

الجمهورية العربية السورية وزارة التعليم العالي والبحث العلمي الجامعة الافتراضية السورية ماجستير التأهيل والتخصص في علوم الحاسوب تقانات تعلم الآلة

Machine Learning

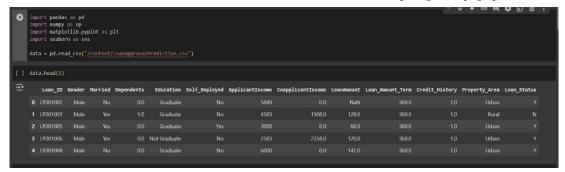
أستاذ ومنسق المقرر: د.عصام سلمان

إعداد الطلاب:

abdulrahman_290065 rand_293968 julia_298202 في إطار سعى المؤسسات المالية والبنوك لتحسين وتبسيط عملية اتخاذ القرارات المتعلقة بالموافقة أو الرفض لطلبات القروض، يهدف هذا المشروع إلى تطوير منصة ويب تدعم اتخاذ هذه القرارات بشكل سريع وفعّال من خلال تحليل البيانات واستخدام تقنيات تعلم الآلة. سيشتمل الموقع على مجموعة من الميزات الأساسية مثل إضافة وإزالة الطلبات، بالإضافة إلى عرض تحليلات استكشافية للبيانات و حل المشكلات المتعلقة بجو دة البيانات.

تتمثل النقطة الرئيسية في هذا المشروع في تطوير نموذج ذكاء اصطناعي باستخدام تقنيات تعلم الآلة بهدف تدريب نموذج قوي وفعّال على مجموعة بيانات محددة. يتضمن هذا المشروع تصميم وتنفيذ عملية تدريب منظم تبدأ من استيراد البيانات ومعالجتها، ثم بناء النموذج وتدريبه، وأخيرًا تصميم موقع ويب يتضمن API للتنبؤ بحالة القروض ، إدارة الطلبات ،وعرض تقارير تحليلية إذاً الخطوة الأُولي هي تحديد الخوار زمية الأفضل لبناء النموذج باستخدام مكتباتPython المتخصصة في تعلم الآلة.

- 1) التحضير المبدئي للبيانات : i استيراد المكتبات الأساسية لتحليل البيانات .i
 - قراءة البيانات .ii
 - عرض أول خمس صفوف .iii



- تحويل المتغيرات الفئوية إلى عددية: في هذه المرحلة، قمنا بمعالجة المتغيرات الفئوية الموجودة في مجموعة البيانات لتحويلها إلى قيم عددية قابلة للاستخدام من قبل نماذج تعلم الآلة.
 - a. تحديد الأعمدة الفئوية.
 - b. تطبیق LabelEncoder على كل عمود فئوى.
 - c. إعادة التحقق من عدد الأعمدة الفئوية المتبقية.

بعد التحويل، يتم التحقق من عدم وجود أعمدة فئوية متبقية تهيئة البيانات لتصبح جاهزة للنماذج الإحصائية والخوار زميات التي لا تدعم القيم النصية مباشرة.

```
[ ] # Import label encoder
    from sklearn import preprocessing
    label_encoder = preprocessing.LabelEncoder()
    obj = (data.dtypes == 'object')
    for col in list(obj[obj].index):
      data[col] = label_encoder.fit_transform(data[col])
    # datatype==object
    obj = (data.dtypes == 'object')
    print("Categorical variables:",len(list(obj[obj].index)))
```

2) تقسيم البيانات إلى تدريب واختبار:

بعد تجهيز البيانات وتحويل القيم الفئوية إلى قيم عددية، قمنا بما يلي:

- . فصل الميزات عن المتغير المستهدف.
- ii. تقسيم البيانات إلى مجموعة تدريب واختبار.
 - iii. التحقق من الأشكال النهائية للمجموعات.

الهدف هو تمكين النماذج من التعلم على جزء من البيانات (التدريب)، ثم اختبار مدى قدرتها على التعميم باستخدام جزء لم تره من قبل (الاختبار).

3) تقييم أولى للنماذج على بيانات التدريب:

بعد تجهيز البيانات وتجزئتها، قمنا بتطبيق أربعة نماذج تصنيف مختلفة لتقييم أدائها على مجموعة التدريب فقط. الهدف من هذه الخطوة هو الحصول على نظرة أولية حول قدرة النماذج على التعلم من البيانات، وليس بالضرورة الحكم على فعاليتها النهائية، النماذج المستخدمة:

- a : Random Forest Classifier .a
 - . 3 = بعدد جبران: K-Nearest Neighbors (KNN) .b
 - .c Support Vector Classifier (SVC) .c
 - . Logistic Regression (LR) : أيضاً تم استخدام الإعدادات الافتر اضية

تم طباعة نسبة الدقة لكل نموذج على بيانات التدريب، وكانت النتائج كما يلي:

```
[ ] from sklearn.neighbors import KNeighborsClassifier
    from sklearn.ensemble import RandomForestClassifier
    from sklearn.svm import SVC
    from sklearn.linear_model import LogisticRegression
    from sklearn import metrics
    knn = KNeighborsClassifier(n_neighbors=3)
    rfc = RandomForestClassifier(n_estimators = 7,
                                   criterion = 'entropy',
                                   random state =7)
    svc = SVC()
    lc = LogisticRegression()
    # making predictions on the training set
    for clf in (rfc, knn, svc,lc):
        clf.fit(X_train, Y_train)
        Y_pred = clf.predict(X_train)
         print("Accuracy score of ",
               clf.__class__.__name__,
               "=",100*metrics.accuracy_score(Y_train,
Accuracy score of RandomForestClassifier = 98.04469273743017
    Accuracy score of KNeighborsClassifier = 78.49162011173185
    Accuracy score of SVC = 68.71508379888269
Accuracy score of LogisticRegression = 79.60893854748603
```

4) التقييم الفعلى للنماذج على مجموعة الاختبار:

في هذه المرحلة، قمنا بتقييم أداء النماذج الأربعة التي تم تدريبها مسبقًا على مجموعة الاختبار هذا التقييم يمثل المؤشر الحقيقي لمدى قدرة كل نموذج على التعميم والتعامل مع بيانات جديدة غير مرئية.

```
[ ] # making predictions on the testing set
    for clf in (rfc, knn, svc,lc):
       clf.fit(X train, Y train)
       Y pred = clf.predict(X test)
       print("Accuracy score of ",
            clf.__class__.__name__,
            100*metrics.accuracy_score(Y_test,
Accuracy score of RandomForestClassifier = 82.5
    Accuracy score of SVC = 69.166666666666666
    Accuracy score of LogisticRegression = 80.833333333333333
```

5) تحليل النتائج:

من خلال نتائج التجربة، تبين أن خوارزمية Random Forest Classifier قدمت أداءً متميزًا مقارنة ببقية الخوارزميات المستخدمة في تصنيف البيانات. يتجسد ذلك في قدرتها على تحقيق توازن جيد بين التعلم والتعميم، مما يجعلها من الخيارات المثلي في المشكلات التي تتطلب تصنيفًا دقيقًا على مجموعات بيانات متنوعة.

- دقة النموذج على مجموعة التدريب: حقق النموذج دقة تصل إلى 99%، مما يشير إلى قدرة الخوارزمية على التعلم بشكل فعّال من البيانات المدخلة
- دقة النموذج على مجموعة الاختبار :حيث أن دقة النموذج على مجموعة الاختبار كانت 82.5%، تعتبر هذه النسبة مرتفعة ،وتعكس قدرة الخوار زمية على تعميم المعرفة المستخلصة من بيانات التدريب.

بناءً على هذا التحليل، سيتم اختيار خوارزمية Random Forest كالنموذج الأمثل لتنبؤ إمكانية قبول القرض، بناءً على مجموعة من المتغيرات (الخصائص) المتعلقة بالمتقدمين. ستشمل هذه العملية التعامل مع البيانات المفقودة، تطبيق تقنيات معالجة القيم الشاذة، ثم ترميز البيانات الفئوية، وذلك لتدريب النموذج على تصنيف المتقدّمين إلى فئتين رئيسيتين:

- قرض معتمد (Loan Approved).
- قرض مرفوض (Loan Rejected).

خ تحليل البيانات وتدريب نموذج التنبؤ بالقروض باستخدام خوارزمية (train_model.py) Random Forest):

- استيراد المكتبات:
 pandas: لإدارة البيانات وتحليلها.
- numpy: لتنفيذ العمليات الرياضية المتقدمة.
 - os: لإدارة المسارات والملفات.
- Joblib : لحفظ النماذج المدربة لتستخدمها لاحقًا.
- sklearn.model_selection: لاستخدام دالة train_test_split:
- sklearn.preprocessing: لاستخدام LabelEncoder لتحويل المتغيرات الفئوية إلى أرقام.
- sklearn.ensemble: لاستخدام نموذج RandomForestClassifier (غابة عشوائية)

إعداد مسارات البيانات والنماذج: يتم تحديد مسارات الملفات التي تحتوي على البيانات والنموذج المُدرَّب باستخدام os.path.join .

- (3) <u>تحميل وتنظيف البيانات:</u>
 تحميل البيانات: يتم تحميل ملف loan_prediction.csv الذي يحتوي على بيانات القروض باستخدام . pandas.read csv()
 - التعامل مع القيم المفقودة :يتم ملء القيم المفقودة في الأعمدة المختلفة باستخدام:
 - القيم الأكثر تكرارًا في الأعمدة الفئوية((mode).
 - القيمة الافتراضية (مثل 360 في Loan_Amount_Term) ، أو القيمة المتوسطة (مثل (LoanAmount

إزالة القيم الشاذة باستخدام IQR: يتم استخدام مقياس المدى بين الرباعيات (Interquartile Range - IQR) الكتشاف القيم الشاذة في بعض الأعمدة مثل: LoanAmount, ApplicantIncome, CoapplicantIncome؛ و استبدال القيم الشاذة بالقيم المتو سطة لتلك الأعمدة.

- للتحضير والترميز:
 إزالة الأعمدة غير الضرورية :مثل عمود Loan_ID لأنه غير مفيد في عملية التدريب.
 - معالجة الأعمدة الفئوية:
- يتم استبدال القيم غير الرقمية في العمود Dependents (مثل "3+") بالقيمة الرقمية 3.
- يتم ترميز الأعمدة الفئوية (Gender, Married, Education, Self_Employed) يتم ترميز الأعمدة الفئوية Property_Area, Loan_Status) باستخدام LabelEncoder لتحويلها إلى قيم عددية لتناسب النموذج.

- 5) تقسيم البيانات وتدريب النموذج:
 تقسيم البيانات يتم تقسيم البيانات إلى مدخلات (x) وأهداف (y) ، حيث أن x يحتوي على الخصائص مثل الدخل والحالة الاجتماعية، بينما يحتوي y على القيمة المستهدفة و هي حالة القرض(Loan_Status) .
- تقسيم البيانات إلى تدريب واختبار :باستخدام train_test_split ، يتم تقسيم البيانات إلى مجموعة تدريب (80%)
 - معالجة عُدم التوازن: يتم حساب الأوزان العكسية لكل فئة بناءً على تواترها في مجموعة التدريب لتقليل تأثير
 - تدريب النموذج : يتم تدريب نموذج الغابة العشوائية باستخدام البيانات المدربة (X_train, Y_train) .
- حفظ النموذج والمُشفرات :بعد تدريب النموذج، يتم حفظه مع المُشفرات باستخدام مكتبة joblib لنتمكن من استخدام النموذج لاحقًا. يتم حفظ النموذج في ملف (model.pkl)، والمشفرات في ملف (encoders.pkl) .

6) <u>تنفیذ العملیة:</u>

 عند تنفیذ الکود، یتم استدعاء دالة ()train_and_save_model ، والتي تقوم بجمیع الخطوات السابقة من تحمیل

7) النتيجة النهائية:

يتم طباعة رسالة تأكيد بعد حفظ النموذج والمُشفّرات بنجاح :النموذج والمشفرات تم حفظها بنجاح ٧٠.

وبذلك قد أصبح النموذج جاهزًا للاستخدام في تنبؤ إمكانية قبول أو رفض القروض بناءً على خصائص المتقدمين، مما يسهم في تحسين عملية أتخاذ القرار داخل المؤسسة المالية.

❖ تصميم وبناء الموقع الالكتروني Trust Bank:

◄ الواجهة الأمامية Frontend: (التقنيات المستخدمة: HTML, CSS, JavaScript)

			Bank		
	Home	Services	About Us	Contact	
Add Loan	Request			Exploraty Data Analysis	
Loan Requ	ests Table			Model Metrics	
		Data I	ssues		

• الصفحة الرئيسية (Home Page):

الصفحة الرئيسية لموقع TrustBank هي الواجهة الأولى ، تحتوي على المكونات الأساسية التي تسهل التنقل داخل الموقع ، و هي عبارة عن خمسة بطاقات رئيسية ، كل بطاقة تحتوي على رابط إلى صفحة مختلفة في الموقع:

- Add Loan Request إضافة طلب قرض
- جدول طلبات القروض Loan Requests Table
- Exploraty Data Analysis تحليل البيانات الاستكشافي
 - معالجة البيانات Data Issues Report
 - مقاییس النموذج Model Metrics

: (Add Loan Request) صفحة طلب القرض

تُعد صفحة "طلب القرض" في موقع Trust Bank واحدة من الصفحات الحيوية التي تتيح تقديم طلبات القروض ، تتضمن النموذج الذي يتطلب تعبئة مجموعة من البيانات التحديد أهلية المتقدم للحصول على القرض . من خلال الضغط على زر "Predict" ، يتم عرض النتيجة التي تحدد ما إذا كان القرض مقبولاً أم مرفوضاً استناداً إلى نموذج تعلم الآلة المُستخدم كما يتم عرض احتمالية الموافقة كنسبة منوية ، وبالضغط على زر الإرسال "Submit" يتم إرسال البيانات لتخزين طلب القرض في قاعدة البيانات ، وفي حال وجود خطأ في البيانات المُدخلة ، سيتم عرض رسالة تحذير تطلب من المستخدم تعبئة كافة الحقول .

• صفحة جدول طلبات القروض (Loan Requests Table):

تُعتبر جزءًا أساسيًا لعرض جميع طلبات القروض التي تم تقديمها عبر الموقع. توفر الصفحة واجهة تفاعلية لعرض تفاصيل كل طلب، وتتيح للمسؤولين إمكانية حذف أي طلب قرض من النظام إذا لزم الأمر.

ضفحة التحليل الاستكشافي (Exploraty Data Analysis) :

تعتمد الصفحة على جلب ملخص البيانات من الواجهة البرمجية الخارجية، وتعرض إحصائيات مثل عدد الطلبات ومتوسط القروض والدخل ونسبة الموافقة. كما تتضمن رسومًا بيانية توضيحية باستخدام Chart.js، مدعومة بمكتبة -chartjs للقروض والدخل yplugin-datalabels

• صفحة تقرير معالجة البيانات (Data Issues Report):

تعرض هذه الصفحة تقريرًا عن مشاكل جودة البيانات مثل القيم المفقودة والقيم الشاذة، يتم جلبها من الواجهة البرمجية الخارجية API ، تُعرض المشاكل في جداول، ويتبعها قسم يوضح طرق المعالجة المستخدمة مثل التعويض أو الحذف في النهاية ، كما تُعرض ملاحظات ختامية حول معالجة البيانات وجودتها.

• صفحة مقاييس أداء نموذج تعلم الآلة المستخدم (Model Metrics):

تعرض صفحة Model Metrics أداء نموذج تعلم الآلة من خلال أربعة مؤشرات رئيسية :الدقة (Accuracy) ، الدقة الإيجابية (Precision) ، الاسترجاع (Recall) ، ودرجة F1 Score) 1 .

يتم جلب هذه القيم تلقائيًا من الواجهة البرمجية الخارجية عند فتح الصفحة، وتُعرض بشكل مبسط وواضح لتقييم جودة النموذج.

الواجهة الخلفية Backend والربط مع نموذج التنبؤ:

تم بناء الواجهة البرمجية باستخدام FastAPI لإدارة طلبات القروض وربطها بنموذج التنبؤ، فيما يلي شرح مختصر مع ذكر المكتبات والدوال الأساسية لكل مرحلة:

1. تصميم البيانات و API:

- o المكتبة: Pydantic
- الدالة الأساسية BaseModel : لتصميم نموذج البيانات.
- o تم تعريف نموذج بيانات موحد للتحقق من صحة البيانات المدخلة وتسهيل التواصل عبر نقاط نهاية API (Endpoints)باستخدام FastAPI .

تحميل النموذج والمشفرات:

- o المكتبة :ioblib
- م الدالة الأساسية (joblib.load : لتحميل النموذج المُدرب و المشفّرات مثل joblib.load :
 - يتم تحميل النموذج المدرب والمشفرات لتحويل القيم النصية إلى تمثيلات رقمية.

3. معالجة البيانات:

- o المكتبة : scikit-learn , Pandas
- o الدالة الأساسية ()fillna : لتعويض القيم المفقودة، ()LabelEncoder لتحويل القيم النصية، و ()drop لحذف الأعمدة غير المفيدة.
 - تتم معالجة البيانات لتعويض القيم المفقودة وتحويل النصوص إلى أرقام، وحذف الأعمدة غير المهمة.

4. التحليل الاستكشافي للبيانات (EDA):

- Matplotlib Seaborn المكتبة:
- الدالة الأساسية :(plt.plot(), sns.histplot : لإنشاء الرسوم البيانية التحليلية.
- يتم عرض تحليلات مرئية لتوزيع البيانات مثل التنبؤات، الجنس، الدخل، والمنطقة السكنية.

5. التنبؤ بحالة القرض:

- scikit-learn المكتبة
- الدالة الأساسية ()model.predict : لإجراء التنبؤ باستخدام النموذج المدرب.
- يتم تمرير البيانات للنموذج المدرب لتحديد ما إذا كان الطلب سيُقبل أو يُرفض مع تحديد نسبة الاحتمالية.

6. تقييم النموذج:

- scikit-learn: المكتبة
- الدالة الأساسية: (accuracy_score(), classification_report لتقبيم دقة النموذج باستخدام مقاييس مثل الدقة و الاسترجاع و F1.
 - يتم تحليل نتائج التنبؤات لتقييم جودة أداء النموذج.

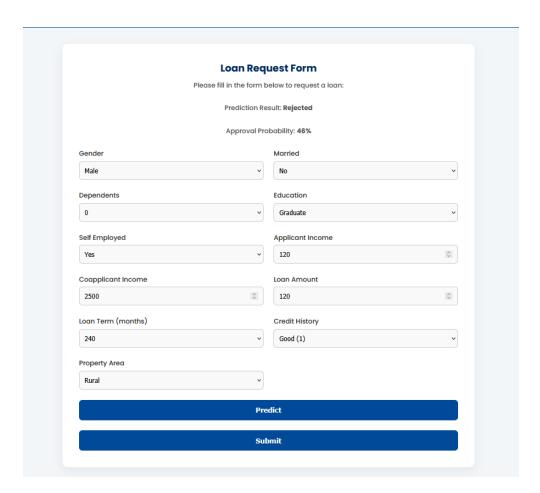
7. تكامل النظام:

- o المكتبة :FastAPI
- الدالة الأساسية: (app.post(), @app.get لتعريف نقاط النهاية الخاصة بالـAPI .
 - يتم تنظيم النظام في وحدات مستقلة تُستدعى عبر APIمما يسهل الصيانة والتطوير.
- https://mlt.vilmo.app/docs#/: يمكن الأطلاع على الواجهة البرمجية للموقع على الرابط التالي://https://mlt.vilmo.app

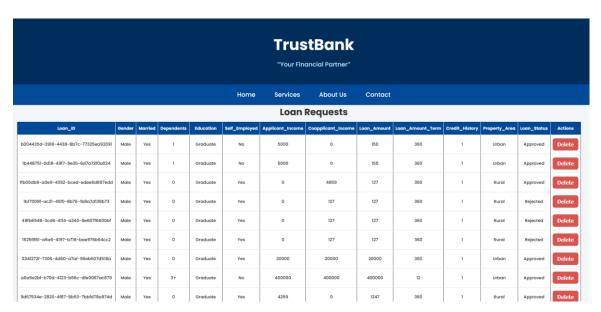
وبذلك أصبح الموقع يدير ويحلل طلبات القروض باستخدام Pandas, scikit-learn, FastAPI، ويقدم التنبؤات مع تحليلات مرئية باستخدام Seaborn, Matplotlib، مع ضمان دقة الأداء وجودة النتائج.

◄ الاختبار والنتائج:

• إضافة طلب قرض وإظهار نتيجة التنبؤ: حيث قمنا إدخال بيانات افتراضية ثم الضغط على زر Predict ، وكانت النتيجة على الشكل التالى:

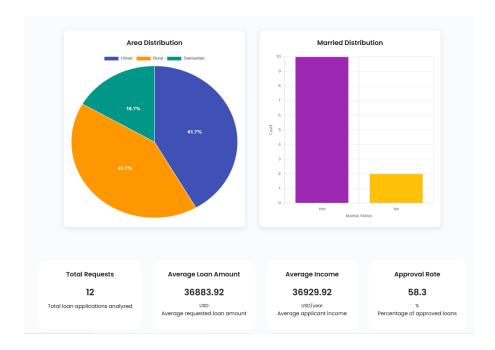


• عرض جدول الطلبات مع إمكانية حذف الطلب: حيث قمنا بإضافة 12 طلب إلى قاعدة البيانات ، وذلك بإدخال البيانات والضغط على زر Submit .

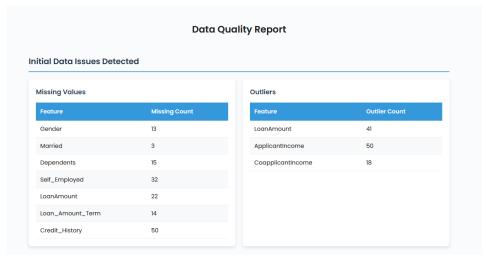


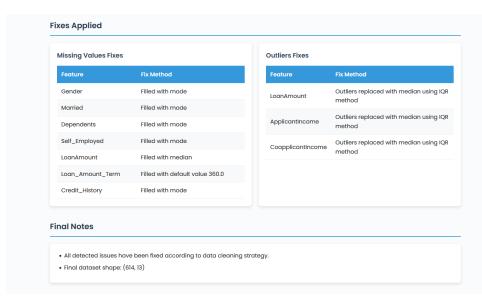
تحليل البيانات الاستكشافي للطلبات المُدخلة: حيث تعرض هذه الصفحة بعض التحليلات الاستنتاجية للطلبات المدخلة إلى قاعدة البيانات ، مثل : معدل القبول والرفض ، عدد الذكور والإناث المتقدمين ، نسبة توزع مناطق المتقدمين جغرافياً ، الحالة الاجتماعية ، كما تعرض عدد الطلبات ومعدل دخل المتقدم وحجم القرض . و بناءً على الطلبات التي قمنا بإدخالها ظهرت النتائج التالية :





• صفحة التقرير الخاصة بمعالجة البيانات الشاذة والمفقودة: تعرض هذه الصفحة تقريراً يوضح المشكلات التي واجهتنا أثناء تحضير البيانات (قيم مفقودة وقيم شاذة) عند تدريب النموذج، كما توضح كيفية معالجة هذه المشكلات.





• صفحة عرض مقاييس أداء النموذج: تعرض هذه الصفحة مقاييس أداء النموذج المُدرّب المستخدم في عملية التنبؤ . بقبول أو رفض طلب المقترض .

حيث كانت الدقة Accuracy (النسبة المئوية للعينات التي صنفت بشكل صحيح) مساوية %76 ، كما أعطى نسبة دقة إيجابية Precision (من بين كل ما تم تصنيفه على انه "إيجابي" كم كان صحيحاً فعلاً ؟) مساوية %75, أما بالنسبة للاسترجاع Recall (من بين كل الإيجابيات الحقيقية في البيانات ، كم منها تم اكتشافه ؟) فأعطى نسبة %94.

مقياس $\mathbf{F1}$ (المتوسط التوافقي بين Precision و Recall) يساوي

Model Performance Metrics

_	
Accuracy:	0.76
Precision:	0.75
Recall:	0.94
F1 Score:	0.83

◄ الخاتمة :

في الختام يمكننا التأكد على أن تقنيات تعلم الآلة أداة أساسية للمؤسسات المالية لتحسين دقة قرار اتها وتقليل المخاطر. لأنها تساهم في تعزيز الكفاءة التشغيلية وتوفير الوقت، مما يعزز قدرة المؤسسة على التكيف والنمو في بيئة اقتصادية ديناميكية.