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 ذات خوارزميات تجميعية مستقر للتنبؤ بمعدلات هطول الامطار بدقة في تنقيب البيانات 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/hnsj397

تاريخ النشر: 2022/09/01م تاريخ القبول: 2022/08/15م

المستخلص

التنبؤ بمعدلات هطول الامطار تأخذ أهمية عالية في تحريك دولاب الإنتاج بالدول، تهدف الدراسة الى بناء نموذج تجميعي من خمس خوارزميات ذات دقة وأكثر استقرارا وتدريبها الكافي باستخدام احجام بيانات كبيرة لحل مشكلة عدم الاستقرار والتدريب غير الكافي، تم اقتراح نموذج للحل يتكون من عدة خطوات؛ تحديد الاهداف، جمع البيانات، تجهيز البيانات، اختيار الميزات، تقييم الخوارزميات الأحادية وعددها عشرة خوارزميات من خلال التدريب والاختبار، اختيار أفضل خمس خوارزميات (شجرة القرار Decision Tree ، الغابات العشوائية Random Forest، أقرب جار k-Nearest Neighbor K، تعزيز التدرج الشديد XGB، والتعبئة Bagging) بناءً على الدقة، بناء النموذج التجميعي بأفضل الخوارزميات للتنبؤ من حيث الدقة، تدريبه، اختباره، لغرض التقييم بعد التنبؤ. توصلت الدراسة الى نتائج أهمها تصميم نموذج للتنبؤ بمعدلات هطول الامطار باستخدام النموذج التجميعي، حقق النموذج معدل دقة Accuracy بلغ 77.6%، Precision بلغ 75.6%، Recall بلغ 77.6%، وF1-score بلغ 76.5%، والتنبؤ بمعدلات هطول الامطار باستخدام فئات محددة بالميليمتر مما يحقق الدقة والاستقرار في التنبؤ بمعدل الهطول.

الكلمات المفتاحية: تنقيب البيانات، هطول الامطار، النماذج التجميعية، الدقة، التصويت الصلب.

RESEARCH TITLE

BUILDING A MODEL WITH STABLE AGGREGATE ALGORITHMS TO ACCURATELY PREDICT RAINFALL RATES IN DATA MINING IN THE 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/hnsj397

Published at 01/09/2022 **Accepted at 15/08/2021**

Abstract

Predicting rainfall rates takes high importance in moving the production wheel in countries. The study aims to build an aggregate model of five more accurate and stable algorithms and train them sufficiently using large data volumes to solve the problem of instability and insufficient training. A model for the solution consisting of several steps has been proposed; Goals setting, data collection, data processing, feature selection, evaluation of the ten unary algorithms through training and testing, selection of the best five algorithms (Decision Tree, Random Forest, K-Nearest Neighbor, XGB, and Bagging) based on accuracy, build an aggregate model with the best prediction algorithms in terms of accuracy, train it and test it for the purpose of evaluation after prediction. The study reached the most important results which includes the design of a model to predict rainfall rates using the aggregation model, the model achieved an accuracy rate of Accuracy of 77.6%, Precision of 75.6%, Recall of 77.6%, and F1-score of 76.5%, and prediction of rainfall rates using specific millimeter categories, which It achieves accuracy and stability in forecasting the rate of precipitation.

Key Words: Data mining, Rainfall, Ensemble Models, Accuracy, Hard voting.

1. مقدمة

لا يزال التنبؤ بهطول الأمطار مصدر قلق كبير وقد اجتذب انتباه الحكومات والصناعات وكيانات إدارة المخاطر، فضلاً عن المجتمع العلمي. هطول الأمطار هو عامل مناخي يؤثر على العديد من الأنشطة البشرية مثل الإنتاج الزراعي، ملء خزانات المياه، البناء وتوليد الطاقة، الغابات، السياحة من بين أمور أخرى [1]. يُعد التنبؤ بمعدلات هطول الأمطار أمرًا ضروريًا لأن هذا المتغير هو الأكثر ارتباطًا بالأحداث الطبيعية المعاكسة مثل الانهيارات الأرضية والفيضانات والحركات الجماعية. هذه الحوادث أثرت على المجتمع لسنوات، لذلك فإن وجود نهج مناسب للتنبؤ بمعدلات هطول الأمطار يجعل من الممكن اتخاذ تدابير وقائية وتخفيفية للتخطيط الجيد للاستفادة منها خاصة في الظواهر الطبيعية المعاكسة.

لحل حالة عدم اليقين استخدمت الدراسة تقنيات ونماذج مختلفة للتعلم الآلي لعمل تنبؤات دقيقة وفي الوقت المناسب. حيث تهدف الدراسة بناء نموذج مستقر وأكثر دقة للتنبؤ بمعدلات هطول الامطار باستخدام النماذج ذات الخوارزميات التجميعية لمعالجة مشكلة انخفاض الدقة في النماذج ذات الخوارزميات الاحادية، وتدريبها باستخدام حجم بيانات كبيرة وعدة خوارزميات لضمان استقراره؛ وذلك من خلال توفير دورة حياة تعلم الآلة من البداية إلى النهاية بدءًا من المعالجة المسبقة للبيانات وحتى تنفيذ النماذج وحتى تقييمها. تتضمن خطوات المعالجة المسبقة للبيانات تحويل المميزات، وتحجيم المميزات، واختيار المميزات المثلى. الدراسة تنفذ نماذج مثل شجرة القرار Decision Tree ، الغابات العشوائية Random Forest، أقرب الجيران K Nearest–k Neighbors، تعزيز التدرج الشديد XGB، والتعبئة Bagging. لأغراض التقييم، استخدم الدراسة معايير هي الدقة Accuracy، الحساسية Precision، Recall وScore1F. ومن أجل ذلك استخدمت الدراسة بيانات هيئة الأرصاد الجوية السودانية وذلك بتحليل بيانات عدد 27 محطة في الفترة بين 2000 /01/01 و 2021/12/31م التي تحتوي على عدد 216.972 سجل مشتملا 35 ميزة من الهيئة العامة للإرصاد الجوية السودانية المخزنة في مستودع عبر الانترنت بموقع ناسا الفضائية.

خوارزمية شجرة القرار Decision Tree هي نوع من التعلم الآلي الخاضع للإشراف المستخدم لتصنيف أو عمل تنبؤات بناءً على كيفية الإجابة على مجموعة سابقة من الأسئلة. النموذج هو شكل من أشكال التعلم تحت الإشراف، بمعنى أن النموذج يتم تدريبه واختباره على مجموعة من البيانات التي تحتوي على التصنيف المطلوب [2].

خوارزمية الغابة العشوائية Random Forest تستخدم الطريقة العشوائية بتوليد عينات للتدريب، ثم تقوم بإنشاء شجرة قرار لكل عينة، وفي الخطوة الأخيرة. تجمع الخوارزمية جميع النتائج من شجرة القرار لعمل تنبؤ بناءً على آلية التصويت. يمكن لخوارزمية الغابة العشوائية دمج المتغيرات الضعيفة والقوية والتعامل مع القيم المتطرفة. الى جانب ذلك، لا تتأثر بالتركيب الزائد [3].

خوارزمية أقرب جار K–Nearest Neighbor هي تعتمد على مقياس التشابه أو المسافة [4]. يمكن استخدام هذه الخوارزمية لحل مشاكل نماذج التصنيف والانحدار. يصنف من خلال إيجاد أقرب نقطة مجاورة [5]، وتطبيق المسافة الإقليدية (Euclidean distance) وجيب التمام للتمييز بين السجلات في التدريب والاختبار [6] [7].

خوارزمية تعزيز التدرج الشديد **XGBoost** هي واحدة من خوارزميّات التّعزيز الشائعة بشكل كبير والمستخدمة على نطاق واسع لأنّها ببساطة قوية جداً، وتعتبر مشابهة لخوارزميّة التّعزيز الاشتقاقيّ Gradient Boost لكنّها تحتوي على بعض الميزات الإضافيّة التي تجعلها أقوى بكثير؛ حيث التّدريب سريع جداً ويمكن موازنته أو توزيعه عبر المجموعات. [8]

خوارزمية التعبئة Bagging تُعرف تقنية Bagging أيضًا باسم Bootstrap Aggregation ويمكن استخدامها لحل مشاكل التصنيف والانحدار. بالإضافة إلى ذلك، تعمل خوارزميات Bagging على تحسين درجة دقة النموذج. تتألف Bagging من ثلاث عمليات: bootstrapping، والتدريب الموازي parallel training، والتجميع aggregation. [9].

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.

1.1 موضوع البحث

تناولت الدراسة تقنيات (وخوارزميات) تعدين البيانات في التنبؤ بمعدلات هطول الامطار، ومعالجة لمشكلة ضعف الاستقرار من خلال بناء النموذج ذات الخوارزميات التجمعية بأسلوب التصويت الصلب اعتمادا على المقارنة من حيث الدقة للخوارزميات المكونة له، واستعرضت الدراسة الحالية الدراسات السابقة في مجال علم الأرصاد الجوي وتحديداً الدراسات التي تناولت منهجيات تحليل البيانات لاكتشاف المشكلات التي تواجه نماذج التنبؤ بهطول الأمطار، وتحليل وتلخيص الدراسات المنشورة في دور النشر الدولية في السنوات الخمس الماضية. ثم اتباع منهجية علمية تطبيقية لتحقيق هدف الدراسة.

2.1 مشكلة البحث

المشكلة هي عدم استقرار نماذج التنبؤ عند بنائها اعتماداً على الخوارزمية الأحادية نتيجة لوجود نقاط ضعف في الخوارزميات المستخدمة للتنبؤ بمعدلات هطول الامطار مثل مناسبة الخوارزمية مع نوع بيانات هطول الامطار في السودان وحجم البيانات الكبيرة، بالإضافة الى عدم كفاية التدريب للنماذج لقلة حجم البيانات المستخدمة.

3.1 حدود البحث

الحدود الزمانية لهذه الدراسة 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

تهدف الدراسة الى بناء نموذج مستقر للتنبؤ بمعدلات هطول الامطار مع مراعاة الدقة وتجنب مشاكل النماذج ذات الخوارزمية الواحدة باستقلال نقاط القوة لكل خوارزمية بفكرة النماذج التجميعية، وضمان استقرار النموذج من خلال استخدام عدة خوارزميات وتدريبه باستخدام بيانات بحجم كبير حتى يتم استخدامه في نظام الكتروني من

قبل الهيئة العامة للأرصاد الجوية السودانية.

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 تصميم نموذج للتنبؤ بمعدلات هطول الامطار مع استخدام المنهج التجريبي والتطبيقي.

6.1 إجراءات البحث

تشمل المراحل التالية: جمع البيانات Data Collection، تجهيز البيانات Data Preparation، مقارنة الخوارزميات خلال التجارب باستخدام طريقة الخوارزميات الأحادية لاستخراج أفضل الخوارزميات واختيارها، ثم بناء النموذج باستخدام طريقة الخوارزميات التجميعية، تدريب النموذج Model Training، اختبار النموذج Model Testing، وتقيم النموذج Model Evaluation حيث يتم قياس مستوى كفاءة نموذج التصنيف بعدد التصنيفات الصحيحة وغير الصحيحة في كل قيمة محتملة للمتغيرات التي يتم تصنيفها. من النتائج المكتسبة.

2. الدراسات السابقة

دراسة Ridwan واخرون [10] ، "نموذج التنبؤ بهطول الأمطار باستخدام أساليب التعلم الآلي: دراسة حالة تيرينجانو، ماليزيا"، هدفت هذه الدراسة الى تطوير ومقارنة العديد من نماذج التعلم الآلي (ML) للعثور على النموذج الأكثر دقة وموثوقية وفعالية للتنبؤ بهطول الامطار باستخدام نوعين من الأساليب، استخدمت الدراسة أربع خوارزميات تعلم آلي مختلفة، توصلت نتائج الدراسة الى ان نموذج BDTR المقترح يعطي أفضل دقة في التنبؤ بهطول الأمطار.

دراسة Hailea واخرون [11] ، بعنوان "تحليل هطول الأمطار والتنبؤ باستخدام تقنية التعلم العميق" تم استخدام منهجية التعلم العميق Deep Learning Approach في هذه الدراسة لتحليل بيانات هطول الأمطار في منطقة كارناتاكا. تم استخدام ثلاث طرق من منهجية التعلم العميق للتنبؤ، ومن ثم تمت مقارنة لهذه التقنيات الثلاثة للتنبؤ بهطول الأمطار شهريًا وتم تقييم أداء التنبؤ لهذه التقنيات الثلاثة. أظهرت النتائج أن نموذج LSTM يُظهر أداءً أفضل مقارنةً بـ ANN وRNN للتنبؤ. يُظهر نموذج LSTM أداءً أفضل مع الحد الأدنى لمتوسط النسبة المئوية للخطأ المطلق (MAPE٪) ومتوسط الجذر التربيعي للخطأ (RMSE٪).

دراسة Gowtham وآخرون [12] ، بعنوان "التنبؤ والتحليل الفعال لهطول الأمطار باستخدام تقنيات التعلم الآلي"، في دراسة شملت مقارنة تقنيات التعلم المستخدمة في التنبؤ الانحدار اللوجستي Logistic regression والغابات العشوائية Random forest، توصلت إلى أنه يمكن إجراء المزيد والعديد من التوقعات من خلال تقييم العديد من طرق التصنيف وبإضافة خصائص المناخ في تواريخ الطقس المختلفة، كما أن الانحدار اللوجستي للتنبؤ بهطول الأمطار فعال للغاية ويوفر نتائج دقيقة.

دراسة Basha واخرون [13] ، بعنوان " التنبؤ بهطول الأمطار باستخدام تقنيات التعلم الآلي والتعلم العميق"، في هذه الدراسة تمت مناقشة استخدام منهجية التعلم العميق Deep Learning في التنبؤ بهطول الأمطار باستخدام تعدد الطبقات بمقارنة المعمارية الحالية مع المعماريات السابقة، تمت الإشارة لأهمية قضايا الدقة في التنبؤ نتيجة

للعلاقات غير الخطية بين العوامل المختلفة المستخدمة في التنبؤ بمعدلات الأمطار باستخدام خوارزميات الذكاء الاصطناعي المختلفة.

دراسة Chatterjee واخرون [14] بعنوان" التنبؤ بهطول الأمطار باستخدام نهج الشبكة العصبية الهجينة"، تم تطوير نموذج للتنبؤ بهطول الأمطار فوق ولاية البنغال الغربية. أداء الشبكة العصبية الهجين من حيث قياس الدقة والاستدعاء مقارنة بالشبكة العصبية متعددة الطبقات(Perceptron–Feedforward (MLP–FFN توقع النموذج المقترح هطول الأمطار بدقة عالية بنسبة 89.54٪.

دراسة Haidar واخرون [15] ، بعنوان" التنبؤ بهطول الأمطار شهريًا باستخدام شبكة عصبية التفافية عميقة أحادية البعد"، تم تطوير نموذج شهري للتنبؤ بهطول الأمطار. ومن ثم مقارنة النموذج المطور بالإصدار الأول من برنامج محاكاة المجتمع الأسترالي للمناخ وعدة أنظمة للأرض، وتوصلت النتائج الى ان النموذج المقترح CNN يقدم أداءً أفضل للتنبؤ بهطول الأمطار.

دراسة kala واخرون [16]، بعنوان" التنبؤ بهطول الأمطار باستخدام الشبكة العصبية الاصطناعية"، في هذه الدراسة تم تطوير نموذج للتنبؤ بهطول الأمطار. وبأخذ أربع عوامل في الاعتبار مثل درجة الحرارة والغطاء السحابي وضغط البخار وهطول الأمطار لتحديد هطول الأمطار مسبقًا. يشير النموذج المقترح المستند إلى ANN إلى دقة مقبولة.

دراسة Sulaiman واخرون [17] ،بعنوان" نموذج التنبؤ بهطول الأمطار الغزيرة باستخدام شبكة عصبية اصطناعية للمنطقة المعرضة للفيضانات"، في هذه الدراسة تم اقتراح نموذج للتنبؤ بهطول. تم جمع بيانات هطول الأمطار من إدارة الأرصاد الجوية المحلية في ماليزيا. تم توقع هطول الأمطار. أظهرت نتيجة هذه الدراسة أن TDNN تفوق في الأداء على نموذج ARIMA.

دراسة Kashiwao واخرون [18] ، 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 للتنبؤ بهطول الأمطار، 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) خلال فترة محددة، 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 Rasel واخرون [19] ، 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، 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.

دراسة Parmar واخرون [20] ، بعنوان "تقنيات التعلم الالي للتنبؤ بهطول الامطار: مراجعة"، هدفت هذه الدراسة الى مراجعة الطرق المختلفة المستخدمة للتنبؤ بهطول الأمطار والمشاكل التي قد يواجها الباحثين أثناء تطبيق

مناهج التنبؤ بهطول الأمطار، استعرضت هذه الدراسة مناهج وخوارزميات مختلفة في الشبكة العصبية الاصطناعية Artificial Neural Network(ANN) للتنبؤ بهطول الأمطار.

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

Authors	Region	Data Set	Algorithm	Measures
Ridwan et al.(2021)	Terengganu Malaysia	1985-2019	BLR, BDTR, DFR, NNR	MAE, RMSE, RAE, RSE, R
Gowtha et al. (2021)	India	2015-2018	LR, RF	Accuracy
Kanchan et al. (2021)	Karnataka-India	1901-217	ANN, RNN, LSTM	MAPE, RMSE
Basha et al. (2020)	India	-	MLP, Auto-Encoders Network	MSE, RMSE
Haidar et al. (2018)	Eastern Australia-Innisfail	Jan 1909-Dec2012	deep CNN, MLP, ACCESS-S1	MAE, RMSE, r, NSE
Sulaiman et al.(2018)	One of district in Malaysia	1965-2015	ANN, TDNN, ARIMA model	RMSE, R2
Chatterjeeet al.(2018)	Southern part of West Bengal India	1989-1995	HNN, K-mean Clustering, MLP-FFN	F-measure, Accuracy, Precision, Recall
Kashiwao et al.(2017)	Japan	2000-2012	MLP, Back-propagation, Random Optimization, RBFN	Total hit rate, Hit rate of precipitation and Hit rate of non-precipitation,Over looking rate, Swing and miss rate,Caching rate, Confusion Matrix
Rasel et al. (2017)	Chittagong Bangladesh	6-years	SVR, ANN	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.

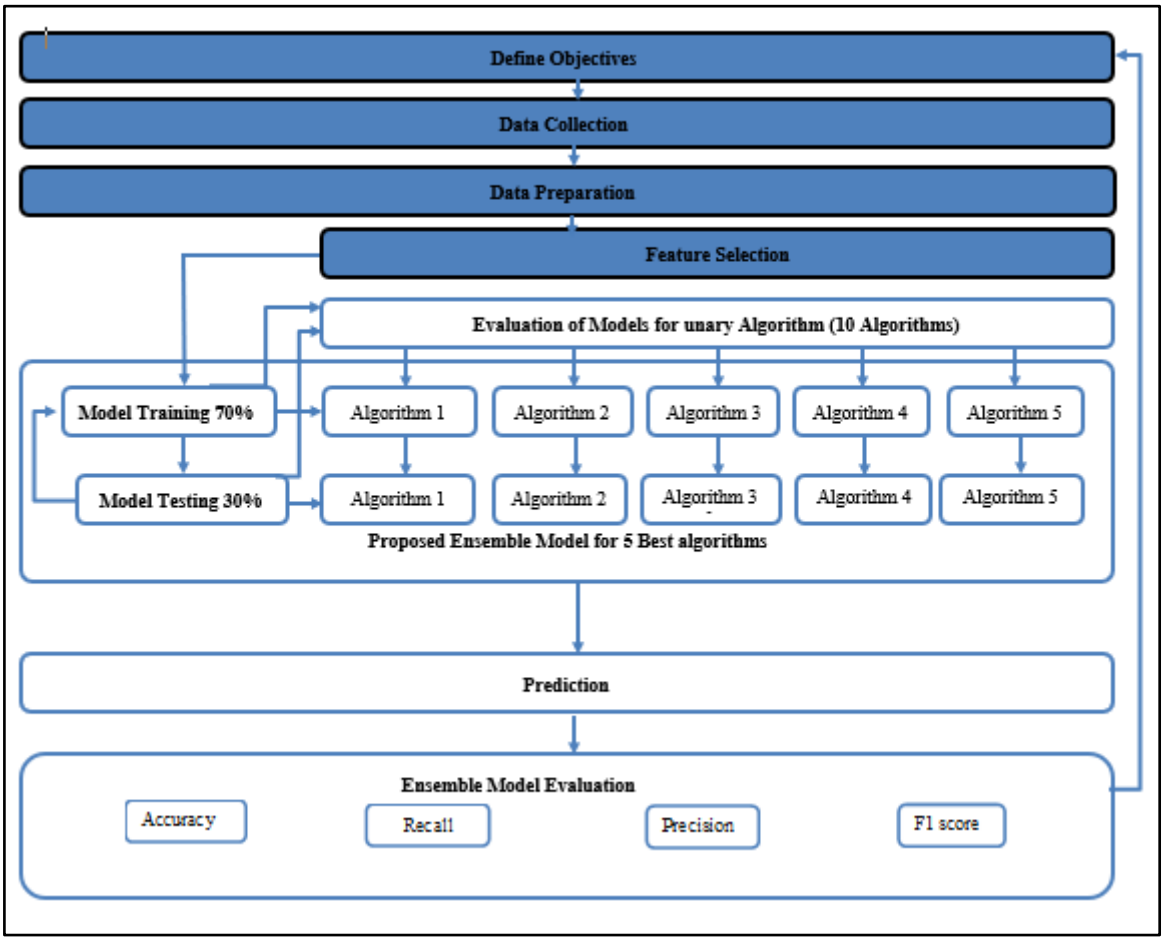
لتحقيق اهدف الدراسة يجب ان تحدد كيفية حل مشكلة الدراسة التي تتمثل في انخفاض الدقة لنماذج التنبؤ ذات

الخوارزمية الاحادية بالإضافة الى عدم كفاية التدريب لقلة حجم البيانات، تم اتباع عدة خطوات هي 10 خطوات؛ اولاً: تحديد الأهداف، ثانياً: جمع بيانات بأحجام كبيرة جدا، ثالثاً: تجهيز البيانات حيث تشمل الاستخلاص، التنظيف، الاستبعاد، التحويل واختيار الميزات المناسبة التي تحقق افضل دقة اعتمادا على بيانات تاريخية عن هطول الامطار، رابعاً: اجراء تجربة للنماذج ذات الخوارزمية الاحادية لمعرفة افضلها وعددها 10 نماذج، خامساً: اختيار افضل 5 نماذج حسب 4 معايير تقييم وهي الدقة Accuracy، الحساسية Precision ، Recall وF1- Score، سادساً: بناء نموذج ذات الخوارزميات التجميعية اعتماداً على النماذج من الخطوة السابقة للاستفادة من نقاط قوتها في الدقة للحصول على أفضل دقة باستخدام التصويت الصلب Hard Voting، سابعاً: تدريب النموذج للحصول على نموذج مستقر، ثامناً: اختبار النموذج للتأكد من صحة النموذج، تاسعاً: تطبيق النموذج للتنبؤ بمعدلات هطول البيانات، وعاشراً: تقييم النموذج للتحقق من مدى تحقيقه للأهداف.

2.3 نموذج الحل المقترح

يوضح الشكل رقم (1) خطوات نموذج الحل المقترح بدءًا من تحديد الاهداف ثم جمع البيانات من المستودع عبر الإنترنت، ثم تليها عملية تمهيدية لأنها جزء أساسي من عملية تصميم التعلم الآلي، ثم بناء النموذج بالاعتماد على تقييم خوارزمياتها الخمسة، ثم عملية تدريب النموذج واختبارها، ثم التنبؤ، ثم التقييم خلال أربع مقاييس أداء لتقييم أداء النموذج.

الخطوة الأولى هي تحديد الاهداف والغرض التي تحققه النموذج، الخطوة الثانية هي جمع البيانات من مصادرها وفقا للأهداف، الخطوة الثالثة هي تجهيز البيانات واختيار الميزات حيث تحتوي على عدة عمليات وأهمها التحويل، الخطوة الرابعة هي تقييم النماذج ذات الخوارزمية الأحادية، الخطوة الخامسة هي اختيار افضل الخوارزميات اعتماداً على تقييمهم، والخطوة السادسة هي بناء النموذج ذات الخوارزميات التجميعية، الخطوة السابعة هي تدريب النموذج بنسبة 70% من حجم البيانات، الخطوة الثامنة هي اختبار النموذج باستخدام 30% من حجم البيانات المتبقية، الخطوة التاسعة هي تنفيذ عملية التنبؤ بالنموذج، الخطوة العاشرة هي تقييم النموذج باستخدام معايير التقييم.



الشكل (1) نموذج الحل المقترح العام

يتم قياس مستوى كفاءة نموذج التصنيف بعدد التصنيفات الصحيحة وغير الصحيحة في كل قيمة محتملة للمتغيرات التي يتم تصنيفها. من النتائج المكتسبة. تُستخدم المعادلات التالية لقياس أداء النموذج Accuracy, Recall, Precision, F1 score [21] و [22]

تم تقييم أداء النموذج باستخدام اربعة مقاييس للأداء:

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

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

2. **الاسترجاع recall** هو عدد النتائج الإيجابية الصحيحة مقسومًا على عدد جميع العينات (عدد الفئات الإيجابية التي يستطيع النموذج التنبؤ بها بشكل صحيح).

$$\text{Recall} = \frac{TP}{TP+FN}$$ المعادلة (2)

3. **الحساسية Precision** هو عدد النتائج الإيجابية الصحيحة مقسومًا على عدد النتائج الإيجابية التي تنبأ بها المصنف .يقيس مدى جودة النموذج عندما يكون التوقع إيجابيًا.

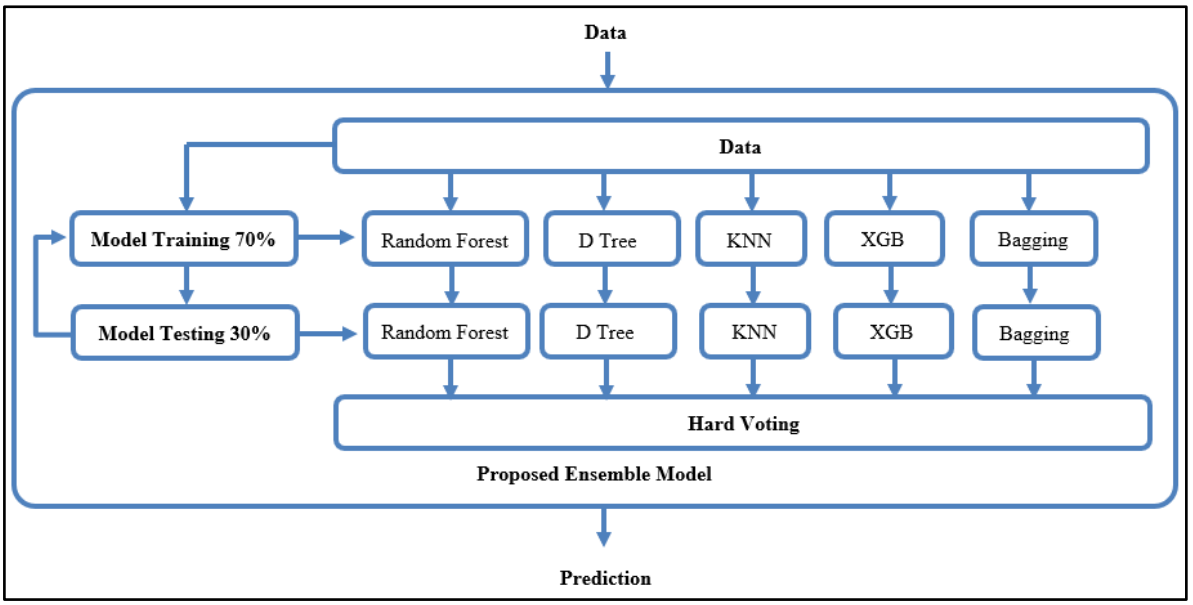
$$Precision = \frac{TP}{TP+FP}$$ المعادلة (3)

4. **F1 score** هو الوسط التوافقي بين precision و Recall

$$F1\ Score = \frac{\text{precision}*\text{Recall}}{\text{precision}+\text{Recall}}$$ المعادلة (4)

الشكل رقم (2) يوضح نموذج التنبؤ ذات الخوارزميات التجميعية التي تم بناءه؛ حيث تبدأ باستقباله

البيانات المجهزة لاختيار الميزات المناسبة للنموذج، ثم استخدام الخوارزميات الخمسة (شجرة القرار Decision Tree، الغابات العشوائية Random Forest، أقرب الجيران k–Nearest Neighbors K، تعزيز التدرج الشديد XGB، والتعبئة Bagging) خلال مرحلة التدريب، ثم اختبار النموذج، ثم استخدام أسلوب التصويت الصلب.



الشكل (2) نموذج التنبؤ ذات الخوارزميات التجميعية

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 H221. وتطبيق الخطوات السابقة:

الخطوة الأولى: تحديد الاهداف

تم تحديد هدفين: تحسين الدقة من خلال بناء نموذج ذات خوارزميات تجميعية للتنبؤ بمعدلات هطول الامطار في دولة السودان، وتدريب النموذج ببيانات ذات حجم كبير.

الخطوة الثانية: جمع البيانات

تم جمع البيانات التي استخدمت في تصميم النموذج من المستودع عبر الإنترنت https://power.larc.nasa.gov/data–access–viewer/ (مجموعة بيانات الإرصاد الجوي)، تتضمن البيانات 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 البلاد، وتعتبر حجم البيانات هي كبيرة جدا ومناسبة للتدريب لجعل النموذج مستقر. والشكل رقم (3) يوضح لقطة من شاشة البيانات الأولية.

	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 (3) لقطة من شاشة البيانات الاولية

الخطوة الثالثة: تجهيز البيانات واختيار المميزات

سيتم إعداد البيانات التي تم جمعها للتحليل بواسطة خوارزميات التعلم الآلي بحيث تصبح البيانات صالحة في الشكل والسياق الصحيحين. وتجرى عدة نشاطات مثل: تنسيق البيانات قبل عملية التحويل في الشكل رقم (3)، حيث يتم تحويل البيانات إلى تنسيق رقمي ليتم التعامل معها بواسطة خوارزميات التعلم الآلي كما موضح في الشكل رقم (4)، وفي حذف القيم المكررة لا توجد قيم مكررة كما في الشكل رقم (5)، وترميز البيانات الفئوية موضح في الشكل رقم (6)، ثم اختيار الميزات المناسبة كما موضح في الشكل رقم (7)، ومسح القيم المتطرفة في الشكل رقم (8)، موازنة الفئات في الشكل رقم (9).

	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

يوضح الشكل رقم (4) أعلاه تنسيق البيانات بعد عملية التحويل لتكون جاهزة للتحليل بواسطة خوارزميات التعلم الآلي حيث تتعرف الخوارزميات على البيانات الرقمية فقط.

يوضح الشكل رقم (5) 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.

```
<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
```

الشكل (5) ملخص البيانات عن كل الميزات

```
#Convert Continous to Category target variable
category = pd.cut(df.PRECTOTCORR  ,bins=[-1,0.1,0.2, 0.4,0.8,1.6,3.2,6.4,12.8,25.6,51.2
                  labels=[1,2,3,4,5,6,7,8,9,10,11,12])
df.insert(12,'rain_group',category)
df.head(10)
```

	station	YEAR	MO	DY	ALLSKY_SFC_SW_DWN	CLRSKY_SFC_SW_DWN	ALLSKY_KT	ALLSKY_SFC_LW_
0	1	2000	1	1	4.74	4.74	0.66	
1	1	2000	1	2	4.76	4.76	0.66	
2	1	2000	1	3	4.74	4.74	0.66	
3	1	2000	1	4	4.41	4.39	0.61	
4	1	2000	1	5	4.62	5.11	0.63	
5	1	2000	1	6	5.08	5.09	0.70	
6	1	2000	1	7	4.34	4.60	0.59	
7	1	2000	1	8	4.95	5.00	0.67	
8	1	2000	1	9	4.81	4.89	0.66	
9	1	2000	1	10	4.38	4.38	0.59	

```
df.rain_group.value_counts()
```

```
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 (6) تحويل البيانات الى 12 فئة

```
In [73]:
X_train, X_test, y_train, y_test = train_test_split(
df[['station','YEAR','MO','DY','ALLSKY_SFC_LW_DWN','ALLSKY_SFC_PAR_TOT',
 'CLRSKY_SFC_PAR_TOT','T2MDEW','RH2M','PS','WS10M_MAX','WD50M']],
   df.rain_group, test_size=0.30, random_state=1)
```

الشكل (7) اختيار الميزات التي تؤثر في الدقة

```
In [9]:  # Handling Outliers

In [10]:
IQR=df.ALLSKY_SFC_SW_DWN.quantile(0.75)-df.ALLSKY_SFC_SW_DWN.quantile(0.25)
lower_bridge=df.ALLSKY_SFC_SW_DWN.quantile(0.25)-(IQR*1.5)
upper_bridge=df.ALLSKY_SFC_SW_DWN.quantile(0.75)+(IQR*1.5)
print(lower_bridge, upper_bridge)

3.9200000000000013 8.879999999999999

In [11]:
df.loc[df['ALLSKY_SFC_SW_DWN']>=8.879999999999999,'ALLSKY_SFC_SW_DWN']=8.879999999999999
df.loc[df['ALLSKY_SFC_SW_DWN']<=3.9200000000000013,'ALLSKY_SFC_SW_DWN']=3.9200000000000013

In [12]:
IQR=df.CLRSKY_SFC_SW_DWN.quantile(0.75)-df.CLRSKY_SFC_SW_DWN.quantile(0.25)
lower_bridge=df.CLRSKY_SFC_SW_DWN.quantile(0.25)-(IQR*1.5)
upper_bridge=df.CLRSKY_SFC_SW_DWN.quantile(0.75)+(IQR*1.5)
print(lower_bridge, upper_bridge)

4.5699999999999985 8.97

In [13]:
df.loc[df['CLRSKY_SFC_SW_DWN']>=8.97,'CLRSKY_SFC_SW_DWN']=8.97
df.loc[df['CLRSKY_SFC_SW_DWN']<=4.5699999999999985,'CLRSKY_SFC_SW_DWN']=4.5699999999999985

In [14]:
IQR=df.ALLSKY_KT.quantile(0.75)-df.ALLSKY_KT.quantile(0.25)
lower_bridge=df.ALLSKY_KT.quantile(0.25)-(IQR*1.5)
upper_bridge=df.ALLSKY_KT.quantile(0.75)+(IQR*1.5)
print(lower_bridge, upper_bridge)

0.47000000000000003 0.8699999999999999
```

الشكل (8) لقطة من مسح القيم المتطرفة

```
In [79]: #imbalanced
smote = SMOTENC(random_state=42, categorical_features=[0,])
X_train_smote, y_train_smote = smote.fit_resample(X_train, y_train)
from collections import Counter
print("before", Counter(y_train))
print("after", Counter(y_train_smote))

before Counter({1: 110713, 6: 6231, 5: 5871, 7: 5526, 4: 4886, 9: 4736, 3: 4221, 2: 3637, 8: 3334, 10: 2636, 11: 79, 12: 10})
after Counter({2: 110713, 1: 110713, 8: 110713, 3: 110713, 6: 110713, 5: 110713, 4: 110713, 7: 110713, 9: 110713, 10: 110713, 11: 110713, 12: 110713})
```

الشكل (9) يوضح عدد عناصر كل فئة قبل وبعد الموازنة

تشير البيانات غير المتوازنة إلى تلك الأنواع من مجموعات البيانات حيث يكون للفئة المستهدفة توزيع غير متساوٍ للعناصر، أي أن تصنيف أحد الفئات يحتوي على عدد كبير جدًا من العناصر والآخر يحتوي على عدد قليل جدًا من العناصر. [23]. والشكل رقم (7 و10) يوضح ذلك، وناتج موازنة الفئات.

الخطوة الرابعة: اجراء التقييم على النماذج ذات الخوارزمية الأحادية

يوجد عشرة نماذج ذات خوارزمية احادية تستخدم للتنبؤ بمعدلات هطول الامطار، تم تقييمهم حسب 4 معايير تقييم وهي الدقة Accuracy، الحساسية Precision، Recall وScore –F1، كما موضح في الشكل رقم (10) وهي لخوارزمية الغابة العشوائية ونفس الاجراء تمت اجراؤها لبقية الخوارزميات التسعة، أما نتيجة التقييم لكل الخوارزميات موضح في الجدول رقم (2) حيث يوضح المقارنة بينهم.

```
In [102…] %%time
DecisionTree(X_train_smote, X_test, y_train_smote, y_test)

Accuracy of DecisionTree classifier on test set: 0.726
precision of DecisionTree classifier on test set: 0.785
Recall of DecisionTree classifier on test set: 0.726
F1-score of DecisionTree classifier on test set: 0.752
Wall time: 1min 21s

In [103…] %%time
KNN(X_train_smote, X_test, y_train_smote, y_test)

Accuracy of KNN classifier on test set: 0.733
precision of KNN classifier on test set: 0.779
Recall of KNN classifier on test set: 0.733
F1-score of KNN classifier on test set: 0.754
Wall time: 21.6 s

In [104…] %%time
RandomForest(X_train_smote, X_test, y_train_smote, y_test)

Accuracy of RandomForest classifier on test set: 0.777
precision of RandomForest classifier on test set: 0.806
Recall of RandomForest classifier on test set: 0.777
F1-score of RandomForest classifier on test set: 0.791
Wall time: 4min 13s

In [105…] %%time
Bagging(X_train_smote, X_test, y_train_smote, y_test)

Accuracy of Bagging classifier on test set: 0.757
precision of Bagging classifier on test set: 0.796
Recall of Bagging classifier on test set: 0.757
F1-score of Bagging classifier on test set: 0.775
Wall time: 4min 37s
```

الشكل (10) نتيجة التقييم الخوارزميات الاحادية (أشجار القرار، أقرب الجيران، الغابة العشوائية، والتعبئة) بالمعايير

الجدول (2) مقارنة بين الخوارزميات المستخدمة بعد اجراء التجربة حسب معايير التقييم

Algorithm	Wall time	Accuracy	Precision	Recall	F1-score
Random Forest	**4min 13s**	**77.7**	**80.6**	**77.7**	**79.1**
Bagging	**4min 37s**	**75.7**	**79.6**	**75.7**	**77.5**
KNN	**21.6 s**	**73.3**	**77.9**	**74.6**	**75.4**
Decision Tree	**1min 21s**	**72.6**	**78.5**	**72.6**	**75.2**
XGB	**35min 5s**	**72.4**	**79.6**	**72.4**	**75.5**
Gradient Boosting	1h 59min 43s	69.2	80.2	69.2	73.8
naive_bayes	779 ms	61.8	76.7	61.8	67.8
Logistic Regression	1min 5s	59.8	76.1	59.8	66.1
SGD	34min 57s	55.1	78.4	55.1	63.2
AdaBoost	3min 12s	55.1	76.4	55.1	59.1

الخطوة الخامسة: اختيار أفضل 5 نماذج من الخطوة السابقة

الاختيار تم بالتركيز على معيار الدقة، فتم اختيار أفضل 5 خوارزميات من حيث معيار الدقة كما موضح في الجدول رقم (2).

الخطوة السادسة: تنفيذ بناء النموذج ذات الخوارزميات التجميعية

تم تصميم النموذج ذات الخوارزميات التجميعية للخوارزميات من أفضل خمس خوارزميات من حيث الدقة الموضحة في الجدول (2)، وتم تنفيذه بأسلوب التصويت الصلب Hard Voting، موضح في الشكل رقم (11).

```
In [37]: rf_clf = RandomForestClassifier(n_estimators=100, random_state=0, n_jobs=-1)
         rf_clf.fit(X_train_smote,y_train_smote)
Out[37]: ▸ RandomForestClassifier

In [38]: dt_clf = tree.DecisionTreeClassifier(random_state=42)
         dt_clf.fit(X_train_smote,y_train_smote)
Out[38]: ▸ DecisionTreeClassifier

In [39]: knn_clf = KNeighborsClassifier(n_neighbors=2)
         knn_clf.fit(X_train_smote,y_train_smote)
Out[39]: ▸ KNeighborsClassifier

In [40]: b_clf = BaggingClassifier()
         b_clf.fit(X_train_smote,y_train_smote)
Out[40]: ▸ BaggingClassifier

In [41]: xgb_clf = XGBClassifier()
         xgb_clf.fit(X_train_smote,y_train_smote)
```

الشكل (11) لقطة من إعداد الخوارزميات للنموذج لمرحلة التدريب

الخطوة السابعة: تدريب النموذج

لا بد من تدريب النموذج لتقييم النموذج بسجلات بنسبة 70%، وهي تساوي 151880 سجل بعدد 13 ميزة، كما موضح في الشكل رقم (12).

حيث التدريب يتم باستخدام 5 خوارزميات مجتمعة لتمثل النموذج التجميعي Ensemble Model،

وموضح في الشكل رقم (13).

In [90]:
```
X_train
```

Out[90]:

	station	YEAR	MO	DY	ALLSKY_SFC_LW_DWN	ALLSKY_SFC_PAR_TOT	CLRSKY_SFC_PAR_TOT	T2MDEW	RH2M	PS	WS10M_MAX	WD50M
102690	1	2017	2	18	351.95	92.02	113.690	11.61	60.50	100.18	6.94	330.62
67940	1	2009	12	31	322.26	111.31	113.480	-1.19	23.44	95.65	9.38	21.56
104638	1	2000	6	19	428.80	121.70	76.475	20.15	66.19	94.65	5.58	187.50
150333	1	2015	7	26	444.82	117.25	136.750	20.06	55.81	95.75	7.90	218.31
127649	1	2019	6	19	424.35	138.73	139.460	16.08	37.25	96.09	8.28	194.94
...	...	...	...	...	...	...	...	...	...	...	...	...
109259	1	2013	2	12	338.73	123.59	124.270	2.08	18.69	94.35	10.52	22.06
50057	1	2005	1	15	310.72	114.26	114.550	-4.57	21.00	94.62	7.10	170.44
5192	1	2014	3	20	331.87	134.42	135.330	-1.07	21.00	97.29	8.44	274.62
208780	1	2021	7	29	412.04	130.65	132.890	6.42	19.19	97.72	7.32	300.94
128037	1	2020	7	11	426.72	139.52	142.440	16.51	38.25	95.91	7.25	232.31

151880 rows × 12 columns

الشكل (12) معلومات عن مميزات وسجلات التدريب (70%) للنموذج

In [42]:
```
voting_clf = VotingClassifier(estimators=[('RFc', rf_clf),
                                          ('DTree', dt_clf),
                                          ('KNNc', knn_clf),
                                          ('Bc', b_clf),
                                          ('XGBc', xgb_clf)],
                              voting='hard')
voting_clf.fit(X_train_smote,y_train_smote)
```

```
C:\Users\Ameen\anaconda3\lib\site-packages\xgboost\sklearn.py:1224: UserWarning: The use of label encoder in XGBClassifier is d
eprecated and will be removed in a future release. To remove this warning, do the following: 1) Pass option use_label_encoder=F
alse when constructing XGBClassifier object; and 2) Encode your labels (y) as integers starting with 0, i.e. 0, 1, 2, ..., [num
_class - 1].
  warnings.warn(label_encoder_deprecation_msg, UserWarning)

[22:12:21] WARNING: D:\bld\xgboost-split_1645118015404\work\src\learner.cc:1115: Starting in XGBoost 1.3.0, the default evaluat
ion metric used with the objective 'multi:softprob' was changed from 'merror' to 'mlogloss'. Explicitly set eval_metric if yo
u'd like to restore the old behavior.
```

Out[42]:

VotingClassifier

RFc	DTree	KNNc	Bc	XGBc
RandomForestClassifier	DecisionTreeClassifier	KNeighborsClassifier	BaggingClassifier	XGBClassifier

الشكل (13) تدريب النموذج ذات الخوارزميات التجميعية بتحديد أسلوب التصويت الصلب Hard Voting

الخطوة الثامنة: اختبار النموذج

لا بد من اختبار النموذج لتقييم النموذج بسجلات بنسبة 30%، وهي تساوي 65092 سجل بعدد 13 ميزة، كما موضح في الشكل رقم (14).

In [83]:
```
X_test.shape, y_test.shape,
```

Out[83]:
```
((65092, 12), (65092,))
```

الشكل (14) معلومات عن مميزات وسجلات الاختبار (30%) للنموذج

الخطوة التاسعة: تطبيق النموذج للتنبؤ بمعدلات هطول الامطار

تطبيق النموذج تنتج عنه التنبؤ برقم تصنيف معدل هطول الامطار التالي بالملمترات. ويتوقع أن تكون نتيجة النموذج عبارة عن أحد الفئات/التصنيفات الافتراضية المكونة من أثني عشر فئة كالآتي:

1. التصنيف 1 يعني أن معدل هطول الامطار محصور بين 1− و1. ملم.
2. التصنيف 2 يعني أن معدل هطول الامطار محصور بين 1. و2. ملم.
3. التصنيف 3 يعني أن معدل هطول الامطار محصور بين 2. و4. ملم.

4. التصنيف 4 يعني أن معدل هطول الامطار محصور بين 4. و8. ملم.
5. التصنيف 5 يعني أن معدل هطول الامطار محصور بين 8.و1.6 ملم.
6. التصنيف 6 يعني أن معدل هطول الامطار محصور بين 1.6 و3.2 ملم.
7. التصنيف 7 يعني أن معدل هطول الامطار محصور بين 3.2 و6.4 ملم.
8. التصنيف 8 يعني أن معدل هطول الامطار محصور بين 6.4 و12.8 ملم.
9. التصنيف 9 يعني أن معدل هطول الامطار محصور بين 12.8 و25.6 ملم.
10. التصنيف 10 يعني أن معدل هطول الامطار محصور بين 25.6 و51.2 ملم.
11. التصنيف 11 يعني أن معدل هطول الأمطار محصور بين 51.2 و102.4 ملم.
12. التصنيف 12 يعني أن معدل هطول الأمطار محصور بين 102.4و 204.8 ملم.

الجدول (3) تطبيق التنبؤ بفئات معدل هطول الامطار (P.C) لعينة من 10 سجلات بيانات متسلسلة لمنطقة زالنجي

station	YEAR	MO	DY	ALLSKY_SFC_SW_DWN	ALLSKY_SFC_PAR_TOT	CLRSKY_SFC_PAR_TOT	T2MDEW	RH2M	PS	WS10M_MAX	WD50M	PRECTOTCORR	R.C	P.C
27	2021	9	1	6.6	126.8	132.36	17.99	67.75	89.33	4.37	105.5	8.4	8	[8]
27	2021	9	2	4.1	78.59	126.3	18.33	71.88	89.4	2.91	189.25	3.55	7	[6]
27	2021	9	3	5.38	104.7	136.87	20.1	86.06	89.39	4.06	161.81	9.13	8	[8]
27	2021	9	4	3.45	68.25	133.52	19.08	78.06	89.42	3.23	103.88	8.1	8	[7]
27	2021	9	5	4.32	84.26	134.14	19.79	84.31	89.52	4.01	148.75	45.17	10	[10]
27	2021	9	6	5.44	105.88	135.34	19.94	86	89.62	4.76	161.44	18.48	9	[9]
27	2021	9	7	4.71	91.48	134	19.71	88.75	89.64	3.7	246.62	6.48	7	[10]
27	2021	9	8	6.3	123.66	143.32	19.29	81.81	89.57	3.48	175.88	6.35	7	[8]
27	2021	9	9	4.95	96.3	129.91	19.53	78.75	89.54	4.82	101.75	10.26	8	[8]
27	2021	9	10	6.35	122.91	130.6	19.37	79	89.62	5.43	116.12	5.88	7	[7]

تم التنبؤ بفئات معدلات هطول الامطار لعينة من سجلات البيانات المتسلسلة من حيث التاريخ في منطقة زالنجي، الفئات الحقيقية رمزت بالكود (R.C) وهي الفئات لمعدلات هطول الامطار الحقيقية في تلك الأيام، والفئات المقابلة التي تم التنبؤ بها رمزت بالكود (P.C)، وموضح في الجدول رقم (3) والشكل رقم (15).

```
In [92]: y_pred = modle.predict([[27,2021,9,1,6.6,126.8,132.36,17.99,67.75,89.33,4.37,105.5]])
         print(y_pred)
         y_pred = modle.predict([[27,2021,9,2,4.1,78.59,126.3,18.33,71.88,89.4,2.91,189.25]])
         print(y_pred)
         y_pred = modle.predict([[27,2021,9,3,5.38,104.7,136.87,20.1,86.06,89.39,4.06,161.81]])
         print(y_pred)
         y_pred = modle.predict([[27,2021,9,4,3.45,68.25,133.52,19.08,78.06,89.42,3.23,103.88]])
         print(y_pred)
         y_pred = modle.predict([[27,2021,9,5,4.32,84.26,134.14,19.79,84.31,89.52,4.01,148.75]])
         print(y_pred)
         y_pred = modle.predict([[27,2021,9,6,5.44,105.88,135.34,19.94,86,89.62,4.76,161.44]])
         print(y_pred)
         y_pred = modle.predict([[27,2021,9,7,4.71,91.48,134,19.71,88.75,89.64,3.7,246.62]])
         print(y_pred)
         y_pred = modle.predict([[27,2021,9,8,6.3,123.66,143.32,19.29,81.81,89.57,3.48,175.88]])
         print(y_pred)
         y_pred = modle.predict([[27,2021,9,9,4.95,96.3,129.91,19.53,78.75,89.54,4.82,101.75]])
         print(y_pred)
         y_pred = modle.predict([[27,2021,9,10,6.35,122.91,130.6,19.37,79,89.62,5.43,116.12]])
         print(y_pred)

[8]
[6]
[8]
[7]
[10]
[9]
[10]
[8]
[8]
[7]
```

الشكل (15) تنفيذ التنبؤ لسجلات البيانات الـ 10 ونتيجته كفئات.

حيث كانت نسبة التطابق بين الفئات الحقيقية والفئات التي تم التنبؤ بها هي 70%، وإذا اخذنا الفئات المتقاربة في السجل الرابع والسجل الثامن بعين الاعتبار تصبح نسبة التطابق حوالي 90%. وإذا تم مقارنتها بمعدلات الدقة نجد أن هنالك تكافؤ.

```
In [87]: # General accuracy score
         print (metrics.accuracy_score(y_pred,y_test))

         0.7763780495298962

In [88]: # General precision score
         print (metrics.precision_score(y_pred,y_test, average='weighted'))

         0.7557720520637343

In [89]: # General Recall score
         print (metrics.recall_score(y_pred,y_test, average='weighted'))

         0.7763780495298962

In [90]: # General f-score score
         print (metrics.f1_score(y_pred,y_test, average='weighted'))

         0.7646628648512404
```

الشكل (16) معدلات معايير التقييم للخوارزمية التجميعية.

الخطوة العاشرة: تقييم النموذج

يتم التقييم بالمقارنة مع الاهداف المحددة بالتركيز على معيار الدقة، بمعدل دقة للنموذج ذات الخوارزميات التجميعية متقاربة لأفضل خوارزمية احادية وموضح في الشكل رقم (16) والجدول رقم (4) مع المقارنة مع أفضل نموذج ذات خوارزمية احادية كما في الشكل رقم (10) وهو أفضل خوارزمية وموضح أيضا في الجدول رقم (2).

الجدول (4) مقاييس التقييم في النموذج ذات الخوارزميات التجميعية

Algorithm	Wall time	Accuracy	Precision	Recall	F1-score
Ensemble Model	43min 48s	77.6	75.6	77.6	76.5

4. النتائج

1. تم تصميم نموذج للتنبؤ بمعدلات هطول الامطار باستخدام النموذج التجميعي ensemble model باستخدام خمس خوارزميات (Random Forest، Decision ،KNN ،Bagging Classifier ،XGB Classifier Tree).
2. حقق النموذج معدل دقة Accuracy بلغ 77.6%، Precision بلغ 75.6%، Recall بلغ 77.6%، وF1-score بلغ 76.5%.
3. من خلال التطبيق لعينة من السجلات نسبة التطابق للتنبؤ لـ 10 أيام متتالية مع المعدل الحقيقي بلغت 70%، وبأخذ الفئات المتقاربة بلغت التطابق 90%.
4. تم التنبؤ بمعدلات هطول الامطار باستخدام فئات محددة بالميليمتر مما يحقق الدقة في التنبؤ بمعدل الهطول.

5. مناقشة النتائج

تم بناء نموذج ادق للتنبؤ بمعدلات هطول الامطار، وحقق النموذج معدل دقة Accuracy بلغ 77.6%، Precision بلغ 75.6%، Recall بلغ 77.6%، وF1-score بلغ 76.5%، ويُلاحظ ان نتائج المقاييس الأربعة المذكورة سابقا متقاربة بدرجة كبيرة، ومتقاربة مع أفضل خوارزمية أحادية من حيث الدقة Accuracy وRecall وF1 بالمقارنة بين الجدول (4) والجدول (2). وبالرغم من ان هذه النتائج مقبولة علميا إلا انها تعتبر منخفضة نسبيا وتحتاج الي تحسين، لضمان استقرار النموذج تم استخدام عدة خوارزميات بحجم بيانات كبيرة في مرحلة التدريب بلغت حوالي 70% من حجم البيانات. وللتحقق من تطبيق نموذج التنبؤ تم اخذ عينة عشوائية لسجلات 10 أيام متتالية ابتداءً من الأول من سبتمبر 2021م حتى العاشر من سبتمبر 2021م في منطقة زالنجي بفئات معدلات هطول الامطار الحقيقية (R.C)، وتنبأ النموذج التجميعي بفئات معدلات هطول الامطار (P.C) الموضحة في الجدول رقم (3) بنسبة تطابق بين الفئات الحقيقية والفئات المتنبئ بها 70%، و اذا اخذنا السجلات رقم 4 و رقم 8 نجد ان الفئات متقاربة جدا بحيث تبلغ نسبة التطابق بالتقارب 90%. نتائج هذا النموذج نتجت من تصميم نموذج ببيانات يومية من محطات قياس في السودان اعتماد هذه النتائج في مناطق اخري يتم عبر مرحلة التحقق ببيانات اخري من مناطق مختلفة.

استخدمت الدراسات السابقة نماذج بخوارزميات أحادية بينما استخدمت هذه الدراسة خمس خوارزميات لتصميم النموذج التجميعي باستخدام أسلوب Hard voting الذي يرفع من اعتمادية نتائج النموذج، واستخدمت بيانات قليلة بينما استخدمت هذه الدراسة بيانات أكبر من حيث عدد السجلات والمتغيرات مما نتج عنه نموذج مدرب بصورة فعالة ونتائج أكثر موثوقية بالإضافة الي ان البيانات المستخدمة لا توجد بها مشاكل مؤثرة في تصميم النموذج.

6. الخاتمة

تم بناء نموذج دقيق نسبيا واكثر استقرارا للتنبؤ بمعدلات هطول الامطار في السودان، حيث خرجت الدراسة بعدة توصيات؛ وهي بناء نموذج يستوعب متغيرات البيئة التي تطرأ، تحسين أكثر لدقة النموذج، تطوير النموذج بحيث يعمل في مناطق مختلفة غير دولة السودان، تطوير نظام معلومات ذكي يستخدم النموذج كتطبيق موبايل، التحقق المستمر عن نقاط ضعف خوارزميات التنبؤ وتحديثها حسب الطلب، تطوير النموذج للتنبؤ بمعدلات هطول الامطار ليعمل بأسلوب التصويت الناعم Soft Voting، تطبيق مفهوم التعلم العميق اذا لم تتوفر بيانات كافية لتطوير النموذج، جمع بيانات ذات احجام كبيرة واستخدام عدة خوارزميات في تدريب النموذج لضمان استقرار أكبر، واستخدام معايير تقييم اضافية لضمان جودة التقييم. تميزت هذه الدراسة ببناء نموذج دقيق وأكثر استقراراً للتنبؤ بمعدلات هطول الامطار في دولة السودان لاستخدامها بالنظام الالكتروني بالهيئة العامة للإرصاد الجوية السودانية.

7. قائمة المصادر والمراجع

1. World Health Organization, "Climate change and human health : risks and responses : summary," World Health Organization, 2003.
2. 2U, Inc, "MastersInDataScience.org is owned and operated," [Online]. Available: https://www.mastersindatascience.org/learning/introduction-to-machine-learning-algorithms/decision-tree/. [Accessed 6 5 2022].
3. B. Boubekeur and S. Güven , "Predicting IPO initial returns using random forest," *Borsa _Istanbul Review*, pp. 13-23, 2020.
4. S. N and G. T, "Analysis of diabetes mellitus for early prediction using optimal features selection," *Journal of Big Data*, vol. 6, 2019.
5. K. H and K. V, "Predictive modelling and analytics for diabetes using a machine learning approach," *Applied Computing Informatics,*, 2018.
6. S. K, K. Z and S. S, "Diagnosis of Diabetes Mellitus using K Nearest Neighbor Algorithm," *International Journal of Computer Science Trends and Technology (IJCST)*, vol. 2, no. 4, pp. 36-43, 2014.
7. k. N. S, P. M and P. G , "Realization of optical Aadder circuit using photonic structure and KNN algorithm," *Optik*, vol. 212, 2020.
8. "خوارزميات تعلم الآلة MACHINE LEARNING ALGORITHMS".
9. G. Joseph , "Bagging algorithms in Python," 22 February 2022. [Online]. Available: https://www.section.io/engineering-education/implementing-bagging-algorithms-in-python/. [Accessed 18 July 2022].
10. W. M. Ridwan, M. Sapitang, A. Aziz, K. F. Kushiar, A. N. Ahmed and A. El-Shafie, "Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia," *Ain Shams Engineering Journal*, vol. 12, no. 2, pp. 1651-1663, 2021.
11. P. Kanchan and N. K. Shardoor, "Rainfall Analysis and Forecasting Using Deep Learning Technique," *Journal of Informatics Electrical and Electronics Engineering (JIEEE)*, vol. 2, no. 2, pp. 1-11, 2021.
12. S. M. Gowtham , S. G. Yenugudhati and . M. A. Mohammad, "Efficient Rainfall Prediction and Analysis using Machine Learning Techniques," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 6, pp. 3467-3474, 2021.
13. 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.
14. S. Chatterjee, B. Datta, S. Sen, N. Dey and C. Narayan , "Rainfall prediction using hybrid neural network approach," in *2018 2nd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom)*, Ho Chi Minh, 2018.
15. A. Haidar and B. Verma, "Monthly Rainfall Forecasting Using One-Dimensional Deep Convolutional Neural Network," *IEEE Access*, vol. 6, pp. 69053 - 69063, 2018.
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. J. Sulaiman and S. H. Wahab, "Heavy Rainfall Forecasting Model Using Artificial Neural Network for Flood Prone Area," in *IT Convergence and Security 2017*, 2018.
18. 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.
19. R. I. Rasel, N. Sultana and P. Meesad, "An Application of Data Mining and Machine Learning for Weather Forecasting," 2017.

20. A. Parmar, M. Sompura and K. Mistree, "Machine Learning Techniques For Rainfall Prediction: A Review," in *2017 International Conference on Innovations in information Embedded and Communication Systems (ICIIECS)*, 2017.
21. A. K. V, Classification Of Diabetes Disease Using Support Vector Machine, vol. 3, 2013, pp. 1797-1801.

N.-A. N and M. R, "Comparison of Classifiers for the Risk of Diabetes Prediction," *Procedia Computer Science*, vol. 69, pp. 132-142, 2015.

M. Saikat , "5 Techniques to Handle Imbalanced Data For a Classification Problem," 2021.