Warith Scientific Journal

استعمال البيانات الضخمة و الخوارزمية الجينية لتنبؤ بسلوك مستخدمين مواقع التواصل الاجتماعي من خلال الانحدار اللوجستي

Using Big Data and Genetic Algorithms to Predict the Behavior of Social Media Users through Logistic Regression

زهراء هلال حمود
zahraa hilal Hamuwd
zhraa.hilal@s.uokerbala.edu.iq
جامعة كربلاء – كلية الإدارة و الاقتصاد
Karbala University/ College of Administration and Economics

أ.م.د مشتاق كريم عبدالرحيم
Mushtaq karim Abd Al-Rahem
mushtag.k@uokerbala.edu.iq
جامعة كربلاء – كلية الإدارة و الاقتصاد
Karbala University/ College of Administration and Economics

المستخلص:

يعد تصنيف الحسابات على منصة إنستغرام إلى حسابات حقيقية ومزيفة أمرًا ضروريًا لفهم ومكافحة ظاهرة النصب والاحتيال الإلكتروني استخدم هذا البحث البيانات الضخمة المتاحة من إنستغرام وبهدف معرفة الحسابات والصفحات الحقيقية أو المزيفة يتم نمذجتها باستعمال الانحدار اللوجستي الثنائي حيث تم استعمال احد اهم الانماذج الانحدار غير الخطي في تقدير معلمات الانحدار اللوجستي وباستعمال طريقة المحاكاة لإيجاد افضل تقدير المعلمات الانحدار اللوجستي الثنائي باستعمال طريقة الإمكان الأعظم الاعتيادية وتم تحسينها باستعمال الخوارزمية الجينية في التقدير لإيجاد افضل تقدير لمعلمات الانحدار اللوجستي و باستعمال المعيار الإحصائي متوسط مربعات الخطأ(MSE) لمقدرات الانحدار اللوجستي لغرض المقارنة بين طرائق تقدير معلمات الانحدار اللوجستي وقد توصلت النتائج ان طريقة الإمكان الأعظم المحسنة هي الفضلى بين طرائق التقدير لامتلاكها اقل متوسط مربعات الخطأ للمقدرات وفي الجانب التطبيقي توصلت النتائج باستعمال طريقة الإمكان الأعظم المحسنة والبيانات الخاصة بالصفحات الحقيقية و المزيفة حيث اتضح ان العوامل التي تودي إلى معرفة الحسابات الحقيقية من عدمها هي (عدد الأشخاص أو الصفحات التي يتابعها المستخدم و طول السيرة الذاتية وتوفر الربط وتوفر الصورة الشخصية ومتوسط عدد الهاشتاج والفاصل الزمني بين المشاركة)وتعد هذا العوامل الأكثر تأثيرا لمعرفة الحساب مزيف أو حقيقيوتم تصنيف الحسابات إلى حقيقية أو مزيفة حيث أظهر النموذج دقة تصنيف بلغت 80% للحسابات الحقيقية و75% للحسابات المزيفة.

الكلمات المفتاحية : البيانات الضخمة, الخوارزمية الجينية ,الإمكان الأعظم , أنموذج الانحدار اللوجستي.

Abstract:

Classifying accounts on Instagram as real or fake is essential for understanding and combating the phenomenon of online scams and fraud. This research utilizes the vast data available from Instagram to distinguish between real and fake accounts using binary logistic regression modeling. One of the prominent non-linear models for estimating logistic regression parameters was employed and enhanced using the Maximum Likelihood Estimation method, further optimized with a genetic algorithm. The Mean Squared Error (MSE) criterion was used to compare parameter estimation methods for logistic regression The results indicated that the enhanced Maximum Likelihood Estimation method performed best due to its lower MSE for estimators. In practical application, using the enhanced Maximum Likelihood Estimation method and data specific to real and fake pages revealed that factors such as the number of followers, bio length, link availability, profile picture availability, average number of hash tags, and time gap between posts were the most influential in determining account authenticity These factors proved most impactful in identifying whether an account was real or fake. The model achieved an accuracy of 80% for real accounts and 75% for fake accounts classification based on the data used.

Keywords: Big Data ,Genetic Algorithm, Maximum Likelihood ,Logistic Regression Model.

1. المقدمة:

في العصر الرقمي الحالي، أصبحت وسائل التواصل الاجتماعي جزءًا أساسيًا من حياتنا اليومية، ومن بين هذه المنصات البارزة انستغرام يستخدم الملايين من الأشخاص حول العالم انستغرام للتواصل الاجتماعي، مشاركة لحظاتهم، واكتساب المعلومات. ومع ذلك، تواجه انستغرام تحديات كبيرة مثل زيادة عدد الحسابات المزيفة التي تسعى للتأثير على المستخدمين ونشر معلومات مضللة. بالتالي، يُعتبر التمييز بين الحسابات الحقيقية والمزيفة أمرًا بالغ الأهمية لضمان سلامة المعلومات في هذا السياق يأتي دور الإحصاء وتقنيات التحليل البياني مثل الانحدار اللوجستي، الذي يُستخدم لتصنيف البيانات الثنائية مثل الحسابات الحقيقية والمزيفة. يعتمد الانحدار اللوجستي على عدة متغيرات مستقلة لتقدير احتمال كون حساب ما حقيقيًا أو مزيفًا، مثل عدد المتابعين ومعدل التفاعل ونوعية المحتوى بالإضافة إلى ذلك، تُعد الخوارزميات الجينية تقنية متقدمة تستفيد من مفاهيم علم الأحياء التطوري لتحسين نماذج الانحدار اللوجستي، وذلك من خلال تحسين قيم المعاملات وزيادة دقة التصنيف ومع تزايد حجم البيانات المتولدة على إنستغرام، يصبح استخدام أدوات معالجة البيانات الكبيرة أمرًا ضروريًا

2. مشكلة البحث

تكمن مشكلة البحث في كيفية التعامل النموذج الانحدار اللوجستي الثنائي مع عملية نمذجة البيانات الضخمة والوصول إلى الطرائق المثلى في تقدير المعلمات الانحدار اللوجستي الثنائي عند استعمال البيانات الضخمة

3. هدف البحث

- استعمال الطرائق الذكية (الخوارزمية الجينية) لتقدير أنموذج الانحدار اللوجستي الثنائي باستعمال البيانات الضخمة ومقارنتها مع طريقة (الإمكان الأعظم)
- بناء أنموذج الانحدار اللوجستي لبيان اهم العوامل التي تؤدي لمعرفة المستخدم الأصلي (الحقيقي) أو المزيف (شراء متابعين) في احد مواقع التواصل الاجتماعي (الانستغرام)

3.1 البيانات الضخمة[1]:

يمكن تفسير البيانات الضخمة على انه مصطلح يشير إلى مجموعة من بيانات المعقدة والمتداخلة و العميقة التي يتم جمعها من مصادر مختلفة وتنسيقات مختلفة مثل النص والصور والصوت والرسائل النصية وحجم تداول الأسهم وأخبار الطقس وتعتبر البيانات الضخمة ليست مجرد كميات متزايدة بل هي بيانات معقدة ولا يمكن التعامل معها بطرائق المعالجة التقليدية وهذا التحدي لا ينحصر على معالجتها فقط بل يشمل أيضا عملية البحث عنها بل جمعها تصنيفها خزنها نقلها لأنها تنمو بوتيرة متسارعة للغاية ويمكن جمع هذه البيانات من وسائل التواصل الاجتماعي مثل فيسبوك وتوتير وانستغرام والمدونات على الأنترنت. ويستخدم مصطلح البيانات الضخمة بشكل عام لوصف التجميع والمعالجة والتحليل وتعاريف ومفاهيم هذا المجال مختلفة و تختلف بحسب الخبراء والشركات والمنظمات المتخصصة حيث يعرف معهد ماكينزي(Mckinsey) العالمي البيانات الضخمة على أنها مجموعة من البيانات التي يفوق حجمها القدرة على معالجتها من حيث التقاط ومشاركة ونقل وتخزين وإدارة وتحليل باستخدام أدوات قواعد البيانات التقليدية ، في خلال فترة زمنية مقبولة شركة جارتنر(Gartner)هي شركة متخصصة في أبحاث واستشارات تكنولوجيا المعلومات تعرف " البيانات الضخمة" بانها أصول المعلومات كبيرة الحجم عالية السرعة تتطلب أشكال جديدة من المعالجات لتعزيز عملية صنع القرار أو الفهم العميق وتحسين العملية

ويقترح (Gacobs) تعريف البيانات الضخم هي البيانات التي يجبرنا حجمها على النظر إلى ابعد من الأساليب المجربة والحقيقية السائدة[5]

يرى الباحث يمكن تعريف البيانات الضخمة بانها كم هائل من البيانات المعقدة التي يصعب تحليلها باستخدام قواعد البيانات التقليدية بسبب حجمها الكبير والمتزايدة بسرعة كبيرة جداً وكذلك تنوع مصادرها لذلك لابد من وجود طرق مبتكرة لمعالجة البيانات الضخمة من اجل تحليل البيانات واستخراج النتائج صحيحة.

❖ أنواع البيانات الضخمة[7]

تنقسم البيانات الضخمة إلى عدة أنواع

1. البيانات المنظمة أو المهيكلة تشير إلى تلك البيانات التي تكون منظمة بشكل جداول أو قواعد بيانات أو على شكل حقول مثبتة ومنسقة في ملف أو سجل ومن الأمثلة على البيانات المنظمة إحصائيات مدونة الويب و نقاط البيع مثل الكمية أو الرموز الشريطية أو الأرقام والتواريخ والعناوين ليسهل الوصول اليها

2.البيانات غير منظمة أو غير مهيكلة: هي عبارة عن بيانات غير مرتبة وغير علائقية وغير منسقة وفوضوية ومثقلة بالنصوص وتمثل البيانات غير مهيكلة معظم البيانات الضخمة ولا يمكن تمثيلها بسهولة في الجداول التقليدية كالصور والرسوم البيانية والفيديوهات والمشاركات المكتوبة في شبكات التواصل الاجتماعي.

3. البيانات شبة المنظمة أو المهيكلة : هي البيانات التي تشابه إلى حد ما البيانات المنظمة وكذلك تحمل مميزات كل من البيانات المنظمة وشبه المنظمة ألا أنها لا تكون على شكل جداول أو قاعدة بيانات من الأمثلة البيانات شبه المنظمة لغة الترميز الموسعة والبريد الإلكتروني و تبادل البيانات الرقمية وغير ذلك

3.2 انموذج الانحدار اللوجستي الثنائي[2][9][5]

يعتبر أنموذج الانحدار اللوجستي الثنائي من احد نماذج الانحدار اللاخطي و يختص بان متغير الاستجابة (y)يتبع توزيع برنولي() و يأخذ القيم (1)و(0) أي ان متغير الاستجابة الفئوي (Y) توجد له حالتين اذ تمثل الحالة الأولى عند وقوع حدث معين عندما (Y=1)) والحالة الثانية هي عدم وقوع الحدث عندما (Y=0)ويكون احتمال وقوع الحدث (النجاح)هو$\left[\left(\hat{Y}_i\right) = \hat{\pi}(X_i)\right]$بالاعتماد على قيم المتغيرات التوضيحية للمشاهدات ويكون احتمال عدم وقوع الحدث (الفشل) هو $[1-\pi(X_i)]$ وبذلك تكون دالة الكثافة الاحتمالية بالصيغة الأتية:

$$P(Y_i \backslash X_i) = [\pi(X_i)^{Y_i}[1-\pi(X_I)^{1-Y_i} \quad \ldots(1)$$

$$Y_i = 0,1$$

$$P(Y=1/X_i) = \pi(X_i)$$

$$P(Y=0/X_i) = 1-\pi(X_i)$$

أنموذج الانحدار اللوجستي الثنائي يحتوي على اكثر من متغير توضيحي واحد(X_i), فيعبر عن توقعه الشرطي لاحتمال متغير الاستجابة (وقوع الحدث) حسب الصيغة الرياضية الأتية:

$$\pi(X_i) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{i1}+\cdots+\beta_P X_{ip}}} \quad \ldots..(2)$$

$$= \frac{e^{(\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip})}}{1+e^{(\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip})}} \quad \ldots..(3)$$

وان احتمال حدوث وقوع الحدث هو:

$$1-\pi(X_i) = \frac{1}{1+e^{(\beta_0+\beta_1 X_{i1}+\cdots+\beta_p X_{ip})}} \quad \ldots.(4)$$

$X_1, X_2, \ldots..X_p$:المتغيرات التوضيحية التي تكون المصفوفة X=Xijبعد$(n*p)$

n ; $i=1,2,\ldots n$:عدد المشاهدين

p ; $j=0,1,\ldots,p$: عدد المتغيرات التوضيحية والمعلمات المجهولة.

$\beta_0, \beta_1, \ldots, \beta_p$.:متجه للمعلمات المراد تقديرها

ويتم تحويل هذا الانموذج الى شكل خطي يتمثل بعلاقة خطية عن طريق المتجه الصفي $\beta_0, \beta_1, \ldots, \beta_p$من المتغيرات التوضيحية مع لوجت الاحتمال $logit\ \pi(X_i)$:وحسب الصيغة الرياضية الاتية

$$Z_i = logit\ \pi(X_i) = \ln\left[\frac{\pi(X_i)}{1-\pi(X_i)}\right] \quad \ldots(5)$$

$$= \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} \quad \ldots(6)$$

$$= \begin{bmatrix} 1 & X_{i1} & \dots & X_{ip} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad = \underline{X'_i}\ \underline{\beta} \qquad \dots . (7)$$

$$Z_i = \underline{X'_i}\ \underline{\beta} + \varepsilon_i$$

$$\pi(X_i) = \frac{e^{\underline{X'_i \beta}}}{1 + e^{\underline{X'_i \beta}}} \qquad \dots (8)$$

$$1 - \pi(X_i) = \frac{1}{1 + e^{\underline{X'_i}\ \underline{\beta}}} \qquad \dots (9)$$

3.3 طريقة تقدير الإمكان الأعظم[10][7]

تعتبر طريقة الإمكان الأعظم من الطرائق التقليدية (الكلاسيكية)واسعة الاستعمال في تقدير معلمات الأنموذج الإحصائية والرياضي لكي يتم الحصول على تقديران المعلمة بواسطة الإمكان الأعظم(MLE) حيث يتم ضرب الحدود في الصيغة (1)لعينة حجمها(N)تعتمد على (X) من المتغيرات التوضيحية ومجموعة متغير الاستجابة(Y)وبهذا فان دالة الإمكان الأعظم في الأنموذج اللوجستي الثنائي ل(n) من المشاهدات هي:

$$P(Y_i/X_i) = [\pi(X_i)]^{Y_i}[1 - \pi(X_i)]^{1-Y_i} \qquad \dots (10)$$

لتسهيل عملية الحل وذلك بأخذ اللوغارتم (Log) على الجانبين للحصول على(MLE)

$$\ln[l(\beta)] = \sum_{i=1}^{n} Y_i \ln[\pi(X_i)] + (1 - Y_i) \ln[1 - \pi(X_i)] \qquad \dots \quad (11)$$

$$\ln[(\beta)] = \sum_{i=1}^{n} \left[Y_i \ln\left(\frac{e^{\acute{X}_i \underline{\beta}}}{1 + e^{\acute{X}_i \underline{\beta}}}\right) + (1 - Y_i) \ln\left(1 - \frac{e^{\acute{X}_i \underline{\beta}}}{1 + e^{\acute{X}_i \underline{\beta}}}\right)\right] \qquad \dots (12)$$

$$= \sum_{i=1}^{n} \left[Y_i ln\left(e^{\acute{X}_i \underline{\beta}}\right) - Y_i ln\left(1 + e^{\acute{X}_i \underline{\beta}}\right) + ln\left(\frac{1}{1 + e^{\acute{X}_i \underline{\beta}}}\right) - Y_i ln\left(\frac{1}{1 + e^{\acute{X}_i \underline{\beta}}}\right)\right] \qquad \dots (13)$$

$$= \sum_{i=1}^{n} \left[Y_i \left(\acute{X}_i \underline{\beta}\right) - \ln\left(1 + e^{\acute{X}_i \underline{\beta}}\right)\right] \qquad \dots (14)$$

ومن اجل الحصول على تقديرات وهي(β) لتعظيم لوغاريتم دالة الإمكان (ln (β)]=L (β)تؤخذ المشتقات من الدرجة الأولى ومساواة الدالة الناتجة بالصفر لكل j من المعلمات أي نستخرج المشتقة الجزئية الأولى لكلβوبحسب ما يأتي:

$$\acute{L}\left(\underline{\beta}\right) = \sum_{i=1}^{n} \left[\left(Y_i - \frac{e^{\acute{X}_i \underline{\beta}}}{1 + e^{\acute{X}_i \underline{\beta}}} \right) \acute{X}_{ij} \right] = \acute{X}\left(Y - P^{(M)}\right) = 0 \qquad \dots (15)$$

$$\acute{L}(\beta) = -\sum_{i=1}^{n} \acute{X}_{ij} [\pi(X_i)(1 - \pi(X_i))] X_{ij} \ = -\acute{X}\, V^{(m)} X \qquad \dots . (16)$$

ونستنتج بانه تكونت لدينا(P+1) من المعادلات غير الخطية ونقوم بحلها بإحدى الطرائق التكرارية التقليدية كطريقة نيوتن رافسون (NR)لا يجاد تقديرات دالة الإمكان الأعظم في المشتقة من الدرجة الأولى ومساواتها للصفر ولذلك فان خوارزمية نيوتن رافسون التكرارية لا يجاد قيم(β) التقديرية لدالة الإمكان الأعظم في الأنموذج اللوجستي الثنائي ستكون في m+1من التكرارات وكالاتي

$$\hat{\beta}^{(m+1)} = \hat{\beta}^{m} + \left(\acute{X}\, V^{m}\ X\ \right)^{-1} \acute{X}\left(Y - P^{(m)}\right) \qquad \dots (17)$$

$$Y=\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_n\end{bmatrix}, P^{(m)}=\begin{bmatrix}{\pi_1}^{(m)}\\{\pi_2}^{(m)}\\\vdots\\{\pi_n}^{(m)}\end{bmatrix}, X=\begin{bmatrix}\acute{X}_1\\\acute{X}_2\\\vdots\\\acute{X}_n\end{bmatrix}=\begin{bmatrix}1 & X_{11} & X_{12} & \cdots & X_{1p}\\1 & X_{21} & X_{22} & \cdots & X_{2p}\\\vdots & \vdots & \vdots & \cdots & \vdots\\1 & X_{n1} & X_{n2} & \cdots & X_{np}\end{bmatrix}$$

$$V^{(m)}=\begin{bmatrix}{\pi_1}^m(1-{\pi_1}^m) & 0 & \cdots & 0\\0 & {\pi_2}^m(1-{\pi_2}^m) & \cdots & 0\\\vdots & \vdots & \cdots & \vdots\\0 & 0 & \cdots & {\pi_n}^m(1-{\pi_n}^m)\end{bmatrix}$$

اذ ان :

(Y):تمثل متجه متغير الاستجابة ذو بعد$(n*1)$

$\widehat{\beta^m}$: (m) هي تقديرات الإمكان الأعظم ذو الرتبة(P+1*n)

$P^{(m)}$: يمثل القيم الاحتمالية لحدوث متغير الاستجابة ذو بعد (n *1)للتكرار (m)

(X: تمثل مصفوفة المتغيرات التوضيحية ذو بعد (n*P +1

$V^{(M)}$: مصفوفة مربعة لتباينات عناصرها قطرها الرئيسي $[{\pi_i}^m(1-{\pi_i}^m)]$مكتسبة من التكرار السابق(m)

3.4 الخوارزمية الجينية[6][3][7]

تعتبر الخوارزمية الجينية احد أساليب الذكاء الاصطناعي وهي من الرسالة العشوائية التي تعالج مشكلة ما من اجل التوصل إلى افضل النتائج الممكنة وبرزت أهميتها في حل المسائل المعقدة لأنها تمتلك كماً هائلا من الحلول البديلة اذ تعتمد على الألية الانتقاء الطبيعي ونظام الجينات الطبيعية واستعمال الخوارزميات الجينية بهدف الحصول على الحلول المثلى للمسائل الرياضية كإحدى الطرائق التكرارية المتبعة حديثاً من اجل اتخاذ القرارات الصحيحة, ان اهم ما هذه الطريقة هو إيجاد علاقة بين المشكلة قيد البحث وبين الخوارزمية وتنشا هذه العلاقة عن طريق الترميز Evaluation Function ودالة التقيمEncoding ويكون ترميز الحلول باستعمال سلسلة من الأرقام الثنائية(0,1) Binary Numbers وهو من افضل الحلول أو استعمال رموز أخرى كأرقام حقيقة وتسمى هذه الأرقام كروموسومات .أما دالة التقييم فتأخذ كل الكروموسوم على جانب وتقييم أدائه في حل المشكلة بإعطائه قيمة معينة وكلما كانت هذه القيمة كبيرة كان الكرموسوم ذات كفاءة عاليةFitness Function وتسمى دالة المفاضلة .تطبق عملية التهجين والطفرة للحصول في النهاية على مجموعة الكروموسومات التي تمثل الجيل الخير وباختيار الأفضل وصولا للحل الأمثل وتنتهي عنده

عمل الخوارزمية الجينية(Work Genetic Algorithms):

تعتبر من التقنيات الهامة في البحث عن الخيار الأمثل من مجموعة حلول متوفرة لتصميم معين توجد في التطبيقات المعلوماتية الأحيائية وتوجد خطوات تعمل بها وهي:

1. يتم تقيم الأبناء الجدد بالاعتماد على الدالة الأصلية
2. تتغير الكروموسومات الأصلية بالاعتماد على تقيم الأبناء وان كل الكرموسوم يتكون من عدد من القيم ويتم تحديد عدد من القيم حسب المسالة وقيمة البداية.
3. يتم اختيار افضل الإباء لكي تتم عملية إنتاج الأبناء.
4. يتم توليد جيل جديد باستخدام الطفرة والعبور.

تطبيق مراحل الخوارزمية الجينية في أنموذج الانحدار اللوجستي الثنائي

يتم تطبيق مراحل الخوارزمية الجينية في معادلة دالة الهدف لكل طريقة لإيجاد تقديرات معلمات أنموذج الانحدار اللوجستي الثنائي وفقا لما يأتي

1. البداية: تكوين الكرموسوم عن طريق قيم (βp) التي تشكل جينات الكروموسوم وان (p=0,1,…,p)ضمن الأعداد الحقيقة.
2. التهيئة: أنشاء الجيل الابتدائي عن طريق إيجاد قيمة أولية للجينات مع القيم العشوائية لمجموعة القيود الأخرى.

3. في دالة الهدف يتم تقيم الكروموسوم من اذ الكفاءة وصولاً إلى الحل الأفضل بتحديد قيمة (βp)
4. أجراء عملية الاختيار للكروموسوم الذي يمتلك قيمة دالة هدف صغيرة باختيار الاحتمال الكبير لها و إيجاد دالة التقييم له عن طريق المعادلة الأتية:

$$fitness\ function = \frac{1}{1 + objective\ funtion} \quad \ldots.(18)$$

f_i:تمثل دالة التقييم.

$o.f_i$:تمثل دالة الهدف.

ومن صيغة دالة التقييم نستطيع إيجاد احتمالية هذه الدالة (افضل القيم)بحسب الصيغة الرياضية الأتية:

$$C_{(i)} = \frac{f_{(i)}}{\sum_{i=1}^{n} f_{(i)}} \quad \ldots.(19)$$

$C_{(i)}$: تمثل احتمال الفرد(الكرموسوم) i

n :يمثل حجم المشاهدات

$f_{(i)}$: تمثل دالة التقييم للفردi

وباستعمال احد معايير الاختبار (roulette wheel)عجلة الروليت بتوليد رقم عشوائي ($R_{(c)}$)محصور في مجال[1,0] فاذا كان تمثل دالة التقييم للفرد $R_C < C_1$ سوف يتم اختيار الكرموسوم الاول (كالام)اويتم الاختيار بحيث يكون الاحتمال محصور وفق $C_{(p-1)}$<$(R_c < C_p)$ يكون الرقم العشوائي محصور وفق

$C_{(p-1)} < R_c < C_{(i)}$وفي كل مرة يتم تحديد كرموسوم واحد للمجتمع الجديد بالاعتماد على دالة التقييم

5. بعد إتمام عملية الاختيار تأتي بعدها عملية التهجين للكروموسومات الجيدة في صفاتها عن طريق التزاوج بين كل كروموسومين وبتطبيق احد معاييره وهو التهجين المنظم بالاعتماد على احتماليةP_c,وتقارن هذه القيمة مع قيمة الجينات للكروموسومين(الآباء) لتكوين الجيل الجديد(الأبناء)ويحدث التبادل عندما تكون قيمة الجين اكبر أو تساوي القيمة الاحتمالية.
6. أخير خطوة ممكن ان تمر بها الكروموسومات هي عملية الطفرة وأيضا تعتمد مقدار احتمالي (P_m) للمعلمات باستبدال جينات منتقاة عشوائياً مع قيمة جديدة أيضا حصلنا عليها بشكل عشوائي

3.5 المحاكاة:

ويمكن تعريف المحاكاة بانها طريقة أو أسلوب تعليمي يستعمل عادة لتمثيل الواقع الحقيقي الذي يصعب الوصول اليه أي يعطي نسخة افتراضية تكون طبق الأصل من العالم الواقعي النظام معين أو أنموذج محدد من الإشارة لهذا الأنموذج بشكل مباشر وهناك العديد من طرائق مختلفة لمحاكاة حيث تعد طريقة مونت كارلو هي الأكثر استعمالا في التحليل الإحصائي التي تستعمل لتوليد البيانات العشوائية من المجتمع النظري المماثل للمجتمع الأصل *مراحل بناء تجربة المحاكاة

بناء تجربة المحاكاة تضم اربع مراحل وهي كالاتي

1. المرحلة الأولى- تحدد القيم الافتراضية وتعد هذه المرحلة من اهم المراحل التي يمكن الاعتماد عليها في المراحل اللاحقة حيث يتم في هذه المرحلة تحديد ثلاث إحجام مختلفة للعينات وكذلك تحديد القيم الافتراضية للمعلمات حيث تم تكرار التجربة(**R=1000**)

جدول(1) القيم الافتراضية للمعلمات في أنموذج الانحدار اللوجستي الثنائي

Para Mod.	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
	0.65	-0.22	-0.48	-0.81	1.03	0.80	0.50	1.51	-0.05	-1.08	-0.47

	β_{11}	β_{12}	β_{13}	β_{14}	β_{15}	β_{16}	β_{17}				
	-1.48	-1.13	0.61	0.73	0.93	-0.33	1.34				

جدول(2) القيم الافتراضية للمعلمات في أنموذج الانحدار اللوجستي الثنائي

Para	B_0	B_1	B_2	B_3	B_4	B_5	B_6	B_7	B_8
	-0.23	1.08	0.22	0.12	0.76	0.13	-2.12	-0.8	-0.67
	B_9	B_{10}	B_{11}	B_{12}	B_{13}	B_{14}	B_{15}	B_{16}	B_{17}
	0.16	-1.79	-2.16	1.32	0.9	-0.2	0.43	-0.06	-0.3

2. المرحلة الثانية - توليد البيانات سيتم توليد (17) متغير توضيحي كما في الجانب التطبيق وذلك باستعمال أسلوب مونت كارلو عن طريق التوزيع المنتظم أما توزيع الخطأ فانة يتبع توزيع برنولي ()وفي هذه المرحلة يتم توليد قيم المتغيرات التوضيحية وقيم متغير الاستجابة لذلك سوف نستعمل طريقة الرفض والقبول لاحتساب المتغيرات ثنائية الاستجابة وتحديد القيم الاحتمالية وفق الاتي:

$$Y = \begin{Bmatrix} 1 if \pi(X_i) \geq 0.5 \\ 0 if \pi(X_i) < 0.5 \end{Bmatrix}$$

3. المرحلة الثالثة التقديرات في هذه المرحلة يتم تقدير معلمات أنموذج الانحدار اللوجستي الثنائي() المعطى في المعادلة وفق الطرائق الاعتيادية وأيضا وفق توظيف الخوارزمية الجينية مع طرائق الاعتيادية التي تم ذكرها في الجانب النظري

4. المرحلة الرابعة: المقارنة بين طرائق التقدير الاعتيادية والحديثة تعد هذا المرحلة الأخيرة من مراحل وصف تجربة المحاكاة حيث يتم في هذه المرحلة المقارنة بين طرائق التقدير بعد ان تم إيجاد تقديرات للمعلمات في المرحلة السابقة باستعمال المقاييس الإحصائية لغرض الحصول على افضل طريقة تقدير للأنموذج قيد البحث لذلك سيتم المقارنة بين طرائق تقدير معلمات الانحدار اللوجستي الثنائي باستعمال متوسط مربعات الخطأ(MSE) للأنموذج المدروسة حسب المعادلة الأتية

$$MSE = \frac{1}{R}\sum_{i=1}^{R} MSE = \frac{1}{R}\sum_{i=1}^{R}\left[\frac{1}{N-P}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2\right]$$

3.5.1 تحليل نتائج المحاكاة:

ستتم عرض نتائج عملية المحاكاة وثم تحليلها للوصول إلى افضل الطرائق لتقدير معلمات أنموذج الانحدار الثنائي بالاعتماد على المقياس الإحصائي متوسط الخطأ (MSE) لمقدرات الأنموذج سوف نوم بعرض نتائج المحاكاة التي تم الحصول عليها عن طريق برنامج (MATLAB)

جدول (3)متوسط مربعات الخطأ (MSE)(لمقدرات الأنموذج الانحدار اللوجستي في الأنموذج (1)للمعلمات باستعمال الطرائق الاعتيادية و الجينية ولكافة أحجام العينات)

sizes	Methods	Classic	Genetic	**Best**
10000	MLE	0.2923	0.0909	**Genetic**
30000	MLE	0.0802	0.0775	**Genetic**
50000	MLE	0.0680	0.0611	**Genetic**

جدول(4) متوسط مربعات الخطأ (MSE)(لمقدرات الأنموذج الانحدار اللوجستي في الأنموذج (2)للمعلمات باستعمال الطرائق الاعتيادية و الجينية ولكافة أحجام العينات)

Size	Classic	Genetic	Best
10000	0.0837	0.0798	Genetic
30000	0.0802	0.0792	Genetic
50000	0.0776	0.0747	Genetic

نلاحظ في الجدول أعلاه تفوق طريقة الإمكان الأعظم المحسنة (MLE.GA)على طريقة الإمكان الاعتيادية وذلك عند أحجام (10000,30000,50000)وكذلك نلاحظ تميز طريقة (MLE.GA)) احتلت المرتبة الأولى من بين الطرائق وذلك لامتلاكها الأفضلية في تقدير المعلمات من حيث امتلاكها اقل متوسط مربعات الخطأ أما في ما يخص الطرائق الاعتيادية تميزت طريقة الإمكان الأعظم عند حجم العينة.(50000)

3.5.2 البيانات الحقيقية:

تم الحصول على البيانات الحقيقية من موقع (kaggle) هو عبارة عن منصة على الأنترنت تقدم مجموعة كبيرة من مجموعات البيانات اذ تم اخذ عينة بحجم(58000) من موقع الانستغرام وذلك لمعرفة الحسابات الحقيقية والمزيفة الخاصة بهذا التطبيق

متغير الاستجابة (Response Variable)

Y_i: مستخدم مواقع التواصل الاجتماعي (الانستغرام) $i = 0,1$

Y=1 :مستخدم حقيقي ,Y=0: مستخدم غير حقيقي(مزيف)

متغيرات التوضيحية (Variables Explanatory)

x_1: متغير كمي عدد المشاركات / أجمالي عدد المشاركات التي نشرها المستخدم على الأطلاق

x_2: متغير كمي عدد الأشخاص أو الصفحات التي يتابعها المستخدم

x_3: متغير كمي عدد المتابعين لدى الحساب المستخدم

x_4: متغير كمي طول السيرة الذاتية

x_5: متغير وصفي توفر صورة

x_6:متغير وصفي توفر الرابط

x_7: متغير كمي متوسط طول التسمية التوضيحية

x_8: متغير كمي التسمية التوضيحية الصفرية نسبة مئوية (0.0 الى1.0)

x_9: متغير كمي نسبة غير الصورة نسبة مئوية(0.0 إلى 1.0) للوسائط غير الصور هناك ثلاث أنوع من الوسائط في الانستغرام هي (الصور, الفيديو, العرض الدائري)

x_{10}: متغير كمي نسبة التفاعل يتم تعرف عليها من خلال عدد الأعجاب مقسوما على عدد الوسائط مقسوما على (عدد المتابعين)

x_{11}: متغير كمي معدل المشاركة يشبه نسبة التفاعل ولكنة مخصص للتعليقات

x_{12}: متغير كمي نسبة علامة الموقع النسبة المئوية (0.0 الى1.0) للمشاركات الموسومة بالموقع

x_{13}: متغير كمي متوسط عدد الهاشتاج

x_{14}: متغير كمي متوسط استخدام الكلمات الرئيسية الترويجية في الهاشتاج

x_{15}: متغير كمي متوسط الكلمات الرئيسية للبحث عن المتابعين في الهاشتاج

x_{16}:متغير كمي متوسط تشابه جيب تمام بين كل زوج من المنشورات لدى المستخدم

x_{17}:متغير كمي متوسط الفاصل الزمني بين المشاركات (بالساعات)

اختبار مربع كأي(Chi square test)

ان أساس هذا الاختبار هو تحديد الفرق بين التكرارات المشاهدة و التكرارات المتوقعة عندما يكون الفرق صغيراً فيكون على إساسه مطابقة البيانات والصيغة الرياضية لاختبار مربع كأي كالاتي

جدول (5) قيمة $\chi 2$ للمتغيرات التوضيحية

المتغيرات	Siq	p-value
X1	0	0.05
X2	0.19	0.05
X3	0.0002	0.05
X4	0	0.05
X5	0	0.05
X6	0	0.05
X7	0	0.05
X8	0	0.05
X9	0.019	0.05
X10	3.13e-07	0.05
1	7.93e-07	0.05
X12	0	0.05
X13	0	0.05
X14	4.10e-09	0.05
X15	0	0.05
X16	0	0.05
X17	0	0.05

$$\chi 2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

$H_0: \chi 2 = 0$

$H_1: \chi 2 \neq 0$

3.5.3 تحليل النتائج التطبيقية:

في هذا القسم سيتم عرض النتائج التطبيقية ومن ثم تحليلها للوصول إلى مدى ملائمة ودقة البيانات الحقيقية مع أنموذج الانحدار اللوجستي الثنائي الذي تم تقديره من خلال أجراء الاختبارات بمقدرات الأنموذج سوف يتم عرض النتائج التطبيقية التي تم الحصول عليها باستعمال برنامج(MAILAB)

جدول (5) يمثل المعلمات المقدرة وكذلك الخطأ المعياري للمتغيرات التوضيحية كافة بطريقة الإمكان الأعظم المحسنة

المعلمات$(\widehat{\beta}_i)$	المعلمات المقدرة	الخطأ المعياري $SE(\widehat{\beta}_i)$	نسبة (Z) $\frac{\widehat{\beta}_i}{SE(\widehat{\beta}_i)}$	المعنوية
$\widehat{\beta}_0$	-1.0625	0.0751	-14.1427	Non-sig
$\widehat{\beta}_1$	-0.0000	1.7e-05	-1.3060	Non-sig
$\widehat{\beta}_2$	0.0000	3.1b e-06	3.6726	Sig
$\widehat{\beta}_3$	-0.0004	5.7e-06	-77.8307	Non-sig
$\widehat{\beta}_4$	0.0049	0.0002	23.4114	Sig

$\hat{\beta}_5$	0.8537	0.0740	11.5330	Sig
$\hat{\beta}_6$	2.0505	0.0311	65.8730	Sig
$\hat{\beta}_7$	-0.0006	0.0001	-10.2281	Non-sig
$\hat{\beta}_8$	-0.0958	0.0409	-2.3422	Non-sig
$\hat{\beta}_9$	0.2276	0.0445	5.1149	Sig
$\hat{\beta}_{10}$	-0.0005	0.0001	-4.9369	Non-sig
$\hat{\beta}_{11}$	0.0492	0.0045	10.8667	Sig
$\hat{\beta}_{12}$	1.4448	0.0402	35.9108	Sig
$\hat{\beta}_{13}$	0.0618	0.0105	5.8799	Sig
$\hat{\beta}_{14}$	-5.0234	0.1849	-27.1622	Non-sig
$\hat{\beta}_{15}$	-0.3871	0.0431	-8.9827	Non-sig
$\hat{\beta}_{16}$	-1.1587	0.0393	-29.4522	Non-sig
$\hat{\beta}_{17}$	0.0005	1.4e-05	31.1031	Sig

من خلال الجدول أعلاه حيث تم تقدير المعلمات وقيم الخطأ المعياري لكل معلمة وتم التقدير بطريقة الإمكان الأعظم الجينية وعن طريق النتائج التي حصلنا عليها نلحظ ان نسبة (Z) التي تتبع توزع الطبيعي القياسي وبمستوى دلالة (α=0,05) وتم مقارنتها مع قيمة (Z) الجدولية 1.96= $Z_{\frac{1}{2}}$(0.05) نحصل على العوامل المؤثرة وكذلك العوامل التي ليس لها تأثير في الأنموذج المقدر ومن تجربة المحاكاة نلحظ المعلمات المقدرة التي ليس لها تأثير تمت أزالتها

($\hat{\beta}_0$ $\hat{\beta}_1$ $\hat{\beta}_3$ $\hat{\beta}_7$ $\hat{\beta}_8$ $\hat{\beta}_{10}$ $\hat{\beta}_{14}$ $\hat{\beta}_{15}$ $\hat{\beta}_{16}$)

بينما معاملات التي لها تأثير هي

($\hat{\beta}_2$ $\hat{\beta}_4$ $\hat{\beta}_5$ $\hat{\beta}_6$ $\hat{\beta}_9$ $\hat{\beta}_{11}$ $\hat{\beta}_{12}$ $\hat{\beta}_{13}$ $\hat{\beta}_{17}$)

ومعامل الانحدار الذي اختص($X_2, X_4, X_5, X_6, X_9, X_{11}, X_{12}, X_{13}, X_{17}$)

لها تأثير في أنموذج الانحدار اللوجستي وذات تاثير معنوي كبير

$$\hat{\pi}(x) = \frac{e(\beta 2xi2 + \beta 4xi4 + \beta 5xi5 + \beta 6xi6 + \beta 9xi9 + \beta 11xi11 + \beta 12xi12 + \beta 13xi13 + \beta 17xi17}{1 + (\beta 2xi2 + \beta 4xi4 + \beta 5xi5 + \beta 6xi6 + \beta 9xi9 + \beta 11xi11 + \beta 12xi12 + \beta 13xi13 + \beta 17xi17}$$

جدول(7) تصنيف البيانات للعينة عن طريق استعمال النموذج المقدر بطريقة الإمكان الأعظم

حالة المشاهدة	التنبؤ		
	المجموع	حدوث 1 $\hat{\pi} \geq 0.5$	عدم حدوث 0 $\hat{\pi} < 0.5$

المشاهدة	فشل Y=0	32866	5270	27596
	نجاح Y=1	25392	19263	6129
	المجموع	58258	24533	33725
دقة النموذج		80.43		
حساسية النموذج		75.86		
نسبة التصنيف الصحيح		83.97		

من الجدول أعلاه يتضح ان النموذج الانحدار اللوجستي الثنائي قام بتصنيف حسابات أو صفحات إلكترونية على أنها حقيقية أو مزيفة كما ياتي

1. تصنيف 27,595 حسابًا مزيفة(غير حقيقي) من بين إجمالي 32,866 حسابًا، حيث بلغت نسبة التصنيف الصحيح للحسابات المزيفة كانت 80.%
2. تم تم تصنيف 19,263 حسابًا حقيقية من بين إجمالي 25,392 حسابًا، مما يشير إلى أن نسبة التصنيف الصحيح للحسابات الحقيقية كانت 75.%
3. النسبة الكلية للتصنيف الصحيح بلغت 84%، مما يعني أن النموذج كان دقيقًا بنسبة 84% في تصنيف جميع الحسابات (سواء كانت حقيقية أم مزيفة)

4.الاستنتاجات والتوصيات:

4.1 الاستنتاجات:

1. طريقة الإمكان الأعظم المحسّنة تتفوق على طريقة الإمكان الأعظم الاعتيادية، كما هو موضح في تجربة المحاكاة التي أجريت على جميع الأحجام المفترضة حيث تم تقدير المعلمات لنموذج الانحدار اللوجستي الثنائي باستخدام هذه الطريقة، وأظهرت نتائج المحاكاة أداءً أفضل عند استخدام الإمكان الأعظم المحسّنة مقارنة بالطريقة الاعتيادية.
2. باستخدام جميع طرق التقدير التقليدية والمحسّنة، يُلاحظ تناقص في قيمة متوسط مربعات الخطأ مع زيادة حجم العينة عند إيجاد مقدرات نموذج الانحدار اللوجستي الثنائي. هذا يعني أنه كلما زاد حجم العينة، تقل احتمالية الوقوع في الخطأ.
3. أما في الجانب التطبيقي أظهرت النتائج ان اكثر العناصر تأثيرا على متغير الاستجابة هي عوامل(عدد الأشخاص أو الصفحات التي يتابعها المستخدم و طول السيرة الذاتية وتوفر الربط وتوفر الصورة الشخصية ومتوسط عدد الهاشتاج والفاصل الزمني بين المشاركة)وتعد هذا العوامل الأكثر تأثيرا لمعرفة الحساب وهمي أو حقيقي.

4.2 التوصيات:

1. توصي الباحثة باستخدام البيانات الضخمة عند تقدير معلمات نموذج الانحدار اللوجستي الثنائي.
2. .استخدام الانحدار اللوجستي المتعدد في الدراسات المستقبلية لتحسين دقة التقديرات.
3. أجراء دراسات يتم استعمال البيانات الضخمة في المجالات الصحية والتجارية والتكنولوجية لتحقيق نتائج دقيقة وفعالة.
4. إجراء دراسات حول الحالات التي تدفع الأفراد لإنشاء صفحات مزيفة على منصات التواصل الاجتماعي، لفهم الدوافع والتحديات التي تواجهها هذه الظاهرة.

المصادر:

1. AL-hinn,N,2018.Analysis the relationship Big data and knowledge Master' dissertation , Management Department Business Faculty Middle East University
2. Ali, S.H., Abood, A.H., Al-Sabbah, S.A.S." Choosing of the best estimate of the parameters of the multiple linear regression model of infertility using the weighted least squares (WLS) and robust M" Indian Journal of Public Health Research & Development, 2019
3. alrudini ,S , A (2019)" "Using genetic algorithms in estimating binary logistic regression parameters with a practical application" Master's thesis in statistics, University of Karbala.
4. Al-Sabbah, S.A.S., Abood, A.H., Mohammed, E.A.A. "The most influential factors for a successful pregnancy after in vitro fertilizatiomodel of probity regression" Annals of Tropical Medicine and Public Health"(2020)

5. Al-Sabbah, S.A.S., Radhy, Z.H., Al Ibrahim, H. Goals programming multiple linear regression model for optimal estimation of electrical engineering staff according load demand (2021) International Journal of Nonlinear Analysis and Applications, 12 (Special Issue), pp. 123-132
6. Mahdavi, I., Paydar, M. M., Solimanpur, M., & Heidarzade, A. (2009). "Genetic algorithm approach for solving a cell formation problem in cellular manufacturing". Expert Systems with Applications, 36(3), 6598-6604
7. Mikhhif,H, k(2022) "Estimating Parameters of Iog-Iogistic Regression Using Genetic Algorithm", Master's Thesis in Statistics, University of Karbala.
8. Nafi,W, 2018,The Impact of big Data on Business In , Master' dissertation, Department of Business Faculty of Business.Abdulwahhb,O,2020,High Dimensions Reduction and Estimation of Nonlinear Models for Big Data with Application, Philosophical Doctorate in Statistic, University of Baghdad
9. Salih,A,(2021), General Linear Regression Model Estimation for Big Data by Using Some of Greedy Algorithms with Application, Philosophical Doctorate in Statistic, University of Baghdad
10. saliha, ahmad saeid(2021)" Bayesian Estimation of Rank-Ordered Logistic Regression", Master's Thesis, College of Management and Economics, University of Karbala.
11. Wei, X., Q. Ling and Z. Han (2018). Robust group lasso :Mode l and recoverability. Linear Algebra and its Applications, 557, 134-173.