Comparison of Machine Learning Algorithms' Performance in Network Traffic Classification

Dr. Jamal Khalifa [*]
Dr. Mohannad Issa **
Faifaa Micaiel***

(Received 6 / 6 / 2024. Accepted 19 / 8 / 2024)

□ ABSTRACT □

Maintaining network availability and improving performance is a primary objective of network management. With the massive growth in network size and traffic load, this task has become increasingly complex. One important area of research is network traffic classification, which offers significant benefits such as reducing traffic congestion and enhancing network management.

This study explores the application of various machine learning algorithms for network traffic classification into large flows and small flows. We implemented and evaluated multiple classifiers on real network traffic "Darknet Dataset", including Support Vector Machine (SVM), Random Forest (RF), k-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Gradient Boosting (GB), and Multilayer Perceptron (MLP). Each classifier was trained and tested on real network traffic data.

Our results indicate that the Random Forest (RF), Decision Tree (DT), Nearest Neighbor (KNN) classifier, Gradient Boosting (GB), and Multilayer Neural Network (MLP) classifiers achieved the highest accuracy, in classification. These results underscore the potential of the models. Machine learning helps in classifying network loads effectively.

Keywords: Software Defined Networking, Machine Learning, Network Classification, Network Management, Quality Of Service

Copyright **:**Tishreen University journal-Syria, The authors retain the copyright under a CC BY-NC-SA 04

[*]**Professor- Department of Communication and Electronics- Faculty of Mechanical and Electrical Engineering- Tishreen University- Lattakia- Syria.**
****Doctor- Ministry of Oil and Mineral Resources, Syria.**
*****Postgraduate Student (PhD)- Department of Communication and Electronics- Faculty of Mechanical and Electrical Engineering- Tishreen University- Lattakia- Syria**. faifaamicaiel@gmail.com

مقارنة أداء خوارزميات التعلم الآلي في تصنيف تدفقات الشبكة

د. جمال خليفة*

د. مهند عيسى**

فيفاء مكائيل***

(تاريخ الإيداع 6 / 6 / 2024. قُبِل للنشر في 19/ 8 / 2024)

□ ملخّص □

يعد الحفاظ على توفر الشبكة وتحسين الأداء هدفا أساساً لإدارة الشبكة. ومع النمو الهائل في حجم الشبكة والحمل، أصبحت هذه المهمة معقدة بشكل متزايد، ويعد تصنيف أحمال الشبكة أحد مجالات البحث الهامة ، والذي يقدم فوائد كبيرة مثل تقليل ازدحام حركة الأحمال، وتعزيز إدارة الشبكة.

تستكشف هذه الدراسة تطبيق خوارزميات التعلم الآلي المختلفة لتصنيف تدفقات الشبكة الى تدفقات ضخمة وتدفقات صغيرة. قمنا بتنفيذ وتقييم مصنفات متعددة على حمل شبكة فعلي "Darknet Dataset"، بهدف التصنيف، بما في ذلك آلة متجه الدعم (SVM)، والغابة العشوائية (RF)، ومصنف الجار الأقرب (KNN)، والانحدار اللوجستي (LR)، وشجرة القرار (DT)، وتعزيز التدرج (GB)، والشبكة العصبونية متعدد الطبقات (MLP). تم تدريب كل مصنف واختباره على بيانات حمل شبكة حقيقية.

تشير النتائج التي توصلنا إليها إلى أن مصنفات الغابة العشوائية (RF) وشجرة القرار (DT) ومصنف الجار الأقرب (KNN) وتعزيز التدرج (GB)، والشبكة العصبونية متعدد الطبقات (MLP) حققت أعلى دقة، في التصنيف، تؤكد هذه النتائج على إمكانات نماذج التعلم الآلي هذه في تصنيف أحمال الشبكة بشكل فعال.

الكلمات المفتاحية: الشبكات المعرفة بالبرمجيات، التعلم الآلي، تصنيف الشبكات، إدارة الشبكات، جودة الخدمة.

حقوق النشر : مجلة جامعة تشرين- سورية، يحتفظ المؤلفون بحقوق النشر بموجب الترخيص CC BY-NC-SA 04

* **أستاذ- قسم هندسة الاتصالات والإلكترونيات- كلية الهندسة الميكانيكية والكهربائية- جامعة تشرين - اللاذقية - سورية.**

** **دكتور مهندس- وزارة النفط - سورية.**

*** **طالبة دراسات عليا (دكتوراه)- قسم هندسة الاتصالات والإلكترونيات- كلية الهندسة الميكانيكية والكهربائية- جامعة تشرين- اللاذقية- سورية.faifaamicaiel@gmail.com**

مقدمة:

إن النمو السريع لعدد الاتصالات والأجهزة في جميع أنحاء العالم، الذي يتجاوز سرعة نمو عدد السكان، وزيادة الاعتماد على الإنترنت الذي أصبح جزءاً أساسياً من الحياة اليومية للأفراد والشركات، قد أدى إلى زيادة كبيرة في حركة حمل الشبكة [1]. بالإضافة إلى اعتماد أجهزة جديدة ذات قدرات كبيرة، مثل الهواتف الذكية، أجهزة التلفزيون الذكية، وحدات تحكم ألعاب الفيديو، وأجهزة المراقبة، جنباً إلى جنب مع انتشار الاتصالات من آلة إلى آلة (M2M) وتطوير خدمات وتطبيقات جديدة، أدى كل ذلك إلى تغيير كبير في أنماط تدفق حركة حمل الشبكة وتتويع حركة البيانات في الشبكات. كما أن ضرورة عدم حصول تأخير في الإرسال، مثل الصوت عبر الإنترنت (VOIP) والمؤتمرات عبر الفيديو (Video Conference)، مقارنة بتطبيقات أخرى مثل تصفح الويب، يتطلب التمييز بين أحمال الشبكة بناءً على عدة معايير مثل عدد الحزم، وقت الوصول، حجم الحزم المرسلة، وغيرها [2].

هذا الوضع وضع تصنيف حركة حمل الشبكة في الوقت الفعلي ضمن مجال مشكلات البيانات الضخمة. ويعد تصنيف حركة حمل الشبكة بسرعة ودقة عالية تحديًا كبيرًا في عصر البيانات الضخمة [3].

لقد حققت خوارزميات التعلم الآلي في السنوات الأخيرة تقدمًا كبيراً، وذلك بفضل توفر كميات هائلة من البيانات وزيادة القدرة الحسابية، مما منحها القدرة على معالجة هذه البيانات، وتحليلها، واكتشاف الأنماط المخفية، واستخلاص تنبؤات قيّمة منها [4]. وأصبحت هذه الخوارزميات فعالة في حل المشكلات المعقدة التي كانت تعتبر ذات يوم تتجاوز قدرات الآلات. وللاستفادة من تقنيات التعلم الآلي والذكاء الاصطناعي، سنتعرف في هذه الدراسة على خوارزميات التعلم الآلي الأكثر شيوعًا في مجال التصنيف.

أهمية البحث وأهدافه:

يُعد التعرف على حركة حمل الشبكة وتصنيفها من الأمور الضرورية لفهم سلوك الشبكة وتحسين تخصيص الموارد. يهدف هذا البحث إلى استخدام تقنيات التعلم الآلي في الشبكات من خلال تدريب عدة نماذج واختبارها على أحمال حقيقية في الشبكة، وتقييم الأداء من أجل العثور على النموذج الذي يناسب التوقعات بشكل أفضل واكتشاف نوع حمل الشبكة.

طرائق البحث ومواده:

اعتمدنا على Darkent Dataset و عدلناها لتكون مناسبة لتصنيف التدفقات و اعتمدنا مجموعة من الخصائص مثل عدد الحزم و حجمها و المدة الزمنية لكل تدفق، وقمنا بتنفيذ النموذج (كود برمجي) باستخدام لغة البرمجة python و تطبيقه على منصة colab (منصة تعمل على أنظمة تشغيل مختلفة، خاصة بلغة بايثون، تؤمن الكثير من التسهيلات لدعم هذه اللغة وتسهيل التعامل معها) و النموذج عبارة عن مصنفات دربت على مجموعة البيانات ثم قارنا هذه المصنفات و طبقنا أفضلها لتصنيف تدفقات الشبكة الفعلية.

الدراسات المرجعية:

اقترح الباحثون في الدراسة [5]، التنبؤ بحجم التدفق عبر الإنترنت لتحسين توجيه الشبكة باستخدام العديد من تقنيات التعلم الآلي، بما في ذلك الشبكات العصبية. يعتمد التنبؤ على معلومات معينة تم جمعها من الحزم الأولى، بما في

ذلك عنوان IP المصدر وعنوان IP الوجهة ومنفذ المصدر ومنفذ الوجهة والبروتوكول وحجم الحزم الثلاث الأولى. بعد التنبؤ بحجم التدفق، يتم توجيه تدفقات الأفيال عبر الطرق الأقل ازدحامًا.

قام الباحثون في الدراسة [6]، بتصنيف الشبكة من خلال التمييز بين خصائص حركة المرور المختلفة مثل عدد الحزم وحجمها، وقت وصول الحزمة، نسبة إرسال واستقبال الحزمة. وقارنوا بين طرق التصنيف التقليدية وثلاثة نماذج من خوارزميات التعلم الآلي.

وفي الدراسة [7]، قام الباحثون بتطبيق ثلاثة نماذج مختلفة من التعلم الخاضع للإشراف، وهي Support Vector Machine (SVM)، Naïve Bayes (NB)، و Nearest Centroid (NC) على تصنيف أحمال الشبكة بناءً على التطبيقات الموجودة في منصة الشبكة المعرفة بالبرمجيات. الدقة التي تم الحصول عليها لـ SVM هي 92.3%، ولـ NB هي 96.79%، و NC هي 91.02%.

اقترح الباحثون في الدراسة [8] طريقة لتصنيف حركة المرور في بيئة شبكة معرفة برمجيًا باستخدام أداة التشفير التلقائي المتغير (VAE). تقوم الطريقة المقترحة بتدريب VAE باستخدام ستة ميزات إحصائية وتستخرج توزيعات الميزات الكامنة للتدفقات في كل فئة خدمة. علاوة على ذلك، تصنف حركة الاستعلام من خلال مقارنة توزيعات الميزات الكامنة لحركة الاستعلام مع التوزيعات المستفادة لفئات الخدمة. بالنسبة للتجربة، تم جمع السمات الإحصائية لتدفقات الشبكة من خدمات الإنترنت المحلية والخارجية في العالم الحقيقي للتدريب والاختبار. وفقًا للنتائج التجريبية، تتمتع الطريقة المقترحة بمتوسط دقة يبلغ 89%.

اقترح المؤلفون في الدراسة [9] طريقة لإنشاء مسارات تدريجية للتدفقات الكبيرة والصغيرة باستخدام أداة خوارزمية خاضعة للإشراف لتصنيف البيانات المتشابهة، وأثبتت النتائج قدرة ESCA على التحديد الدقيق مع عينات أقل بكثير تم جمعها في فترة كشف قصيرة.

اقترح الباحثون في الدراسة [10] تصنيفًا قائماً على التطبيقات لحركة المرور مع وبدون اتصال بالإنترنت،استناداً إلى آليات التعلم العميق، عبر اختبار ثلاثة نماذج للتعلم العميق، وهي Multilayer Perceptron (MLP)، و Stacked Autoencoder (SAE)، و Convolutional Neural Network (CNN). باستخدام أداة TCP replay، حققت نتائج التدريب دون اتصال بالإنترنت دقة تزيد عن 93% في تحديد سبعة تطبيقات شائعة لجميع النماذج الثلاثة. علاوة على ذلك، تحققت دقة بنسبة 87% للتنبؤ بالاختبار عبر الإنترنت القائم على التطبيق.

في الدراسة [11]، قام الباحثون بتصميم وتقييم مصنف لشبكة SDN باستخدام ثلاث تقنيات للتعلم الآلي، وتم التصنيف حسب البروتوكولات TCP و UDP. ثم قدموا تقنية قادرة على الجمع بين نماذج تعلم الآلة المختلفة وتسريع الدقة بما يصل إلى 3.5%.

في الدراسة [12]، عمل الباحثون على بناء مصنف اعتماداً على عدة بروتوكولات، بما في ذلك WWW، و DNS، و FTP، و ICMP، و P2P، والصوت عبر بروتوكول الإنترنت. مع معدل دقة يصل إلى 99.8%، برزت شجرة القرار في هذه الدراسة.

في الدراسة [13]، تم اقتراح حل لاكتشاف تدفق الفيل عن طريق معالجة الحزمة الأولى من التدفق. يتم تحقيق الاكتشاف من خلال تدريب خوارزميات التعلم الآلي، والحصول على نتائج أداء أعلى مقارنة باستخدام نموذج أخذ عينات الحزمة.

في الدراسة [14]، اقترح الباحثون تصنيف حركة المرور على أساس التعلم الآلي باستخدام خوارزميتين للتعلم الآلي: الانحدار اللوجستي (خاضع للإشراف) وتجميع وسائل K (غير خاضع للإشراف) للتصنيف باستخدام مولد حركة مرور الإنترنت الموزع (D-ITG). بلغت الدقة 78.89٪، مما يشير إلى أن نماذج الانحدار اللوجستي تعمل بشكل أفضل عند مقارنتها بتجميع الوسائل K.

1 الأحمال ضمن الشبكة:

عند مراقبة الشبكة، نجد أن هناك مجموعة واسعة من التطبيقات ذات المتطلبات والقيود المختلفة على موارد الشبكة. نظرًا لأن التدفق هو مجموعة من الحزم التي لها نفس عناوين IP المصدر والوجهة، ونفس أرقام منفذ المصدر والوجهة، ونفس نوع البروتوكول، فإن معظم حركة أحمال الشبكة تتكون من عدد صغير نسبيًا من التدفقات الكبيرة، تسمى هذه التدفقات بتدفقات الفيل (Elephant Flow). أما التدفقات المتبقية، وهي كبيرة العدد ولكنها تحمل حركة أحمال قليلة جدًا، فتسمى تدفقات الفئران (Mice Flow) [15]. تحمل تدفقات الفيل معظم حركة المرور بالبايت، بينما تكون معظم التدفقات هي تدفقات الفئران. ومن الشائع أن نلاحظ أن 95% من التدفقات هي تدفقات الفئران، لكن تدفقات الفيل تشغل أكثر من 95% من الحجم الإجمالي. على سبيل المثال، يعتبر تصفح الويب تدفق فئران، في حين أن عمليات نقل الملفات غالبًا ما تكون تدفقات فيل. يُعرف هذا السلوك بظاهرة الفيلة والفئران، وفي الإحصائيات، تسمى هذه الظاهرة أيضًا "التفاوت في العدد الشامل" (mass-count disparity). يمكن أن يؤدي هذا التفاوت إلى ضعف أداء الشبكة، حيث يشغل تدفق الفيل بشكل كبير النطاق الترددي، بينما يحتاج تدفق الفئران إلى تأخير منخفض. لذلك، من الضروري تنظيم حركة هذه التدفقات ضمن الشبكة، مما يتطلب تصنيفها في البداية. [16]

2 طرق تصنيف الأحمال في الشبكات:[17]

2.1 الفحص العميق للحزم(DPI) :

يتضمن DPI فحص محتوى كل حزمة لتحديد التطبيق أو البروتوكول الذي ينشئ حركة المرور. تعد هذه الطريقة فعالة لتصنيف أنواع مختلفة من حركة المرور، مثل تصفح الويب، وتدفق الفيديو، وVoIP، ونقل الملفات.

2.2التصنيف القائم على التدفق:

يقوم التصنيف القائم على التدفق بتصنيف حركة المرور بناءً على تدفق الحزم بين عناوين المصدر والوجهة والمنافذ والبروتوكولات. تعتبر هذه الطريقة مناسبة لتحديد أنواع معينة من تدفقات حركة المرور وتحديد أولوياتها، مثل الاتصال في الوقت الفعلي أو نقل البيانات المجمعة.

2.3التصنيف القائم على البيانات الوصفية:

يمكن استخدام استخراج البيانات الوصفية من رؤوس الحزم أو الحمولات لتصنيف حركة المرور بناءً على سمات محددة، مثل نوع التطبيق أو هوية المستخدم أو متطلبات الأمان.

2.4التصنيف القائم على التعلم الآلي:

يمكن استخدام خوارزميات التعلم الآلي لتحليل أنماط حركة المرور وسلوكها لتصنيف حركة المرور تلقائياً إلى فئات مختلفة. يمكن لهذا الأسلوب التكيف مع ظروف الشبكة المتغيرة وتحديد أنواع جديدة من حركة المرور دون تكوين يدوي.

3 خوارزميات التعلم الآلي (Machine Learning) :

التعلم الآلي هو طريقة لتحليل البيانات تتعلم من البيانات المدخلة وتتخذ القرارات بناءً على المعلومات التي تم جمعها. تتضمن العملية عموماً المعالجة المسبقة، التدريب، ومرحلة الاختبار. تتضمن المعالجة المسبقة إجراءات مثل إعداد

البيانات، تصفيتها، وإسنادها وضبطها بشكل محدد. بمجرد معالجة البيانات مسبقاً، يتم استخدام أساليب التعلم الآلي لتدريب النماذج. بعد ذلك، يتخذ النظام قرارات بناءً على المدخلات الواردة من مرحلة التدريب.[18] إن الهدف الرئيسي من هذا المجال تطوير أنظمة ذكية قادرة على تحسين نفسها تدريجياً وبشكل آلي عند التعرف على البيانات والمدخلات الجديدة بدون الحاجة للإعداد اليدوي. يتم استخدام التعلم الآلي في عدة مجالات مثل التصنيف، التنبؤ، والتحكم الآلي ويمكن تبسيط عملية التعلم الآلي بالنموذج التالي:

الشكل(1) نموذج مبسط لعمل التعلم الآلي

يختلف التعلم الآلي عن البرمجة التقليدية من حيث طبيعة الخرج. في حالة برمجة النظام الحاسوبي بخوارزمية ما، يتم إدخال قيم معينة لتعطي قيم مقابلة على الخرج. أما في حالة خوارزميات التعلم الآلي، فإن خرج خوارزمية التعلم الآلي يكون نموذجاً (خوارزمية) قادراً على استقبال دخل معين (x) وإعطاء خرج مقابل لهذا الدخل.$y = f(x)$

يمكن تقسيم خوارزميات التعلم الآلي ML (Machine Learning) إلى ثلاث فئات:

التعلم الآلي الخاضع للإشراف:(Supervised Machine Learning) :

في هذا النوع، يتم بناء نماذج تصنف الحالات الجديدة إلى فئات معروفة. هناك مرحلتان في التعلم الخاضع للإشراف: مرحلة التدريب، التي تبني نموذج التصنيف من خلال تحليل مجموعة بيانات التدريب، ومرحلة الاختبار (أو التصنيف)، التي تستخدم النموذج المدمج في مرحلة التدريب لتصنيف الحالات الجديدة. يتم تكييف النظام بطريقة تمكنه من التنبؤ بالمخرجات الصحيحة للمدخلات الجديدة بناءً على ما تعلمه من بيانات التدريب.

التعلم الآلي غير الخاضع للإشراف:(Unsupervised Machine Learning) :

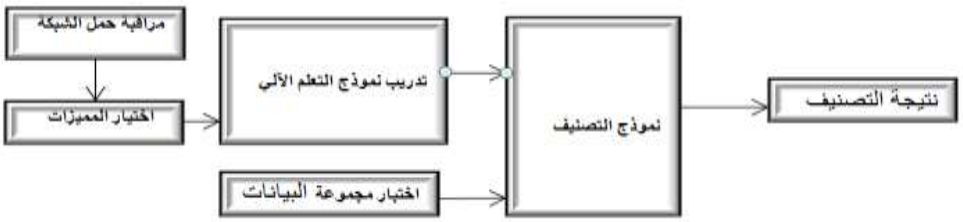
في هذا النوع، يتم تدريب النماذج بدون وجود بيانات معروفة مسبقاً. بدلاً من ذلك، يقوم النظام بالبحث عن الأنماط والهيكليات داخل البيانات المدخلة.

التعلم الآلي شبه الخاضع للإشراف:(Semi-Supervised Machine Learning) :

هذا النوع يجمع بين التعلم الخاضع للإشراف والتعلم غير الخاضع للإشراف. يتم استخدام كمية صغيرة من البيانات المعلمة مع كمية كبيرة من البيانات غير المعلمة لتحسين دقة النموذج.

ستكون هو خوارزميات التعلم الآلي الخاضع للإشراف مجال بحثنا هذا، حيث أن هذا النوع يقوم ببناء نماذج تصنف الحالات الجديدة إلى فئات معروفة. في مرحلة التدريب، يتم بناء نموذج التصنيف من خلال تحليل مجموعة بيانات التدريب. في مرحلة الاختبار (أو التصنيف)، يتم استخدام النموذج المدمج في مرحلة التدريب لتصنيف الحالات الجديدة، مما يمكن النظام من التنبؤ بالمخرجات الصحيحة للمدخلات الجديدة بناءً على ما تعلمه من بيانات التدريب.

يتضمن التعلم الآلي الخاضع للإشراف البيانات المصنفة، حيث تحتوي على متغيرات الإدخال، أو الميزات (features)، ومتغير الإخراج المقابل لها، أو السمة (label). تستخدم الخوارزميات هذه البيانات المصنفة لتعلم رسم المخططات بين متغيرات الإدخال والإخراج، والتنبؤ بالبيانات الجديدة غير المرئية. يوضح الشكل (2) المخطط الصندوقي لخوارزمية التعلم الآلي تحت الإشراف.[19]



الشكل (2) مخطط صندوقي لخوارزمية التعلم الآلي تحت الإشراف

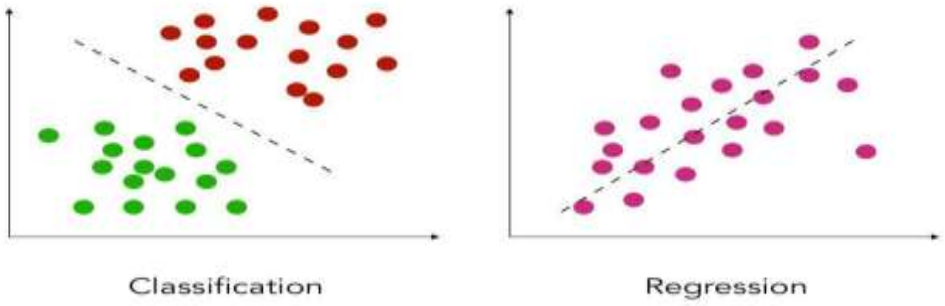
ويستخدم التّعلم الخاضع للإشراف بشكل أساسي في:

آ- التصنيف (Classification):

يتم في التصنيف توقع قيمة أو متغير فئوي حيث يكون فضاء الاحتمالات مقسماً إلى مجموعة محدودة من الصفوف. يُسمى التصنيف ثنائياً (Binary Classification) في حالة وجود صفين، وفي بقية الحالات يُسمى التصنيف متعدد الصفوف (Multi-Class Classification).

ب- الانحدار (Regression):

الهدف في هذه الحالة هو توقع قيمة مستمرة تقابل معطيات الدخل. مثلاً، يمكن توقع سعر منزل في منطقة معينة وبمساحة معينة. يختلف نوع الانحدار باختلاف شكل النموذج الناتج عن خوارزمية التعلم الآلي والذي يعبر عن العلاقة بين الدخل والخرج. يكون الانحدار خطيًا (Linear Regression) إذا كان النموذج الناتج يتمثل بعلاقة خطية بين الخرج والدخل، بالإضافة إلى أنواع أخرى مثل الانحدار المتعدد الحدود (Polynomial Regression) والانحدار اللوجستي (Logistic Regression).ويبين الشكل(3) الفرق بين الانحدار الخطي والتصنيف الثنائي:



الشكل(3) الفرق بين التصنيف و الانحدار

نستعرض فيما يلي بعض خوارزميات التعلم الآلي الخاضعة للإشراف:

3.1 آلة متجه الدعم (SVM) Support Vector Machine:

تستخدم خوارزمية التعلم الخاضع للإشراف البيانات المصنفة لتدريب النموذج حيث نرسم كل عنصر بيانات كنقطة في الفضاء ذي البعد (m) حيث m هو عدد الميزات التي لدينا مع قيمة كل ميزة هي قيمة إحداثيات معينة. [20],[21]

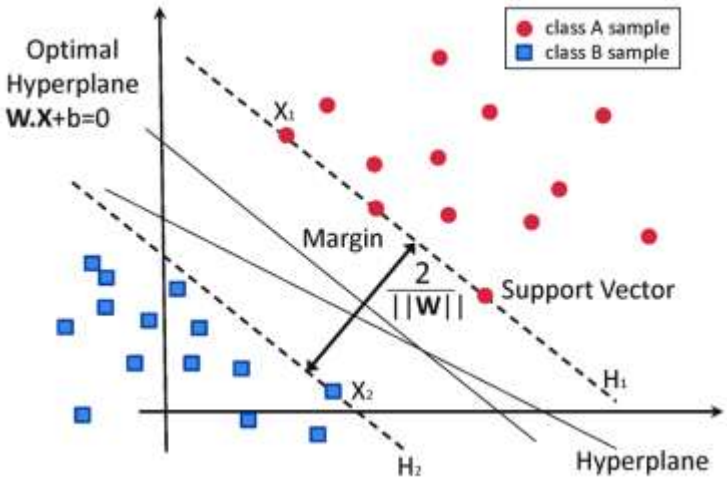
يتم إنشاء حدود القرار بين البيانات المسماة بواسطة نموذج SVM، ويعمل مصنف SVM الخطي البسيط عن طريق انشاء خط مستقيم (فاصل) بين فئتين. أي أن جميع نقاط البيانات على جانب واحد من الخط تمثل فئة وأن نقاط البيانات على الجانب الآخر من الخط تمثل فئة مختلفة، أي أنه يمكن تحديد عدد لا حصر له من الخطوط. ويمكن تمثيل الخط الفاصل للفصل في فضاء العينة بالمعادلة الخطية التالية:

$$w^T + b = 0 \quad (1)$$

حيث: $w=\{w_1, w_2, \ldots, w_d\}$ هو متجه طبيعي يتحكم في اتجاه الخط الفاصل و b هو الانحياز الذي يتحكم في المسافة بين الخط الفاصل والأصل. يحدد المتجه الطبيعي w والتحيز b الخط الفاصل للفصل يمكن كتابة المسافة من أي نقطة x في مساحة العينة إلى الخط الفاصل على النحو التالي:

$$r = \frac{|w^T + b|}{\|w\|} \quad (1)$$

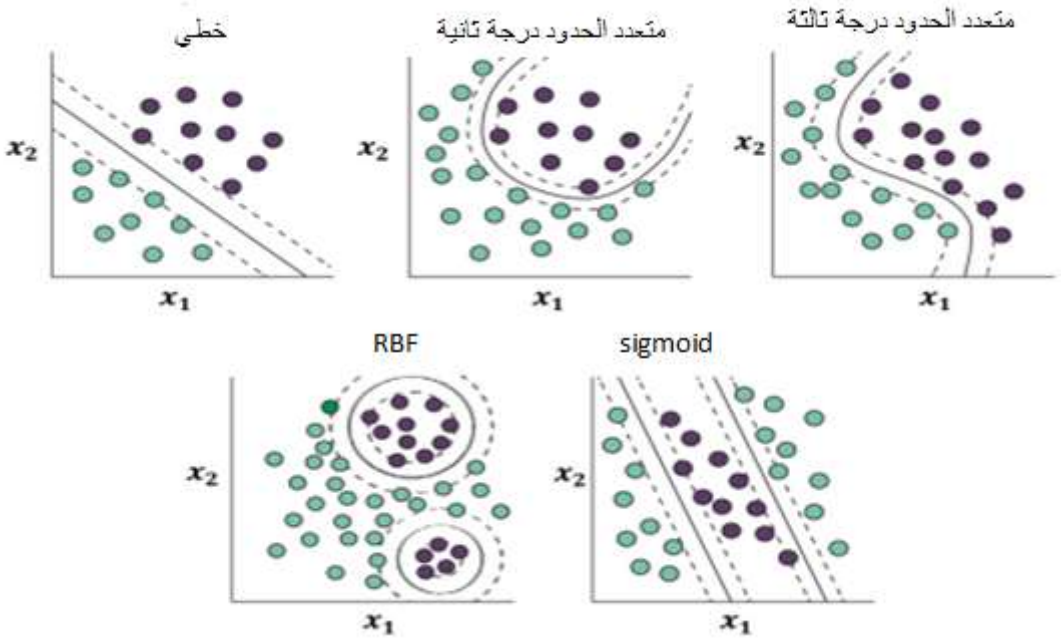
ويبين الشكل(4) مخطط التصنيف بخوارزمية svm



الشكل(4)مخطط يوضح التصنيف بخوارزمية svm

ويمكن أن يوجد عدة فواصل بين صفوف المعطيات لكن أفضلها هو الذي يعطي أكبر هامش (المسافة بين الفاصل وأقرب نقاط الصف إليه.

تصنف هذه الخوارزميّة ضمن الخوارزميّات التي تستخدم توابع النّواة (Kernel methods) حيث يتم استخدام عدة توابع مختلفة لتحسين حدود القرار. وأكثر هذه التوابع استخداماً هي التابع السيني، ومتعددة الحدود، وRBF، والخطية. تتوفر بيانات العالم الحقيقي أحادية البعد أو متعددة الأبعاد، في بعض الأحيان، من الممكن أيضاً فصل مجموعات البيانات هذه خطياً. قد يعمل التابع الخطي مع مجموعات البيانات القابلة للفصل خطياً.



الشكل(5) أمثلة على الدوال

3.2 شجرة القرار (Decision Tree DT):

شجرة القرار هي طريقة تعلم خاضعة للإشراف تتميز بقدرتها على التعامل مع المتغيرات الفئوية (categorical) والمستمرة .(continuous) تعتمد هذه الطريقة على بناء نموذج شجري يستند إلى القرارات المتخذة من خلال فحص قيم السمات.

تقوم شجرة القرار بتحليل البيانات وتقسيمها إلى فئات (أو تصنيفات) باستخدام سلسلة من القرارات المبنية على القيم الموجودة في مجموعة من المتغيرات. يتم تمثيل هذه القرارات في شكل هيكل شجري، حيث يتم تقسيم البيانات إلى فروع (Branches)وتصنيفات (Leaves) بناءً على معايير محددة [22].

تُستخدم هذه الخوارزمية في العديد من التطبيقات مثل التصنيف والتنبؤ، حيث تُستخدم لتصنيف العناصر إلى فئات مختلفة بناءً على متغيرات معينة. كما يمكن استخدامها في تحليل البيانات واكتشاف العلاقات السببية بين المتغيرات. تعمل شجرة القرار عن طريق تحليل البيانات واستخراج العلاقات بين المتغيرات المختلفة للوصول إلى قرار تصنيف دقيق، وذلك بالخطوات التالية:

- اختيار المتغيرات: نضع في الجذر أفضل سمة لمجموعة البيانات.
- تقسيم مجموعة التدريب: يتم تقسيم مجموعة التدريب إلى مجموعات فرعية تحتوي كل منها على بيانات بنفس قيمة السمة.
- تقسيم البيانات: يتم تقسيم البيانات إلى مجموعة صغيرة من الأقسام (subsets) باستخدام المتغيرات المختارة، حيث يتم تحليل القيم المختلفة لكل متغير وتقسيم البيانات إلى فروع مختلفة بناءً على قيمة المتغيرات.
- حساب الاختلافات: يتم حساب الاختلافات بين فروع البيانات المختلفة، ويتم اختيار الفرع الذي يعطي أفضل تمثيل للبيانات.

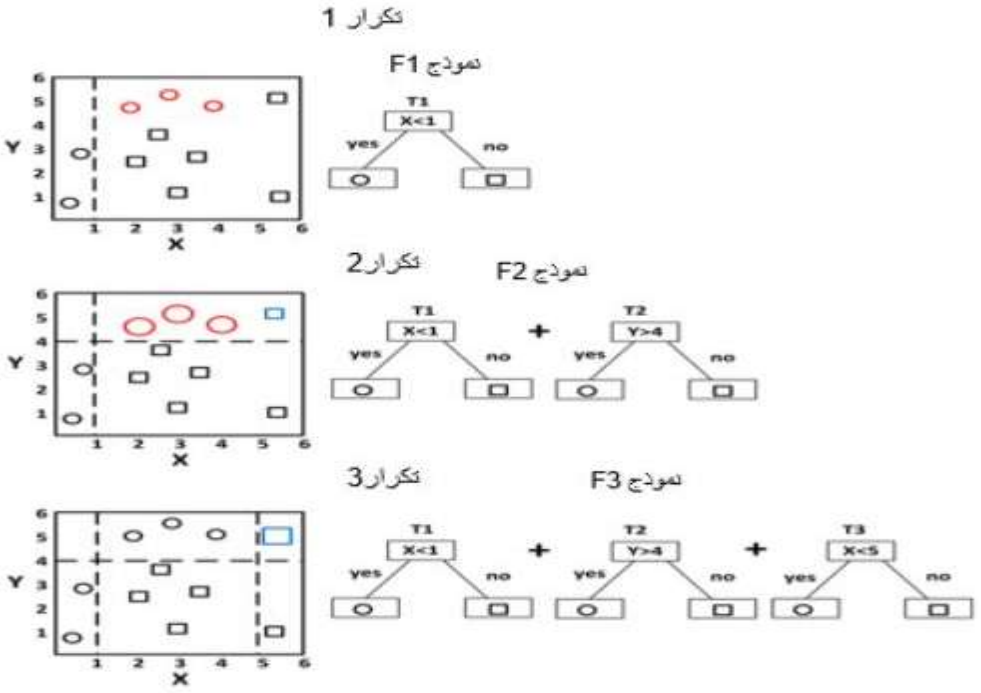
- إنشاء الشجرة: يتم إنشاء شجرة القرار عن طريق تكرار الخطوات السابقة حتى يتم تقسيم البيانات إلى مجموعات فرعية مختلفة، ويتم تمثيل ذلك في شكل شجرة تتكون من فروع وأوراق.
- التصنيف: يتم استخدام شجرة القرار المنشأة لتصنيف البيانات الجديدة، حيث يتم اتباع الفروع المناسبة في الشجرة لتحديد التصنيف الذي تنتمي إليه البيانات.

بهذه الطريقة، تساعد شجرة القرار في تحليل البيانات بفعالية واستخلاص الأنماط والعلاقات بين المتغيرات لتحقيق تصنيف دقيق وموثوق.

3.3 تعزيز التدرج Gradient Boosting (GB):

هي خوارزمية تعزيز تُستخدم عندما نتعامل مع الكثير من البيانات لعمل تنبؤ بقوة تنبؤ عالية. التعزيز Boosting هو في الواقع مجموعة من خوارزميات التعلم التي تجمع بين توقع العديد من التنبؤات الأساسية من أجل تحسين المتانة على متنبئ واحد. فهو يجمع بين عدة متنبئين ضعيفين أو متوسطين لبناء متنبئ قوي.

تعمل هذه الخوارزميّة من خلال إضافة نماذج التنبّؤ بشكل متتابع، وكلّ نموذج يصحّح النموذج السّابق وتحاول أن تقوم بملاءمة نموذج التنبّؤ الجديد مع الأخطاء المتبقية من النّموذج الذي سبقه (بمعنًى آخر يركز تعزيز التدرج على الفرق بين التنبّؤ والحقيقة الأساسيّة) [23]



الشكل(6) مخطط عمل الخوارزمية (GB)

3.4 الغابة العشوائية Random Forest(RF) :

هي تقنية تعليمية خاضعة للإشراف يمكنها التعامل مع مشكلات التصنيف والانحدار. وهي عبارة عن مزيج من أساليب شجرة القرار المختلفة، وكلما زاد عدد الأشجار المضمنة، كلما كان النموذج أكثر دقة. وهي تعمل بشكل مشابه لشجرة القرار التي تعتمد على المعلومات المكتسبة. كل شجرة قرار في التصنيف سوف تصنف نفس المشكلة، وسيتم تحديد النتيجة النهائية من خلال الأخذ في الاعتبار غالبية النتائج و يتم اختيار عشوائي لمجموعة فرعية من ميزات البيانات و

تعد قدرة هذا النموذج على التعامل مع مجموعات البيانات الكبيرة والتعامل مع القيم المفقودة هي الميزة الأكثر أهمية له.[24][25]

الشكل(7) مخطط عمل الخوارزمية (RF)

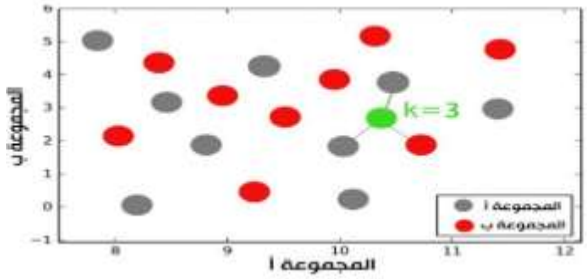
3.5 الجار الأقرب Nearest Neighbor (KNN) :

هي طريقة تعلم خاضعة للإشراف تعتمد على المثيل. يشير البارامتر k في نموذج KNN إلى عدد الجيران الذين يجب أخذهم في الاعتبار من أجل التصنيف. يعتمد مبدأ عمل هذه الخوارزمية على حساب المسافة بين النقاط،

لتحديد أقرب جار لنقطة بيانات، يجب أن نستخدم معيار التشابه أو الاختلاف بين نقاط البيانات، و يتم تخزين جميع عينات التدريب في مساحة نمط p هناك العديد من معايير التشابه أو الاختلاف، منها المسافة الإقليدية (هي مقياس للاختلاف بين نقطتي بيانات x_i, x_j) يتم تحديد المسافة الإقليدية على النحو التالي:

$$d(xi, xj) = \sqrt{\sum_{i=1}^{p} (x_{i,1} - x_{j,1})^2}, i \qquad i \neq j \quad (3)$$

حيث كلما قلت المسافة بين نقطتين زاد احتمال انتماء النقطتين بعضهما لبعض و من هنا جاءت تسمية هذه الخوارزمية، أي إذا افترضنا أن k = 3 فإن الخوارزمية سوف تقوم بقياس المسافة بين النقطة المستهدفة و أقرب ثلاث نقاط إليها فإذا كانت أقرب نقطتين تنتمي الى المجموعة (أ) والنقطة الثالثة وحدها تنتمي الى لمجموعة (ب) فإن النقطة المستهدفة سيتم تصنيفها على أساس أنها تنتمي الى المجموعة(أ). كما في الشكل:[26][27]

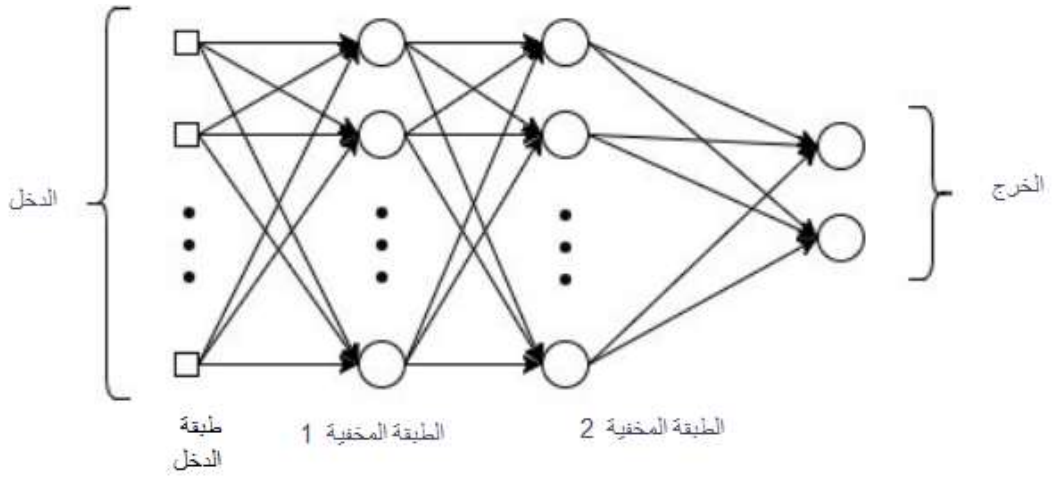


الشكل(8) طريقة عمل الخوارزمية (KNN)

3.6 الشبكة العصبونية MLP Multilayer perceptron :

هي نوع من الشبكات العصبية التي تستخدم على نطاق واسع في التعلم الآلي و الذكاء الصنعي و هي شبكة عصبية متجهة إلى الأمام ، أي أن المعلومات تتدفق في اتجاه واحد من طبقة الإدخال إلى طبقة الإخراج . تتكون بنية ال

MLPمن ثلاث طبقات أو أكثر: طبقة الإدخالInput layer ، طبقة مخفية واحدة أو أكثرHidden layer ، و طبقة الإخراج Output layer نقوم بتغذية طبقة الإدخال ببيانات الإدخال الخاصة بنا ونحصل على النتائج من طبقة الإخراج. يمكننا زيادة عدد الطبقة المخفية بقدر ما نشاء، لجعل النموذج أكثر دقة و تعقيدا وفقاً للمهمة التي نريد انجازها.**[29] [28]**



الشكل(9)

3.7 الانحدار اللوجستي Logistic Regression (LR):

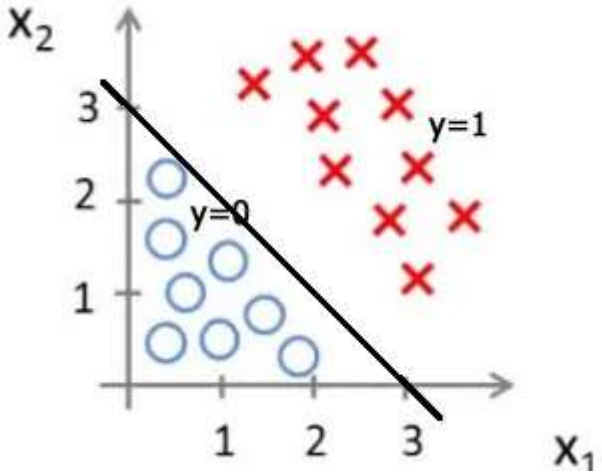
هو مصنف ML بسيط خاضع للإشراف، ويرسم تنبؤات لاحتمالات الأحداث باستخدام دالة تكلفة هي دالة سينية. بناءً على القيمة المختارة للحد يتم استخدام مخرجات الوظيفة، التي يتراوح نطاقها من 0 إلى 1، لتعيين الملاحظات لفئات منفصلة و امكانية تعميمه الى فئات متعددة.

و هو خوارزمية إحصائية تستخدم لنمذجة العلاقة بين متغيرين و تفترض أن هناك علاقة خطية بين المتغير التابع (المتغير الذي يتم توقعه) و متغير واحد أو أكثر من المتغيرات المستقلة (تلك المستخدمة لإجراء التنبؤ) .الهدف منه هو العثور على الخط الأنسب الذي يصف العلاقة بين المتغيرات.[30]

يمكن تمثيل الانحدار البسيط بالمعادلة ادناه :

$$Y = W_1 X_1 + W_2 X_2 + b \quad (4)$$

حيث: y : متغير تابع w :الميل X : متغير مستقل b : ثابت.



الشكل (10) مخطط يوضح التصنيف باستخدام (LR)

4 مقاييس الأداء الأكثر شيوعًا المستخدمة في مسائل التصنيف:[32][31]

عندما تكون مهمة نماذج التعلم الآلي هي التصنيف يتم استخدام **confusion matrix** (و تسمى في بعض المراجع مصفوفة الالتباس أو الارتباك) لتقييم أداء نماذج التصنيف، هذه المصفوفة تقوم بحساب TP و FP و FN و TN وتعطينا مصفوفة تعبر عنهم بأبعاد (n_classes, n_classes) تلخص مدى نجاح نموذج التصنيف و توضح عدد التنبؤات الصحيحة و الخاطئة التي قدمها النموذج مقارنة بالنتائج الفعلية و يبين الجدول التالي هذه المصفوفة :

		Predicted Class	
		0	*1*
Actual Class	*0*	*True Negative (TN)*	*False Negative (FN)*
	1	*False Positive (FP)*	*True Positive (TP)*

الجدول(1) مصفوف الارتباك

يحتوي الجدول على عدد من الصفوف والأعمدة يساوي عدد الفئات الموجودة في النموذج.

يتم وضع عينات فئة معينة محددة بشكل صحيح بواسطة المصنف في الإيجابية الحقيقية (TP)، يتم وضع عينات الفئات الأخرى المحددة بدقة في المؤشرات السلبية الحقيقية (TN) وبالمثل، يتم وضع العينات التي تنبأ بها المصنف بشكل غير صحيح في التصنيف الإيجابي الكاذب (FP) والخطأ في التصنيف السلبي الكاذب (FN) .

حيث أن:

الإيجابيات الحقيقية (TP): الفئة الفعلية إيجابية، والفئة المتوقعة إيجابية.

السلبيات الحقيقية (TN): الفئة الفعلية سلبية، والفئة المتوقعة سلبية.

الإيجابيات الكاذبة (FP): الفئة الفعلية سلبية، والفئات المتوقعة إيجابية.

السلبيات الكاذبة (FN): الفئة الفعلية إيجابية، والفئة المتوقعة سلبية.

و من هذه المعاملات يتم حساب معاملات جديدة تعبر بدقة عن أداء نماذج التصنيف: [19]

- **الدقة (Accuracy)** هي مقياس التقييم الأكثر استخدامًا و تحدد بنسبة عدد المكونات المصنفة بشكل صحيح الى العدد الاجمالي لمكونات العينة و تحسب بالعلاقة:

$$\text{Accuracy}=\frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (5)$$

- **Precision** :يقيس قدرة النموذج على التنبؤ بالعينات الإيجابية بشكل صحيح و يشير إلى عدد التوقعات الموجبة فعليا من التوقعات الايجابية و يعطى بالعلاقة:

$$Precision=\frac{TN}{TP+FP} \quad (6)$$

- **Recall**: يقيس قدرة النموذج على تحديد العينات الإيجابية بشكل صحيح ويسمى هذا المقياس أيضا بالحساسية أو بالمعدل الإيجابي الحقيقي True Positive Ratio ويشير إلى عدد التوقعات الموجبة التي توقعها النموذج بشكل صحيح على أنها صحيحة، ويعطى بالعلاقة التالية:

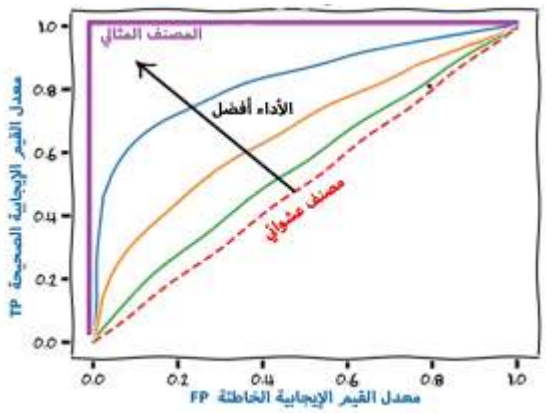
$$Recall=\frac{TP}{TP+FN} \quad (7)$$

- **F1-Score**: هي المتوسط التوافقي للدقة و الاستدعاء. كذلك ، تعتبر درجة F1 مقياس شائع الاستخدام عندما يكون كل من الدقة والاستدعاء مهمين ويعرف بالعلاقة التالية:

$$F1.score=\frac{TP}{TP+0.5(FP+FN)} \quad (8)$$

منحي (ROC): Receiver Operating Characteristic Curve

هو تمثيل رسومي لأداء نموذج مصنف ثنائي يتم إنشاؤه عن طريق رسم معدل القيم الإيجابية الحقيقية (TPR) مقابل معدل القيم الإيجابية الخاطئة (FPR) عند عتبات تصنيف مختلفة ، و تعبر المنطقة الواقعة تحت المنحني عن قدرة المصنف على التمييز بين الفئات الايجابية و السلبية كما هو مبين بالشكل التالي.



الشكل(11) منحني ROC

التحقق المتقاطع (Cross-validation) :

هو أسلوب تقييم نموذجي يقوم بتقسيم بيانات التدريب إلى مجموعات أصغر. ثم يتم تقسيم كل مجموعة إلى عينات للتدريب والتحقق من صحة النموذج. يتدرب النموذج أولاً مع العينة المخصصة ويتم اختباره مع العينات المتبقية بعدد (k) من المرات (k-times cross-validation) حيث k رقم يحدده المستخدم (5 أو 10)

و عادة يتم استخدام هذا الأسلوب للتأكد من عدم وجود Overfitting التي تمثل مشكلة في التعلم الآلي والإحصائيات حيث يتعلم النموذج أنماط مجموعة بيانات التدريب بشكل جيد للغاية، ويشرح مجموعة بيانات التدريب بشكل مثالي ولكنه يفشل في تعميم قوته التنبؤية على مجموعات أخرى من البيانات.

النتائج و المناقشة:

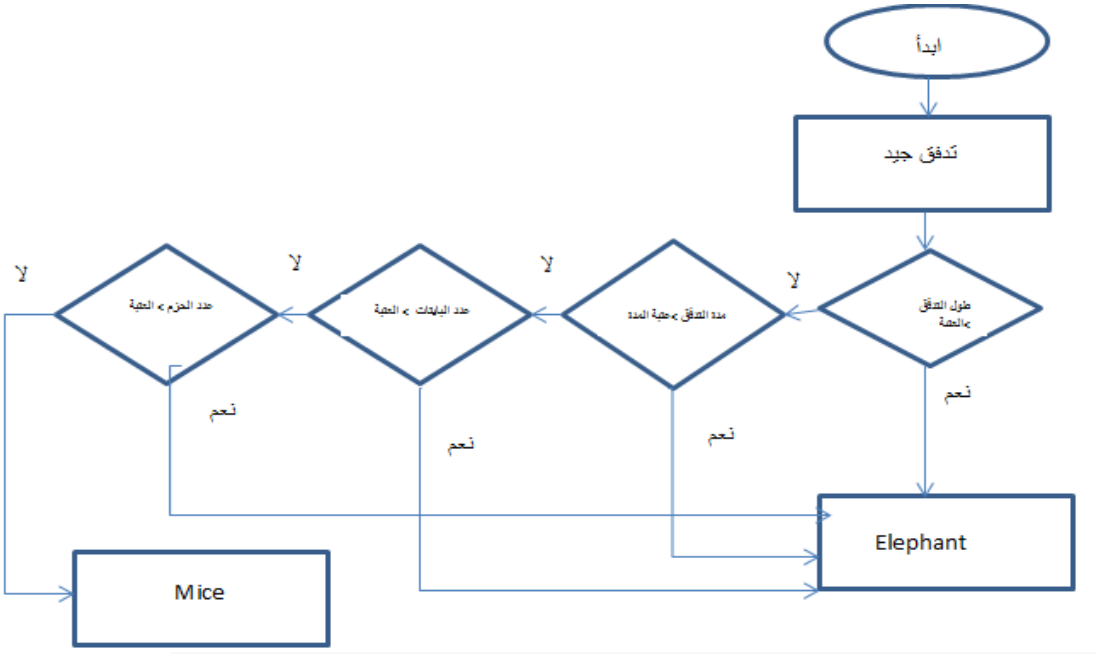
سنطبق نماذج لتعلم الآلة بلغة بايثون Python باستخدام مكتبة Scikit-learn ، وهذه المكتبة هي أداةٌ لتطبيق التعلم الآلي بلغة البايثون و مبنية على بعض التقنيات مثل NumPy و pandas و Matplotlib .

كما سنستخدم المُصنّفات KNN و RF و GB و DT و MLP و LR و SVM مع قاعدة بياناتٍ حقيقية، و بعد استدعاء قاعدة البيانات سيتم انشاء نموذج تنبؤي للتعلم الآلي بلغة بايثون حيث يوجد لدينا نوعي تصنيف اما أن يكون حجم التدفق كبير (Elephant) أو أن يكون حجم التدفق صغير (Mice) و بالتالي يعتبر هذا التصنيف ثنائي و ان المصنفات التي سبق ذكرها تدعم التصنيف الثنائي ، و يتم تحديد عتبات (thresholds) لمعرفة نوع التدفق بناء على طول التدفق مدة التدفق و عدد البايتات و عدد الحزم خلال الزمن كالتالي:

```
length_threshold = 400000
duration_threshold = 60e6  # 60 seconds in microseconds
bytes_per_s_threshold = 4e5  # 400 KB/s

packets_per_s_threshold = 100  # 100 packets/s
```

و تم اختيار هذه القيم لتتناسب مع بيئة الاختبار ويمكن تعديلها بحسب التطبيقات العاملة بالشبكة، و يتم المقارنة حيث اذا كانت طول التدفق أكبر من عتبة الطول أو مدة التدفق أكبر عتبة المدة أو عدد بايتات التدفق في الثانية أكبر العتبة المحددة أو عدد حزم التدفق في الثانية أكبر من عد الحزم المحدد يتم تصنيف هذا التدفق على تدفق كبير و الا يعتبر تدفق صغير

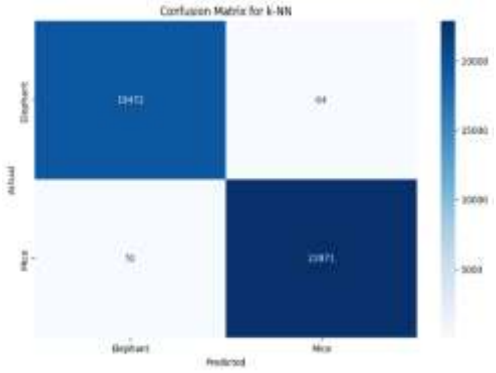


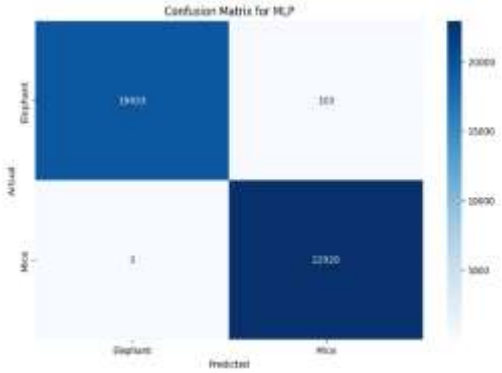
الشكل (12) مخطط خوارزمية العمل

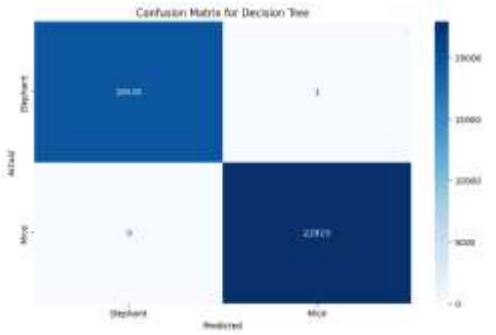
من اجل تطبيق نماذج التعلم الآلي يجب أولاً تنظيم البيانات في مجموعات اذ نقسم البيانات الخاصة بنا إلى جزأين قبل بناء النموذج، بحيث تكون هناك مجموعة للتدريب ومجموعة للاختبار. نستطيع استخدام المجموعة المخصصة للتدريب من أجل تدريب وتقييم النموذج أثناء مرحلة التطوير. حيث ستمنحنا منهجية تنبؤات هذا النموذج المُدرّب على المجموعة المخصصة للاختبار غير المرئية، فكرة دقيقة عن أداء النموذج وقوته.

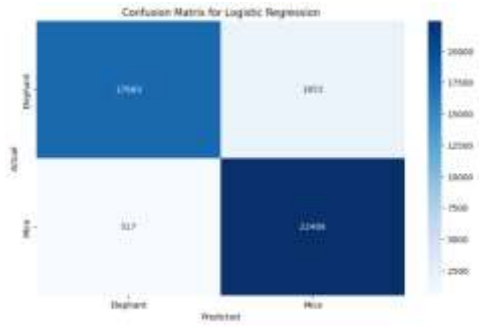
في مثالنا لدينا الآن مجموعة مخصصة للاختبار test تُمثّل 30% من مجموعة البيانات الأصلية، وسيشكل الجزء المتبقي من البيانات المجموعة المخصصة للتدريب train.

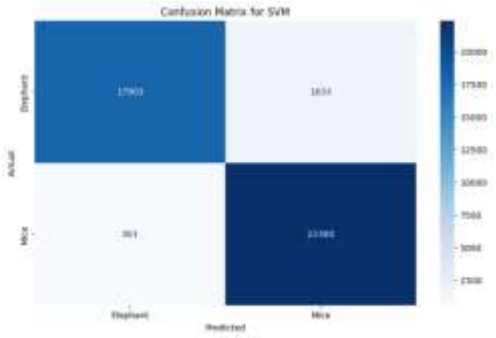
ثم نقوم ببناء النموذج و نُدّربه للتنبؤ على المجموعة المخصصة للاختبار ثم نقيمه بواسطة مقاييس الأداء و تبين المخططات التالية تقيم أداء النماذج عن طريق **(Confusion Matrix)** التي توضح عدد التنبؤات الصحيحة و الخاطئة التي قدمها النموذج مقارنة بالنتائج الفعلية:

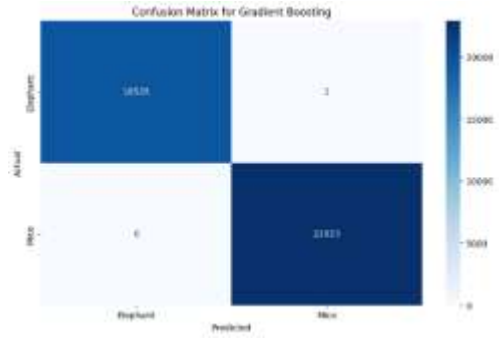


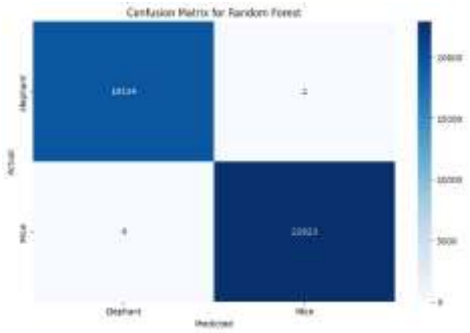






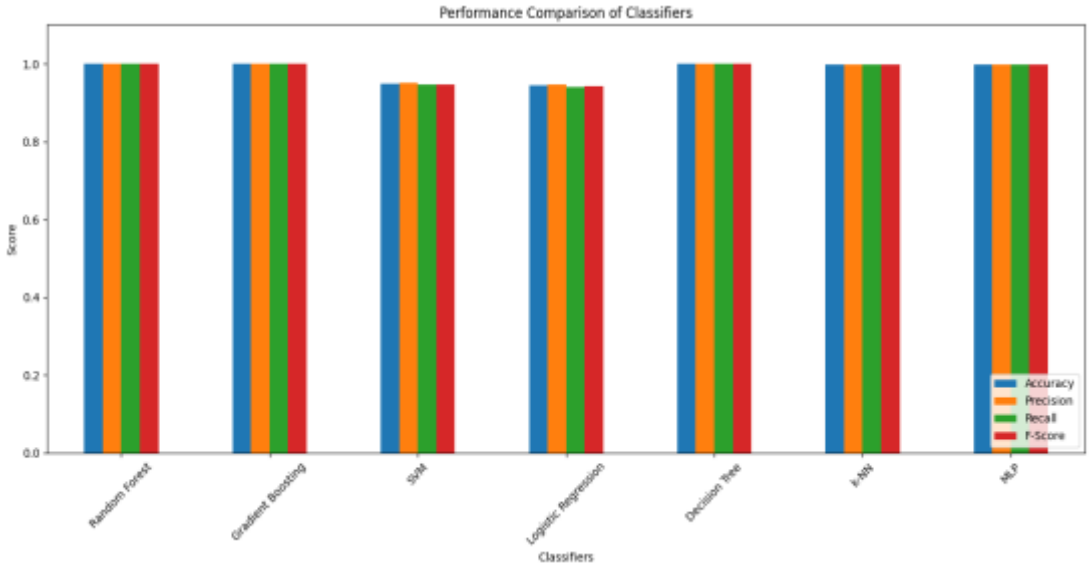






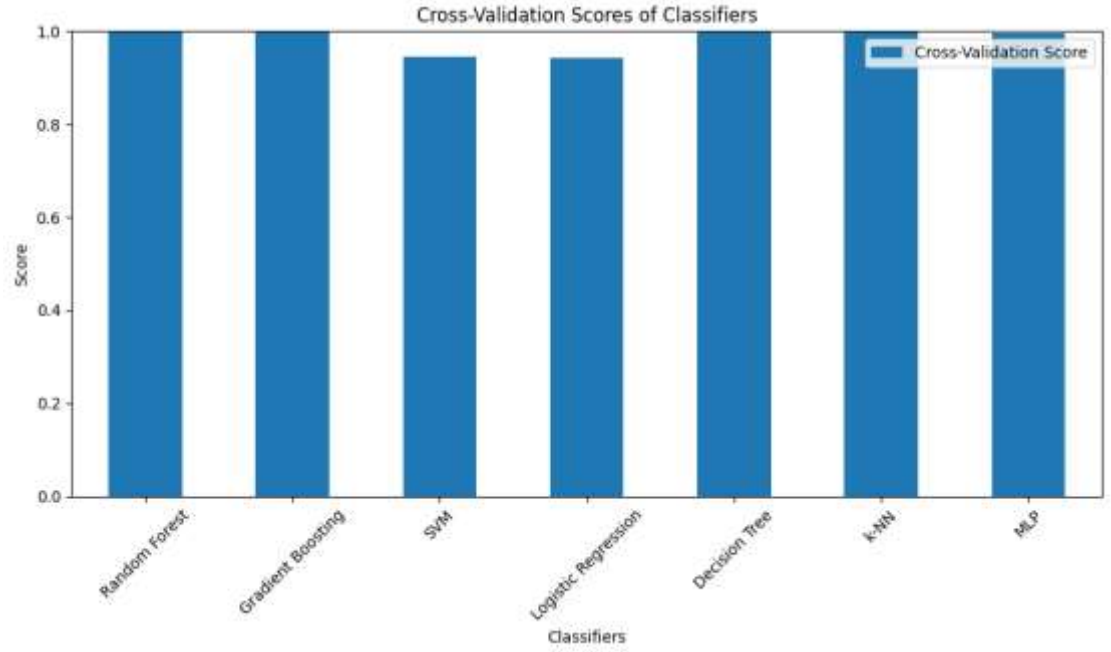
الشكل(13)confusion matrix

و قمنا بتطبيق معايير الأداء البقية على النماذج فكانت النتائج كالتالي:



الشكل(12) المقارنة بين خوارزميات التعلم الآلي

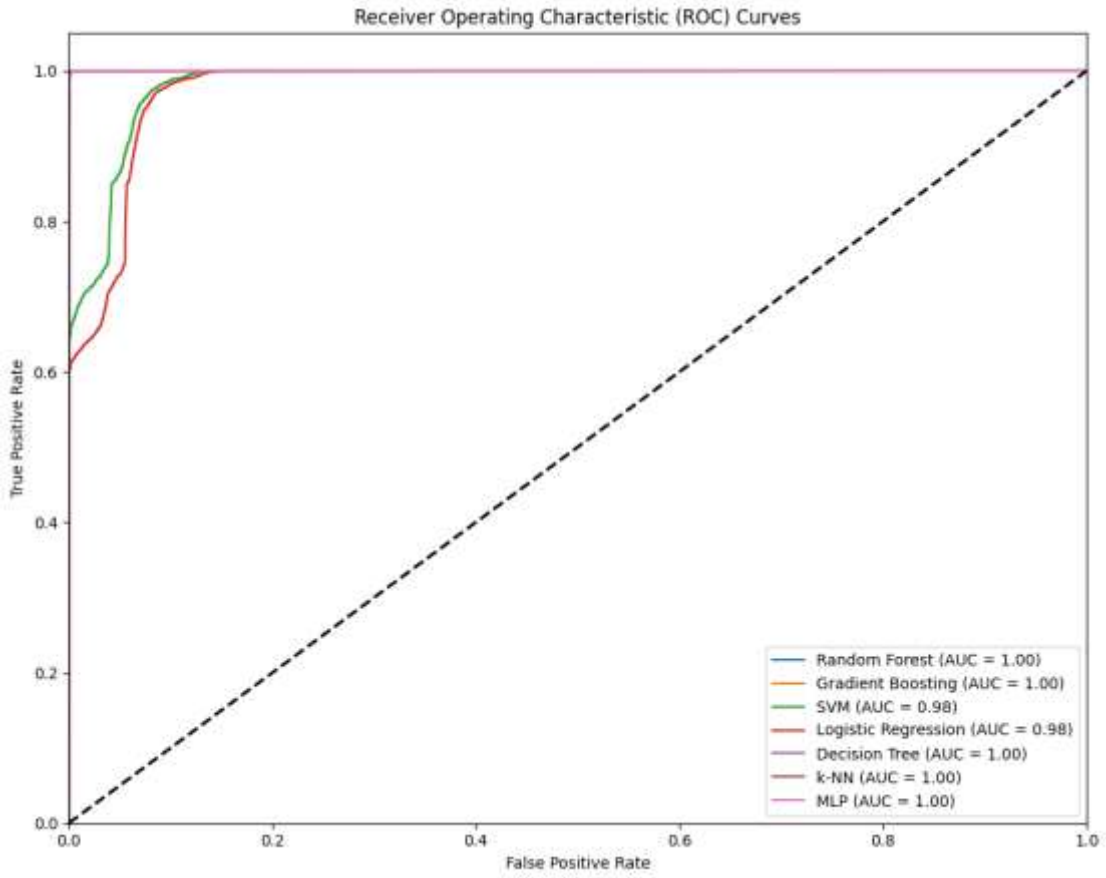
و بناءً على هذه النتائج نجد أن التصنيف باستخدام الخوارزميات (GP) (K-NN) (MLP) (RF) و (DT) حقق القيم المثالية للتصنيف و من أجل التأكد أن هذه المثالية التي أعطتها النماذج هي ليست مشكلة (over fitting) قمنا بتقييم أداء هذه النماذج بالتحقق المتقاطع **(Cross-validation)** فكانت النتائج كالتالي:



الشكل(14) المقارنة بين خوارزميات التعلم الآلي

حيث تم تقسيم بيانات التدريب الى مجموعات ثم تقسيم كل مجموعة الى عينات للتدريب و تم الاختبار في مثالنا هنا خمسة مرات أي K=5.

و تم تقييم نماذج التصنيف السابقة عن طريق رسم معدل القيم الإيجابية الحقيقية (TPR) مقابل معدل القيم الإيجابية الخاطئة (FPR) عند عتبات تصنيف مختلفة بما يسمى منحني (ROC)، و تعبر المنطقة الواقعة تحت المنحني عن قدرة المصنف على التمييز بين الفئات الايجابية و السلبية كما هو مبين بالشكل التالي :

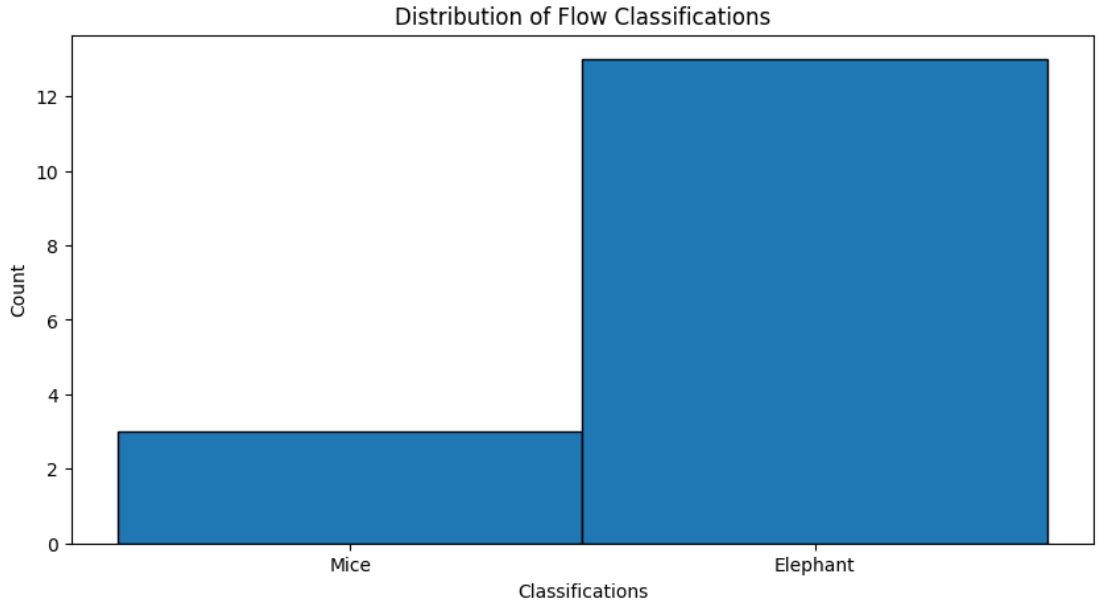


الشكل(15) منحني (ROC)

و من أجل اختبار نماذج التصنيف قمنا بتطبيق نموذج التصنيف المقترح على أحمال فعلية في الشبكة فكانت النتائج كالتالي:

```
2024-05-09 22:07:22,718 - INFO - Flow ('192.168.5.4', '62.72.160.114', 42298,
443, 6) classified as Mice
Flow ('192.168.5.4', '62.72.160.114', 42298, 443, 6) classified as Mice
2024-05-09 22:07:22,728 - INFO - Flow ('62.72.160.104', '192.168.5.4', 443,
42513, 6) classified as Elephant
Flow ('62.72.160.104', '192.168.5.4', 443, 42513, 6) classified as Elephant
2024-05-09 22:07:22,733 - INFO - Flow ('192.168.5.4', '62.72.160.104', 42513,
443, 6) classified as Elephant
Flow ('192.168.5.4', '62.72.160.104', 42513, 443, 6) classified as Elephant
2024-05-09 22:07:22,743 - INFO - Flow ('62.72.160.114', '192.168.5.4', 443,
42385, 6) classified as Elephant
Flow ('62.72.160.114', '192.168.5.4', 443, 42385, 6) classified as Elephant
2024-05-09 22:07:22,748 - INFO - Flow ('192.168.5.4', '62.72.160.114', 42385,
443, 6) classified as Elephant
Flow ('192.168.5.4', '62.72.160.114', 42385, 443, 6) classified as Elephant
2024-05-09 22:07:22,753 - INFO - Flow ('62.72.160.114', '192.168.5.4', 443,
42305, 6) classified as Elephant
Flow ('62.72.160.114', '192.168.5.4', 443, 42305, 6) classified as Elephant
2024-05-09 22:07:22,761 - INFO - Flow ('192.168.5.4', '62.72.160.114', 42305,
443, 6) classified as Elephant
Flow ('192.168.5.4', '62.72.160.114', 42305, 443, 6) classified as Elephant
2024-05-09 22:07:22,765 - INFO - Flow ('62.72.160.114', '192.168.5.4', 443,
42296, 6) classified as Elephant
Flow ('62.72.160.114', '192.168.5.4', 443, 42296, 6) classified as Elephant
2024-05-09 22:07:22,768 - INFO - Flow ('192.168.5.4', '62.72.160.114', 42296,
443, 6) classified as Elephant
Flow ('192.168.5.4', '62.72.160.114', 42296, 443, 6) classified as Elephant
```

```
2024-05-09 22:07:22,814 - INFO - Flow ('62.72.160.104', '192.168.5.4', 443,
42329, 6) classified as Mice
Flow ('62.72.160.104', '192.168.5.4', 443, 42329, 6) classified as Mice
2024-05-09 22:07:22,832 - INFO - Flow ('192.168.5.4', '62.72.160.104', 42329,
443, 6) classified as Elephant
Flow ('192.168.5.4', '62.72.160.104', 42329, 443, 6) classified as Elephant
2024-05-09 22:07:22,859 - INFO - Flow ('62.72.160.114', '192.168.5.4', 443,
42315, 6) classified as Elephant
Flow ('62.72.160.114', '192.168.5.4', 443, 42315, 6) classified as Elephant
2024-05-09 22:07:22,879 - INFO - Flow ('192.168.5.4', '62.72.160.114', 42315,
443, 6) classified as Elephant
Flow ('192.168.5.4', '62.72.160.114', 42315, 443, 6) classified as Elephant
2024-05-09 22:07:22,886 - INFO - Flow ('62.72.160.104', '192.168.5.4', 443,
42388, 6) classified as Elephant
Flow ('62.72.160.104', '192.168.5.4', 443, 42388, 6) classified as Elephant
2024-05-09 22:07:22,896 - INFO - Flow ('192.168.5.4', '62.72.160.104', 42388,
443, 6) classified as Elephant
Flow ('192.168.5.4', '62.72.160.104', 42388, 443, 6) classified as Elephant
2024-05-09 22:07:22,904 - INFO - Flow ('62.72.160.114', '192.168.5.4', 443,
42298, 6) classified as Mice
Flow ('62.72.160.114', '192.168.5.4', 443, 42298, 6) classified as Mice
```



الاستنتاجات والتوصيات:

قمنا في هذا البحث بتدريب مجموعة خوارزميات تعلم آلي على تصنيف أحمال الشبكة . ومع تنفيذ نماذج التعلم الخاضعة للإشراف المختلفة، كان من الممكن تصنيف حركة أحمال الشبكة بمستوى دقة أعلى من 98% لجميع النماذج. ومع ذلك، كانت الخوارزميات الأفضل أداءً هي (GP) (K-NN) (MLP) (RF) و (DT) ، بمستوى دقة 100%. وأظهرت الدراسة أداء هذه الخوارزمية أنها تتمتع بمستوى عالٍ من الدقة والحساسية، مما يعني أنها تستطيع التنبؤ بدقة بنوع حركة المرور (الفئران أو الفيل)،و قمنا بتطبيق نموذج التصنيف على أحمال شبكة فعلية حقيقية و أظهرت نتائج جيدة و في الأبحاث القادمة سنتقوم بتطبيقها على شبكات معرفة برمجيا (SDN).

References:

[1] Al Neyadi E, Al Shehhi S, Al Shehhi A, Al Hashimi N, Mohammad QH, Alrabaee S." *Discovering public wi-fi vulnerabilities using raspberry pi and kali linux"*. In2020

12th Annual Undergraduate Research Conference on Applied Computing (URC) 2020 Apr 15 (pp. 1-4). IEEE.

[2] Azab A, Khasawneh M, Alrabaee S, Choo KK, Sarsour M." *Network traffic classification: Techniques, datasets, and challenges*". Digital Communications and Networks. 2022 Sep 18.

[3] Eissa, M.E., Mohamed, M.A. & Ata, M.M. "*A robust supervised machine learning based approach for offline-online traffic classification of software-defined networking*". *Peer-to-Peer Netw. Appl.* **17**, 479–506 (2024).

[4] Salman O, Elhajj IH, Kayssi A, Chehab A." A review on machine learning–based approaches for Internet traffic classification". Annals of Telecommunications. 2020 Dec;75(11):673-710.

[5] Ö. Tonkal and H. Polat, 'Traffic Classification and Comparative Analysis with Machine Learning Algorithms in Software Defined Networks', *Gazi Üniversitesi Fen Bilim. Derg. Part C Tasar. Ve Teknol.*, vol. 9, no. 1, pp. 71–83, Mar. 2021

[6] Poupart P, Chen Z, Jaini P, Fung F, Susanto H, Geng Y, Chen L, Chen K, Jin H." *Online flow size prediction for improved network routing*" In2016 IEEE 24th International Conference on Network Protocols (ICNP) 2016 Nov 8 (pp. 1-6). IEEE.

[7] A.Ahmed ,G. Agunsoye "*A Real-Time Network Traffic Classifier for Online Applications Using Machine Learnin*"Algorithms 2021, 14, 250.

[8] Raikar MM, Meena SM, Mulla MM, Shetti NS, Karanandi M.*" Data traffic classification in software defined networks (SDN) using supervised-learning*" Procedia Computer Science". 2020 Jan 1;171:2750-9.

[9] Jang Y, Kim N, Lee BD." *Traffic classification using distributions of latent space in software-defined networks: An experimental evaluation*". Engineering Applications of Artificial Intelligence. 2023 Mar 1;119:105736.

[10] Tang F, Zhang H, Yang LT, Chen L. "*Elephant flow detection and load-balanced routing with efficient sampling and classification*". IEEE Transactions on Cloud Computing". 2019 Feb 26;9(3):1022-36.

[11] Lin-Huang C, Tsung-Han L, Hung-Chi C, Cheng-Wei S." *Application-based online traffic classification with deep learning models on SDN networks*". Advances in Technology Innovation". 2020 Aug 31;5(4):216.

[12] Mondal PK, Aguirre Sanchez LP, Benedetto E, Shen Y, Guo M.*" A dynamic network traffic classifier using supervised ML for a Docker-based SDN network*". Connection Science. 2021 Jul 3;33(3):693-718.

[13] Ashour MM, Yakout MA, AbdElhalim E. "*Traffic Classification in Software Defined Networks based on Machine Learning Algorithms*". International Journal of Telecommunications. 2024 Feb 8;4(01):1-9.

[14] Durner R, Kellerer W."*Network function offloading through classification of elephant flows*". IEEE Transactions on Network and Service Management. 2020 Feb 27;17(2):807-20.

[15] Joshi S, Budihal SV. "*Traffic Classification of Software-Defined Networks Using Machine Learning*". InInternational Conference on ICT for Sustainable Development 2023 Aug 3 (pp. 1-9). Singapore: Springer Nature Singapore.

[16] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "*Evaluating the Quality of Machine earning Explanations: A Surveyon Methods and Metrics*", Electronics, vol. 10, no. 5, p. 593, Mar. 2021

[17] Khalifah J, Issa M ."*Classification Of Network Load Traffic By taking advantage of Software Defined Network*" .Tishreen University Journal-Engineering Sciences Series,. 2023;45(6):115-27.

[18] Suryadevara CK. "*DIABETES RISK ASSESSMENT USING MACHINE LEARNING: A COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS*". IEJRD-International Multidisciplinary Journal. 2023 Aug;8(4):10.

[19] Pallathadka H, Wenda A, Ramirez-Asís E, Asís-López M, Flores-Albornoz J, Phasinam K. "*Classification and prediction of student performance data using various machine learning algorithms*". Materials today: proceedings. 2023 Jan 1;80:3782-5.

[20] Patle A, Chouhan DS. "*SVM kernel functions for classification*". In2013 International conference on advances in technology and engineering (ICATE) 2013 Jan 23 (pp. 1-9). IEEE.

[21] Yekkehkhany B, Safari A, Homayouni S, Hasanlou M. *"A comparison study of different kernel functions for SVM-based classification of multi-temporal polarimetry SAR data*". The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. 2014 Oct 22;40:281-5.

[22] Patel HH, Prajapati P. Study and analysis of decision tree based classification algorithms. International Journal of Computer Sciences and Engineering. 2018 Oct 31;6(10):74-8.

[23] Bentéjac C, Csörgő A, Martínez-Muñoz G.'*A comparative analysis of gradient boosting algorithms*". Artificial Intelligence Review. 2021 Mar;54:1937-67.

[24] Breiman L. Random forests. Machine learning. 2001 Oct;45:5-32.

[25] Liu K, Hu X, Zhou H, Tong L, Widanage WD, Marco J." *Feature analyses and modeling of lithium-ion battery manufacturing based on random forest classification*". IEEE/ASME Transactions on Mechatronics. 2021 Jan 5;26(6):2944-55.

[26] S. Shekhar and H. Xiong, "*Nearest Neighbor*," Encycl. GIS, vol. I, pp. 771–771, 2008,

[27] Mucherino A, Papajorgji PJ, Pardalos PM, Mucherino A, Papajorgji PJ, *Pardalos PM." K-nearest neighbor classification*". Data mining in agriculture. 2009:83-106.

[28] Taud H, Mas JF. "*Multilayer perceptron (MLP)*". Geomatic approaches for modeling land change scenarios. 2018:451-5.

[29] Bisong E, Bisong E. *"The multilayer perceptron (MLP). Building Machine Learning and Deep Learning Models on Google Cloud Platform*": A Comprehensive Guide for Beginners. 2019:401-5.

[30] Hosmer Jr DW, Lemeshow S, Sturdivant RX. "*Applied logistic regression*". John Wiley & Sons; 2013 Feb 26.

[31] TONKAL Ö, POLAT H. "*Traffic Classification and Comparative Analysis with Machine Learning Algorithms in Software Defined Networks*". Gazi University Journal of Science Part C: Design and Technology. 2021 Mar 3;9(1):71-83.

[32] Gómez J, Riaño VH, Ramirez-Gonzalez G. "*Traffic classification in IP networks through Machine Learning techniques in final systems*". IEEE Access. 2023 May 4;11:44932-40.