مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (5) العدد(7) 2021

Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (5) No. (7) 2021

مقارنة أداء خوارزميّات التعلّم الألي للتنبؤ ببيانات تحسسيّة

رهام رجب حسن *

(تاريخ الإيداع 9 /6 /2021 . قُبِل للنشر في 18 /8 /2021)

□ ملخّص □

تطوّرت تكنولوجيا الذكاء الصنعي بشكل كبير في السنوات الأخيرة وأصبحت أداة رئيسيّة في جميع القطّاعات ،ومن ميّزات الذكاء الصنعي أنه يستطيع معالجة كميات هائلة من البيانات دون أي عناء، وبسرعة وفعالية. يعدّ التعلّم الألي أحد أهم أبواب الذكاء الصنعي لذلكتضمنت هذه المقالة دراسة لمجموعة من خوارزميّات التعلّم الألي بهدف مقارنة قدرتها على التنبؤ خلال مجموعة من القراءات التحسسيّة تم الحصول عليها عبر نشر مجموعة من حساسات الرطوبة والحرارة الثنائيّة Dht22 (Dual humidity and temperature), في محاولة للاستفادة من تقنيّات التعلّم الألي في تقديم معالجة إضافية تحتاجها البيانات الضخمة (Big Data) والتي تعدّ شبكات الحسّاسات اللاسلكيّة من أبرز مصادرها .

قمنا بإنجاز هذا العمل عبر مجموعة من خوارزميات التعلّم الألي ، وهي خوارزمية العنقدة والتجميع (K-means) وخوارزميّة الجوار الأقرب KNN (K nearest neighbor) ، خوارزميّة الغابة العشوائيّة (Random forest) وخوارزميّة متجهات الدعم Support Vector Machine(SVM)وخوارزميّة الانحدار الخطي LR (Linear Regression)والمقارنة بينها من ناحية الدّقة في التنبؤ ،علما أن التنبؤ كان في حالة التنبؤ بالقيم المستمرة (كما في حالة الـ linear Regression) وفي حالة التصنيف (في الخوارزميّات الأخرى المتبقيّة)

الكلمات المفتاحية: التعلّم الألي ، مجموعة بيانات تحسسيّة , خوارزميّاتK-means ، RF LR, KNN, SVM ,تحليل البيانات التحسسية، Python

*حاصلة على درجة الماجستير في هندسة تكنولوجيا الاتصالات–من قسم هندسة تكنولوجيا الاتصالات – في كليّة هندسة تكنولوجيا المعلومات والاتصالات –جامعة طرطوس– سوريا

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (5) العدد(7) 2021

Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (5) No. (7) 2021

Performance comparison of machine learning algorithms to predict sensing data

Reham Rajab Hassan *

(Received 9 / 6/ 2021 . Accepted 18 / 8/ 2021)

□ ABSTRACT □

Artificial intelligence technology has grown significantly in recent years and has become a major tool in all sectors .One of the advantages of artificial intelligence is that it can process huge amounts of data effortlessly, quickly and effectively.Machine learning is one of the most important sections of artificial intelligence. Therefore,this paper presents using of a set of machine learning algorithms in order to comparing their predictability over a range of sensors readings, it was obtained by deploying dual temperature and humidity sensors, In an attempt to take advantage of machine learning techniques to provide additional processing needed by Big Data, of which wireless sensor networks are one of the most prominent sources.

We accomplished this work with a set of machine learning techniques , k-means , k Nearest Neighbor , Random Forest , Support Vector Machine , and Linear Regression algorithm .Comparing them in terms of accuracy in predicting, Note that the prediction was in the case of predicting continuous values (as in the case of linear regression) and in the case of classification(in the rest of algorithms).

Key Words: Machine Learning**,** Sensing Dataset, K-means ,KNN, RF, SVM,LR algorithms ,Sensed data analysis, Python

* Master Degree ,from Communication Technology Engineering Department, Information and communication Technology Engineering , Tartous University, Syria .

1-مقدّمة :

تطوّرت شبكات الحساسات اللاسلكيّة لأجل عمليّات المراقبة وغيرها عبر الأجيال المختلفة، وتضمن نظام المراقبة اللاسلكي المكوّنات الأساسيّة الأتية : وحدات التحسس وشبكات الاتصال اللاسلكية وأنظمة دعم القرار DSS(Decision Support System) . ركّزت الحلول وطرق الانتشار المقترحة في السنوات الأخيرة على مكوّنات محدّدة في شبكات الحسّاسات اللاسلكيّة,لكن تتطلّب تقنيّة الـWSN أدوات و نطاق يتعامل مع عدد واسع من التحدّيات المختلفة [1].

تقوم أجهزة مثل الحساسات بإنتاج كميات ضحمة من البيانات بشكل مستمر يصبح معها التعامل مع البيانات أصعب ، حيث تتسم البيانات المولّدة في الوقت الحالي بأنها بيانات غير مهيكلة(Unstructured Data) في غالبيتها وهذا يستدعي عمليّات تحليل ومعالجة للتعامل مع هذه البيانات

تقوم شبكات الـWSNبتجميع كميات ضخمة من بيانات خلال الزمن الحقيقي وعبر حساسات مختلفة (حرارة ، رطوبة ، تحديد الموقع ...) لذلك أصبحت شبكات WSN تقنية مهمة لدعم جمع البيانات الكبيرة .في البيئات الداخلية (indoor environments) [2]. تعدّ عملية إيجاد وبناء النماذج عبر مجموعات البيانات الكبيرة عمليّة صعبة وهو تماماً ما تقوم به خوارزميّات التعلم الألي والتي تطبّق تقنيّات إحصائيّة على كميات كبيرة جدا من البيانات بهدف إيجاد النموذج الأفضل لحل مشكلة ما .

2- هدف البحث :

تم التوجّه نحو علم البيانات للتعرف على فرص وتحديّات البيانات الكبيرة الأمر الذي سيعيد تشكيل العديد من المجالات مثل إدارة الأعمال وعلم الاجتماع والهندسة وغيرها. ومن ضمن هذه المجالات تكنولوجيا شبكات الحساسات اللاسلكيّة والّتي تعدّمن المصادر الضخمة للبيانات حيث أن كميات كبيرة من البيانات يتم توليدها من قبل عدة حساسات وبشكل أسي واسع النطاق [3].

هدف هذا البحث إلى مقارنةأداء مجموعة من خوارزميّات التعلّم الالي للتنبؤ وذلك من خلال مجموعة بيانات تحسسية تحوي بالمجمل حوالي 17000قراءة درجة حرارة تم جمعها بواسطة شبكة حساسات مكوّنة من 6 حساسات رطوبة وحرارة (Dht 22) وخلال أوقات مختلفة من العام.

3- مواد وطرق البحث

تم في هذا البحث باستخدام حزمة الـ Anacondaتحديدا الـJupyter Notebookوتطبيق مجموعات من خوارزميّات التعلّم الألي (K-means , KNN, RF,SVM, Linear Regression) لمقارنة أدائها في التنبؤ عبر حوالي 1200 قراءة درجة حرارة من حساسات الحرارة والرطوبة الثنائيّة (Dht 22) بعد تقسيم البيانات إلى مجموعة اختبار (test set بنسبة 20%) ومجموعة تدريب (training set بنسبة 80%)وفي حالات أخرى تم اعتماد مجموعة الاختبار بنسبة 30%علما أن التنبؤ نوعين وهما تنبؤ بالقيم المستمرة (كما في حالة الـ linear Regression)وتنبؤ بالقيم المتقطّعة أو ما يُعرف "بالتصنيف" (مع اعتماد ثلاثة أصناف لدرجات الحرارة) .

3-1 علم البيانات (التعلّم الألي):

إن علم البيانات هو عبارة عن مزيج من مهارات المعلوماتية وعلم الرياضيات والجبر والإحصاء إضافة للخبرة الموضوعية على سبيل المثال، خبرة طبية عند تحليل بيانات مريض، ويتضمن هذا العلم فن استعمال النظريات العلمية لاستخراج الأفكار والمعرفة من البيانات

يرتبط علم البيانات مع ثلاث مفاهيم رئيسية علماً أنّه لا يقتصر على أي منها لا بل أنه يتخطاها في كثير من الأحيان وهذه المفاهيم عبارة عن :

- ☒ الذكاء الصناعي (Artificial Intelligence) : وهو تمكين الحاسب من محاكاة الذكاء البشري .
- ☒ التعلم الألي (Machine Learning): عبارة عن تقنيات إحصائية لتحسين الأداء من خلال التجربة .
- ☒ التعلم العميق (Deep Learning): وهو تدريب الألات لنفسها على الأداء مثل التعرف على الصوت والصورة باستخدام الشبكات العصبونية .

تتكوّن دورة علم البيانات من المراحل التالية:

- جمع البيانات (Data Collection)
- تنظيف البيانات وتحضيرها (Data Cleaning and Preparing)
- استكشاف البيانات وعرضها (Data Exploration & Visualization)
- التحليلات التنبؤية (Predictive Analysis)
- بناء منتجات تعتمد على البيانات (Data Product-ionization).

3-2 البيانات الكبيرة والتعلّم الألي (Big Data & Machine Learning)

يعدّ إنترنت الأشياء (IOT) والبيانات الكبيرة (Big Data) من المصطلحات الأكثر انتشارا في السنوات الأخيرة الماضية وتشكل البيانات الضخمة الركن الأساسي في عالم الIOT حيث يرتبط عدد كبير من الألات (machines) مع بعضها البعض. يتم تعريف البيانات الكبيرة بواسطة مصطلح معروف ب (3V)وهي ثلاث خصائص تتميز بها هذه البيانات:

- ➢ الحجم(Volume) ويشير إلى الحجم الهائل للبيانات .
- ➢ والتنوع (Variety) الذي يشير إلى الصيغة المختلفة التي تكون عليها البيانات : مهيكلة (structured) ، شبه مهيكلة (semi-structure) غير مهيكلة (unstructured).
- ➢ السرعة (Velocity) : يشير إلى السرعة في تحليل البيانات المتدفقة .

يتمّ في الوقت الحالي تنفيذ العديد من النُهج عند التعامل مع البيانات الكبيرة وبشكل خاص عند اختيار الميّزات والعنقدة والتصنيف ...إلخ والتي تلعب دور مهم في تحليل البيانات الضخمة عندما يكون من الضروري استرجاع البيانات أو البحث عنها أو تصنيفها باستخدام مجموعات البيانات الضخمة [4].

3-2-1خوارزميّة الـK-means:

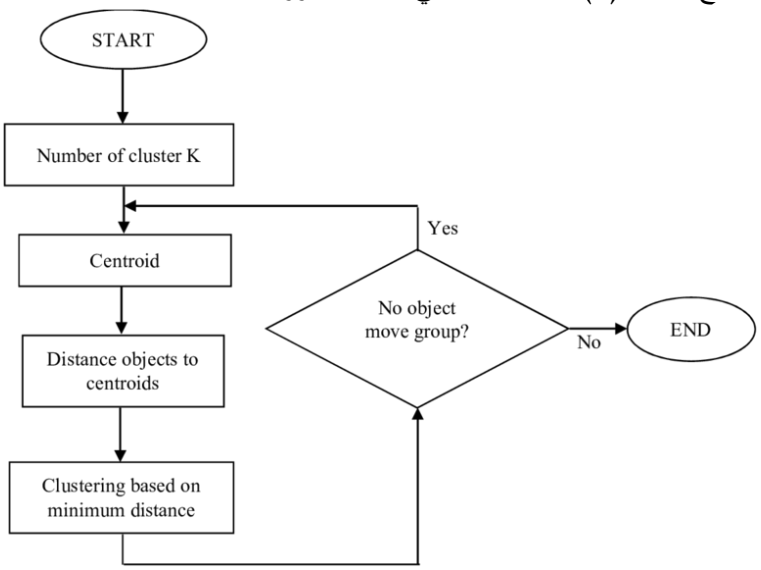
تعدّ خوارزميّات التعلم الألي الخاضعة للإشراف (supervised Machine Learning) والغير خاضعةللإشراف (Unsupervised Machine Learning) من أنواع خوارزميّات التعلّم الألي الرئيسيّة حيث في

النوع الأوّل (الخاضعة للإشراف) يتم استخدام بيانات معنونة (labelled data)أو ما يسمى بيانات تدريب لتدريب الخوارزميّة ويُستخدم هذا النوع مثلا في مشاكل التصنيف (classification) والانحدار. أما في الخورازميّات الغير خاضعة للإشراف تكون مهمةنموذج التعلم الوصول إلى الاستنتاجات من خلال بيانات غير معنونةويُستخدم هذا النوع عادة في تحليل البيانات واستكشافها وفي العنقدة (clustering).

يهدف التنقيب في البيانات إلى استخدام التقنيّات والبرمجيّات من أجل التعرّف على الأنماط والنماذج في مجموعات البيانات المختلفة ،وتعدّ العنقدة (Clustering) واحدة من أهم المواضيع التي تم دراستها ضمن هذا المجال . إن الـk-meansهي خوارزميّة لتجميع الكائنات بناء على خصائص أقسام المتغيّر k وهي من أنواح خوارزميّات التعلّم الألي الغير قابلة للإشراف ، وتهدف هذه الخوارزميّة إلى تقليل التباين الكلي داخل العنقود (intra – cluster variation) أو إلى تقليل دالة الخطأ (squared error function)[5].

تعتمد هذه الخوارزميّة على نقاط المعدّل أو المتوسّط (mean) لمجموعة بيانات و Kهو عدد العناقيد أو المجموعات التي نهدف إلى اكتشافها . وبعد أن يتمّ تشكيل وتوليد المجموعات بواسطة خوارزميّة k-means يمكن بعدئذ تمرير البيانات إليها ليتمّ تشكليها ، ومن ثمّ يمكن إدخال البيانات غير المصنّفة إلى هذه الخوارزميّة ومن ثم تقرّر الخوارزميّة بأي بنية وراثيّة يمكن تجزئة البيانات

تم استخدام خوارزميّة الـk–meansفي هذا المقال لتصنيف القراءات التحسسيّة الواردة من شبكة حساسات ويوضّح الشكل (1) المخطط التدفقي لهذه الخوارزميّة .



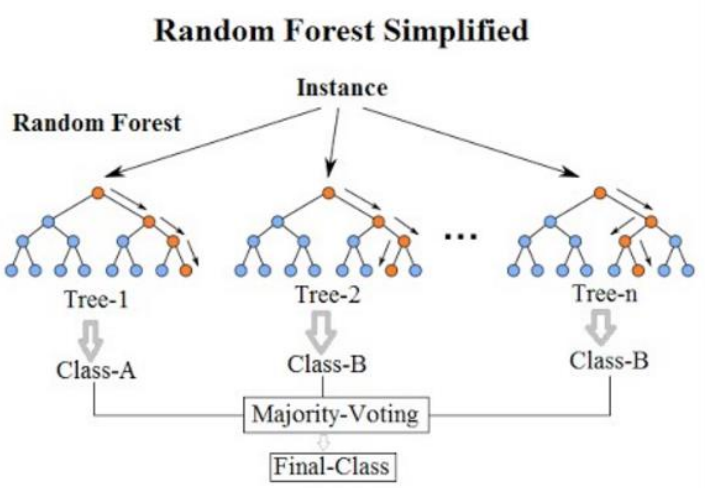
الشكل (1): المخطّط التدفقيّ لـ[5]K–means

3–2–2خوارزميّة الجوار الأقرب (KNN) k–Nearest Neighbor :

هي من خوارزميّات التعلّم الأليالقابلة للإشراف وهي خوارزميّة توصف "بالكسولة " وتتميّز بفعاليتها رغم بساطتها. تقوم هذه الخوارزميّة بحساب المسافة بين نقاط البيانات الجديدة ونقاط مجموعة التدريب بغض النظر عن نوع المسافة الّتي قد تكون أي نوع (إقليديّة ،Manhattan ،...إلخ)ويتم اختيار نقاط البيانات الأقرب حيث k ممكن أت تكون أي عدد صحيح ، وفي النهاية يتمّ تعيين نقاط البيانات إلى الصنف (Class)حيث معظم نقاط البيانات k تنتمي . سنقوم بعرض نتائج تنفيذ هذه الخوارزميّة على مجموعة البيانات التحسسية لدينا باستخدام pythons scikit –learn library[6].

3-2-3خوارزميّة الغابة العشوائيّة الـRandom Forest:

هي عبارة عن خوارزميّة تعلّم ألي تحت الإشراف وتعدّ من أكثر الخوارزميّة المستخدمة نظراً لبساطتها وتنوّعها حيث تستخدم لكلّ من خوارزميّات التصنيف والانحدار . يتم التدريب عادة عبر عمليّة التعبئة (bigging) التي تعتمد علىدمج مجموعة من نماذج التعلّم مع بعضها البعض بهدف تحسين النتيجة الاجمالية ، حيث تنشئ مجموعة الـrandom forestالعديد من أشجار القرار يتم دمجها معا للحصول على تنبؤ أكثر دقّة واستقراراً .وتضيف هذه الخوارزميّة مزيد من العشوائيّة إلى النموذج أثناء بناء الأشجار . وبدلا من البحث عن الميّزات (features) الأكثر أهميّة عند تقسيم العقدة يتم البحث عن الميّزة الأفضل بين مجموعة المزايا العشوائيّة. لذلك في هذه الخوارزميّة يتم أخذ مجموعة فرعيّة فقط من الميّزات في الحسبان من خلال خوارزميّة تقسيم العقدة .ومن المزايا الهامة جدا لهذه الخوارزميّة هي قياس الأهمية النسبيّة لكل ميّزة في عمليّة التنبؤ . يوضّح الشكل (2) بشكل مبسّط هذه الخوارزميّة[7] .



الشكل (2) : خوارزميّة الـRandom Forest

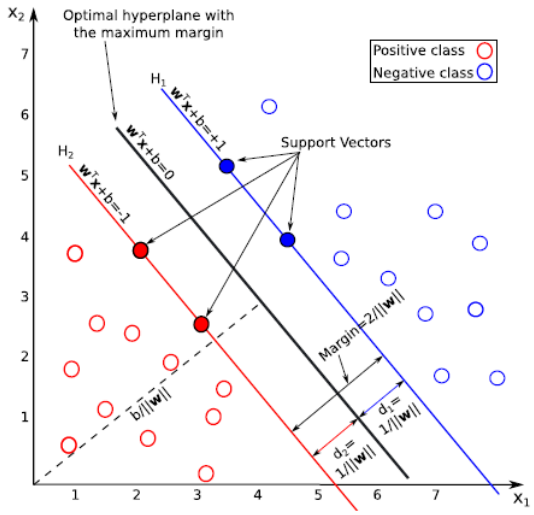
3-2-4خوارزمية متّجهات الدعم Support Vector Machine:

هي خوارزميّة تعلّم ألي قابلة للتعلّم والإشراف تقوم بعمليّة تصنيف البيانات من خلال إيجاد ما يُعرف" بحد القرار الأمثلي" (Optimal Decision Boundary) الذي يملك المسافة الأعظميّة بالنّسبة لأقرب النّقاط لكل الـClasses.

تعمل خوارزميّات الـSVM بشكل جيّد مع البيانات المفصولة بشكل خطيّ ، أمّا في حالة البيانات غير القابلة للفصل ، فيمكن استخدام الـKernel functionsالتي تعمل على تحويل البيانات إلى فضاء متعدّد الأبعاد بهدف إمكانيّة فصل البيانات [8]

ألية عمل الـSVM في حالات الفصل الخطيّة :

نختار ما يسمّىseparate hyper lineللتأكّد من أنّ كل ملاحظة هي على الجانب الصحيح منه ومن ثمّ حساب المسافة العامودية بينه وبين هذه النقاط واختيار المسافة الأقصر لنحصل على الهامش المنشود كما يوضّح الشكل (3) والّذي نسعى أن يكون أعظميّا من أجل مثاليّة التصنيف . يبين الشكل (3) مثال عن استخدام الـSVM في حال كانت البيانات مفصولة بشكل خطي وفق المعادلة $w^t x + b$ حيثd_1, d_2 تمثل المسافة بين المستوى الأول والثاني (متّجهات الدعم) على التوالي والـhyper plane(الذي يمثّل الحد الفاصل المثالي)[8]



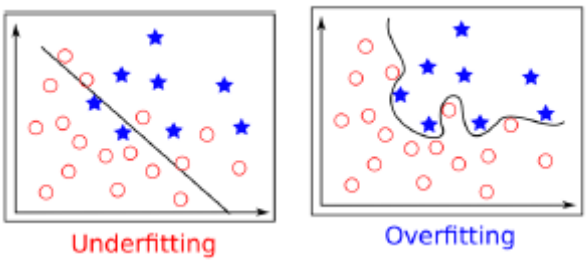
الشكل (3): خوارزميّة متّجهات الدّعم .

<u>أليّة عمل الـSVM في حالات الفصل اللاخطيّة :</u>

تعتمدالالية هنا على بارومترين رئيسين ، تؤثر كيفيّة اختيارهما على أداء النموذج و هما Kernel parameter(الذي سنوضحه بالتفصيل فيما يلي) و Penalty Parameter (C) الذي يعبّر عن عمليّة المقايضة بين تقليل الخطأ وتكبير هامش التصنيف حيث زيادة قيمة C تتناسب عكسا مع عدد متّجهات الدعم (SV) وعرض هامش الـSVMوتشكّل خطر الـOverfitting(والذي يعني مبالغة بالخط الفاصل بين الصنفين) كما يوضّح الشكل (4–a) والعكس صحيح القيم القليلة تسبّب مشكلة الـ underfittig (والتي تعني البساطة في رسم الحد الفاصل) كما يوضّح الشكل (4–b).

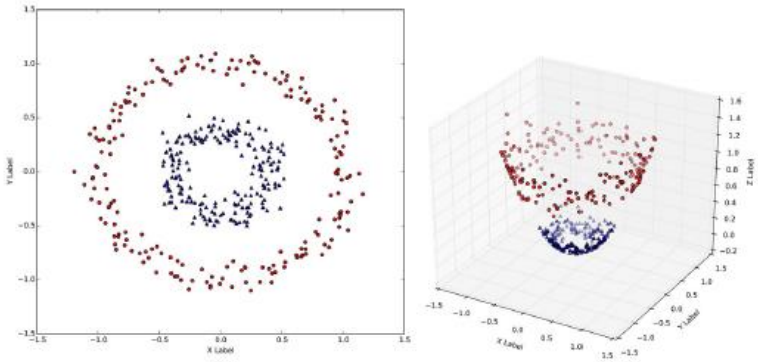
في الواقع لا يمكن أن تكون البيانات مفصولة خطيّا عن بعضها البعض بنسبة 100%[8]وهذا يجعل عمليّة تصنيف فصل وتوزيع البيانات أكثر صعوبة وتعالج الـSVM هذه الحالة بالاعتماد على مفهومين هما :

- ✓ Soft Margin : محاولة إيجاد خط يفصل بين نقاط البيانات لكن مع تحديد النقاط التي حصل خطأ في تصنيفها .
- ✓ Kernel Tricks : وهي الطريقة التي قمنا بتطبيقها على مجموعة البيانات الخاصّة بنا والتي تنطوي على محاولة إيجاد (non linear boundary) عن طريق إجراء بعض التحويلات على الميّزات الموجودة في مجموعة البيانات (Features transformation) للحصول بالنتيجة على ميّزات جديدة تمثّل المفتاح لإيجاد الـnon linear boundary.حيث يستخدم تابع التحويل non –linear الذي ينقل البيانات إلى فضاء ذو مستوى أعلى تسهل معه عمليّة التصنيف كما يبيّن الشكل (5)[8]



(b) (a)

الشكل (4): الـOverfitting& underfittig



الشكل (5): non-linear function

تم فيما يلي توضيح أكثر أنواع النوى غير الخطيّة حيث x هو عينات التدريب و$K(x_i, x_j)$ هو تابع التخطيط (mapping function) لبيانات الدخل ب n بعد $(x \epsilon R^n)$[8]

- Polynomial kernel :

تمثّل المعادلة (1) الصيغة الرياضيّة لهذا التابع حيث d يعبّر عن درجة التصنيف

$$K(x_i, x_j) = \langle x_i, x_j \rangle^d \qquad (1)$$

- Radial Basis Function (RBF) :

يعتمد هذا التابع على نسبة المسافة بين العيّنات σويُعرف أيضا بالGaussian Kernelويتم التعبير عنه وفق المعادلة التالية (2):

$$K(x_i, x_j) = \exp(-\left|\left|x_i - x_j\right|\right|^2 / 2\sigma^2 \qquad (2)$$

حيث كلما زادت σ قلّت المسافة بين العيّنات وقل عدد أشعة الدعم وهو ما يعرف بحالة الـ Under fittingوعلى العكس كلما قلّت σ زاد عدد متّجهات الدعم وهو ما يقابل الـ Overfitting.

- Sigmoid Kernel:

يأتي هذا التابع من حقل الشبكات العصبونيّة وهو تابع ثنائي يُستخدم كتابع تفعيل في الذكاء الصنعي .

ينطوي كل من هذه التوابع على طريقة لمعالجة / تحويل الميّزات لأجل إنتاج مزايا جديدة تسهّل عمليّة تصنيف البيانات. ويتم استخدام هذه البارامترات كصناديق سوداء دون معرفة التفاصيل الخاصّة بها[8]

3-2-5 الانحدار الخطّي Linear Regression :

يعتبر التعرف الخطي من تقنيّات تحليل السلاسل الزمنيّة (Time Series Analysis) وهو عبارة عن نموذج تعلم ألي أساسي يهدف إلى الحصول على مقاربة بين متغيرين، أما الخسارة في هذه الحالة فهي الفرق بين القراءات الحقيقة والقراءات المتوقعة والتي نحاول تقليلها قدر الإمكان للحصول على أفضل مقاربة لتمثيل نموذج التعلم الالي الذي نبنيه. لذلك فإن الهدف هو إيجاد قيمة الوزن (w)،الذي يقلل الفارق بين القيم المتوقعة وتلك الحقيقة ونحصل على الوزن الذي يحقق المقاربة المناسبة من خلال عمليّة التجريب [9].

فإذا كان y متغير غير مستقل و xمتغير مستقل عندئذ يوفّر نموذج التعرف الخطي توقعا ل(y) انطلاقا من (x) وفق الصيغة الموضحة في المعادلة (3) [9].حيث يُستخدم للتنبؤ بقيمة المتغيّر الهدف بناء على متغيرات ذات قوّة تنبؤيّة.

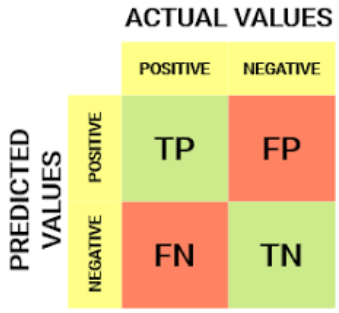
$$Y=\alpha + \beta x + \varepsilon \qquad (3)$$

3-3 تقييم أداء الخوارزميّات :

اعتمدنا الدّقة (accuracy) كمعيار لتقييم أداء الخوارزميّات السابقة واستخلصناها منمصفوفة الارتباك (Confusion Matrix)وهي عبارة عن جدول يسمح بتصوّر أداء خوارزميّات التصنيف غالبايوضّح الشكل (6) شكل مصفوفة الخطأ في حالة تصنيف ثنائي مثلا (إيجابي أو سلبي 0) حيث P هي عدد الحالات الإيجابيّة في البيانات و N هي عدد الحالات السلبيّة في المجموعة و TP(True Positive) هي عدد الحالات الإيجابيّة المتوقّعة بشكل صحيح و TN (True Negative) هي عدد الحالات السلبيّة المتوقّعة بشكل صحيح من قبل الخوارزميّة و FN (False Negative) هي عدد الحالات السلبيّة التي فشلت الخوارزميّة في تصنيفها بشكل صحيح وهكذا وانطلاقا من مكوّنات هذا الجدول يتم حساب عدّة مؤشرات على الأداء ,نذكر من ضمنها الدّقة (accuracy) ودقّة الاختبار (f-score) والحساسيّة (recall) ومدى دقّة تصنيف الحالات الإيجابيّة(Precision) والتي تُحسب في هكذا حالة كما هو موضّح في مجموعة المعادلات (4)[10]. يوجد نوعان لمصفوفة الخطأ وهما المصفوفة الخاصّة بصنفين (2 class confusion matrix) ومصفوفة الأصناف المتعدّدة (multiclassconfusion matrix).

$$accuracy = \frac{TP+TN}{P+N} \quad Precision = \frac{TP}{TP+FP} \quad recall = \frac{TP}{P}$$

$$f-score = \frac{tp}{tp + \frac{1}{2(FP+FN)}} \qquad (4)$$



الشكل (6) : مصفوفة الخطأ (confusion matrix)

وفي حالات أخرى مثل التنبؤ بالقيم المستمرّة كما في حالة الانحدار الخطيّ (Linear Regression) استخدمنا ثلاث طرق شائعة وهي معدّل الخطأ المطلق الـ Mean Absolute Errorومتوسّط مربّع الخطأ Mean Square Error(MSE)والجذر التربيعي لمتوسط مربع الخطأ Root Mean Square Error (RMSE)و توضّح مجموعة المعادلات التالية (5) طريقة حسابها حيث Nعدد العينات الكلي و القيم الحقيقيّة (y_i) والقيم المتوقّعة (y_j) .

$$MAE = \frac{1}{N}\sum|y_i - y_j| \quad (5) MSE = \frac{1}{N}\sum(y_i - y_j)^2 \quad RMSE = \sqrt{MSE = \frac{1}{N}\sum(y_i - y_j)^2}$$

4-3 تحليل البيانات باستخدام لغة البرمجة (بايثون) :

على الرغم من ظهور العديد من التجهيزات والأنظمة التي تتميّز بمميّزات عالية [11]إضافة إلى وجود تطبيقات تأخذ بالحسبان تحليل البيانات الكبيرةإلّا أن حجم البيانات المولّدة قد يتجاوز قدرتها على المعالجة خاصّة في الأنظمة والتطبيقات التي تعمل على منع حدوث الكوارث والتي يجب أن تُنفَّذ وفق تسلسل زمني دقيق .

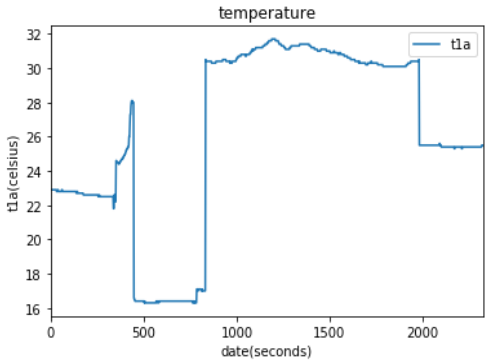
تتطلّب خوارزميّات معالجة وهيكلة البيانات نقل تعليمات مفصّلة إلى الألة وهو ما توفّره لغة عالية المستوى مثل البايثون وهي لغة كائنيّة التوجه OOP (Object Oriented Language) تحتوي على عدد كبير من المكتبات الجاهزة التي تمكّن من إنجاز العمل المطلوب بعدد قليل من الأسطر البرمجيّة[6] .

4- النتائج والمناقشة :

عرضنا ضمن الجدول التالي (1) جزء من مجموعة البيانات التحسسيّة ،حيث تم بداية استكشاف وتصوّر البيانات كما هو موضّح في الشكل (7) . وذلك بعد القيام بتحضير البيانات (data Preprocessing)فيما يتعلّق بالتعامل مع السجلات الفارغة والقيم الشاذة إن وجدت لتكون البيانات جاهزة لإدخالها إلى خوارزميّات التعلّم الآلي .

الجدول (1) : قراءات الحسّاسات

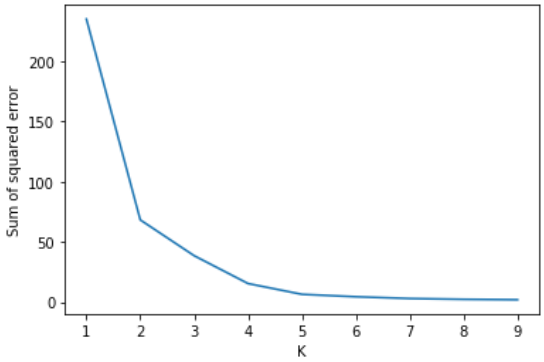
16.5	16.4	16.2	22.3	22.1	22.9	20	19.9	19.7
16.5	16.4	16.2	22.3	22.1	22.9	20.1	19.9	19.8
16.5	16.4	16.2	22.3	22.1	22.9	20.1	19.9	19.8
16.5	16.4	16.1	22.2	22.1	22.8	20	19.9	19.8
16.5	16.4	16.1	22.2	22.1	22.8	20	19.9	19.8
16.5	16.4	16.1	22.2	22.1	22.8	20	19.9	19.8
16.5	16.4	16.1	22.2	22.1	22.8	20	19.9	19.8
16.5	16.4	16.1	22.2	22.1	22.8	20	19.9	19.8
16.5	16.4	16.1	22.2	22	22.7	20	19.9	19.8
16.5	16.4	16.1	22.1	22	22.7	20	19.9	19.8
16.5	16.4	16.1	22.1	22	22.6	20	19.9	19.8
16.5	16.4	16.2	22.1	22	22.6	20	19.9	19.8
16.5	16.4	16.2	22	22	22.6	20	19.9	19.8
16.5	16.4	16.1	22	21.9	101	20	19.9	19.8
16.5	16.4	16.1	22	21.9	22.5	20	19.9	19.8
16.5	16.4	16.1	22	21.9	22.5	20	19.9	19.8
16.5	16.4	16.1	22	21.9	22.4	20	19.9	19.8
16.5	16.4	16.1	21.9	21.9	22.4	20	19.9	19.8
16.5	16.4	16.1	21.9	21.9	22.4	20	19.9	19.8
16.5	16.4	16.1	21.9	21.8	22.4	20	19.9	19.8
16.5	16.4	16.1	21.9	21.8	22.3	20	19.9	20.1
16.5	16.4	16.1	21.9	21.8	22.3	20	19.9	20.1
16.5	16.4	16.1	21.9	21.8	101	20	19.9	20.5
16.5	16.4	16.1	21.8	21.8	101	20	19.9	20.5



الشكل (7): تصوّر البيانات

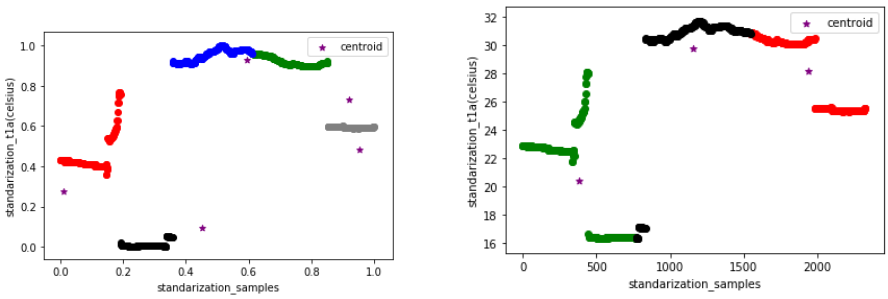
4-1 تطبيق خوارزميةK-means على العينات المجمعة:

تم اختيارقيمة Kعشوائيا بالتجريب وبالمقارنة مع قيمة الجذر التربيعي لمجموع الأخطاء المقابلةلهاSSE(Sum Squared Error) . لاحظناباستخدام مخطط (elbow plot)الموضّح في الشكل (8) أنّه مع زيادة قيمةK تقل قيمة الـSSE .



الشكل (8): elbow plot

تم اختيار k=3كعدد للعناقيد ومن ثم تم تحديد ثلاث نقاط مميزة تمثّل مراكز هذه العناقيد وبعدهاتم بتصنيف القراءات التحسسية في ثلاثة مجموعات تمّ ترميزها ب0 و 1 و2حيث (مرتفعة (0) (بين 26 و 35)– منخفضة (2) (من 9 حتى 17 درجة) – معتدلة (1)(بين 20 و 25) ومن ثم جُمعت البيانات ضمن عناقيد وذلك من خلال حساب المسافة الإقليدية بين كل منها وبين هذه المراكز وهكذا تم تحديد لأي عنقود تنتمي كل نقطة من خلال اختيار المسافة الأقصر كما يوضح الشكل (9) . وعليه تم تحديث المواقع البدائيّة للـ means وتحريكها للحصور على أفضل تجزئة ممكنة،وتم اختيار قيمة جديدة لـK كما يوضح الشكل (10) . وذلك بعد توحيد الميّزات(features) وهي في حالتنا هذه الزمن ودرجات الحرارة عبر ما يُعرف بعمليّة الـfeature scaling



الشكل (9) :k=3 الشكل (10) : k=4

4-2 تطبيق خوارزميةKNN على العينات المجمعة:

تبيّن بعد تطبيق خوارزميّة الKNN في تصنيف درجات الحرارة (وباختيار k=3 بدايةً)حيث كانت درجات الحرارة هي الميّزة المعتمدة ,أعطت الخوارزميّة دقة بنسبة 96% عند تطبيقها على مجموعة الاختبار المكوّنة من 79سجل وبزيادة k لتصبح 4 تعطي دقة حوالي 100%على نفس مجموعة الاختبار كما يوضّح الجدول (2) وتبيّنه مصفوفة الخطأ أيضا في الشكل (11) وهذا يشكّل أداء ممتاز بالنسبة لمجموعة البيانات المستخدمة

الجدول (2) :التقييم(KNN Evaluation)

K=4	K=3	KNN
99.9%	96%	accuracy
0.03	0.03	MAE

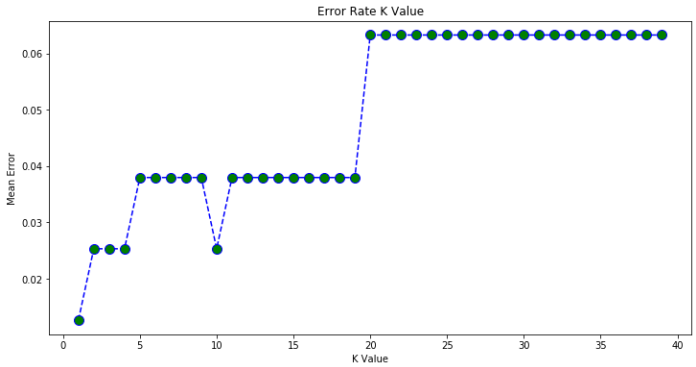
```
[0  0  76]]
[0  0  2 ]
[[0  0  1 ]
precision    recall  f1-score   support

76        0.98      1.00      0.96       0
2         0.00      0.00      0.00       1
1         0.00      0.00      0.00       2

accuracy                           0.96        79
macro avg       0.32      0.33      0.33        79
weighted avg       0.93      0.96      0.94        79
```

الشكل (11) : مصفوفة الخطأ

ضمن إطار تقييم أداء الخوارزميّة لمعرفة قيمة k المناسبة للاستخدام تمحساب متوسط الخطأ المقابل لكلّ قيمة k، حيث لاحظنا أن قيم k الفعالة كانت بين (5 و 20). كما يوضّح الشكل (12)



الشكل (12) : تحديد القيمة الفعّالة لـK

3-4 تطبيق خوارزميةRandom Forest على العينات المجمعة:

على الرغم من وجود خمس ميّزات مأخوذة بالحسبان (Year-month-sample-average-t1a) كما يوضّح الجدول (3-a) لكن الخط الرئيسي لعمليّة التنبؤ في حالتنا كان ميّزة معدّل القيم السابقة average ofhistorical data).بالتجريب والمقارنة مع الدّقة ومعدّل الخطأ المطلق ، حصلنا على تنبؤ من الـRF بدقّة عالية ضمن مجموعة البيانات المستخدمة لديناباستخدام ألف شجرة اتّخاذ قررا (Decision Estimator) حيثلاحظنا من النتائج أن خوارزميّة الـRF قدّمت دقة تقريبا 100% وبخطأ مطلق (MAE) صغير جدا كما يوضّح الجدول(3-b).

الجدول (3-a): الصفوف الأولى في مجموعة البيانات

	year	month	sample	average	t1a
0	2020	spring	1	21	22.9
1	2020	spring	3	21	22.9
2	2020	spring	5	21	22.9
3	2020	spring	7	21	22.9
4	2020	spring	9	21	22.9

الجدول (3-b): التقييم (RF Evaluation)

test set 30%	test set 20%	RF
99.9%	99.79%	accuracy
0.04	0.05	MAE

4-4 تطبيق خوارزميةSVM على العينات المجمعة:

تم استخدام لغة البايثون عبر المكتبة Sklearn المتضمّنة أدوات خاصة (SVM library)، والتي تحوي بدورها على خوارزميّات SVM مختلفة ، ومع الأخذ بالحسبان أنّنا نقوم بعمليّة تصنيف ، قمنا باستدعاء الـSVC (Support Vector Classifier) الذي يأخذ الـKernel. Type كمتغيّر وتم عرض في الجدول التالي دقّة خوارزميّة الـSVM في تصنيف مجموعة البيانات التحسسيّة مع تغيير الـ kernel. Type إلى (poly(d=8و rbf و sigmoid على التوالي وعبر طريقتين لتقسيم الـ dataset إلى بيانات اختبار وتدريب

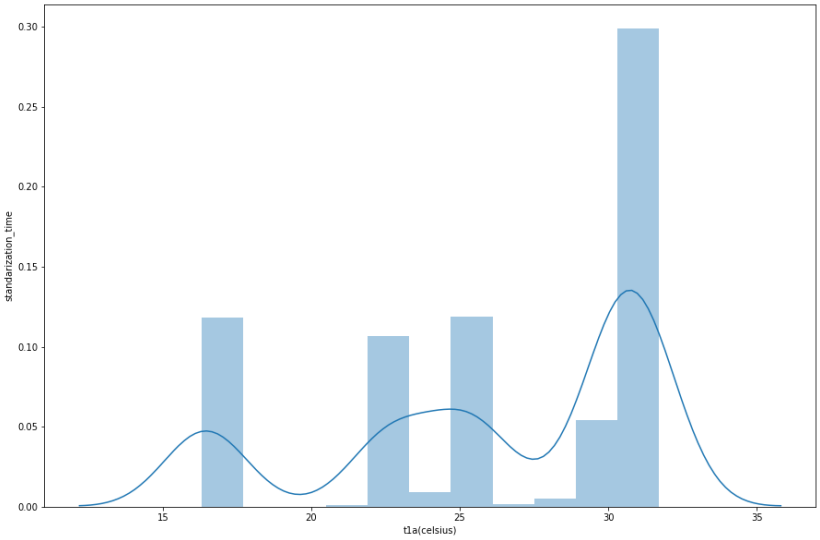
الجدول (4) :التقييم (SVM Evaluation)

sigmoid	rbf	poly	Kernel / SVM
49%	73%	16%	Accuracy (20% test set)
50%	68%	17%	Accuracy (30% test set)

نلاحظ من النتائج كما يبيّن الجدول (4) أعلاه أن التابع الغوصي (RBF) قدّم أفضل دقة بين الـ kernels بينما انخفضت الدقة إلى حوالي 50% مع الـsigmoid وهذا بسبب أنّ هذا التابع يفيد أكثر في حال كان التصنيف ثنائي فقط (صنفين فقط) بينما الـPolynomial Kernel قّم اسوء دقة عبر مجموعة البيانات المستخدمة لدينا

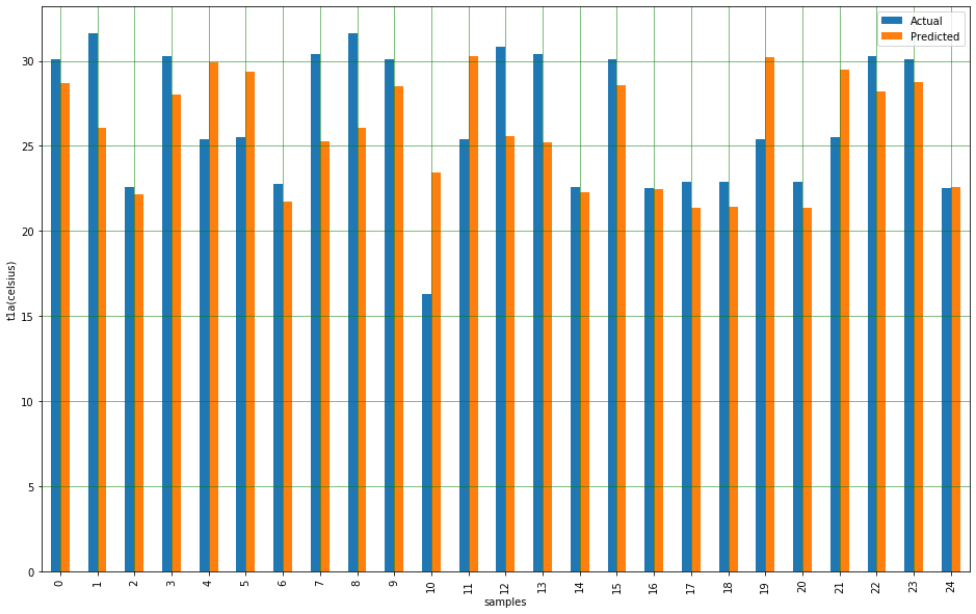
4-5 تطبيق خوارزمية الانحدار الخطّيLinear Regression:على العينات المجمعة:

تم اعتماد خوارزميّة الانحدار الخطي (Linear Regression Algorithm) مع الأخذ بالحسبان أنّ الزمن هو المتغير المستقل ودرجة الحرارة هي المتغير غير المستقلّ وتم تدريب الخوارزميّة بداية بتقسيم البيانات إلى 80% كبيانات تدريب و 20% بيانات اختبار .



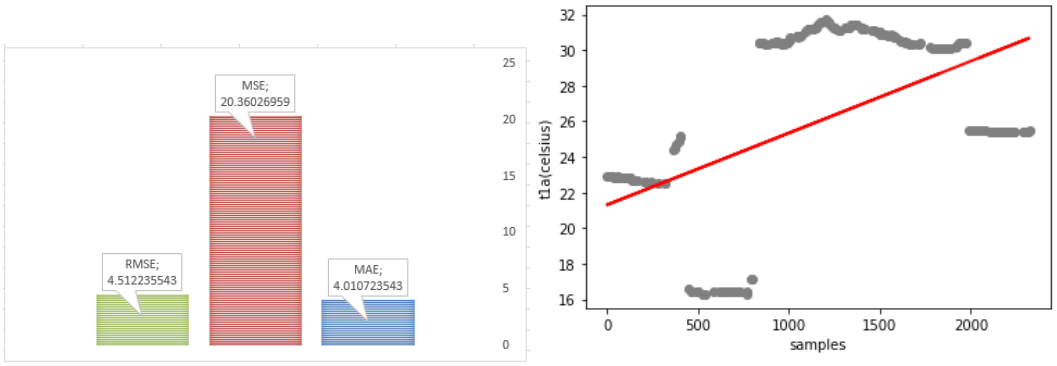
الشكل (13): تصوّر القراءات كمعدّل.

يبيّن الشكل (13) أعلاه تصوّر المعدّل لهذه القراءات وبعد استخدام الانحدار الخطي للتنبؤ عرضنا فيما يلي المقارنة بين القيم الحقيقيّة والقيم المتوقّعة والذي يمكننا أن نستخلص منه أنه هنالك تقارب إلى حد ما بين القيم الحقيقيّة والمتوقّعة كما يبيّن الشكل (14)



الشكل (14): القيم المتوقّعة والقيم الفعليّة (actual Vs. Predicted)

يبيّن الشكل (15) خط الاتجاه مقارنة مع بيانات الاختبار وهو ما يُعرف بعمليّة الـfitting

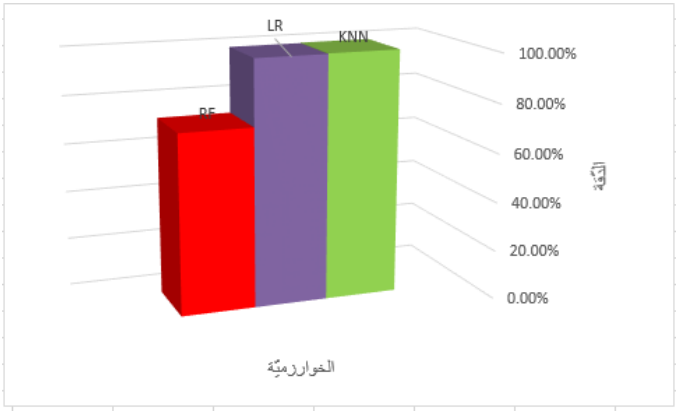


الشكل (15): خط الاتجاه مع بيانات الاختبار الشكل (16) : تقييم الخوارزميّة (LR Evaluation)

أما الخطوة الأخيرة والمهمّة فهي تقييم النموذج حيث نلاحظ في الشكل (16) أن قيمة الـMSE (20.3)أكبر بحوالي خمسة أضعاف من قيمة الـMAE وهذا بالنتيجة يعني أنّ الخوارزميّة لم تعط كثير من الدقّة ولكن لازال بإمكانها أن تقوم بتنبؤات جيدة إلى حد ما.

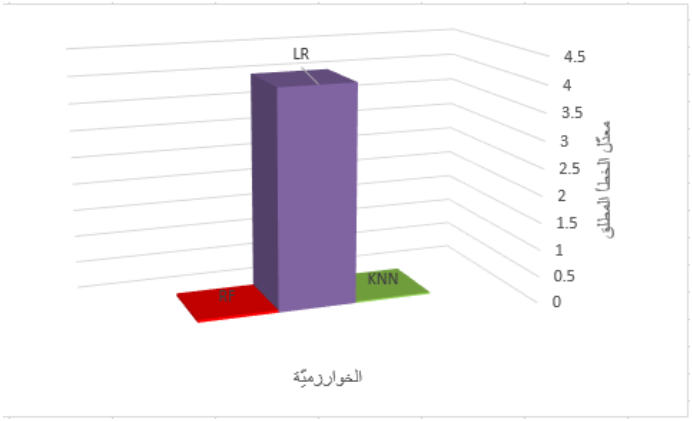
5- الاستنتاجات والتوصيات:

يوضّح الشكل التالي (17) مقارنة من حيثالدقّة بين كل من الـRF و SVM(RBF) و KNN



الشكل (17) : مقارنة من حيث الدّقة

يبيّن الشكل (18) مقارنة من حيث معدّل الخطأ المطلق بين كل من الـKNN و LR و RF :



الشكل (18) : مقارنة من حيث الخطأ المطلق (MAE)

بناء على ما سبق يمكن أن نستنتج أنّه ضمن مجموعة البيانات المستخدمة :

- ✓ بسبب عدم قوّة العلاقة الخطيّة بين درجات الحرارة والزمن بما فيه الكفاية لتحقيق تنبؤ عالي الدّقة ، حقق الانحدار الخطي دقّة منخفضة مقارنة بباقي الخوارزميّات وأعلى خطأ مطلق (حوالي 4).
- ✓ مع عدد كبير نسبيّا من القراءات وفضاء منخفض من الميّزات في مجموعة البيانات قدمت الـKNNو RF أعلى دقّة بالتصنيف.
- ✓ قدّمت الـSVM دقّةجيّدة جدّا فقط عند استخدام (Kernel RBF) أما في حالتي الـ polynomialو sigmoid تراجعت الدّقة ويرجَّح أن يكون سبب هذا التراجع أن التابعين الأخيرين خاصّة الـ sigmoid يعملان بفعاليّة كبرى مع حالات التصنيف الثنائيّة وفي حالتنا كنا نتعامل مع ثلاثة أصناف .

التوصيات :

1. لنحدّد المعيار في اختيار خوارزميّة تعلم ألي للقيام بتنبؤ عبر مجموعة بيانات يجب أن نأخذ بعين الاعتبار عدّة نقاط نذكر منها :
   - عدد العيّنات الموجودة في مجموعة التدريب .
   - عدد الميّزات (features)المُعتمدة .
   - تبيان وجود مشكلتي الـunderfittig و الـOverfitting
2. يمكن نلجأ إلى طرق التعلّم الألي العميق مثل الشبكات العصبونيّة ودراسة أدائها عبر مجموعة البيانات هذه .

6- المراجع :

[1]LIAO, Y.; Mollineaux, M.; Hsu, R.; Bartlett, R.; Singla, A.; Raja, A.,& Rajagopal, R. (2014),*Snow fort: An open source wireless sensor network for data analytics in infrastructure and environmental monitoring*. *IEEE Sensors Journal*,*14*(12),4253-4263

[2] AZIZ, Fayeem; CHALUP, Stephan K.; JUNIPER, James.(2019), Big Data in IoT Systems. *Internet of Things (IoT): Systems and Applications*,The University of Newcastle, Callaghan, NSW 2308, Australia

[3] KIM, B. S; KIM, K. I; Shah, B; Chow, F; & Kim, K. H.(2019), *Wireless sensor networks for big data systems*. *Sensors*. *19*(7), 1565.

[4]DIMEDOVA , Liliya;NIKULCHEV,Evgeny;SOKOLOVA,Yulia(2016),Big Data Classification Using the SVM Classifiers with the Modified Particle Swarm Optimization and the SVM Ensembles,*(IJACSA) International Journal of Advanced Computer Science and Applications*,*Vol. 7, No. 5,* Russia .

[5] CHEN, Zhang; XIA, Shixiong(2009). K-means clustering algorithm with improved initial center. In: *2009 Second International Workshop on Knowledge Discovery and Data Mining*. IEEE, p. 790-792.

[6]PEDREGOSA, Fabian, et al(2011), Scikit-learn: *Machine learning in Python*. *the Journal of machine Learning research*, 12: 2825-2830.

[7]IZQUIERDO-VERDIGUIER, Emma; ZURITA-MILLA, Raúl(2020),*An evaluation of Guided Regularized Random Forest for classification and regression tasks in remote sensing*. *International Journal of Applied Earth Observation and Geoinformation*, 88: 102051.

[8]THARWAT, A. (2019), Parameter investigation of support vector machine classifier with kernel functions. *Knowledge and Information Systems*, *61*(3), 1269-1302.

[9] BERK, Richard A (2008). Statistical learning from a regression perspective. New York: Springer.

[10]RAHMAN, Md Akizur; MUNIYANDI, Ravie Chandren (2020),An enhancement in cancer classification accuracy using a two-step feature selection method based on artificial neural networks with 15 neurons. *Symmetry*, 12.2: 271, Malaysia.

[11]SALEH,Susi.,2017,*Electronic platform Design for control systems and data acquisition using the LPT and the COM ports*. Vol. (1) No. (1),Tartous University Journal for Research and Scientific Studies,Syria,pp.10-11.