استخدام خوارزمية الغابة العشوائية وخوارزمية الانحدار الخطي للتنبؤ بمبيعات الأدوية

Using Random Forest and linear regression algorithms for Medicine Sales Prediction

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:

سلطت هذه الورقة الضوء على التطور الذي لحق بتقنيات التنقيب في البيانات، ففي السنوات الأخيرة ومع الكم الهائل من البيانات وجب الحصول على تقنيات للتنقيب عن المعلومات من مناجم البيانات وهذه التقنيات تتطور يوماً بعد يوم لتواكب هذا العصر.الهدف من هذه الورقة هو استخدام تقنيات تنقيب البيانات لزيادة أداء المنظمة والتنبؤ بدقة بالمبيعات المستقبلية للدواء في مدن السودان المختلفة باستخدام تقنيات التنقيب في البيانات، واتخذت شركة بينا للأدوية نموذجاً لها. وأخذت 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. استخدمت تقنيات الانحدار وتقنيات السلاسل الزمنية لبناء نماذج التنبؤ بالمبيعات، واختبرت عدة خوارزميات وتم اختيار الخوارزميات التي أظهرت أفضل أداء من حيث الدقة وأقل نسبة خطأ لاستخدامها وحفظ نماذج التنبؤ الخاصة بها. وتم التوصل الى نتائج للتنبؤ بالمبيعات

حسب المدينة ونتائج للتنبؤ بالمبيعات حسب الصنف باستخدام النموذج الأفضل بنسبة خطأ 0.09% لنموذج الغابة العشوائية، ونسبة خطأ 0.34% لنموذج الانحدار الخطي.

الكلمات المفتاحية: تنقيب البيانات، مبيعات الأدوية، الانحدار، السلاسل الزمنية.

Abstract:

This paper spotted light on the development of data mining techniques. In recent years, with the huge amount of data, it is necessary to obtain techniques for data mining from data mines, and these techniques are developing day after day to keep pace with this era. The aim of this paper is to use data mining techniques to increase the organization's performance and accurately predict future sales of medicine in different cities of Sudan using data mining techniques, and Bina Pharmaceutical Company was taken as a model. The data was taken from Bina Pharmaceutical Company in Khartoum, and the study included 2,700 records representing three years' worth of data (2017-2018-2019) that were processed to forecast 2020 sales and compare them with actual sales of 2020 to determine its accuracy. Regression techniques and time series techniques were used to build sales prediction models, and several algorithms were tested, and the algorithms that showed the best performance in terms of accuracy and lowest error rate were selected to use and save their prediction models. Results for forecasting sales by city and sales forecasting by category were obtained using the best model, with an error rate of 0.09% for the random forest model, and an error rate of 0.34% for the linear regression model.

Keywords: Datamining, Drug sales, Regression, Time series.

1. المقدمة

يعد التنبؤ بالمبيعات جزءًا مهمًا من إدارة عمليات الأعمال، على الرغم من الصعوبات في تنفيذ عمليات التنبؤ عبر الشركات المختلفة ، يبقى الغرض المقصود من التنبؤ بالمبيعات نفس الشيء: الحصول على تقدير دقيق تقريبًا لمنتج أو خدمة، على ضوء البيانات السابقة والحالية وحالة البيئة (على سبيل المثال ، السياسية والاجتماعية والاقتصادية)، لتخطيط وتنظيم الأعمال التجارية وفقًا لذلك [1]. وشركات الأدوية كأي من المؤسسات تواجه مشكلة تراكم البيانات دون الاستفادة منها والدواء سلعة لها خصوصية وعملية بيعها تتم عبر مجموعة من الشروط والأحكام وتتأثر بعوامل مختلفة وتظهر مشاكل بيع الدواء في ظهور نقص في بعض الأصناف نتيجة لعدم استيرادها في الوقت المناسب وانتهاء صلاحية بعض الأصناف نتيجة لاستيرادها أكثر من الحاجة الفعلية كما تظهر مشكلة التوزيع حيث يتم توزيع الأدوية المعينة في مناطق معينة بكميات قد لا تكون في حاجة لها أو كميات أقل مما هو مطلوب أي عدم توفر الكمية المناسبة في الوقت المناسب والمكان المناسب ووجود الكم الهائل من البيانات دون الاستفادة منها يؤدي الى خسائر فادحة تزداد يوماً بعد يوم وعدم حصول المرضى على الأدوية التي يحتاجون البها في الوقت المناسب اذا لم تتم معالجة هذه المشاكل لذا وجب الاستفادة من التقنية والبرمجيات الحديثة. وتنقيب البيانات هي عملية البحث داخل كميات كبيرة من البيانات للكشف عن العلاقات التي لم يتم كشفها سابقا بين عناصر البيانات؛ والمعروفة أيضا باسم اكتشاف المعرفة في قواعد البيانات [2]. يمكّن التحليل الدقيق للمبيعات الشركات من اتخاذ قرارات تجارية مستنيرة وتنفيذ تنبؤات قصيرة المدى [1]. الهدف من هذه الورقة هو استخدام تقنيات تنقيب البيانات للتنبؤ بدقة بالمبيعات المستقبلية للدواء في مدن السودان المختلفة، والتنبؤ بمبيعات الأصناف المختلفة لتوفير الكميات المناسبة في الأوقات المناسبة في الأماكن المناسبة. وتم استخدام تقنيات الانحدار ومنها تم اختيار خوارزمية الغابة العشوائية لبناء نموذج التنبؤ، وخوارزمية الانحدار الخطي لبناء نموذج السلاسل الزمنية.

2. الإطار النظري:

1.2 خوارزمية الغابة العشوائية (Random forest):

الغابة العشوائية أو الغابات العشوائية عبارة عن خوارزمية للتعلم الآلي تم تطويرها بناءً على مجموعة من أشجار القرار .تستخدم هذه الخوارزمية للتصنيف، الانحدار ومهام أخرى. عادةً ما تتمتع خوارزمية الغابة العشوائية بدقة أفضل مقارنة بشجرة القرار [3].

2.2 الانحدار الخطّيّ Linear Regression :

الانحدار الخطّيّ البسيط Simple Linear Regression: هو أبسط أشكال الانحدار وأكثرها شيوعاً، ويقوم بتقدير العلاقة التي تربط بين متغيّرين أحدهما المتغيّر التّابع والآخر المتغيّر المستقل، وتكون العلاقة التي تربطهما علاقة خطّيّة، وينتج من هذا النّموذج معادلة خطيّة يمكن استخدامها لتفسير العلاقة بينهما أو لتقدير قيمة المتغير التّابع عند معرفة قيمة المتغير المستقلّ [4].

3. المنهجية

1.3 جمع البيانات:

تم أخذ البيانات من شركة بينا للأدوية في مدينة الخرطوم قسم المبيعات، وتم اختيار بيانات أكبر عشر صيدليات من كل مدينة من التسع مدن التي تتعامل معها الشركة وشملت الدراسة عدد 2700 سجل تم أخذها من قسم المبيعات تمثل بيانات ثلاث سنوات 2017–2018–2019.

2.3 وصف البيانات:

هذا الجزء يتضمن وصف للمتغيرات والميزات والعوامل التي تشكل البيانات التي تم جمعها وهي: الرقم (.No)، اسم الدواء (Drug Name)، استخدام الدواء (General Use)، المدينة (City)، الصيدلية (Pharmacy)، السعر (Price)، المبيعات في الربع الأول (Q1)، المبيعات في الربع الثاني (Q2)، المبيعات في الربع الثالث (Q3)، المبيعات في الربع الرابع (Q4)، مجموع المبيعات (Total Sales)، المبيعات (Sales)، السنة (Year).

3.3 تهيئة وتنظيف البيانات:

البيانات التي تم أخذها من قسم المبيعات في شركة بينا للأدوية هي بيانات أولية لا يمكن استخدامها مباشرة في عملية التنقيب، يجب أن يتم معالجتها في عدة خطوات لتصير قابلة للاستخدام وهذه الخطوات هي

في هذه الخطوة تم:

1. تصحيح أخطاء البيانات المتمثلة في أخطاء الإدخال
2. تم إزالة البيانات المكررة.

3. تم إزالة البيانات غير المكتملة.

4. تم أيضا استبعاد بعض الحقول الغير مهمة مثل اسم المندوب واليوم والشهر لعدم أهميتها في البحث.

5. تحويل البيانات لصيغة csv.

4.3 البرمجيات التي تم استخدامها في الورقة:

تم استخدام مجموعة من أحدث البرمجيات للوصول للنتائج وهي:

لغة البايثون:python [9]

الإصدار المستخدم: python 3.10.2

مدير البيئة Anaconda: Environment Manager [10]

الإصدار المستخدم: Anaconda3 2021.11

جوبيتر نوت بوك :Jupyter notebook [11]

الإصدار المستخدم: Jupyter notebook 6.4.5

بيئة التطوير :Pycharm [12]

الإصدار المستخدم: Pycharm 2021.2.2

5.3 مكتبات بايثون المستخدمة:

pandas 1.4.0

numpy 1.22.1

pickle 4.0

scikit learn 1.0.2

matplotlib 3.5.1

plotly 5.5.0

streamlit 1.4.0

sqlite3 3.37.0

4. التنفيذ والنتائج:

تم أخذ بيانات أكبر عشر صيدليات من كل مدينة من التسع مدن التي تتعامل معها الشركة ثم تمت معالجتها وعرضها وتجهيزها لتطبيق خوارزميات التنقيب في البيانات عليها، وقد تم استخدام تقنيات الانحدار (Regression) وتقنيات السلاسل الزمنية (Time series) لبناء نماذج التنبؤ بالمبيعات.

تم استخدام لغة البرمجة بايثون الإصدار: python 3.10.2 ومكتبات تعلم الآلة للتنقيب في البيانات واستخراج النتائج وعرضها ولغة بايثون من لغات البرمجة القوية والنشطة في عدة مجالات من ضمنها تعلم الآلة والتنقيب في البيانات.

بالنسبة للبيانات التي تم استخدامها تم اختيار خوارزمية منحدر الغابة العشوائية (Random Forest Regressor) لبناء نموذج التنبؤ الانحداري المبني على المدخلات لأنها أعطت أدق النتائج وأقل نسبة خطأ، أما بالنسبة لتنبؤ السلاسل الزمنية (Time Series Forecasting) فتم اختيار خوارزمية الانحدار الخطي (Linear Regression) لأنها كانت الأفضل أداء من حيث الدقة وأقل نسبة خطأ.

والجدول رقم (1) يوضح نتائج خوارزمية منحدر الغابة العشوائية التي استخدمت لبناء نموذج التنبؤ الانحداري:

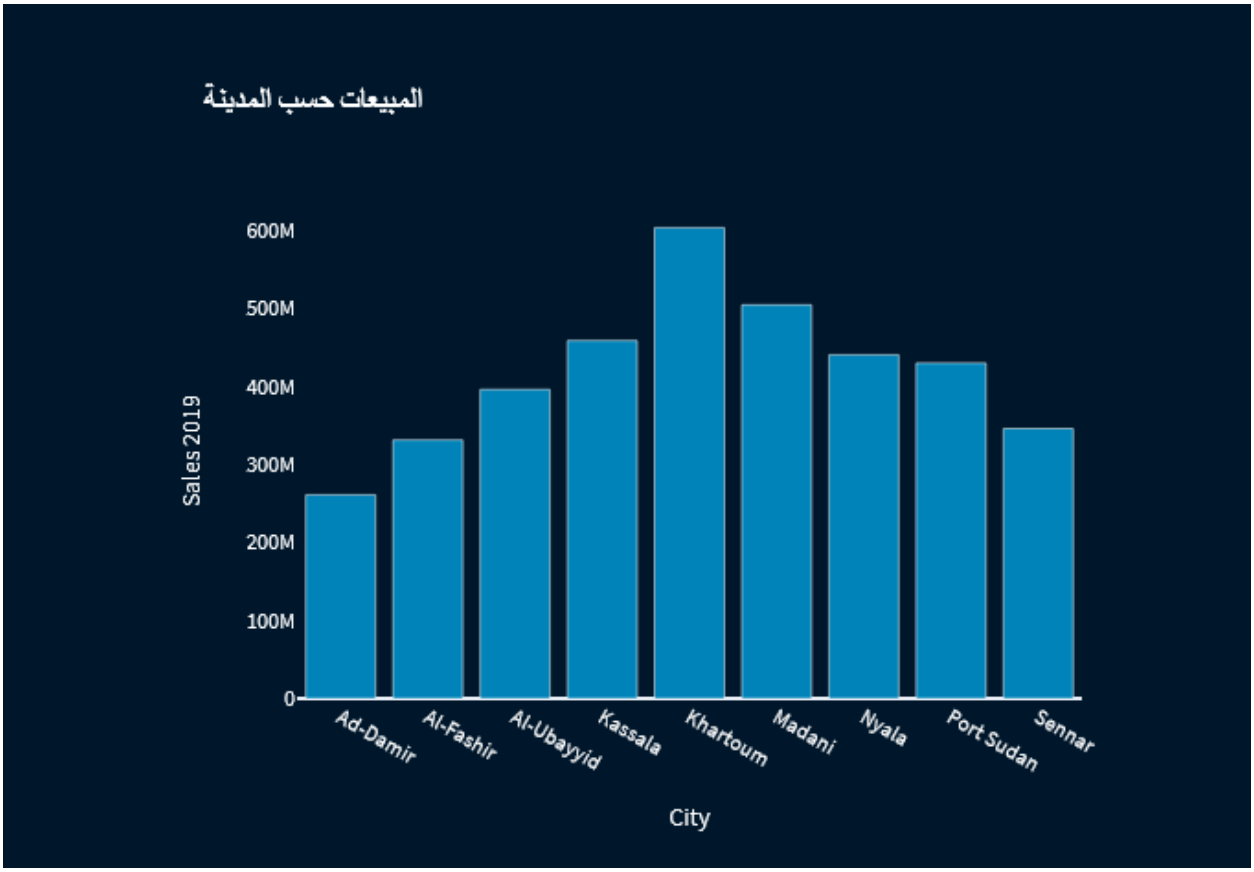
مقياس الانحدار / الخوارزمية	Sqrt Mean Squared Error	Mean Absolute Error	Explained Variance Score	Coefficient of Determination R^2 Score
Random Forest Regressor	3,468,218 0.09%	2,458,127 0.06%	0.98	0.98

والجدول رقم (2) يوضح نتائج خوارزمية الانحدار الخطي التي استخدمت لبناء نموذج السلاسل الزمنية:

مقياس الانحدار / الخوارزمية	Sqrt Mean Squared Error	Mean Absolute Error	Explained Variance Score	Coefficient of Determination R^2 Score
Linear Regression	16,545,918 0.34%	9,171,076 0.20%	0.99	0.99

من خلال تحليل بيانات الأدوية الخاصة بشركة بينا واستخدام تعلم الآلة وتقنيات التنقيب في البيانات تم التوصل للنتائج التالية:

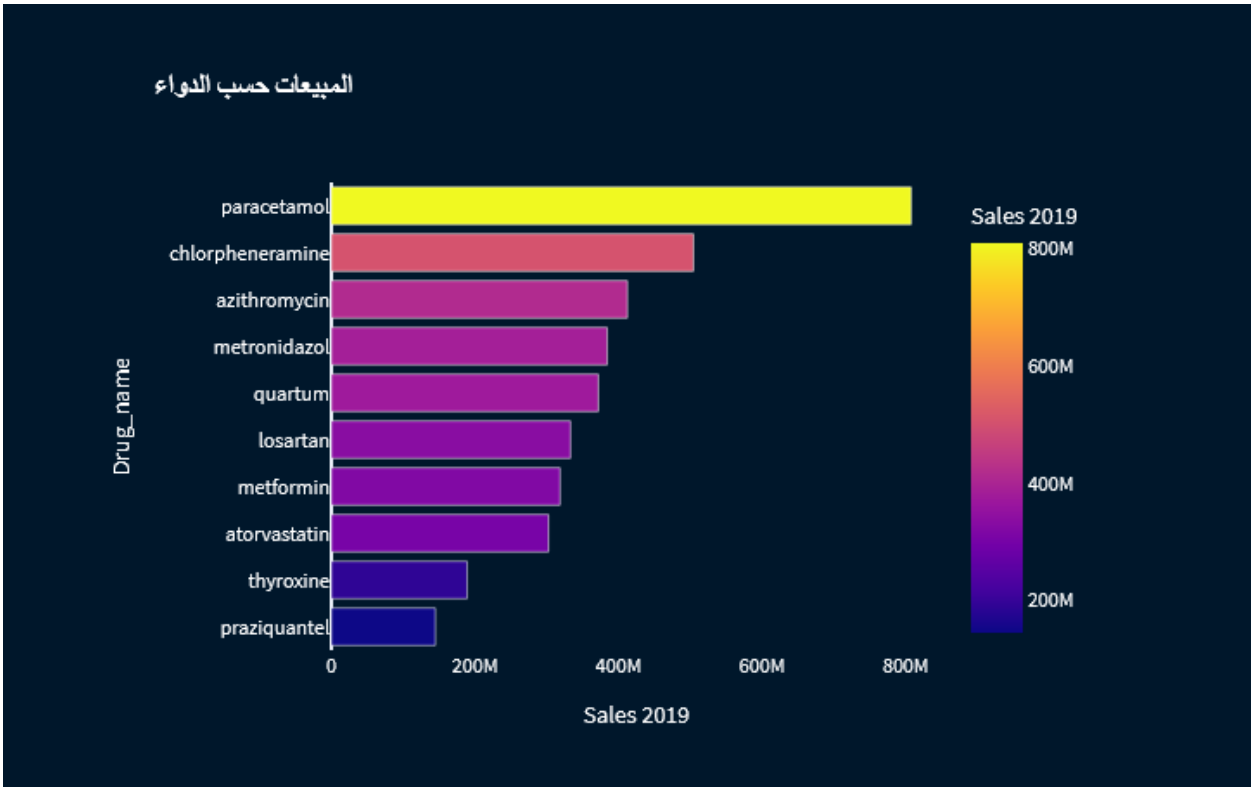
. أعلى المبيعات كانت في مدينة الخرطوم وأقلها في مدينة الدامر.



شكل رقم (1) يوضح مبيعات المدن

حيث يوضح شكل رقم (1) تقرير عن المدن ومبيعاتها في شكل أعمدة بيانية وكانت أعلى مبيعات في مدينة الخرطوم وأقل مبيعات في مدينة الدامر.

2. أكثر الأدوية مبيعاً هو Paracetamol وأقلها مبيعاً هو Praziquantel.



شكل رقم (2) يوضح الأدوية الأكثر مبيعاً

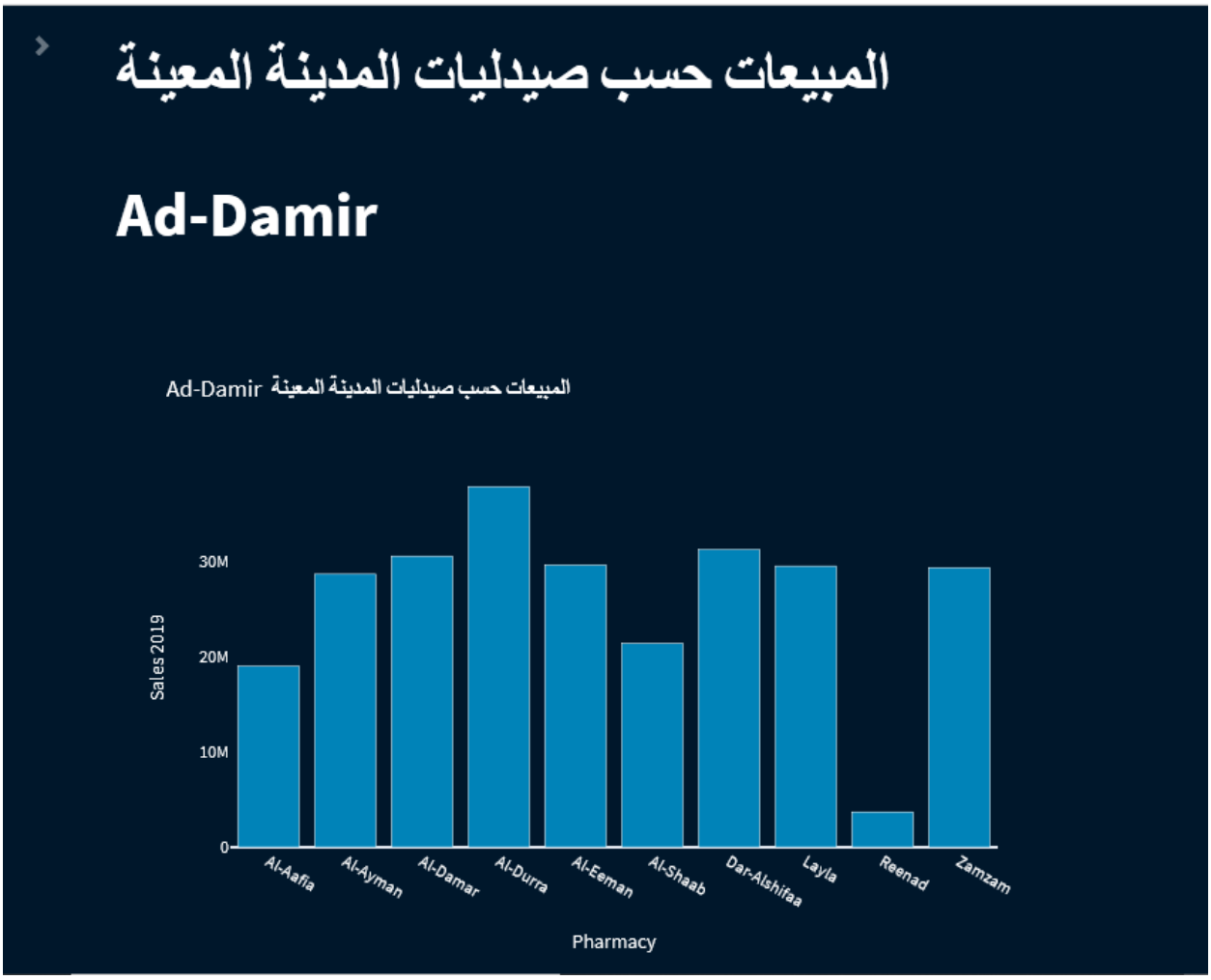
حيث يوضح شكل رقم (2) تقرير عن الأدوية ومبيعاتها في شكل أعمدة أفقية بيانية متدرجة وكان الأعلى مبيعاً Paracetamol والأقل Praziquantel.

3. أعلى مبيعات في مدينة الخرطوم كانت في صيدلية الخوارزمي وأقلها في صيدلية البركة.

شكل رقم (3) يوضح الصيدليات الأكثر مبيعاً في مدينة الخرطوم

حيث يوضح شكل رقم (3) تقرير عن الصيدليات ومبيعاتها في مدينة الخرطوم وكانت أعلى مبيعات في صيدلية الخوارزمي وأقل مبيعات في صيدلية البركة.

4. أعلى مبيعات في مدينة الدامر كانت في صيدلية الدرة وأقلها في صيدلية ريناد.



شكل رقم (4) يوضح الصيدليات الأكثر مبيعاً في مدينة الدامر

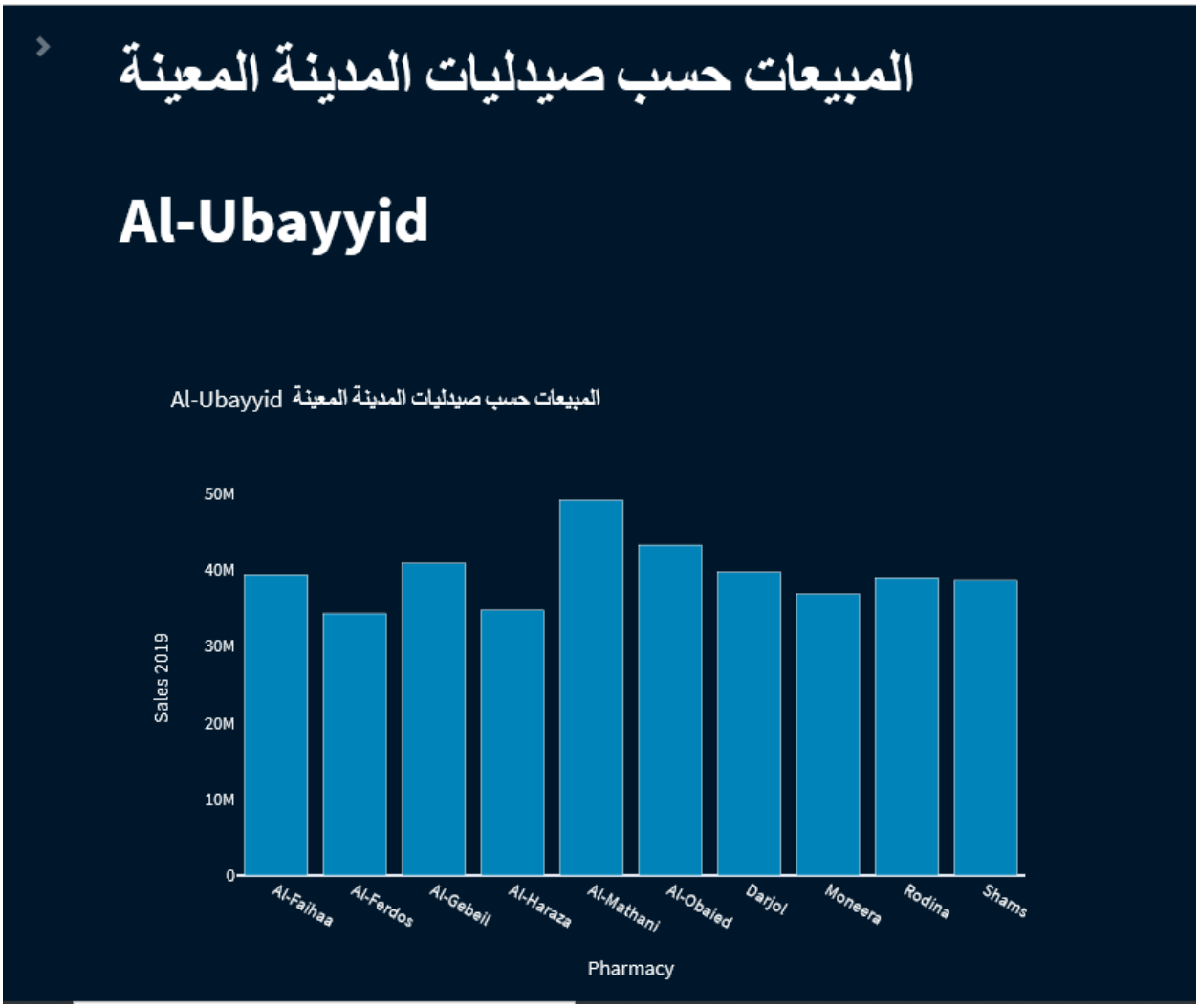
حيث يوضح شكل رقم (4) تقرير عن الصيدليات ومبيعاتها في مدينة الدامر وكانت أعلى مبيعات في صيدلية الدرة وأقل مبيعات في صيدلية ريناد.

5. أعلى مبيعات في مدينة الفاشر كانت في صيدلية يعقوب وأقلها في صيدلية اسماعيل.

شكل رقم (5) يوضح الصيدليات الأكثر مبيعاً في مدينة الفاشر

حيث يوضح شكل رقم (5) تقرير عن الصيدليات ومبيعاتها في مدينة الفاشر وكانت أعلى مبيعات في صيدلية يعقوب وأقل مبيعات في صيدلية اسماعيل.

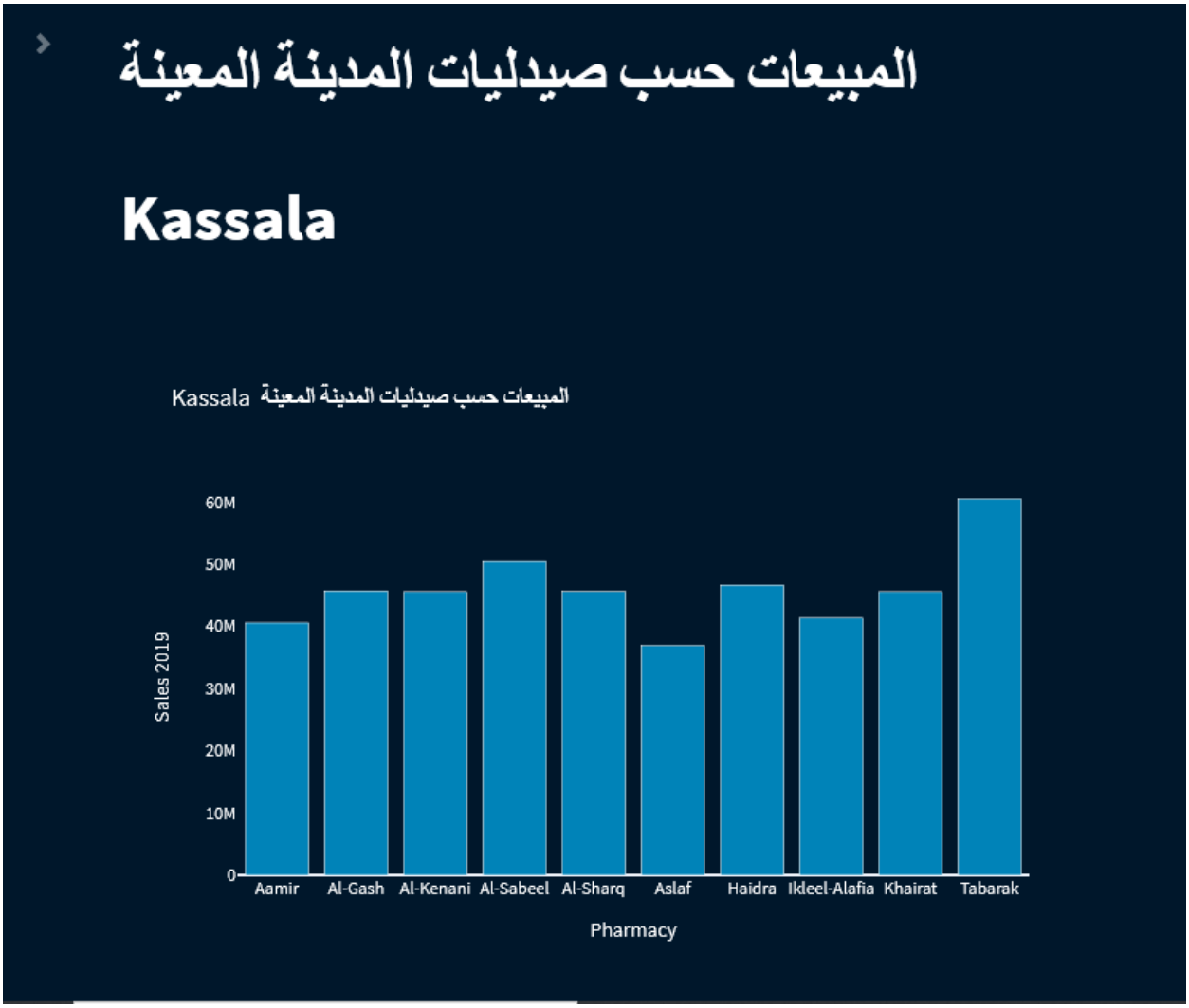
6. أعلى مبيعات في مدينة الأبيض كانت في صيدلية المثاني وأقلها في صيدلية الفردوس.



شكل رقم (6) يوضح الصيدليات الأكثر مبيعاً في مدينة الأبيض

حيث يوضح شكل رقم (6) تقرير عن الصيدليات ومبيعاتها في مدينة الأبيض وكانت أعلى مبيعات في صيدلية المثاني وأقل مبيعات في صيدلية الفردوس.

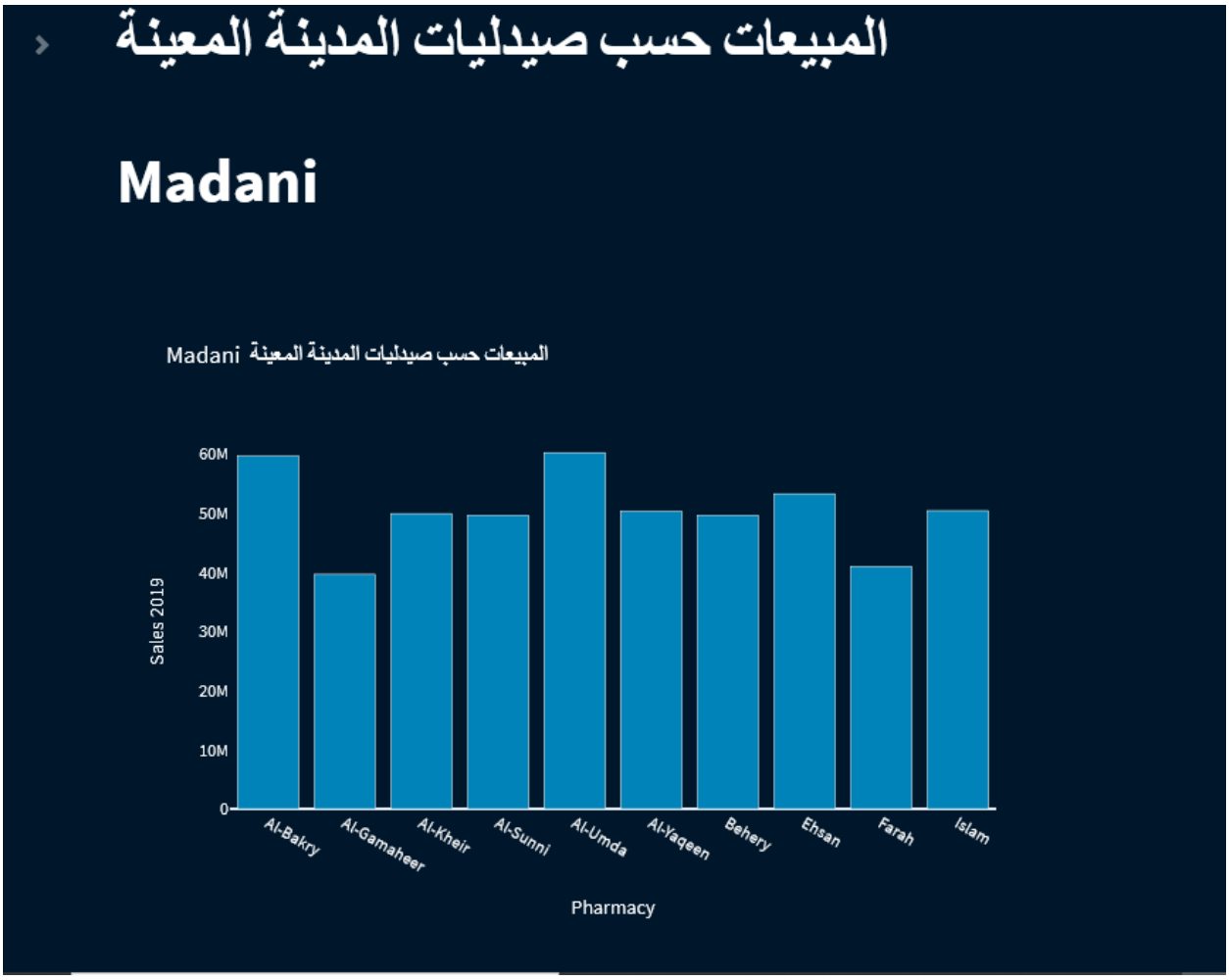
7. أعلى مبيعات في مدينة كسلا كانت في صيدلية تبارك وأقلها في صيدلية أسلاف.



شكل رقم (7) يوضح الصيدليات الأكثر مبيعاً في مدينة كسلا

حيث يوضح شكل رقم (7) تقرير عن الصيدليات ومبيعاتها في مدينة كسلا وكانت أعلى مبيعات في صيدلية تبارك وأقل مبيعات في صيدلية أسلاف.

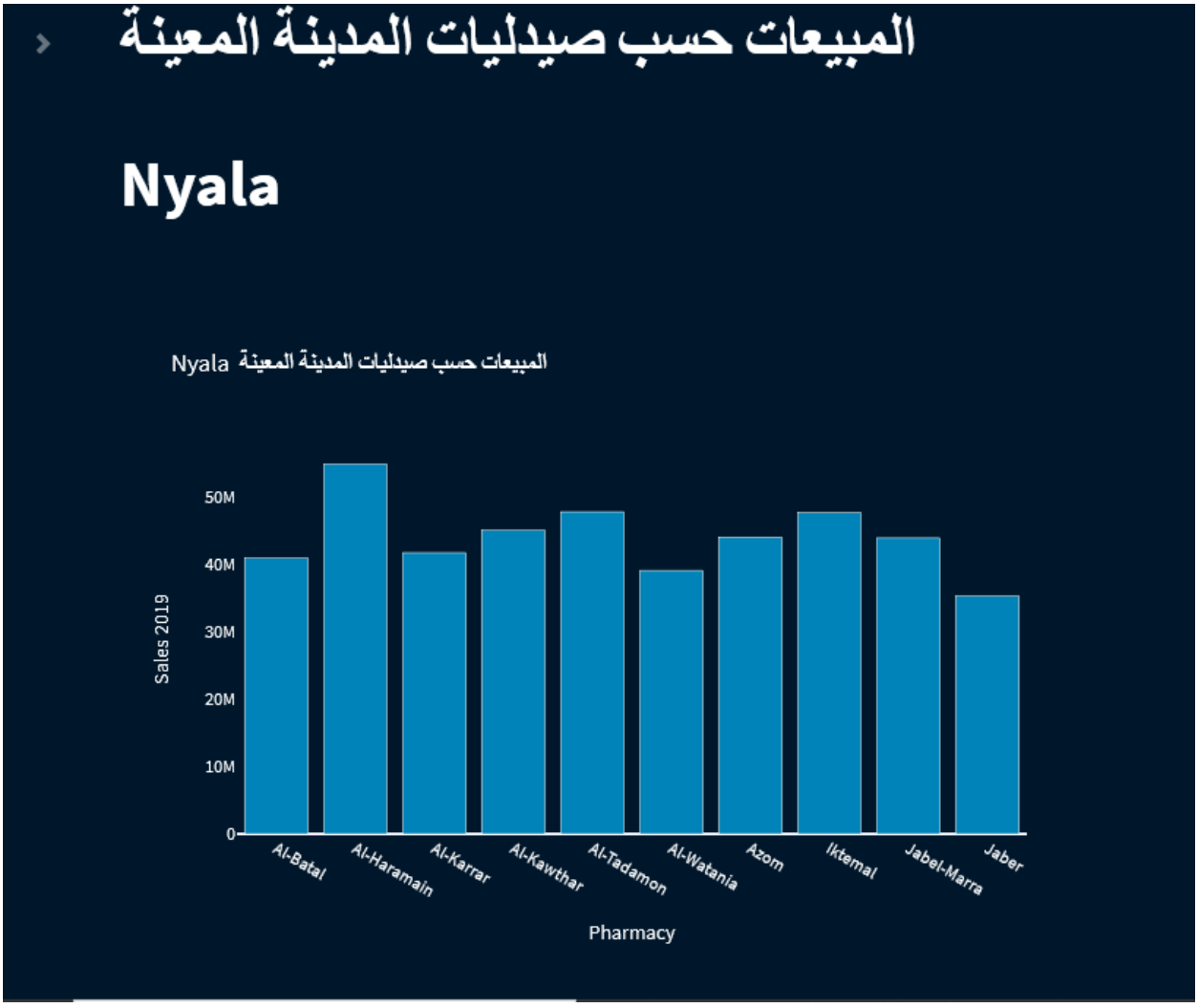
8. أعلى مبيعات في مدينة ود مدني كانت في صيدلية العمدة وأقلها في صيدلية الجماهير.



شكل رقم (8) يوضح الصيدليات الأكثر مبيعاً في مدينة مدني

حيث يوضح شكل رقم (8) تقرير عن الصيدليات ومبيعاتها في مدينة مدني وكانت أعلى مبيعات في صيدلية العمدة وأقل مبيعات في صيدلية الجماهير.

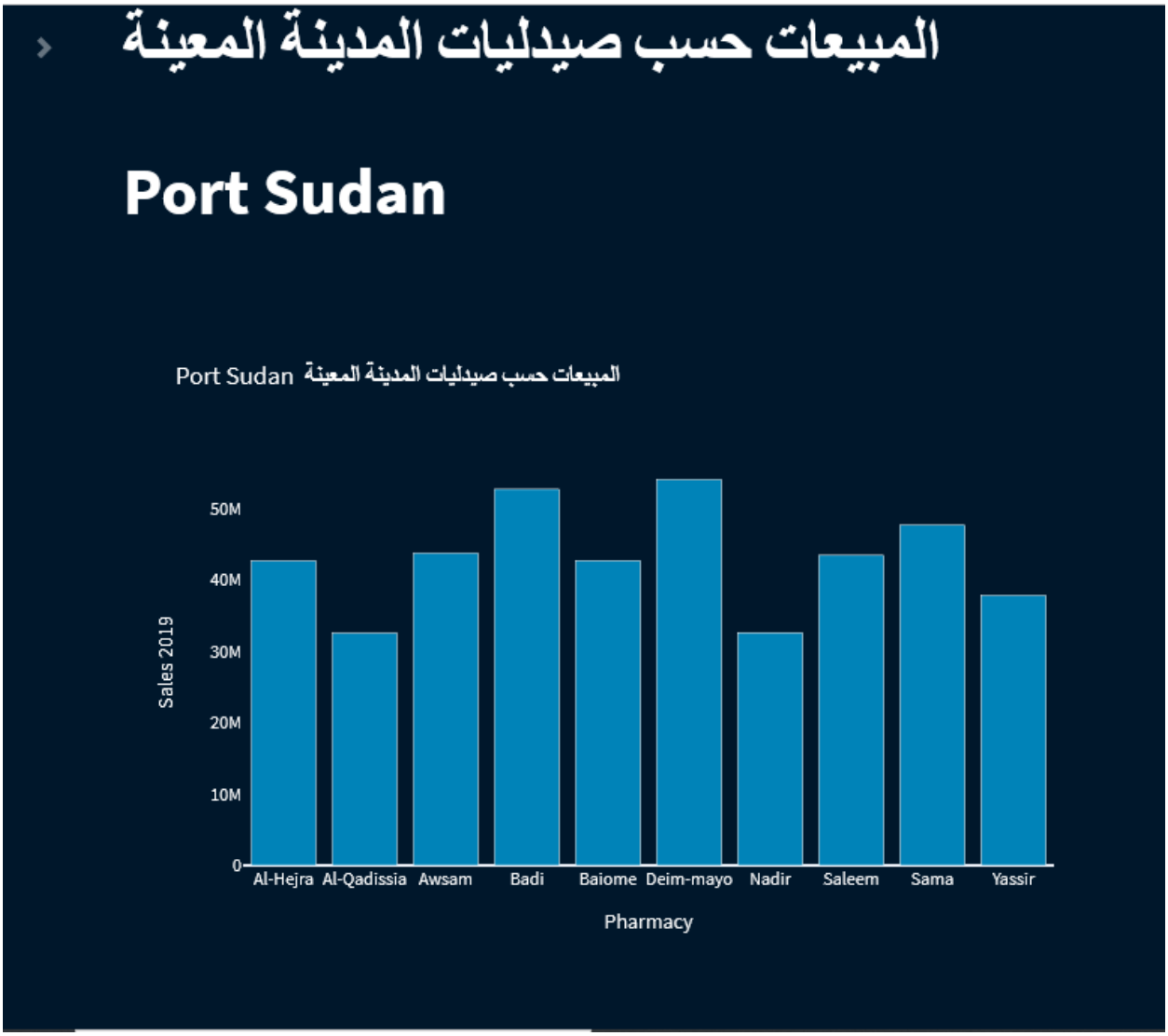
9. أعلى مبيعات في مدينة نيالا كانت في صيدلية الحرمين وأقلها في صيدلية جابر.



شكل رقم (9) يوضح الصيدليات الأكثر مبيعاً في مدينة نيالا

حيث يوضح شكل (9) تقرير عن الصيدليات ومبيعاتها في مدينة نيالا وكانت أعلى مبيعات في صيدلية الحرمين وأقل مبيعات في صيدلية جابر.

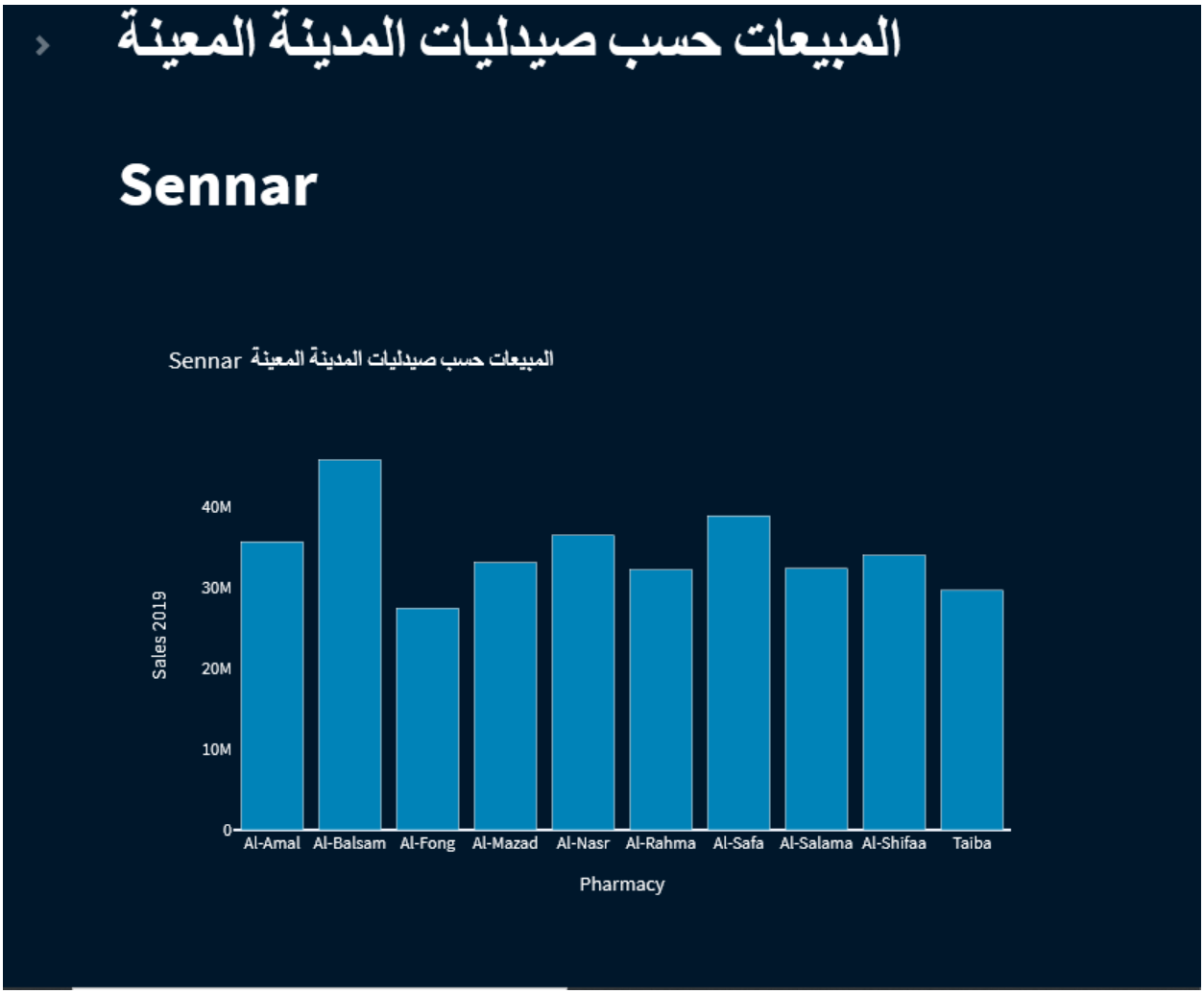
10. أعلى مبيعات في مدينة بورتسودان كانت في صيدلية ديم مايو وأقلها في صيدلية القادسية.



شكل رقم (10) يوضح الصيدليات الأكثر مبيعاً في مدينة بورتسودان

حيث يوضح شكل رقم (10) تقرير عن الصيدليات ومبيعاتها في مدينة بورتسودان وكانت أعلى مبيعات في صيدلية ديم مايو وأقل مبيعات في صيدلية القادسية.

11. أعلى مبيعات في مدينة سنار كانت في صيدلية البلسم وأقلها في صيدلية الفونج.

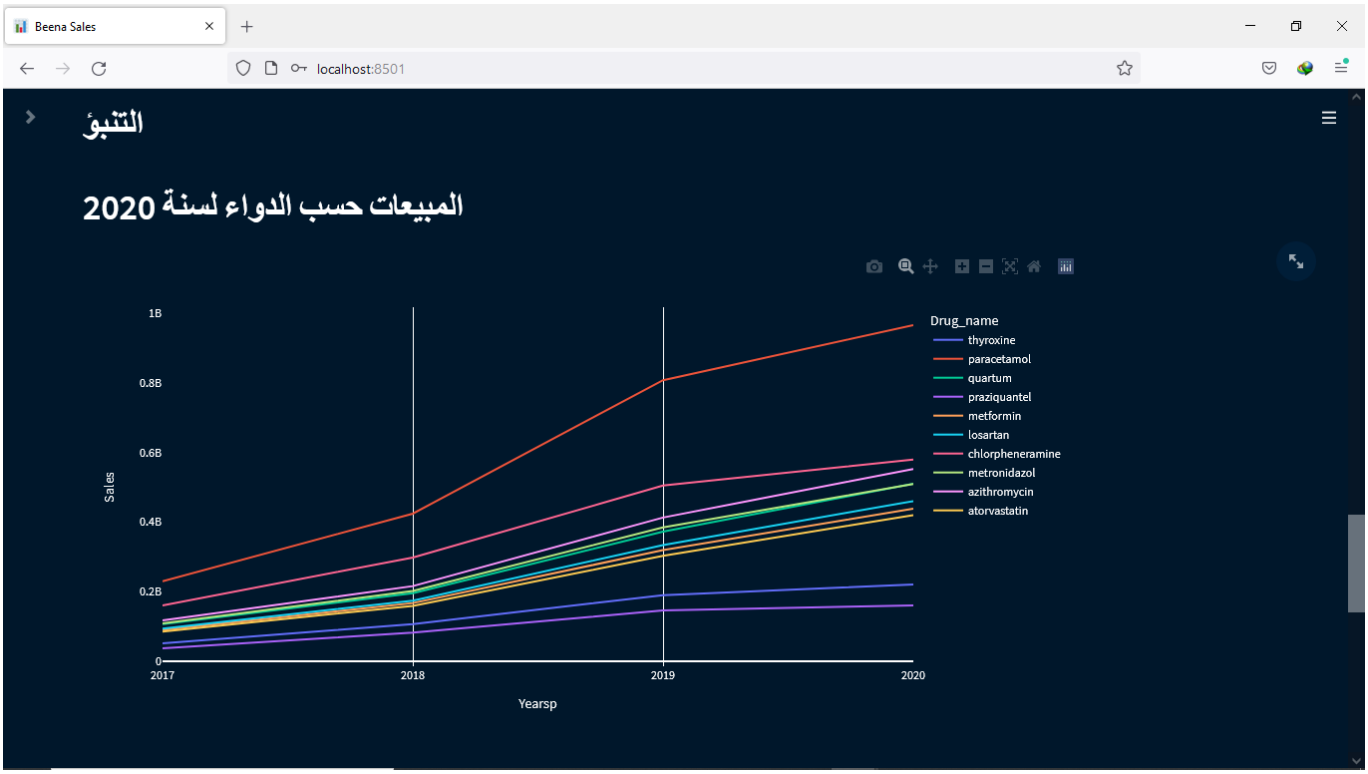


شكل رقم (11) يوضح الصيدليات الأكثر مبيعاً في مدينة سنار

حيث يوضح شكل رقم (11) تقرير عن الصيدليات ومبيعاتها في مدينة سنار وكانت أعلى مبيعات في صيدلية البلسم وأقل مبيعات في صيدلية الفونج.

12. أعلى مبيعات دواء thyroxine كانت في مدينتي نيالا والفاشر، أعلى مبيعات دواء Quartum كانت في مدينتي مدني وسنار، أعلى مبيعات دواء Praziquantel كانت في مدينتي مدني وسنار، أعلى مبيعات لأدوية Paracetamol, Chlorpheneramine, Azithromycin, Metronidazol, Losartan, Metformin, Atorvastatin كانت في مدينة الخرطوم، المبيعات المتوقعة لعام 2020 لدواء Paracetamol هي 966 مليون جنيه، المبيعات المتوقعة لعام 2020 لدواء Chlorpheneramine هي 580 مليون جنيه، المبيعات

المتوقعة لعام 2020 لدواء Azithromycin هي 550 مليون جنيه، المبيعات المتوقعة لعام 2020 لدواء Metronidazol هي 509 مليون جنيه، المبيعات المتوقعة لعام 2020 لدواء Quartum هي 510 مليون جنيه، المبيعات المتوقعة لعام 2020 لدواء Losartan هي 460 مليون جنيه، المبيعات المتوقعة لعام 2020 لدواء Metformin هي 440 مليون جنيه، المبيعات المتوقعة لعام 2020 لدواء Atorvastatin هي 420 مليون جنيه، المبيعات المتوقعة لعام 2020 لدواء Thyroxine هي 220 مليون جنيه، المبيعات المتوقعة لعام 2020 لدواء Praziquantel هي 160 مليون جنيه.



شكل رقم (12) يوضح المبيعات المتوقعة حسب الدواء لسنة 2020

حيث يوضح الشكل رقم (12) تقرير عن الأدوية ومبيعاتها الحالية والمتوقعة للعام 2020 في شكل بياني خطي حيث يمثل كل دواء بلون خاص يوضح مبيعات الدواء والنمو الخاص بها.

13. المبيعات المتوقعة لعام 2020 في مدينة الخرطوم هي 815 مليون جنيه، المبيعات المتوقعة لعام 2020 في مدينة ود مدني هي 665 مليون جنيه، المبيعات المتوقعة لعام 2020 في مدينة كسلا هي 575 مليون جنيه، المبيعات المتوقعة لعام 2020 في مدينة نيالا هي 573 مليون جنيه، المبيعات المتوقعة لعام 2020 في مدينة بورتسودان هي 560 مليون جنيه، المبيعات المتوقعة لعام 2020 في مدينة الأبيض هي 500

مليون جنيه، المبيعات المتوقعة لعام 2020 في مدينة سنار هي 420 مليون جنيه، المبيعات المتوقعة لعام 2020 في مدينة الفاشر هي 380 مليون جنيه، المبيعات المتوقعة لعام 2020 في مدينة الدامر هي 325 مليون جنيه.

شكل رقم (13) يوضح المبيعات المتوقعة حسب المدينة لسنة 2020

حيث يوضح الشكل رقم (13) تقرير عن المدن ومبيعاتها الحالية والمتوقعة للعام 2020 في شكل بياني خطي حيث يمثل كل دواء بلون خاص يوضح مبيعات المدن والنمو الخاص بها.

5. الخلاصة:

تم أخذ بيانات أكبر عشر صيدليات من كل مدينة من التسع مدن التي تتعامل معها شركة بينا للأدوية في هذه الورقة العلمية التي تعني باستخدام تقنيات التنقيب في البيانات لزيادة أداء المنظمة والمساعدة في اتخاذ القرار (Decision support). ثم تمت معالجتها وعرضها وتجهيزها لتطبيق خوارزميات تنقيب البيانات عليها، وقد تم استخدام تقنيات الانحدار (Regression) وتقنيات السلاسل الزمنية (Time series) لبناء نماذج التنبؤ

بالمبيعات. وتم اختبار النماذج التي تم بناؤها باستخدام مقاييس الانحدار وكانت قدرة النموذج للتنبؤ بالمبيعات جيدة وبنسبة خطأ قليلة نسبياً وفق مقايس الانحدار بالمقارنة مع القيم الحقيقية للمبيعات.

ويوضح الجدول رقم 3 مقارنة مع بعض نتائج الدراسات السابقة:

Algorithm	Mean absolute error	Sqrt mean squared error
Support vector machine for regression [5]	2.1269	2.6592
Radial Basis Function [6] Network	2.077	3.0499
Gradient Boost [7]	187.32	344.71
Bayesian Linear Regression [8]	2469.54	4361.40
Random forest regressor	**2,458,127**	**3,468,218**

6. التوصيات والعمل المستقبلي:

1. استخدام تقنيات تنقيب البيانات الخاصة بالتعلم غير المشرف عليه (unsupervised learning) مثل تقنيات العنقدة (clustering) وقاعدة الارتباط (Association rule) لإيجاد علاقة بين المدخلات بدلاً عن التركيز فقط على المخرجات لزيادة كمية المعلومات المتحصل عليها من البيانات.

2. اختبار خوارزميات انحدار أخرى غير التي تم استخدامها في البحث مثل خوارزمية انحدار الشبكة العصبية وانحدار لاسو، مقارنة نتائجها وحساب الخطأ والدقة.

المراجع:

[1]

Smriti Keny, Sagarika Nair, Silka Nandi and Deepak Khachane May 2021 Sales Prediction for a Pharmaceutical Distribution Company International Journal of Engineering and Applied Physics (IJEAP)Vol. 1, No. 2, pp. 186~191 ISSN: 2737-8071

[2]

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

تاريخ الدخول: 18/4/2022

[3]

https://www.javatpoint.com/machine-learning-random-forest-algorithm

https://coursee.org/blog/artificial-intelligence/random-forest-algorithm/

تاريخ الدخول: 28/4/2022

[4]

https://towardsdatascience.com/introduction-to-regression-analysis-9151d8ac14b3

https://aiinarabic.com/regression/

تاريخ الدخول: 1/5/2022

[5]

Waleed Hamed Ahmed Eisa Nov 2017 Prediction of the Generated Amount of Power from a Thermal Power Plant Using Data Mining,

[6]

Ricciardi C, Ponsiglione AM, Scala A, Borrelli A, Misasi M, Romano G, Russo G, Triassi M, Improta G. Apr 2022 Machine Learning and Regression Analysis to Model the Length of Hospital Stay in Patients with Femur Fracture. Bioengineering (Basel). 14;9(4):172. doi: 10.3390/bioengineering9040172. PMID: 35447732; PMCID: PMC9029792.

[7]

C. Castillo-Bot´on a, D. Casillas-P´erez b,*, C. Casanova-Mateo c, S. Ghimire d, E. Cerro-Prada e, P.A. Gutierrez f, R.C. Deo d, S. Salcedo-Sanz (2022) Machine learning regression and classification methods for fog events prediction Atmospheric Research 272106157

[8]

C. CATAL, K. ECE, B. ARSLAN and A. AKBULUT January 2019 Benchmarking of Regression and Time Series Analysis Techniques for Sales Forecasting BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING, Vol. 7, No. 1,

[9]

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 ،

[10]

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

[11]

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

[12]

Haagsman, Ernst (4 April 2019). "Collaboration with Anaconda, Inc". *PyCharm Blog*. Retrieved 26 May 2019