Analysis of Clustering and Classification Algorithms to Forecast Electric Power Consumption

Dr. Mariam Saii*
Dr.George Isber **
Dr.Oulfat Jolaha ***
Jina Mhanna****

(Received 27 / 3 / 2022. Accepted 18 / 9 / 2022)

□ ABSTRACT □

Our era is characterized by the wide spread of data of all kinds to the extent that it has become impossible for analysts to extract meaningful information by resorting only to traditional approaches to preliminary data analysis. With the presence of large amounts of stored data, the need has increased to develop tools that are characterized by strength and speed to analyze it, and extract information and knowledge from it. Hence, data mining techniques emerged as techniques aimed at extracting knowledge from huge amounts of data (known in turn as Big Data) [2].

Forecasting electric energy consumption requires knowledge of daily consumption quantities, consumption times and other influencing factors that constitute large amounts of data that can be analyzed using data mining algorithms. The accurate prediction of electrical load is still a challenging task due to many problems such as the non-linear nature of the time series or the seasonal patterns it displays, which are very time consuming and also affect the accuracy of the prediction performance. The process can be improved by using the support beam (SVM) algorithm and other algorithms. Initially, the climatic factors, type, location and period of consumption and a number of other influencing factors were studied in order to improve the performance of the electric power consumption forecasting process.

The following clustering and classification algorithms were also analyzed to predict the electric power consumption, where it was suggested to use the SVM algorithm to solve the electric load prediction problem, which provided a solution to the classification and regression problems and also helped classify the categorical data and gave an independent optimization solution of the model compared to previous studies. A group of classifiers such as random forests, support vector machine, neural networks, and others were also used to reach accurate results that help to take appropriate decisions in the field of electricity consumption prediction.

In the last stage, based on the prediction values resulting from this study, work was done to distribute electrical energy in the most appropriate manner and in line with the importance of higher usage so that we have the ability to operate energy sources at specific times and in appropriate quantities to try to reduce the waste caused by operating unnecessary sources.

Keywords: Random Forest, Support Vector Machine, energy rationalization, consumption prediction and Neural Network

*Professor; Faculty of Electrical & Mechanical Engineering; Tishreen University; Lattakia; Syria.
**Professor; Faculty of Electrical & Mechanical Engineering; Tishreen Universsity; Lattakia; Syria.
***Associate Professor; Faculty of Electrical & Mechanical Engineering; Tishreen University; Lattakia; Syria.
****PHD student; Faculty of Electrical & Mechanical Engineering; Tishreen University; Lattakia; Syria.

تحليل خوارزميات العنقدة والتصنيف للتنبؤ باستهلاك الطاقة الكهربائية

د. مريم ساعي*

د. جورج اسبر **

د. ألفت جولحة ***

جينا مهنا ****

(تاريخ الإيداع 27 / 3 / 2022. قُبِل للنشر في 18/ 9 / 2022)

□ ملخّص □

يتميز عصرنا الراهن بالانتشار الواسع النطاق للبيانات على اختلاف أنواعها حتى أضحى من المستحيل على المحللين استخلاص معلومات ذات معنى باللجوء فقط إلى المداخل التقليدية للتحليل التمهيدي للبيانات.

مع وجود كميات كبيرة من البيانات المخزنة ازدادت الحاجة إلى تطوير أدوات تمتاز بالقوة والسرعة لتحليلها، واستخراج المعلومات والمعارف منها، ومن هنا ظهرت تقنيات التنقيب في البيانات (Data Mining) كتقنيات تهدف إلى استخراج المعرفة من كميات هائلة من البيانات (تعرف بدورها بالبيانات الضخمة Big Data) [1].

يتطلب التنبؤ باستهلاك الطاقة الكهربائية معرفة كميات الاستهلاك اليومية وأوقات الاستهلاك وغيرها من العوامل المؤثرة والتي تشكل كميات كبيرة من البيانات يمكن تحليلها باستخدام خوارزميات

تنقيب البيانات. ولا يزال التنبؤ الدقيق بالحمل الكهربائي يمثل مهمة صعبة بسبب العديد من المشاكل مثل الطابع غير الخطي للسلسلة الزمنية أو الأنماط الموسمية التي يعرضها، والتي تستغرق وقتاً كبيراً كما تؤثر على دقة الأداء في التنبؤ. يمكن تحسين العملية باستخدام خوارزمية شعاع الدعم (SVM) [2].

بدايةً، تم دراسة العوامل المناخية ونوع وموقع وفترة الاستهلاك ومجموعة أخرى من العوامل المؤثرة وذلك لتحسين أداء عملية التنبؤ باستهلاك الطاقة الكهربائية.

كما تم تحليل خوارزميات العنقدة والتصنيف التالية للتنبؤ باستهلاك الطاقة الكهربائية، حيث تم اقتراح استخدام خوارزمية SVM لحل مشكلة التنبؤ بالحمل الكهربائي، والتي قدمت حلاً لمشاكل التصنيف والانحدار كما أنها ساعدت في تصنيف البيانات الفئوية وأعطت حلاً أمثلياً مستقل من النموذج بالمقارنة مع الدراسات السابقة.

تم أيضاً استخدام مجموعة من المصنفات مثل الغابات العشوائية، آلة شعاع الدعم والشبكات العصبية وغيرها للوصول إلى نتائج دقيقة تساعد على اتخاذ القرارات المناسبة في مجال التنبؤ باستهلاك الكهرباء.

وفي المرحلة الأخيرة، بناءً على قيم التنبؤ الناتجة عن هذه الدراسة تم العمل على توزيع الطاقة الكهربائية بالشكل الأنسب وبما يتوافق مع أهمية الاستخدام الأعلى بحيث يصبح لدينا القدرة على تشغيل مصادر الطاقة في أوقات محددة وبكميات مناسبة لمحاولة تقليل الهدر الناجم عن تشغيل المصادر الغير ضرورية.

الكلمات المفتاحية: الغابات العشوائية، آلة شعاع الدعم، ترشيد استهلاك الطاقة، التنبؤ بالاستهلاك والشبكات العصبية.

* **أستاذ– كلية الهندسة الكهربائية والميكانيكية اختصاص تحليل نظم التعرف باستخدام معالجة الصورة– جامعة تشرين– اللاذقية– سورية.**

** **أستاذ – كلية الهندسة الكهربائية والميكانيكية اختصاص برمجة نظم الطاقة الكهربائية – جامعة تشرين– اللاذقية– سورية.**

*** **أستاذ مساعد – كلية الهندسة الكهربائية والميكانيكية اختصاص التحكم الآلي والتسيير الذاتي– جامعة تشرين– اللاذقية– سورية.**

**** **طالبة دكتوراه – كلية الهندسة الكهربائية والميكانيكية اختصاص تحكم آلي– جامعة تشرين– اللاذقية– سورية.**

مقدمة:

تعتبر الطاقة الكهربائية شريان الحياة في جميع البلدان، حيث تقوم معظم التجهيزات الصناعية والمنزلية على استخدام هذه الطاقة بصورة رئيسية. يبدأ انتاج الطاقة الكهربائية من خلال التوليد في مراكز مختلفة اعتماداً على نماذج مختلفة من مصادر الطاقة مثل طاقة الرياح والطاقة الشمسية والوقود الأحفوري وغيرها من نماذج الطاقة المتوفرة في الطبيعة.
مع ازدياد التوسع العمراني وارتفاع عدد السكان، اتجهت العديد من البلدان لترشيد استهلاك الطاقة الكهربائية وتوفيرها ضمن الحدود المعقولة وحتى في البلدان التي لا زالت تعتمد على الوقود الأحفوري والتي تستخدم الطاقات المتجددة أو الطاقة النووية. تسمح عمليات الترشيد بالحفاظ على استهلاك الطاقة ضمن الحدود الدنيا الأمر الذي يفيد في مجالين رئيسيين، يتمثل الأول في تخفيض كميات الاستهلاك من مصادر الطاقة المختلفة وبالتالي تخفيض تكاليف انتاج الطاقة الكهربائية في حين يتمثل الثاني في تخفيض درجات التلوث على البيئة الطبيعية.
تعتبر مشكلة تأمين الطاقة وتوفيرها لاحتياجات الإنسان إحدى المشاكل الكبيرة في الوقت الراهن ويعتبر حلها في غاية الأهمية، ومن المعلوم بأن الطاقة الكهربائية لا تخزن إلا بكميات محدودة تمثل بمصادر طبيعية، لذلك فإن إنتاج الطاقة الكهربائية بكميات أكبر من الطاقة المطلوبة للاستهلاك، يجعل بعض مولدات الطاقة تعمل دون فائدة بالإضافة إلى الاستهلاك العالي للموارد، وبنفس الطريقة فإن نقص الطاقة المولدة يسبب أضراراً كبيرة للشبكة الكهربائية والمستهلك، تتمثل هذه الأضرار بانخفاض التردد وزيادة في هبوط الجهد مما يؤثر سلباً على نوعية القدرة المقدمة للمستهلك.
عمليات التنقيب بالبيانات تقنية حديثة فرضت نفسها بقوة في عصر المعلوماتية، حيث يوفر استخدامها القدرة على الاستكشاف والتركيز على أهم المعلومات التي تهتم بها القطاعات والمنظمات في جميع المجالات، كما تركز تقنيات التنقيب على بناء التنبؤات المستقبلية واستكشاف السلوك مما يسمح باتخاذ القرارات الصحيحة في الوقت المناسب.
يهدف البحث إلى التنبؤ باستهلاك الطاقة الكهربائية باستخدام خوارزميات العنقدة والتصنيف، كما يهدف إلى توزيع الطاقة الكهربائية بالشكل الأمثل، إن مدينة اللاذقية إحدى المراكز العمرانية الرئيسية في الجمهورية العربية السورية وتتميز بموقعها على البحر الأبيض المتوسط الأمر الذي يجعلها أحد المراكز السياحية الرئيسية مما يتطلب الحاجة إلى استهلاكٍ عالٍ في الأشهر السياحية المتمثلة بالصيف. بالإضافة إلى ذلك، فإن العديد من مناطق اللاذقية تعاني من انخفاض درجات الحرارة شتاءً الأمر الذي يتطلب أيضاً تغذية مقبولة للمستهلكين في فصل الشتاء. لذلك تكمن أهمية البحث في التوزيع الفعال للطاقة الكهربائية بناءً على تحليل بيانات الاستهلاك والتنبؤ به، وتوفير استهلاك الطاقة الكهربائية وعدم هدرها وفقاً لنتائج توزيع الطاقة الكهربائية بالشكل الأمثل.

الدراسات المرجعية:

استخدم الباحثين Martín M- Acera M [4] خوارزمية تستفيد من الطابع المحلي للسلسلة الزمنية، كان الهدف الرئيسي من هذه الدراسة هو التنبؤ بالحد الأقصى للطلب اليومي على الكهرباء باستخدام خوارزمية SVM. عند تطبيق تلك الخوارزمية تم الحصول على نتائج قريبة جداً من القيم الحقيقية مما يدل على فعالية هذه الخوارزمية مع إنشاء الشبكة الذكية.
للتغلب على مشكلة التنبؤ قصير الأمد اقترحت الدراسة المنجزة من قبل Zhang P., Wu X., Wang X., Bi Sh [5]إطار جديد للتنبؤ بالأحمال على المدى القصير يعتمد على تقنيات البيانات الضخمة، حيث تم إجراء تحليل العنقدة لتصنيف أنماط الحمل اليومية لبيانات الاستهلاك الكهربائي، بعد ذلك تم تطبيق خوارزمية شجرة القرار وخوارزمية SVM لوضع قواعد التصنيف. بينت النتائج التي تم الحصول عليها أن خطأ التنبؤ بالحمل الكهربائي باستخدام إطار

عمل المقترح هو 1٪ من حمل النظام الحقيقي، كما أنه قادر على التنبؤ بكمية كبيرة من البيانات في الوقت الحقيقي مما ساهم في تحسين في الدقة أكبر بكثير من النهج التقليدي.

كما استخدمت الدراسة المنجزة من قبل Hambali M., Akinyemi A., Oladunjoye J., Yusuf N. [6] خوارزميات تنقيب البيانات وهي خوارزمية CART وخوارزمية REPTree للتنبؤ بالحمل الكهربائي، حيث أن التنبؤ مهمة معقدة لأن الاستهلاك يتأثر بعدة عوامل مثل نوع اليوم، العوامل الاقتصادية والظروف الجوية. توصلت النتائج إلى أن خوارزميات شجرة القرار تقلّل من تأثير فقدان البيانات وعدم توازنها كما وضحت أن تقنية REPTreeهي مناسبة للتنبؤ بالأحمال الكهربائية عند وجود عدد قليل من العوامل المؤثرة كعامل المناخ، كما تفوقت في الأداء على خوارزميات شجرة القرار الأخرى المستخدمة في هذه الدراسة مع مقاييس خطأ أقل وقدرة تصنيف أعلى.

اعتمد أيضاً الباحثين Arai and Barakbah على طريقة الخوارزميات الفائقة من أجل التوصل إلى أمثلية مراكز العناقيد الأولية في خوارزمية K-means، استخدمت هذه الطريقة نتائج خوارزمية K-means ومن ثم تحويل جميع المراكز الناتجة عن العنقدة عن طريق دمجها مع الخوارزمية الفائقة لتحديد المراكز الأولية للعناقيد، تفوقت هذه الخوارزمية بالأداء من أجل العنقدة المعقدة في حال وجود مجموعة بيانات ضخمة والعديد من الواصفات. كما تميزت بامتلاكها لخصائص السرعة والدقة المتوافرة لدى خوارزمية K-means [8].

كما قدم Huang et al. طريقة automatic variable weighing ضمن خوارزمية K-means والتي يمكنها التنبؤ بشكل تلقائي بأوزان المتحولات. تمكنت متحولات الأوزان التي تم توليدها بهذه الطريقة من التوقع الملحوظ للمتحولات في العناقيد كما سمحت باختيار المتحولات في العديد من تطبيقات التنقيب في البيانات حيث نستخدم بيانات واقعية وضخمة عادة، من خلال إزالة المتحولات غير الهامة، مما أدى إلى رفع دقة عملية العنقدة [7].

طرائق البحث ومواده:

1-حدود البحث:

تم إجراء عمليات العنقدة لبيانات الطاقة الكهربائية في المنطقة الساحلية، ولتحقيق ذلك تم الاعتماد على بيانات الاستهلاك لثلاثة سنوات ولمناطق مختلفة من المحافظة حيث تضمنت الدراسة جميع المناطق ضمن مدينة اللاذقية، يعتمد البحث بصورة رئيسية على مجموعة من السمات الرئيسية التي تتضمن المنطقة والتي تم اعتمادها كمعرف ID يتبع لكل منطقة وهي التاريخ، قيمة الاستهلاك، درجة الحرارة، سرعة الرياح، الفصل، الرطوبة، أيام الدوام الرسمي وأيام العطلات، بالإضافة إلى ذلك فقد تم الأخذ بعين الاعتبار استمرارية توافر الطاقة الكهربائية في مركز الاستهلاك حيث يجب أن تتوافر دون انقطاع على مدار الساعة في المستشفيات وبشكل جزئي في المصانع خلال أوقات العمل (نظم الورديات المعتمدة) بالإضافة إلى الأخذ بعين الاعتبار استهلاك المجمعات السكنية.

تم في البحث اعتماد البيانات المقدمة من الشركة العامة للكهرباء حيث تتضمن هذه البيانات توزيع الطاقة الكهربائية بدءاً من محطات 66 كيلو فولت وحتى مراكز الاستهلاك (منازل، محلات. الخ). حيث قسمت المناطق إلى 17 قطاع مستقل يذكر منها على سبيل المثال (جبلة – القرداحة – الجامعة). ومن ثم تم تجزئة كل قطاع إلى مناطقه الجزئية المختلفة وتم ترميز كل منطقة برمز مستقل. تتضمن هذه البيانات 249869 سجل.

2-أدوات البحث:

تعد خوارزميات وتقنيات التنقيب خوارزميات فعالة تهدف إلى التمييز بين النماذج المختلفة للبيانات وفقاً لسمات محددة، حيث تتعامل بعض التقنيات بصورة مباشرة مع أشعة السمات في حين تحتاج أخرى إلى عمليات معالجة أولية من أجل ضبط البيانات لتناسب دخل المصنف كما تتفاوت هذه التقنيات بطريقة معالجة البيانات المقدمة ودقة الخرج وفقاً لنمط بيانات الدخل.

إن الفارق الرئيسي بين العنقدة والتصنيف يكمن في كون العنقدة عملية تعليم بدون معلم لمجموعة من البيانات لأصناف تعتبر مجهولة حيث يتم تحديد معدل التقارب بين السمات ومن ثم استنتاج الأصناف الخاصة بها، في حين يكون التصنيف لأصناف معروفة وبمعلم.

اعتمدت الدراسة بصورة رئيسية على عمليات العنقدة والتصنيف لتحديد المجالات المختلفة لاستهلاك الطاقة الكهربائية ولفترات مختلفة ضمن العام مع تحديد عمليات الاستهلاك في أيام الدوام الرسمي وأيام العطلات والمناطق التي تقوم بالاستهلاك حسب أهميتها والحرارة والرياح.

تم الاعتماد على أهم تقنيات التصنيف الفعالة وهي خوارزمية آلة شعاع الدعم SVM، الغابات العشوائية Random Forest والشبكات العصبية Neural Networks.

2-1 خوارزمية Kmeans:

تعتبر خوارزمية Kmeans[9] إحدى الخوارزميات الفعالة في مجال العنقدة، حيث تبدأ باختيار مراكز عشوائية للسمات المتاحة لدينا (تاريخ الاستهلاك ومقدار الاستهلاك على سبيل المثال)، من أجل تحديد معدل التشابه بين السمات المختلفة المدروسة تم الاعتماد على حساب المسافة الاقليدية وهو عبارة عن معيار إحصائي يقوم بحساب المسافة بين العينات المختلفة بحيث يتم تجميع العينات ذات المسافة الأقل من قيمة محددة ثابتة فيما بينها بعنقود مستقل يمثل مجموعة متشابهة في الصفات، بعد ذلك يتم تحديد مركز كل عنقود بأخذ المتوسط الحسابي لعيناته وحساب المسافة بين جميع عينات الأصناف (والممثلة بالاستهلاك ضمن فترات زمنية محددة) وهذه المراكز ومن ثم تحديد معدل التقارب، يمكن أخذ هذه المراكز في البداية عشوائياً وحساب متوسط السمات بعد إضافة العينة المدروسة إلى الصنف لإيجاد مركز جديد حيث يستمر التكرار حتى الوصول إلى ثبات في المراكز وعندها يتم الانتهاء من التصنيف.

تعطى علاقة المسافة الإقليدية التي تستخدم في خوارزمية K-means لإيجاد درجة الترابط بين العناصر بالشكل:

$$dist\left(x،y\right) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

حيث x_i و y_i الواصفات الخاصة بالعناصر المدروسة، كما يتم اختيار المراكز بشكل عشوائي باستخدام عنقدة K-means وذلك وفق الخطوات التالية [11]:

1- تحديد قيمة k المطلوبة حيث تمثل k عدد العناقيد المرغوبة.

2- تحديد المراكز الرئيسية. يتم تحديد المركز الأولي بصورة عشوائية من البيانات الموجودة حيث يكون عدد المراكز الأولية مساوٍ لعدد العناقيد.

3- إيجاد النقاط الأقرب إلى كل مركز عنقود من خلال حساب المسافة الاقليدية.

4- ترتيب البيانات من خلال المسافة الدنيا، حيث تصبح البيانات تابعة للعنقود إذا كانت أقرب لمركز العنقود.

5- إيجاد المراكز الجديدة اعتماداً على متوسط البيانات في كل عنقود.

6- اختيار حدوث أي تغييرات، اذا حدث تغيير يعود للخطوة 3 أو التوقف في حال عدم حدوث أي تغييرات في الاسناد إلى العناقيد.

2-2 خوارزمية آلة شعاع الدعم (Support Vector Machine) SVM:

تستخدم آلة شعاع الدعم الموجه [13] من أجل عمليات التصنيف، الانحدار وتمييز الأنماط. الهدف الرئيسي منها هو إيجاد أفضل دالة تصنيف فيما بين السمات المدروسة، كما تهدف إلى التمييز بين أعضاء فئتين من بيانات التدريب. الفكرة من الخوارزمية هي إيجاد مستوى مثالي يفصل بين الفئتين والذي يستخدم لتصنيف وتحديد كل نمط، حيث تتميز هذه الخوارزمية بالدقة العالية في التصنيف وتطبق في مجالات واسعة منها تحديد فئات النص، تصنيف الصورة وفي التطبيقات الطبية.

3-2 خوارزمية الغابات العشوائية Random Forest

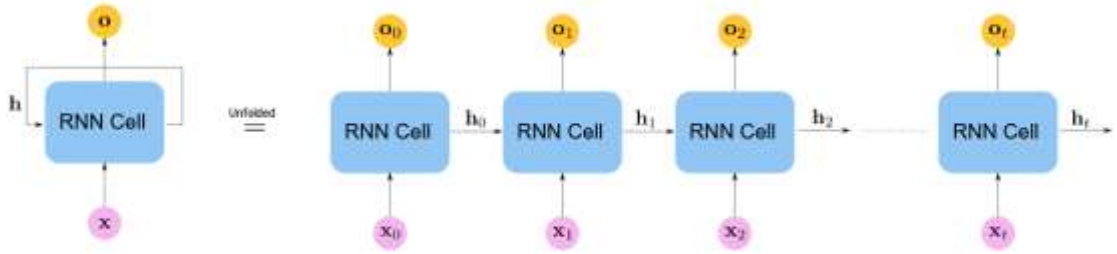
تعتبر خوارزمية الغابات العشوائية [15] من خوارزميات التعلم الآلي تم تطويرها بناءً على مجموعة من أشجار القرار. تستخدم هذه الخوارزمية للتصنيف، الانحدار ومهام أخرى. عادةً ما تتمتع خوارزمية الغابة العشوائية بدقة أفضل مقارنة بشجرة القرار.

4-2 الشبكات العصبونية Neural Networks:

تم تطوير الشبكة العصبية الاصطناعية [16] كتقليد للشبكة العصبية لدى الانسان، حيث تتكون الشبكة العصبية الاصطناعية من خلايا عصبية أو عقد صناعية. تتميز بقدرتها العالية لحل مشاكل وقضايا الذكاء الاصطناعي حيث مازالت هذه الشبكات في مرحلة التطوير المستمر، بما أن البحث يعتمد على مجموعة من السلاسل الزمنية فإن الشبكة الأفضل في هذا المجال هي الشبكات العصبية التكرارية RNN (Recurrent Neural Networks) والتي تعتمد على تسلسل زمن ما للتنبؤ بالقيم المستقبلية.

تعتبر شبكات RNN إحدى أهم شبكات التعلم العميق من أجل معالجة البيانات ذات النمط التتابعي، كما يعتبر هذا النمط من الشبكات الأفضل في معالجة البيانات التتابعية ضمن فترة من الزمن حيث يحتاج النموذج العميق ذو التغذية الأمامية لمجموعة من البارامترات لكل معامل في التتابع.

تقوم هذه الشبكات بتطبيق الوزن نفسه على جميع المعاملات في التتابع كما تقوم بتقليل عدد المعاملات وتسمح للنموذج بالتعميم على كامل طول المتحولات ضمن التتابع. تسمح RNN بتعميم النموذج إلى بنى المعطيات أكثر من البيانات التتابعية كما في البيانات المكانية أو الجغرافية.

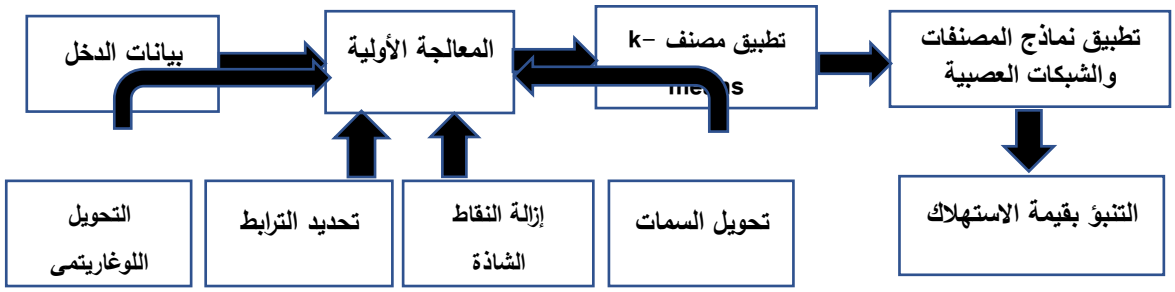


الشكل (1) الشبكات من نمط Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN)

وتعتبر الشبكات من نمط Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) إحدى أقوى أدوات التصنيف الديناميكية المعتمدة، يرتكز نموذج تعلم الآلة على تطوير الخوارزميات التي تسمح بتطوير

الأداء بشكل ذاتي ديناميكي اعتماداً على التدريب. أي كلما زادت عمليات التدريب للخوارزمية سيكون أداء هذه الخوارزمية أكثر فعالية، تتم تلك العملية من خلال خلق تابع تصنيف من بيانات التدريب المتاحة. يتم بعد ذلك قياس الأداء للمصنف المصمم عن طريق تطبيق البيانات غير الملاحظة ومطابقة الخرج.

3- نظام التنبؤ باستهلاك الطاقة الكهربائية المقترح:



الشكل (2) المخطط الصندوقي لنظام التنبؤ

لا تتمتع البيانات التي تم الحصول عليها من مصادر مختلفة ببنية تسمح باستخدامها بصورة مباشرة من أجل عمليات العنقدة والتصنيف، مما أدى إلى الحاجة لتعديل هذه البيانات في البداية من خلال إجراء عمليات تحسين وضبط أولية تحقق الشروط التي يتطلبها مصنف K-means. على سبيل المثال، إحدى الشروط الرئيسية ليعمل مصنف K-means بصورة فعالة ألا تتضمن البيانات على قيم صفرية أو فارغة حتى لا ينتج أي أخطاء لاحقاً في عملية التصنيف، كما يجب ألا تتضمن بيانات مكررة منعاً لإعطاء أي وزن إضافي لبيانات محددة دون أخرى.

وكذلك الأمر من أجل عنقدة البيانات المرتبة وفق سماتها تم تحويل البيانات باستخدام إحدى أدوات استخلاص السمات أو باستخدام النماذج البارامترية، على سبيل المثال يمكن استخدام الانحدار الديناميكي dynamic regression أو الشبكات العصبية. أما في حال البيانات المرتبة سطرياً يتم التصنيف بصورة مباشرة عبر أشعة السلاسل الزمنية من دون أي عملية تحويل مسبقة لعملية العنقدة [18] [17].

تتضمن الخوارزمية المقترحة في نظام التنبؤ باستهلاك الطاقة الكهربائية مجموعة من المراحل التي تهدف إلى ضبط البيانات لتطابق نموذج الدخل المطلوب من كل مصنف والتي تعرف بعمليات المعالجة الأولية للبيانات.

3-1 المعالجة الأولية للبيانات:

تتضمن المعالجة الأولية للبيانات إجرائيتين هما إجراء التحويل اللوغاريتمي وإزالة النقاط الشاذة .

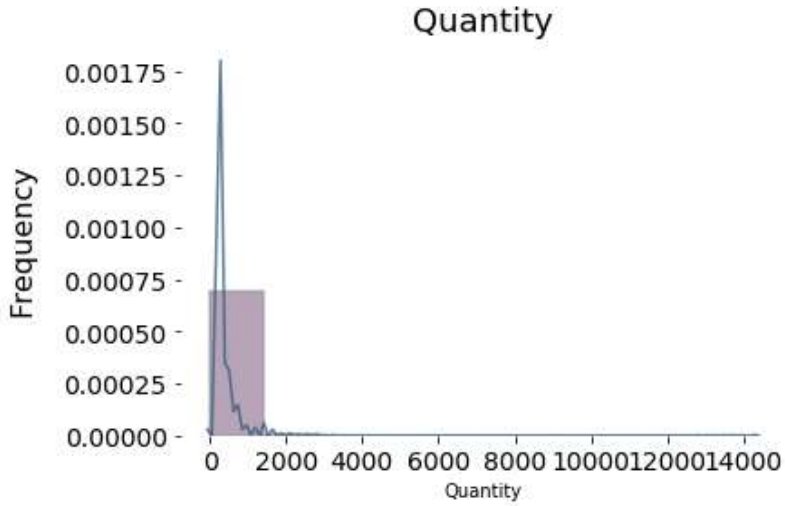
أ- إجراء التحويل اللوغاريتمي:

لملاحظة صفات وخصائص البيانات الموجودة لدينا نحتاج لاستخدام عدد من الواصفات الإحصائية كالمتوسط الحسابي والانحراف المعياري والتباين. عند حساب التباين بالنسبة للمنطقة وكمية الاستهلاك في المناطق المختلفة نتج لدينا المخططات الموضحة في الشكل (3) و(4) التي تبين التباين في استهلاك الطاقة الكهربائية بالكيلو واط ساعي والتباين بالنسبة لتوزيع الاستهلاك حسب المناطق المدروسة، على التوالي.

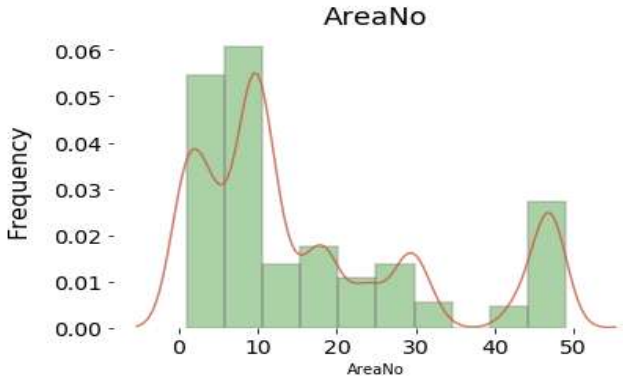
يلاحظ من المخططات السابقة أن توزع البيانات لا يحقق شروط التوزع التي تتطلبها عمليات العنقدة باستخدام K-means والتي تتضمن:

- شكل العنقود: إن تباين التوزيع يجب أن يكون كروياً أي أن البيانات يجب أن تكون موزعة طبيعياً ولها نفس التباين.

- حجم العنقود: يجب أن يكون لكل العناقيد نفس العدد من السمات.
- العلاقات بين المتحولات: يجب أن تكون أقل ما يمكن أو لا يوجد أي ترابط بين المتحولات.

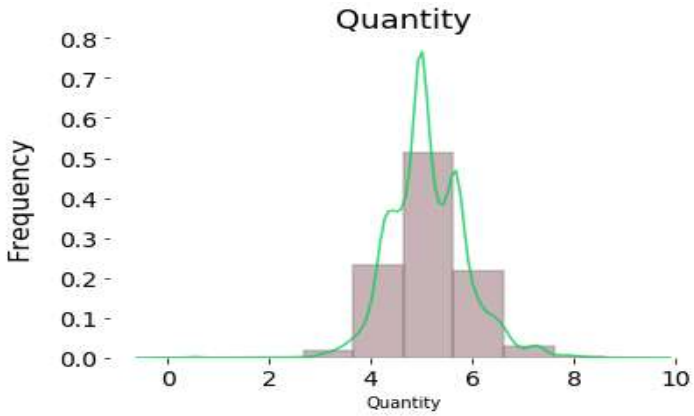


الشكل (3) التباين لاستهلاك الطاقة الكهربائية بالكيلو واط ساعي

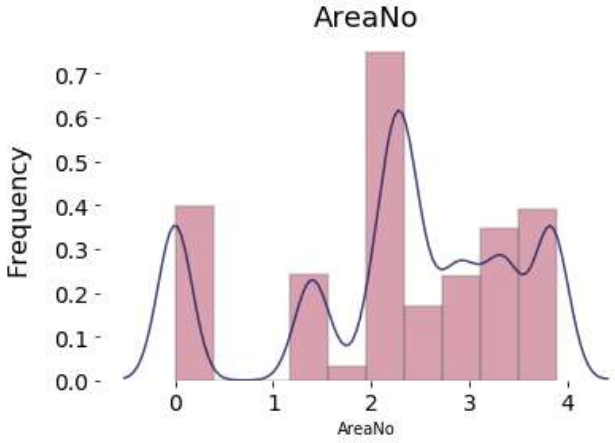


الشكل (4) التباين بالنسبة لتوزع الاستهلاك حسب مناطق الدراسة

كما أن البيانات لدينا ليست موزعة بشكل طبيعي والتباينات ليست قريبة من بعضها البعض لذلك يجب القيام بعمليات التعديل على البيانات. لإجراء تلك العملية تم الاعتماد على التحويل اللوغاريتمي الذي يسمح بعملية ضبط مجال البيانات، حيث يسمح هذا التحويل بمعالجة حالات التوزع غير الطبيعي كما يسمح بمعالجة حالات التباين غير المتقاربة، يوضح الشكلين (5)، (6) نتائج التباين في توزيع البيانات بعد إجراء عملية التعديل باستخدام التحويل اللوغاريتمي.



الشكل (5) التباين في استهلاك الطاقة الكهربائية بالكيلو واط ساعي بعد تعديل البيانات

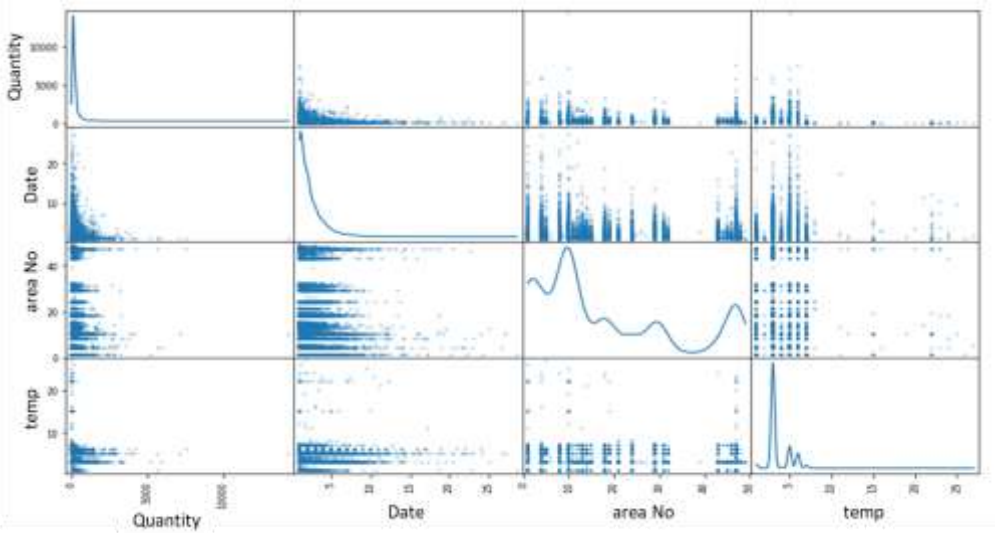


الشكل (6) التباين للمناطق المدروسة بعد تعديل البيانات

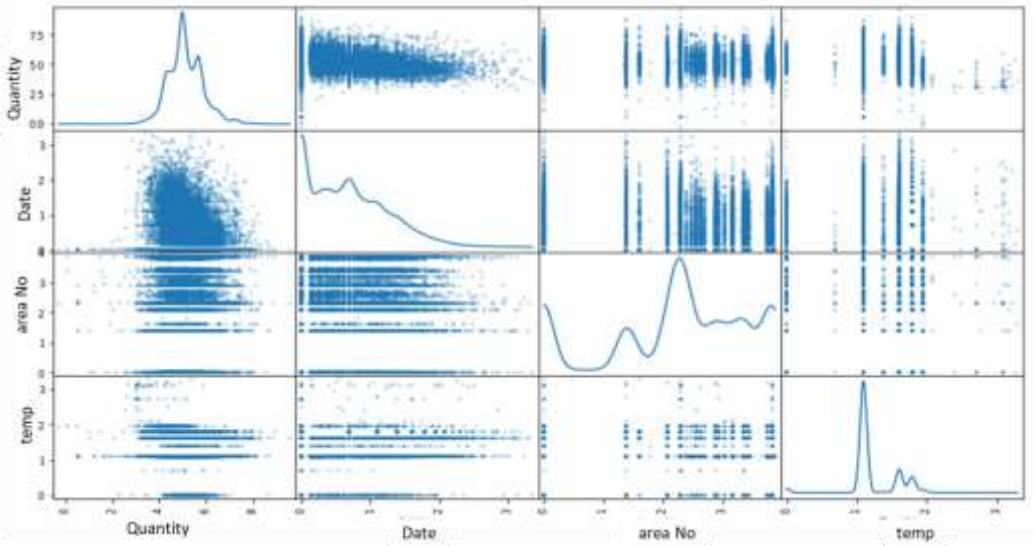
يُلاحظ أن مخططات التباين في استهلاك الطاقة الكهربائية وفي المناطق المدروسة أخذت شكل التوزع الطبيعي المناسب لعمليات التصنيف الأمر الذي يساعد على رفع دقة المصنف.

تم دراسة ترابط السمات في وضعها الطبيعي ونتج الشكل (7) الذي يبين أن الترابط بين السمات لا يتمتع بالتوزع الغاوصي الطبيعي بالإضافة إلى عدم وضوح الترابط بين السمات المختلفة ومن ثم تم تطبيق التحويل اللوغاريتمي على البيانات ودراسة ترابط السمات، حيث يلاحظ من الشكل (8) أن البيانات بعد التحويل اللوغاريتمي قد أصبحت ذات توزع غوصي بالإضافة إلى زيادة درجة الوضوح بين الأزواج الرئيسية مثل المنطقة والكمية والمنطقة والتاريخ وانخفاض الترابط بين الحرارة والكمية ضمن البيانات المتاحة لنا وذلك بسبب تقارب المجال لدرجات الحرارة في المنطقة المدروسة.

كذلك تم دراسة ترابط السمات في وضعها الطبيعي ومن ثم تمت دراستها بعد التحويل اللوغاريتمي حيث بينت النتائج أن هذا التحويل قد سمح بتوضيح الترابط بين السمات بصورة فعالة بالإضافة إلى تحويلها إلى التوزع الطبيعي.



الشكل (7) درجة الترابط بين السمات



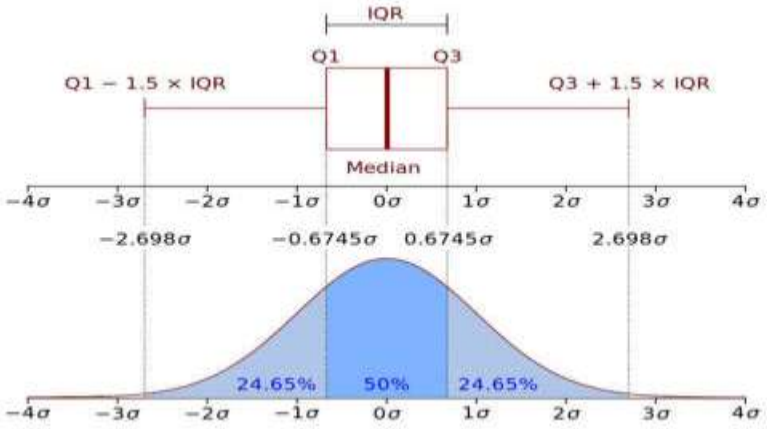
الشكل (8) درجة الترابط حسب مناطق الدراسة

يوضح الشكلين (7) و (8) درجة الترابط بين السمات المختلفة المعتمدة في الدراسة. إن درجة الترابط بين السمات المختلفة مهم لأنه يجب الاعتماد على السمات الأكثر ترابطاً وجمعها ضمن مجال وحيد في حين يجب أخذ السمات ذات الترابط الأقل بصورة مباشرة.

ب- إزالة النقاط الشاذة:

وهي مرحلة بالغة الأهمية في حالات المعالجة الأولية للبيانات إذ يمكن أن يسبب وجود النقاط الشاذة outlier حالات انحراف skew لعدد من النقاط والتي يجب أن نأخذها بعين الاعتبار. تم الاعتماد على تقنية Tukey's Method [21] من أجل كشف هذه الحالات ودرجة استمراريتها بمجال معياري بمقدار 1.5 مرة من interquartile range (IQR) جميع النقاط التي تتوضع خارج هذا المجال تسمى نقاط غير طبيعية (والتي تمثل حالات استهلاك غير طبيعية). يجب الأخذ بعين الاعتبار ألا يتم حذف جميع هذه النقاط تجنباً لخسارة كم من البيانات وإنما فقط حذف النقاط التي تحدث لأكثر من سمة.

الهدف الرئيسي من اعتماد طريقة Tukey's Method حساب مجالات موثوقة تتعلق بدقة البيانات حيث تهتم بدراسة مجال الخطأ بين كل زوج من البيانات. في حين يسمح معيار interquartile range (IQR) بتحقيق التوزع الطبيعي للبيانات ضمن مجالات محددة مختارة للبيانات. يوضح الشكل (9) مبادئ استخدام هذه الطريقة [22].

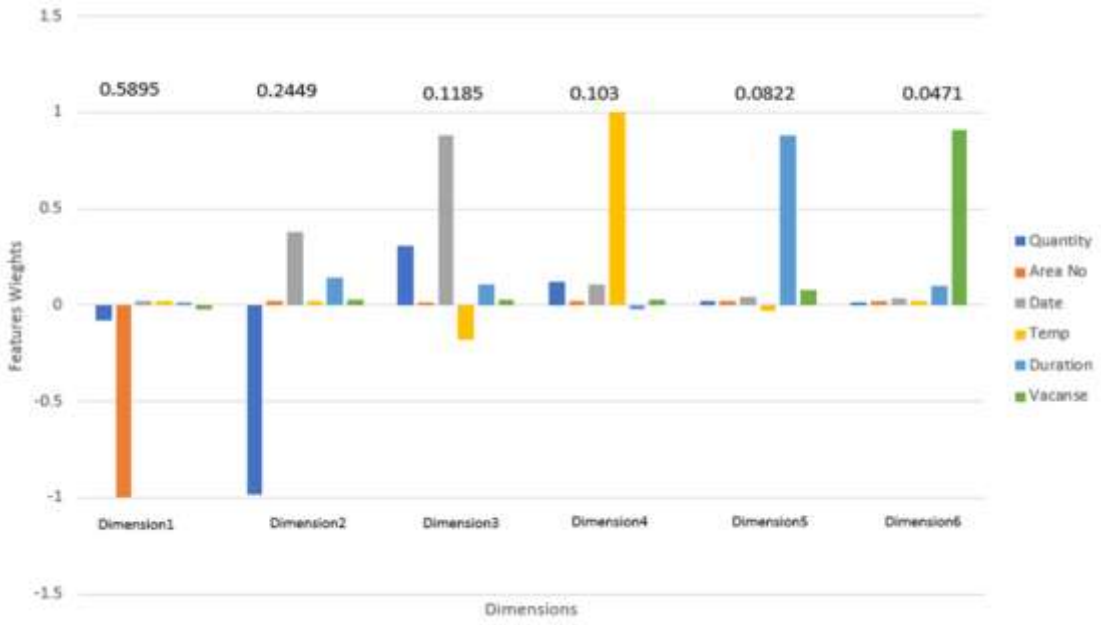


الشكل (9) حساب معامل Tukey's Method باستخدام IRQ

3-2 تحويل السمات Feature Transformation:

بعد أن تم تجهيز البيانات في مرحلة المعالجة الأولية يتم إجراء تحويل السمات باستخدام خوارزمية PCA [23] بتحديد السمات التي يجب زيادة وزنها والسمات التي يجب تقليلها. يعطي هذا التحويل مقدار معدل التباين لكل بعد من الأبعاد. بالإضافة إلى ذلك يحدد تحويل السمات عدد التباينات المتاحة ضمن البيانات والذي يمكن توضيحه ضمن بعد وحيد. ويمثل كل عنصر component يتم تحديده من هذه العملية على أنه سمة في فضاء العينات.

يتم استخدام PCA نظراً لكون أحد الأهداف الرئيسية لها هو تقليل عدد الأبعاد وبالتالي تقليل تعقيد المسألة. يأتي هذا التحويل بكلفة استخدام ابعاد أقل والتي تستخدم قيم تباين أقل في المعطيات المدروسة ويسمح التباين التراكمي بتحديد عدد الأبعاد المناسبة للمسألة المدروسة حيث يوضح الشكل (10) نتائج استخدام هذه الخوارزمية الذي يبين أن سمتي كمية الاستهلاك ورقم المنطقة هما السمتان الأكثر تأثيراً في نتائج البيانات التي سيتم تمريرها إلى مصنف K-means في المرحلة التالية.

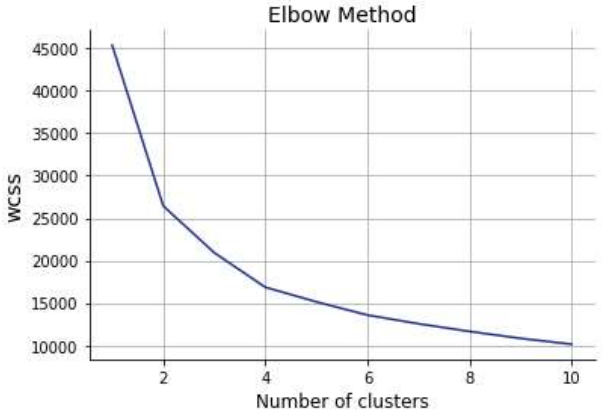


الشكل (10) تحديد درجة ترابط السمات باستخدام PCA

3-3-اختيار عدد العناقيد في مصنف K-means:

عند القيام بعملية العنقدة يجب تحديد عدد العناقيد المناسبة. تم استخدام خوارزمية elbow and silhouette methods [24] بالاعتماد على قيم total intra-cluster variation (total within-cluster sum of square ((WSS) بحيث يكون الهدف الرئيسي هو تقليل قيمة WSS.

تبحث خوارزمية Elbow method عن درجة تغير WSS ضمن عدد من العناقيد وبالتالي سيتم حساب K-means لعدد من القيم الخاصة بـ k ومن ثم رسم مخطط WSS مع عدد العناقيد. بعد رسم مخطط عدد العناقيد مع معامل WCSS تحدد نقطة الانحراف (المعصم) تمثل عدد العناقيد المناسب كما في الشكل (11).



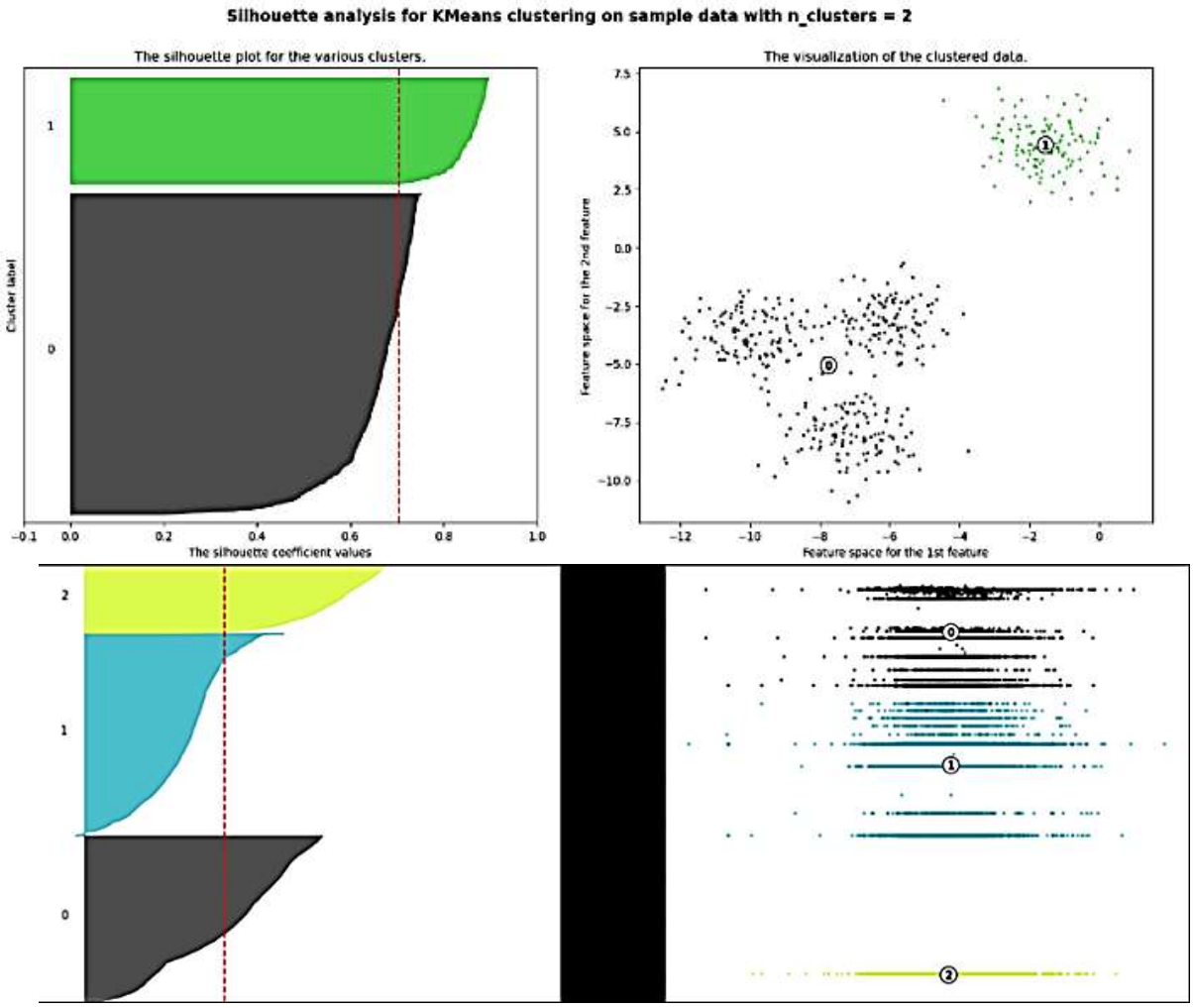
الشكل (11) استخدام خوارزمية Elbow methods

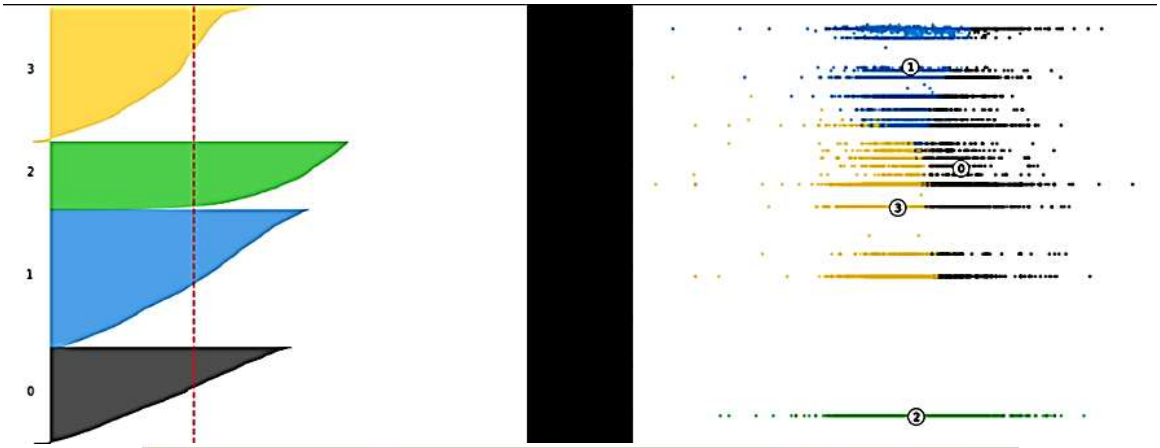
تعمل طريقة silhouette analysis مع خوارزمية المعصم Elbow method لتحديد العدد الصحيح والمناسب للعناقيد حيث تستخدم لدراسة مسافة الفصل بين العناقيد الناتجة. تقوم هذه التقنية بتحديد المسافة بين نقطة ما من عنقود إلى العناقيد المجاورة.

عند إجراء العمليات الحسابية وفق هذه التقنية فإنها تمتلك قياس يتراوح بين 1 و1− حيث تمثل القيم القريبة من 1 أن العينات بعيدة عن العناقيد المجاورة، في حين تمثل القيمة 0 كون العينة قريبة من الحد الفاصل بين العنقودين. أما القيم السالبة تعني أن العينات تم تصنيفها بصورة خاطئة.

تسمح هذه الخوارزمية بإجراء الاختبارات على عدد العناقيد من أجل اختيار عدد العناقيد الأفضل لتطبيق المصنفات بفعالية كي لا يتم اختيار هذه العناقيد بصورة عشوائية وذلك وفقاً للسمات التي تم اختيارها. مع تخفيض عدد العناقيد يتم تجميع العديد من السمات ضمن عنقود واحد وبالتالي وفي حالة اختيار عنقودين نلاحظ في الشكل (12) حيث تتباعد المناطق 1 و2 إلا أن المنطقة الأولى تتضمن العديد من السمات التي قد تتداخل وبالتالي قد تسبب انخفاض دقة المصنف.

بنفس الطريقة وفي حال وجود أربعة عناقيد، نلاحظ تداخلاً بين المناطق ولكن هذا يقابل أخذ تباعد السمات بعين الاعتبار وبالتالي الحصول على مجال أكبر من الفصل.






```
For n_clusters = 2 The average silhouette_score is : 0.7049787496083262
For n_clusters = 3 The average silhouette_score is : 0.5882004012129721
For n_clusters = 4 The average silhouette_score is : 0.6505186632729437
For n_clusters = 5 The average silhouette_score is : 0.56376469026194
For n_clusters = 6 The average silhouette_score is : 0.4504666294372765
```

الشكل (12) استخدام خوارزمية silhouette methods للضبط الدقيق لعدد العناقيد

نلاحظ من تطبيق خوارزمية elbow أن أفضل نتائج للترابط حدثت في حالة عنقودين بقيمة 0.704 وأربعة عناقيد بقيمة 0.65 وبالتالي فإنه يمكننا اختيار أي من هذين المجالين، يسمح اختيار العناقيد الأربعة باختبار تباعد أكبر بين العينات والتفريق فيما بينها بالنسبة للسمات المختلفة.

النتائج والمناقشة:

تمثل عمليات ترشيد الطاقة الكهربائية إحدى أهم العمليات الضرورية في الجمهورية العربية السورية، حيث تؤمن توفير الفائض من الطاقة الكهربائية عن طريق عزل النقاط الشاذة والتي تمثل حالات الاستهلاك غير الشرعي.

يعتمد البحث على إجراء عمليات التصنيف للقطاعات حسب الاستهلاكات ضمنها ومن ثم إجراء عمليات رفع أوزان الاحتياجات للمناطق الأعلى أهمية وتحديد مجال الاستهلاك وطرقة التوزيع للطاقة الكهربائية لهذه المواقع باستخدام مجموعة من الخوارزميات. كما تم تحديد عدد العناقيد المناسبة باستخدام خوارزمية Elbow method حيث تم اختيار أربعة عناقيد بما يتناسب مع الدراسة.

يوضح الشكل (13) مجموع قيم استهلاك الطاقة الكهربائية في المراكز الرئيسية (20 كيلو فولت) في مدينة اللاذقية وفقاً لجميع الأشهر المختلفة ولثلاثة أعوام، تم أخذ مجموع الواردات إلى كل مركز من المراكز الرئيسية الموجودة في محافظة اللاذقية (17 مركز) وبنفس اليوم من العام حيث يوضح المخطط مجموع الواردات إلى جميع المراكز (بالواط) خلال الفترة ما بين 2018/1/1 وحتى 2020/1/1 حيث تم رسمها ببرنامج Excel نظراً لكون البيانات مخزنة على شكل ملف csv، حيث يلاحظ أن الحد الأدنى للاستهلاك يحدث في الفترة ما بين شهري آذار ونيسان حيث يكون الاستهلاك أصغرياً وهو ما يتوافق مع الوضع في سورية نظراً لانتهاء فصل الشتاء وكون هذه الفترة خارج الموسم السياحي، مما يدل على نجاح الخوارزمية المستخدمة في التنبؤ.

الشكل (13) استهلاك الطاقة الكهربائية في مدينة اللاذقية لثلاث سنوات

الاستنتاجات والتوصيات:

عند تحديد التصنيفات للقطاعات حسب الاستهلاك تم الأخذ بعين الاعتبار زيادة وزن وأهمية بعض القطاعات (كالمشافي على سبيل المثال) مقارنةً مع الاستهلاك المنزلي التقليدي وتم أخذه بعين الاعتبار عند إجراء عمليات التصنيف كما تم الأخذ بأهمية مجموعة من الظروف الأخرى الهامة كحالة الطقس والرياح والرطوبة والفترة بالعام (حسب الفصل) حيث يزداد الاستهلاك في فصلي الصيف والشتاء وينخفض خلال الخريف والربيع.

تم الاعتماد على عدد من خوارزميات التصنيف والتنبؤ بدقتها حيث اعتمدنا في دراستنا على مصنفات الغابات العشوائية Random forest وأجهزة أشعة الدعم support vector machine بالإضافة إلى الشبكات العصبية neural networks. تم أيضاً تحديد عدد العناقيد المناسبة باستخدام خوارزمية elbow and silhouette methods حيث تم اختيار أربعة عناقيد بما يتناسب مع الدراسة. بعد الانتهاء من العنقدة تم إضافة العينات إلى عناقيدها وبعد ذلك تطبيق أنواع المصنفات الثلاثة للتأكد من انتماء العينات بصورة صحيحة وتقريبية للصنف الذي تتبع له، عند إجراء المطابقة بين القيم الفعلية والقيم المتنبأ بها لكل طريقة من الطرق تم حساب نسبة الخطأ في أداء كل مصنف من المصنفات وبالتالي الدقة الناتجة اعتماداً على برمجية python ومجموعة التوابع المضمنة في مكتبة sklearn حيث نتجت لدينا قيم الدقة كما هو مبين في الجدول (1) الآتي:

الجدول (1) نتائج معدل دقة المصنفات المستخدمة

الطريقة	الدقة
Random Forest Classifier (مصنف الغابات العشوائية)	0.9930318984205636
Support Vector Machine Classifier (مصنف آلة شعاع الدعم)	0.9335707649427067
Neural Network Classifier (مصنف الشبكات العصبية)	0.9907091978940848

تتمتع البيانات المدروسة بالتسلسل الزمني المتتالي واستقلالية بياناتها، مما يسمح ببناء أشجار القرار بشكل فعال لذلك نجد أن الغابات العشوائية هي الأكثر دقة لتصنيف البيانات، في هذه الطريقة نعتمد على مجموعة من قوانين اتخاذ القرار وفقاً لمجموعة السمات التي قمنا بدراستها (الاستهلاك، المنطقة، الفترة الزمنية، الحرارة، الرطوبة وغيرها من

السمات). من خلال تطبيق الشبكات العصبية وجدنا أنها تتمتع بمجال دقة مرتفع وتزداد مع زيادة عدد الدورات وإن كانت هذه الطريقة محكومة بكمية البيانات المدروسة. بالإضافة إلى ذلك، يمكن لنا استخدام السلاسل الزمنية في شبكات RNN (Recurrent Neural Network) للتنبؤ بمجال الاستهلاك المستقبلي بصورة مرتفعة.

من ناحية أخرى، نجد أن تقنية SVM أعطت دقة مقبولة إلا أنها أقل من التقنيتين السابقتين نظراً لأنها تعتمد على تحديد الفاصل بين مجالين من العينات وتحديد الخط الفاصل الأقل مما يرفع من نسبة الخطأ في هذه الطريقة لتقارب قيم الاستهلاك في بعض المناطق وصعوبة قيام هذه الخوارزمية بتحديد خط فاصل مقبول بين عينتين متقاربتين. تسمح الطريقة المتبعة عند دمجها مع السلاسل العصبية التكرارية بتصنيف المناطق إلى مجموعات مستقلة حسب قيمة الاستهلاك ضمنها والتنبؤ بقيم الاستهلاك المستقبلية والذي يسمح بتحديد ساعات تقنين متغيرة ودقيقة حسب قيم وأوقات الاستهلاك في المناطق المختلفة.

يمكن توسيع العمل بحيث يشمل توزيع هذه الاستهلاكات على المواقع الخدمية الأكثر فعالية (كالمصانع ومحطات المياه والمشافي) حيث أن متوسط استهلاك محافظة اللاذقية للطاقة الكهربائية يتراوح ما بين 125 وحتى 175 ميغاوات في حين أن الوارد الفعلي لمدينة اللاذقية لا يتعدى 100 ميغاوات.

يمكن أيضاً توسيع الدراسة أيضاً لضبط عدد ساعات الاستهلاك حسب ساعات اليوم ودراسة عوامل مختلفة أخرى كالرطوبة وسرعة الرياح وتأثيرها على استهلاك الطاقة الكهربائية.

كما يمكن تطبيق نموذج الحفاظ على الاستهلاك على بيانات محلية، ورفع فعالية تزويد الطاقة اعتماداً على وزن مكان الاستهلاك، الأمر الذي يسمح بعمليات التوفير لموارد الفيول والغاز في هذه الفترة وتخزينها في حين ترتفع في أكثر من مناسبة أخرى (مواسم الشتاء والمواسم السياحية وفق الاستهلاك الوارد إلى مدينة اللاذقية) مما يتطلب المزيد من الترشيد للطاقة الكهربائية.

References:

1.Han، J.; Kamber، M.; Pie، J. Data Mining Concepts and Techniques; Academic Press; Morgan Kaufmann Publisher: Waltham، MA 02451، USA، 2012.

2. Hiroyuki Mori, Kaoru Nakano, An Efficient Hybrid Intelligent Method for Electricity Price Forecasting, Dept. of Network Design, Meiji Unversity, Tokyo 164-8525, Japan, Procedia Computer Science 95 (2016) 287 – 296.

3.Jiawei H., Kamber M. DATA MINING, Concepts and Techniques, ELSELVIER, Second Edition, (2016) Pp.769.

4.Martín M., Acera M. Electricity Load Forecasting Using Machine Learning Techniques. University Pontificia of Salamanca. 2010, (318-320).

5. Zhang P., Wu X., Wang X., Bi Sh، Short-term load forecasting based on big data technologies, CSEE Journal of Power and Energy Systems (Volume: 1, Issue: 3, Sept. 2015).

6. Hambali M., Akinyemi A., Oladunjoye J., Yusuf N,Electric Power Load Forecast Using Decision Tree Algorithms, Computing, Information Systems, Development Informatics & Allied Research Journal, Vol. 7 No. 4, December, 2016.

7.Huang، J.Z. Automated Variable Weighing in K-Means Type Clustering. IEEE Trans. Pattern Anal. Mach. Intell. 2005، 27، 657–668.

8.Arai، K.; Barakbah، A.R. Hierarchial K-means: An Algorithm for Centroids Initialization for K-means. Rep. Fac. Sci. Eng. Saga Univ. 2007، 36، 25–31.

9.Khan، S.S.; Ahmad، A. Cluster Centre Initialization Algorithm for K-means clustering. Pattern Recognit. Lett. 2004، 25، 1293–1302.

10.Yedla، M.; Pathakota، S.R.; Srinivasa، T.M. Enhancing K-means Clustering Algorithm with Improved Initial Center. Int. J. Comput. Sci. Inf. Technol. 2010، 1، 121–125.
11.Shakti، M.; Antony، S.T. An Effective Determination of Initial Centroids in K-means Clustering Using Kernel PCA. Int. J. Comput. Sci. Inf. Technol. 2011، 2، 955–959.
12.Fahim، A.M. An Efficent Enhanced k-means Clustering algorithm. J. Zhejiang Univ. Sci. A2006، 7، 1626–1633.
13. Yangyang Fua, Zhengwei Li, Hao Zhang, Peng Xua, Using Support Vector Machine to Predict Next Day Electricity Load of Public Buildings with Sub-metering Devices, Science Direct, Volume 121, 2015, (1016-1022).
14.Turkay B., Demren D. Electrical Load Forecasting Using Support Vector Machine. Istanbul Technical University. (2011) (49-53).
15. Adele Cutler, D. Richard Cutler and John R. Stevens, Random Forests, Ensemble Machine Learning: Methods and Applications (pp.157-176).
16.Javeed S., Nizami A., ZAl G, Forecasting electric energy consumption using neural networks, Science Direct, Volume 23, Issue 12, December 1995, Pages 1097-1104.
17.Kotsiantis، S.B.; Pintelas، P.E. Recent Advances in Clustering: A Brief Survey. WSEAS Trans. Inf. Sci. Appl. 2004، 1، 73–81.
18.Jain، A.K. Algorithm for Clustering Data; Prentice-Hall: Englewood Cliffs، NJ 07632، USA،1988.
19.Amri، Y.; Fadhilah، A.L.; Setani، N.; Rani، S. Analysis Clustering of Electricity Usage Profile Using K-Means Algorithm. In IOP Conference Series: Materials Science and Engineering; IOP Publishing: city، US state abbrev. if applicable، country، 2016; Volume 105، p. 012020. https://iopscience.iop.org/article/10.1088/1757-899X/105/1/012020، Accessed on 12 April 2019.
20.Santamouris، M.; Mihalakakou، G.; Patargias، P.; Gaitani، N.; Sfakianaki، K.; Papaglastra، M.; Pavlou، C.; Doukas، P.; Primikiri، E.; Geros، V.; et al. Using Intelligent Clustering Technique to Classify the energy performance of School Buildings. Energy Build. 2007، 39، 45–51.
21. https://www.sfu.ca/~jackd/Stat302/Wk04-2_Full.pdf.
22. Dewey Lonzo Whaley, The Interquartile Range: Theory and Estimation, East Tennessee State University, 2005.
23. Lindsay I Smith, A tutorial on Principal Components Analysis, February 26, 2002.
24. Selecting the number of clusters with silhouette analysis on KMeans clustering , Available online: https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html, (accessed on 6 April 2019).
25. Umargono E., Suseno J., Gunawan V., K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula Advances in Social Science, Education and Humanities Research, volume 474, 2019.
26.Damayanti، R.; Abdullah، A.G.; Purnama، W.; Nandiyanto، A.B. Electricity Load Profile Analysis Using Clustering Techniques. In IOP Conference Series: Materials Science and Engineering; IOP Publishing، 2017; Volume 180، p. 012081. Available online:https://iopscience.iop.org/article/10.1088/1757-899X/180//012081 (accessed on 26 March 2019).
27.Prahastono، I.; King، D.J.; Ozveren، C.S. A review of Electricity Load Profile Classification methods. In Proceedings of the 42nd Universities Power Engineering Conference، Brighton، UK، 4–6 September 2007; pp. 1187–1191.

28.Molina-Solana، M.; Ros، M.; Ruiz، M.D.; Gómez-Romero، J.; Martin-Bautista، M.J. Data Science for Building Management: A review. Renew. Sustain. Energy Rev. 2017، 70، 598–609.
29.Kim، S.S. Variable Selection and Outlier Detection for Automated K-means clustering. Commun. Stat. Appl. Methods2015، 22، 55–67.
30.Bessa، R.J.; Trindade، A.; Mirinda، V. Spatial-Temporal Solar Power Forecasting for Smart Grids. IEEE Trans. Ind. Inform. 2014، 11، 232–241.
31.Ceci، M.; Corizzo، R.; Malerba، D.; Rashkovska، A. Soatial Autocorrelation and Entropy for Renewable Energy Forecasting. Data Min. Knowl. Discov. 2019، 33، 698–729.