المجلة المصرية للسكان وتنظيم الأسرة

مجلد (55) العدد الثاني – ديسمبر 2023
Homepage:.mskas.journals.ekb.eg
Print ISSN 1110-1156 – Online ISSN 2786-0078

جامعة القاهرة

المقارنة ما بين تعلم الالة والتعلم العميق في بناء نموذج قادر على التنبؤ بخروج العملاء من الشركات الخدمية

حنان خضارى مهدى محمود

مدرس بمعهد النيل العالى لتكنولوجيا المعلومات والحاسب الآلي

تم الأستلام :2023-07-12 تم المراجعة: 2023-08-14 تم القبول: 2023-08-17

الملخـــص

ان ظهور الثورة الرقمية ومعها ظهور تقنيات تعلم الالة والتعلم العميق في الساحة كإضافة حديثة للقائمة الطويلة من المناهج والممارسات والأدوات العلمية، التي زودتنا بمنظور جديدة للكثير من مشاكلنا التقليدية وساهمت في زيادة فهمنا لمنظومات عالمنا المعاصر ومشاكل هذه المنظومات بقصد يمكننا من التحكم بها .حيث تعتمد الشركات تلك المعنية بالخدمات على وجه الخصوص على تقنيات تعلم الالة والتعلم العميق للقيام بعملياتها وإدارتها، والتفاعل مع عملائها وموردیها.

يهدف هذا البحث الى بناء نموذج قادر على التنبؤ بخروج العملاء من الشركات الخدمية، باستخدام خوارزميات تعلم الالة والتعلم العميق والمقارنة بينهم واختيار الاكثر دقة.

وتم تطبيق ذلك على قاعدة بيانات عملاء تابعين لشركة اتصالات، وبعد تطبيق 8 خوارزميات من خوارزميات تعلم الالة خوارزمية الانحدار اللوجستى (Logistic regression) وخوارزمية الغابة العشوائية (Random Forest) وخوارزمية الجار الاقرب K-Nearest Neighbor(KNN) وخوارزمية بايز (Naive Bayes) وخوارزمية الة المتجهات الداعمة Support vector machine (SVM) وخوارزمية شجرة القرار(Decision Tree) وخوارزمية تعزيز التدرج (Gradient Boosting) بجانب خوارزمية التعلم العميق Deep Learning (DNN) ،وتمت المقارنة بينهم حسب المعايير(Recall ، Classification Accuracy، Precision، Area Under the Curve) ،وتم اختيار خوارزمية تعزيز التدرج التي أعطت أعلى دقة حيث وصلت نسبة الدقة إلى 84.7% ومن خلال تحليل البيانات والنتائج والمعلومات المستخرجة، والتي طبقت عبر خوارزميات تعلم الالة والتعلم العميق فى البيانات امكن معرفة بعض العوامل المؤثرة والاسباب وراء خروج العملاء من شركات الاتصالات.

الكلمات الدالة

برنامج (orange) ، خوارزميات تعلم الالة ، خوارزميات التعلم العميق

مقدمة

يعد الاحتفاظ بالعملاء من أهم المسائل التي تشغل جميع الشركات وخاصة الشركات الخدمية، حيث تسعى هذه الشركات إلى تحسين جودة الخدمات التي تقدمها وتوفير تجربة استخدام ممتازة للعملاء. ومع ذلك، فإن العديد من العملاء يقررون الرحيل عن الشركات الخدمية لأسباب مختلفة، وهو ما يؤثر بشكل سلبي على أداء الشركة وأرباحها.

وعليه، إن التوجه السائد في العالم حاليا يتجه نحو استثمار تقنيات الذكاء الاصطناعي التي تساهم في تحقيق أهدافها، وهذا ما يقدمه البحث الحالي من خلال استخدام تقنيات الذكاء الاصطناعي للتنبؤ بخروج العملاء في شركة خدمية

ويأتي دور التنبؤ بخروج العملاء كأداة حيوية للمساعدة في توقع العملاء الذين قد يرحلون عن الشركة، وذلك بحيث يمكن للشركة اتخاذ الإجراءات اللازمة للحفاظ على هؤلاء العملاء. ومن أهم الأدوات المستخدمة لتحليل وتنبؤ سلوك العملاء هي تقنيات تعلم الآلة والتعلم العميق.

وفي هذا البحث، سنقوم بدراسة مقارنة بين خوارزميات تعلم الآلة وخوارزميات التعلم العميق في التنبؤ بخروج العملاء من الشركات الخدمية. وتهدف هذه الدراسة إلى تحليل ومقارنة أداء الخوارزميات المختلفة، بما في ذلك الدقة والكفاءة، وتحليل أي تباين أو اختلاف في أداء الخوارزميات المختلفة.

أهمية البحث:

ترجع اهمية البحث الى :

1) بناء عدة نماذج قادرة على التنبؤ بخروج العملاء
2) اختيار الخوارزمية الأفضل القادرة على إعطاء أفضل نتائج للتنبؤ بخروج العملاء ،
3) مساعدة الشركات على فهم الأسباب الرئيسية التي تدعو العملاء إلى ترك الشركة وخاصة الشركات الخدمية وعلى سبيل المثال شركة الاتصالات،
4) ندره الكتابات في مجال التنبؤات الدقيقة عن خروج العملاء أو الاحتفاظ بالعميل،
5) تقديم توصيات تساهم في تحسين إدارة العلاقة مع العملاء في المنظمات بشكل عام، وفي قطاع الاتصالات بشكل خاص

مشكلة البحث:

تعد صناعة الاتصالات المحرك الأساسي المهيمن في قطاع التكنولوجيا، والبيئة الأكثر تنافسا في القرن الحادي والعشرين، بقيمة بلغت 7.2 تريليون دولار في عام 2019، وقد صنفت School Business Harvard استراتيجية الاحتفاظ بالعملاء الحاليين بأنها الاستراتيجية الأكثر في تحقيق ربح للشركات، حيث ان زيادة معدلات الاحتفاظ بالعملاء بنسبة 5 %ستؤدي إلى زيادة الأرباح بنسبة 25 %إلى 95 %

(**B**eers 2021).

يتجلى نجاح هذه الاستراتيجية باستخدام تقنيات الذكاء الاصطناعي بالاعتماد على مجموعة من الخوارزميات الرياضية الشهيرة والعمليات الاحصائية والتحليلية، عبر جمع وتنظيم وتحليل بيانات العملاء والاحتفاظ بهم وذلك للتنبؤ المسبق للعملاء

المحتمل خروجهم والوقت المحتمل لهذا الخروج، وتكريس كافة الجهود للحفاظ على العملاء عبر اكتشاف المعرفة والتنبؤ بالمستقبل

وعادة ما تركز الشركات بشكل أكبر على اكتساب العملاء الجدد، بينما الاحتفاظ بالعملاء الحاليين كان يأتي كأولوية ثانويةً دائما، ومع ذلك يمكن أن يكلف جذب عميل جديد خمسة أضعاف تكلفة الاحتفاظ بعميل حالي (عبيد 2018).، ولتعظيم عدد العملاء اتجهت الشركات للاحتفاظ بالعميل القديم بدلا من جذب عملاء جدد للشركة ، كما ان الاتجاه نحو العميل القديم افضل لانه يوجد لدى الشركة البيانات اللازمة حول تفاعل العميل مع الخدمات، و بالتالي يمكن التنبؤ بالعميل الذي سيغادر، ومعرفة الرد المناسب لإبقائه، بتقديم بعض العروض أو الباقات التي تهمه. لهذا يسعى هذا البحث إلى بناء نموذج للتنبؤ بمشكلة خروج العملاء باستخدام تقنيات الذكاء الاصطناعي (تعلم الالة والتعلم العميق)، حيث تعتبر هذه المشكلة من أخطر المشاكل التي تسعى جميع الشركات وخصوصا الخدمية منها لإيجاد الحلول لها.

أهداف البحث:

1. يهدف البحث إلى بناء نموذج قادرا على التنبؤ باستخدام خوارزميات تعلم الالة والتعلم العميق
2. المقارنة بين بعض خوارزميات الذكاء الاصطناعي من تعلم الالة والتعلم العميق فى التنبؤ
3. الاستفادة من خوارزميات تعلم الالة والتعلم العميق لبناء نموذج تنبؤي قادر على التنبؤ بخروج العملاء لتمكين العاملين بالشركات الخدمية لمعرفة فيما إذا كان العملاء سيجددون عقودهم أم لا فى ضوء استخدام هذا النموذج فى الحصول على تنبؤات دقيقة عن خروج العملاء
4. محاولة اكتشاف السمات المشتركة بين العملاء المتسربين للحد من الأسباب الداخلية في الشركات الخدمية التي تدعو العملاء إلى ترك خدمة الشركة، مما يكبد هذهً الشركات خسائر كبيرة، نظرا لأهمية العميل للشركات الخدمية.
5. محاولة البحث عن أهم الأسباب التي تؤثر على العملاء، وتؤدي لتركهم الشركة.

منهجية البحث:

اعتمد هذا البحث على المنهج التنبؤي وهو الذي يستطلع بدراسة" المستقبل "وفق برامج وآليات التنبؤ، من خلال دراسة التغيرات في سلوك الماضي وخط سيره، وإسقاطاتها على المقابل المستقبلي لها. وتم ذلك من خلال دارسة وتحليل بعض سمات واتجاهات العملاء، واستخدام خوارزميات التعلم الالى والتعلم العميق ومن ثما اقتراح النموذج الملائم بناءً على المعطيات الموجودة (حمود 2020)

بيانات البحث

تم تحميل البيانات من موقع Kaggle الخاص بنشر مجموعات البيانات التابعة للمنظمات والجامعات والشركات والتي تنشر بغرض وضع بعض المقترحات من قبل الباحثين لحل هذه المشكلة، وبغرض البحث العلمي

أدوات تحليل البيانات

برنامج **Orange** هو بيئة تطوير مفتوحة المصدر مخصصة لتحليل البيانات وتصورها بشكل بصري. يستخدم البرنامج لإنشاء نماذج تعلم آلي، وتحليل البيانات، وإجراء تحليل استكشافي، وتصور النتائج بواجهة سهلة الاستخدام تعتمد على سحب وإفلات (Drag-and-Drop)، مما يجعلها سهلة الاستخدام حتى لأولئك الذين ليس لديهم خلفية قوية في مجال تحليل البيانات ويدعم برنامج Orange مجموعة متنوعة من الوظائف والتقنيات، بما في ذلك:

- تنظيم وتنظيف البيانات.

- إنشاء نماذج تعلم آلي وتحليلها.
- تصنيف وتنبؤ البيانات.
- تجميع وتصفية البيانات.
- تصور البيانات باستخدام رسوم بيانية مختلفة.
- إجراء تحليل استكشافي لاستكشاف العلاقات بين المتغيرات.

وتعتمد واجهة برنامج Orange على السحب والإفلات..

الدراسات السابقة:

الدراسة الأولى :عبد الرحيم قاسم احمد (2018) "توقع خروج العملاء من شركات الاتصالات باستخدام تعلم الالة في بيئة البيانات الكبيرة"

استخدمت هذه الدراسة نظام يعتمد على خوارزميات تعلم الآلة للتنبؤ بالعملاء المحتمل تسربهم من شركات الاتصالات، حيث قدمت طريقة جديدة لتحضير البيانات واختيار المتغيرات المناسبة لمعالجة البيانات الكبيرة، على قاعدة بيانات عالمية من شركة Orange حيث اشارت النتائج التي تم الحصول عليها في هذا العمل إلى دقة عالية مقارنة بالأعمال الأخرى المنشورة باستخدام نفس قاعدة البيانات، وتمت مقارنة عدة خوارزميات شجرية وهي (شجرة القرار والغابات العشوائية وخوارزمية شجرة التعزيز التدريجي وخوارزمية شجرة التعزيز التكيفى)

وتوصلت الدراسة الى أن التنبؤ بالتسرب الوظيفي يعد من أحد أهم المصادر التي تحقق عائدا للشركات من خلال الحفاظ على العملاء، وتضمنت البيانات المستخدمة للوصول إلى النموذج الأمثل على شقين القسم الأول من شركة Orange وتم تجريب هذا النموذج على شركة سيرتيل وتم وضع القاعدة التالية :إذا لم يقم أحد العملاء بإجراء او استقبال أي مكالمات خلال فترة زمنية تبلغ 30 يوم فقد تم اعتباره متسربا، وحقق نموذج شجرة التعزيز التكيفي أفضل النتائج حيث حقق Roc بنسبة % 75.00 يليها خوارزمية شجرة التعزيز التدريجي ثم الغابات العشوائية واخيرا شجرة القرار

الدراسة الثانية: Dost, el. (2019)

The Classification of Customers' Sentiment using Data Mining Approaches

تسلط هذه الدراسة الضوء على أهمية التنقيب في البيانات وتحليل المشاعر للمديرين وصناع القرار لمراقبة التقدم والحفاظ على جودة منتجاتهم أو خدماتهم ومراقبة أحدث اتجاهات السوق لدعم الأعمال. واقترحت الدراسة إطارًا لتحليل المشاعر لآراء العملاء باستخدام تقنيات استخراج البيانات مستخدمه ثلاث خوارزميات للتعلم الآلي خاضعة للإشراف وهى آلة المتجهات الداعمة (Support Vector Machines) SVM ، شجرة القرارات (Decision Tree) ، وبايز Naive Bayes وتوصلت الدراسة الى ان المتجهات الداعمة اكثر دقة بنسبة 90.30% وتوصلت الدراسة الى ان المنهجية المقترحة مفيدة للباحثين ومقدمي الخدمات وصناع القرار.

الدراسة الثالثة Kavitha el (2020)

Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms.

اقترحت هذه الدراسة نموذجًا يستخدم تقنيات التعلم الآلي للتنبؤ بخروج العملاء من شركات الاتصالات. واستخدمت الدراسة خوارزميات مثل Decision Tree و Random Forest و XGBoost لتحليل سلوك العملاء وتحديد أولئك الذين من المحتمل أن يلغوا اشتراكهم. وتحديد اكثر الخدمات تفضيلا لهؤلاء العملاء، حتى تخفيض معدل التغيير، مما يجعل خدمات الاتصالات أكثر

ربحية. وتؤكد هذه الدراسة أيضًا على أهمية اختيار المتغيرات لضمان استخدام المتغيرات المهمة فقط في التحليل، مما قد يؤدي إلى دقة أفضل.

الدراسة الرابعة Homa, el. (2021)

Customer Churn Prediction in Irancell Company Using Data Mining Methods .

ناقشت هذه الدراسة تحدي خروج العملاء في الشركات والمؤسسات، حيث يمكن أن يؤدي فقدان العملاء إلى خسائر مالية وحتى الإفلاس. واقترحت الدراسة طريقة هجينة تعتمد على الخوارزمية الجينية Genetic Algorithm والشبكة العصبية المعياريةModular Neural Network للتنبؤ بخروج العملاء باستخدام بيانات من شركة Irancell. وقد بلغت دقة الطريقة المقترحة حوالي 95.5٪ مقارنة بالطرق الأخرى في هذا المجال. وسلطت هذه الدراسة الضوء على استخدام الشبكة العصبية المعيارية مع وحدتين من الشبكة العصبية المغذية والخوارزمية الجينية للحصول على البنية المثلى لوحدات الشبكة العصبية كأهم مؤشرات هذه الطريقة.

الدراسة الخامسة: Mahmoud, el. (2021)

Customer Retention: Detecting Churners in Telecoms Industry using Data Mining Techniques

هدفت هذه الدراسة إلى تحليل تأثير جودة الخدمة ونماذج التنبؤ التي تعتمد على تقنيات التنقيب فى البيانات للكشف عن الاضطرابات في صناعة الاتصالات. يصنف النموذج المقترح بيانات العملاء المتذبذبين باستخدام خوارزميات التصنيف لتحديد الأسباب الجذرية لخروج العملاء. وكشفت النتائج أن هناك خمس ارتباطات بين سلوك العملاء والمتغيرات التي تؤثر على جودة الخدمة، مما يؤثر على ولاء العملاء. حيث قامت الدراسة بتقييم أداء ثلاث خوارزميات، NN وSVM وRF، وحساب مصفوفة الشك لكل منها. واستخدمت الدراسة برنامج WEKA للحصول على النتائج وبشكل عام، قدمت الدراسة رؤى حول استخدام تقنيات التنقيب في البيانات للتنبؤ بخروج العملاء في قطاع الاتصالات.

الدراسة السادسة Fujo, el (2022)

Predicting Customer Churn in the Telecommunication Industry Using Machine Learning Techniques

توصلت دراسة "" إلى أن تقنيات تعلم الآلة يمكن استخدامها بنجاح في توقع خروج العملاء من شركات الاتصالات. وقد استخدمت الدراسة مجموعة من الخوارزميات الشائعة في تعلم الآلة مثل الشبكات العصبية (Neural Networks) **وآلة المتجهات الداعمة** (Support Vector Machines) وشبكات العصب الانتقائي (Selective Neural Networks) ، وقد حصلت على دقة جيدة في التنبؤ بخروج العملاء. كما توصلت الدراسة إلى أن استخدام تقنيات التعلم العميق يمكن أن يؤدي إلى نتائج أفضل في بعض الحالات، ولكن يتطلب ذلك المزيد من البيانات والموارد الحاسوبية.

الدراسة السابعة Dalli, A. (2022)

Impact of Hyper parameters on Deep Learning Model for Customer Churn Prediction in Telecommunication Sector

توصلت هذه الدراسة إلى أن استخدام تقنيات التعلم العميق يمكن أن يكون فعالاً في توقع احتمالية خروج العميل في صناعة الاتصالات، وأن تحسين الهايبرباراميترز (Hyper parameters) يمكن أن يؤدي إلى تحسين أداء نماذج التنبؤ بانسحاب العميل وتوصلت الدراسة إلى أنه توجد عوامل متعددة تؤثر بشكل كبير على دقة نماذج التنبؤ منها:

1. نوع الشبكة العصبية: يجب اختيار النموذج العصبي الأنسب لمشكلة التنبؤ المحددة.

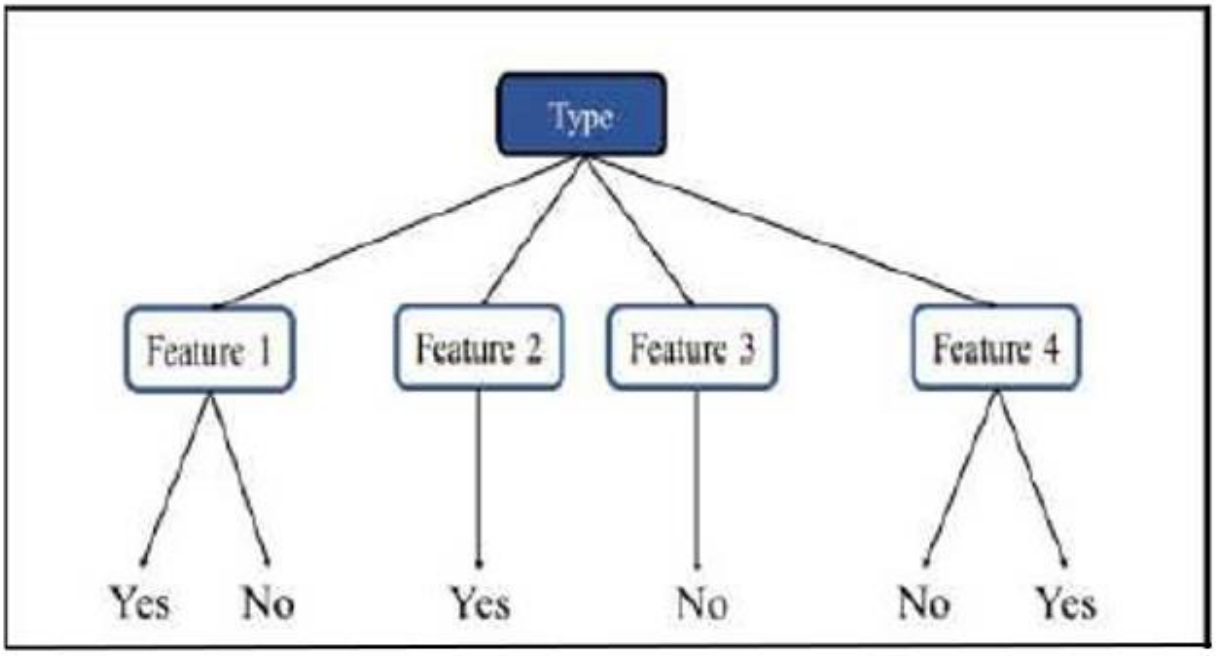
2. حجم البيانات: يجب جمع حجم كافٍ من البيانات لتدريب نموذج التنبؤ العميق بشكل صحيح.
3. تحسين الهايبربراميترز: يجب تحسين الهايبربراميترز لتحسين أداء نموذج التنبؤ، مثل عدد الطبقات العصبية وعدد الوحدات العصبية وسرعة التعلم.
4. طريقة تقسيم البيانات: يجب استخدام طريقة مناسبة لتقسيم البيانات إلى مجموعات التدريب والتحقق والاختبار لتحقيق أفضل أداء لنموذج التنبؤ.

الخوارزميات المستخدمة في التنبؤ:

1) شجرة القرار (Decision Tree)

تعد من أبرز خوارزميات التصنيف وهي نموذج استكشافي يظهر على شكل شجرة كما يعبر اسمها، وبشكلٍ دقيق يمثل كل فرع من فروع الشجرة سؤال، وتمثل أوراقها أجزاء من قاعدة البيانات تنتميً تصنيفينا للتصنيفات التي تم بنائها (Dean,2014).

وتعتبر خوارزمية شجرة القرار واحدة من الخوارزميات المشهورة في التصنيف، وهي مكونة من مجموعة من العقد والوصلات، وكل عقدة يمكن ان تعبر عن سمة من السمات، وكل وصلة تعبر عن قيمة ما لهذه السمة، أما عقد الأوراق فإنها تعبر عن الهدف المتنبأ به وهو حالة خروج العملاء في هذا البحث وتكمن شهرة هذه الخوارزميات بسبب سهولة تفسيرها وكونها قادرة على التعامل مع السمات الفئوية ويمكن توسيعها لتتعامل مع أكثر من صنفين ويوضح الشكل التالى شجرة القرار



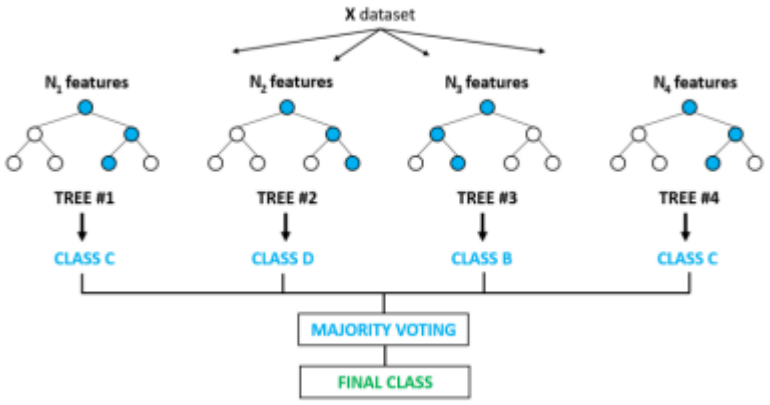
شكل توضيحى (1) لخوارزمية **Decision Tree**

2) الغابة العشوائية Random Forest

هي أحد النماذج الإحصائية المستخدمة في تعلم الآلة وتحليل البيانات، وتعتبر تطويرًا لنموذج شجرة القرار (Decision Tree). وتتألف الغابة العشوائية من مجموعة من الأشجار العشوائية، والتي تتكون كل منها من شجرة قرار مستقلة تعتمد على تقنية العينات العشوائية.

تتميز الغابة العشوائية بالدقة العالية في التنبؤ والتصنيف والتحليل، وتعد من أكثر النماذج استخدامًا في تطبيقات تعلم الآلة، وذلك لأنها تتيح للمستخدمين الحصول على تصنيف دقيق للبيانات بشكل سريع.

ويتم بناء الغابة العشوائية عن طريق إنشاء مجموعة من الأشجار العشوائية، وذلك بتحديد عدد معين من الأشجار وتدريب كل شجرة بطريقة مختلفة، وذلك باستخدام عينات عشوائية من البيانات التدريبية، ومن ثم دمج نتائج كل شجرة لتوفير تصنيف شامل ودقيق للبيانات (Smith 2017) وتستخدم الغابة العشوائية في العديد من التطبيقات، مثل التصنيف الآلي والتنبؤ وتحليل البيانات وتحليل الصور وتحليل النصوص والشكل التالي يوضح شكل الغابة العشوائية



شكل توضيحى (2) لخوارزمية الغابة العشوائية

حيث يتم تدريب 4 أشجار عشوائية في الغابة العشوائية، وفيها يتم توزيع البيانات التدريبية بشكل عشوائي على كل شجرة. ويتم تصنيف البيانات في نهاية كل شجرة إلى فئة معينة، ويتم دمج نتائج التصنيف لكل شجرة للحصول على تصنيف شامل ودقيق للبيانات.

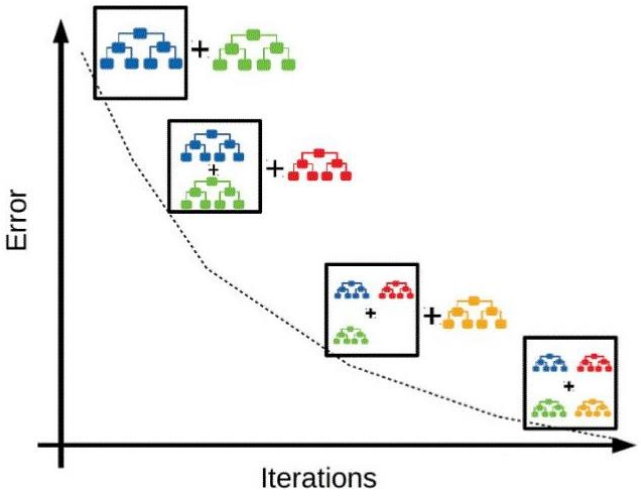
3) الانحدار اللوجيستي Logistic regression

هو نوع من أنواع الانحدار، وهو يعد من أكثر النماذج شيوعاً في تحليل البيانات الوصفية، حيث أنه أسلوب إحصائي لفحص العلاقة بين المتغير التابع ذي المستوي الوصفي ومتغير واحد أو أكثر من المتغيرات المستقلة (رزوق 2013) ، وهو نموذج احصائي ينتمي لنماذج الانحدار الخطي يمكن من نمذجة متغير ثنائي الحد بدلالة مجموعة من المتغيرات العشوائية المتوقعة، رقمية كانت او فئوية، ويستخدم الانحدار اللوجيستي للتنبؤ باحتمالية وقوع حدث ما بمعرفة إضافية لقيم متغيرات يمكن ان تكون مفسرة أو مرتبطة بهذا الحدث (دعيش واخرون 2018).

4) خوارزمية تعزيز التدرج Gradient Boosting Algorithm

تعتبر خوارزمية تعزيز التدرج Gradient Boosting Algorithm واحدة من أشهر الخوارزميات المستخدمة في تعلم الآلة للتنبؤ بالقيم المستقبلية وتحسين دقة النماذج ويتم استخدام هذا الخوارزمية في مجالات مختلفة مثل التحليل الاحصائي وتحليل البيانات والذكاء الاصطناعى (Candice et al. 2021).

وتتميز هذه الخوارزمية بالقدرة على التعامل مع مجموعة كبيرة من البيانات والمتغيرات وتعلم العلاقات بينها بطريقة فعالة حيث تستند الخوارزمية على فكرة التعلم التدريجي، حيث يتم بناء النماذج التالية على أساس نماذج سابقة اى زيادة الدقة عن طريق تقليل دالة الخسارة (الخطأ الذى ينشأ من اختلاف القيمة المتوقعة عن القيمة الفعلية) وجعل هذا الخطأ هدفا للتكرار التالى والشكل التالي يوضح شكل الخوارزمية

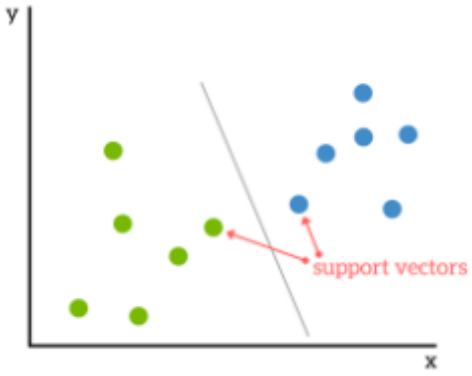


شكل توضيحى (3) لخوارزمية Gradient Boosting Algorithm

5) **آلة المتجهات الداعمة Support vector machine (SVM)**

هي خوارزمية تستعمل بشكل أساسي في عمليات التصنيف. وتعتمد **SVM** على فكرة إيجاد مستوي **Hyperplane** والذي يقوم بتقسيم البيانات إلى عدة إجزاء منفصلة بحيث يمثل كل جزء منهم **Class** أو تصنيف معين من البيانات. وتعتبر أقرب نقطتين للمستوي الفائق الذي نقوم برسمه الـــ **SVM** هي المتجهات الداعمة **Support Vector**.وتستعمل **SVM** في مهام تصنيف النصوص مثل تصنيف المواضيع وتمييز الرسائل المزعجة وتحليل المشاعر**(Kelleherd & Tierney, 2018)** ، وفي التعرف على الصور، وفي مجالات تمييز الأرقام المكتوبة يدويًا.

وتمتاز **SVM** بالدقة والنتائج المرتفعة، وتعمل جيدًا على مجموعات البيانات الصغيرة، وكذلك تظل فعالة في الحالات التي يكون فيها عدد الأبعاد أكبر من عدد العينات.



شكل توضيحى (4) لخوارزمية SVM

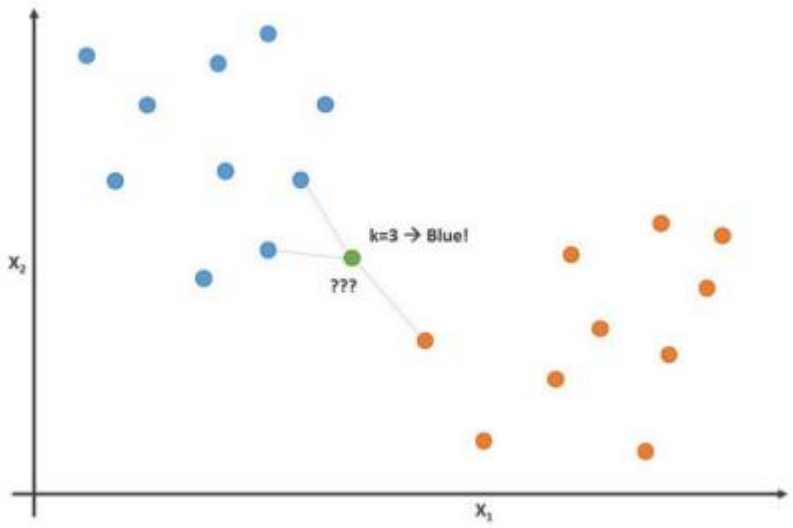
6) **بايز Naive Bayes**

هي طريقة تصنيف تعتمد علي نظرية بايز والافتراض المستقل للظروف المميزة وتمتاز **Naive Bayes** بأنها نشأت من النظرية الكلاسيكية ولها كفاءة تصنيف ثابتة، وتؤدي أداءً جيدًا للبيانات الصغيرة، يمكنها التعامل مع مهام متعددة التصنيفات، وهى خوارزمية مناسبة للتدريب المتزايد خاصةً عندما يتجاوز حجم البيانات الذاكرة، وليست حساسة جدًا للبيانات المفقودة(احمد، 2018)، خوارزمية بسيطة نسبيًا، كما أنها غالبًا ما تستخدم لتصنيف النص

7) **الجار الأقرب K-Nearest Neighbor (KNN)**

هي تقنية تعلم آلي تستخدم لتصنيف العينات الجديدة. تعتمد هذه التقنية على اقتراب العينة الجديدة من أقرب جيرانها في مساحة الخصائص.

وهي من خوارزميات التصنيف والتنبؤ التي تهدف للتنبؤ عن طريق مقارنة السجلات الشبيهة بالسجل المراد التنبؤ له وتقدير القيمة المجهولة لهذا السجل بناء علي معلومات لتلك السجلات. ومن العمليات التي يمكن أن تساعد في تطوير وزيادة فاعلية خوارزمية الجار الأقرب هو الأخذ بعين الاعتبار عدد أكبر من الجوار في محيط السجل المراد استكشافه Nabipour (2020).



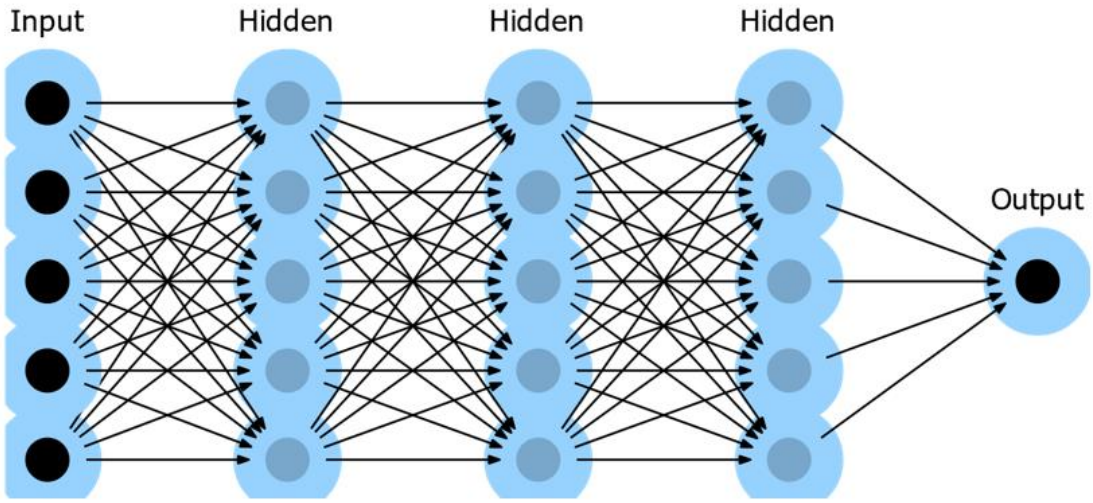
شكل توضيحى (5) لخوارزمية KNN

8) الشبكات العصبية العميقة Deep neural networks (DNN)

احد خوارزميات التعلم العميق فالشبكات العصبية Neural Networks هي نموذج رياضي يستند إلى عمل الدماغ البشري، حيث تعتبر من أهم التقنيات المستخدمة في مجال الذكاء الاصطناعي. وتتكون الشبكات العصبية من مجموعة من العقد (neurons)الاصطناعية التي تقوم بمعالجة الإشارات الواردة إليها، ويتم توصيل هذه العقد بطريقة معينة تسمى الهيكل العصبي Neural Network Architecture.،وتتميز الشبكات العصبية بقدرتها على التعلم من الأخطاء، حيث يتم تدريب الشبكة العصبية على مجموعة من البيانات المدخلة Input Data ، والمخرجات المطلوبة Output Data ، ويتم تحسين أداء الشبكة العصبية بتكرار عملية التدريب حتى يتم الحصول على أفضل النتائج.

ويتم استخدام الشبكات العصبية في عدة تطبيقات مثل التعرف على الصور والفيديو والأصوات، والتنبؤ بالمعلومات، وتصنيف البيانات، وتحليل النصوص، وتحسين أداء الروبوتات والمتحكمات الآلية، وتحسين أداء النظم الهيدروليكية والكهربائية والإلكترونية، وغيرها من التطبيقات (Zhao. et al. 2017) .

ويتم تصميم هذه الشبكات بشكل متنوع، وتختلف في هيكلها العام وعدد طبقاتها وطريقة توصيل العقد بين الطبقات، وكل نوع من هذه الشبكات يستخدم لحل مشكلات محددة. ويعتمد تصميم الشبكات العصبية على فهم البيانات المدخلة والمتوفرة وتحديد المخرجات المطلوبة وطبيعتها ومن بين خوارزميات التعلم العميق الشائعة الشبكات العصبية المتعددة Multilayer Perceptron (MLP):وتتألف هذه الشبكات من طبقات متعددة من العقد(neurons) ، وتتميز بقدرتها على تحليل المتغيرات المعقدة والتنبؤ بالبيانات المستقبلية.



شكل توضيحى (6) لخوارزمية MLP

معايير المقارنة بين خوارزميات التنبؤ:

- **دقة التصنيف Classification Accuracy**

هي نسبة التوقعات التي حصل عليها النموذج بشكل صحيح بالنسبة للعدد الكلى (Nicolas et al 2016) ، حيث أن:

$$\text{الدقة} = \frac{\text{عدد التنبؤات الصحيحة}}{\text{العدد الإجمالي للتنبؤات}}$$

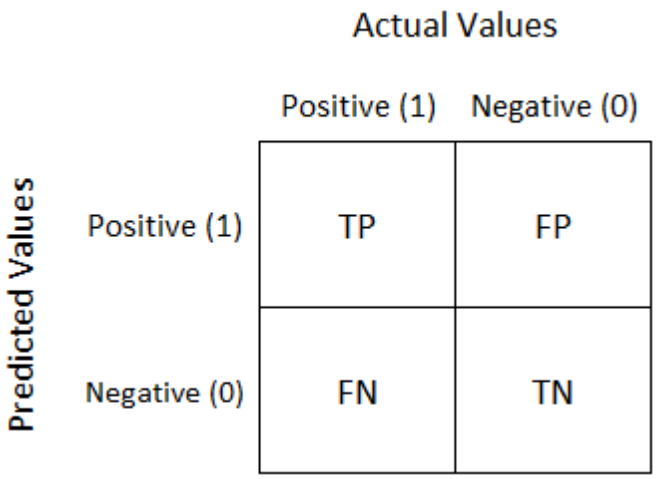
وتأخذ شكل المعادلة التالية:

$$\textbf{Classification } \boldsymbol{Accuracy} = \frac{\boldsymbol{TP + TN}}{\boldsymbol{TP + TN + FP + FN}}$$

حيث أن:

- $\boldsymbol{TP}$: توقع صحيح إيجابي **True Positive**
- $\boldsymbol{TN}$: توقع صحيح سلبي **True Negative**
- $\boldsymbol{FP}$: توقع خطأ إيجابي **False Positive**
- $\boldsymbol{FN}$: توقع خطأ سلبي **False Negative**

ويتم تطبيق المعادلة السابقة على **"مصفوفة الشك Confusion Matrix"**



رسم توضيحى لمصفوفة الشك Confusion Matrix

- $P(e)$: تشير إلي نسبة عدد الحالات التصادفية.

- **Recall**

هو أحد مقاييس التقييم الأكثر استخدامًا لمجموعة **"البيانات غير المتوازنة"**، ويحسب عدد الإجابات الصحيحة الفعلية التي توقعها المصنف على أنها صحيحة.

ويطبق من خلال المعادلة التالية:

$$Recall = \frac{TP}{TP + FN}$$

- **Precision**

يصف دقة نموذج استخراج البيانات، اى من بين تلك الحالات المتوقعة الايجابية، كم منها إيجابي بالفعل؟

ويطبق من خلال المعادلة التالية:

$$Precision = \frac{TP}{TP + FP}$$

وقد تم تحديد دقة مقارنة نماذج التنبؤ بعد تدريبها بالاعتماد على أسلوب الاختبار **(Validation Cross)** حيث تم استخدام نسبة **80%** من البيانات للتدريب ونسبة **20%** للاختبار بنسبة تقسيم متساوية **10** وذلك لجميع النماذج

- **المساحة تحت المنحنى (Area Under the Curve - AUC)**

يُعتبر AUC مقياسًا هامًا لقياس دقة التنبؤ وتقييم أداء النماذج في المشكلات ذات الطبيعة الثنائية، ويستخدم AUC لتقييم أداء نماذج التصنيف الثنائي عند استخدام تقنيات التصنيف الثنائي، ويتم إنشاء منحنى الاستجابة (ROC curve) الذي يرتبط بتصنيف البيانات بين فئتين. وتحتوي المساحة تحت المنحنى على قيمة تمثل قدرة النموذج على التمييز بين الفئتين المختلفتين،

وتتراوح قيمة AUC بين 0 و1، حيث يكون AUC = 1 للنموذج المثالي الذي يمكنه تمامًا التمييز بين الفئتين، وAUC = 0.5 للنموذج الذي يصنف بشكل عشوائي.

البيانات

تتكون البيانات من 22 متغير،13 من هذه المتغيرات لها قيم محدد وتختلف من عميل لأخر، 5 متغيرات ثنائية لها قيمتان "1 " و "0" و3 متغيرات تأخذ قيم رقمية والمتغير" CHURN" وهو المتغير التابع.Variable Target والجدول التالي يوضح ذلك

جدول (1) متغيرات الدراسة

مسلسل	Variables	Content	Information
.1	Unnamed	5986	غير معرف.
.2	Customer ID	5986	رقم تعريف للعميل.
.3	Gender	Male Female	جنس العميل.
.4	Senior Citizen	No Yes	سواء كان العميل متقاعدا أم لا
.5	Partner	No Yes	ما إذا كان العميل متزوجا.
.6	Dependents	0 1	هل العملاء مستقلين ماديا.
.7	Tenure	From 0 To 72	مدة الخدمة: وهي عدد الأشهر التي كان بها العميل عميلاً للشركة.
.8	Phone Service	No Yes	هل الخدمة الهاتفية مفعلة.
.9	Multiple Lines	No Yes	ما إذا كانت خطوط الهاتف المتعددة.
.10	Internet Service	DSL Fiber Optic No	مزود الإنترنت للعميل.
.11	Online Security	No No Internet Service Yes	هل تم تمكين خدمة الأمان عبر الإنترنت.
.12	Online Backup	No No Internet Service Yes	هل خدمة النسخ الاحتياطي عبر الإنترنت مفعلة.
.13	Device Protection	No No Internet Service Yes	هل لدي العميل تأمين على المعدات.
.14	Tech Support	No No Internet Service Yes	هل خدمة الدعم الفني مفعلة.
.15	Steaming TV	No No Internet Service Yes	هل خدمة البث التلفزيوني مفعلة.
.16	Steaming Movies	No No Internet Service Yes	هل تم تنشيط خدمة السينما والأفلام.
.17	Contract	Month-to-month One year Two years	نوع عقد العميل.
.18	Paperless Billing	No Yes	ما إذا كان العميل يستخدم الفواتير غير الورقية.

.19	Payment Method	Bank Transfer (Automatic) Credit Card (Automatic) Electronic Check Mailed Check	طريقة الدفع.
.20	Monthly Charges	From 18.25 To 118.75	الدفعة الشهرية الحالية.
.21	Total Charges	From 18.8 To 8684.8	المبلغ الإجمالي الذي دفعه العميل مقابل الخدمات طوال الوقت.
.22	CHURN	No Yes	هل هناك تسرب بالزبائن أم لا.

المصدر: موقع Kaggle

التحليل الاحصائي

أولاً: تحضير البيانات

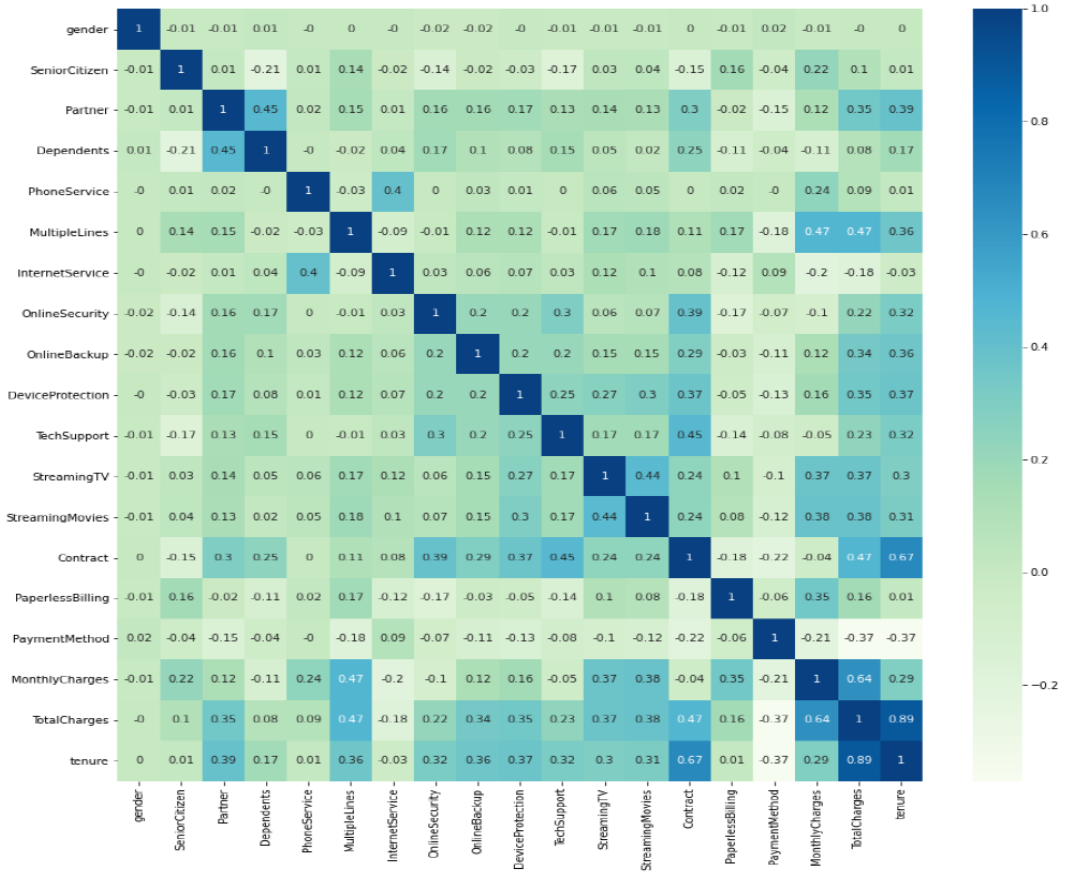
قبل البدء بعملية التحليل الاحصائي لابد من تنقية البيانات وذلك لان التضخم الكبير الذي أصبحت عليه قواعد البيانات في هذا العصر يجعلها عرضة لاحتواء الكثير من البيانات الغير متناسقة أو حتى فقد بعض البيانات الموجودة فيها، وذلك بسبب ضخامتها وتدفقها من مصادر متعددة .كما أن البيانات ذات الجودة المنخفضة سوف تؤدي بطبيعة الحال إلى عدة نتائج بجودة منخفضة أيضا تحليلها والتنقيب فيها .ومنً أجل ذلك ينبغي رفع جودة البيانات أولا وبالتالي يمكن أن نتوقع ارتفاع كفاءة التحليل والتنقيب فيها وتيسير عملياتها لتكون بالشكل الأمثل

1. احتوت البيانات على متغيرين (2 Variables) كل قيمة فيهما هي تعتبر قيمة فريدة لا تعطي أي نوع من المعلومات، ولا يمثلان أي إضافة للنموذج، أو أي مساعدة مهمة في التنبؤ، بل والعكس، سيكونون بمثابة تشويش لنتائج البيانات المتغير الأول (unname) والمتغير الثانى costumer ID) ليصبح عدد متغيرات البيانات.(20 Variables)
2. تم اختبار درجة اتساق البيانات داخل نفس مجموعة البيانات أو عبر مجموعات بيانات متعددة، للتأكد من خلوها من التعارض وتم التأكد من أن كل متغير لا يحتوي على أي قيم مشابهة أو متشابهة ولكن يمكن تصنيفها بشكل منفصل، والتي قد تكون بسبب أخطاء املائية أو القيم التي تتم كتابتها بشكل مختلف .و نتيجة لذلك رصد تعارض المتغير (Citizen Senior) مع اتساق باقي المتغيرات، و تم تحويلها من متغير ثنائي (0,1) إلى متغير فئوي(No , Yes) ، ليصبح باتساق المتغيرات الاخرى، و كتصنيف القيم في كل من المتغيرات الفئوية الاخرى
3. values Missing تم التدقيق في مجموعات البيانات للتأكد من خلوها من القيم المفقودة وتم اكتشاف 22 صف للبيانات كانت مجرد مسافات فارغة، في قيم المتغير (Charge Total) وقد تم حذف الصفوف التي تحتوى على بيانات مفقودة وهى 10 صفوف وهى كمية صغيرة جداً مقارنة بعدد الصفوف في مجموعة البيانات

ثانياً: تحليل مدى الارتباط

للتنبؤ بخروج العميل من الشركة فلابد أولا من معرفة العلاقات بين المتغيرات المختلفة، مع متغير الهدف وهو خروج العميل لمعرفة الاسباب التي تؤدي إلى خروجه، ولمعرفة علاقة الارتباط بين العوامل التي قد تؤثر في قرار العميل في ترك شركة الاتصالات أو في البقاء فيها، لذا تم تحويل جميع البيانات من بيانات اسمية، إلى بيانات رقمية، وحساب مصفوفة الارتباط بين متغيرات الدراسة، كانت النتائج كما يلي:

مصفوفة الارتباط بين متغيرات الدراسة

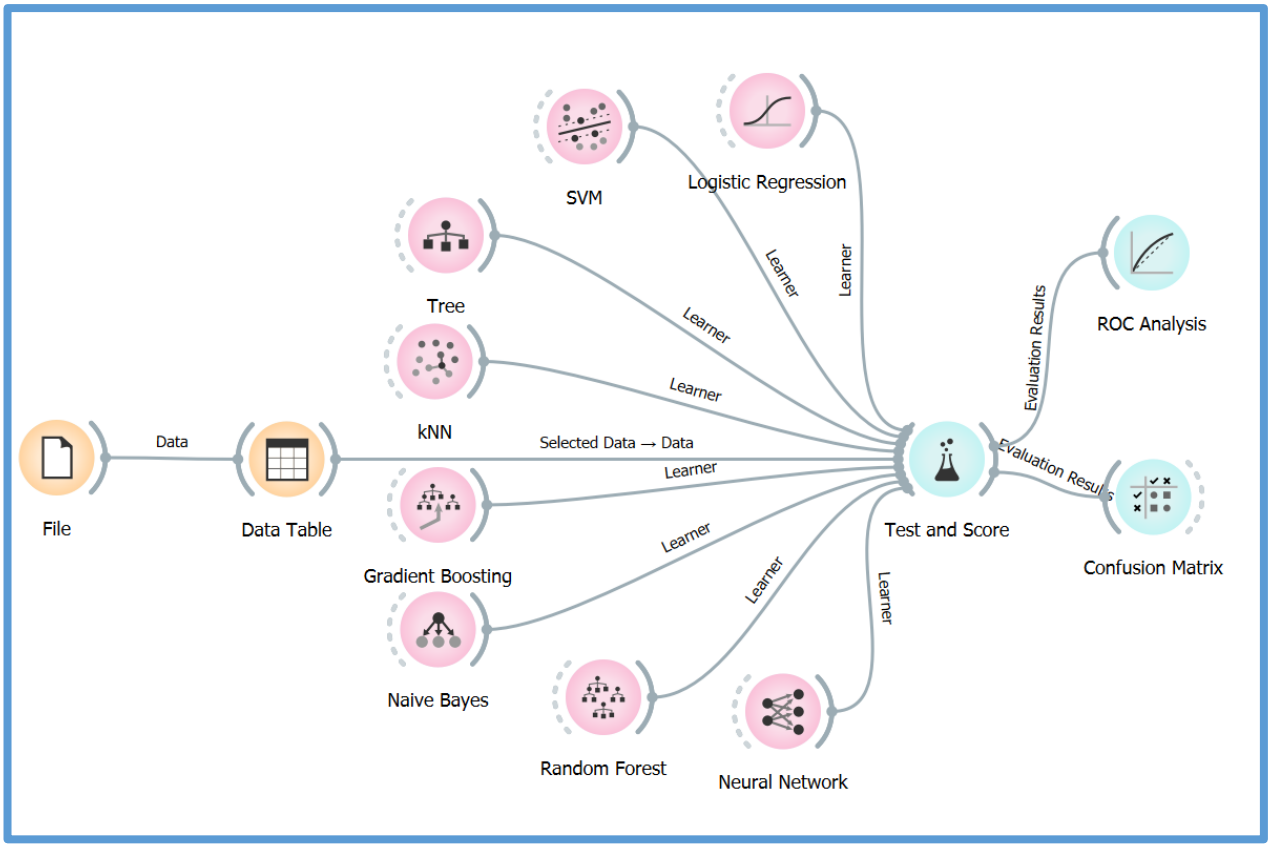


المصدر: مخرجات برنامج Orange

ومن خلال مصفوفة الارتباط نلاحظ ان معظم المتغيرات الفئوية الأخرى لديها نوع من العلاقة مع المتغيرات الفئوية الأخرى، على الرغم من ضعف هذه العلاقات، كما نلاحظ أيضا وكما هو متوقع ليس للجنس أي ارتباط مع أي من المتغيرات الفئوية الأخرى، ونلاحظ وجود علاقة قوية للعقد Contract مع مدة العقد .Tenure وايضاً وجود ارتباط قوى بين متغير الرسوم الاجمالية (Charge Total) وبين الرسوم الشهرية (Charge Monthly) والمدة (Tenure) لذا يجب أن نحذف المتغيرين(Tenure&Charge Total) من قائمة المتغيرات المتماثلة، ليصبح عدد المتغيرات 18 متغير.

ثالثاً: بناء النموذج

وقد تم بناء نموذج تنبؤي من خلال استخدام خوارزميات التصنيف التي تم توضيحها من قبل وتم استخدام برنامج أورانج (Orange) لتنفيذ وبناء النموذج التنبؤي، وقد تم الاعتماد على أسلوب الاختبار(Cross Validation) بنسبة تقسيم 10 لجميع خوارزميات التنبؤ والشكل التالي يوضح شكل النماذج على برنامج Orange.



المصدر: مخرجات برنامج Orange

رابعاً: المقارنة واختيار الخوارزمية الأفضل

- بعد تطبيق خوارزميات التنبؤ تمت المقارنة بينهم لاختيار المصنف الأفضل النهائي، وتم ذلك من خلال مقارنة (Accuracy،Recall، AUC، Precision) الخاصة بكل خوارزمية، والجدول التالى يوضح نتائج المقارنة Nicolas:

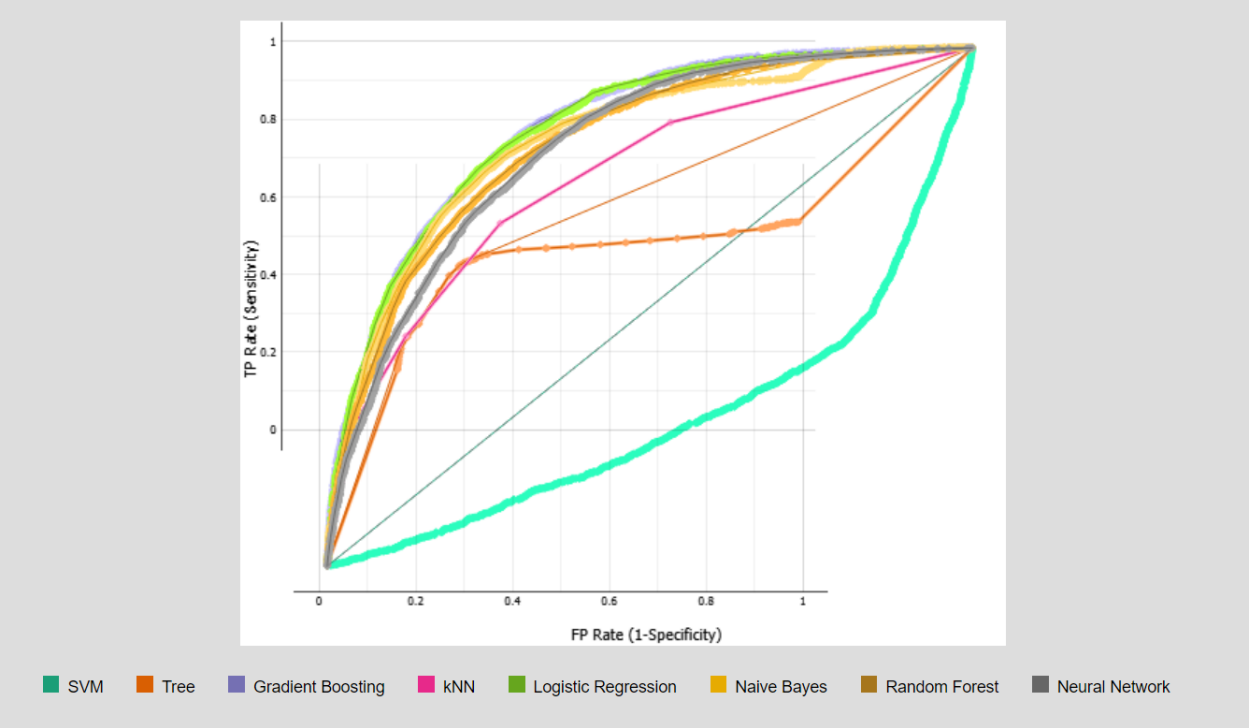
- **جدول (2) يوضح المقارنات بين المصنفات من حيث الدقة فى التصنيف**

Scores

Model	AUC	CA	Prec	Recall
Logistic Regression	0.844	0.803	0.795	0.803
Gradient Boosting	0.847	0.804	0.795	0.804
Random Forest	0.814	0.784	0.773	0.784
Neural Network	0.811	0.777	0.770	0.777
Naive Bayes	0.822	0.734	0.799	0.734
kNN	0.744	0.758	0.743	0.758
SVM	0.283	0.740	0.749	0.740
Tree	0.637	0.747	0.737	0.747

المصدر: مخرجات برنامج Orange

كما هو واضح بالجدول السابق فإن خوارزمية تعزيز التدرج Gradient Boosting Algorithm تعطي تفوقًا على جميع المقاييس فلو أردنا المقارنة بناء على الدقة تكون هي الأعلى، ليكون هي النموذج الأفضل بدرجة دقة 84.7% كما هي الأعلى على جميع المعايير الاخرى Precision و Recall والشكل التالي يوضح منحنى AUC للخوارزميات والذى يوضح ايضا دقة خوارزمية Gradient Boosting Algorithm



شكل (7) يوضح المساحة تحت المنحنى AUC لجميع الخوارزميات

خامساً: النتائج

1. تقدم تقنيات تعلم الالة والتعلم العميق إمكانات واسعة لتطبيقات الذكاء الصناعي لشتى أنواع البيانات ومن بينها البيانات عن عملاء شركة اتصالات، مما يسمح لنا بفهم أوسع لسلوكيات العملاء، والقيام بخطوات استباقية للحفاظ على العملاء.
2. النوع متغير ليس له اى ارتباط مع أي من المتغيرات الفئوية الأخرى
3. وجود علاقة قوية بين متغير العقد ومدة العقد.
4. متغير الرسوم الاجمالية (Charge Total) بجانب خدمة الأمان، النسخ الاحتياطي، حماية الجهاز، والدعم الفني، اكثر المتغيرات تاثيراً في خروج العملاء
5. تعتمد دقة نتائج خوارزميات تعلم الالة والتعلم العميق في البيانات بصورة أساسية على مدى دقة وشمولية البيانات المستخدمة.
6. يمثل برنامج (orange) أداة فعالة للاستفادة من تقنيات الذكاء الاصطناعي من خلال ما يمنحه التطبيق من خيارات وإمكانيات متنوعة، حيث يسمح التطبيق بتجهيز البيانات وفق صيغ متعددة
7. إن عملية تحضير البيانات قبل القيام بعملية التنبؤ، تعد خطوة جوهرية وأساسية في عملية التنبؤ حيث أن بناء نموذج من دون تحضير البيانات يعطي نتائج غير دقيقة.
8. .إن استخدام الخوارزميات الشجرية يعطي أفضل النتائج في التصنيف، والتي يمكن أن تطبق على البيانات، والتي كان أفضلها خوارزمية تعزيز التدرج Gradient Boosting Algorithm

سادساً: التوصيات:

1. إعطاء خوارزميات تعلم الالة والتعلم العميق في البيانات أهمية أكبر داخل الشركات لما تحمله من أدوات تساعد على معرفة نتائج ومعلومات كان من الصعب الوصول اليها سابقا
2. الاهتمام باستمرارية السعي بجمع المعلومات والبيانات عن العملاء لما له من أهمية للتنبؤ بخروجهم.
3. الاهتمام أكثر بالتسويق لخدمة الأمان، النسخ الاحتياطي، حماية الجهاز، والدعم الفني، لما لهم من تأثير في خروج العملاء.

4. لا يوجد خوارزمية واحدة تكون الأكثر فعالية في تنبؤ خروج العملاء من الشركات الخدمية، وذلك يعتمد على البيانات المتاحة والمشكلة المحددة.
5. الاستفادة من أدوات أخرى للتنبؤ في البيانات، لمعالجة هذا النوع من البيانات، وعلى وجوه الخصوص استخدام خوارزميات وأدوات أخرى غير التصنيف، للوصول الى افضل النتائج في تحليل البيانات
6. الاستفادة في الدراسات المستقبلية من استخدام عملية دمج مجموعة من الخوارزميات معًا لتحقيق أفضل النتائج

المراجع

المراجع العربية

1) عبد الرحيم احمد.(2018). توقع تسرب الزبائن في شركات الاتصالات باستخدام تعلم الآلة في بيئة المعطيات الكبيرة

2) راكان رزوق. (2013) التنقيب في البيانات الأسس النظرية والتطبيقية.

3) عبد الرحيم احمد (2018) توقع تسرب الزبائن في شركات الاتصالات باستخدام تعلم الالة في بيئة المعطيات الكبيرة

4) محسن حمود (2020) التمثيل المرئي للبيانات data science .

5) محمد دعيش، و محمد ساري (2017) نموذج الانحدار اللوجستي: مفهومه، خصائصه، تطبيقاته

6) محمد مصطفى عبيد (2018) كتاب التحليل المتقدم وتنقيب البيانات. القاهرة: دار الفكر العربي.

المراجع الأجنبية

7. Alharbi, M. G., and Khalifa, H.A. (2021). Enhanced fuzzy Delphi method in forecasting and decision-making. Advances in Fuzzy Systems (Hindawi), Vol.2021, Article ID 2459573, 6 pages.

8. Ahola, j., & Rinta, R. (2001). *Data mining case studies in coustomer profiling*
9. Brian beers (2021).what is the telecommunications sector .investopedia.
10. Candice, Bentéjac., Anna, Csörgő., Gonzalo, Martínez-Muñoz. (2021). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3):1937-1967
11. Chris Smith .(2017).Decision Trees and Random Forests.
12. Dalli, A. (2022). Impact of hyperparameters on deep learning model for customer churn prediction in telecommunication sector. Mathematical Problems in Engineering.
13. Data mining tutorial (2020)..EDucba.
14. Dost, Muhammad, Khan., Tariq, Aziz, Rao., Faisal, Shahzad. (2019). The Classification of Customers' Sentiment using Data Mining Approaches. 4(4):146-156. doi: 10.31703/GSSR.2019(IV-IV).19
15. Dean, J. (2014). Big Data, Data Mining, and Machine Learning.

16. Fujo, S. W., Subramanian, S., & Khder, M. A. (2022). Customer churn prediction in telecommunication industry using deep learning. Information Sciences Letters, 11(1), 24.
17. Homa, Meghyasi., Abas, Rad. (2020). Customer Churn Prediction in Irancell Company Using Data Mining Methods.
18. Kelleherd, J., & Tierney, B. (2018). DATA SCIENCE.
19. Mahmoud, Ewieda., Essam, M, Shaaban., Mohamed, Roushdy. (2021). Customer Retention: Detecting Churners in Telecoms Industry using Data Mining Techniques. International Journal of Advanced Computer Science and Applications.
20. Nabipour, M., Nayyeri, P., Jabani, H., S, S., & Mosavi, A. (2020). Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis. IEEE Access, 8,pp. 150199-150212
21. Nicolas, Auger., Cyril, Nicaud., Carine, Pivoteau. (2016). Good predictions are worth a few comparisons. 47:14-. doi: 10.4230/LIPICS.STACS.2016.12
22. Reza, M. M., Nahar, S., & Akter, T. (2018). Segmentation of Mobile Customers using Data Mining Techniques. International Journal of Engineering and Technical Research.
23. Shreyas, Rajesh, Labhsetwar. (2020). Predictive analysis of customer churn in telecom industry using supervised learning.
24. Chris Smith (2017) .Decision Trees and Random Forests .
25. V., Kavitha., G., Hemanth, Kumar., S., V, Mohan, Kumar., M, R, Harish. (2020). Churn Prediction of Customer in Telecom Industry using Machine Learning Algorithms. International Journal of Engineering Research.
26. Zhang, S., Omar, A. H., Hashim, A.S., Alam, T., Khalifa, H.A., and Elkotb, M. A. (2023). Enhansing groundwater management and prediction of water quality in the urban environment using optimized algorithm of least square SVM.. Urban Climate, (49): 101487
27. Zhao Y, Li J, Yu L.(2017) A deep learning ensemble approach for crude oil price forecasting. Energy Economics. 66:9-16.