# التنبـؤ بإيرادات الافلام باستخدام خوارزميات التعليم الآلي

تم نشر هذا البحث بتاريخ 15\7\2025 من قبل :

نوري السعدون - حمزة محارب - لين جمّال - هناء دياب - هبة سلاّم


## ملخص البحث

يهدف هذا البحث إلى تطوير نماذج قوية للتعلم الآلي للتنبؤ بإيرادات وربحية الأفلام، وهي مهمة اساسية لصناعة السينما. تم تحقيق ذلك من خلال منهجية شاملة تضمنت دمج ومعالجة مصادر بيانات متنوعة وغنية بالميزات من منصات مثل Kaggle وIMDb، بالإضافة إلى بيانات الممثلين وجوائز الأوسكار. تم تطبيق مجموعة واسعة من نماذج التعلم الآلي، بما في ذلك الانحدار الخطي، أشجار القرار، Random Forest، XGBoost، وCatBoost. أظهرت التجارب أن نموذج Random Forest قدم أداءً فعالاً في التنبؤ بإيرادات الأفلام، في حين برز نموذج CatBoost بتحقيق دقة ممتازة في مهمة تصنيف ربحية الفيلم (رابح/خاسر)، وذلك مع معالجة ناجحة لتحديات البيانات مثل القيم الصفرية. تؤكد النتائج المنجزة على مساهمة البحث في هذا المجال، حيث تُظهر النماذج أداءً تنافسيًا وقويًا مقارنة بالأعمال السابقة، مما يوفر أسساً علمية لدعم اتخاذ القرارات الاستراتيجية في صناعة السينما.


## I. مقدمة

تُعد صناعة الأفلام من القطاعات الاقتصادية الكبرى على مستوى العالم، حيث تتسم بمخاطر استثمارية كبيرة. في ظل التنافس المتزايد وارتفاع تكاليف الإنتاج، أصبح فهم العوامل التي تساهم في نجاح الفيلم التجاري أمراً بالغ الأهمية للمنتجين، المستثمرين، وصناع القرار. يتوقف تحقيق الأرباح بشكل كبير على القدرة على توقع إيرادات الفيلم المحتملة قبل إصداره أو في مراحله المبكرة، مما يسمح باتخاذ قرارات صائبة حول جوانب الإنتاج والتسويق وتوزيع الميزانية. علاوة على ذلك، يُعرف أن اختيار طاقم الممثلين يلعب دوراً محورياً في جذب الجمهور وتحديد القيمة الفنية والتجارية للفيلم، مما يجعل عملية اختيار الممثلين المناسبين تحدياً معقداً يتطلب تحليلاً دقيقاً.

تهدف هذه الدراسة إلى معالجة هذه التحديات من خلال تطوير نموذج شامل للتعلم الآلي وتنفيذ تنقيب بيانات متكامل، يركز على توقع إيرادات الأفلام وتقديم نظام توصية ذكي لاختيار الممثلين. على عكس الدراسات التقليدية التي قد تعتمد على مجموعات بيانات محدودة أو ميزات سطحية، يعتمد هذا المشروع على بناء مجموعة بيانات فريدة وموسعة. تم دمج بيانات متعددة المصادر، تشمل معلومات الأفلام الأساسية، وتقييمات المستخدمين، وبيانات تفصيلية عن الممثلين من قواعد بيانات مثل IMDB، بما في ذلك تقييمات الممثلين ، بالإضافة إلى ميزات هندسية متقدمة مستخلصة من بيانات الميزانية والإيرادات الأولية. هذه المجموعة الغنية من الميزات تتيح استكشافاً أعمق للعلاقات المعقدة بين عوامل الإنتاج المختلفة والأداء المالي للفيلم.

تتمثل المساهمات الرئيسية لهذه الورقة في النقاط التالية:

- بناء وتجميع مجموعة بيانات غير تقليدية وشاملة لإيرادات الأفلام، تتضمن ميزات غنية ومُهندسة خصيصاً عن الممثلين وأداء الأفلام.
- إجراء تحليل استكشافي معمق للبيانات للكشف عن الأنماط الخفية والمشكلات المحتملة مثل القيم المفقودة والتوزيعات غير المتوازنة.
- تطبيق تقنيات متقدمة لهندسة الميزات لإنشاء مقاييس جديدة ذات تأثير كبير على توقع الإيرادات
- تطبيق ومقارنة عدة خوارزميات للتعلم الآلي لتوقع إيرادات الأفلام، مع ضبط دقيق للمعاملات الفائقة وتقييم الأداء باستخدام مقاييس متعددة.
- تطوير نظام توصية بالممثلين يعتمد على البيانات المستخلصة، بهدف مساعدة صناع الأفلام في اختيار الكادر المناسب.
- تقديم تحليل مقارن مفصل لنتائج النماذج مع خطوط الأساس، ومناقشة أداء النماذج بناءً على تعقيدها ووقت تدريبها، وتحديد العوامل الأكثر تأثيراً.

يتم تنظيم بقية هذه الورقة على النحو التالي: يناقش القسم الثاني الأعمال ذات الصلة. يستعرض القسم الثالث وصف مجموعة البيانات والتحليل الاستكشافي للبيانات. يعرض القسم الرابع خطوات معالجة البيانات المسبقة وهندسة الميزات. يفصل القسم الخامس نماذج التعلم الآلي والمنهجيات التجريبية. يقدم القسم السادس النتائج وتحليل الأداء. يناقش القسم السابع مقارنة العمل مع خط الأساس أو أحدث التقنيات. وأخيراً، يقدم القسم الثامن الخاتمة والأعمال المستقبلية.

## II. اعمال ذات صلة

تُعد صناعة الأفلام مجالاً ذا أهمية اقتصادية وثقافية كبرى، مما يدفع الحاجة الملحة لتطوير نماذج دقيقة لتوقع إيرادات الأفلام ونجاحها. تناولت العديد من الدراسات هذه المشكلة باستخدام تقنيات متنوعة ومجموعات بيانات مختلفة.

ركزت بعض الدراسات على توقع الإيرادات بناءً على عوامل ما قبل الإصدار. فعلى سبيل المثال، قامت Xu (2024) [1] بتوقع إيرادات شباك التذاكر باستخدام ميزات مثل الميزانية، والشعبية، والنوع، ومدة العرض، محققةً دقة تصل إلى 73% باستخدام نموذج الغابات العشوائية. ورغم أن هذه الدراسة أكدت أهمية الميزانية والشعبية، إلا أنها أشارت إلى الحاجة لتحسين جودة وكمية البيانات. وبالمثل، استكشفت دراسة Udandarao و Gupta [2] (2024)مجموعة واسعة من خوارزميات التعلم الآلي للانحدار، بما في ذلك النماذج التجميعية، وتوصلت إلى أداء قوي( R-squared - 0.86 ، RMSE - 8.26 مليون دولار) على مجموعة بيانات أكبر. هذه الدراسة أشارت صراحة إلى أن النظر في عوامل مثل المخرج، الكاتب، والممثلين يمثل مجالاً للعمل المستقبلي.

من ناحية أخرى، تناولت دراسات أخرى جوانب مختلفة لنجاح الأفلام أو استخدمت تقنيات أعمق. قام Agarwal وآخرون (2021) [3] بدراسة شاملة لتوقع "معدل نجاح الفيلم" كمسألة تصنيف بناءً على "الميتاكور" من IMDb ، مستخدمين مجموعة بيانات كبيرة وميزات مفصلة لتقييمات المستخدمين، مع إظهار الشبكات العصبية لأفضل أداء (دقة 86%). وفي سياق التعلم العميق، قامت Zheng (2024) [4] بتوقع إيرادات شباك التذاكر باستخدام نماذج مثل Bidirectional LSTM و XGBoost ، مؤكدة على "التأثير الكبير للممثلين النجوم والمخرجين" على الإيرادات.

على الرغم من التقدم الذي أحرزته هذه الدراسات، يبرز مشروعنا الحالي كونه يكمل هذه الجهود ويسد فجوات رئيسية. ففي حين ركزت الدراسات السابقة على توقع الإيرادات [1، 2، 4] أو تصنيف النجاح [3]، إلا أنها غالباً ما افتقرت إلى دمج ميزات متعمقة عن الممثلين أو تحليل تأثيرهم بشكل كمي ومفصل في نماذج التوقع. كما لم تتناول هذه الدراسات جانب أنظمة التوصية بالممثلين، والذي يُعد ذا أهمية عملية كبيرة لصناع الأفلام. لذلك، يتميز هذا العمل بتجميع وبناء مجموعة بيانات فريدة وشاملة تتضمن مقاييس متقدمة ومُهندسة خصيصاً لتقييم الممثلين (مثل متوسط تقييماتهم و معلومات اضافية عنهم)، بالإضافة إلى دمج مهمة توقع الإيرادات مع تطوير نظام توصية بالممثلين، مما يوفر أداة تحليلية وتوصية أكثر تكاملاً لقطاع صناعة الأفلام.

## III. وصف مجموعة البيانات و التحليل الاستكشافي

### 1. اختيار وجمع مجموعة البيانات

لضمان شمولية ودقة توقع إيرادات الأفلام ونظام توصية الممثلين، تم بناء مجموعة بيانات فريدة وموسعة من مصادر متعددة. تمحورت عملية جمع البيانات حول دمج مجموعات بيانات أفلام عامة مع بيانات مفصلة عن الممثلين، وذلك لإنشاء قاعدة بيانات غنية قادرة على التقاط تعقيدات صناعة السينما و معايير النجاح.

تمثلت النقطة الأساسية للبيانات في مجموعة " Kaggle daily update dataset" [5]، والتي توفر معلومات أساسية عن الأفلام مثل الميزانية، والإيرادات، والشعبية، والأنواع، وتواريخ الإصدار. تكميلاً لهذه البيانات، تم استخراج معلومات إضافية عن الأفلام والممثلين من قاعدة بيانات IMDB [6] تم تصفية هذه البيانات بعناية لضمان ملاءمتها لأهداف المشروع. حيث تم استبعاد الأفلام الوثائقية والمسلسلات التلفزيونية، وقصر التركيز على الأفلام الناطقة باللغة الإنجليزية، لضمان تجانس وتركيز مجموعة البيانات.

الجزء الأكثر تميزاً في عملية جمع البيانات وهندستها هو إنشاء مجموعة بيانات مخصصة للممثلين و استخراج تقييم للافلام . فمن خلال تحليل البيانات المستخلصة مجموعة IMDB ، تم تطوير ثلاثة مقاييس تقييم فريدة لكل ممثل، تضاف كميزات جديدة إلى مجموعة البيانات الرئيسية:

1. **متوسط تقييم الممثلين :** يمثل متوسط تقييم IMDB لجميع الأفلام التي شارك فيها الممثل طوال مسيرته الفنية.
2. **تقييم الممثلين بمتوسط بايزي :** تم استخدام متوسط بايزي لتنعيم تقييمات الممثلين وتوفير تقدير أكثر استقراراً. تعتمد هذه الطريقة على ترجيح متوسط تقييم الممثل بالمتوسط العام، مع إعطاء وزن أكبر للممثلين ذوي العدد الأكبر من الأفلام. يضمن هذا النهج تقييمات موثوقة وعادلة، خاصة للممثلين ذوي السجلات الأقل، بتقريب تقييماتهم نحو المتوسط العام.

3. **تقييم الممثلين بنظام العقوبة :** تم تطوير نظام عقوبة مخصص لتعديل متوسط تقييم الممثل بناءً على إجمالي عدد الأصوات التي حصلت عليها أفلامه. يتم تطبيق معامل خصم (مقام) على متوسط التقييم، حيث يصبح الخصم أكبر كلما قل إجمالي عدد الأصوات للممثل، مما يعطي وزناً أقل للممثلين الذين لديهم عدد أصوات ضئيل في أفلامهم.

تم دمج هذه المقاييس الجديدة للممثلين مع مجموعة بيانات الأفلام الرئيسية ، هذه العملية نتج عنها إضافة ميزة غنية تُعزز بشكل كبير من قدرة النماذج على توقع الإيرادات بناءً على الكادر.

المجموعة النهائية للبيانات، بعد دمج هذه المصادر ومعالجة البيانات المسبقة، تتكون من **37093 سجلاً** وتضم **44 ميزة** هذه المجموعة توفر قاعدة بيانات قوية وشاملة تتضمن معلومات ديموغرافية عن الأفلام، بيانات مالية، تقييمات الجمهور، ومقاييس مُهندسة خصيصاً عن أداء الممثلين، مما يجعلها مناسبة تماماً لمهام توقع الإيرادات وإنشاء نظام توصية.

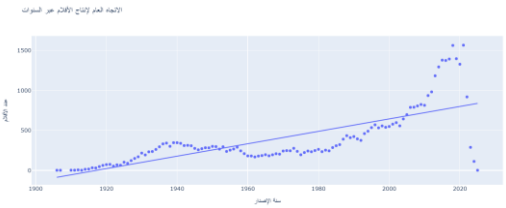
## 2. التحليل الاستكشافي للبيانات

- توزيع الميزات الرقمية الرئيسية :


*رسم توضيحي 1 تزويع الميزات الرقمية*

كشف التحليل الاستكشافي لتوزيعات الميزات الرقمية عن خصائص هامة تتطلب معالجة مسبقة للبيانات. لوحظ أن ميزات مثل الميزانية (budget) والإيرادات (revenue)والشعبية (popularity) وعدد الأصوات (vote_count) تظهر انحرافاً إيجابياً شديداً. يشير هذا إلى أن غالبية الأفلام تقع ضمن نطاق قيم منخفضة لهذه الميزات، مع وجود عدد قليل جدا من الأفلام ذات قيم مرتفعة للغاية، مثل الأفلام ذات الإيرادات الضخمة أو الميزانيات الكبيرة. تم تطبيق تحويل لوغاريتمي على الميزات شديدة الانحراف، مما ساعد في تقريب توزيعاتها من التوزيع الطبيعي وتسهيل الكشف عن الأنماط الكامنة.

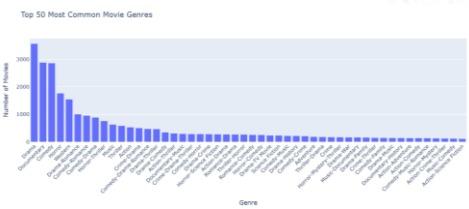
- تحليل الاتجاه العام لإنتاج الأفلام عبر السنوات:




*رسم توضيحي 2 الاتجاه العام لانتاج الافلام عبر السنوات*

أظهر تحليل السلاسل الزمنية لعدد الأفلام المنتجة سنوياً اتجاهاً تصاعدياً واضحاً على المدى الطويل، خاصة منذ بداية التسعينيات. شهدت الفترة من عام 2000 فصاعداً قفزات سريعة في الإنتاج، وهو ما يمكن ربطه بتطور التكنولوجيا وانتشار منصات العرض . بينما كان النمو تدريجياً وبطيئاً بين عامي 1900 و 1950 (مراحل بدايات صناعة السينما)، ولوحظ انخفاض في الإنتاج بين 1940-1980، ربما بسبب تأثيرات الحرب العالمية. ومع نهاية التسعينيات وبداية الألفية، حدثت طفرة. ومع ذلك، لوحظ انخفاض حاد في عدد الأفلام المنتجة بعد عام 2020، والذي يظهر تأثير جائحة كوفيد-19 وتأجيل أو إلغاء العديد من الإصدارات، بالإضافة إلى احتمالية عدم اكتمال البيانات للسنوات الأخيرة و زيادة الاقبال على منصات البث على الانترنت .

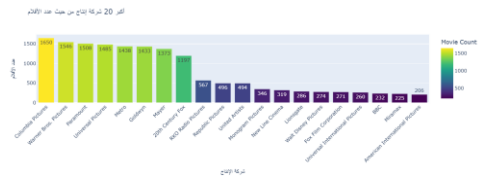
- تحليل الأنواع الأكثر شيوعاً والأنواع الأكثر ربحية:



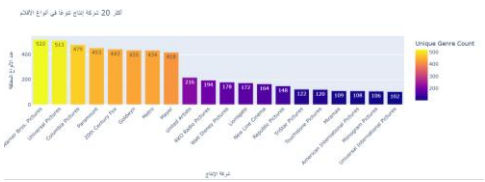

*رسم توضيحي 3 اكثر 50 نوع افلام متكرر*

كشف التحليل أن توزيع إنتاج الأفلام حسب الأنواع غير متوازن بشكل كبير، مع تركيز واضح على الدراما، والكوميديا، والأفلام الوثائقية كأكثر الأنواع إنتاجاً ضمن البيانات . وعند النظر إلى الربحية الصافية، تبين أن الأنواع التي تجمع بين الأكشن والإثارة (Action-Thriller)، أو الكوميديا مع الرومانسية (Comedy-Romance)، أو الكوميديا والدراما والرومانسية (Comedy-Drama-Romance)، بالإضافة إلى الأكشن والدراما (Action-Drama) والرومانسية والكوميديا-(Romance Comedy) تحقق أعلى الأرباح الصافية. هذه النتائج تشير إلى أن دمج عناصر من هذه الأنواع غالباً ما يجذب جماهير أوسع ويحقق عوائد استثمارية ممتازة. كما لوحظ أن الأنواع ذات الميزانيات المنخفضة نسبياً مثل الرعب والوثائقي يمكن أن تكون مربحة جداً مقارنة بتكلفتها. بينما تركزت الأنواع الأقل ربحية (أو الخاسرة) في فئات مثل الأكشن، والموسيقى، والخيال العلمي، مما يشير إلى ضرورة التدقيق في المشاريع ضمن هذه الفئات. يُستنتج أن هناك علاقة واضحة بين الاستثمار الذكي في ميزانيات متوسطة وتحقيق أرباح كبيرة، وأن الاستثمار المرتفع للغاية يمكن أن يحقق أرباحاً ضخمة في أنواع معينة مثل Action-Thriller.

- تحليل أفضل 20 شركة إنتاج من حيث عدد الأفلام المنتجة و انواعها:



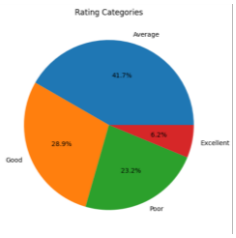

*رسم توضيحي 4 اكثر 20 شركة من حيث عدد الافلام*




*رسم توضيحي 5 اكثر 20 شركة تنوعا في انواع الافلام*

تبين أن الشركات الأكثر إنتاجاً للأفلام هي أيضاً الأكثر تنوعاً من حيث أنواع الأفلام التي تنتجها، مثلWarner Bros. Pictures ، Universal Pictures، Columbia Pictures، Paramount، و 20th Century Fox، والتي تنتج مئات الأنواع المختلفة من الأفلام. هذا يشير إلى أن الشركات الكبرى تسعى إلى الموازنة بين "الإنتاج الضخم" و"تنوع الأنواع" لاستهداف جماهير مختلفة وزيادة الأرباح. ومع ذلك، لا توجد علاقة مباشرة وحتمية بين التنوع وحجم الإنتاج؛ فبعض الشركات المتخصصة قد تفضل التركيز على أنواع محددة لبناء هوية قوية في سوق معين، مما يمكن أن يعزز فرص ربحها في ذلك المجال.

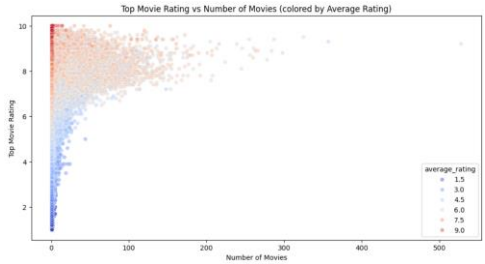
- توزيع تقييمات الأفلام :




*رسم توضيحي 6 توزع تقييمات الأفلام*

أظهر تحليل توزيع تقييمات الأفلام أن غالبية الأفلام تقع ضمن فئة "متوسط"، تليها فئة "جيد". بينما تشكل الأفلام "الممتازة" نسبة صغيرة ,و "الضعيفة" نسبة لا يستهان بها يشير هذا التوزيع إلى أن معظم الأفلام تحقق تقييمات مقبولة إلى جيدة، مع وجود أقلية قليلة جداً من الأفلام عالية التميز

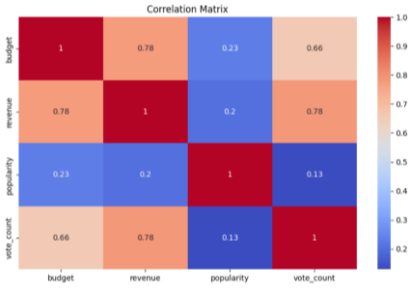
- العلاقة بين عدد أفلام الممثل وأعلى تقييم لفيلم شارك فيه:




*رسم توضيحي 7 التقييمات امام عدد الأفلام*

كشف التحليل عن علاقة مثيرة للاهتمام: الممثلون ذوو عدد الأفلام القليل غالباً ما يظهرون "أعلى تقييم لفيلم" عالي جداً (اكثر من 8)، ويميل متوسط تقييمهم العام ليكون مرتفعاً. بينما كلما زاد عدد الأفلام التي شارك فيها الممثل، يميل التوزيع إلى أن يصبح أكثر تجانساً، حيث يستمر وجود أفلام ذات تقييمات عالية، ولكن يصبح وجود الممثلين ذوي "أعلى تقييم فيلم" شديد الارتفاع (9-10) معدوم و الافلام العالية (اعلى من 8) تصبح اقل كثافة. بمعنى آخر، من الأسهل على ممثل ذي عدد أفلام قليل أن يمتلك فيلماً واحداً بتقييم عالٍ جداً مقارنة بممثل شارك في عدد كبير من الأفلام، حيث يصبح متوسط تقييمه العام أكثر استقراراً وقد يقلل من احتمالية مشاركته بفلم بتقييم عالٍ جداً بشكل استثنائي ضمن مسيرته الفنية الواسعة.

- مصفوفة الارتباط :




*رسم توضيحي 8 مصفوفة الارتباطات*

*رسم توضيحي 9 الارتباطات بشكل اوضح*

أظهرت مصفوفة الارتباط علاقات قوية وواضحة بين الميزات المالية وميزات التقييم:

- **علاقة قوية جداً بين الميزانية والإيرادات:** بلغ معامل الارتباط الإيجابي 0.78، مما يشير إلى أن زيادة الميزانية ترتبط بشكل كبير بزيادة الإيرادات المحققة.
- **علاقة قوية جدًا بين الإيرادات وعدد الأصوات :** بلغ معامل الارتباط الإيجابي 0.78، مما يدل على أن الأفلام ذات الإيرادات العالية تميل أيضاً إلى الحصول على عدد كبير من التقييمات من الجمهور.
- **علاقة جيدة بين الميزانية وعدد الأصوات :** بلغ معامل الارتباط الإيجابي 0.66، مما يشير إلى أن الأفلام ذات الميزانيات الكبيرة غالباً ما تجذب اهتماماً أكبر من الجمهور وتتلقى المزيد من الأصوات و هذا طبيعي كون جزء كبير من الميزانية سيذهب نحو التسويق.
- **علاقات ضعيفة مع الشعبية :** على عكس المتوقع، أظهر ارتباطاً ضعيفًا مع الميزانية (0.23)، والإيرادات (0.20)، وعدد الأصوات (0.13). هذا يشير إلى أن مقياس "الشعبية" في هذه البيانات قد يعتمد على عوامل أخرى أو له تعريف مختلف لا ينعكس بقوة في هذه المتغيرات المالية وتقييمات الجمهور المباشرة.

يتضح من التحليل الاستكشافي أن الميزانية وعدد الأصوات هما عاملان رئيسيان لهما تأثير كبير وارتباط قوي بالإيرادات في هذه المجموعة من البيانات. كما أن هندسة ميزات تقييم الممثلين أضافت أبعاداً جديدة للتحليل، مما يُظهر أهمية هذه الميزات في فهم ديناميكيات نجاح الأفلام.

# IV. معالجة البيانات و هندسة السمات

تعد مرحلة معالجة البيانات و هندسة الميزات حاسمة لضمان جودة البيانات، وتحسين أدائها في نماذج التعلم الآلي، واستخراج أقصى قيمة ممكنة منها .نظرًا للطبيعة المعقدة لمجموعة البيانات ودمج مصادر متعددة، تضمنت هذه المرحلة عمليات مكثفة من التنظيف، والتحويل، وإنشاء ميزات جديدة .سيتم تقسيم الشرح حسب المكونات الرئيسية لمجموعة البيانات لتقديم رؤية واضحة للخطوات المتخذة.

## 1. معالجة مجموعة بيانات الأفلام الأساسية

شكلت مجموعة بيانات الأفلام الأصلية من Kaggle نقطة البداية، وخضعت للخطوات التالية:

### أ. معالجة البيانات الأولية والتنظيف

1. **تصفية الأفلام حسب اللغة:**تم الاحتفاظ فقط بالأفلام التي كانت لغتها الأصلية هي الإنجليزية
2. **حذف الأعمدة غير الضرورية :**تم التخلص من عدة أعمدة لم تكن ذات صلة مباشرة بمهام التنبؤ أو التوصية، أو كانت تحتوي على بيانات غير قابلة للاستخدام بسهولة. هذه الأعمدة شملت: backdrop_path keywords, poster_path, و recommendations.
3. **معالجة القيم المفقودة والصفوف المكررة:**
   - **حذف الصفوف ذات القيم المفقودة و المكررة :** نظراً لوجود نسبة عالية من القيم المفقودة في أعمدة حساسة و هامة للتحليل، فقد تم اتخاذ القرار بحذف جميع الصفوف التي تحتوي على أي قيم فارغة لضمان الاعتماد فقط على سجلات بيانات مكتملة و تم حذف أي صفوف مكررة مع الاحتفاظ بالنسخة الأولى
   - **حذف التكرارات بناءً على العنوان وتاريخ الإصدار :**لمنع تكرار الأفلام التي قد تظهر عدة مرات بسجلات مختلفة قليلاً ولكن بنفس العنوان وتاريخ الإصدار، تم حذف التكرارات
   - **النتيجة :**بعد هذه الخطوات، استقر عدد الأفلام المتبقية في مجموعة البيانات على **48654 فيلماً** .
4. **معالجة قيم وقت التشغيل الصفري والقيم المتطرفة(runtime) :**
   - تم استبعاد الأفلام التي كان وقت تشغيلها runtime يساوي صفراً
   - تم إجراء فحص إضافي للأفلام ذات أوقات التشغيل القصيرة جداً (أقل من 15 دقيقة) والطويلة جداً (أكثر من 400 دقيقة). بعد التحقق من أنها لا تنتمي لأنواع محددة قد تبرر هذه القيم (مثل الأفلام القصيرة جداً أو الوثائقية)، تم اتخاذ قرار بتصفية مجموعة البيانات للاحتفاظ فقط بالأفلام التي يتراوح وقت تشغيلها بين 15و 400 دقيقة يضمن هذا التركيز على الأفلام التقليدية ويزيل القيم المتطرفة التي قد تؤثر سلباً على النمذجة.

### ب. هندسة الميزات

تم اشتقاق عدة ميزات جديدة من الأعمدة الموجودة لإثراء مجموعة البيانات وزيادة قدرة النماذج على التعلم و التقاط العلاقات:

1. **ميزة الربح(profit) :** تم حساب الربح الصافي لكل فيلم بطرح قيمة الميزانية من الإيرادات
2. **تفصيل تاريخ الإصدار(release_date) :** تم تقسيم عمود release_date إلى مكوناته الزمنية لإنشاء ميزات جديدة:
   - release_year: سنة إصدار الفيلم.
   - release_month. شهر إصدار الفيلم.
   - release_day: يوم إصدار الفيلم .يساعد هذا في الكشف عن الأنماط الموسمية أو الاتجاهات المتعلقة بسنة الإصدار.
3. **عدد الممثلين الفريدين (unique_actors_count):** تم استخلاص عدد الممثلين الفريدين المشاركين في كل فيلم من عمود credits الذي يحتوي على أسماء الممثلين،
4. **تحليل المشاعر لجملة الترويج (tagline_sentiment) :** تم تطبيق تحليل المشاعر على جمل الترويج الخاصة بالأفلام لإنشاء ميزة جديدة تعكس النبرة العاطفية للجملة (إيجابية، سلبية، أو محايدة). تهدف هذه الميزة إلى تزويد نموذج التعلم الآلي بمعلومة إضافية قد تكون مؤشراً مهماً على جاذبية الفيلم للجمهور وبالتالي على إيراداته المحتملة. على سبيل المثال، قد تعزز جملة ترويجية إيجابية اهتمام المشاهدين وتزيد الإيرادات.
5. **تصنيف حقبة الفيلم(movie_era_classification) :** تم إنشاء ميزة فئوية جديدة تُصنف الفيلم إلى حقب زمنية مختلفة بناءً على سنة إصداره، وذلك لالتقاط تأثير التغيرات التاريخية والثقافية في صناعة السينما:
   - 'Classic': قبل عام 1980
   - 'Old_School': من 1980 إلى قبل 2000
   - 'Modern': من 2000 إلى قبل 2015
   - 'New_Era': من 2015 فصاعداً
6. **تصنيف فئة الميزانية (budget_category):** تم تحويل الميزانية إلى فئات لتسهيل التحليل والفهم، وربما لالتقاط العلاقات غير الخطية:
   - 'low_budget': أقل من 1 مليون دولار
   - 'medium_budget': من 1 مليون إلى أقل من 20 مليون دولار
   - 'high_budget': من 20 مليون إلى أقل من 100 مليون دولار
   - 'blockbuster': 100 مليون دولار فأكثر

## 2. معالجة مجموعة بيانات IMDb الكاملة

خضعت مجموعة بيانات IMDb ، والتي تُستخدم لتوفير معلومات إضافية حول الأفلام والممثلين، لعمليات تنظيف و هندسة ميزات محددة لضمان جودتها وملاءمتها للمهمة:

### أ. معالجة البيانات الأولية والتنظيف

1. **معالجة القيم المفقودة:** تم التعامل مع القيم المفقودة في عمود genres بشكل خاص عن طريق استبدالها بالقيمة "Unknown"، لضمان اكتمال هذا العمود الهام. وتم فحص الأعمدة الأخرى للقيم المفقودة، وتم التعامل معها ضمن استراتيجية شاملة.
2. **حذف الأعمدة غير الضرورية :**تم التخلص من الأعمدة الغير ضرورية مثل isAdult و endYear لأن التركيز كان على الأفلام العامة
3. **تصفية الأفلام فقط :**لضمان أن مجموعة البيانات تحتوي على الأفلام التقليدية فقط، تم تصفية السجلات للاحتفاظ بتلك التي يكون فيها titleType يساوي 'movie'
4. **حذف الأنواع غير المرغوبة :**تم استبعاد الأفلام التي تنتمي إلى أنواع معينة غير ذات صلة بالتنبؤ بإيرادات شباك التذاكر للأفلام التجارية التقليدية. شملت هذه الأنواع : 'Reality-TV', 'News', 'Documentary', 'Biography', 'Adult', 'Game-Show', 'Short', 'Sport', 'music', و'Talk-Show'.

5. **معالجة السنوات غير المنطقية :**تم فحص عمود startYear لتحديد أي قيم غير منطقية (مثل السنوات التي تسبق 1900 أو تتجاوز 2025). تم تحديد هذه السجلات لضمان أن جميع الأفلام ضمن نطاق زمني معقول.
6. **معالجة قيم وقت التشغيل غير الطبيعية(runtimeMinutes) :**
   - لوحظ وجود قيم متطرفة وغير منطقية ، تتراوح من دقيقة واحدة إلى ما يقرب من 59460 دقيقة (حوالي 41 يوماً).
   - لضمان التركيز على الأفلام التقليدية ذات وقت تشغيل معقول، تم تصفية مجموعة البيانات للاحتفاظ فقط بالأفلام التي يتراوح وقت تشغيلها بين 30 و 300 دقيقة
7. **إضافة متوسط تقييم الفيلم :**تم دمج متوسط التقييم(averageRating) لكل فيلم من مجموعة بيانات التقييمات إلى مجموعة بيانات IMDb الأساسية باستخدام المعرف الفريد للفيلم (tconst) كعمود ربط

**ب. هندسة الميزات**

تم اشتقاق ميزات جديدة من الأعمدة الموجودة في مجموعة بيانات IMDb لتعزيز قدرتها التحليلية:

1. **تصنيف وقت التشغيل (runtime):** تم تحويل وقت تشغيل الفيلم الرقمي إلى فئات وصفية لتسهيل التحليل والنمذجة:
   - 'short': أقل من 75 دقيقة.
   - 'standard': من 75 إلى 120 دقيقة.
   - 'long': أكثر من 120 دقيقة.
2. **عمر الفيلم (movie_age):** تم حساب عمر الفيلم حتى العام الحالي (2025) عن طريق طرح سنة الإصدار(startYear) من 2025:
3. **فئة التقييم (rating_bucket):**تم إنشاء ميزة فئوية جديدة بناءً على متوسط تقييم الفيلم لتصنيف جودته:
   - 'Excellent': 8 أو أعلى.
   - 'Good': من 6.5 إلى أقل من 8.
   - 'Average': من 5 إلى أقل من 6.5.
   - 'Poor': أقل من 5.

## 3. معالجة بيانات الممثلين وهندسة الميزات

**أ. استخلاص وتجميع معلومات الممثلين**

1. **تصفية وتوحيد معلومات الممثلين:**
   - تم البدء بتصفية جدول الذي يحتوي على أدوار الممثلين في الأفلام للاحتفاظ فقط بالسجلات التي تخص الممثلين
   - بعد ذلك، تم دمج هذا الجدول مع جدول أسماء الأشخاص( باستخدام عمود nconst معرّف الشخص) للحصول على الأسماء الأساسية للممثلين المرتبطة بكل دور في فيلم.
2. **تجميع الممثلين لكل فيلم:**
   - تم تجميع أسماء جميع الممثلين المشاركين في كل فيلم ضمن عمود واحد جديد تم ذلك عن طريق التجميع حسب معرّف الفيلم مع إزالة أي أسماء مكررة من بيانات IMDB مع اضافة عمود افضل فلم
3. **تنظيف بيانات الممثلين المدمجة:**
   - تم تحويل عمود actors_in_movie إلى نوع stringو تم ملء القيم مفقودة
   - تم فلترة الصفوف التي لا تحتوي على معلومات ممثلين

**ب. هندسة الميزات**

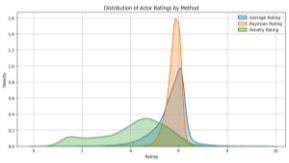
لتمثيل جودة أداء الممثلين بشكل أكثر دقة، تم تطوير ثلاثة مقاييس تقييم مختلفة لكل ممثل، بالإضافة إلى معلومات عن أفضل فيلم له. تطلبت هذه العملية أولاً دمج عدد الأصوات من مجموعة بيانات افلام IMDB، حيث أن vote_count أمر بالغ الأهمية لنظام العقوبة.

1. **المتوسط البسيط لتقييم الممثلين(average_rating) :**
   - يمثل هذا المتوسط الحسابي المباشر لجميع تقييمات الأفلام التي شارك فيها الممثل. يوفر هذا المقياس رؤية سريعة لأداء الممثل بناءً على جميع أعماله.
2. **المتوسط البايزي لتقييم الممثلين(bayesian_avg_rating) :**
   - يهدف هذا المقياس إلى توفير تقييم أكثر استقراراً وموثوقية للممثلين، خاصة أولئك الذين لديهم عدد قليل من الأفلام. يتم حسابه كمتوسط مرجح بين المتوسط العام لتقييمات جميع الممثلين (global_avg) والمتوسط البسيط للممثل الفردي.
   - تُطبق الصيغة

$$(C * global\_avg + sum\_ratings) / (C + num\_movies)$$

حيث C هو معامل الثقة (تم تعيينه بـ 5، وهو قيمة قابلة للضبط)، global_avg هو المتوسط العام لتقييمات جميع الأفلام، sum_ratings هو مجموع تقييمات الأفلام التي شارك فيها الممثل، و num_movies هو عدد الأفلام التي شارك فيها الممثل.

   - هذا النهج يسحب تقييم الممثلين ذوي عدد الأفلام القليل نحو المتوسط العام، مما يمنع المبالغة في تقدير أو التقليل من قيمتهم بناءً على عينة صغيرة.
3. **تقييم الممثلين مع عقوبة على قلة عدد الأصوات**
   - هذا النظام يهدف إلى معاقبة (خفض) متوسط تقييم الممثل إذا كان إجمالي عدد الأصوات التي حصلت عليها أفلامه منخفضاً، مما يشير إلى قلة تعرضه لتقييم الجمهور.
   - يتم حساب مجموع الأصوات الكلي (total_votes) لجميع الأفلام التي شارك فيها الممثل.
   - يتم تطبيق معامل خصم (divisor) على المتوسط البسيط للممثل (average_rating) بناءً على total_votes وفقاً للعتبات التالية (التي تم تحديدها بناءً على تحليل توزيع الأصوات):
     - إذا كان أقل من 10: divisor = 4
     - إذا كان بين 10 و 30: divisor = 3
     - إذا كان بين 31 و 2:100 = divisor
     - إذا كان أكبر من 100: لا يوجد خصم
   - penalty_avg_rating = average_rating / divisor
   - هذا يضمن أن الممثلين الذين لا تحظى أفلامهم بالكثير من التقييمات يظهرون بتقييم أكثر تحفظاً.




*رسم توضيحي 10 مقارنة توزع التقييمات بالطرق الثلاث*

4. **معلومات أفضل فيلم للممثل:**
   - لكل ممثل، تم أيضاً استخلاص اسم الفيلم ذو التقييم الأعلى له (top_movie_name)وتقييم هذا الفيلم.(top_movie_rating)

**النتائج :**نتج عن هذه العمليات جدول مجموعة بيانات خاصة للممثلين تم دمجها لاحقا مع بيانات الافلام لتكوين سمة تقييم الكادر التمثيلي

- من خلال الإحصائيات الوصفية، لوحظ أن المتوسط البايزي كان له انحراف معياري أقل بكثير مقارنة بالمتوسط البسيط ، مما يؤكد فعاليته في تنعيم التقييمات.
- في المقابل، أظهر تقييم العقوبة متوسطاً أقل وتشتتاً أكبر، مما يعكس تطبيق الخصومات على الممثلين ذوي التقييمات القليلة.

## 4. معالجة بيانات جوائز الأوسكار وهندسة الميزات

لإضافة بُعد يتعلق بالجوائز المرموقة، تم دمج بيانات جوائز الأوسكار [7] في التحليل، مع عمليات معالجة وهندسة ميزات لربطها بدقة ببيانات الأفلام والممثلين.

**أ. معالجة البيانات الأولية والتنظيف**

- **توحيد العناوين للمطابقة:** تم تحديد عدد الأفلام التي تطابقت بين المجموعتين و ذلك بعد القيام بتحويلات لتوحيد الاسماء (ازالة الفراغات و التحويل لاحرف صغيرة)، مما وفر أساساً لربط بيانات الأوسكار و كان عددها 3097

**ب. هندسة الميزات**

تم اشتقاق ثلاث ميزات رئيسية تتعلق بجوائز الأوسكار من بياناتها الخام:

1. **حالة الأوسكار للفيلم (movie_oscar):**
   - تم تحويل عمود إلى قيمة رقمية: 2 إذا كان الفيلم فائزاً ، و 1 إذا كان مرشحاً فقط.
   - تم بعد ذلك إنشاء خريطة تربط كل فيلم عدد حالات الأوسكار الذي حققها (ترشيح أو فوز).
   - تم تطبيق هذه الخريطة على عمود عناوين الافلام ,هذه الميزة تعكس مستوى تقدير الأكاديمية للفيلم نفسه.
2. **حالة الأوسكار لطاقم العمل (movie_credits_oscar):**
   - لتقييم تأثير الممثلين أو المخرجين الحائزين على جوائز أوسكار في الفيلم، تم إنشاء ميزة movie_credits_oscar.
   - أولاً، تم توحيد أسماء الأشخاص (الممثلين، المخرجين) في oscars_dfعن طريق تحويلها إلى أحرف صغيرة وإزالة المسافات والأحرف غير الأبجدية الرقمية.
   - تم تحديد حالة الأوسكار الرقمية لكل شخص
   - تم إنشاء خريطة تربط كل اسم شخص موحّد بأعلى حالة أوسكار حققها (فوز أو ترشيح).
   - بعد ذلك، لكل فيلم ، تم استخراج أسماء الممثلين من عمود credits (بعد توحيدها وتقسيمها) واستخدام الخريطة لتحديد عدد حالات الأوسكار بين جميع الأفراد الرئيسيين المشاركين في الفيلم.
   - تهدف هذه الميزة إلى قياس "ثقل الأوسكار" الذي يجلبه طاقم العمل إلى الفيلم.

3. **مجموع جوائز الأوسكار للشركات المنتجة(company_oscars) :**
   - بنفس طرق المعالجة السابقة تم حساب مجموع حالات الأوسكار (1 للترشيح، 2 للفوز) لكل شركة إنتاج عبر جميع الأفلام التي أنتجتها والتي ظهرت في بيانات الأوسكار.
   - تم إنشاء دالة تجمع هذه القيم لكل فيلم بناءً على شركات الإنتاج المرتبطة به.
   - تعكس هذه الميزة الخبرة والتقدير العام لشركات الإنتاج في الفوز أو الترشح لجوائز الأوسكار، مما قد يكون مؤشراً على جودتها وقدرتها على إنتاج أفلام ناجحة.

### 5. معالجة بيانات الممثلين النجوم

لإنشاء ميزة تميز "نجوم السينما" الأكثر شهرة ودخلاً، تم تحليل ودمج بيانات [8] من مصدرين رئيسيين.

**أ. وصف مصادر البيانات**

1. **Celebrity.csv:**
   - يحتوي على معلومات عامة عن المشاهير
   - الحجم الأولي: 9980 صفاً و 8 أعمدة.
2. **forbes_celebrity_100.csv:**
   - يحتوي على قائمة فوربس لأعلى المشاهير دخلاً.
   - الحجم الأولي: 1647 صفاً و 4 أعمدة.

**ب. تنظيف بيانات المشاهير(Celebrity.csv)**

1. **حذف الأعمدة غير الضرورية :**تم إزالة الأعمدة Unnamed, adult, gender و id لعدم الحاجة إليها في هذا التحليل.
2. **تصفية حسب مجال التمثيل :**تم الاحتفاظ فقط بالسجلات التي ينتمي فيها المشهور إلى مجال التمثيل
3. **التخلص من الأسماء غير الإنجليزية**
4. **إزالة التكرارات**
5. **تحليل الشعبية وتحديد أفضل 150 :**بعد تنظيف البيانات، تم إجراء تحليل وصفي لسمة popularity. ثم تم ترتيب البيانات ترتيباً تنازلياً حسب popularity، ثم تم اختيار أفضل 150 شخصية.
6. **الحجم النهائي :**أصبحت مجموعة بيانات المشاهير النظيفة تحتوي على 5946 صفاً و 4 أعمدة.

**ج. تنظيف بيانات فوربس(forbes_celebrity_100.csv)**

1. **تصفية فئة التمثيل :**تم تصفية البيانات للاحتفاظ فقط بالممثلين نتج عن ذلك 176 ممثلاً و 107 ممثلات.
2. **معالجة التكرارات :**عن طريق استخراج أحدث ظهور فقط لكل شخصية
3. **الترتيب حسب الدخل :**تم ترتيب النتائج تنازلياً حسب قيمة Pay لتحديد المشاهير الأعلى دخلاً.
4. **الحجم النهائي :**أصبحت مجموعة بيانات فوربس النظيفة تحتوي على 524 صفاً و 4 أعمدة.

**د. مطابقة ودمج الملفين**

1. **توحيد الأسماء للمطابقة :**لتسهيل عملية المطابقة الدقيقة بين المجموعتين، تم تحويل أسماء المشاهير في كلتا المجموعتين إلى صيغة موحدة (أحرف صغيرة، بدون مسافات أو أحرف خاصة).
2. **استخراج التقاطع :**تم تحديد الأسماء المشتركة بين مجموعتي البيانات بعد عملية التوحيد. بلغ عدد الأسماء المتطابقة الفريدة 73 اسماً.
3. **التصفية والدمج :**تم الاحتفاظ فقط بالسجلات التي تحتوي على هذه الأسماء المتطابقة في كل من مجموعتي بيانات المشاهير وفوربس. بعد ذلك، تم دمج المجموعتين باستخدام العمود الموحد Name كعمود ربط.
4. **معالجة التكرارات بعد الدمج :**تم الاحتفاظ بأحدث سجل دخل لكل شخصية بعد عملية الدمج، مما أدى إلى قائمة نهائية من 73 شخصية فريدة.

- **ميزة is_superstar:** تم إضافة ميزة (is_superstar) وتم تعيين قيمتها إلى True لجميع السجلات التي فيها نجم في بيانات الافلام لتمييز هذه الشخصيات كنقاط بيانات "نجوم سينما."

## V. نماذج التعليم الآلي و التجارب

في هذا القسم، سيتم استعراض النماذج المستخدمة، المنهجية التجريبية، والنتائج المحققة في محاولة التنبؤ بإيرادات وربحية الأفلام، مع تسليط الضوء على التحديات والحلول المتبعة.

### 1. تحضير البيانات لتدريب النماذج

قبل البدء بتدريب النماذج، تم إجراء خطوة أساسية لإزالة الميزات التي اعتُبرت غير مفيدة بشكل مباشر لعملية التنبؤ بالإيرادات أو الربح او التي اضعفت اداء النموذج. شملت هذه الميزات:

- معرفات البيانات id, tconst, primaryTitle, originalTitle, normalized_title, normalized_primary_title.
- معلومات وصفية أو زمنية قد لا تساهم بشكل مباشر كمتغيرات مستقلة قوية في التنبؤ بعد هندسة الميزات : title, tagline, release_date, startYear, runtimeMinutes, genres_y, status.
- ميزات تتعلق بالتقييم أو الأوسكار averageRating, movie_oscar

تم تقسيم مجموعة البيانات الناتجة إلى ثلاثة أقسام: بيانات تدريب ، بيانات تحقق ، وبيانات اختبار(Test) ، لضمان تقييم موضوعي لأداء النماذج.

### 2. اختيار المتغير المستهدف

تم إجراء تجربة أولية لتقييم أي من المتغيرين المستهدفين (الربح أو الإيرادات) يقدم أداءً أفضل للتنبؤ. أظهرت النتائج على مجموعة الاختبار النهائية ما يلي:

| المتغير المستهدف | أفضل نموذج | R2 |
|---|---|---|
| **الربح** | Random Forest | 0.6005 |
| **الإيرادات** | Random Forest | 0.7761 |

*جدول 1 اختيار المتغير الهدف*

بناءً على هذه النتائج، تم اتخاذ قرار بالاعتماد على **الإيرادات (Revenue)** كمتغير مستهدف رئيسي نظراً لأدائها التنبؤي المتفوق.

### 3. النماذج المستخدمة والنتائج الأولية

تم تدريب واختبار مجموعة من نماذج التعلم الآلي المتنوعة للتنبؤ بالإيرادات, النتائج الأولية للتنبؤ بالإيرادات (على بيانات التدريب):

| النموذج | $R^2$ Score |
|---|---|
| Linear Regression | 0.778514 |
| Decision Tree | 0.686897 |
| Bagging | 0.791188 |
| Random Forest | 0.820138 |
| XGBoost | 0.801316 |
| Gradient Boosting | 0.781913 |
| LightGBM | 0.793551 |
| CatBoost | 0.812622 |

*جدول 2 النماذج المستخدمة*

أظهر نموذج **Random Forest** أفضل أداء مبدئي بقيمةR2 تبلغ 0.82 على بيانات التدريب، و 0.77 على مجموعة الاختبار المخصصة للإيرادات.

### 4. التحديات والتجارب المتكررة

واجهت عملية التنبؤ تحدياً كبيراً يتمثل في وجود عدد كبير من الأفلام ذات الإيرادات والميزانيات الصفرية، وهي قيم غير واقعية وتؤثر سلباً على توزيع البيانات وأداء النماذج. لمواجهة هذا التحدي، تم إجراء عدة تجارب:

1. تطبيق التحويل اللوغاريتمي والقص (Challenges and Iterative Experiments)

- **الدافع :**لمعالجة التوزيع غير الطبيعي لبيانات الإيرادات وتقليل تأثير القيم المتطرفة والصفرية. تم تطبيق تحويل لوغاريتمي على المتغير المستهدف، مع قص القيم الأقل من 1 إلى صفر قبل التحويل.
- **مقاييس التقييم :**تم تقييم أداء النماذج باستخدام $R^2$ وخطأ المتوسط المطلق (MAE) والجذر التربيعي لمتوسط مربعات الأخطاء (RMSE).

| النموذج | $R^2$ | MAE | RMSE |
|---|---|---|---|
| Linear Regression | -33588.1 | 3.58952e+08 | 1.42802e+10 |
| Decision Tree | 0.6433 | 1.0905e+07 | 4.6536e+07 |
| Bagging | 0.7499 | 7.70838e+06 | 3.89656e+07 |
| Random Forest | 0.775 | 7.25141e+06 | 3.69612e+07 |
| XGBoost | 0.7611 | 7.58782e+06 | 3.80872e+07 |
| Gradient Boosting | 0.6681 | 7.95966e+06 | 4.4891e+07 |

| | | | |
|---|---|---|---|
| LightGBM | 0.7581 | 7.3896e+06 | 3.83229e+07 |
| CatBoost | 0.7405 | 7.27423e+06 | 3.96886e+07 |

*جدول 3 تطبيق التحويل اللوغاريتمي و القص*

ظل نموذج **Random Forest** هو الأفضل في هذه التجربة أيضاً، محققاً $R^2$ قدره 0.775

2. الانحدار الشرطي بعد التصنيف الثنائي (Conditional Regression after Binary Classification)

**الدافع :**لمعالجة مشكلة القيم الصفرية بشكل أكثر مباشرة، تم استخدام نهج من خطوتين:

- **الخطوة الأولى :**بناء نموذج تصنيف ثنائي للتنبؤ ما إذا كانت الإيرادات ستكون صفراً أم غير صفرية.
- **الخطوة الثانية :**إذا كانت الإيرادات غير صفرية، يتم تمريرها إلى نموذج انحدار منفصل للتنبؤ بالقيمة الفعلية للإيرادات.

**النتائج:**

| النموذج | $R^2$ | MAE | RMSE |
|---|---|---|---|
| Linear Regression | -19306.370 | 2303481380 | 26840524324 |
| Random Forest | 0.753 | 48206994 | 95985253 |
| XGBoost | 0.692 | 53355225 | 107153341 |
| Gradient Boosting | 0.654 | 53359829 | 113568278 |
| LightGBM | 0.766 | 47414590 | 93502160 |
| CatBoost | 0.739 | 48193353 | 98650595 |

*جدول4 تطبيق الانحدار الشرطي بعد التصنيف الثنائي*

**الاستنتاج :**على الرغم من هذا النهج، لوحظ أن النموذج لا يزال يواجه صعوبة في التنبؤ ببعض القيم بدقة، مما يشير إلى وجود أخطاء كبيرة في بعض الحالات.

3. التصنيف الثنائي للربح/الخسارة ( Binary Classification for Profit/Loss)

**الدافع :**نظراً للتحديات في التنبؤ بقيمة الإيرادات الدقيقة بسبب القيم الصفرية، تم تحويل المشكلة إلى مسألة تصنيف: هل الفيلم رابح أم خاسر؟ (إذا كانت الإيرادات أكبر من الميزانية، فهو رابح).

**معالجة عدم توازن الفئات :**لوحظ أن 80% من الأفلام كانت "خاسرة"، مما يمثل عدم توازن كبير في الفئات. تم استخدام معلمة weight في نموذج XGBoost لموازنة الأهمية بين الفئات) على غرار إعادة أخذ العينات (resampling لمعالجة هذه المشكلة.

**مقاييس التقييم :**تم استخدام مقاييس تصنيف مثل الدقة(Accuracy) ، الاستدعاء(Recall) ، الدقة(Precision) ، ودرجة(F1-score) F1 .

**النتائج(XGBoost) :**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 6276 | 0 | 0.93 | 0.96 | 0.94 |
| 1143 | 1 | 0.74 | 0.58 | 0.65 |
| accuracy | | | 0.90 | 7419 |
| macro avg | 0.83 | 0.77 | 0.80 | 7419 |
| weighted avg | 0.90 | 0.90 | 0.90 | 7419 |

*جدول5 نتائج XGBoost*

مصفوفة الارتباك[[6044 232] ,[479 664]] :

عدد الأخطاء: 711 من 7419

**النتائج CatBoost (أفضل أداء في التصنيف):**

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| **0** | 0.93 | 0.97 | 0.95 | 6276 |
| **1** | 0.78 | 0.63 | 0.70 | 1143 |
| **accuracy** | | | 0.92 | 7419 |
| **macro avg** | 0.86 | 0.80 | 0.82 | 7419 |
| **weighted avg** | 0.91 | 0.92 | 0.91 | 7419 |

*جدول 6 نتائج CatBoost*

أظهر نموذج **CatBoost** أفضل أداء في مهمة التصنيف الثنائي للربح/الخسارة، محققًا دقة إجمالية 0.92

لتقييم مدى فعالية وأهمية النموذج المقترح في هذا المشروع، تم إجراء مقارنة شاملة مع مجموعة من الأبحاث السابقة في مجال التنبؤ بإيرادات الأفلام ونجاحها. تهدف هذه المقارنة إلى وضع النتائج المحققة في سياق المعرفة الحالية وتحديد نقاط القوة والضعف مقارنة بالأساليب والتقنيات الأساسية والحديثة.

## VI. النتائج و تحليل الاداء

في هذا المشروع، تم التركيز على مهمتين رئيسيتين:

1. **التنبؤ بإيرادات الفيلم(Revenue Prediction) :**
   - **أفضل نموذج** : Random Forest
   - **الأداء (بدون تحويل لوغاريتمي):** $R^2 = 0.7761$ على مجموعة الاختبار.
   - **الأداء (بعد التحويل اللوغاريتمي)** : $R^2 = 0.775$، MAE = 7.25m، RMSE = 36.96m.
2. **تصنيف ربحية الفيلم(Profit/Loss Classification) :**
   - **أفضل نموذج:** CatBoost
   - **الأداء** :Accuracy = 0.92، F1-score = 0.70(للفئة الإيجابية – الرابحين)

## VII. مقارنة العمل مع خط الأساس أو أحدث التقنيات

### ب. مقارنة بالدراسات المرجعية (Comparison with Referenced Studies)

1. **"A Comprehensive Study on Various Statistical Techniques for Prediction of Movie Success" (Agarwal et al.)**

- **هدف الورقة :**التنبؤ بنجاح الفيلم (تصنيف إلى ناجح، فاشل، متوسط) بناءً على Metascore.
- **النماذج وأداؤها (الدقة) :**الشبكة العصبية (ANN) هي الأفضل بدقة 86% تليها Logistic Regressionبدقة 76%
- **المقارنة :**تركز هذه الورقة على مهمة تصنيف النجاح، وهي مشابهة لمهمة تصنيف الربح/الخسارة في مشروعنا. حقق نموذج CatBoost في مشروعنا دقة إجمالية بلغت 92% لتصنيف الربح/الخسارة. تُظهر هذه النتيجة أن أداء نموذج CatBoost لدينا يتجاوز بشكل ملحوظ الدقة المبلغ عنها لأفضل النماذج في هذه الدراسة (86% للشبكة العصبية)، مما يشير إلى فعالية النهج المتبع في مشروعنا لمهمة التصنيف الثنائي.

### 2. "Movie Revenue Prediction Using Machine Learning Models" (Udandarao & Gupta)

- **هدف الورقة :**التنبؤ بإيرادات الأفلام.
- **النماذج وأداؤها :** Gradient Boosting (0.8242) و XGBoosting (0.8102) و Random Forest (0.7786)
- **المقارنة :**تتشابه هذه الدراسة بشكل وثيق مع مهمة التنبؤ بالإيرادات في مشروعنا. نموذج Random Forest في مشروعنا حقق $R^2$ قدره 0.7761. بالمقارنة، حقق Random Forestفي هذه الدراسة $R^2$ قدره 0.7786، مما يظهر تقارباً كبيراً في الأداء بين النموذجين لنفس النوع. ومع ذلك، تجاوزت نماذج Boosting في هذه الدراسة أداءRandom Forest لدينا. هذا يشير إلى أن هناك مجالاً لتحسين أداء نماذج الانحدار لدينا، ربما من خلال تحسين المعاملات الفائقة أو استكشاف ميزات إضافية.

### 3. "Long–Range Movie Box Office Prediction Based on Machine Learning" (Xu)

- **هدف الورقة :**التنبؤ بإيرادات شباك التذاكر على المدى الطويل قبل إصدار الفيلم.
- **النماذج وأداؤها** : Random Forest (0.738) , linear regression (0.608)، والشبكة العصبونية(0.625) .
- **المقارنة :**تركز هذه الدراسة على التنبؤ بالإيرادات، وتشبه مهمتنا في هذا الصدد. نموذج Random Forest في مشروعنا (0.7761 ) يتفوق على أداء Random Forestفي هذه الدراسة(0.738 ) هذا يشير إلى أن مجموعة الميزات الغنية وعملية المعالجة المسبقة للبيانات في مشروعنا قد ساهمت في تحقيق أداء تنبؤي أفضل لإيرادات الأفلام مقارنةً بهذه الدراسة.

### 4. "Predicting Movie Box Office Based on Machine Learning, Deep Learning, and Statistical Methods" (Zheng)

- **هدف الورقة :**التنبؤ بإيرادات شباك التذاكر العالمية.
- **النماذج وأداؤها:** Bidirectional LSTM (0.7364) ، XGBoost (0.7084)، Random Forest (0.6724).
- **المقارنة :**تتناول هذه الدراسة أيضاً التنبؤ بالإيرادات وتستخدم نماذج تعلم آلة وتعلم عميق. مقارنة بأفضل نموذج لدينا للانحدار(Random Forest 0.7761) ، فإن أداء Random Forestو XGBoost في هذه الورقة (0.6724 و 0.7084 ) أقل من أدائنا. ومع ذلك، حقق نموذج التعلم العميق Bidirectional LSTM في هذه الورقة $R^2$ قدره 0.7364، والذي يقترب من أداء Random Forest لدينا. هذا يسلط الضوء على أن نماذج التعلم العميق قد توفر بدائل قوية، خاصة عند التعامل مع بيانات معقدة.

### ج. الاستنتاجات من المقارنة (Conclusions from Comparison)

تظهر هذه المقارنة أن العمل المنجز في هذا المشروع يحقق أداءً تنافسياً، وفي بعض الحالات، يتفوق على النتائج المبلغ عنها في الأبحاث السابقة، لا سيما في مهمة تصنيف ربحية الفيلم، حيث حقق نموذج CatBoost دقة عالية جداً. في مهمة التنبؤ بالإيرادات، أظهر نموذج Random Forest أداءً قوياً يتجاوز بعض الأبحاث المرجعية ويقترب من أفضل النتائج المبلغ عنها في دراسات أخرى تستخدم نماذج مشابهة.

ومع ذلك، تشير مقارنتنا مع الأبحاث التي استخدمت نماذج التعلم المعزز(مثل Gradient Boosting و XGBoost أو التعلم العميق مثل(LSTM ) إلى أن هناك دائماً مجالاً لمزيد من التحسين. قد يتضمن العمل المستقبلي استكشافاً أعمق لتعديلات المعاملات الفائقة (hyperparameter tuning)لهذه النماذج، أو دمج ميزات إضافية، أو تجربة معماريات تعلم عميق أكثر تعقيداً لربما تتجاوز الأداء الحالي.

## VIII. الخاتمة و الاعمال المستقبلية

يمثل هذا المشروع جهداً شاملاً للتنبؤ بإيرادات وربحية الأفلام، وقد نجح في دمج ومعالجة مجموعات بيانات متعددة ومعقدة كأساس قوي للنمذجة. أظهرت التجارب فعالية نموذج Random Forest في التنبؤ بالإيرادات، محققاً أداءً تنافسياً حتى بعد معالجة تحدي القيم الصفرية بتحويلات لوغاريتمية. كما برز نموذج CatBoost في مهمة تصنيف ربحية الفيلم، محققاً دقة عالية تجاوزت 90%، مما يدل على قدرته الفائقة في التعامل مع عدم توازن الفئات. بشكل عام، تضع هذه النتائج المشروع في مكانة تنافسية مقارنة بالأبحاث السابقة في المجال. للأعمال المستقبلية، يُقترح تحسين دقة نماذج الانحدار من خلال ضبط المعاملات الفائقة واستكشاف نماذج التعلم العميق، بالإضافة إلى تطوير أساليب أكثر تقدماً لتعزيز قابلية تفسير النماذج لفهم أعمق للعوامل المؤثرة في نجاح الأفلام.

## IX. مراجع


1. *Long–Range Movie Box Office Prediction Based on Machine Learning*, Highlights in Science, Engineering and Technology, *vol. 92, pp. 309–310, 2024. [Online]. Available: https://drpress.org/ojs/index.php/HSET/article/view/19900/19480*
2. *V. Udandarao and P. Gupta, "Movie Revenue Prediction using Machine Learning Models,"* arXiv preprint arXiv:2405.11651, *May 2024. [Online]. Available: https://arxiv.org/abs/2405.11651*
3. *M. Agarwal, S. Venugopal, R. Kashyap, and R. Bharathi, "A Comprehensive Study on Various Statistical Techniques for Prediction of Movie Success,"* arXiv preprint arXiv:2112.00395, *Dec. 2021. [Online]. Available: https://arxiv.org/abs/2112.00395*
4. *Y. Zheng, "Predicting Movie Box Office Based on Machine Learning, Deep Learning, and Statistical Methods,"* Applied and Computational Engineering, *vol. 94, pp. 20–32, 2024. [Online]. Available: https://ewadirect.com/proceedings/ACE/article/view/16032*
5. *A. Pawar, "Movies Daily Update Dataset," Kaggle. [Online]. Available: https://www.kaggle.com/datasets/akshaypawar7/millions-of-movies. [Accessed: Jul. 15, 2025].*
6. *IMDb, "IMDb Non-Commercial Datasets," updated Mar. 18, 2024. [Online]. Available: https://developer.imdb.com/non-commercial-datasets/. [Accessed: Jul. 15, 2025].*
7. *unanimad, "The Oscar Award, 1927 - 2025," Kaggle, updated Apr. 2025. [Online]. Available: https://www.kaggle.com/datasets/unanimad/the-oscar-award. [Accessed: Jul. 15, 2025].*
8. *Ö. F. Eker, "Forbes Celebrity 100 since 2005," Kaggle, updated Jun. 9, 2020. [Online]. Available: https://www.kaggle.com/datasets/slayomer/forbes-celebrity-100-since-2005. [Accessed: Jul. 15, 2025].*