطبيق دالة softmax لتحويل المخرجات إلى احتمالات يمكن تفسيرها في سياق تحليل المعنويات.

Pandas وNumpy: تُستخدم هاتان المكتبتان لتحليل البيانات المالية المُدخلة ولتحليل معجم Loughran-McDonald. كما تُستخدم Pandas لإدارة البيانات وتحليلها بفعالية.

Matplotlib: تُستخدم لإنشاء الرسوم البيانية التي تساعد في توضيح النتائج وتحليل البيانات بصريًا.

PyPDF2: تُستخدم هذه المكتبة لاستخراج النصوص من ملفات PDF، مما يتيح تحليل المستندات المالية مباشرة.

النتائج :

المشاعر العامة **إيجابية**، مما يشير إلى نظرة مالية إيجابية

المشاعر العامة **سلبية**، مما يشير إلى مخاوف محتملة

المشاعر العامة **محايدة**، مما يشير إلى نظرة مالية مستقرة

التطويرات والتحديثات المستقبلية:

في إطار السعي لتحسين النظام وتوسيعه ليشمل المزيد من الأسواق المالية وتقديم تحليلات أكثر دقة وعمقاً، نخطط لإدخال عدة تحسينات مستقبلية تشمل:

1. **دراسة سوق دمشق للأوراق المالية :**سيتم توسيع نطاق المشروع ليشمل تحليل تقارير الشركات المدرجة في سوق دمشق للأوراق المالية. سيشمل ذلك تكامل بيانات السوق المحلي مع النماذج الإحصائية المستخدمة حالياً، مما يساعد في توقع الأداء المالي للشركات السورية بدقة أكبر.

2. **استخدام نماذج إحصائية تعتمد على بيانات سلاسل زمنية :**سيتم تطوير نماذج إحصائية تستند إلى بيانات سلاسل زمنية متعددة السنوات، مما يسمح بتحليل التغيرات والتوجهات على المدى الطويل. هذه النماذج ستكون قادرة على تحسين دقة التوقعات الخاصة بالإفلاس والأداء المالي المستقبلي.

3. **تحليل المشاعر باستخدام تقنيات الذكاء الاصطناعي المتقدمة :**سيتم تعزيز قدرة النظام على تحليل مشاعر النصوص المالية باستخدام تقنيات الذكاء الاصطناعي المتقدمة مثل التحليل الدلالي العميق. هذا سيمكن من تقييم أدق للهجة التقارير المالية وتأثيرها على قرارات المستثمرين.

4. **التكامل مع البيانات الاقتصادية الكلية :**سنعمل على دمج بيانات اقتصادية كلية (مثل معدلات التضخم، أسعار الفائدة، وتقلبات العملة) مع التحليلات المالية للشركات. هذا سيمكن النظام من تقديم رؤى أوسع حول التأثيرات الخارجية على الأداء المالي للشركات.

5. **توسيع قاعدة البيانات لتشمل شركات ناشئة وصغيرة :**التركيز الحالي على الشركات الكبرى سيشمل في المستقبل الشركات الناشئة والصغيرة، حيث سنقوم بتطوير نماذج مخصصة تناسب طبيعة هذه الشركات، مع الأخذ بعين الاعتبار التحديات الفريدة التي تواجهها.

References :

1 - “Financial Ratios - Complete List and Guide to All Financial Ratios” by the Corporate Finance Institute (CFI)

https://corporatefinanceinstitute.com/resources/accounting/financial-ratios/

2 - “Introduction to Financial Analysis” by Kenneth S. Bigel

https://open.umn.edu/opentextbooks/textbooks/1221

3 - “Financial Ratios Guide” by the Corporate Finance Institute (CFI)

https://corporatefinanceinstitute.com/resources/accounting/financial-ratios-definitive-guide/

4 - The Finance Book: Understand the numbers even if you're not a finance professional

https://www.amazon.co.uk/Finance-Book-Understand-numbers-professional/dp/1292123648/ref=as_li_ss_tl?dchild=1&keywords=financial+statements&qid=1610189075&s=books&sr=1-1-spons&psc=1&spLa=ZW5jcnlwdGVkUXVhbGlmaWVyPUEzQ1dOOFQ3SEdTSkJUJmVuY3J5cHRlZElkPUEwODU4NzYzMzBOMDBMOEdFTDdVMiZlbmNyeXB0ZWRBZElkPUEwNzE3NTUzMTRFMlcyUFhaNjhMUiZ3aWRnZXROYW1lPXNwX2F0ZiZhY3Rpb249Y2xpY2tSZWRpcmVjdCZkb05vdExvZ0NsaWNrPXRydWU%3D&linkCode=sl1&tag=sucatlif-21&linkId=37fde83366dd8aca1375fd1344c787aa&language=en_GB

5 - FT.Warner Hussein: Finance Book 2e: Understand the numbers even if you're not a finance professional:

https://www.amazon.co.uk/Finance-Book-Stuart-Warner/dp/1292401982/ref=bmx_dp_l9n2xe2o_d_sccl_2_1/257-9115590-

9700018?pd_rd_w=T6Bl5&content-id=amzn1.sym.bc00e763-e2be-4cd5-b205-cc33263061df&pf_rd_p=bc00e763-e2be-4cd5-b205-cc33263061df&pf_rd_r=5QFGM705J7BTBE9DP89E&pd_rd_wg=YloKe&pd_rd_r=07767695-662c-499b-96a8-c82dd36362ef&pd_rd_i=1292401982&psc=1