التنبؤ بأسعار السيارات باستخدام تقنيات التنقيب في البيانات
- دراسة تطبيقية على بيانات مجموعة من السيارات بالمملكة المتحدة
Car Prices Using Data Mining Techniques - An applied study on a UK cars dataset -

توفيق بنين[1] *، مصطفى طويطي[2]، ذهيبة بن عبد الرحمان[3]

[1] مخبر التطبيقات الكمية والنوعية للارتقاء الاقتصادي والاجتماعي والبيئي بالمؤسسات الجزائرية، جامعة غرداية (الجزائر) (Benine.toufik@univ-ghardaia.dz)

[2] مخبر التطبيقات الكمية والنوعية للارتقاء الاقتصادي والاجتماعي والبيئي بالمؤسسات الجزائرية، جامعة غرداية (الجزائر) (Touaiti.mustapha@univ-ghardaia.dz)

[3] مخبر التطبيقات الكمية والنوعية للارتقاء الاقتصادي والاجتماعي والبيئي بالمؤسسات الجزائرية، جامعة غرداية (الجزائر) (Benabderrahmane.dehiba@univ-ghardaia.dz)

تاريخ الاستلام: 2025/03/20؛ **تاريخ القبول**: 2025/05/01؛ **تاريخ النشر**: 2025/06/01

ملخص: تهدف الدراسة إلى فحص فعالية خوارزمية XGBoost الانحدارية ومقارنتها بخوارزمية الغابة العشوائيةRF ، خوارزمية آلة دعم المتجه الانحدارية SVR وخوارزمية الشبكات العصبية العميقة DNN للتنبؤ بأسعار السيارات في المملكة المتحدة، توصلت الدراسة إلى أن خوارزمية الغابات العشوائية (RF) ملائمة ومناسبة لتقدير أسعار السيارات بدقة، وبالتالي يمكن الاعتماد عليها في تحسين وترشيد قرارات مشتري وبائعي السيارات، وذلك لكونها حققت أكبر معامل تحديد R^2 بقيمة 95.90 % وأقل قيمة للجذر التربيعي لمتوسط مربع الأخطاء RMSE والتي قدرت بـــ 1946.07 مقارنة بخوارزمية XGBoost وخوارزمية آلة دعم المتجه الإنحدارية(SVR) والشبكة العصبية العميقة (DNN).

الكلمات المفتاح : تنقيب في البيانات؛ خوارزمية XG Boost الانحدارية؛ خوارزمية الغابات العشوائية RF؛ خوارزمية آلة دعم المتجه الانحدارية SVR؛ خوارزمية الشبكات العصبية العميقة DNN.

تصنيف JEL : C8؛ C88؛ D81

Abstract: The study aims to examine the effectiveness of the XGBoost regression algorithm and compare it with the Random Forest (RF) algorithm, the Support Vector regression (SVR) algorithm, and the Deep Neural Networks (DNN) algorithm to predict car prices in the United Kingdom, the study found that the Random Forest (RF) algorithm is appropriate and effective for accurately estimating car prices, and therefore can be relied upon to improve and rationalize decisions of car buyers and seller, This conclusion is based on the RF algorithm achieving the highest coefficient of determination (R^2) of 95.90 % and the lowest Root Mean Squared Error (RMSE) of 1946.07, compared to the XGBoost regression algorithm, the Support Vector regression (SVR) algorithm, and the Deep Neural Networks (DNN) algorithm.

Keywords: Data mining; XGBoost regression algorithm; Random Forest (RF) algorithm; Support Vector regression (SVR) algorithm; Deep Neural Networks (DNN) algorithm.

Jel Classification Codes : C8; C88; D81.

* المؤلف المرسل.

I– تمهيد :

تعد صناعة السيارات من الركائز الأساسية للاقتصادات الحديثة، نظرا لدورها المحوري في دعم الناتج المحلي الإجمالي، توفير فرص العمل، تعزيز الابتكار والاستدامة. كما تؤثر بشكل مباشر وغير مباشر على الديناميكية الاقتصادية، من خلال مدخلي الاقتصاد الكلي والجزئي.

ويشكل قرار التسعير الدقيق استراتيجية حاسمة بالنسبة لمصنعي السيارات، التجار والمستثمرين في هذا القطاع، إذ يُسهم في تحسين القدرة التنافسية وتحقيق هوامش ربح ملائمة، فضلا عن انعكاسه المباشر على سلوك المستهلكين وقراراتهم الشرائية.

ومع تطور الأساليب الكمية والتقنيات الحديثة في معالجة البيانات، باتت عملية اتخاذ القرار قابلة للتحسين من خلال الاستفادة من تقنيات الذكاء الاصطناعي والتعلم الآلي. وتعد تقنيات التنقيب في البيانات من بين الأدوات الفعّالة التي يعتمد عليها المحللون الاقتصاديون ومهندسو البيانات لاستخلاص رؤى دقيقة وتوقعات قيّمة.

سنقوم في هذه الدراسة باختبار فعالية خوارزمية XGBoost الانحدارية، ومقارنتها بكل من خوارزمية الغابات العشوائيةRF، خوارزمية آلة دعم المتجهSVR وخوارزمية الشبكات العصبية العميقةDNN، وذلك بغرض التنبؤ بأسعار السيارات وتحديد المتغيرات الأكثر تأثيرا في تشكيل تلك الأسعار.

I–1–إشكالية الدراسة: ومن خلال ما سبق يمكن طرح الإشكالية التالية: ماهي أفضل خوارزمية للتنبؤ بأسعار السيارات؟

I–2–الفرضيات: للإجابة على إشكالية الدراسة، تم اقتراح الفرضيات التالية:

- خوارزمية XGBoost الانحدارية تتفوق على خوارزمية الغابة العشوائيةRF من حيث دقة التنبؤ؛
- خوارزمية XGBoost الانحدارية تتفوق على خوارزمية آلة دعم المتجه الانحداريةSVR وخوارزمية الشبكات العصبية العميقة DNN من حيث دقة التنبؤ؛
- المتغيرات مثل شكل السيارة، نوعها، سنة الصنع، الوقود المستهلك، نوع علبة السرعات، حجم المحرك وعدد الأميال المقطوعة تؤثر بشكل كبير في تحديد سعر السيارة.

I–3–أهداف الدراسة: تهدف هذه الدراسة إلى استكشاف أفضل خوارزمية لعملية التنبؤ بأسعار السيارات من بين الخوارزميات الآتية: XGBoost الإنحدارية، خوارزمية الغابة العشوائية RF، خوارزمية آلة دعم المتجه الانحدارية SVR وخوارزمية الشبكات العصبية العميقة DNN، ومقارنة أدائهن مع تحديد أهم المتغيرات التي تؤثر في أسعار السيارات وبالتالي التأثير على قرارات مشتري أو بائعي السيارات.

I–4–أهمية الدراسة: تكمن أهمية الدراسة في استخدام التنقيب في البيانات كأحد الطرق الحديثة التي يمكن الإعتماد عليها في ترشيد القرارات.

I–5–منهج الدراسة والأدوات المستخدمة: اعتمدنا في الدراسة على المنهج الوصفي في الجانب النظري لتوضيح بعض المفاهيم والأدوات الخاصة بالتنقيب في البيانات، بالإضافة إلى المنهج الكمي في الجانب التطبيقي من أجل التنبؤ بأسعار السيارات باستخدام خوارزمية XGBoost الانحدارية، خوارزمية آلة دعم المتجه الانحدارية SVR، خوارزمية الغابة العشوائيةRF وخوارزمية الشبكات العصبية العميقة DNN. واستخدمنا لغة البرمجة بايثون لتطبيق العمل المطلوب من خلال بيئة العمل google colab.

I–6–الدراسات السابقة: قدمت الدراسات السابقة مجموعة من الفوائد في تحديد أهم المتغيرات المؤثرة والمتنوعة في تحديد أسعار السيارات باستخدام العديد من الخوارزميات المعتمدة في التنقيب في البيانات والتي يمكن عرضها كما يلي:

–دراسة (MUTİ & YILDIZ, 2023) بعنوان: **"Using Linear Regression For Used Car PricePrediction "**، هدفت الدراسة إلى فحص فعالية نموذج الانحدار الخطي للتنبؤ بأسعار السيارات المستعملة حيث تم استخدام بيانات سيارات مستعملة في تركيا لسنة 2020 وتحتوي البيانات على 15 متغير يفترض بأنها تؤثر على سعر السيارات المستعملة (نوع السيارة، الطراز، سنة الصنع، قوة المحرك، نوع الهيكل، المسافة المقطوعة، نوع علبة السرعات.....)، وتمت معالجة مجموعة البيانات مسبقًا لإزالة القيم غير المنطقية وإزالة كل المتغيرات التي تحتوي على عدد كبير من المتغيرات المفقودة، ثم تدريب النموذج باستخدام 70% من مجموعة البيانات واختباره باستخدام 30% من مجموعة البيانات. توصلت الدراسة إلى أن خوارزمية الانحدار الخطي ملائمة ومناسبة لتقدير أسعار السيارات المستعملة بناءا على قيمة معامل التحديد (R^2 = 73%).

–دراسة (Bukvić, Pašagić, Škrinjar, Fratrović, & Abramović, 2022) بعنوان: **"Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning"**، هدفت الدراسة إلى التنبؤ بأسعار المركبات المستعملة في السوق الكرواتية باستخدام تقنيات التنقيب عن البيانات والتعلم الآلي الخاضع للإشراف واستكشاف العلاقة بين السمات مثل سنة الإنتاج والكيلومترات المقطوعة للتنبؤ باتجاهات الأسعار. من خلال استخدام بيانات مجمعة من البائع "Njuškalo" عبر الإنترنت حيث تضمنت سمات مثل الشركة المصنعة، الطراز، سنة التصنيع، الكيلومترات المقطوعة وسعر بيع المركبات. وتم استخدام SAS Studio

لوصف وتحليل المتغيرات وMatlab لتدريب النموذج والتحقق من نتائج التنبؤ. وتوصلت الدراسة إلى أن دقة التنبؤ بالنموذج زادت مع التدريب، مما حقق دقة تنبؤ بنسبة 95٪ مع الانحدار الخطي.

-**دراسة** (Chen , Hao, & Xu, 2017) بعنوان: **"Comparative Analysis of Used Car Price Evaluation Models"**، هدفت الدراسة إلى مقارنة أداء خوارزمية الانحدار الخطي وخوارزمية الغابة العشوائية في التنبؤ بأسعار السيارات المستعملة في ظل تعقيدات الطراز المختلفة، من أجل الوصول إلى النموذج الأمثل تم استخدام 100000 سجل لسيارات مستعملة في جميع أنحاء الصين، تم جمعها من العديد من أسواق السيارات في شنغهاي وموقع ويب لمزاد السيارات المستعملة على مستوى البلاد، يغطي الصفقات بين أكتوبر 2015 وأكتوبر 2016. تم في هذه الدراسة بناء ثلاثة: النموذج الأول بسيط عن طريق استخدام عينة بيانات لسيارات من نفس الطراز ومتغيرات محددة، النموذج الثاني متوسط التعقيد تم بناءه باستخدام عينة بيانات لسلسلة سيارات محددة ومتغيرات محددة والنموذج الثالث (سمي بالنموذج العالمي) تم الوصول لهذا النموذج عن طريق استخدام كل البيانات المجمعة وكل المتغيرات الخاصة بالسيارات (19 متغيرًا، 102959 سيارة). توصل المقال إلى أن خوارزمية الغابة العشوائية كانت جيدة في جميع النماذج وكانت هي الأمثل للنموذج الثالث والمعقد (النموذج العالمي) الذي يمكنه تفسير 95.06٪ من تغيرات الأسعار، أما الانحدار الخطي تحسنت دقته عند زيادة حجم العينات في النماذج البسيطة

I-7-الإطار النظري للتنقيب في البيانات: سنتناول في هذا الجزء من الدراسة مفهوم التنقيب في البيانات، أنواعه ومراحله.

I-7-1-مفهوم التنقيب في البيانات: هناك العديد من الدراسات التي حاولت تعريف التنقيب في البيانات نذكر منها ما يلي:

يعتبر التنقيب في البيانات عملية تستخلص المعرفة من كميات كبيرة من البيانات من خلال القيام بتحليل وتفسير البيانات لاكتشاف العلاقات والأنماط والاتجاهات غير المعروفة. وذلك عن طريق استخدام الذكاء الاصطناعي، التعلم الآلي، الإحصاء والرياضيات، وينتج عن التنقيب في البيانات رؤى وتنبؤات مستقبلية مساعدة في صنع القرار (العيفة، 2023، صفحة 94).

والتنقيب في البيانات هو مجال بحث وتطبيق متعدد التخصصات يهدف إلى اكتشاف معرفة جديدة ومفيدة من قواعد البيانات الضخمة، من خلال استخدام أساليب تتراوح بين الذكاء الاصطناعي والإحصاءات وقواعد البيانات (قوت و بلحاج، 2023، صفحة 64)

وهو عملية تحليلية لفحص مجموعات البيانات الموجودة في مستودعات البيانات للكشف عن رؤى محجوبة وصالحة وقابلة للتطبيق. وبالتالي استخراج المعرفة القيمة والكشف عن المعلومات الدقيقة والعملية والتي تمكننا من صنع قرارات فعالة (بوهرين و لكحل، 2023، صفحة 559).

ومنه يمكننا تعريف التنقيب في البيانات بأنه عملية تمكننا من استخراج معلومات ذات مغزى من مجموعة واسعة من البيانات باستخدام خوارزميات تعلم الآلة والذكاء الاصطناعي المعتمدة على الإحصاء والرياضيات من خلال الكشف عن العلاقات، الأنماط والاتجاهات غير المعروفة داخل البيانات واجراء تنبؤات قيّمة تساهم في صنع القرارات الرشيدة.

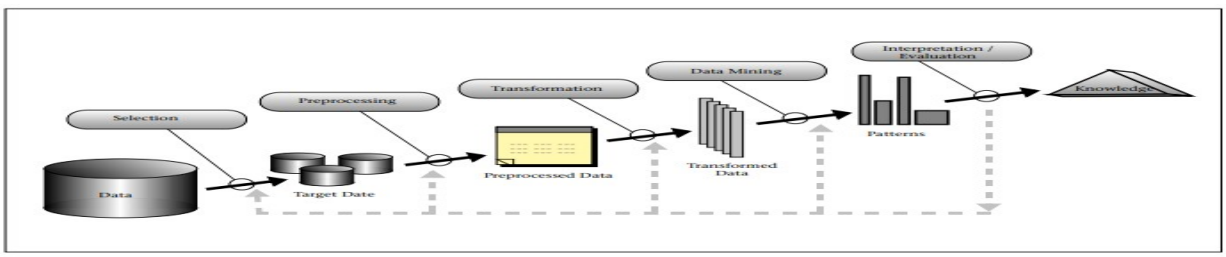
I-7-2-أنواع التنقيب في البيانات: يمكن تقسيم التنقيب في البيانات إلى قسمين كما يلي:

I-7-2-1-التنقيب الاستكشافي (الوصفي): يقدم نظرة عامة موجزة وشاملة لمجموعة من البيانات، ويسلط الضوء على ميزاتها دون تحقيق هدف محدد. حيث تركز هذه التقنية على فهم البنية الجوهرية للبيانات، العلاقات، والاتصالات فيما بينها. وتقوم بتحليل البيانات المتاحة لكشف كيفية ترابط العناصر داخلها، موفرة بذلك وصفا مفصلا للبيانات (Sumiran, 2018, p. 949).

I-7-2-2-التنقيب التنبؤي: ويقصد به الوصول إلى قيمة المتغير الناتج في المستقبل باستخدام ما حصل للمتغير في الماضي، أي تكوين نموذج تنبؤي لمتغير الهدف، مع واحد أو أكثر من المتغيرات المفسرة عن طريق استخدام معلومات معروفة في الماضي للوصول لما هو غير معروف في المستقبل. وتزداد الفاعلية ودقة النموذج التنبؤي كلما كانت البيانات نقية وموثوقة وذات حجم كبير (عمار عبد اللطيف، 2008، صفحة 310).

I-7-3-مراحل التنقيب في البيانات: يمكن تلخيص مراحل التنقيب في البيانات في الشكل التالي:

الشكل رقم 1: مراحل التنقيب في البيانات



المصدر: (Fayyad, Piatetsky-Shapiro, & Smyth , 1996, p. 41)

I-7-3-1-مرحلة جمع البيانات (Data collection): جمع البيانات أولى مراحل التنقيب في البيانات، وتتضمن تحديد وجمع البيانات المستهدفة الأولية التي سيتم استكشافها والتنقيب فيها(Fayyad, Piatetsky-Shapiro, & Smyth , 1996, p. 42) .

I-7-3-2-اختيار البيانات (Data Selection) وتكامل البيانات (Data integration): تتم في هاته الخطوة اختيار واستيراد عينة البيانات المراد التنقيب فيها واستخدامها في عملية التحليل (Fayyad, Piatetsky-Shapiro, & Smyth , 1996, p. 42). أما تكامل البيانات أي تجميعها في مصدر واحد نستعملها في حالة كانت هناك مصادر متعددة للبيانات (Sumiran, 2018, p. 948).

I-7-3-3-مرحلة تحضير البيانات (Data preprocessing): تسمى هذه المرحلة أيضا **تنظيف البيانات (Data Cleaning)** ويتم فيها المعالجة المسبقة للبيانات وتحضيرها للعمل عليها (Fayyad, Piatetsky-Shapiro, & Smyth , 1996, p. 42) وذلك عن طريق تنقية البيانات من الأخطاء، البيانات الغير المتسقة، القيم المفقودة، القيم المتطرفة، التكرارات والتناقضات. وتعتبر من المراحل المهمة قبل تنفيذ التنقيب في البيانات (Sumiran, 2018, p. 948).

I-7-3-4-مرحلة تحويل البيانات (Data Transformation): ويقصد بها تحويل البيانات للأشكال الملائمة للتنقيب فيها (Sumiran, 2018, p. 948).

I-7-3-5-مرحلة التنقيب في البيانات (Data mining): يتم في هذه المرحلة تطبيق خوارزميات لاستكشاف البيانات وتحليلها بهدف استخراج الأنماط(Fayyad, Piatetsky-Shapiro, & Smyth , 1996, pp. 39-43) . كذلك الوصول إلى تنبؤات مستقبلية عن طريق التنقيب في البيانات المستهدفة.

I-7-3-5-1-خوارزمية (Extreme Gradient Boosting – XGBoost): تعتبر أحد خوارزميات التعلم الآلي الشائعة التي تم تصميمها لكل من مهام التصنيف والانحدار وتشتهر بكفاءتها وأدائها الجيد(Knechtle, et al., 2024, p. 3) . وفيما يلي بعض المميزات الرئيسية لـــها (Crossman-Smith, 2024) :

- إطار عمل تعزيز التدرج (Gradient Boosting Framework): تقوم خوارزمية XGBoost ببناء نماذج بطريقة متتابعة، حيث يقوم كل نموذج جديد بتصحيح الأخطاء التي ارتكبها النموذج السابق. يتم ذلك من خلال تحسين (loss function) باستخدام الهبوط التدريجي (gradient descent).
- التنظيم (Regularization): تتضمن تقنيات تنظيم (Lasso) L1 و (Ridge) L2، والتي تساعد في منع الإفراط في (overfitting) وتحسين تعميم النموذج.
- التعامل مع القيم المفقودة: يمكن لـــ XGBoost التعامل تلقائيًا مع القيم المفقودة، مما يجعلها قوية في السيناريوهات الواقعية التي قد يكون فيها البيانات غير مكتملة.
- المعالجة المتوازية (Parallel Processing): تم تحسين الخوارزمية من حيث السرعة ويمكنها الاستفادة من عدة أنوية (Central Processing Unit - CPU) للحساب بشكل أسرع، وهو ما يعد مفيدًا بشكل خاص مع مجموعات البيانات الكبيرة.
- أهمية الميزات (Feature Importance): توفر XGBoost رؤى حول أهمية الميزات المختلفة في اتخاذ القرارات التنبؤية، مما يساعد في فهم البيانات الأساسية بشكل أفضل.
- المرونة (Flexibility): تدعم الخوارزمية العديد من دوال الهدف، بما في ذلك الانحدار والتصنيف والترتيب، مما يجعلها مرنة لأنواع مختلفة من المشكلات.

I-7-3-5-2-خوارزمية الغابات العشوائية (Random Forest – RF): تستخدم بشكل رئيسي في مهام التصنيف والانحدار (Du, et al., 2024, p. 5)، وهي خوارزمية مرنة للغاية في مجال تعلم الآلة، تستخدم تقنية التجميع (Bagging) باستخدام العينة العشوائية (Bootstrap Aggregation)، والتي تعتمد على مجموعة كبيرة من أشجار الانحدار غير المرتبطة، ويتم حساب متوسط نتائجها للحصول على دالة تنبؤ تقديرية (Pasta, Carapellese , Faedo, & Brandimarte, 2023) . وتعتبر تقنية التجميع إحدى تقنيات التعلم الجماعي التي تستفيد من نهج شجرة التصنيف والانحدار (Classification and Regression Trees - CART) لإنشاء مجموعة كبيرة من أشجار القرار المنفصلة، مما يعزز من قوة ودقة التنبؤات بشكل جماعي. حيث تجمع بين نماذج ضعيفة متعددة لتشكيل نموذج قوي وموحد. يتم تحقيق ذلك من خلال ثلاث خطوات رئيسية: إعادة أخذ العينات من بيانات التدريب لإنشاء مجموعات بيانات فرعية متنوعة، بناء أشجار قرار فردية باستخدام هذه المجموعات، وأخذ متوسط مخرجات هذه الأشجار للحصول على النتيجة النهائية. توفر خوارزمية الغابة العشوائية العديد من المزايا، بما في ذلك مقاومتها لمجموعات البيانات غير المتوازنة، دقتها العالية، وسهولة تدريبها بفضل قدرتها على

حساب أوزان المتغيرات بشكل تلقائي. ويمكن أن تتأثر فعاليتها بعوامل مثل عدد أشجار القرار، عمقها، وحجم مجموعة الميزات المختارة (Du, et al., 2024, p. 5). تُستخدم الغابات العشوائية على نطاق واسع في مجالات متعددة، بدءًا من تحسين استراتيجيات التحكم في أنظمة طاقة الأمواج، وصولاً إلى التنبؤ بسلوك المستخدمين في التسويق، وتعزيز الكشف عن الاحتيال المالي من خلال التقييم الدقيق للمخاطر الائتمانية وتحديد حالات الاحتيال (Du & Zhai, 2024, p. 1).

I-7-3-5-3-خوارزمية آلة دعم المتجه الانحدارية (Support Vector Regression – SVR): طريقة قوية لتحليل البيانات التنبؤية، تم تصميمها خصيصا لتقريب الوظائف المعقدة بدقة عالية. يتمثل أحد العناصر الأساسية في SVR هو تحديد "هامش التسامح" (margin of tolerance) الذي تهمل ضمنه الانحرافات عن القيم المتوقعة. يتيح هذا الأسلوب للنموذج التركيز على الحالات التدريبية الضرورية، التي تعرف باسم متجهات الدعم. وتنبع مرونة هذه الخوارزمية من قدرتها على استخدام وظائف نواة متنوعة، مثل النواة الخطية (linear)، النواة متعددة الحدود (polynomial) ونواة الأساس الشعاعي الغاوسي (Gaussian radial)، مما يجعلها مناسبة لمجموعة واسعة من المجالات. وتستخدم SVR على نطاق واسع في العديد من التخصصات والموضوعات، بما في ذلك ضغط الصور، التنبؤ بالسلاسل الزمنية المالية وتقدير أوقات السفر (Basak, Pal, & Chandra Patranabis, 2007, pp. 3-8).

I-7-3-5-4-خوارزمية الشبكة العصبية العميقة (Deep Neural Network – DNN): هي نموذج متقدم للتعلم الخاضع للإشراف مصمم لتعزيز استخراج الميزات من خلال عدة طبقات معالجة. يتكون هيكلها من طبقة إدخال، وطبقات خفية متعددة، وطبقة إخراج، حيث يلعب كل منها دورا حيويا في تمكين الشبكة من التعرف على الأنماط المعقدة (Adnan , et al., 2021, p. 4) . وتعتبر الشبكات العصبية العميقة فعالة للغاية في تطبيقات متنوعة، مثل التعرف على الصور، النمذجة التنبؤية، تحليل الصور الطبية، ورؤية الحاسوب، أين يكون التفسير الدقيق لأنماط البيانات المعقدة أمرا بالغ الأهمية (Reddy Gadekallu, et al., 2023, p. 2).

I-7-3-6-مرحلة تقييم الأنماط، النماذج وعرضها (Pattern Evaluation and Présentation): يتم في هاته المرحلة عرض النماذج والأنماط في شكل مرئي وإزالة الأنماط الزائدة عن الحاجة (Sumiran, 2018, p. 949).

I-7-3-7-مرحلة اكتشاف المعرفة (Knowledge Discovery): تتم في هذه المرحلة اكتشاف المعرفة والأنماط المخفية واستعمالها في صنع القرارات (Sumiran, 2018, p. 949).

I-7-4-مقاييس تقييم أداء الخوارزميات: هناك العديد من المقاييس المعتمدة لقياس أداء الخوارزميات على البيانات المدروسة من بينها:

I-7-4-1-مقياس معامل التحديد (R^2): يعتبر من أفضل المقاييس في تحليلات الانحدار في أي مجال علمي. ويشير إلى النسبة المئوية للتغيرات التي تحدث في متغير الهدف التي يمكن التنبؤ بها باستخدام المتغيرات المستقلة (متغيرات الإدخال) (Chicco, Warrens, & Jurman, 2021, p. 5)، ويمكن حسابه بالصيغة التالية:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

Y_i: القيم الفعلية

$\hat{Y}_i$: القيم المقدرة

$\overline{Y}$: المتوسط الحسابي للقيم الفعلية

I-7-4-2-مقياس الجذر التربيعي لمتوسط مربع الأخطاء (RMSE): [illegible] ما يمكن ويحسب كما يلي(Chicco, Warrens, & Jurman, 2021, p. 5) :

Y_i: القيم المقدرة

X_i: القيم الفعلية

m: عدد القيم الفعلية

II-الطريقة والأدوات:

نظرا للتطور التكنولوجي والمعلوماتي وكذا التدفق الهائل للبيانات في شتى المجالات أصبح التنقيب في البيانات أحد أهم التقنيات المعتمدة في المجالات البحثية المختلفة، وخاصة في المجال الاقتصادي وتعتبر الخوارزميات: XG Boost الإنحدارية، آلة دعم المتجه الانحدارية SVR، الغابة العشوائيةRF والشبكات العصبية العميقة DNN من بين الخوارزميات التي أثبتت دقتها وفعاليتها في التعامل مع الكثير من المشكلات البحثية. وسنحاول في هذه الدراسة الإعتماد على هذه الخوارزميات من أجل التنبؤ بأسعار السيارات بدقة عالية.

II-1-جمع البيانات (Data collection):

تم تنزيل بيانات هذه الدراسة من موقع Kaggle (https://www.kaggle.com/datasets/adityadesai13/used-car-dataset-ford-and-mercedes) على شكل ملف CSV، تحتوي على 10 متغيرات وهي : طراز السيارة (model)، السنة (year)،

عدد الأميال المقطوعة (mileage)، ناقل الحركة (transmission)، السعر (price)، نوع الوقود (fuel type)، الضرائب (tax)، الكفاءة في استهلاك الوقود بالأميال لكل جالون (mpg)، حجم المحرك (engine size)، لـ 108540 سيارة مستعملة في المملكة المتحدة. عينة من البيانات موضحة في الجدول رقم 1 كما يلي:

الجدول رقم 1: عينة من البيانات الخام للدراسة

	model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize	tax(£)
0	A1	2017	12500	Manual	15735	Petrol	150.0	55.4	1.4	NaN
1	A6	2016	16500	Automatic	36203	Diesel	20.0	64.2	2.0	NaN
2	A1	2016	11000	Manual	29946	Petrol	30.0	55.4	1.4	NaN
3	A4	2017	16800	Automatic	25952	Diesel	145.0	67.3	2.0	NaN
4	A3	2019	17300	Manual	1998	Petrol	145.0	49.6	1.0	NaN

المصدر: من إعداد الباحثين باستخدام لغة البرمجة بايثون

II-1-1-التعريف بمتغيرات الدراسة:

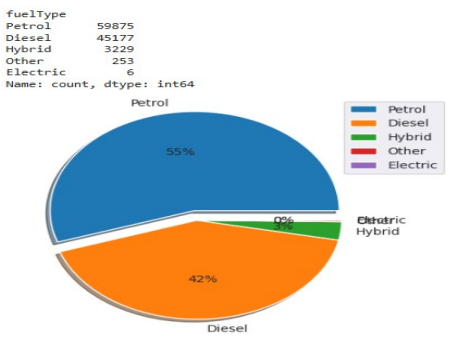
الجدول رقم 2: تحديد نوع وطبيعة متغيرات الدراسة

اسم المتغير	نوع المتغير	طبيعة المتغير
Price	رقمي	متغير الهدف (Target)
Model	اسمي	متغيرات الإدخال (Features)
Year	رقمي	
Transmission	اسمي	
Mileage	رقمي	
fuel type	اسمي	
Tax	رقمي	
Mpg	رقمي	
engine size	رقمي	

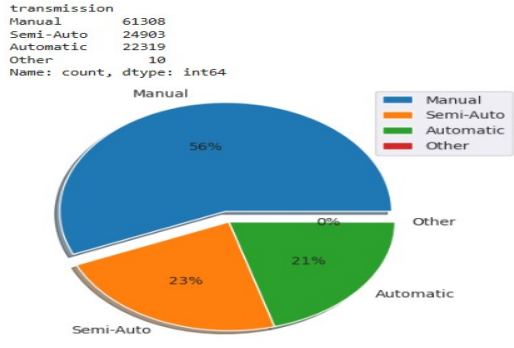
المصدر: من إعداد الباحثين

II-1-2-عرض بيانات متغيرات الدراسة: يعتبر العرض المرئي للبيانات مهما لإعطاء نظرة أولية لشكل البيانات. حيث يوضح الشكل رقم 2 وصفا للمتغيرات الاسمية كما يلي:

الشكل رقم 2: النسب المئوية لكل صفة التي تتكون منها البيانات الوصفية



Distribition of fuelType



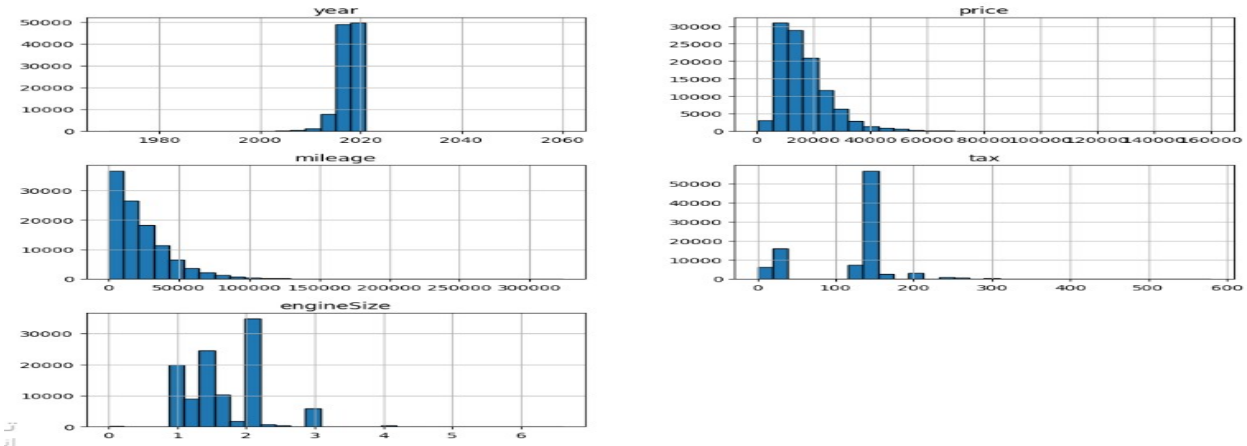
Distribition of transmission

المصدر: من إعداد الباحثين باستخدام لغة البرمجة بايثون

نلاحظ من خلال الشكل رقم 2 وبالنسبة لناقل الحركة (Transmission) في مجموعة البيانات تمتلك 56 % من المركبات ناقل حركة يدوي(Manual)، تليها المركبات ذات الناقل شبه الأوتوماتيكي (Semi-auto) 23 %، ثم الأوتوماتيكي(Automatic) 21 %. وتمثل الفئات الأخرى (Other) أقل من 0.01 % من مجموعة البيانات.

أما بالنسبة لنوع الوقود (fuel type) في مجموعة البيانات فنلاحظ بأن 55 % من المركبات تستخدم البنزين (petrol) والتي تستخدم الديزل (Diesel) تمثل 42 %، في حين تشكّل المركبات الهجينة (Hybrid) نسبة صغيرة تبلغ 3 %. وتعتبر الفئات الأخرى والكهربائية (Other and Electric) شبه معدومة، حيث تشكل كل منهما أقل من 1% من مجموعة البيانات.

الشكل رقم 3: عرض قيم البيانات الرقمية



المصدر: من إعداد الباحثين باستخدام لغة البرمجة بايثون

يوضح الشكل رقم 3 تمثيلا بيانيا للتوزيع التكراري لبيانات المتغيرات الرقمية، تحديداً سنة الصنع (year)، السعر(price)، عدد الأميال المقطوعة (mileage)، الضريبة (tax)، وحجم المحرك (engine Size). ويظهر تحليل هذه المتغيرات خصائص إحصائية متباينة تنعكس في أنماط توزيعها.

فيما يخص متغير سنة الصنع، يُلاحظ أن هناك تركيز كبير في الفترة ما بين 2010 و2020، مما يظهر تفرطح مرتفع (Leptokurtosis) تعكس هيمنة المركبات الحديثة ضمن العينة. أما توزيع متغير السعر، فيتسم بانحراف موجب واضح (Positive Skeweness) وبذيل ممتد نحو القيم المرتفعة، مما يشير إلى وجود عدد قليل من السيارات ذات الأسعار العالية التي تؤثر بشكل غير متناسب على المتوسط الحسابي، ويصاحب هذا التوزيع تفرطح إيجابي ناتج عن تمركز كبير حول الأسعار المنخفضة.

وبالنسبة لمتغير عدد الأميال المقطوعة، فإنه يتبع نمطا منحرفا نحو اليمين، حيث تظهر معظم المركبات مسافة مقطوعة أقل من 100,000 مع تسجيل بعض القيم القصوى، مما يعكس استخداما معتدلا للمركبات.

أما متغير الضريبة، فتعكس توضعا كثيفا حول القيم المنخفضة من 100 إلى 200، مع انحراف طفيف نحو اليمين وتفرطح حاد، مما يشير إلى أن مجموعة البيانات تحتوي أساساً على مركبات ذات ضرائب منخفضة إلى معتدلة.

وأخيرا، يتسم توزيع حجم المحرك بطابع متعدد القيم مع ذروة حوالي 1.0 إلى 2.0 لتر، وعدد قليل من المركبات ذات المحركات الكبيرة تصل إلى 3.0 لتر.

الشكل رقم 4: مصفوفة الارتباط بين متغيرات الدراسة الرقمية بعد المعالجة المسبقة للبيانات

المصدر: من إعداد الباحثين باستخدام لغة البرمجة بايثون

تعرض النتائج الموضحة في الشكل رقم 4 مصفوفة الارتباط بين المتغيرات الرقمية في بيانات الدراسة وتشمل السنة (year)، السعر (price)، عدد الأميال المقطوعة (mileage)، الضريبة (tax)، الكفاءة في استهلاك الوقود بالأميال لكل جالون (mpg) وحجم المحرك (engine Size) يمكن التعليق على النتائج كما يلي:

- قيمة الارتباط بين السنة والسعر (0.50) مما يشير الى ارتباط إيجابي متوسط بين السنة والسعر، أي أنه كلما كانت السيارة حديثة كلما ارتفع سعرها.
- قيمة الارتباط بين عدد الأميال والسعر (0.43-) وتعني ارتباط سلبي متوسط بين عدد الأميال المقطوعة والسعر، أي أنه كلما زادت المسافة المقطوعة للسيارة انخفض سعرها.
- قيمة الارتباط بين حجم المحرك والسعر (0.63) مما يدل على ارتباط إيجابي قوي بين حجم المحرك والسعر، أي أنه كلما زاد حجم المحرك للسيارة كلما كان السعر أعلى.
- قيمة الارتباط بين الضريبة والسعر (0.3) وتظهر ارتباط إيجابي ضعيف بين الضريبة والسعر، أي أنه كلما كان سعر السيارة أعلى كلما كانت الضريبة أعلى.
- قيمة الارتباط بين الكفاءة في استهلاك الوقود والسعر (0.29-) وتظهر ارتباط سلبي ضعيف بين الكفاءة في استهلاك الوقود والسعر، أي أنه كلما زادت كفاءة استهلاك الوقود انخفض سعر السيارة بشكل طفيف. هذا يشير إلى أن المركبات الأعلى كفاءة في استهلاك الوقود قد تكون ضمن الفئات الاقتصادية أو ذات المحركات الأصغر حجمًا، مما يفسر انخفاض متوسط أسعارها.
- قيمة الارتباط بين عدد الأميال المقطوعة والسنة (0.74-) وتدل على ارتباط سلبي قوي بين عدد الأميال المقطوعة والسنة، أي أنه كلما كانت السنة حديثة، كلما كانت المسافة المقطوعة قليلة.
- قيمة الارتباط بين الكفاءة في استهلاك الوقود وحجم المحرك (0.24-) وهي تمثل ارتباط سلبي ضعيف بين الكفاءة في استهلاك الوقود وحجم المحرك، أي أنه كلما كان حجم المحرك أكبر، كلما قلت الكفاءة في استهلاك الوقود أقل.
- قيمة الارتباط بين عدد الأميال المقطوعة والضريبة (0.21-) وتدل على ارتباط سلبي ضعيف بين عدد الأميال المقطوعة والضريبة، إذ تعني أنه كلما كانت الأميال مقطوعة أكثر كلما كانت قيمة الضرائب أقل.

الجدول رقم3: الإحصاءات الوصفية للمتغيرات الرقمية

	min	median	mean	max	std
price	450.000000	14698.000000	16890.120000	159999.000000	9756.270000
mileage	1.000000	17265.000000	23025.930000	323000.000000	21176.420000
tax	0.000000	145.000000	120.260000	580.000000	63.400000
mpg	0.300000	54.300000	55.170000	470.800000	16.140000
engineSize	0.000000	1.600000	1.660000	6.600000	0.560000

المصدر: من إعداد الباحثين باستخدام لغة البرمجة بايثون

أظهرت الإحصاءات الوصفية أن متوسط سعر السيارات في العينة قُدِّر بـــ 16890.12، في حين بلغ الحد الأقصى للسعر 159999.00والحد الأدنى 450.00، مما يشير إلى تباين كبير في الأسعار (الانحراف المعياري = 9756.27) .ويشير هذا التفاوت إلى عوامل متعددة مثل نوع السيارة، فئتها، القوة الحصانية، وسنة الصنع.

أما بالنسبة لعدد الأميال المقطوعة(mileage) ، فقد بلغ المتوسط 23025.93 ميلا، بينما تراوحت القيم بين 1.00 و 323000.00 ميل، مع انحراف معياري مرتفع نسبيًا(21176.42) ، مما يعكس تباينًا كبيرًا في استخدام المركبات في البيانات المدروسة.

فيما يخص قيمة الضريبة(tax) ، بلغ متوسطها 120.26، وكانت القيم تتراوح بين 0.00 و580.00، مع تشتت واضح (الانحراف المعياري = 63.40)، ما قد يعكس اختلافًا في السياسات الضريبية حسب نوع، قوة وحداثة السيارة.

بالنسبة لكفاءة استهلاك الوقود(mpg) ، فإن المتوسط بلغ 55.17 ميل/غالون، بينما كانت القيم الدنيا والعليا 0.30 و470.80 على التوالي، ما يدل على وجود تباين كبير في الكفاءة، تُفسَّر باختلاف نماذج السيارات، نوع الوقود، وحجم المحرك.

أخيرًا، بلغ متوسط حجم المحرك 1.66لتر، في حين تراوحت القيم بين 0.00 و6.60 لتر. ورغم أن الانحراف المعياري (0.56) لم يكن مرتفعًا، فإن وجود القيمة الدنيا "0.00" قد يشير إلى سيارات كهربائية أو بيانات ناقصة تحتاج إلى مراجعة.

II-2-اختيار البيانات وتكاملها: قمنا في هاته الخطوة باختيار واستيراد البيانات المراد التنقيب فيها وذلك بدمج بيانات 11 علامة من السيارات في ملف واحد، حيث كانت لكل علامة بياناتها الخاصة وتم تحضيرها للتنقيب فيها.

II-3-تحضير البيانات (تنظيف البيانات): في هاته الخطوة المهمة قمنا بما يلي:

- حذف عمود المتغير tax(£)لأنه يحتوي على 95 % من البيانات المفقودة.
- تحتوي المتغيرات mpg وtax على8.61 % و13.09 % على التوالي من البيانات المفقودة، حيث قمنا بتعويضهم بالمتوسط الحسابي (mean).

II-4-تحويل البيانات: تم تحويل الميزات الوصفية مثل transmission، fuel type وmodel إلى ميزات رقمية. والجدول الموالي يوضح عينة من البيانات بعد تحويلها:

الجدول رقم4: تحويل البيانات الوصفية إلى بيانات رقمية

transmission_Automatic	transmission_Manual	transmission_Other	transmission_Semi-Auto	fuelType_Diesel	fuelType_Electric	fuelType_Hybrid	fuelType_Other	fuelType_Petrol
0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

المصدر: من إعداد الباحثين باستخدام لغة البرمجة بايثون

يبين الجدول أعلاه الأعمدة الخاصة بالمتغيرات الاسمية التالية:

ناقل الحركة (Transmission): يصف هذا المتغير نوع ناقل الحركة في المركبة، ويتم ترميز فئاته عبر أربعة أعمدة:

- transmission_Automatic
- transmission_Manual
- transmission_Other
- transmission_Semi-Auto

نوع الوقود (Fuel Type): يصف هذا المتغير نوع الوقود الذي تستخدمه المركبة، ويتم ترميز فئاته أيضًا عبر خمسة أعمدة:

- fuelType_Diesel
- fuelType_Electric
- fuelType_Hybrid
- fuelType_Other
- fuelType_Petrol

أين تم تحويلها إلى متغيرات رقمية عن طريق الترميز الأحادي (One-Hot Encoding) ، حيث يمثل كل عمود متغيرا ثنائيا (0 أو 1) لفئات مختلفة من ميزتين. كل صف يحتوي على قيمة 1 في عمود واحد فقط لكل ميزة (نوع ناقل الحركة ونوع الوقود)، على سبيل المثال:

- العمود الذي يحتوي على (transmission_Manual=1) يشير إلى مركبة بناقل يدوي ويشير (transmission_Manual = 0) إلى أن ناقل الحركة ليس يدوي.
- العمود الذي يحتوي على (fuelType_Petrol =1) يشير إلى مركبة تعمل بالبنزين ويشير (fuelType_Petrol = 0) إلى أن المركبة لا تعمل بالبنزين.

وبالتالي:

- الصف الذي يحتوي على (transmission_ Manual = 1) و(fuelType_ Petrol = 1) يدل على مركبة تعمل بالبنزين بناقل حركة يدوي.
- الصف الذي يحتوي على (transmission_Automatic = 1) و(fuelType_Electric = 1) يدل على مركبة كهربائية بناقل حركة أوتوماتيكي.

III–النتائج ومناقشتها:

بعد تحضير البيانات وتنظيفها قمنا بتدريب كل من خوارزمية XG Boost الانحدارية، خوارزمية الغابة العشوائيةRF، خوارزمية آلة دعم المتجه الانحدارية SVR وخوارزمية الشبكات العصبية العميقة DNN على 80 % من بيانات الدراسة بهدف التنبؤ بأسعار السيارات واختبارهم باستخدام 20 %. واعتمدنا على مقياس معامل التحديد (R2) ومقياس الجذر التربيعي لمتوسط مربع الأخطاء (RMSE) للتحقق من دقة النتائج وتقييم أدائهم. والجدول رقم 5 يلخص أداء الخوارزميات كما يلي:

الجدول رقم 5: تلخيص نتائج الدراسة

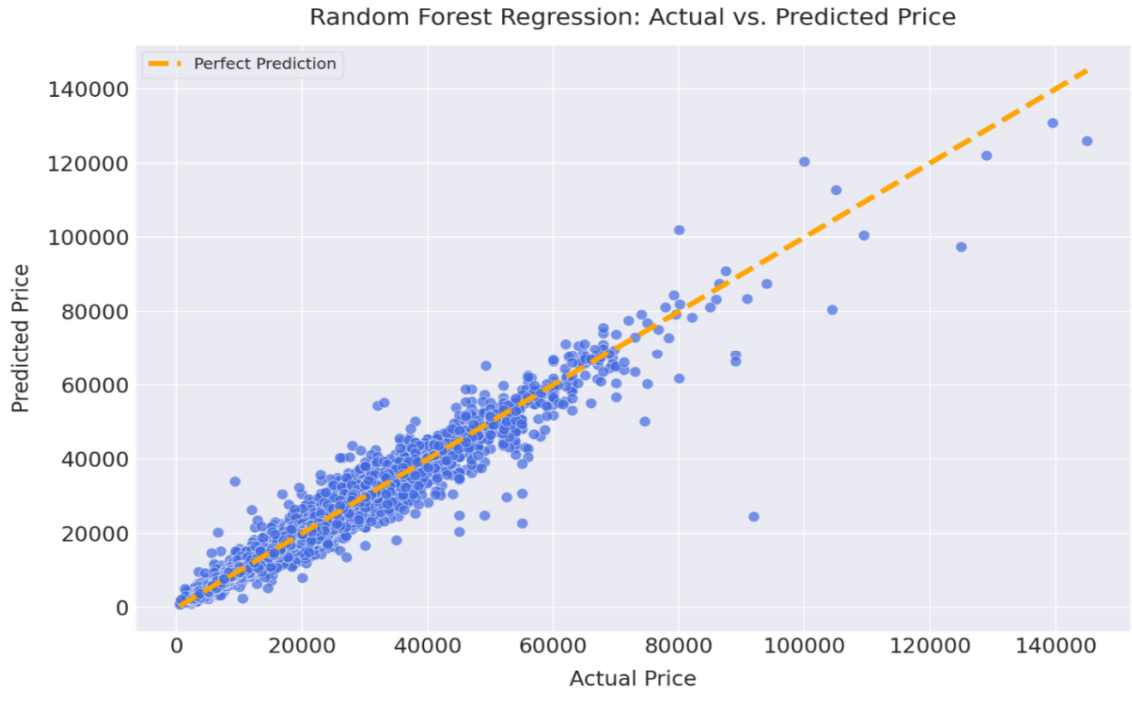
الخوارزميات المستخدمة في الدراسة	R-squared (R^2)	Root Mean Squared Error (RMSE)
XGBoost Regression	94.92%	2173.27
Random Fores (RF)	**95.90%**	1946.07
Deep Neural Network (DNN)	90.14%	3129.04
Support Vector Regression (SVR)	85.79%	3633.59

المصدر: من إعداد الباحثين

يلخص الجدول رقم 5 قيم معامل التحديد (R^2) والجذر التربيعي لمتوسط مربع الأخطاء (RMSE) للخوارزميات المستخدمة في الدراسة. يمكن استخلاص الملاحظات التالية من النتائج:

- التقدير باستخدام خوارزمية الغابات العشوائية (Random Forest Regressor) حقق أعلى قيمة لمعامل التحديد (R^2 = 95.90 %) وأقل قيمة لـــلجذر التربيعي لمتوسط مربع الأخطاء (RMSE = 1946.07)، مما يشير إلى أداء متفوق من حيث الدقة التنبؤية وتقليل الخطأ.
- التقدير باستخدام خوارزمية XGBoost أظهر أداءً جيداً بمعامل تحديد بلغ R^2 = 94.92 % وRMSE قيمته 2173.27. ورغم أنه أقل دقة قليلا مقارنة بخوارزمية الغابات العشوائية، فإن النتائج المتقاربة تشير إلى متانته.
- التقدير باستخدام خوارزمية الشبكات العصبية العميقة (DNN) أظهر نتائج متوسطة، حيث بلغ معامل التحديد 90.14 % والجذر التربيعي لمتوسط مربع الأخطاء بقيمة 3129.04 يشير ذلك إلى فعالية أقل وربما يكون ذلك نتيجة لحساسيته لضبط المعاملات الفائقة أو لتوزيع البيانات.
- التقدير باستخدام خوارزمية آلة الدعم المتجه الانحدارية (SVR) حققت أقل قيمة لمعامل التحديد (85.79 %) وأعلى قيمة لـــ RMSE (3633.59)، مما يشير إلى صعوبة في التعميم وتقليل الخطأ.

الشكل رقم 5: انتشار القيم الحقيقية والمقدرة للأسعار



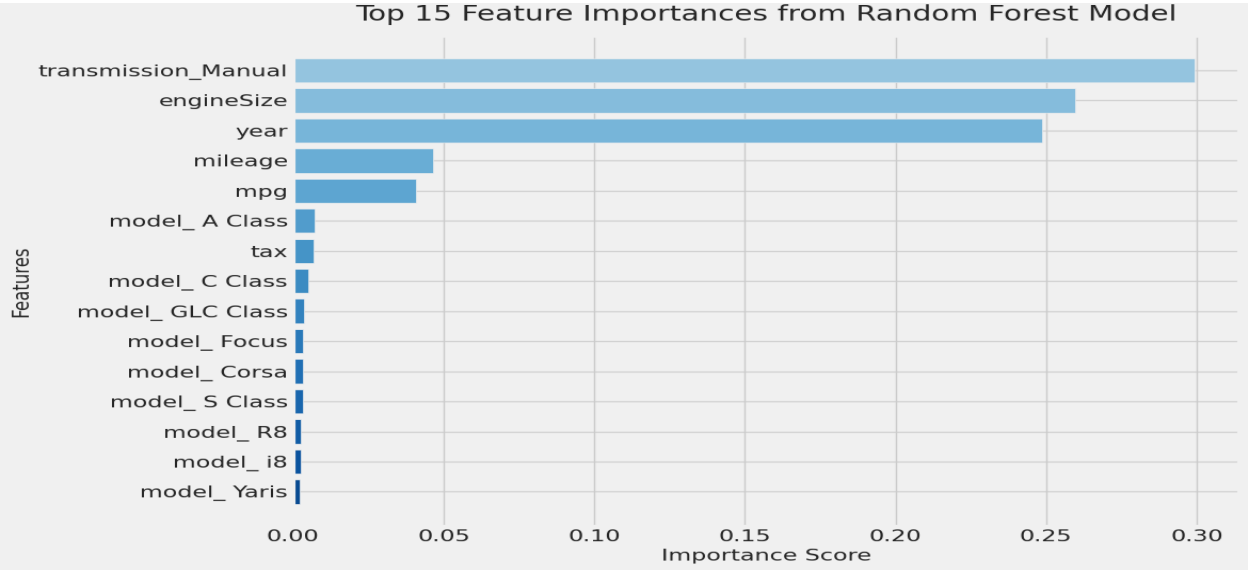
المصدر: من إعداد الباحثين باستخدام بايثون

يعرض الشكل رقم 5 الرسم البياني لنتائج نموذج الانحدار باستخدام خوارزمية الغابة العشوائيةRF، حيث تم مقارنة الأسعار المتوقعة (predicted Price) بالأسعار الفعلية (Actuel Price) ويمكن أن نخلص إلى ما يلي:

- وجود ارتباط إيجابي قوي بين الأسعار الفعلية والأسعار المتوقعة، مما يشير إلى أن خوارزمية الغابة العشوائيةRF تعمل بشكل جيد.
- تظهر الكثافة العالية للنقاط المجمعة حول الخط البرتقالي المتقطع (يمثل خط التنبؤ المثالي) أن خوارزمية الغابة العشوائيةRF تمكنت من التنبؤ بأسعار السيارات بدقة عالية في نسبة كبيرة من الحالات.
- هناك درجة من التباعد عند مستويات الأسعار العالية من 90000 فما فوق، حيث تنتشر النقاط بعيداً عن خط التنبؤ المثالي. يشير هذا إلى ضعف تقدير الأسعار المرتفعة جداً. أي أن التباين في دقة التنبؤ يزداد مع زيادة الأسعار.
- بعض النقاط بعيدة بشكل ملحوظ عن خط التنبؤ المثالي، مما يدل على وجود نقاط شاذة محتملة.

بشكل عام، يظهر التنبؤ باستخدام خوارزمية الغابة العشوائية RF أداء قوي في التنبؤ بالأسعار الأقل من 90000، حيث أن غالبية التنبؤات قريبة جدا من القيم الحقيقية. بينما تعكس الأخطاء الطفيفة في التنبؤات عند مستويات الأسعار العالية (الأكبر من 90000) مجالات للتحسين، ربما من خلال تقنيات النمذجة المتقدمـــــــــــــة باستخدام خوارزميات أقوى أو المعالجة المسبقة الإضافية للبيانات.

الشكل رقم 6: أفضل 15 متغير يأثر في سعر السيارات باستخدام خوارزمية الغابة العشوائيةRF



المصدر: من إعداد الباحثين باستخدام بايثون

من الشكل رقم 6 الذي يقدم أهم 15 متغيرا مؤثرا على أسعار السيارات باستخدام خوارزمية الغابة العشوائيةRF أين يظهر المحور الأفقي درجة الأهمية والمحور الرأسي المتغيرات المستخدمة في الخوارزمية يمكن أن نستنتج ما يلي:

- المتغيرات ذات درجة الأهمية الأعلى هي ناقل الحركة اليدوي(transmission_Manual)، بنسبة 30 %، يليه حجم المحرك (engine size) بدرجة أهمية تصل إلى 26 %، ثم سنة الصنع (year) بدرجة أهمية تقارب 25 %، بعده تأتي كل من الأميال المقطوعة (mileage) و الكفاءة في استهلاك الوقود (mpg) بدرجة أهمية أقل من 5 % وباقي المتغيرات تأتي في الأخير بدرجة أهمية أقل من 2 %، مما يشير إلى أن ناقل الحركة اليدوي، حجم المحرك، وسنة الصنع هي المتغيرات الرئيسية التي تؤثر في أداء خوارزمية الغابة العشوائيةRF. يشير ذلك إلى أن هذه المتغيرات تحتوي على القدر الأكبر من المعلومات ذات الصلة بمتغير الهدف (سعر السيارات).

ومن منظور اقتصادي، يمكن تفسير هيمنة ناقل الحركة اليدوي كأحد أكثر المتغيرات تأثيرًا على السعر، بكون هذا النوع من السيارات أكثر شيوعًا ضمن الفئات الاقتصادية. وهو ما يعكس توجهات شرائية لدى مستهلكين يتميّزون بحساسية مرتفعة تجاه الأسعار. كما يُعدّ كل من حجم المحرك وسنة الصنع من العوامل الأساسية في تقدير القيمة السوقية للسيارة، إذ تميل الفئات ذات الدخل المتوسط إلى تفضيل السيارات الحديثة ذات المحركات المتوسطة، لما توفره من توازن بين الأداء وتكاليف التشغيل والضرائب.

ويُظهر هذا الترتيب في درجة الأهمية مدى وعي المشترين بالجوانب التقنية والوظيفية التي تؤثر على الكلفة الإجمالية للملكية، مثل استهلاك الوقود، مستوى الضرائب، وتكاليف الصيانة. وعليه، تُعد هذه المتغيرات حاسمة في تحديد الأسعار النهائية، لاسيما في الأسواق التي تتسم بشدة المنافسة وارتفاع مستوى الوعي المالي، مثل سوق السيارات المستعملة.

علاوة على ما سبق، تُمثل العلامة التجارية إحدى الركائز غير الرقمية التي تؤثر على تشكيل أسعار السيارات. فهي ترتبط في ذهن المستهلك بمفاهيم الجودة، الموثوقية، وتكاليف الصيانة المتوقعة، مما يجعلها عاملًا غير مباشر لكنه فاعل في قرارات الشراء. وقد بيّنت دراسات متعددة أن العلامة التجارية تؤثر في مخرجات النماذج التنبؤية، لا سيما عند اعتماد بيانات تركز على علامات بعينها (Bukvić, Pašagić, Škrinjar, Fratrović, & Abramović, 2022) (Chen , Hao, & Xu, 2017). ورغم تضمينها في متغير النوع (model)، إلا أن طريقة التشفير الأحادي المستخدمة قد تقلّل من أهميتها الإحصائية، خصوصًا في الخوارزميات القائمة على الأشجار. لذلك، من المفيد مستقبلاً إعادة تصنيف العلامات ضمن شرائح اقتصادية (فاخرة، متوسطة، اقتصادية) لتحسين دقة النماذج وتفسيرها، لا سيما في الأسواق الناشئة كالسوق الجزائري، حيث تلعب العلامات التقليدية مثل Renault وToyota دورًا رمزيًا قويًا في سلوك المستهلك المحلي.

وفي الأخير يمكن القول بأن خوارزمية الغابة العشوائيةRF أفضل خوارزمية من بين الخوارزميات المدروسة ويمكنها أن تساعد في ترشيد القرارات لدقتها العالية، ولكن وجب زيادة البحث فيها أكثر لزيادة فعاليتها في التعامل مع الأسعار العالية والقيم الشاذة.

IV –الخلاصة:

يعتبر التنقيب في البيانات أحد التقنيات الحديثة في تحليل البيانات التي تعتمد على تعلم الآلة، والذي يتم عن طريق نوعين أساسين التنقيب التنبؤي والتنقيب الاستكشافي، حيث تم الاعتماد في هذه الدراسة على التنقيب التنبؤي عن طريق مقارنة خوارزمية XG Boost الانحدارية بالخوارزميات: الغابة العشوائيةRF، آلة دعم المتجه الإنحداريةSVR والشبكات العصبية العميقةDNN في التنبؤ بأسعار السيارات في المملكة المتحدة وتوصلنا إلى النتائج التالية المتعلقة بالفرضيات:

- تم رفض الفرضية الأولى التي تنص على أن خوارزمية XG Boost تتفوق على خوارزمية الغابة العشوائيةRF من حيث دقة التنبؤ، وذلك لتحقيق هذه الأخيرة أعلى قيمة لمعامل التحديد (R2 = 95.90 %) وRMSE قيمته 1946.07؛
- تم تأكيد الفرضية الثانية التي تنص على أن خوارزمية XG Boost تتفوق على كل من خوارزمية آلة دعم المتجه الإنحدارية SVR وخوارزمية الشبكات العصبية العميقة DNN من حيث دقة التنبؤ، وذلك لتحقيقها معامل تحديد R^2 قدّر بـــ 94.92 % وRMSE قيمته 2173.27؛
- تم رفض الفرضية الثالثة التي تنص على أن المتغيرات كشكل السيارة، نوعها، سنة الصنع، الوقود المستهلك، نوع علبة السرعات، حجم المحرك وعدد الأميال المقطوعة تؤثر بشكل كبير في تحديد سعر السيارات. ومن خلال النتائج تبين بأن المتغيرات: ناقل الحركة اليدوي، حجم المحرك، سنة الصنع كان تأثيرهم كبير في التنبؤ بسعر السيارات بحصولهم على 80 % من درجة الأهمية أما كل من الأميال المقطوعة والكفاءة في استهلاك الوقود لم يكن تأثيرهم كبير، لإستحواذهم على درجة أهمية 10 % وباقي المتغيرات العشرة كان تأثيرهم ضعيف تحصلوا على درجة أهمية 10 %.

وفي الأخير من خلال هذه الدراسة التطبيقية يمكننا القول بأنه على الرغم من عدم تحقيق خوارزمية XG Boost الانحدارية لأفضل النتائج لكنها كانت قريبة جدا من خوارزمية الغابة العشوائيةRF التي حققت أعلى دقة وأفضل نتيجة، وبالتالي يمكننا الاعتماد على خوارزمية الغابات العشوائية للمساعدة في ترشيد القرارات بالنسبة لمصنعي السيارات والتجار والمستثمرين في مجال السيارات وكذا مشتري أو بائعي السيارات.

وبناء على ما تم التوصل إليه يمكننا تقديم الاقتراحات التالية:

– رغم أن الدراسة تمت على بيانات من المملكة المتحدة، إلا أن النتائج قابلة للتطبيق في السياق الجزائري في حال توفر بيانات مشابهة. كما يمكن توظيف المنهجية المعتمدة لتطوير نموذج مبدئي لتقييم أسعار السيارات المستعملة في الجزائر، بالاعتماد على بيانات منصات البيع المحلية مثل Ouedkniss؛

– انشاء تطبيق يعتمد على أهم خوارزميات التنقيب في البيانات للتنبؤ بأسعار السيارات يستخدم عن طريق الأندرويد في الهاتف النقال.

- الإحالات والمراجع:

1. أصالة العيفة. (2023). التنبؤ باستخدام التنقيب في البيانات لتحسين عملية اتخاذ القرار. الباحث الاقتصادي، الصفحات 92-102.
2. زين العابدين عمار عبد اللطيف. (2008). الاتجاهات الحديثة للإدارة العلمية في المكتبات و مراكز المعلومات : الإدارة بالأهداف نموذجا. آداب الرافدين، الصفحات 1-30.
3. صلاح الدين قوت، و طارق بلحاج. (2023). استخدام تقنيات التنقيب عن البيانات مع نموذج RFM لتقدير القيمة الدائمة للزبون: دراسة حالة. مجلة ميلاف للبحوث والدراسات، الصفحات 57-73.
4. فتيحة بوهرين، و فاطمة الزهراء لكحل. (2023). استخدام تقنيات التنقيب عن البيانات لتحليل مؤشرات السيولة لمصرف الكويت الوطني باستخدام برنامج DATAtab خلال الفترة 2017-2021. مجلة المنهل الاقتصادي، الصفحات 555-576.
5. BIBLIOGRAPHY Adnan , S., Shahid Khan, A., Nisar, K., Haider, I., Hassan, R., Reazul Haque, M., . . . J. P. C. Rodrigues, J. (2021). Anomaly Detection Using Deep Neural Network for IoT Architecture. applied sciences, pp. 1-19.
6. Reddy Gadekallu, T., Khare, N., Bhattacharya, S., Singh, S., Reddy Maddikunta, P., & Srivastava, G. (2023). Deep neural networks to predict diabetic retinopathy. Journal of Ambient Intelligence and Humanized Computing, pp. 1-14.
7. Basak, D., Pal, S., & Chandra Patranabis, D. (2007). Support Vector Regression. Neural Information Processing – Letters and Reviews, pp. 203-2024.
8. Bukvić, L., Pašagić, J., Škrinjar, P., Fratrović, T., & Abramović, B. (2022). Price Prediction and Classification of Used-Vehicles Using Supervised Machine Learning. Sustainability, pp. 1-17.
9. Chen , C., Hao, L., & Xu, C. (2017). Comparative Analysis of Used Car Price Evaluation Models. AIP Conference Proceedings, pp. 1-6.
10. Chicco, D., Warrens, M., & Jurman, G. (2021). The coefficient of determination R-squared. peerj computer science, pp. 1-24.
11. Crossman-Smith, J. (2024, 03 24). XGBoost Explained: A Beginner's Guide. Retrieved from DATA SCIENCE THEORY | AI & MACHINE LEARNING | DEEP DIVES: https://medium.com/low-code-for-advanced-data-science/xgboost-explained-a-beginners-guide-095464ad418f
12. Du, J., Jacinthe, P.-A., Song, K., Zhang, L., Zhao, B., Liu, H., & Wang, Y. (2024). Maize crop residue cover mapping using Sentinel-2 MSI data and random forest algorithms. International Soil and Water Conservation Research, pp. 1-14.
13. Du, Q., & Zhai, J. (2024). Application of artificial intelligence Sensors based on random forest algorithm in financial recognition models. Measurement: Sensors 101245, pp. 1-9.
14. Fayyad, U., Piatetsky-Shapiro, G., & Smyth , P. (1996). From Data Mining to Knowledge Discovery in Databases. AI MAGAZINE, pp. 37-54.
15. Knechtle, B., Villiger, E., Weiss, K., Valero, D., Gajda, R., Scheer, V., . . . Thuany, M. (2024). Analysis of the 72-h ultramarathon using a predictive XG Boost model. Sport Sciences for Health, pp. 1-24.
16. MUTİ, S., & YILDIZ, K. (2023). Using Linear Regression For Used Car Price Prediction . International Journal of Computational and Experimental Science and ENgineering , pp. 11-16.
17. Pasta, E., Carapellese , F., Faedo, N., & Brandimarte, P. (2023). Data-driven control of wave energy systems using random forests and deep neural networks. Applied Ocean Research, pp. 1-15.

18. Sumiran, K. (2018). An Overview of Data Mining Techniques and Their Application in Industrial Engineering. Asian Journal of Applied Science and Technology (AJAST), pp. 947-953.

كيفية الاستشهاد بهذا المقال حسب أسلوب APA:

توفيق بنين & مصطفى طويطي & ذهيبة بن عبد الرحمان (2025)، التنبؤ بأسعار السيارات باستخدام تقنيات التنقيب في البيانات - دراسة تطبيقية على بيانات مجموعة من السيارات بالمملكة المتحدة -، مجلة الدراسات الاقتصادية الكمية، المجلد 11 (العدد 01)، الجزائر: جامعة قاصدي مرباح ورقلة، ص.ص: 89-102.

يتم الاحتفاظ بحقوق التأليف والنشر لجميع الأوراق المنشورة في هذه المجلة من قبل المؤلفين المعنيين وفقا لـــ **رخصة المشاع الإبداعي نَسب المُصنَّف – غير تجاري – منع الاشتقاق 4.0 دولي** (CC BY-NC 4.0).

مجلة الدراسات الاقتصادية الكمية مرخصة بموجب **رخصة المشاع الإبداعي نَسب المُصنَّف – غير تجاري – منع الاشتقاق 4.0 دولي** (CC BY-NC 4.0).

The copyrights of all papers published in this journal are retained by the respective authors as per the **Creative Commons Attribution License**.

Journal Of Quantitative Economics Stadies **is licensed under a** Creative Commons Attribution-Non Commercial license **(CC BY-NC 4.0).**