HNSJ Humanities & Natural Sciences Journal
ISSN: (e) 2709-0833
www.hnjournal.net

مجلة العلوم الإنسانية والطبيعية
مجلة علمية محكمة (التصنيف: NSP)
معامل التأثير للعام 2022 = 4.91

عنوان البحث

بناء نموذج اختيار الميزات المؤثرة بخوارزمية الاختيار الأمامي المتسلسل للتنبؤ بأفضل دقة بمعدلات هطول الأمطار بدولة السودان

الامين عبدالله الامين طه[1] مرتضى مالك ادم الحاج[2] عاطف معاوية الطيب[3]

[1] محاضر بالجامعة السعودية الالكترونية.
[2] استاذ مساعد في تقانة المعلومات جامعة افريقيا العالمية
[3] استاذ مساعد شركة أكاديمية الجزيرة العالمية.

HNSJ, 2022, 3(9); https://doi.org/10.53796/hnsj396

تاريخ النشر: 2022/09/01م تاريخ القبول: 2022/08/10م

المستخلص

هدفت الدراسة الى بناء نموذج لاختيار أفضل الميزات المؤثرة للتنبؤ بأفضل دقة بمعدلات هطول الامطار في دولة السودان، لأن بعض النماذج المستخدمة في عملية التنبؤ تم تطويرها باستخدام ميزة وحيدة بحيث لا يتم الأخذ بالميزات الأخرى التي تؤثر في نتائج النموذج، كما لا توجد خوارزمية محددة لاختيار أفضل الميزات تناسب بيانات هطول الامطار للتنبؤ بمعدلها من خلال الدراسات السابقة التي اعتمدت عليها الدراسة. استخدمت الدراسة 10 خوارزميات (importance of random forest, Lasso, Persons Correlation Coefficient, ANOVA, Forward selection, Backward selection, Recursive Feature Elimination, Information gain, Correlation وImportance Features) لاختيار أفضل الميزات من حيث دقة التنبؤ وتم تجربتها على مجموعة بيانية مكونة من 35 ميزة و216792 سجل وتقييمها باستخدام معيار الدقة من خلال أربع خوارزميات تصنيف. خلصت الدراسة الى أن خوارزمية الاختيار الأمامي المتسلسل هي الأفضل من حيث الدقة بمعدلات دقة 78.6% باستخدام خوارزمية الغابة العشوائية، ثم 77.6% باستخدام خوارزمية أقرب الجيران K، ثم 76.6% باستخدام خوارزمية التعبئة، ثم 73.8% باستخدام خوارزمية شجرة القرار، وتوصي الدراسة بتجربة خوارزميات اختيار ميزات تحقق دقة أعلى في التنبؤ بمعدلات هطول الامطار في داخل وخارج دولة السودان.

الكلمات المفتاحية: اختيار الميزات، الدقة، التنبؤ، التقييم، هطول الامطار.

RESEARCH TITLE

BUILDING A MODEL FOR SELECTING THE INFLUENTIAL FEATURES USING THE SEQUENTIAL FORWARD SELECTION ALGORITHM TO PREDICT THE BEST ACCURACY OF RAINFALL RATES IN SUDAN

Alameen abdallah Alameen Taha[1], Murtada Malik Adam Elhaj[2], Atif Muawia Eltaib[3]

[1] Lecturer at the Saudi Electronic University.
[2] Assistant Professor of Information Technology, International University of Africa
[3] Assistant Professor, Al Jazeera International Academy Company.

HNSJ, 2022, 3(9); https://doi.org/10.53796/hnsj396

Published at 01/09/2022 **Accepted at 10/08/2021**

Abstract

The study aimed to build a model to choose the best features affecting the best accuracy to predict rainfall rates, because some models used in the forecasting process were developed using a single feature so that other features that affect the model results in terms of accuracy are not considered, and there is no specific algorithm to choose the best The factors fit the rainfall data to predict its rate through previous studies on which the study relied. The study used 10 algorithms (importance of random forest, Lasso, Persons Correlation Coefficient, ANOVA, Forward selection, Backward selection, Recursive Feature Elimination, Information gain, Correlation, and Importance Features) to choose the best features in terms of prediction accuracy and they were tested on a data set consisting of 35 features and 216,792 records and evaluated using an accuracy criterion through four classification algorithms. The study concluded that the sequential forward selection algorithm is the best in terms of accuracy with accuracy rates of 78.6% using the random forest algorithm, then 77.6% using the K-nearest neighbor algorithm, then 76.6% using the Bagging algorithm, then 73.8% using the decision tree algorithm. And the study recommends trying Feature selection algorithms with higher accuracy to predicate rainfall rates inside and outside Sudan.

Key Words: Feature Selection, Accuracy, Prediction, Rain Fall.

1. مقدمة

عادةً ما تحتوي نماذج التصنيف على عدد كبير من الميزات في البيانات، ولكن ليست جميعها مهمة للتنبؤ[1] يمكن أن يؤدي تحديد مجموعة محددة من الميزات إلى زيادة أداء النموذج بشكل كبير لتنقيب البيانات، وتسهيل فهم النموذج تنقيب البيانات، وجعل النموذج أكثر وضوحًا [2].

1.1 منهجيات اختيار الميزات

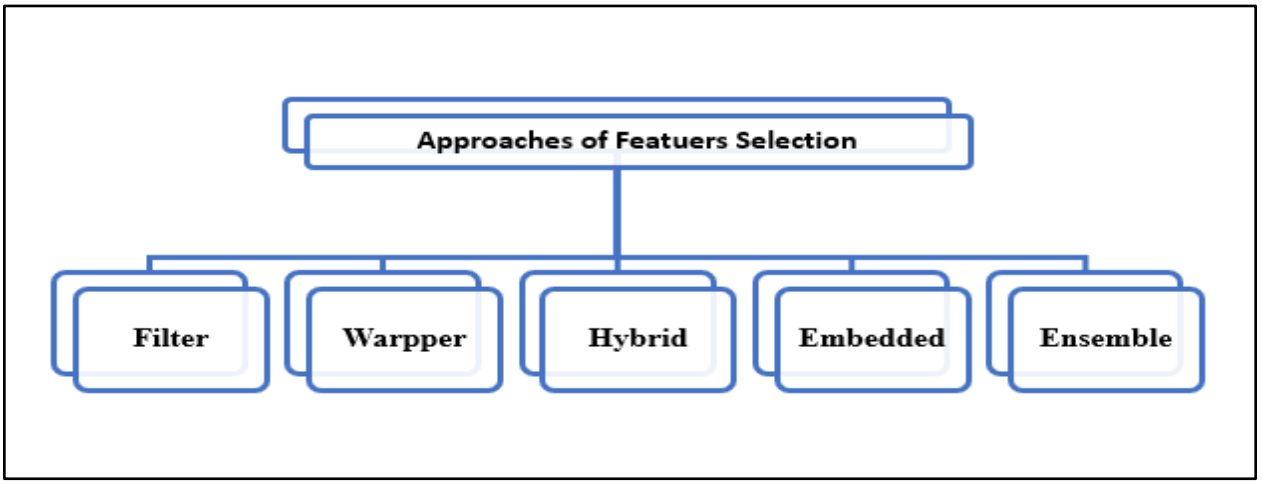
منهجيات اختيار الميزات Feature Selection Methodologies تنقسم منهجيات اختيار الميزات إلى خمسة أنواع:

1.1.1 طريقة التصفية

تعتمد طريقة التصفية Filter Method على تقييم أي ميزة بشكل فردي بناءً على نموذج إحصائي، حيث يتم تقييم كل ميزة ثم يتم ترتيب الميزات وفقًا لتقييم الميزات، ثم اختيار الجزء العلوي من الميزات ذات أعلى تقييم ليتم اعتمادها في النموذج النهائي [3]. تتميز طريقة التصفية ببساطتها وسرعة تنفيذها [4].

2.1.1 طريقة التغليف

تعتمد طريقة Wrapper Method على خوارزمية التنبؤ لتحديد مجموعة الميزات التي سيتم استخدامها في تصميم النموذج بناءً على مجموعة الميزات التي تعطي أعلى دقة مع الخوارزمية [5]. تستخدم هذه الطريقة على نطاق واسع وخاصة في التطبيقات التي تهتم بالدقة أكثر من السرعة، بعد كل شيء، فإنه يعطي نتائج أفضل، ولكن الأمر يستغرق وقتًا طويلاً في المعالجة للحصول على أفضل نتيجة [6]. ويمكن أيضًا استخدامه في تطبيقات الوقت الفعلي عندما يكون لدينا عدد قليل من الميزات [7].



الشكل (1) منهجيات اختيار الميزات

3.1.1 الطريقة الهجينة

الطريقة الهجينة Hybrid Method هي تتكون من مرحلتين: الأولى تقييم السمات وترتيبها وفقًا لمعيار معين، والمرحلة الثانية يتم فيها اختيار مجموعة الميزات التي تعطي أفضل نتيجة [8]. تزيل هذه الطريقة الميزات التي لا تزيد من دقة النموذج [9].

4.1.1 الطريقة المضمنة

تعتمد طريقة Embedded Method على خوارزمية تصنيف مثل طريقة التغليف، لكن الارتباط في

الطريقة المضمنة أقوى [10]. حيث تكون هذه الطريقة عبارة عن مزيج بين طريقة التصفية وطريقة التغليف، حيث يتم دمج عملية اختيار الميزات في مرحلة تدريب النموذج، ومن هذه العملية يتم إرجاع نتيجة تدريب النموذج ومجموعة الميزات المختارة. دمج عملية الاختيار في مرحلة التدريب النموذجي يحسن أداء النموذج [11].

5.1.1 طريقة المجموعة

تعتمد طريقة المجموعة Ensemble Method على استخدام أكثر من طريقة لاختيار مجموعة من الميزات التي تعطي أفضل أداء، فهذه الطريقة جيدة مقارنة باستخدام طريقة اختيار واحدة لتجنب ضعف الطريقة. لذلك فإن استخدام طريقة ثانية يعطي نتائج أفضل وموثوق بها، بالإضافة إلى أن استخدام عدة طرق يؤدي إلى توليد طريقة أكثر استقرارًا خاصة مع البيانات ذات الأبعاد العالية [12].

الدراسة تم تقسيمها الى سبعة اقسام؛ أولاً: المقدمة حيث تحتوي على مقدمة، المشكلة، حدود البحث، أهداف البحث، منهجية البحث، واجراءات البحث. ثانياً: الدراسات السابقة حيث تشمل 11 دراسة سابقة ومقارنة بينهم. ثالثاً: فكرة ونموذج وتطبيق الحل المقترح. رابعاً: النتائج. خامساً: مناقشة النتائج. سادسا: الخاتمة. وسابعاً: قائمة المصادر والمراجع.

2.1 مشكلة البحث

تتمثل المشكلة في أن بعض النماذج المستخدمة في عملية التنبؤ تم تطويرها باستخدام عامل [14] ، لا يتم الأخذ بالميزات الأخرى التي تؤثر في نتائج النموذج، كما لا توجد خوارزمية محددة لاختيار أفضل الميزات تناسب بيانات هطول الامطار للتنبؤ بمعدلها.

3.1 حدود البحث

الحدود الزمانية لهذا البحث هي الفترة الممتدة من (ديسمبر 2016م وحتى اغسطس 2022م)، وتم جمع البيانات الأولية في شهر مارس وابريل 2017م، الحدود المكانية هي الهيئة العامة للأرصاد الجوية السودانية بدولة السودان.

4.1 أهداف البحث

تشمل اهداف البحث: تحديد الميزات المؤثرة في عملية التنبؤ بمعدلات هطول الامطار، تحديد أفضل الميزات باستخدام أفضل الخوارزميات وبعد مقارنة نتائج الخوارزميات، بناء نموذج لتحديد أفضل الميزات، واختيار أفضل خوارزمية لتحديد الميزات للتنبؤ بأفضل دقة بمعدلات هطول الامطار بالسودان.

5.1 منهجية البحث

المنهجية العلمية المتبعة لإجراء هذا البحث تشمل المنهج التحليلي حيث تم جمع بيانات ومسح الدراسات السابقة وتحليلها وتصنيفها ومن ثم استخلاص الفجوة العلمية لغرض بناء نموذج لتحديد أفضل الميزات للتنبؤ بمعدلات هطول الامطار باستخدام المنهج التجريبي والتطبيقي.

6.1 إجراءات البحث

تهدف الإجراءات الى بناء نموذج يتألف من المراحل التالية: جمع البيانات Data Collection، تجهيز البيانات Data Preparation، تجربة ومقارنة بعض طرق من منهجيات مختلفة لاختيار الميزات وتقييمها من

حيث الدقة باستخدام خوارزميات تصنيف للتنبؤ (أقرب الجيران K–Nearest Neighbor ، شجرة القرار Decision Tree، الغابة العشوائية Random Forest ، والتعبئة Bagging)، اختيار أفضل طريقة من حيث الدقة، وتحديد الميزات لأفضل طريقة لغرض استخدامها في التنبؤ بمعدلات هطول الامطار.

2. الدراسات السابقة

دراسة Nikhil، (2021) [13] ، بعنوان " توقع هطول الأمطار باستخدام تقنيات التعلم الآلي"، استعرضت الدراسة مناهج وخوارزميات التعلم الالي للتنبؤ بهطول الأمطار، استخدمت الدراسة مجموعة خوارزميات التعلم الالي (Logistic Regression، Decision Tree، K – Nearest Neighbour، Random Forest، AdaBoost، Gradient Boosting) للتنبؤ بهطول الامطار، واشتملت بيانات الدراسة على متغيرات الطقس اليومية في المدن الكبرى في استراليا، تقدم النتائج مقارنة لمقاييس التقييم المختلفة لتقنيات التعلم الآلي ومدى صلتها بالتنبؤ بهطول الأمطار من خلال تحليل بيانات الطقس.

دراسة Basha واخرون (2020) [14] ، بعنوان " التنبؤ بهطول الأمطار باستخدام تقنيات التعلم الآلي والتعلم العميق"، في هذه الدراسة تمت مناقشة استخدام منهجية التعلم العميق Deep Learningفي التنبؤ بهطول الأمطار باستخدام تعدد الطبقات بمقارنة المعمارية الحالية مع المعماريات السابقة، تمت الإشارة لأهمية قضايا الدقة في التنبؤ نتيجة للعلاقات غير الخطية بين الميزات المختلفة المستخدمة في التنبؤ بمعدلات الأمطار باستخدام خوارزميات الذكاء الاصطناعي المختلفة.

دراسة Poornima واخرون (2019) [15] ، بعنوان "التنبؤ بهطول الأمطار باستخدام شبكة عصبية متكررة قائمة على LSTM مع وحدات خطية مرجحة " في هذه الدراسة تم اقتراح نموذجًا للتنبؤ بهطول الأمطار باستخدام RNN القائم على تقنية LSTM، تم الدراسة في منطقة حيدر أباد باستخدام مجموعة بيانات هطول الأمطار، تم استخدام الحد الأدنى والأقصى لدرجة الحرارة، وسرعة الرياح، وأشعة الشمس، والرطوبة النسبية الدنيا والقصوى، وميزات التبخر. وبمقارنة أداء نموذج LSTM مقارنة بأساليب RNN وLSTM وELM و–Holt Winters و ARIMAتظهر نتيجة هذه الدراسة أن تقنية LSTM تعطي نتائج أفضل مقارنة بالطرق الأخرى المستخدمة في التنبؤ بمعدلات هطول الأمطار.

دراسة kala واخرون (2018) [16]، بعنوان" التنبؤ بهطول الأمطار باستخدام الشبكة العصبية الاصطناعية"، في هذه الدراسة تم تطوير نموذج باستخدام الشبكة العصبية الاصطناعية (ANN)مثل شبكة التغذية العصبية الأمامية (FFNN) للتنبؤ بهطول الأمطار. وبأخذ أربع ميزات في الاعتبار مثل درجة الحرارة والغطاء السحابي وضغط البخار وهطول الأمطار لتحديد هطول الأمطار مسبقًا. تم استخدام جذر متوسط الخطأ التربيعي (RMSE) ومصفوفة الارتباك لقياس دقة التنبؤ. يشير النموذج المقترح المستند إلى ANN إلى دقة مقبولة.

دراسة Tharun واخرون، (2018) [17] ، بعنوان " التنبؤ بهطول الأمطار باستخدام تقنيات التنقيب في البيانات" هدفت هذه الدراسة الى مقارنة تقنيات الانحدار المختلفة القائمة على الخطأ النسبي، استخدمت هذه الدراسة تقنيات دعم الانحدار المتجه Support Vector Regression (SVR)، الغابة العشوائية Random forest (RF)، شجرة القرار Decision Tree (DT)، اشتملت الدراسة على بيانات الطقس اليومية (درجة

الحرارة، سرعة الرياح، اتجاه الرياح) في مدينة كونور لمدة 9 سنوات في الفترة من 2005 وحتى 2014، توصلت الدراسة الى ان نموذج RF أفضل وأكثر كفاءة مقارنة نماذج SVR و DT.

دراسة Aftab واخرون، (2018) [18] ، بعنوان " التنبؤ بهطول الأمطار في مدينة لاهور باستخدام تقنيات التنقيب عن البيانات"، هدفت هذه الدراسة إلى تحليل أداء تقنيات التنقيب عن البيانات للتنبؤ بهطول الأمطار في مدينة لاهور باستخدام إطار تصنيف، استخدمت هذه الدراسة تقنيات Support Vector Machine (SVM)، Naïve Bayes (NB) ، k Nearest Neighbor (KNN) ، Decision Tree (J48)، Multilayer Perceptron (MLP)، اشتملت بيانات البحث التي تم جمعها من مواقع ويب للتنبؤ بالطقس على العديد من سمات الغلاف الجوي (درجة الحرارة، الضغط الجوي على سطح الارض، الضغط الجوي على سطح البحر، ميل الضغط، الرطوبة النسبية، سرعة الرياح، أدنى درجة حرارة، أقصى درجة حرارة، الرؤية، مقياس معدل الرطوبة) في مدينة لاهور لمدة 12 سنة في الفترة من 2005 وحتى 2017، وفقًا للنتائج ، كان أداء تقنيات التصنيف المستخدمة جيدًا بالنسبة لفئة عدم هطول الأمطار ولكن بالنسبة لفئة المطر، لم تعمل التقنيات بشكل جيد، أوصت الدراسة إجراء المزيد من التنبؤات من خلال استكشاف المزيد من تقنيات التصنيف والسمات المناخية على بيانات الطقس المختلفة.

دراسة Kashiwao واخرون، (2017) [19] ، بعنوان "دراسة مبنية على الشبكات العصبية لهطول الأمطار المحلية باستخدام بيانات الإرصاد الجوي الموجودة على الإنترنت، دراسة حالة وكالة الإرصاد الجوي اليابانية"، هدف النظام المقترح إلى استخدام البيانات الموجودة على الإنترنت كـ "بيانات ضخمة" للتنبؤ بهطول الأمطار، استخدمت الدراسة نهجين للتنبؤ Radial Basis Function Network (RBFN)، و Multi-layer Perceptron(MLP) ، وقد اشتملت الدراسة على استخدام ثمانية أنواع من بيانات الأرصاد الجوية في اليابان (الضغط الجوي في الموقع، الضغط الجوي على سطح البحر، التساقط، درجة الحرارة ، درجة حرارة الهواء الطلق، ضغط البخار، الرطوبة، سرعة الرياح) في الفترة من 2000 وحتى 2012، توصلت نتائج الدراسة ان نهج (MLP) افضل في التنبؤ بهطول الامطار، تمت مقارنة نتائج التنبؤ مع نتائج وكالة الأرصاد الجوية اليابانية وأن الطريقة المقترحة تفوقت على تنبؤات وكالة الأرصاد الجوية اليابانية.

دراسة Qiu واخرون، (2017) [20] ، بعنوان " نموذج التنبؤ بهطول الأمطار على المدى القصير باستخدام الشبكات العصبية التلافيفية متعددة المهام"، اقترحت الدراسة نموذج الشبكة العصبية الالتفافية متعددة المهام للتنبؤ بهطول الأمطار، استخدمت الدراسة تقنيات التعلم متعدد المهام والتعلم العميق Multi-Task Convolutional Neural Networks (MT-CNN) للتنبؤ بكمية هطول الأمطار على المدى القصير، وقد اشتملت الدراسة على ثمانية أنواع من متغيرات الطقس بناءً على ميزات متعددة المواقع (حالة المطر، ارتفاع المرصد، سرعة الرياح، اتجاه الرياح، درجة الندى، درجة الحرارة، الضغط الجوي، الرطوبة)، في الفترة من 2002 وحتى 2015، أظهرت النتائج أن النموذج المقترح يتفوق بشكل كبير على مجموعة واسعة من النماذج الأساسية بما في ذلك نظام المركز الأوروبي للتنبؤات الجوية (ECMWF).

دراسة Rasel واخرون، (2017) [21] ، بعنوان " تطبيق التنقيب في البيانات والتعلم الآلي للتنبؤ بالطقس"، هدفت الدراسة الى مراقبة أداء التنبؤ بالطقس لمختلف تقنيات التعلم الآلي واستخراج البيانات واقتراح

نموذج للتنبؤ بالطقس بدقة عالية، واستخدمت الدراسة تقنيات Support Vector Regression (SVR) و Artificial Neural Network (ANN) لاستخراج البيانات، اشتملت بيانات الدراسة على نوعين من بيانات الطقس (هطول الامطار ودرجة الحرارة) لمدة ستة سنوات من منطقة العاصمة شيتاغونغ من إدارة الأرصاد الجوية في بنغلاديش، اظهرت نتائج هذه الدراسة أظهرت نتائج SVR أفضل للتنبؤ بهطول الأمطار، وأن ANN اظهرت نتائج افضل للتنبؤ بدرجة الحرارة.

الجدول رقم (1) يظهر مقارنة بين الدراسات السابقة حول التنبؤ بهطول الامطار والدراسة الحالية

الجدول رقم (1) يظهر مقارنة بين الدراسات السابقة حول التنبؤ بهطول الامطار والدراسة الحالية

Authors	Region	Dataset	Features	Measure
Nikhil et al. (2021) [13]	Australian	-	Date, Location, Min Temp, Max Temp, Rainfall, Evaporation, Sunshine, Wind Gust Direction, Wind Gust Speed, Wind Dir 9 am, Wind Dir 3 pm, Wind Speed 9 am, Wind Speed 3 pm, Humidity 9 am, Humidity 3 pm, Pressure 9 am, Pressure 3 pm, Cloud 9 am, Cloud 3 pm, Temp 9 am, Temp 3 pm, Rain Today, RISK, Rain Tomorrow	Accuracy, Precision, Recall, F1score, AUC
Basha et al. (2020) [14]	India	-	Rainfall	MSE, RMSE
Poornima et al.(2019) [15]	Hyderabad	1980-2014	Max and Min Temperature, Wind Speed, Sunshine, Minimum and Maximum Relative Humidity, Evapotranspiration, Rainfall	Accuracy, RMSE, loss, LR: Learning rate of network, No. of epochs
Tharun et al.(2018) [17]	Coonoor-India	2005-2014	daily Temperature, daily humidity, daily cloud speed, daily windspeed, daily wind direction	R-square, adjusted R-square
Aftab et al. (2018) [18]	Lahore	2005-2017	Temperature, Atmospheric Pressure (weather station), Atmospheric Pressure (sea level), Pressure Tendency, Relative Humidity, Mean Wind Speed, Minimum Temperature, Maximum Temperature, Visibility, Dew Point Temperature	Precision, recall, f-measure
Kashiwao et al.(2017) [19]	Japan	2000-2012	Temperature, Humidity, Atmospheric Pressure, Amount of Precipitation, Vapor Pressure and Wind Velocity	Total hit rate, Hit rate of precipitation and Hit rate of non precipitation, Overlooking rate, Swing and miss rate, Caching rate, Confusion Matrix
Qiu et al. (2017) [20]	China	2002-2015	Rain condition, Observatory height, Wind speed, Wind direction, Dew point, Temperature, Air pressure, Humidity	MSE, MSE, Correlation, CSI: Critical Success Index
Rasel et al. (2017) [21]	Chittagong Bangladesh	6-years	Rainfall, Temperature	RMSE, MAE

3. الحل المقترح

هذا القسم يشمل ثلاث مواضيع؛ فكرة الحل المقترح، ونموذج الحل المقترح العام التي توضح خطوات الحل وفقا للفكرة، ثم تطبيق الحل المقترح وفقا للنموذج.

1.3 فكرة الحل المقترح

الحل المقترح هو بناء نموذج لتحديد افضل الميزات لاستخدامها في عملية التنبؤ بمعدلات هطول الامطار، ويتألف النموذج من عدة خطوات، أولا تحديد الهدف وتشمل تحديد مصدر البيانات وتحديد الطرق التي تستخدم لاختيار الميزات، ثانيا جمع البيانات، ثالثا تجهيز البيانات لتناسب الطرق التي تم تحديدها في الخطوة الأولى، رابعا تحديد خوارزميات التقييم للطرق حسب الدقة، خامسا تجربة كل طريقة و تقييمها بواسطة خوارزميات التقييم حسب الدقة، سادسا اختيار افضل الطرق اعتمادا على نتائج التقييم في الخطوة السابقة، و أخيرا تحديد أفضل الميزات حسب افضل طريقة تم اختيارها في الخطوة السابقة.

2.3 نموذج الحل المقترح

يوضح الشكل رقم (1) خطوات الحل المقترح في النموذج بدءًا من تحديد الاهداف ثم جمع البيانات من المستودع عبر الإنترنت، حتى الخطوة السابعة والأخيرة وهو اختيار أفضل الميزات.

الخطوة الأولى هي تحديد الاهداف و تشمل تحديد مصدر البيانات وطرق اختيار الميزات، الخطوة الثانية هي جمع البيانات من مصادرها، الخطوة الثالثة هي تجهيز البيانات واختيار الميزات حيث تحتوي على عدة عمليات وأهمها التحويل، الخطوة الرابعة هي تحديد خوارزميات التقييم للطرق وهي 4 خوارزميات، الخطوة الخامسة هي تجربة كل طريقة وتقييمها من خلال خوارزميات التقييم حسب معيار الدقة، والخطوة السادسة هي اختيار أفضل طريقة حسب نتيجة التقييم، الخطوة السابعة والأخيرة هي تحديد أفضل الميزات المؤثرة في التنبؤ بمعدلات هطول الامطار.

الشكل (1) نموذج الحل المقترح العام

يتم التقييم باستخدام معيار الدقة في خوارزميات التقييم وهي خوارزميات/نماذج تصنيف تُستخدم المعادلة التالية لقياس دقة طرق اختيار الميزات/الميزات Accuracy [22] و [23]:

دقة التصنيف Accuracy هي عدد العينات التي صنفت بشكل صحيح إلى العدد الكلي للعينات.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$ المعادلة (1)

3.3 تطبيق الحل المقترح

تم استخدام لغات وبرامج لتجهيز البيانات وهي لغة Python من خلال محرر Jupyter Notebooks لتنفيذ التعليمات البرمجية في برنامج Anaconda Navigator V2.1.4، والذي يستخدم مكتبات pandas، وNumPy، وScikit-learn Python library. وتم التنفيذ على جهاز حاسوب محمول Laptop شركة لينوفو بذاكرة 4 جيجابايت، ومعالج انتل Corei5-8250U 1.60 قيقا هيرتز، ونوع النظام 64 بت، ونظام تشغيل ويندوز 10 برو نسخة 21H2. وتطبيق الخطوات في النموذج السابق:

الخطوة الأولى: تحديد الاهداف

تم تحديد مصدر بيانات هطول الامطار في السودان وهو من مستودع بيانات وكالة ناسا الفضائية عبر الإنترنت https://power.larc.nasa.gov/data-access-viewer/ .

تم تحديد طرق او خوارزميات اختيار الميزات التالية لتوفرها وسهولة تطبيقها و مناسبتها مع نوعية بيانات هطول الامطار و هي حوالي 10 طرق وهي: importance of random forest classifier, Lasso, Persons Correlation Coefficient, ANOVA , Forward selection, Backward selection, Recursive Feature Elimination، Information gain ، Correlation، و Importance Features.

الخطوة الثانية: جمع البيانات

تم تنزيل مجموعة البيانات من المصدر المحدد مسبقا وتتضمن 216.972 سجلاً و35 ميزة والتي تمثل البيانات اليومية لعناصر الأرصاد الجوي في الفترة من يناير 2000م وحتى ديسمبر 2021م لـ 27 محطة إرصاد جوية على مستوى دولة السودان، وموضح في الشكل رقم (2).

	station	YEAR	MO	DY	ALLSKY_SFC_SW_DWN	CLRSKY_SFC_SW_DWN	ALLSKY_KT	ALLSKY_SFC_LW_DWN	ALLSKY_SFC_PAR_TOT	CLRSKY_SFC
0	Abu Hamad	2000	1	1	4.74	4.74	0.66	343.80	92.40	
1	Abu Hamad	2000	1	2	4.76	4.76	0.66	342.10	93.40	
2	Abu Hamad	2000	1	3	4.74	4.74	0.66	344.60	93.10	
3	Abu Hamad	2000	1	4	4.41	4.39	0.61	353.70	88.00	
4	Abu Hamad	2000	1	5	4.62	5.11	0.63	373.20	93.30	
...	...	...	...	...	...	...	...	...	...	
216967	Zalengei	2021	12	27	4.63	5.75	0.56	356.91	84.66	
216968	Zalengei	2021	12	28	5.70	5.79	0.69	353.03	106.71	
216969	Zalengei	2021	12	29	4.98	5.69	0.60	357.78	91.94	
216970	Zalengei	2021	12	30	6.15	6.13	0.74	331.33	113.66	
216971	Zalengei	2021	12	31	6.15	6.15	0.74	332.70	113.35	

216972 rows × 35 columns

شكل (2) لقطة من شاشة البيانات الاولية

الخطوة الثالثة: تجهيز البيانات

سيتم إعداد البيانات التي تم جمعها للتحليل بواسطة خوارزميات التعلم الآلي بحيث تصبح البيانات صالحة في الشكل والسياق الصحيحين. يوضح الشكل رقم (2) تنسيق البيانات قبل عملية التحويل، حيث يتم تحويل البيانات إلى تنسيق رقمي ليتم التعامل معها بواسطة خوارزميات التعلم الآلي كما موضح في الشكل رقم (4).

يوضح الشكل رقم (3) أعلاه معلومات حول البيانات، بما في ذلك نوع بنية البيانات، إطار البيانات (Data Frame)، كما يعرض أيضًا الميزات وأطوالها وعددها ونوع البيانات في كل ميزة بالإضافة إلى عدد السجلات وما إذا كانت هناك قيم مفقودة في البيانات.

وتجرى عدة نشاطات مثل: التحويل في الشكل رقم (4)، وأيضا تبين أنه لا توجد قيم مفقودة كما موضح في الشكل رقم (5)، وفي حذف القيم المكررة تبين أنه لا توجد قيم مكررة موضحة في الشكل رقم (6)، وفي إزالة القيم المتطرفة وتطبيع البيانات موضحة في الشكل رقم (7)، تحويل البيانات الى فئات موضح في الشكل رقم (8)، وترميز البيانات الفئوية موضح في الشكل رقم (9)، حيث يتم الاحتياج للفئات لغرض استخدامها في خوارزميات التصنيف لتقييم طريقة اختيار الميزات، وموازنة الفئات في الشكل رقم (10).

الخطوة الرابعة: تحديد خوارزميات تقييم طرق اختيار الميزات

تم تحديد أربع خوارزميات لتقييم طرق او خوارزميات اختيار الميزات وهي خوارزميات تصنيف للتنبؤ بمعدلات هطول الامطار كفئة من ضمن الفئات وهي: أقرب الجيران K–Nearest Neighbor (KNN)، شجرة القرار Decision Tree (DT) ، الغابة العشوائية Random Forest (RF)، والتعبئة Bagging (B).

تم اختيار هذه الخوارزميات نسبة لسرعتها في التدريب والاختبار لتقييم طرق اخيار الميزات من حيث معيار الدقة.

الخطوة الخامسة: تجربة طرق اختيار الميزات وتقييمها حسب الدقة

تم تجربة طرق اختيار الميزات في الخطوة الأولى، وكل طريقة حددت عدد من الميزات، ولاختبار مستوى الدقة لهذه الميزات لكل طريقة تم تقييمها باستخدام خوارزميات التقييم (تصنيف حسب الفئات) في الخطوة السابقة من خلال معيار الدقة للتقييم للتنبؤ بمعدلات هطول المطار، ونتائج هذه التجربة موضحة في الجدول رقم (2).

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 216972 entries, 0 to 216971
Data columns (total 35 columns):
 #    Column                Non-Null Count    Dtype
---   ------                --------------    -----
 0    station               216972 non-null   object
 1    YEAR                  216972 non-null   int64
 2    MO                    216972 non-null   int64
 3    DY                    216972 non-null   int64
 4    ALLSKY_SFC_SW_DWN     216972 non-null   float64
 5    CLRSKY_SFC_SW_DWN     216972 non-null   float64
 6    ALLSKY_KT             216972 non-null   float64
 7    ALLSKY_SFC_LW_DWN     216972 non-null   float64
 8    ALLSKY_SFC_PAR_TOT    216972 non-null   float64
 9    CLRSKY_SFC_PAR_TOT    216972 non-null   float64
 10   ALLSKY_SFC_UVA        216972 non-null   float64
 11   ALLSKY_SFC_UVB        216972 non-null   float64
 12   ALLSKY_SFC_UV_INDEX   216972 non-null   float64
 13   WS2M                  216972 non-null   float64
 14   T2M                   216972 non-null   float64
 15   T2MDEW                216972 non-null   float64
 16   T2MWET                216972 non-null   float64
 17   TS                    216972 non-null   float64
 18   T2M_RANGE             216972 non-null   float64
 19   T2M_MAX               216972 non-null   float64
 20   T2M_MIN               216972 non-null   float64
 21   QV2M                  216972 non-null   float64
 22   RH2M                  216972 non-null   float64
 23   PS                    216972 non-null   float64
 24   WS10M                 216972 non-null   float64
 25   WS10M_MAX             216972 non-null   float64
 26   WS10M_MIN             216972 non-null   float64
 27   WS10M_RANGE           216972 non-null   float64
 28   WD10M                 216972 non-null   float64
 29   WS50M                 216972 non-null   float64
 30   WS50M_MAX             216972 non-null   float64
 31   WS50M_MIN             216972 non-null   float64
 32   WS50M_RANGE           216972 non-null   float64
 33   WD50M                 216972 non-null   float64
 34   PRECTOTCORR           216972 non-null   float64
dtypes: float64(31), int64(3), object(1)
memory usage: 57.9+ MB
```

الشكل (3) ملخص البيانات عن كل الميزات

	station	YEAR	MO	DY	ALLSKY_SFC_SW_DWN	CLRSKY_SFC_SW_DWN	ALLSKY_KT	ALLSKY_SFC_LW_DWN	ALLSKY_SFC_PAR_TOT	CLRSKY_SFC_
0	1	2000	1	1	4.74	4.74	0.66	343.80	92.40	
1	1	2000	1	2	4.76	4.76	0.66	342.10	93.40	
2	1	2000	1	3	4.74	4.74	0.66	344.60	93.10	
3	1	2000	1	4	4.41	4.39	0.61	353.70	88.00	
4	1	2000	1	5	4.62	5.11	0.63	373.20	93.30	
...	...	...	...	...	...	...	...	...	...	
216967	27	2021	12	27	4.63	5.75	0.56	356.91	84.66	
216968	27	2021	12	28	5.70	5.79	0.69	353.03	106.71	
216969	27	2021	12	29	4.98	5.69	0.60	357.78	91.94	
216970	27	2021	12	30	6.15	6.13	0.74	331.33	113.66	
216971	27	2021	12	31	6.15	6.15	0.74	332.70	113.35	

216972 rows × 35 columns

الشكل (4) البيانات بعد عملية التحويل

Missing Value

```
In [7]: missing = pd.DataFrame({'missing':df.isnull().sum()})
        missing
```

Out[7]:

	missing
station	0
YEAR	0
MO	0
DY	0
ALLSKY_SFC_SW_DWN	0
CLRSKY_SFC_SW_DWN	0
ALLSKY_KT	0
ALLSKY_SFC_LW_DWN	0
ALLSKY_SFC_PAR_TOT	0
CLRSKY_SFC_PAR_TOT	0
ALLSKY_SFC_UVA	0
ALLSKY_SFC_UVB	0
ALLSKY_SFC_UV_INDEX	0
WS2M	0
T2M	0

الشكل (5) عدد القيم المفقودة لبعض الميزات في المجموعة البيانية

Duplicates

```
In [8]: df1 = pd.DataFrame(df)
        newdf = df1.drop_duplicates()

In [9]: newdf.shape
Out[9]: (216972, 35)
```

الشكل (6) التحقق من وجود السجلات المتكررة في المجموعة البيانية

Feature scaling

```
In [45]: from sklearn.preprocessing import StandardScaler
         sc = StandardScaler()
         X_train = sc.fit_transform(X_train)
         X_test = sc.transform(X_test)

In [46]: X_train
Out[46]: array([[ 0.        ,  0.39558198,  1.58920867, ..., -1.67701414,
                 -0.62276962,  0.49928345],
                [ 0.        , -1.18204016, -0.7310636 , ...,  1.39698879,
                 -0.16998365, -0.40225334],
                [ 0.        , -1.02427794, -0.44102957, ...,  1.35299248,
                  0.30983431, -0.8303804 ],
                ...,
                [ 0.        ,  0.55334419, -1.02109764, ..., -1.11271367,
                 -0.79751218,  0.55691594],
                [ 0.        ,  1.65767969,  0.42907253, ...,  0.86903309,
                 -0.88826246,  0.84919499],
                [ 0.        ,  1.49991748,  0.1390385 , ...,  0.60122947,
                  0.67476628,  0.03822354]])
```

الشكل (7) تحجيم البيانات وتطبيعها ومسح القيم المتطرفة في المجموعة البيانية

Convert Continous to Category target variable

```
In [16]: category = pd.cut(obj_df.PRECTOTCORR  ,bins=[-1,0.1,0.2,0.4,0.8,1.6,3.2,6.4,12.8,25.6,51.2,102.4,204.8],
                           labels=[1,2,3,4,5,6,7,8,9,10,11,12])
         obj_df.insert(12,'rain_group',category)
         obj_df.head(10)
```

Out[16]:

	station	YEAR	MO	DY	ALLSKY_SFC_SW_DWN	CLRSKY_SFC_SW_DWN	ALLSKY_KT	ALLSKY_SFC_LW_DWN	ALLSKY_SFC_PAR_TOT	CLRSKY_SFC_PAR_
0	1	2000	1	1	4.74	4.74	0.66	343.8	92.4	-9
1	1	2000	1	2	4.76	4.76	0.66	342.1	93.4	-9
2	1	2000	1	3	4.74	4.74	0.66	344.6	93.1	-9
3	1	2000	1	4	4.41	4.39	0.61	353.7	88.0	-9
4	1	2000	1	5	4.62	5.11	0.63	373.2	93.3	-9
5	1	2000	1	6	5.08	5.09	0.70	357.4	102.8	-9
6	1	2000	1	7	4.34	4.60	0.59	351.6	88.3	-9
7	1	2000	1	8	4.95	5.00	0.67	339.1	100.0	-9
8	1	2000	1	9	4.81	4.89	0.66	334.5	97.4	-9
9	1	2000	1	10	4.38	4.38	0.59	327.6	87.6	-9

10 rows x 36 columns

الشكل (8) تحويل البيانات الى 12 فئة

```
In [20]: obj_df.rain_group.value_counts()

Out[20]: 1     158087
         6       9032
         5       8420
         7       7907
         4       6943
         9       6748
         3       6021
         2       5160
         8       4809
         10      3723
         11       106
         12        16
         Name: rain_group, dtype: int64
```

الشكل (9) يوضح ترميز فئات البيانات وعدد عناصرها

```
Imbalanced

In [47]: from imblearn.over_sampling import SMOTE

In [48]: smote = SMOTE()

In [49]: X_train_smote, y_train_smote = smote.fit_resample(X_train.astype('float'),y_train)

In [50]: from collections import Counter
         print("before", Counter(y_train))
         print("after", Counter(y_train_smote))

         before Counter({1: 110730, 6: 6224, 5: 5911, 7: 5472, 4: 4881, 9: 4726, 3: 4241, 2: 3596, 8: 3371, 10: 2640, 11: 77})
         after Counter({1: 110730, 2: 110730, 9: 110730, 3: 110730, 6: 110730, 8: 110730, 4: 110730, 7: 110730, 5: 110730, 10: 110730, 1
         1: 110730})
```

الشكل (10) يوضح موازنة الفئات

الجدول رقم (2) نتائج تجربة وتقييم طرق اختيار الميزات من حيث الدقة

Method	No of Features	No & Selected Features	KNN	DT	RF	B
1. Forward selection (Supervised)	**12**	**'station', 'YEAR','MO','DY','CLRSKY_SFC_SW_DWN', 'ALLSKY_SFC_LW_DWN', 'ALLSKY_SFC_UVA','WS2M','T2MWET','T2M_MIN','RH2M','PS'**	**77.6**	**73.8**	**78.6**	**76.7**
2. Information gain **(Supervised)**	**14**	'MO', 'CLRSKY_SFC_SW_DWN', 'ALLSKY_KT', 'ALLSKY_SFC_LW_DWN', 'T2M','T2MDEW', 'T2MWET', 'T2M_RANGE', 'T2M_MAX', 'T2M_MIN', 'QV2M', 'RH2M','PS', 'WD10M']	73.8	73.1	76.5	75.5
3. Pearson Correlation Coefficient (Unsupervised)	17	'ALLSKY_SFC_LW_DWN','ALLSKY_SFC_PAR_TOT','ALLSKY_SFC_UVA','ALLSKY_SFC_UVB','CLRSKY_SFC_SW_DWN','QV2M','RH2M','T2MWET', 'T2M_MAX','T2M_MIN','TS','WD50M',	72.6	70.5	76.3	74.6

		'WS10M_MAX','WS10M_MIN', 'WS50M','WS50M_MAX','WS50M_MIN'				
4. Correlation (Unsupervised)	16	‘ALLSKY_SFC_LW_DWN', 'ALLSKY_SFC_PAR_TOT', 'ALLSKY_SFC_UVA', 'ALLSKY_SFC_UVB','QV2M', 'RH2M','T2M_MAX','T2M_MI’, 'TS','WD50M','WS10M_MAX',' WS10M_MIN','WS50M','WS50M_MAX','WS50M_MIN'	72.2	70.3	76.1	74.3
5. Importance of Random Forest (Unsupervised)	12	'ALLSKY_KT', 'ALLSKY_SFC_LW_DWN', 'T2M', 'T2MDEW','T2MWET', 'TS','T2M_RANGE', 'T2M_MAX', 'T2M_MIN', 'QV2M', 'RH2M', 'PS'	72.8	72.7	75.8	75.1
6. Backward Selection (Supervised)	12	ALLSKY_SFC_SW_DWN', 'CLRSKY_SFC_SW_DWN', 'ALLSKY_SFC_LW_DWN', 'ALLSKY_SFC_PAR_TOT', 'T2MDEW','T2MWET','T2M_RANGE','QV2M','RH2M','PS','WS10M','WS50M'	73	71	75.8	74.3
7. Recursive Feature Elimination (Supervised)	11	'ALLSKY_SFC_LW_DWN', 'CLRSKY_SFC_PAR_TOT', 'T2M', 'T2MDEW', 'T2MWET','TS', 'T2M_RANGE', 'T2M_MAX', 'T2M_MIN', 'QV2M'RH2M', 'PS'	72.8	71	75.5	74.3
8. Feature Importance (Supervised)	12	'QV2M', 'RH2M','T2MDEW', 'T2MWET','T2M_MAX','T2M','PS',' T2M_RANGE','T2M_MIN', 'MO','TS','WD10M'	73.5	72.2	75.5	74.8
9. ANOVA (Supervised)	12	'ALLSKY_SFC_LW_DWN','ALLSKY_SFC_UVB','T2MWET',' TS','QV2M','RH2M','WS10M',' WS10M_MAX','WS10M_MIN',' WD10M','WS50M_RANGE','WD50M'	71.6	69.7	75.3	73.8
10. Lasso (Unsupervised)	7	['T2MWET', 'ALLSKY_SFC_UVA', 'CLRSKY_SFC_PAR_TOT', 'ALLSKY_SFC_UV_INDEX',' WD50M','T2MDEW','RH2M']	69.7	68.6	73.1	71.8

الخطوة السادسة: اختيار أفضل طريقة حسب نتيجة التقييم بمعيار الدقة

حسب الجدول رقم (2) اتضح بأن أفضل خوارزمية هي خوارزمية التسلسل الأمامي Forward selection، وهي أحرزت حسب التجارب أعلى معدلات دقة في كل خوارزميات التقييم (التصنيف).

الخطوة السابعة: تحديد أفضل الميزات المؤثرة في التنبؤ بمعدلات هطول الأمطار حسب الدقة

وفقا للجدول رقم (2)، تبين أن أفضل الميزات عددها 12 عامل وهي:

'station', 'YEAR','MO','DY','CLRSKY_SFC_SW_DWN', 'ALLSKY_SFC_LW_DWN', 'ALLSKY_SFC_UVA','WS2M','T2MWET','T2M_MIN','RH2M','PS'.

الجدول رقم (3) معاني الميزات المستخدمة في التنبؤ بمعدلات هطول الأمطار

كود الميزة	المعنى
Station	اسم المحطة
YEAR	العام
MO	الشهر
DY	اليوم
ALLSKY_SFC_SW_DWN	كل اشعاع الموجات القصيرة الهابطة من سطح السماء
CLRSKY_SFC_SW_DWN	اشعاع الموجات القصيرة الصافي الهابطة من سطح السماء
ALLSKY_KT	كل مؤشر صفاء تشمس السماء
ALLSKY_SFC_LW_DWN	كل اشعاع الموجات الطويلة الهابطة من سطح السماء
ALLSKY_SFC_PAR_TOT	مجموع كل الـ PAR من سطح السماء
CLRSKY_SFC_PAR_TOT	مجموع الـ PAR الصافي من سطح السماء
ALLSKY_SFC_UVA	كل اشعاع UVA من سطح السماء
ALLSKY_SFC_UVB	كل اشعاع UVB من سطح السماء
ALLSKY_SFC_UV_INDEX	كل مؤشر UV من سطح السماء
WS2M	سرعة الرياح عند 2 متر
T2M	درجة الحرارة عند 2 متر
T2MDEW	درجة الندى / الصقيع عند 2 متر
T2MWET	درجة حرارة المصباح المبتل عند 2 متر
TS	درجة حرارة سطح الارض
T2M_RANGE	درجة الحرار عند مدى 2 متر
T2M_MAX	درجة الحرار عند 2 متر كحد أقصى
T2M_MIN	درجة الحرارة عند 2 متر كحد أدنى
QV2M	الرطوبة النوعية عند 2متر
RH2M	الرطوبة النسبية عند 2 متر
PRECTOTCORR	معدل هطول الامطار الحقيقي
PS	ضغط السطح
WS10M	سرعة الرياح عند 10 متر
WS10M_MAX	سرعة الرياح عند 10 متر كحد أعلى
WS10M_MIN	سرعة الرياح عند 10 متر كحد أدنى
WS10M_RANGE	سرعة الرياح عند مدى 10 متر
WD10M	اتجاه الرياح عند 10 متر
WS50M	سرعة الرياح عند 50 متر
WS50M_MAX	سرعة الرياح عند 50 متر كحد أعلى
WS50M_MIN	سرعة الرياح عند 50 متر كحد أدنى
WS50M_RANGE	سرعة الرياح عند مدى 50 متر
WD50M	اتجاه الرياح عند 50 متر

4. النتائج

1. تم بناء نموذج لاختيار أفضل ميزات للتنبؤ بمعدلات هطول الامطار من حيث الدقة.
2. تبين من التجارب لـ 10 خوارزميات من منهجيات مختلفة أن أفضل خوارزمية من حيث الدقة لاختيار الميزات هي خوارزمية الاختيار الأمامي المتسلسل Forward selection من منهجية الـ Wrapper.

3. أعلى معدلات دقة تم الوصول اليها لخوارزمية الاختيار الأمامي المتسلسل من خلال خوارزميات التقييم(التصنيف) المستخدمة هي 78.6% باستخدام خوارزمية الغابة العشوائية (Random Forest)، ثم 77.6% باستخدام خوارزمية أقرب الجيران (KNN)، ثم 76.6% باستخدام خوارزمية التعبئة (Bagging)، ثم 73.8% باستخدام خوارزمية شجرة القرار (Decision Tree).
4. تم تحديد أفضل الميزات من خلال تحقيقها لأفضل معدل دقة في التنبؤ بمعدلات هطول الامطار، وعددها 12 وهي (CLRSKY_SFC_SW_DWN, DY, MO, YEAR, Station ALLSKY_SFC_LW_DWN,ALLSKY_SFC_UVA,WS2M,T2MWET,T2M_MIN,RH2M,PS)، ومعاني الميزات موضح في الجدول رقم (3).

5. مناقشة النتائج

تم بناء نموذج يتعامل مع عدة ميزات من بيانات هطول الامطار التي تم جمعها، وتجهيزها، ثم تحديد عدد 10 خوارزميات لتجربتها على هذه الميزات لاختيار أفضلها من خلال استخدام أفضل الميزات لقياس معدل الدقة باستخدام خوارزميات التصنيف.

خوارزمية الاختيار الأمامي المتسلسل Forward selection من منهجية الـ Wrapper تم استخدامها وتم اختيار 12 ميزة باعتبارها حققت أعلى دقة وعند اختيار 11 أو 13 ميزة قلت الدقة لأنها خوارزمية تحت الاشراف Supervised وكذلك الخوارزميات الأخرى تحت الاشراف التي جربت، كما يوجد بعض الخوارزميات غير الخاضعة للأشرف Unsupervised وموضحة في الجدول رقم (2). وتم تقييمها بمعيار الدقة Accuracy من خلال أربع خوارزميات تصنيف (أقرب الجيران K–Nearest Neighbor (KNN)، شجرة القرار Decision Tree (DT)، الغابة العشوائية Random Forest (RF)، والتعبئة Bagging (B))، وأحرزت اعلى معدلات دقة. الخوارزمية التي تليها لاختيار الميزات هي Information gain، ثم خوارزمية Correlation persons، وكلاهما من منهجية التصفية Filtering. والخوارزمية التي أحرزت أقل معدلات دقة هي خوارزمية Lasso وهي من المنهجية المضمنة Embedded. والجدول رقم (2) يوضح خوارزميات اختيار الميزات المجربة بالترتيب التنازلي حسب معدل الدقة لكل خوارزمية تقييم (تصنيف).

تم تحديد الميزات التي تحقق اعلى معدل دقة في التنبؤ بمعدلات هطول الأمطار وعددها 12 ميزة: (CLRSKY_SFC_SW_DWN, DY, MO, YEAR, Station
LLSKY_SFC_LW_DWN,ALLSKY_SFC_UVA,WS2M,T2MWET,T2M_MIN,RH2M,
PS)، ومعاني الميزات موضح في الجدول رقم (3)، وهذه الميزات تم تحديدها من خلال خوارزمية الاختيار الأمامي المتسلسل Forward selection وهي تعتبر أنسب وأدق خوارزمية من بين عشرة خوارزميات من خلال التجارب.

6. الخاتمة

تم بناء نموذج لتحديد أفضل الميزات المؤثرة في التنبؤ بمعدلات هطول الامطار في السودان وهذه الميزات تم تحديدها من خلال خوارزمية الاختيار الأمامي المتسلسل Forward selection وهي تعتبر أنسب

وأدق خوارزمية من بين عشرة خوارزميات من خلال التجارب ، حيث خرجت الدراسة بعدة توصيات؛ وهي بناء نموذج يستوعب عدة ميزات او ميزات في البيئة التي تطرأ، تطوير النموذج بحيث يعمل في مناطق مختلفة غير دولة السودان بدقة أو أكثر دقة، التحقق المستمر عن نقاط ضعف خوارزميات اختيار الميزات وتحديثها حسب الطلب، في المستقبل تجربة طرق أخرى أحدث تحقق أعلى دقة. تميزت هذه الدراسة ببناء نموذج لتحديد أفضل الميزات المؤثرة في التنبؤ بمعدلات هطول الامطار في دولة السودان باختيارها من أفضل خوارزمية اختيار ميزات بعد تجربة 10 خوارزميات باستخدام معيار الدقة في التقييم وباستخدام 4 خوارزميات تقييم.

7. قائمة المصادر والمراجع

1. X. B, Z. M, M. S and N. B. W, "Particle Swarm Optimization for Feature Selection in Classification : A Multi-Objective Approach," *IEEE Explore,* vol. 43, pp. 1656 - 1671, 13 12 2012.
2. A. S. M and L. J, "Feature selection based on mutual information for machine learning prediction of petroleum reservoir properties," *International Conference on IT in Asia (CITA),* pp. 2-7, 4 5 2015.
3. C. R and K. U. A, "A novel filter feature selection method using rough set for short text data," *Journal Pre-proofs,* Vols. 16,1, 2020.
4. H. E, X. B and Z. M, "Differential evolution for filter feature selection based on information theory and feature ranking," *Knowledge-Based Systems,* vol. 140, pp. 103-119, 2018.
5. M. K. M, M. I. M and M. K, "A new wrapper feature selection approach using neural network," *Neurocomputing,* vol. 73, pp. 3273-3283, 2010.
6. V. F. Rodriguez-Galiano, J. A. Luque-Espinar, M. Chica-Olmo and M. P. Mendes , "Feature selection approaches for predictive modelling of groundwater nitrate pollution: An evaluation of filters, embedded and wrapper methods," *Science of the Total Environment,* vol. 624, pp. 661-672, 2018.
7. G. Jesús , O. Julio , D. Miguel , M.-S. Pedro and Q. G. John , "A new multi-objective wrapper method for feature selection – Accuracy and stability analysis for BCI," *Neurocomputing,* 2019.
8. J. D and S. V, "“Feature selection and classification systems for chronic disease prediction: A review," *Egyptian Informatics Journal,* vol. 19, pp. 189-189, 2018.
9. P. Jamshid , A. Mohsen, E. A. Tahereh and . H. O. Mohammad, "An Efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets," *scientific Reports,* pp. 10-15, 2019.
10. L. S. H, M. Z. Member, Q. I and L. G, "An Embedded Feature Selection Method for Imbalanced Data Classification," *IEEE/CAA JOURNAL OF AUTOMATICA SINICA,* pp. 1-13, 2019.
11. L. Meng, "Embedded feature selection accounting for unknown data heterogeneity," *Expert Systems With Applications,* vol. 119, pp. 350 - 361, 2018.
12. . B.-C. V and A.-B. A, "Ensembles for feature selection: A review and future trends," *Information Fusion,* vol. 52, pp. 1 - 12, 2018.
13. Nikhil Oswal, "Predicting Rainfall using Machine Learning Techniques," *TechRxiv,* 2021.
14. C. Z. Basha, N. Bhavana, B. Ponduru and V. Sowmya , "Rainfall Prediction Using Machine Learning & Deep Learning Techniques," in *Proceedings of the International Conference on Electronics and Sustainable Communication Systems (ICESC 2020),* 2020.
15. S. Poornima and M. Pushpalatha, "Prediction of Rainfall Using Intensified LSTM Based Recurrent Neural Network with Weighted Linear Units," *Atmosphere,* vol. 10, no. 11, pp. 1-18, 2019.
16. A. Kala and S. G. Vaidyanathan, "Prediction of Rainfall Using Artificial Neural Network," in *International Conference on Inventive Research in Computing Applications (ICIRCA),* Coimbatore,India, 2018.

17. V. Tharun, P. Ramya and S. R. Devi, "Prediction of Rainfall Using Data Mining Techniques," in *Proceedings of the 2nd International Conference on Inventive Communication and Computational Technologies (ICICCT 2018)*, 2018.
18. S. Aftab, M. Ahmad, N. Hameed, M. S. Bashir, I. Ali and Z. Nawaz, "Rainfall Prediction in Lahore City using Data Mining Techniques," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 4, 2018.
19. T. Kashiwao, K. Nakayama, S. Ando and K. L. Ikeda, "A neural network-based local rainfall prediction sys-tem using meteorological data on the Internet: A case study using data from the Japan Meteorological Agency," in *Applied Soft Computing*, 2017.
20. M. Qiu , P. Zhao, K. Zhang, J. Huang and X. Shi, "A short-term rainfall prediction model using multitask convolutional neural networks," in *2017 IEEE International Conference on Data Mining*, 2017.
21. R. I. Rasel, N. Sultana and P. Meesad, "An Application of Data Mining and Machine Learning for Weather Forecasting," 2017.
22. A. K. V, Classification Of Diabetes Disease Using Support Vector Machine, vol. 3, 2013, pp. 1797-1801.
23. N.-A. N and M. R, "Comparison of Classifiers for the Risk of Diabetes Prediction," *Procedia Computer Science*, vol. 69, pp. 132-142, 2015.