Journal of Education and Science (ISSN 1812-125X)

www.edusj.mosuljournals.com

Determining the Quality of Dairy Products Using Machine Learning Techniques

Nour A. Fadel [1] Baydaa S. Bahnam [2]

[1,2] Department of Software Engineering, College of Computer Science & Mathematics, University of Mosul, Mosul, Iraq

Article information

Article history:
Received November 01, 2023
Accepted December 17, 2023
Available online March 01, 2024

Keywords:
Machine Learning (ML)
Support vector Machine (SVM)
K nearest neighbor (KNN)
Decision Tree (DT)
Bagging Dairy Product Quality

Correspondence:
Baydaa Sulaiman Bahnam
baydaa_sulaiman@uomosul.edu.iq

Abstract

Food products are an essential source of human life, so they have a very important place, and it will be important to monitor and determine their quality in a short time. Our study deals with a very important and indispensable food product, which is the milk product, which is considered the main and important element in people's lives, especially children, because it is the main source of their growth, building their bones, and strengthening. Because it is a highly perishable product, it must be monitored and its specifications must be monitored, because any gram of milk that is of low or poor quality may cause tons of milk to spoil, and also cause major financial losses. Therefore, a study was conducted to determine the quality of dairy product through machine learning algorithms (ML), which are support vector machine (SVM) algorithm, nearest neighbors (KNN) algorithm, decision tree (DT) algorithm and Bagging algorithm using milk dataset taken from data warehouse Kaggle. This data consists of 1059 samples and seven features. The proposed models were trained and tested with the aim of finding the best and most accurate model for detecting milk quality and were evaluated using the evaluation metrics: accuracy, precision, recall, f1_score and confusion matrix. According to the evaluation results three models: SVM, KNN, and DT outperformed Bagging algorithm, as they obtained the highest level for all metrics 100%. The SVM algorithm was the most efficient because its execution time was 0.146 seconds, which was less than the other models.

DOI: 10.33899/edusj.2023.144324.1405, ©Authors, 2024, College of Education for Pure Science, University of Mosul.
This is an open access article under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

1. المقدمة

تعد جودة الألبان من الامور ذات الأهمية الكبيرة في صناعة الأغذية والتغذية، حيث تعتبر الألبان من المصادر الرئيسية للبروتينات، والكالسيوم، والفيتامينات، والمعادن التي يحتاجها جسم الإنسان. وتتواجد الألبان في مجموعة متنوعة من المنتجات مثل الحليب، واللبن، والأجبان، ومشتقاتها [1]. يلعب الحليب دورًا مهمًا في النظم الغذائية على مستوى العالم [2]. فالحليب ومنتجات الألبان هي أغذية حيوانية ميسورة التكلفة في البلدان منخفضة إلى متوسطة الدخل [3]. ويعد إنتاج الحليب مصدرًا مهمًا لكسب الرزق لصغار مزارعي الألبان [4] . حيث يتزايد الطلب على منتجات الألبان في البلدان المنخفضة والمتوسطة الدخل، مدفوعًا بالنمو السكاني، وارتفاع الدخل، وتغير أنماط الحياة. هناك ضرورة ملحة للمزارعين أصحاب المزارع الصغيرة لإنتاج الحليب الذي يفي بمعايير سلامة الأغذية للاستفادة من هذا الطلب المتزايد على الألبان ومنتجات الألبان[5] .

ان الحليب منتج قابل للتلف. كل جرام حليب ذو جودة رديئة يمكن أن يتسبب في تلف أطنان من الحليب، مما يتسبب في خسائر مالية كبيرة. ملايين البكتيريا يمكن أن تتكون في الحليب الفاسد في وقت قصير جدا. بهذه الطريقة، إذا كان الناس يستهلكون الحليب أو منتجات الألبان قد تحدث حالات تعرض صحة الإنسان للخطر. تؤثر جودة الألبان على القيمة الغذائية للمنتجات الألبانية وعلى الصحة العامة، لذلك تُعتبر معايير الجودة والسلامة ضرورية لضمان توفير منتجات صحية ومأمونة للمستهلكين [6].

ومع التطورات الحاصلة في التكنولوجيا وبشكل واسع دعت الحاجة الى استخدام تقنيات التعلم الآلي في عمليات التصنيف لجودة منتجات الالبان لما لها مكانة مهمة لفائدتها. لذا من الضروري جدًا متابعة جودة هذه المنتجات التي تم انتاجها وتحديدها في وقت قصير باستخدام هذه التقنيات. لذلك تم في هذا العمل استخدام طرائق التعلم الالي: خوارزمية آلة ناقل الدعم (SVM) وخوارزمية الجيران الاقرب (KNN) وخوارزمية شجرة القرار (DT) وخوارزمية التعبئة Bagging لتقييم جودة منتج الحليب.

فيما يلي كيفية تنظيم هذه الورقة. الفقرة الثانية تتضمن البحوث السابقة. توضح الفقرة الثالثة العمل المقترح والخوارزميات المستخدمة في تقييم جودة منتج الحليب، النتائج والمناقشة موضحة في الفقرة الرابعة واخيرا تقدم الفقرة الخامسة الاستنتاجات وآفاق العمل المستقبلي.

2.الاعمال السابقة

قدمت دراسات متنوعة بالاونة الاخيرة في مجال جودة منتجات الالبان بالاعتماد على طرق التعلم الآلي. فقد قدم الباحثون (Xiao et al. 2019) ثلاث نماذج: نموذج الغابة العشوائية (RF) ، نموذج الانحدار اللوجستي (LR) ، ونموذج التعزيز التكيفي ((AdaBoosting(AB) وأجرى اختبارات للعثور على أنسب نموذج تصنيف بيانات الحليب، اهمها أظهر أن لون ورائحة وطعم الحليب كصفات لها القدرة على التعرف على جودة الحليب. وفقًا للنتائج التي تم حصلوا عليها، فقد كانت اعلى نسبة نجاح 96.8٪ باستخدام خوارزمية الغابة العشوائية [7]. اجرى الباحثون (Sambasivam et al. , 2020) تقدير معدل نقص الفيتامين د (VDD) من خلال استخدام عدة خوارزميات للتعلم الآلي: خوارزمية الجيران الاقرب (KNN)، شجرة القرار(DT) ، الغابة العشوائية ، التعزيز التكيفي AdaBoost (AB)، مصنف التعبئة (BC) ، الأشجار الإضافية (ET)، هبوط التدرج العشوائي (SGD)، التدرج التعزيز (GB) Gradient Boosting، والة متجه الدعم (SVM) ، وخوارزمية بيرسيبترون متعدد الطبقات (MLP). وقد تم تطبيق هذه الخوارزميات على بيانات تم جمعها لـ 3044 طالبًا جامعيًا تتراوح أعمارهم بين 18 و21 عامًا. وفقًا للنتائج التي حصلوا عليها فقد كانت اعلى نسبة 96 ٪ باستخدام خوارزمية الغابة العشوائية[8] . قدم الباحثان (Kavitha and Deepa, 2021) تحليل مقارن للكشف عن الحليب النقي من المغشوش باستخدام عدة خوارزميات للتعلم الآلي: الانحدار اللوجستي LR، Naive Bayes، الغابة العشوائية، آلة المتجهات الداعمة، والتدرج التعزيز (GB)[12]. استخدم الباحثون (Frizzarin et al., 2021) عدة طرق احصائية للتعلم الالي من اجل التنبؤ بسمات الحليب البقري الخام اعتمادا على التحليلات الكيميائية. وتبين ان النهج المخصص لمتوسط النموذج Model Averaging (MA) كان الافضل لقدرته على التنبؤ ب 6 من السمات الـ 14 التي تم فحصها[9]. استخدم الباحث (ÇELIK, 2022) الشبكة العصبية (NN) والتعزيز التكيفي AdaBoost (AB) لتصنيف جودة الحليب الى ثلاث مستويات (high, medium, low) وقد استخدام منصة Orange، وهي مفتوحة المصدر ومكتوبة بلغة بايثون ، كنظام أساسي للتطبيق. وفقا للنتائج التي تم حصل عليها؛ فقد كانت اعلى دقة للتصنيف بنسبة 99.9٪ باستخدام لخوارزمية AdaBoost و95.4٪ باستخدام خوارزمية الشبكة العصبية[10] . طور الباحثون (Zhang et al., 2022) نموذج أنف إلكتروني لتمييز مصدر الحليب، وتقدير محتوى الدهون والبروتين في الحليب، والتعرف على أصالة ودقة الحليب وتقييم جودة الحليب. واستخدام ثلاث خوارزميات للتعلم الآلي مثل الانحدار اللوجستي (LR)، وآلة ناقل الدعم (SVM)، والغابة العشوائية (RF) لبناء مصدر الحليب (مزرعة الأبقار) وتقييم ومقارنة تأثيرات التصنيف. اثبت النتائج التجريبية أن تأثير التصنيف لنموذج الدمج SVM-LDA المعتمد عل التحليل التمييزي الخطي LDA أفضل من النماذج الفردية الأخرى، فقد وصلت دقة مجموعة الاختبار إلى 91.5%[11].

يمكن أن يؤدي تحديد جودة الحليب بالطرق اليدوية إلى خطأ كبير أو ضياع الوقت. وان تحديد جودة الحليب بميزة واحدة او ميزتين او ثلاثة فقط لن يعطي نتائج دقيقة، لذا يجب تحديد هذه الجودة من خلال مراعاة العديد من الميزات. لهذا السبب، سيكون من الجيد تحديد جودة منتوج الحليب بنسب عالية الدقة بواسطة تقنيات التعلم الالي وباستخدام عدة ميزات. تم في هذه الرسالة تحديد جودة الحليب من خلال سبع ميزات وباستخدام اربع خوارزميات للتعلم الآلي: خوارزمية الة متجه الدعم Support Vector Machine (SVM) ، خوارزمية الجيران الاقرب k-nearest neighbor (KNN)، خوارزمية شجرة القرار Decision Tree (DT) وخوارزمية Bagging. حيث ان هذه الخوارزميات عالجت عدة مسائل في التصنيف والانحدار والعمليات الغير خطية وايضا تتوقع الاخراج بشكل سريع.

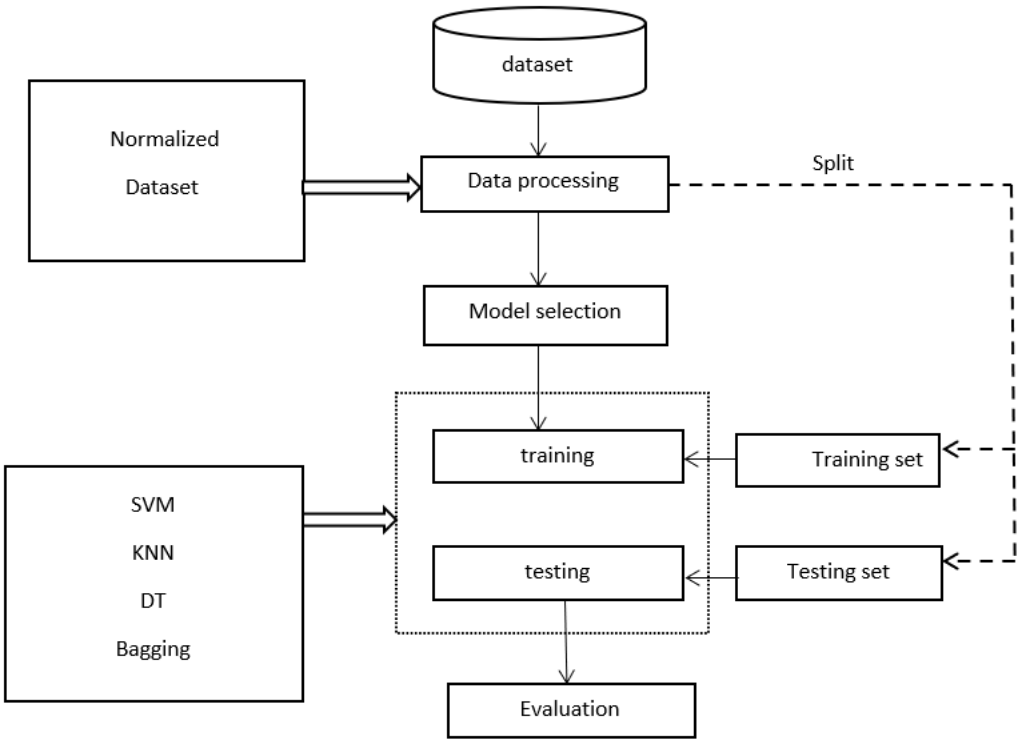
3.العمل المقترح

في هذه الفقرة تم توضيح خطوات العمل المقترح لتصميم وتنفيذ تصنيف جودة منتج الالبان باستخدام خوارزميات التعلم الآلي، حيث يتكون النظام المقترح في هذا البحث من خمس خطوات كما هو مبين في الشكل (1). وفيما يلي خطوات عمل النموذج المقترح:

الخطوة الاولى (المعالجة المسبقة للبيانات): تم استخدام تقنية تطبيع البيانات كمعالجة اولية في النطاق (0 و 1) من خلال طريقة (Min-Max).

الخطوة الثانية (تقسيم مجموعة البيانات المستخدمة): في هذا العمل تم تقسيم مجموعة البيانات إلى 80% لتدريب الخوارزميات و20% الاختبار.
الخطوة الثالثة (التطبيق): تطبيق اربع خوارزميات من تقنيات التعلم الآلي على مجموعة بيانات الحليب وذلك في خطوتين تدريب واختبار.

الخطوة الاخيرة (التقييم): هي عملية تقييم نماذج التعلم الالي التي تم تدريبها.



الشكل 1. مخطط العمل المقترح

1.3 البيانات المستخدمة

لغرض تدريب واختبار النماذج التي تم بناؤها للتصنيف الثنائي لجودة الحليب تم استخدام مجموعة البيانات الموجودة على موقع Kaggle [13]. حيث تتكون مجموعة بيانات الحليب من 1059. ويتم التصنيف الثنائي لكل عينة على انها ذات جودة عالية (high) او جودة منخفضة (low). والشكل (2) يوضح عينة مختارة من هذه البيانات.

pH	Temprature	Taste	Odor	Fat	Turbidity	Colour	Grade
6.6	35	1	0	1	0	254	high
6.6	36	0	1	0	1	253	high
8.5	70	1	1	1	1	246	low
9.5	34	1	1	0	1	255	low
6.6	37	0	0	0	0	255	medium
6.6	37	1	1	1	1	255	high
5.5	45	1	0	1	1	250	low
4.5	60	0	1	1	1	250	low
8.1	66	1	0	1	1	255	low
6.7	45	1	1	0	0	247	medium
6.7	45	1	1	1	0	245	medium
5.6	50	0	1	1	1	255	low
8.6	55	0	1	1	1	255	low
7.4	90	1	0	1	1	255	low
6.8	45	0	1	1	1	255	high
6.5	38	1	0	0	0	255	medium
4.7	38	1	0	1	0	255	low
3	40	1	1	1	1	255	low
9	43	1	0	1	1	250	low
6.8	40	1	0	1	0	245	medium
6.6	45	0	1	1	1	250	high
6.5	36	0	0	1	0	255	medium
4.5	38	0	1	1	1	255	low
6.6	45	1	1	1	1	245	high
6.8	35	1	0	1	0	246	medium

الشكل 2. عينة من مجموعة بيانات الحليب

وتتألف هذه البيانات من سبع سمات وهي:

1.الرقم الهيدروجيني (pH):

الرقم الهيدروجيني (pH) هو مقياس لدرجة الحموضة أو القاعدية في الحليب. حيث ان قيمة pH تتراوح عادة ما بين 6.4 و 6.8 للحليب الطبيعي. وان تغييرات في قيمة pH يمكن أن تؤثر على خصائص البروتينات والقوام والتخزين واستقراريه الحليب. على سبيل المثال، ارتفاع في الحموضة قد يؤدي إلى تخثر البروتينات في الحليب وتكوين اللبن.

2.درجة الحرارة (Temperature):

درجة الحرارة هي مؤشر آخر على جودة الحليب. الحفاظ على درجة حرارة منخفضة بعد الحصول على الحليب من الحيوان وأثناء التخزين يمكن أن يقلل من نمو البكتيريا وتدهور الحليب. درجة الحرارة تؤثر أيضًا على الخصائص الفيزيائية والكيميائية للحليب.

3.الذوق (Taste):

ان الذوق يعتبر عاملًا مهمًا لقبول الحليب. حيث ان طعم الحليب يمكن أن يكون مؤثرًا بشكل كبير على مدى تناوله، وهذا يتأثر بالعديد من العوامل منها توازن النكهات المختلفة ومدى وجود أي ملوثات أو تدهور. وتأخذ هذه الميزة في مجموعة البيانات القيمة 1 (طعم جيد) او 0 (طعم سيئ).

4.الرائحة (Odor):

الرائحة لها أيضًا تأثير قوي على تجربة الاستهلاك. وان رائحة الحليب يجب أن تكون نقية وممتعة، وأي تغيير في الرائحة يمكن أن يشير إلى تدهور أو تلف في الحليب. وتأخذ هذه الميزة القيمة 1 (رائحة جيدة) او 0 (رائحة سيئة).

5.الدهون (Fat):

تؤثر الدهون على الحليب، فالدهون تسهم في إعطاء الحليب نكهته وقوامه الكريمي، بالإضافة إلى دورها في تحسين قيمته الغذائية. وتأخذ القيم اما 1 (دسم) أو 0 (قليل الدسم).

6.العكارة (Turbidity):

ان العكارة تعكس شفافية الحليب. فالحليب ذو جودة عالية يجب أن يكون شفافًا وخاليًا من الجسيمات المعلقة. درجة العكارة تؤثر على المظهر العام ويمكن أن تكون مؤشرًا للتلوث أو التدهور. وتأخذ هذه السمة القيم الممكنة اما 1 (عكارة جيد) أو 0 (عكارة سيئة).

7.اللون (Color):

ان لون الحليب يُعتبر مؤشرًا مهمًا للجودة أيضًا. اللون الأبيض النقي هو اللون المرجو للحليب الطازج والصحيح. أي تغيير في اللون قد يشير إلى تدهور أو وجود مشكلة في الحليب.

تلك السمات السبعة تعمل سويًا لتحديد جودة وملائمة الحليب للاستهلاك البشري أو استخدامه في منتجات أخرى. تُعتبر هذه السمات مؤشرات هامة لمعرفة ما إذا كان الحليب يعتبر منتجًا صالحًا للأكل وذو جودة عالية. وهذا يمثل سمة الاخراج درجة الجودة (Grade) وفي عملنا هذا سيتم الكشف عن جودة الحليب اما الحليب ذات جودة عالية (high) او ذات جودة منخفضة (low).

2.3 معالجة البيانات

تم معالجة البيانات المستخدمة بخطوتين رئيسيتين:

1.عملية تصنيف الثنائي للبيانات حيث يتم في هذه المرحلة تم اجراء عملية التصنيف الثنائي على الخصائص السبعة من مجموعة بيانات الحليب وبهذا تم تصنيف كل عينة من عينات الحليب اما حليب عالي الجودة (high) او منخفض الجودة (low) كما في الجدول (1).

الجدول (1): التصنيف الثنائي لمجموعة بيانات الحليب

قيمة خاصية درجة الجودة قبل التصنيف	قيمة خاصية درجة الجودة قبل التصنيف
0	Low, Medium
1	High

2. تطبيع البيانات

من اهم التقنيات في معالجة البيانات هي تطبيع البيانات كونها تحسن أداء نظام التعلم الآلي ذلك لان بعض مجموعات البيانات تتالف من ميزات بقيم ونطاقات ومقاييس مختلفة جدًا. في هذا البحث تم استخدمت تقنية Min -Max لتطبيع سمات البيانات بنطاق (0-1) باستخدام المعادلة (1)، حيث $z \epsilon Z$(max ,min) و z′ينتمي الى(0 ,1)[14]:

$$\acute{z} = \frac{z - min}{max - min} \qquad (1)$$

3.3 تقسيم البيانات للتدريب والاختبار

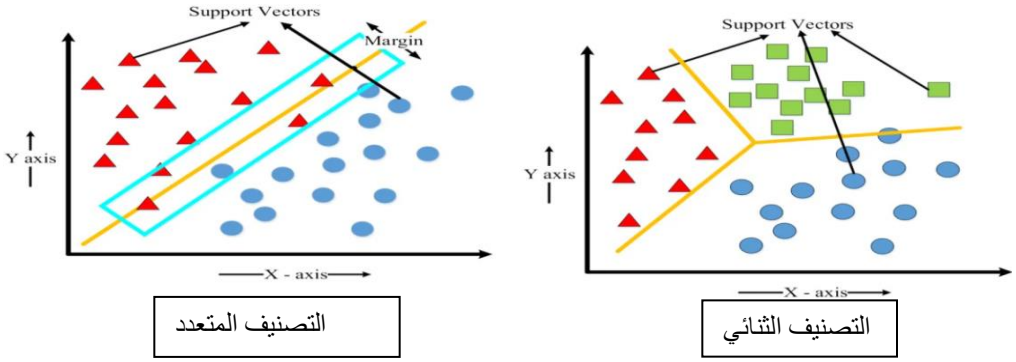
في هذه الخطوة من النموذج المقترح يتم تقسيم مجموعة بيانات الحليب الى 80% كبيانات تدريب وعددها 848 عينة و20% كبيانات الاختبار وعددها212 عينة.

4.3 تقنيات التعلم الآلي المقترحة

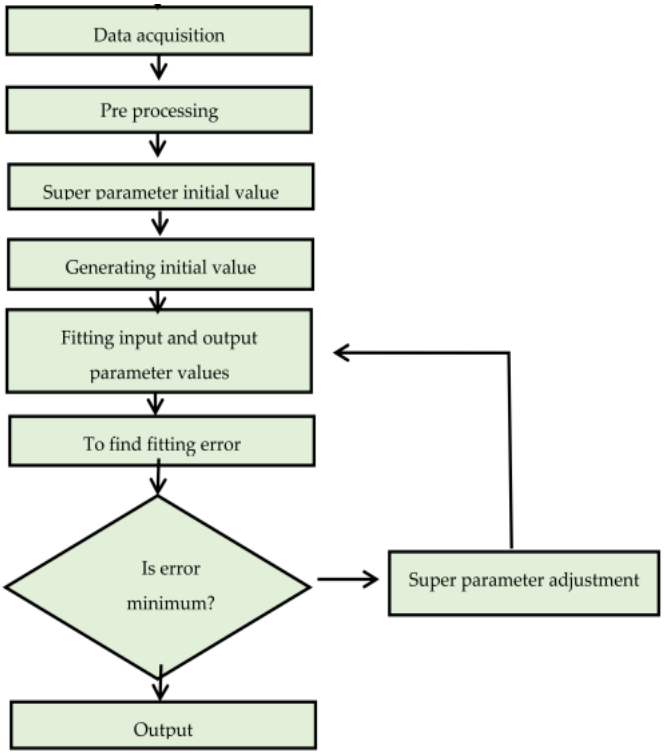
تم استخدام في هذا العمل اربع من خوارزميات التعلم الألي للكشف عن جودة منتج الحليب وهي خوارزمية الة متجه الدعم Support Vector Machine (SVM) ، خوارزمية الجيران الاقرب k-nearest neighbor (KNN)، خوارزمية شجرة القرار Decision Tree (DT) وخوارزمية Bagging .

1.خوارزمية الة متجه الدعم (SVM)

تعد الة متجه الدعم SVM طريقة من طرائق تعلم الالة الموجه المستعملة في التصنيف والتي تستند الى نظرية التعلم الاحصائي وتُستخدم لحل مشكلة تصنيف البيانات والتحكم في الانحدار[15]. تقوم هذه الخوارزمية بإيجاد الخطوط أو الأسطح التي تفصل بين الفئات المختلفة في مجموعة البيانات. تبحث هذه الخوارزمية عن مستوى فاصل مثالي يكون متساوي الابعاد لفصل البيانات الى فئتين two classesفي حالة التصنيف الثنائي الذي يصنف التدريب الى فئتين فقط. اما تطبق التصنيف المتعددmulticlass svm اذا كانت بيانات التدريب تصنف لأكثر من فئتين[16] . وهنا تكمن الفكرة الرئيسة من svm في بناء مستوى مثالي optimal hyperplane في فضاء يعمل على حل مسائل التصنيف وتمييز النماذج وكلما ازداد حجم المستوى المثالي زادت كفاءة الخوارزمية ودقة التصنيف، ولتحديد الهامش الاكبر maximum margin يتم بناء مستويين متوازيين يقعان على جانبي الهامش الذي يمثل المسافة الواقعة بين المستويين المتوازيين. لذا فان خوارزمية تعلم الالة svm تستند الى مفهوم مستويات القرار التي تعرف حدود القرار ، ومستوى القرار هو الحد الذي يفصل بين مجموعة الكيانات التي تمتلك انتماءات قوية او تصنيفية مختلفة وخوارزمية svm تجد المستوى المثالي ذو الهامش الاكبر لفصل الفئات. والفائدة من الهامش هو تجنب وقوع في المطب محلي والحصول على افضل تصنيف[17][18] والشكل(3) يوضح المستوى المثالي لعملية تصنيف لمسالة ثنائية (أ) والتصنيف المتعدد (ب) بالاعتماد على خوارزمية الة متجه الدعم [19] . اما الشكل (4) يوضح المخطط الانسيابي لخوارزمية متجه الة الدعم [20] .



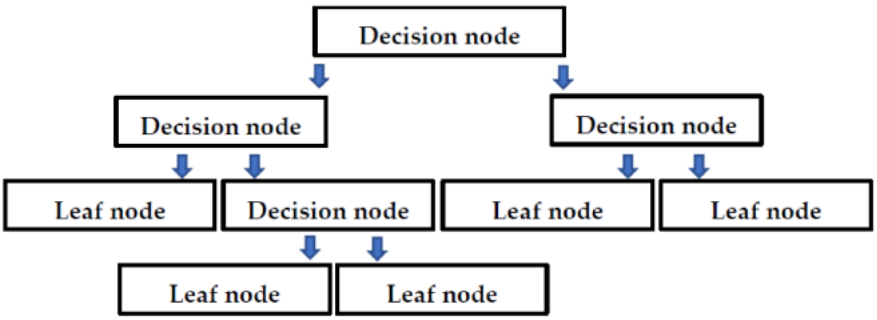
الشكل (3) المستوى المثالي لعمليتي التصنيف الثنائي والمتعدد بالاعتماد على خوارزمية الة متجه الدعم[19]



الشكل 4. المخطط الانسيابي لخوارزمية الة متجه الدعم [20]

2. خوارزمية شجرة القرار.

شجرة القرار هي تقنية في تعلم الآلة تُستخدم لحل مشكلات التصنيف والتنبؤ. تعتمد شجرة القرار على تقسيم البيانات إلى فئات أو تصنيفات مختلفة باستخدام سلسلة من القرارات الثنائية. تعتمد شجرة القرار على مجموعة من عبارات if-else ويتم تصنيفها وفقًا للشروط. وتتكون من عدة عقد، تعرف أيضًا بالأوراق، كما هو موضح في الشكل (5). ويتم تطبيق اختبار على كل ورقة، مما يرسل استعلامًا إلى فروع تلك العقدة. تستمر هذه الحلقة حتى يصل الاستعلام إلى ورقة النهاية. والقيمة التي يتم إرجاعها مرتبطة بكل عقدة ورقية بالشجرة. يجب أن تركز هذه العقدة الورقية على بناء أصغر شجرة من خلال التركيز على السمات الرئيسية أولاً. من السمات المهمة تنظيم العينات في مجموعات. بعد أن تقوم السمة الأولى بتقسيم العينات، تصبح العينات المتبقية مشاكل DT نفسها ولكن مع عينات أقل وسمة واحدة أقل، يمكن لهذه الأشجار الفرعية ذات السمات الأقل أهمية التغلب على التعقيد. كلما زاد عدد العينات في العقدة يعني مستوى تعقيد أعلى. تحتوي العقدة المتجانسة على عينة من فئة واحدة، مما يقلل من التعقيد. هدف العقدة هو زراعة الأشجار من خلال المحاولة المتكررة للحصول على عقد ورقية واضحة قدر الإمكان عن طريق تقليل فئات العينة [21][22].

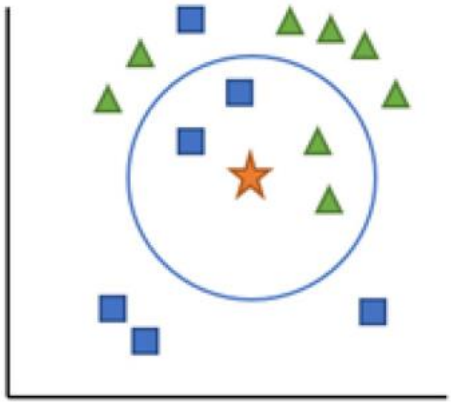


الشكل 5.المخطط الانسيابي لشجرة القرار[20]

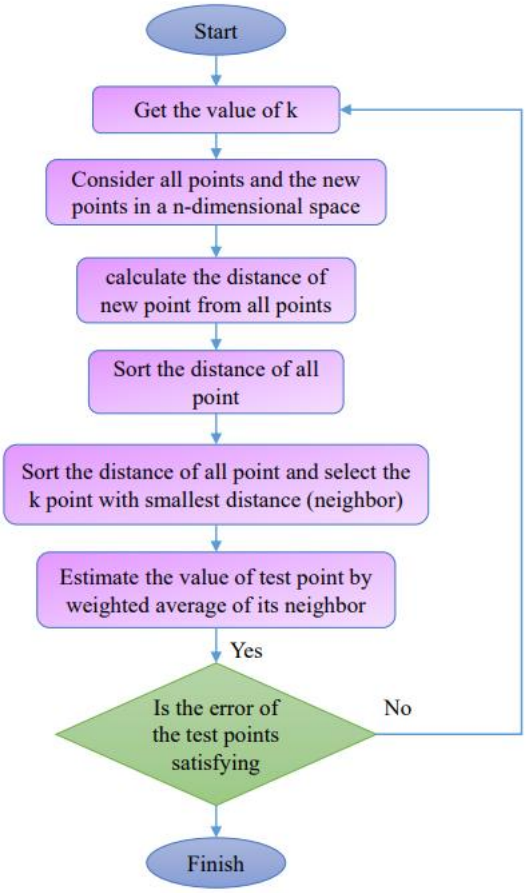
3. خوارزمية الجيران الاقرب k-Nearest Neighbors (KNN)

يعد مصنف KNN واحدًا من أبسط المصنفات وأكثرها استخدامًا في خوارزميات التصنيف . تم اقتراح KNN في عام 1951 بواسطة Fix and Hodges وتم تعديله بواسطة Cover and Hart. يمكن استخدام هذه التقنية لكل من التصنيف والانحدار. يعتمد المفهوم الرئيسي لـ KNN على حساب المسافات بين العينات التي تم اختبارها وعينات بيانات التدريب من أجل تحديد أقرب جيرانها. يتم بعد ذلك تعيين العينة التي تم اختبارها ببساطة إلى فئة أقرب جار لها [23].

في KNN ، تمثل قيمة K عدد أقرب الجيران. هذه القيمة هي العامل الحاسم الأساسي لهذا المصنف بسبب قيمة k التي تحدد عدد الجيران الذين يؤثرون على التصنيف. عندما k=1 سيتم تعيين كائن بيانات الجديد إلى فئة أقرب جار له. يتم أخذ الكائنات المجاورة من مجموعة من كائنات بيانات التدريب حيث يكون التصنيف الصحيح معروفًا بالفعل. تعمل خوارزمية KNN بشكل طبيعي مع البيانات الرقمية وتستخدم مفهوم المسافة لتصنيف كائنات البيانات مثل المسافات الإقليدية، ومانهاتن. وتعد المسافة الإقليدية هي الأكثر استخدامًا على نطاق واسع مع KNN [24]. والشكل (6) يوضح المصنف الجيران الاقرب KNN عند(k=4)[24] اما الشكل (7) يوضح المخطط الانسيابي للخوارزمية [25].



الشكل 6. المصنف الجيران الاقرب KNN (k=4)[24]

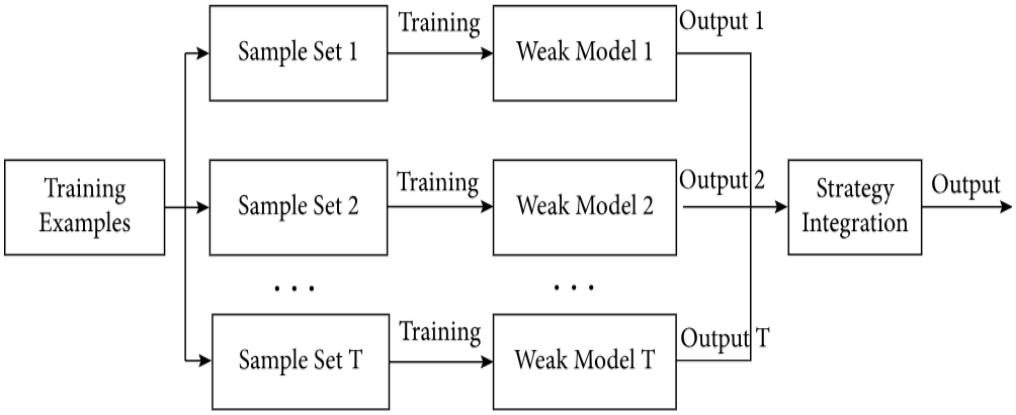


الشكل 7. المخطط الانسيابي لخورزمية الجيران الاقرب [25]

4.خوارزمية Bagging

خوارزمية Bagging هي اختصار لـ"Bootstrap Aggregating" ، وهي تقنية في تعلم الآلة تستخدم لتحسين أداء النماذج التنبؤية. تُعرف هذه التقنية أيضًا باسم التجميع التمهيدي. إنها واحدة من أقدم وأبسط تقنيات تعلم الآلة، ولكنها أكثر فعالية. ويحسب تنبؤاته النهائية عن طريق حساب متوسط النتائج من جميع أشجار القرار المبنية على مجموعات فرعية من التدريب التمهيدي [26][27]. في طريقة المجموعة هذه، يتم استخدام إجراء أخذ عينات التمهيد للحصول على مجموعات فرعية عشوائية من مجموعة التدريب الخاصة بالنموذج الأصلي مع الاستبدال. يتم استخدامه في الانحدار والتصنيف لتحسين دقة أساليب تعلم الآلة من خلال المساعدة في استنتاج التباين وتعزيز قوة النموذج باستخدام أشجار القرار. تتمثل المعلمات الفائقة في تقنيات التعبئة بشكل أساسي في اختيار النماذج الأساسية مثل أشجار القرار وعدد النماذج الأساسية المستخدمة [28] .

الفائدة الرئيسية للخوارزمية Bagging هي زيادة الاستقرار وتقليل التباين في أداء النماذج التنبؤية، مما يساعد في تحسين الأداء العام للنموذج وتقليل مشكلة الانحياز(Over fitting) [29]. والشكل (8) يوضح المخطط الانسيابي لخوارزمية Bagging [30].



الشكل 8. المخطط الانسيابي لخورزمية Bagging [30]

5.3 مقاييس التقييم

ان الغاية من المقاييس معرفة مدى دقة خوارزميات التعلم الآلي في تقييم جودة الالبان وذلك من خلال المقارنة بالتقييم الحاصل بالبيانات مع التقييم المتوقع من خلال الخوارزميات باستخدام مقاييس التقييم [31][32] وهناك العديد من مقاييس التقييم وقد استخدمت خمسة مقاييس وهي الأكثر شيوعا واستخداما ولغرض المقارنة مع اعمال سابقة تم استخدام نفس المقاييس، وان الرموز المستخدمة في المعادلات موضحة بالشكل الاتي :

- True positive (TP):هو التوقع الايجابي الحقيقي الجودة.
- True negation (TN): هو التوقع السلبي الحقيقي الجودة.
- False positive (FP): هو التوقع الايجابي الكاذب للجودة.
- False negation (FN): هو التوقع السلبي الكاذب الجودة.

وفيما يلي مقاييس التقييم المستخدمة [31][32] :

1.5.3 الدقة (Accuracy)

هي العدد الكلي للتنبؤات الصحيحة مقسوما على العدد الكلي للتنبؤات التي تم اجراؤها على مجموعة البيانات، ادق دقة هي 1 بينما اقل دقة هي 0 ، والتي يمكن حسابها عن طريق المعادلة التالية:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

2.5.3 الضبط (Precession)

هي نسبة التوقعات الإيجابية الصحيحة (TP) مقسوما على العدد الكلي للتوقعات الإيجابية وافضل دقة هي 1 واسوأ دقة هي 0 ، والتي يمكن حسابها عن طريقة المعادلة التالية:

$$Precession = \frac{TP}{TP + FP} \quad (3)$$

3.5.3 معدل الاستدعاء(Recall)

هي نسبة التوقعات الإيجابية الصحيحة (TP) مقسوما على التوقعات الإيجابية الصحيحة مع التوقعات السلبية الكاذبة وكما في المعادلة التالية:

$$Recall = \frac{TP}{TP + FN} \quad (4)$$

4.5.3 F1_score

هو المتوسط التوافقي للدقة والاستدعاء والتي يمكن حسابها عن طريق المعادلة التالية:

$$f1_{_score} = 2 * \frac{precision * recall}{precision + recall} \quad (5)$$

5.5.3 مصفوفة الارتباك (confusion matrix)

هو عبارة عن جدول يعرض فيه نتائج التوقع للتصنيف حيث يلخص فيه قيم التوقع الصحيحة والخاطئة من خلال المقارنة مع قيم التدريب والتي يتم وصفها بالصواب والخطأ مع قيم التنبؤ التي يتم وصفها بالإيجابية والسلبية [33] . في هذا البحث استخدمت مصفوفة الارتباك لمعرفة تقييم جودة الحليب حيث يوضح الجدول (2) نموذج مصفوفة الارتباك.

الجدول 2. مصفوفة الارتباك

Actual	Prediction		
	Positive (1)	True Positive (TP)	False Positive (FP)
	Negative (0)	False Negative (FN)	False Positive (FP)

4.مناقشة النتائج

لقد تم تدريب واختبار النماذج الاربعة المقترحة للكشف عن جودة الحليب اما حليب ذات جودة عالية (high) او ذات جودة منخفضة (low) باستخدام مجموعة بيانات الحليب الموجودة على موقع Kaggle. والجدول (3) يوضح نتائج اختبار النماذج المقترحة اعتمادا على الخوارزميات الاربعة.

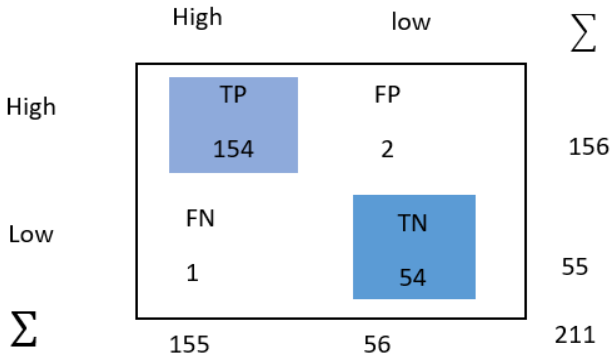
الجدول (3). نتائج تنفيذ النماذج المقترحة لمجموعة بيانات منتج الحليب.

Model	Accuracy	Precision	Recall	F1	Time
SVM	100	100	100	100	0.146
KNN	100	100	100	100	0.850
DT	100	100	100	100	0.814
Bagging	99.905	100	99.583	99.789	1.571

بناءً على النتائج التي تم الحصول عليها من الجدول اعلاه، يتبين إن الخوارزميات الاربعة (SVM، KNN،DT وBagging) تعمل بشكل ممتاز على مجموعة بيانات بالحليب وبدقة عالية. هذا يشير إلى أن هذه الخوارزميات ربما تكون مناسبة لمهمة الكشف عن جودة الحليب. حيث نلاحظ أن الخوارزميات الثلاثة (SVM، KNN، DT) حصلوا على أعلى مستوى لجميع المقاييس حيث كانت نتائج هذه الخوارزميات الثلاثة (100%) لمقياس Accuracy ، Precision ، Recall ومقياس f1 . في حين تفاوت وقت تنفيذ كل منهم المستغرق بالثانية. فقد ابدت خوارزمية SVM اقل وقت تنفيذ تليها DT ثم KNN . اماخوارزمية Bagging فكان ادائها اقل حيث حصلت على دقة (99.905%) ، مقياس الاحكام (100%) ، مقياس الاستدعاء (99.583%) ومقياس f1(99.789%).

وقد تم اجراء تقييم كمي للبيانات عن طريق مقياس مصفوفة الارتباك وعمل مقارنة بين النماذج حيث يمثل الشكل (9) مصفوفة الارتباك للخوارزميات الثلاثة SVM وKNN وDT.اما والشكل (10) يبين مصفوفة الارتباك لخوارزمية Bagging.

الشكل (9). مصفوفة الارتباك للخوارزميات الثلاثة SVM,DT,KNN



والشكل (10). مصفوفة الارتباك الخوارزمية Bagging.

ولتوضيح نتائج العمل المقترح مع اعمال الدراسات السابقة في مجال تقيم جودة الحليب سواء بالكشف عن جودة الحليب او تصنيفه وباستخدام مجاميع مختلفة من البيانات، فقد تم عمل ملخص عن كل دراسة في نفس المجال موضحة في الجدول (4).

الجدول (4). نتائج العمل المقترح مع اعمال البحوث والدراسات السابقة في مجال تقييم جودة الحليب.

ت	البحوث	مجموعة بيانات	الخوارزميات المستخدمة	المقاييس	النتائج	ملاحظات
1	Xiao et al., 2019[7]	مجموعة بيانات الحليب	1.Random Forest (RF) 2. Linear Regression (LR) 3. AdaBoost (AB)	Accuracy	اعلى نسبة نجاح تم الحصول عليها باستخدام الغابة العشوائية 96.8%	استخدام ثلاث سمات لون ورائحة وطعم الحليب لتصنيف جودة الحليب
2	Sambasivam et al., 2020[8]	جمع البيانات الأولية التي تحتوي على مستويات فيتامين د في الدم من إجمالي 3044 طالبًا جامعيًا تتراوح أعمارهم بين 18 و21 عامًا.	1. K nearest Neighbor (KNN) 2. Decision tree (DT) 3. RF 4. AdaBoost (AB) 5. Bagging Classifier (BC) 6. Extra Tree (ET) 7. Stochastic Gradient Descent (SGD) 8. Gradient Boosting (GB) 9. Support Vector Machine (SVM) 10. Multi-Layer Perceptron (MLP)	Accuracy Precision Recall F1-score	وفقًا للنتائج التي حصلوا عليها ، فقد كانت اعلى نسبة 96 % باستخدام خوارزمية الغابة العشوائيةRF	تم استخدام 11 معلمة للتنبؤ بخطورة نقص فيتامين د مع فئات عمرية محددة
3	Kavitha and Deepa, 2021[9]	جمع الباحث عينات من الحليب النقي والمغشوش	1.LR 2. Naive Bayes 3.RF 4.SVM 5. Gradient Boosting (GB)	Accuracy Precision	تبين ان الغابة العشوائية كانت الاكفأ	تحليل مقارن للكشف عن الحليب النقي من المغشوش

4	Frizzarin et al., 2021[10]	عينات الحليب الماخوذة من 622 بقرة	1. Partial least squares Regression (PLSR) 2. ridge regression (RR) 3. least absolute shrinkage and selection operator (LASSO) 4. Model Averaging (MA) 5. Neural Network (NN) 6. Partial least squares discriminant analysis (PLSDA) 7. RF 8.Boosting Decision trees 9. SVM	Accuracy Root mean square error (RMSE)	تبين ان النهج المخصص Model (MA)Averaging كان الافضل لقدرته على التنبؤ بجودة الحليب البقري الخام ولـ 6 من السمات الـ 14.	في التحليلات الكيميائية تعتبر الطرق الاحصائية افضل من غيرها كون نتائجها قابلة للتفسير اكثر بالاضافة الى ضبط معلماتها اقل من بقية الطرق الاخرى
5	ÇELIK, 2022[11]	مجموعة بيانات الحليب	1. Neural Network (NN) 2. Adaptive Boosting (AB)	Accuracy Precision Recall F1-score	اعلى دقة للتصنيف بنسبة 99.9% باستخدام خوارزمية AB بينما 95.4% باستخدام NN	استخدام منصة Orange، وهي مفتوحة المصدر ومكتوبة بلغة بيثون لتصنيف جودة الحليب الى ثلاث اصناف وباستخدام سمتين درجة حموضة وحرارة الحليب
6	Zhang et al., 2022[12]	تم جمع 1000 مجموعة من بيانات الحليب من 10 مزارع في اماكن مختلفة	1. logistic Regression (LR) 2. RF 3. SVM	Accuracy	نموذج الدمج SVM-LDA المعتمد عل التحليل التمييزي الخطي LDA أفضل من النماذج الفردية الأخرى، فقد وصلت دقة مجموعة الاختبار 91.5%.	تطوير نموذج أنف إلكتروني لتمييز مصدر الحليب، وتقدير محتوى الدهون والبروتين في الحليب، والتعرف على أصالة ودقة الحليب وتقييم جودة الحليب.
7	**العمل المقترح في دراستنا**	**مجموعة بيانات الحليب من موقع كاجل**	**1.SVM 2. KNN 3. DT 4. Bagging**	**Accuracy Precision Recall F1-score**	**نموذج خوارزمية SVM اكفأ النماذج الاخرى لحصولها على أعلى مستوى 100% لجميع المقاييس وباقل وقت تنفيذ 0.146 ثانية**	**استخدام سبع سمات للحليب: الرقم الهيدروجيني، درجة الحرارة، الذوق، الرائحة، الدهون ولون الحليب للكشف عن جودة الحليب**

ويتضح من الجدول اعلاه ان دراسة [7] و[11] صنفت بيانات الحليب الى ثلاث اصناف (high, medium, low) واستخدمت دراسة [7] ثلاث سمات لون ورائحة وطعم الحليب بينما [11] استخدم سمتين هما درجة الحموضة ودرجة الحرارة لتصنيف الحليب. اما في دراستنا استخدمنا سبع ميزات للحليب : الرقم الهيدروجيني، درجة الحرارة، الذوق، الرائحة، الدهون ولون الحليب للكشف عن جودة الحليب اما حليب ذات جودة عالية او ذات جودة منخفضة وتم الحصول على افضل النتائج 100% لجميع المقاييس المستخدمة لنموذج SVM وباقل وقت تنفيذ 0.146 ثانية. اما دراسة [8] حددت خطورة نقص فيتامين د لـ 3044 طالب جامعي وحصلت خوارزمية الغابة العشوائية على 96 %. وايضا تبين من دراسة [9]، ان خوارزمية الغابة العشوائية كانت الافضل في الكشفت عن جودة الحليب اما حليب نقي او حليب مغشوش. وتنبأت دراسة [10] بجودة الحليب البقري الخام من خلال استخدام 6 سمات وبالاعتماد على التحليلات الكيميائية وكان نموذج MA الافضل. واستخدام [12] نموذج انف الكتروني لتمييز محتوى الدهون والبروتين في الحليب وتقييم جودته وكانت اعلى دقة 91.5% للنموذج SVM-LDA.

5.الاستنتاجات

تم في هذه الدراسة استخدم اربع خوارزميات التعلم الآلي لتدريب واختبار مجموعة بيانات الحليب الماخوذة من موقع كاجل من اجل الكشف عن جودة الحليب اما حليب ذات جودة عالية او ذات جودة منخفضة. فقد تم استخدام خوارزمية SVM، KNN، DT و Bagging وكانت دقة الخوارزميات الثلاثة 100% لجميع المقاييس المستخدمة في حين دقة خوارزمية Bagging هي 99.905 %. نستنتج من هذا البحث ان افضل خوارزمية هي SVM لحصولها على أعلى دقة باقل وقت تنفيذ في الكشف عن جودة الحليب. نقترح بالمستقبل كشف وتصنيف جودة الحليب باستخدام خوارزميات اخرى بالتعلم اللالي والعميق ومقارنتها مع نتائج الخوارزميات المستخدمة.

شكر وتقدير

يود المؤلفون أن يشكروا جامعة الموصل / كلية علوم الحاسوب والرياضيات على تسهيلاتهم، مما ساعد على تحسين جودة هذا العمل.

Resources:

[1] J. Han and J. J. A. Wang, "Dairy Cow Nutrition and Milk Quality," Agriculture, vol. 13, 702,, 2023, https://doi.org/10.3390/agriculture13030702.

[2] S. Alonso, E. Muunda, S. Ahlberg, E. Blackmore, and D. Grace, "Beyond food safety: Socio-economic effects of training informal dairy vendors in Kenya," Global Food Security , vol. 18, pp. 86-92, 2018, https://doi.org/10.1016/j.gfs.2018.08.006

[3] E. Muunda, N. Mtimet, F. Schneider, F. Wanyoike, P. Dominguez-Salas, and S. J. F. P. Alonso, "Could the new dairy policy affect milk allocation to infants in Kenya? A best-worst scaling approach," vol. 101, p. 102043, 2021 https://doi.org/10.1016/j.foodpol.2021.102043.

[4] G. Msalya, "Contamination levels and identification of bacteria in milk sampled from three regions of Tanzania: evidence from literature and laboratory analyses," Veterinary Medicine International, 1–10. https://doi.org/10.1155/2017/9096149, 2017.

[5] R. Ramsing et al., "Dairy and Plant-Based Milks: Implications for Nutrition and Planetary Health," Curr Envir Health Rpt 10, 291–302, 2023, https://doi.org/10.1007/s40572-023-00400-z

[6] H. Lemma D, A. Mengistu, T. Kuma, B. J. F. Q. Kuma, and Safety, "Improving milk safety at farm-level in an intensive dairy production system: relevance to smallholder dairy producers," vol. 2, no. 3, pp. 135-143, 2018, https://doi.org/10.1093/fqsafe/fyy009.

[7] L. Xiao, K. Xia, and H. Tian, "Research on Classification Model of Fermented Milk Quality Control Based on Data Mining," in 2019 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), 2019, pp. 324-327: IEEE.

[8] G. Sambasivam, J. Amudhavel and G. Sathya, "A Predictive Performance Analysis of Vitamin D Deficiency Severity Using Machine Learning Methods," in IEEE Access, vol. 8, pp. 109492-109507, 2020, https://doi.org/10.1109/ACCESS.2020.3002191.

[9] P. V. KavithaP. V. Deepa, "A comparative analysis of the machine learning methods for milk adulteration detection", AIP Conference Proceedings 2408, 030008 (2021), vol 2408, Issue1.

[10] M. Frizzarin et al., "Predicting cow milk quality traits from routinely available milk spectra using statistical machine learning methods,"Journal of Dairy Science, vol. 104, no. 7, pp. 7438-7447, 2021.

[11] A. ÇELİK, "Using Machine Learning Algorithms to Detect Milk Quality," Eurasian Journal of Food Science and Technology, vol. 6, no. 2, pp. 76-87, 2022.

[12] Y. Zhang, L. Zhang, Y. Ma, J. Guan, Z. Liu, and J. Liu, "Research on dairy products detection based on machine learning algorithm," MATEC Web Conf., vol. 355, p. 03008, 2022. [Online]. Available: https://doi.org/10.1051/matecconf/202235503008.

[13] Kaggle Inc. 2022. Kaggle. https://www.kaggle.com/general

[14] P. J. Muhammad Ali, "Investigating the Impact of Min-Max Data Normalization on the Regression Performance of K-Nearest Neighbor with Different Similarity Measurements", ARO, vol. 10, no. 1, pp. 85-91, Jun. 2022.

[15] A. Kulkarni, V. Hundekar, S. Sannakki, and V. Rajpurohit, "Survey on Opinion Mining Algorithms and Applications," International Journal of Computer Techniques, vol. 4, no. 3, pp. 9, 2017.

[16] B. S. Bahnam and S. Abd Dawwod, "A proposed model for diabetes mellitus classification using coyote optimization algorithm and least squares support vector machine," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 11, no. 3, p. 1164, 2022, https://doi.org/10.11591/ijai.v11.i3.pp1164-1174

[17] F.M. Siemers and J. Bajorath," Differences in learning characteristics between support vector machine and random forest models for compound classification revealed by Shapley value analysis.," Sci Rep 13, 5983, 2023. https://doi.org/10.1038/s41598-023-33215-x

[18] Khairnar J. and Kinikar M. "Machine Learning Algorithms for Opinion Mining and Sentiment Classification", International Journal of Scientific and Research Publications, vol. 3, Issue 6, pp.1-6, 2013.

[19] R. Muzzammel and A. Raza, "A Support Vector Machine Learning-Based Protection Technique for MT-HVDC Systems," Energies, vol. 13, no. 24, p. 6668, Dec. 2020, https://doi.org/10.3390/en13246668.

[20] A. Nafees, et al.," Modeling of Mechanical Properties of Silica Fume-Based Green Concrete Using Machine Learning Techniques," Polymers, vol.14, no.30, 2022, https://doi.org/10.3390/polym14010030

[21] S. Chithra, S.S. Kumar, K. Chinnaraju, F.A.A. Ashmita," comparative study on the compressive strength prediction models for High Performance Concrete containing nano silica and copper slag using regression analysis and Artificial Neural Networks," Constr. Build. Mater. 2016, 114, 528–535. https://doi.org/10.1016/j.conbuildmat.2016.03.214

[22] H. Tanyildizi, " Prediction of the Strength Properties of Carbon Fiber-Reinforced Lightweight Concrete Exposed to the High Temperature Using Artificial Neural Network and Support Vector Machine," Adv. Civ. Eng. 2018, 5140610, | https://doi.org/10.1155/2018/5140610

[23] M. Mohammed, M.B. Khan and E.B.M. Bashier," Machine learning: algorithms and applications,". CRC Press, Boca Rato, 2016.

[24] N. Ali, D. Neagu, & P.Trundle," Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets,". SN Appl. Sci. 1, 1559, 2019, https://doi.org/10.1007/s42452-019-1356-9

[25] M.R.Mahdiani, E. Khamehchi, S. Hjirezaei and A. Hemmati-Sarapardeh, "Modeling viscosity of crude oil using k-nearest neighbor algorithm,"ADVANCES IN GEO-ENERGY RESEARCH, 2020, https://doi.org/10.46690/ager.2020.04.08

[26] Zhang, C.; Ma, Y. Ensemble Machine Learning: Methods and Applications; Zhang, C., Ma, Y., Eds.; Springer: Boston, MA, USA, 2012; ISBN 978-1-4419-9325-0. [**Google Scholar**]

[27] J. Rocca, " Ensemble Methods: Bagging, Boosting and Stacking," Available online: https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205 (accessed on 25 October 2023).

[28] G.E. Adjovu, H.Stephen and S., "Ahmad, A Machine Learning Approach for the Estimation of Total Dissolved Solids Concentration in Lake Mead Using Electrical Conductivity and Temperature," Water, vol.15, no.13, pp.2439, 2023, https://doi.org/10.3390/w15132439

[29] X. Wang, S. Gao, S. Zhou, Y. Guo, Y. Duan, and D. Wu, "Prediction of House Price Index Based on Bagging Integrated WOA-SVR Model," Mathematical Problems in Engineering, vol. 2021, p. 3744320, 2021/10/29 2021, https://doi.org/10.1155/2021/3744320

[30] C. Yang, Y. Zhao, X. Cai, W. Wei, X. Feng, and K. Zhou, "Path Planning Algorithm for Unmanned Surface Vessel Based on Multiobjective Reinforcement Learning," Computational Intelligence and Neuroscience, vol. 2023, p. 2146314, 2023/02/15 2023, https://doi.org/110.1155/2023/2146314

[31] Ž. Đ. Vujovic," Classification Model Evaluation Metrics," International Journal of Advanced Computer Science, vol.12, no. 6, 2021, https://doi.org/10.14569/IJACSA.2021.0120670

[32] R. Yilmaz and F. H. YAĞIN, "Early detection of coronary heart disease based on machine learning methods," Medical Records, vol. 4, no. 1, pp. 1-6, 2022. https://doi.org/10.37990/medr.1011924

[33] A. Tharwat, "Classification assessment methods," Applied Computing and Informatics," vol.17, no.1, pp.168–192, 2018, https://doi.org/10.1016/J.ACI.2018.08.003 enic, 4-hydroxy benzoic.

تحديد جودة منتجات الألبان باستخدام تقنيات التعلم الآلي

نور عبد الخالق[1]، بيداء سليمان بهنام[2]

[2، 1] قسم البرمجيات، كلية علوم الحاسوب والرياضيات، جامعة الموصل، الموصل، العراق

الخلاصة:

تعتبر المنتجات الغذائية مصدراً أساسياً لحياة الإنسان، لذلك لها مكانة مهمة جداً، وسيكون من المهم مراقبة جودتها وتحديدها في وقت قصير. تتناول دراستنا منتجاً غذائياً مهماً جداً ولا يمكن الاستغناء عنه، وهو منتج الحليب، الذي يعتبر العنصر الأساسي والمهم في حياة الناس، وخاصة الأطفال، لأنه المصدر الرئيسي لنموهم، وبناء عظامهم، وتقويتها. ولأنه منتج سريع التلف، فلا بد من مراقبته ومراقبة مواصفاته، لأن أي جرام من الحليب يكون منخفض الجودة أو رديء الجودة قد يتسبب في فساد أطنان من الحليب، ويتسبب أيضًا في خسائر مالية كبيرة. لذلك تم إجراء دراسة لتحديد جودة منتج الألبان من خلال خوارزميات التعلم الآلي (ML) وهي خوارزمية آلة ناقل الدعم (SVM) وخوارزمية الجيران الاقرب (KNN) وخوارزمية شجرة القرار (DT) وخوارزمية التعبئة باستخدام مجموعة بيانات الحليب مأخوذة من مستودع البيانات Kaggle . تتكون هذه البيانات من 1059 عينة وسبع ميزات. تم تدريب النماذج المقترحة واختبارها بهدف إيجاد النموذج الأفضل والأكثر دقة للكشف عن جودة الحليب وتم تقييمها باستخدام مقاييس التقييم: الدقة، الضبط، الاستدعاء، f1_score ومصفوفة الارتباك. ووفقاً لنتائج التقييم، تفوقت النماذج الثلاثة: SVM، وKNN، و DT على خوارزمية التعبئة، حيث حصلت على أعلى مستوى لجميع المقاييس بنسبة 100%. وكانت خوارزمية SVM هي الأكثر كفاءة لأن وقت تنفيذها كان 0.146 ثانية، وهو أقل من النماذج الأخرى.