دراسة مقارنة بين خوارزمية الانحدار الخطي وخوارزميتي منحدر شجرة القرار ومنحدر الغابة العشوائية في التنبؤ بمبيعات الأدوية

A Comparative study between Linear regression, Decision tree and random forest Algorithms in the Prediction of medicine sales

سامية محمد الحسن ابراهيم
أستاذ محاضر
جامعة المغتربين
samia.mohamed1234@gmil.com

أ.د احمد صلاح الدين عبد الله
بروفيسور
جامعة السودان المفتوحة
asalahous @gmil.com

الخرطوم ، السودان

المستخلص:

يوجد العديد من الخوارزميات لكل منها ميزات وعيوب وبالتالي قد يجد الباحث صعوبة في تحديد أي الخوارزميات أنسب لنوع الدراسة المعين، والذي هو مبيعات الأدوية على سبيل المثال الذي تناولته هذه الورقة، فالحل الأمثل لمشكلة صعوبة الاختيار هو التطبيق والاختبار المباشر للخوارزميات التي تم ترشيحها لبناء نموذج التنبؤ، ويتم هذا الاختبار باستخدام مقاييس معينة تسمى مقايس الانحدار.التنقيب في البيانات مجال واسع يتضمن العديد من التقنيات، وهذه التقنيات يتم اختيارها حسب نوع المعلومات المراد استخراجها من مجموعة البيانات، وفي هذه الورقة استخدمت بيانات الأدوية للتنبؤ بالمبيعات، مما تطلب استخدام تقنيات الانحدار في التنقيب في البيانات. الهدف من هذه الورقة هو مقارنة بين ثلاث من أشهر خوارزميات الانحدار وهي خوارزميات الانحدار

الخطي، منحدر شجرة القرار ومنحدر الغابة العشوائية لايجاد الخوارزمية التي تعطي أفضل نتائج في التنبؤ بمبيعات الأدوية حسب الدقة ومعدل الخطأ وفق مقاييس الانحدار. تم أخذ البيانات من شركة بينا للأدوية في مدينة الخرطوم وشملت عدد 2700 سجل تمثل بيانات ثلاث سنوات (2017–2018–2019) تمت معالجتها للتنبؤ بمبيعات 2020 ومقارنتها مع المبيعات الحقيقية لعام 2020 لتحديد دقتها. وفي هذه الورقة أفضت نتيجة الاختبار والتقييم الى اختيار خوارزمية منحدر الغابة العشوائية في بناء نموذج التنبؤ الانحداري وفق معطيات محددة لأنها كانت الأفضل بين الخوارزميات المختبرة، وفق المقاييس المعينة وبنسبة خطأ لا تتجاوز **0.09%** ، أما بالنسبة لتنبؤ السلاسل الزمنية أظهرت نتائج التقييم والاختبار أن خوارزمية الانحدار الخطي هي التي أظهرت أفضل النتائج بالمقارنة مع الخوارزميات الأخرى التي تم اختبارها وبنسبة خطأ لا تتجاوز **0.34%** وفق مقاييس الانحدار المعينة.

الكلمات المفتاحية: تنقيب البيانات، مبيعات الأدوية، المقارنة، تقنيات الانحدار.

Abstract:

Data mining is a broad field that includes many techniques, and these techniques are selected according to the type of information to be extracted from the data set. This paper deals with drug data for sales forecasting, which required the use of regression techniques in data mining. There are many algorithms, each with its own advantages and disadvantages, and therefore the researcher may find it difficult to determine which algorithms are most suitable for the particular type of study, which is drug sales, for example, which was dealt with in this paper. This test is done using specific scales called regression metrics. The aim of this paper is to compare three of the most famous regression algorithms which are linear regression algorithms, decision tree slope and random forest slope to find the algorithm that gives the best results in predicting drug sales according to accuracy and error rate according to regression measures. The data was taken from Bina Pharmaceutical Company in Khartoum city and included 2,700 records representing

three years (2017-2018-2019) data that were processed to forecast sales of 2020 and compared with the real sales of 2020 to determine its accuracy. In this paper, the result of the test and evaluation led to the selection of the random forest regressor algorithm in building a regression prediction model according to specific data inputs because it was the best among the tested algorithms, with not more than **0.09% error** according to the specific metrics compared to other tested regression algorithms. As for the prediction of time series, the results of evaluation and testing showed that the linear regression algorithm showed the best results compared to other algorithms tested according to regression metrics with not more than **0.34% error**.

Keywords: data mining, drug sales, comparison, regression techniques.

1.المقدمة:

التنقيب في البيانات هو عملية تحليل البيانات لتحديد العلاقات التي لم تكتشفها التحليلات السابقة من قبل، كما أنها تحليل البيانات لإقامة علاقات وتحديد أنماط [1]. فاليوم لا تجد مجالاً إلا استخدم فيه التنقيب في البيانات، فاختيار تقنية التنقيب يتطلب معرفة نوع المعلومات التي يراد استخراجها في حالة تعلم الآلة المشرف عليه، ومن ثم يختلف أداء كل خوارزمية باختلاف عدد المدخلات والعلاقة بينها ونوع البيانات المراد التنقيب فيها مما يؤدي لزيادة أو نقصان دقة نفس الخوارزمية في البيانات المختلفة، الأمر الذي أدى لضرورة اختبار عدة خوارزميات والمقارنة بينها واختيار الخوارزمية الأفضل التي تتناسب ونوع البيانات وتعطي أدق النتائج وأقل نسبة خطأ. و عدم القيام بعملية التقييم والاختبار والمقارنة قد يؤدي الى نقصان في دقة النموذج مما يؤثر سلباً على جودة النتائج المطلوبة وبالتالي لا يتم الحصول على الدعم المناسب في اتخاذ القرار (Decision support) وهو أحد أهم الأهداف للتنقيب في البيانات، لذا كان من الضوري اختبار وتقييم نماذج الخوارزميات (model evaluation) وفي هذه الورقة العلمية تم استخدام بيانات مبيعات الأدوية واستخدام تقنيات التنقيب في البيانات للتنبؤ بالمبيعات حسب معطيات محددة وأيضاً التنبؤ بالمبيعات المستقبلية في صورة سلاسل زمنية. والمبيعات هي عبارة عن قيم مستمرة (continuous values) مما يتطلب استخدام تقنيات الانحدار

(Regression techniques) المتخصصة في هذا النوع من المخرجات وهي من تقنيات تعليم الآلة المشرف عليه وتتضمن العديد من الخوارزميات، وتحليل الانحدار هو طريقة تحليل إحصائي لتحديد العلاقة الكمية بين متغيرين أو أكثر [2] لذلك يتم استخدامها في التنبؤ بالقيم المستمرة، وتقنيات الانحدار يمكن استخدامها أيضاً في تنبؤ السلاسل الزمنية، حيث تعرف السلسلة الزمنية بأنها مجموعة القياسات المسجلة لمتغير واحد أو أكثر مرتبة حسب زمن وقوعها [3]، وهذا المبدأ يمكن تطبيقه على مبيعات الأدوية التي هي مخرجات هذه البيانات. الهدف من هذه الورقة هو مقارنة بين ثلاث من أشهر خوارزميات الانحدار وهي خوارزميات الانحدار الخطي، منحدر شجرة القرار ومنحدر الغابة العشوائية لايجاد الخوارزمية التي تعطي أفضل نتائج في التنبؤ بمبيعات الأدوية حسب الدقة ومعدل الخطأ وفق مقاييس الانحدار، وعناصر المقارنة بين الخوارزميات المستخدمة هي متوسط الخطأ المطلق، متوسط الخطأ التربيعي، نتيجة التباين الموضح ومعامل التحديد.

3.المنهجية:

1.3 البرمجيات التي تم استخدامها في الورقة العلمية:

لغة البرمجة بايثون:python الإصدار المستخدم: python 3.10.2، [10]

مدير البيئة Environment Manager: Anaconda الإصدار المستخدم: Anaconda3 2021.11، [11] جوبيتر نوت بوك :Jupyter notebook الإصدار المستخدم: Jupyter notebook 6.4.5 [12]

2.3 مقاييس الانحدار (Regression metrics):

هي مقاييس تستخدم لتقييم أداء الخوارزميات (model evaluation) بمقارنة نتائج التنبؤ مع القيم الفعلية الاختبارية وقد تم استخدام أربعة من مقاييس الانحدار (Regression metrics) وهي:

1.2.3 متوسط الخطأ المطلق (Mean absolute error):

في الإحصاء، متوسط الخطأ المطلق هو مقياس للأخطاء بين الملاحظات المقترنة التي تعبر عن نفس الظاهرة. تتضمن أمثلة Y مقابل X مقارنات بين القيم المتوقعة مقابل المشاهدة، والوقت اللاحق مقابل الوقت الأولي، وتقنية قياس واحدة مقابل تقنية قياس بديلة.

تقوم وظيفة mean_absolute_error بحساب الخطأ المطلق المتوسط، وهو دالة خطر تقابل القيمة المتوقعة لفقدان الخطأ المطلق. [4]

إذا كانت $\hat{y}_i$ هي القيمة المتوقعة للعينة رقم (i) وy_i هي القيمة الحقيقية المقابلة يمكن تعريف متوسط الخطأ المطلق[4] (MAE) ل (n) عينة على النحو التالي كما في المعادلة (1):

$$\mathrm{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i| . \quad (1)$$

2.2.3 متوسط الخطأ التربيعي (Mean squared error):

في الإحصاء، متوسط الخطأ التربيعي يقيس متوسط مربعات الأخطاء – أي متوسط الفرق التربيعي بين القيم المقدرة والقيمة الفعلية.

تحسب دالة mean_squared_error متوسط الخطأ التربيعي، وهو دالة مخاطرة تقابل القيمة المتوقعة لخسارة الخطأ التربيعي أو الخسارة التربيعية. [5]

إذا كانت $\hat{y}_i$ هي القيمة المتوقعة للعينة رقم (i) وy_i هي القيمة الحقيقية المقابلة يمكن تعريف متوسط الخطأ التربيعي[5] (MSE) ل (n) عينة على النحو التالي كما في المعادلة (2):

$$\mathrm{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 \quad (2)$$

3.2.3 نتيجة التباين الموضح (Explained variance score):

في الإحصاء، يقيس التباين الموضح النسبة التي يمثل بها النموذج الرياضي تباين (تشتت) مجموعة بيانات معينة. [6]

إذا كان $\hat{y}$ هو الناتج المستهدف المقدر وكان y هو الناتج المستهدف (الصحيح) المقابل، فسيتم تقدير التباين الموضح[6] على النحو التالي كما في المعادلة (3):

$$\texttt{explained_variance}(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}} \quad (3)$$

4.2.3 معامل التحديد (R^2 score – Coefficient of determination):

في الإحصاء، يشير معامل التحديد، الذي يرمز له بـ R^2 أو r^2 أو"R squared" ، إلى نسبة التباين في المتغير التابع الذي يمكن التنبؤ به من خلال المتغير (أو المتغيرات) المستقلة .

وهو يستخدم في النماذج الإحصائية التي يكون هدفها الرئيسي التنبؤ بالنتائج المستقبلية أو اختبار الفرضيات، على أساس المعلومات الأخرى ذات الصلة. يوفر معامل التحديد مقياسًا لمدى تكرار النتائج التي تمت ملاحظتها في النموذج، استنادًا إلى نسبة التباين الكلي للنتائج التي أوضحها النموذج.

الدالة r2_score تحسب R^2، معامل التحديد. توفر مقياسًا لمدى جودة توقع العينات المستقبلية بواسطة النموذج. [7] [8] [9]

إذا كانت $\hat{y}_i$ هي القيمة المتوقعة للعينة رقم (i) وy_i هي القيمة الحقيقية المقابلة يمكن تعريف معامل التحديد[7] (R^2) ل (n) عينة على النحو التالي كما في المعادلة رقم (4):

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \hat{y}_i)^2}{\sum_{i=0}^{n_{\text{samples}}-1}(y_i - \bar{y})^2} \quad (4)$$

where $\bar{y} = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} y_i$.

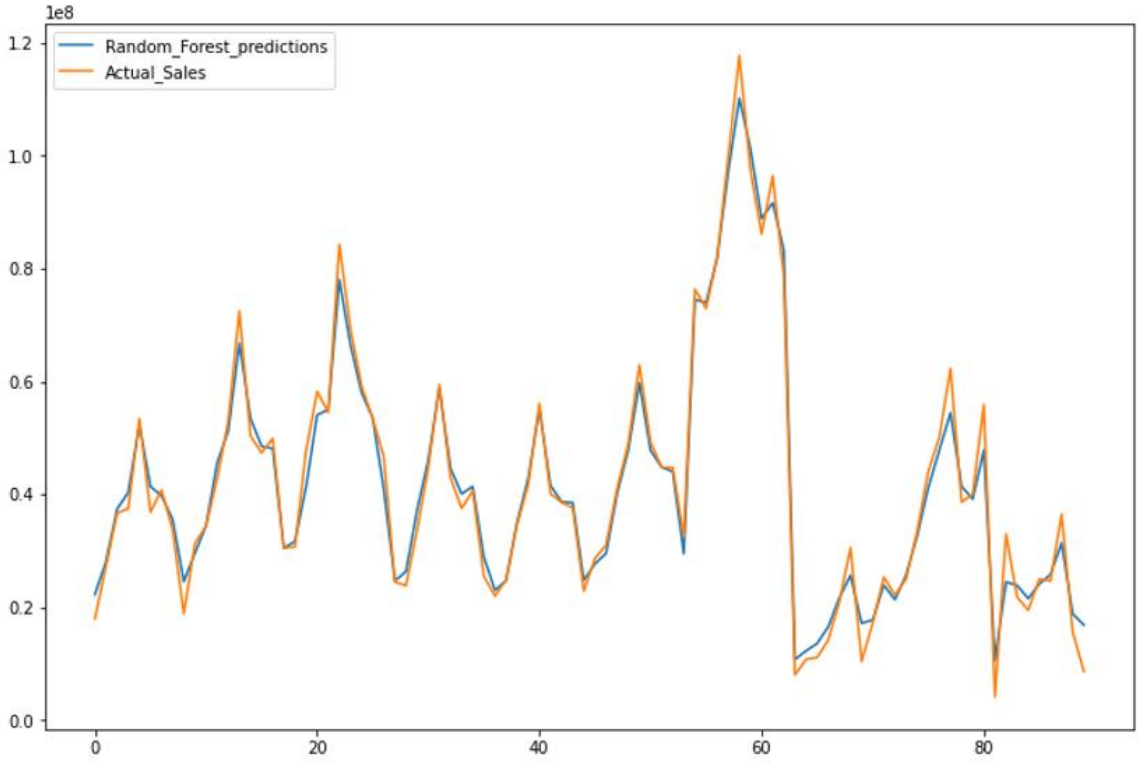
4. التنفيذ والنتائج:

تم اختبار ثلاث من خوارزميات الانحدار وهي خوارزمية الانحدار الخطي (Linear Regression) خوارزمية منحدر شجرة القرار (Decision Tree Regressor) خوارزمية منحدر الغابة العشوائية (Random forest Regressor).

1.4 نتائج خوارزميات نموذج التنبؤ الانحداري وفق معطيات محددة:

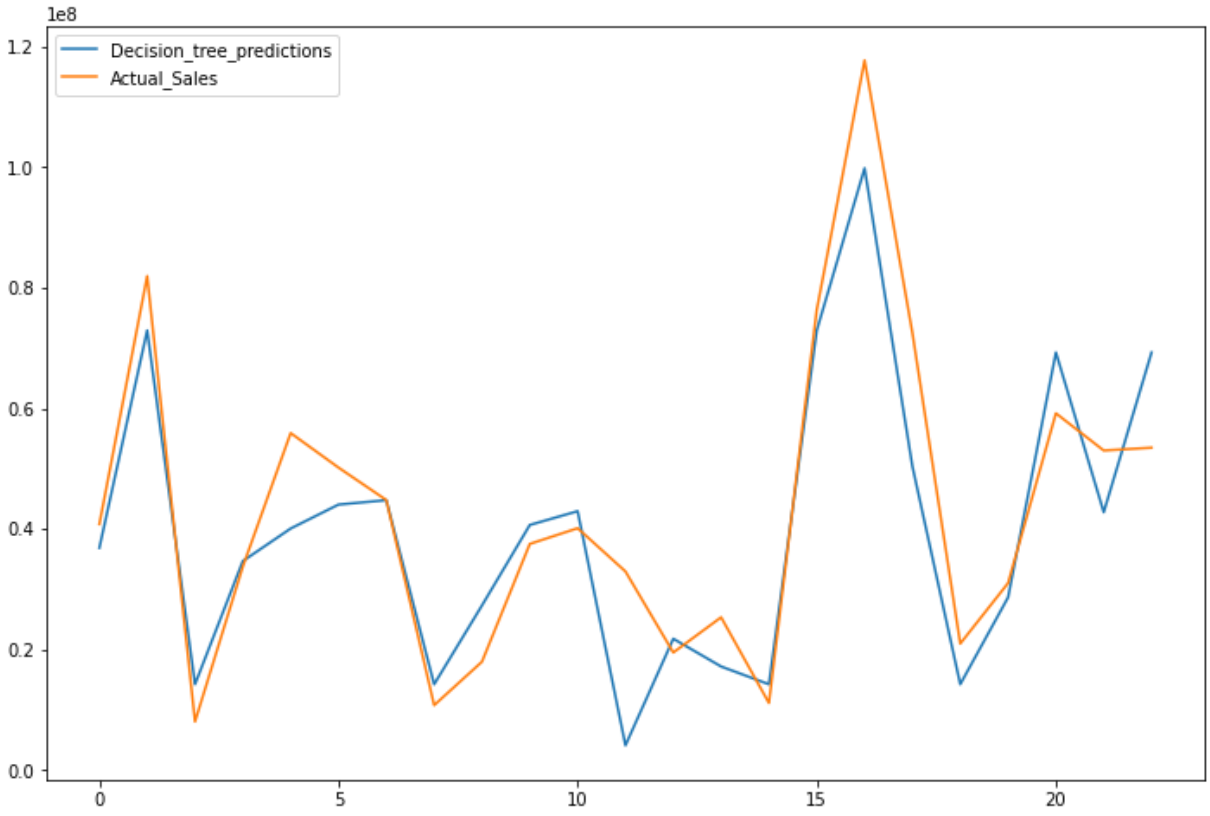
النتائج التالية توضح أداء خوارزميات الانحدار التي تم اختبارها ومقارنة نتائج تنبؤها مع القيم الحقيقية لبناء نموذج التنبؤ الانحداري حسب معطيات محددة وهي اسم الدواء والمدينة.

الشكل رقم (1) التالي يوضح مقارنة نتائج تنبؤ خوارزمية Random Forest Regressor مع القيم الحقيقية الاختبارية بيانياً:



شكل رقم (1) يوضح اختبار خوارزمية منحدر الغابة العشوائية

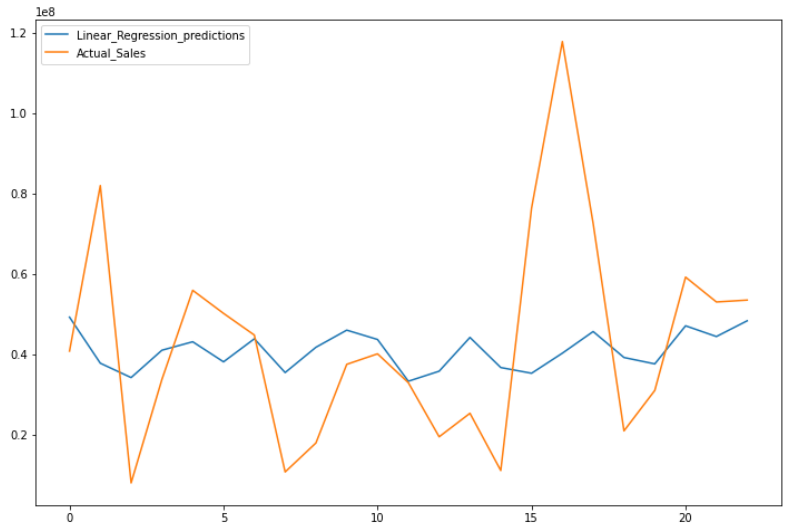
الشكل رقم (2) التالي يوضح مقارنة نتائج تنبؤ خوارزمية Decision Tree Regressor مع القيم الحقيقية الاختبارية بيانياً:



شكل رقم (2) يوضح اختبار خوارزمية منحدر شجرة القرار

الشكل رقم (3) التالي يوضح مقارنة نتائج تنبؤ خوارزمية Linear Regressor مع القيم الحقيقية الاختبارية بيانياً:

```
In [65]: plt.rcParams['figure.figsize'] = (12,8)
         plt.plot(y_lpred,label='Linear_Regression_predictions')
         plt.plot(y_test,label='Actual_Sales')
         plt.legend(loc="upper left")
         plt.show()
```



شكل رقم (3) يوضح اختبار خوارزمية الانحدار الخطي

الجدول رقم (1) التالي يلخص المقارنة بين الخوارزميات من حيث الأداء وفق مقاييس الانحدار المستخدمة:

Coefficient of Determination R^2 Score	Explained Variance Score	Mean Absolute Error	Sqrt Mean Squared Error	مقياس الانحدار / الخوارزمية
0.98	0.98	2,458,127 0.06%	3,468,218 0.09%	Random Forest Regressor
0.82	0.84	8,344,667 0.22%	11,002,306 0.29%	Decision Tree Regressor
0.05	0.06	18,672,903 0.50%	25,223,485 0.67%	Linear Regression

جدول رقم (1) مقارنة بين خوارزميات الانحدار

وتم اختيار الخوارزمية التي أظهرت أفضل أداء من حيث الدقة وأقل نسبة خطأ لاستخدامها وحفظ نموذج التنبؤ الخاص بها وهو نموذج التنبؤ وفقاً لمدخلات محددة وهي المدينة واسم الدواء وفي هذه الحالة الخوارزمية

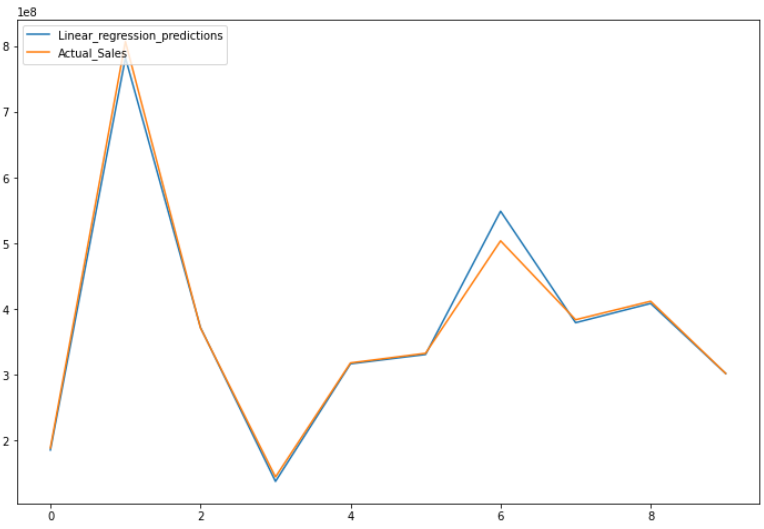
الأفضل أداء هي خوارزمية ال Random Forest Regressor وفقاً لمقاييس الانحدار حيث كانت نسبة متوسط الخطأ المطلق **0.06%** ونسبة الجذر التربيعي لمتوسط الخطأ التربيعي **0.09%** والتباين الموضح **0.98** ومعامل التحديد **0.98** وتليها خوارزمية Decision tree regressor حيث كانت نسبة متوسط الخطأ المطلق **0.22%** ونسبة الجذر التربيعي لمتوسط الخطأ التربيعي **0.29%** والتباين الموضح **0.84** ومعامل التحديد **0.82** وفي الأخير خوارزمية Linear regression حيث كانت نسبة متوسط الخطأ المطلق **0.50%** ونسبة الجذر التربيعي لمتوسط الخطأ التربيعي **0.67%** والتباين الموضح **0.06** ومعامل التحديد **0.05** .

2.4 نتائج نموذج التنبؤ الانحداري للسلاسل الزمنية:

تم تقييم أداء الخوارزميات (model evaluation) بمقارنة نتائج تنبؤ السلاسل الزمنية مع القيم الفعلية باستخدام مقاييس الانحدار المذكورة أعلاه (Regression metrics) والمقارنة البيانية لاختيار الخوارزمية الأفضل أداء لتنبؤ السلاسل الزمنية كما توضح الأشكال التالية نتائج المقارنة:

الشكل رقم (4) التالي يوضح مقارنة نتائج تنبؤ خوارزمية Linear Regressor مع القيم الحقيقية الاختبارية بيانياً:

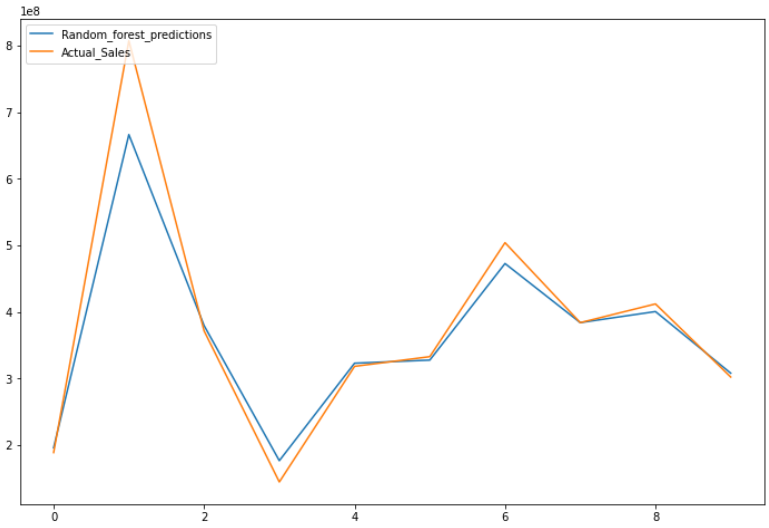
```
In [90]: plt.rcParams['figure.figsize'] = (12,8)
         plt.plot(y_pred1,label='Linear_regression_predictions')
         plt.plot(y,label='Actual_Sales')
         plt.legend(loc="upper left")
         plt.show()
```



شكل رقم (4) اختبار خوارزمية الانحدار الخطي للسلاسل الزمنية

الشكل رقم (5) التالي يوضح مقارنة نتائج تنبؤ خوارزمية Random Forest Regressor مع القيم الحقيقية الاختبارية بيانياً:

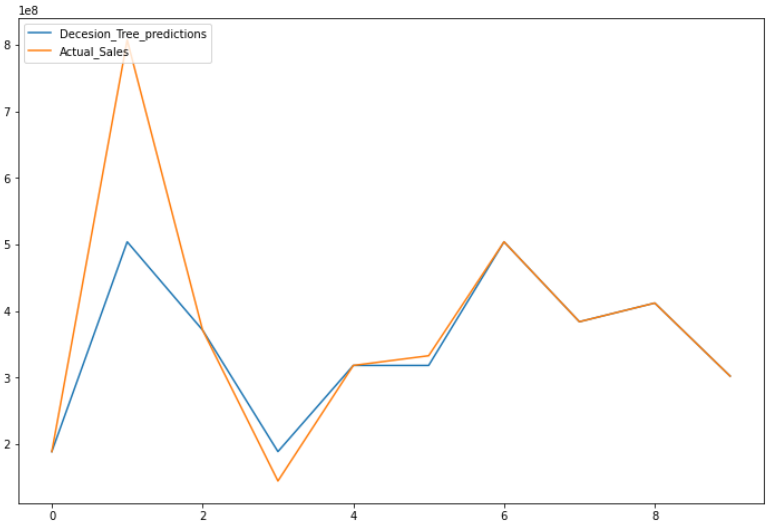
```
In [93]: plt.rcParams['figure.figsize'] = (12,8)
         plt.plot(y_predf,label='Random_forest_predictions')
         plt.plot(y,label='Actual_Sales')
         plt.legend(loc="upper left")
         plt.show()
```



شكل رقم (5) يوضح اختبار خوارزمية الغابة العشوائية للسلاسل الزمنية

الشكل رقم (6) التالي يوضح مقارنة نتائج تنبؤ خوارزمية Decision Tree Regressor مع القيم الحقيقية الاختبارية بيانياً:

```
In [97]: plt.rcParams['figure.figsize'] = (12,8)
         plt.plot(y_predt,label='Decesion_Tree_predictions')
         plt.plot(y,label='Actual_Sales')
         plt.legend(loc="upper left")
         plt.show()
```



شكل رقم (6) يوضح اختبار خوارزمية شجرة القرار للسلاسل الزمنية

الجدول رقم (2) التالي يلخص المقارنة بين الخوارزميات من حيث الأداء وفق مقاييس الانحدار المستخدمة:

Coefficient of Determination R^2 Score	Explained Variance Score	Mean Absolute Error	Sqrt Mean Squared Error	مقياس الانحدار / الخوارزمية
0.99	0.99	9,171,076 0.20%	16,545,918 0.34%	Linear Regression
0.93	0.93	24,661,087 0.51%	47,090,596 0.98%	Random Forest Regressor
0.69	0.71	36,203,395 0.75%	97,022,938 2.02%	Decision Tree Regressor

جدول رقم (2) مقارنة بين خوارزميات السلاسل الزمنية

وتم اختيار الخوارزمية التي أظهرت أفضل أداء من حيث الدقة وأقل نسبة خطأ لاستخدامها وحفظ نموذج التنبؤ للسلاسل الزمنية الخاص بها وفي هذه الحالة هي خوارزمية ال Linear Regression وفقاً لمقاييس الانحدار

حيث كانت نسبة متوسط الخطأ المطلق **0.20%** ونسبة الجذر التربيعي لمتوسط الخطأ التربيعي **0.34%** والتباين الموضح **0.99** ومعامل التحديد **0.99** وتليها خوارزمية Random forest regressor حيث كانت نسبة متوسط الخطأ المطلق **0.51%** ونسبة الجذر التربيعي لمتوسط الخطأ التربيعي **0.98%** والتباين الموضح **0.93** ومعامل التحديد **0.93** وفي الأخير خوارزمية Decision tree regressor حيث كانت نسبة متوسط الخطأ المطلق **0.75%** ونسبة الجذر التربيعي لمتوسط الخطأ التربيعي **2.02%** والتباين الموضح **0.71** ومعامل التحديد **0.69** .

5. الخلاصة:

في هذه الورقة التي تعني بمقارنة ثلاث من خوارزميات الانحدار للتنبؤ بمبيعات الأدوية، واختيار الخوارزمية الأكثر دقة وأقل نسبة خطأ وفق مقاييس الانحدار، تم اختيار خوارزمية منحدر الغابة العشوائية في بناء نموذج التنبؤ الانحداري وفق معطيات محددة لأنها كانت الأفضل بين الخوارزميات المختبرة، أما بالنسبة لتنبؤ السلاسل الزمنية فتم اختيار خوارزمية الانحدار الخطي التي أظهرت أفضل النتائج بالمقارنة مع الخوارزميات الأخرى التي تم اختبارها وفق مقاييس الانحدار.

6. التوصيات:

التوسيع في عملية اختبار الخوارزميات باختبار خوارزميات انحدار أخرى مثل خوارزمية انحدار ريدج Ridge وانحدار ايلاستسك نت Elastic net ومقارنة نتائجها وحساب الخطأ والدقة.

المراجع:

[1]

http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/data mining.htm

تاريخ الدخول: 20/4/2022

[2]

https://arabicprogrammer.com/article/3086224934/

تاريخ الدخول: 13/5/2022

[3]

"معلومات عن متسلسلة زمنية على موقع"id.ndl.go.jp ، id.ndl.go.jp، مؤرشف من الأصل في 11 مايو 2020

[4]

Willmott, Cort J.; Matsuura, Kenji (December 19, 2005). "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance". Climate Research. **30**: 79–82. doi:10.3354/cr030079.

[5]

Bickel, Peter J.; Doksum, Kjell A. (2015). Mathematical Statistics: Basic Ideas and Selected Topics. Vol. I (Second ed.). p. 20. If we use quadratic loss, our risk function is called the mean squared error (MSE) ..

[6]

https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics

تاريخ الدخول: 24/5/2022

[7]

 Steel, R. G. D.; Torrie, J. H. (1960). Principles and Procedures of Statistics with Special Reference to the Biological Sciences. McGraw Hill.

[8]

Glantz, Stanton A.; Slinker, B. K. (1990). Primer of Applied Regression and Analysis of Variance. McGraw-Hill. ISBN 978-0-07-023407-9.

[9]

Draper, N. R.; Smith, H. (1998). Applied Regression Analysis. Wiley-Interscience.

[10]

Kuhlman, Dave، January June 2012 "A Python Book: Beginning Python, Advanced Python, and Python Exercises"،.
Venners, Bill January 2003 "The Making of Python" ،Artima Developer ،Artima ،

[11]

"About Anaconda". Archived from the original on 19 April 2020. Retrieved 27 April 2020.
"Anaconda Navigator". docs.anaconda.com

[12]

"Jupyter kernels > List of (some) IPython compatible kernels". GitHub