مقارنة بين بعض طرائق تقدير حجم العينة لتقدير معلمات أنموذج التصنيف في حالة وجود بيانات عالية الابعاد باستعمال المحاكاة(*)

الباحث: بلاسم سعدون جاسم
مديرية تربية ديالى
وزارة التربية
blasm5517@gmail.com

أ.د. دجلة إبراهيم مهدي
كلية الإدارة والاقتصاد
جامعة بغداد
dr.dejela.mahdi@gmail.com

المستخلص:

تم في هذا البحث استعمال عدة طرائق لتحديد حجم العينة الامثل لتقدير معالم البيانات ذات الابعاد العالية (High dimensional: HDD) التي يكون فيها عدد المتغيرات التوضيحية اكبر من حجم العينة $(P > n)$. وهذه الطرائق هي طريقة متباينة بونفروني وهي حالة خاصة من التقريب الطبيعي وطريقة متباينة بيرشتاين. يتم تقدير انموذج الانحدار اللوجستي الثنائي اللاخطي بحجم عينة كل طريقة في حالة بيانات عالية الابعاد باستعمال طريقة الذكاء الاصطناعي وهي طريقة الشبكة العصبية الاصطناعية (ANN) كونها تعطي تقدير عالي الدقة بما يتناسب مع نوع البيانات ونوع الدراسة الطبية. يتم توظيف القيم الاحتمالية التي تم الحصول عليها من الشبكة العصبية الاصطناعية في حساب مؤشر اعادة التصنيف الصافي (NRI)، تم كتابة برنامج لهذا الغرض باستعماله لغة البرمجة الاحصائية (R) حيث تم الاعتماد على معيار متوسط اكبر خطا مطلق (MME) لمؤشر شبكة اعادة التصنيف الصافي (NRI) للمقارنة بين طرائق تحديد حجم العينة وبوجود عدد المعلمات الافتراضية مختلفة في ظل قيمة هامش خطا معين (ε). للتحقق من اداء الطرائق باستعمال معايير المقارنة اعلاه حيث كانت اهم الاستنتاجات هي ان طريقة متباينة بيرشتاين هي الافضل في تحديد حجم العينة الامثل باختلاف عدد المعلمات الافتراضية وقيمة هامش الخطأ.

الكلمات المفتاحية: متباينة بونفروني، الشبكة العصبية الاصطناعية، متباينة بيرشتاين، متوسط أكبر خطأ مطلق.

A comparison between some methods of estimating the sample size to estimate the parameters of the classification model in the case of high-dimensional data using simulations

Researcher: Balasim Saadoun Jassim
Diyala education directorate
Ministry of Education

Prof. Dr. Dejela Ibrahim Mahdi
College of Administration and Economics
University of Baghdad

Abstract:

In this research, several methods were used to determine the optimum sample size to estimate the parameters of high-dimensional data (HDD). Where the number of explanatory variables is greater than the sample size (P> n). These methods are the Bonferroni inequality method a special case of normal approximation and the

(*) بحث مستل من رسالة ماجستير.

Bernstein's inequality method. The non-linear logistic regression model is estimated in the sample size of each method in the case of high-dimensional data using the artificial intelligence method, which is the artificial neural network method (ANN), as it gives a high-precision estimate commensurate with the data type and type of medical study. The probabilistic values obtained from the artificial neural network are used in the calculation of the net reclassification index. A program was written for this purpose using the statistical programming language (R) where the mean maximum absolute error criterion (MME) of the net reclassification network index (NRI) was used to compare the methods of specifying the sample size and the presence of a number of different default parameters under the given margin of error value (ε). To verify the performance of the methods using the comparison criteria above where the most important conclusions were that the Bernstein's inequality method is the better in determining the optimal sample size according to the number of default parameters and the error margin value.

Keywords: Bonferroni inequality, Artificial neural network, Bernstein's inequality, Mean maximum absolute error.

١. **المقدمة:**

زاد اهتمام الباحثين في السنوات الاخيرة بتحديد حجم العينة الامثل للحصول على دقة وتقدير كافيين وللحصول على معالم عالية الدقة وذلك لتقييم عدد كبير من الاختبارات في مجال التشخيص في ان واحد.

من اجل دراسة اداء دقة التشخيص للحصول على مجموعة بيانات مناسبة بحجم عينة معقول. قد لا يمكننا الحصول على حجم العينة ذات كفاءة احصائية لتحقيق نتائج معنوية ذات دقة عالية، من ناحية اخرى لا يمكن اجراء دراسة بحجم عينة كبيرة جدا يؤدي الى فقدان الدقة في النتائج المرجوة وكذلك الجهد والوقت والكلفة المادية. فهناك عدة طرق لحساب حجم العينة الامثل في مجالات العلوم الاجتماعية والاقتصادية والطبية وخصوصا في مجال الطب التشخيصي.

لذلك فان مسألة تقدير حجم العينة في هذه الحالة أصبح أصعب واعقد لذلك لابد من استعمال طرق حساب حجم عينة تتناسب مع البيانات عالية الابعاد وكذلك تتناسب مع معايير الدقة، اذ اقترح (Pencina et al.,) في عام (٢٠٠٨) مؤشر اعادة التصنيف الصافي (Net reclassification index: NRI) اذ يستعمل هذا المؤشر في تحسين دقة التشخيص عند اضافة علامات حيوية جديدة الى العلامات الحيوية الاساسية (6). والحصول على أنموذج جديد يضمن المتغيرات الاساسية بإضافة الى العلامات الجديدة، ولتكن العلامات الحيوية الجديدة على سبيل المثال الجينات التي تخص بروتين معين الخاصة بمرضى سرطان الثدي حيث يتم اضافة هذه العلامات الى العلامات الحيوية الاساسية مثلا العمر والجنس وعدد ضربات القلب ودرجة الحرارة للجسم و ضغط الدم ومعدل التنفس وغيرها لغرض تحسين دقة التشخيص للمصابين بهذا المرض وللحكم عن كل هذه العملية نستعمل معيار دقة وهو مؤشر اعادة التصنيف الصافي (NRI) لذلك لابد من تحديد حجم عينة مناسب قياس دقة التشخيص باستعمال هذا المؤشر.

٢. **هدف البحث:** يهدف البحث الى جانبين اساسيين هما:

أولاً. تحديد أفضل حجم عينة ليعطي أفضل النتائج من الممكن الاعتماد عليها في اتخاذ القرار الاحصائي والطبي.

ثانياً. ايجاد أفضل طريقة لتقدير معلمات النموذج والتي تمثل أفضل المعالم الافتراضية للبيانات عالية الابعاد.

٣. **الجانب النظري:**

٣-١. طرائق تحديد حجم العينة (Sample Size Determination Methods): وفي هذا البحث سوف نستعمل عدة طرائق لتحديد حجم العينة الامثل في ظل وجود بيانات ذات الابعاد العالية لتقدير مؤشر اعادة التصنيف الصافي (Net reclassification index) (NRI) باستعمال أنموذج الانحدار اللوجستي اللاخطي الثنائي ويمكن سرد هذه الطرق على النحو الاتي:

٣-١-١. طريقة التقريب الطبيعي (Normal Approximation Method): تفترض هذه الطريقة أن المقدر يمتلك توزيعاً طبيعياً محاذي (Asymptotical normal):

$$\sqrt{n}(\hat{\beta}_j - \beta_j) \sim N(0, \sigma_j^2), \quad k = 1, 2, \ldots\ldots, p \qquad \ldots(1)$$

وحيث ان متجه الأخطاء يفترض ان يتوزع توزيعاً طبيعيا بمتوسط صفر وتباين ثابت (σ_j^2) (Jiang and Li, 2018: 2)، يمكن تحقيق ذلك من خلال خواص تقدير المعلمات. تستخدم مثل هذه النتائج المقاربة لحساب حجم العينة عندما يكون حجم العينة n كبير فاذا كان لدينا:

$$Pr\left(\frac{\sqrt{n}|(\hat{\beta}_j - \beta_j)|}{\sigma_j} > Z_{\frac{\alpha}{2}}\right) < \alpha \qquad \ldots\ldots\ldots(2)$$

$$n^* = \frac{Z^2{}_{\alpha/2}\sigma^2{}_j}{\epsilon^2} \qquad \ldots\ldots\ldots(3)$$

اذ ان:

n^*: حجم العينة.

$Z_{\alpha/2}$: القيمة الجدولية للتوزيع الطبيعي القياسي التي تقابل مستوى المعنوية (قيمة الثقة) والتي تستخرج من جداول التوزيع الطبيعي القياسي.

$\sigma^2{}_j$: تباينات المجتمع غير المعلومة.

ان استعمال الصيغة (3) لحساب حجم العينة تضمن ان خطا التقدير لمعلمة المجتمع لـ (β_k) ضمن حدود حد الخطأ التقدير(ϵ) ، باحتمالية $(1-\alpha)$ هذه الصيغة هي مناسبة إذا كنا ندرس فقط معلمة واحدة $p = 1$ (يعني مؤشر حيوي واحد) .

٣-١-١. متباينة بونفروني (Bonf.)، (The Bonferroni Inequality): نستعمل متباينة بونفروني من اجل التوسع في حساب حجم العينة للصيغة (3) في حالة اكثر من معلمة واحدة $(P > 1)$ حيث تم تعريفها من قبل (Janos Galambos) التي تمثل مجموع الحوادث الاحتمالية التي تتوزع حسب توزيع برنولي للعزوم. (Simes, 1986: 751)

اذ ان:

$$Pr\left(\max_j|\hat{\beta}_j - \beta_j| > \varepsilon\right) \leq \sum_{j=1}^{p} Pr\left(\max_j|\hat{\beta}_j - \beta_j| > \varepsilon\right) \quad \ldots(4)$$

حيث ان مجموع الاحتمالات للقيم العليا لخطا التقدير اكبر من ε وهي قيمة صغيرة جدا حيث ان $\varepsilon > 0$. وحدود احتمال الخطأ لكل معلمة تساوي α/p . لـ jth من المقدرات حسب الصيغة الاحتمالية التالية:

$$Pr\left(\frac{\sqrt{n}\left|\widehat{\beta}_j - \beta_j\right|}{\sigma_j} > Z_{\alpha/(2p)}\right) < \alpha/p \qquad \dots\dots\dots (5)$$

ولتحقيق أكثر عمومية الى احتمال الخطأ والحصول على صيغة لحساب حجم العينة في حالة البيانات عالية الابعاد. نقوم بتعميم صيغة حجم العينة رقم (٣) باستعمال هذه المتباينة حيث نحصل على الصيغة الاتية:

$$n^* = \frac{Z^2{}_{\alpha/(2p)}\,\nu}{\epsilon^2} \qquad \dots\dots\dots (6)$$

اذ ان:

$Z_{\alpha/(2p)}$: تمثل القيمة المعيارية بمستوى ثقة معين والتي تستخرج من جداول التوزيع الطبيعي المعياري.

$$\nu = \max_j {\sigma_i}^2$$

اذ ان:

ν : تمثل القيم العظمى لتباينات المجتمع لكل j.

٣-٢. طريقة المتباينات (Inequalities Method): في بعض الأحيان يكون تطبيق التقريب الطبيعي باستخدام نظرية الغاية المركزية غير مرغوب فيه وقد لا يكون التوزيع احادي القمة المتماثل مناسبًا لشكل التوزيع وذلك لوصف المعلمات المقدرة في العينات غير المحدودة وخاصة عندما تكون المعلمات الحقيقية ضمن الحدود الطبيعية .

٣-٢-١. متباينة بير شتاين (Bernstein inequality) (Berns): هي طريقة لحساب حجم العينة تم اقتراحها من قبل الباحث (Van der Laan and Bryan) في عام (٢٠٠١)م باستعمال متباينة بيرن شتاين (Bernstein inequality). (Van der Laan and Bryan, 2001: 445)

$$Pr\left(|\widehat{\beta}_j - \beta_j| > \epsilon\right) = pr\left(\left|\sum_{i=1}^{n}[\psi_{ij} - \beta_j]\right| > n\epsilon\right) \leq 2exp\left(\frac{-n\varepsilon^2}{2\nu_j + 2\frac{M\epsilon}{3}}\right)$$

ولتحقيق حدود احتمال الخطأ تحت مستوى معنوية (α).

نستعمل أسلوب (*Bonferroni* Correction) يعتبر أحد الطرق المهمة للتصدي لمشكلة الاختبارات المتعددة.

من خلال (Bonferroni Correction) يمكن ان نعمل على تجزئة لمستوى المعنوية الاختبار لكـل معلمـة في النمـوذج وتقليص عدد المعـالم التي تكون تحت مستـوى المعنوية المخصص لكل منها وجعلها مساوية للصفر تماما، من خلال ذلك نحدد النموذج الملائم للبيانات تحت الدراسة.

وذلك بأخذ الطرف الايمن للمتباينة الذي يساوي تقدير احتمال الخطأ (α/P) (Van Der vaart and Wallner, 1996: 102).

$$2exp\left(\frac{-n\varepsilon^2}{2v_j + 2\frac{M\epsilon}{3}}\right) = \frac{\alpha}{p} \qquad \cdots\cdots\cdots (7)$$

من المعادلة اعلاه نحصل على صيغة لحساب حجم العينة التالية:

$$n^* = \frac{2v + \frac{2M\epsilon}{3}}{\varepsilon^2}\left(\log p + \log\frac{2}{\alpha}\right) \qquad \cdots\cdots\cdots (8)$$

اذ ان:

p: تمثل عدد المعلمات. $M > 0$: قيمة ثابتة اكبر من الصفر.

ε: وتمثل مقدار الخطأ المسموح به في تقدير المتوسط أو النسبة (ويساوي مضروب القيمة المعيارية عند مستوى ثقة معين في الخطأ المعياري للمتوسط أو مضروباً في الخطأ المعياري للنسبة في حالة تقدير النسبة).

٣-٣. نموذج الانحدار اللوجستي الثنائي: الانحدار اللوجستي هو أحد عناصر مجموعة من النماذج تسمى بمجموعة النماذج الخطية العامة ويستعمل انموذج الانحدار اللوجستي لوصف العلاقة بين متغير الاستجابة الذي يكون من النوع المتقطع والمتغيرات التوضيحية تكون من النوع المتقطع او المستمر او الخلط بينهما كما يعتبر من الاساليب المهمة في تصنيف البيانات وتخصيصها لفئات معينة وتكون فيه العلاقة بين متغير الاستجابة (y) والمتغيرات التوضيحية (x_i) اياً كان نوعها كمياً او نوعياً غير خطية حيث يكون متغير الاستجابة (y) ثنائي الاستجابة يتوزع برنولي (Bernoulli distribution) مفترضاً إحدى القيمتين (1,0)، وان احتمال حدوث الاستجابة (π_i) واحتمال عدم حدوث الاستجابة $(1-\pi_i)$. كما يعتبر حالة خاصة من نماذج الانحدار العامة (GLRM) فأن دالة الكثافة الاحتمالية يمكن كتابتها على النحو الاتي: (Scott, 2005: 5)

$$p(Y = y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} \qquad \cdots\cdots\cdots (9)$$

اذ ان:

$$y_i = 0, 1$$

y_i : متغير الاستجابة الثنائي.

π_i : احتمال حدوث الاستجابة عندما $y_i = 1$.

فان توقع متغير الاستجابة يمثل احتمال حدوث الاستجابة.

$$E(y_i) = p(Y = 1) = \pi_i \qquad \cdots\cdots\cdots (10)$$

فيكون تباين متغير الاستجابة حسب توزيع برنولي (Bernoulli distribution) هو.

$$V(y_i) = \pi_i(1-\pi_i) \qquad \cdots\cdots\cdots (11)$$

وليكن $x_1, x_2, \ldots\ldots x_k$ مجموعة من المتغيرات التوضيحية وان n تمثل عدد المشاهدات لهذه المتغيرات التي تكون المصفوفة X .

$$X = (x_{ij})_{n\times k}$$

اذ إن:

$\mathbf{i = 1, 2, \cdots\cdots, n}$ ، $\mathbf{(n)}$ تمثل حجم العينة.

$\mathbf{j = 1, 2, \cdots\cdots, k+1}$ ، $\mathbf{(k)}$ عدد المتغيرات التوضيحية , $\boldsymbol{p = k+1}$ يمثل عدد المعلمات.

فاذا كان $\mathbf{y_i = [y_1, y_2, \ldots\ldots y_n]}$ عينه عشوائية من المتغير ثنائي الاستجابة وأن $\mathbf{y_i \in \{0, 1\}}$.

وبذلك تكون صيغة أنموذج الانحدار اللوجستي كالاتي:

$$\mathbf{y_i = \pi_i + e_i} \qquad \cdots\cdots\cdots (12)$$

اذ إن:

$\boldsymbol{\pi_i}$ تمثل دالة الانحدار اللوجستي (احتمال الاستجابة).

$$\boldsymbol{\pi_i = \frac{e^{\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j}}{1 + e^{\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j}}} \qquad \cdots\cdots\cdots (13)$$

اذ ان:

$\boldsymbol{\beta_j}$: متجه المعلمات من درجة $\boldsymbol{(1 \times p)}$ اذا كان:

$\boldsymbol{i = 1, 2, \ldots, n\ ,\ j = 1, 2, \ldots, p}$

$\boldsymbol{\underline{X_{ij}} = \{x_{i1}, x_{i2}, \ldots x_{ij}\}}$: مصفوفة من المتغيرات التوضيحية من درجة $\boldsymbol{(n \times p)}$

$\boldsymbol{n}$: يمثل عدد المشاهدات.

$\boldsymbol{p}$: يمثل عدد المتغيرات التوضيحية.

حيث ان حد الخطأ العشوائي $\mathbf{(e_i)}$ يتوزع توزيع برنولي بمتوسط صفر كما في الصيغة:

$$\mathbf{e_i = y_i - \pi_i} \qquad \cdots\cdots\cdots (14)$$

وبأخذ التوقع لطرفي المعادلة $\mathbf{(14)}$ نحصل على:

$$\mathbf{E(e_i) = \pi_i - \pi_i = 0} \qquad \cdots\cdots\cdots (15)$$

أما تباين حد الخطأ العشوائي فانه يساوي تباين متغير الاستجابة الثنائي.

$$\mathbf{V(e_i) = \pi_i(1 - \pi_i)} \qquad \cdots\cdots\cdots (16)$$

ولذا فأن حد الخطأ العشوائي يكون له متوسط صفر وتباين $\mathbf{\pi_i(1 - \pi_i)}$ ومن الملاحظ ان تباين حد الخطأ يعتمد على قيم احتمال الاستجابة $\mathbf{(\pi_i)}$ (Shen and Gao, 2008: 4).

٤-٣. التحويل الخطي الى دالة الانحدار اللوجستي الثنائي: بسبب الانحناءات السلبية في معلمات الانحدار اللوجستي والتي تؤثر هذه الانحناءات على خصائص مقدرات المعلمات وقيم حدوث الاستجابة التي يمكن التنبؤ بها لذلك يلجأ الكثير من الاحصائيين الى التحويل الخطي من خلال استعمال تحويل دالة (Logit) (Logit function) من اجل ازالة انحناء معلماتها حيث تكون نتائج الاختبار مضللة بسبب ان المعلمات لا تتوزع طبيعيا وتكون متحيزة وتبايناتها لا تكون اقل ما يمكن. و قد قام الباحث (Berkson) عام (١٩٤٤) م بإيجاد علاقة لوغاريتمية من اجل تحويل العلاقة بين المتغيرات التوضيحية و احتمال حدوث الاستجابة $\mathbf{(\pi_i)}$ إلى علاقة خطية عن طريق

تحويل الاحتمال $\mathbf{pr(Y=1|X)}$ إلى دالة يكون مدها من $(-\infty,+\infty)$ وهذه الدالة تسمى نسبة الارجحية والتي تعطى بالصيغة الاتية:

$$\frac{\mathbf{pr(Y=1|X)}}{\mathbf{pr(Y=0|X)}}=\frac{\mathbf{pr(Y=1|X)}}{\mathbf{1-pr(Y=1|X)}} \qquad \cdots\cdots\cdots(17)$$

حتى نحصل على دالة مداها من $(-\infty,+\infty)$ نأخذ اللوغاريتم الطبيعي لنسب الأرجحية (Odde) ويكون بالصيغة الاتية:

$$\mathbf{logit}\,(\pi_i)=\mathbf{Ln}\left(\frac{\pi_i}{1-\pi_i}\right) \qquad \cdots\cdots\cdots(18)$$

$$\mathbf{Ln}\left(\frac{\pi_i}{1-\pi_i}\right)=\mathbf{Ln}\left(e^{\beta_0+\sum_{j=1}^{p}X_{ij}\beta_j}\right) \qquad \cdots\cdots\cdots(19)$$

$$\mathbf{Ln}\left(\frac{\pi_i}{1-\pi_i}\right)=\beta_0+\sum_{j=1}^{p}X_{ij}\,\beta_j \qquad \cdots\cdots\cdots(20)$$

ويمكن اعادة كتابة الأنموذج كما في الشكل الاتي:

$$\mathbf{Ln}\left(\frac{\pi_i}{1-\pi_i}\right)=\beta_0+\beta_u U+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_j x_{ij}\,, \qquad \cdots(21)$$

اذ ان:

$$\mathbf{i=1,2\ldots,n} \quad , \quad \mathbf{j=1,2,\ldots,p}$$

$\mathbf{U}$: متغير الاساسي

$\mathbf{X_{i1},X_{i2},\ldots,X_{ij}}$: تمثل المشاهدات الخاصة بالمتغيرات التوضيحية.

$\mathbf{\beta_0,\beta_1,\beta_2,\ldots,\beta_j}$: معلمات نموذج الانحدار اللوجستي.

وهذا يعني ان أنموذج الانحدار اللوجستي الثنائي أصبح خطي باستعمال تحويل دالة (Logit) (Kleinbaum and Klein, 2002: 17-20)

٥-٣. الشبكة العصبية الاصطناعية (Artificial Neural Net Work (ANN)) الشبكات العصبية الاصطناعية (ANN) عبارة عن شبكات من الخلايا العصبية الاصطناعية متصلة ببعضها البعض لتعمل كعناصر معالجة بسيطة من أجل القيام بمهمة محددة. ان المعرفة التي اكتسبتها الشبكة من بيئتها تشبه معالجة الخلايا العصبية في الدماغ البشري، حيث يتم الحصول على المعلومات من عملية التعلم وتخزينها من نقاط القوة في الاتصال بين الخلايا العصبية، والتي تسمى بالأوزان ايضاً.

هناك انواع مختلفة من الشبكات فمنها شبكات متعددة الطبقات (Multilayer Net Work) والتي استعملت بشكل واسع لحل مشاكل الانحدار اذ يحتوي نموذج الشبكة العصبية متعددة الطبقات على طبقة إدخال (عقد المصدر)، أو طبقة مخفية واحدة أو أكثر (عقد حسابية) وطبقة الإخراج (عقد حساب). حيث ان طبقة الإدخال لها وظيفة استقبال إشارات الداخلة وأن الطبقات المخفية مسؤولة عن معالجة ونشر الإشارات المستقبلة (الداخلة) واخراجها. (Haykin, 2008: 27) فقد تكون لدى بعض الشبكات العصبية متعددة الطبقات اتصالات خاصة تسمى (Skip Layer)،

حيث يكون لبعض إشارات الإدخال اتصال مباشر بطبقة الإخراج أي تنفيذ الطبقة المخفية. (Riple, 1996: 144)

$$y_i = \phi_i\left(b_i + \sum_{l=1}^{m} W_{kl}X_l + \sum_{i=1}^{n} W_{il}.\phi_k\left(\sum_{j=1}^{m} W_{kj}X_j + b_j\right)\right) \cdots\cdots (22)$$

تبين المعادلة (٩) الشبكة العصبية الاصطناعية مع تخطي طبقة الاتصالات حيث تتم معالجة نتيجة عقدة الادخال (k) بواسطة عقدة الاخراج (l) وسوف يتم استعمال (**دالة التنشيط اللوجستية**) والتي تأخذ الصيغة الاتية:

$$\phi(z) = \frac{1}{1+e^{-z}} \cdots\cdots\cdots (23)$$

واعتماداً على دالة التنشيط (Activation function) سوف تكون مخرجات الشبكة العصبية قيم ثنائية محصورة بين (٠، ١) وكما ان هذه الشبكة تمثل الاتصال المباشر بين عقدة الادخال وعقد الاخراج وتسمى هذه الشبكة بالشبكة العصبية أحادية الطبقة المخفية مع تخطي طبقة الاتصالات. (Habibnia and Maasoumi, 2019: 6).

٦-٣. مقاييس دقة التشخيص والتصنيف

(Measures of Diagnostic and classification accuracy):

يعد التشخيص والتصنيف اساساً في الممارسة الطبية في حالة مرض معين فقد تكون الحاجة الى التمييز بين الوضع الحالي للمرض وحالة غياب المرض، فقد يشمل التصنيف على أكثر من فئتين من المرضى يلزم معالجتهما بشكل منفصل، وغالباً ما تستعمل المؤشرات الحيوية (Biomarker) للتنبؤ بحالة مرض معين وذلك بالاعتماد على مجموعة من الادوات الاحصائية لتقييم دقة التشخيص (Li and Fine, 2013: 383)، وسوف نستعمل أحد هذه المقاييس المهمة وهو مقياس مؤشر أعادة التصنيف الصافي (NRI) ومن ثم الحصول على أفضل حجم عينة يعطي أدق تشخيص للمؤشرات الحيوية الجديدة المضافة الى المؤشرات الحيوية الأساسية بهدف التشخيص او الفحص الطبي. (Pencina D'Agostino and Steyerberg, 2011: 12)

٧-٣. مؤشر أعادة التصنيف الصافي (Net reclassification index: NRI):

هو مقياس تم اقتراحه من قبل (Pencina and others) في العام (٢٠٠٨) لتقييم التنبؤ بالعلامة الجديدة المضافة الى الأنموذج الأساسي في حالة وجود فئات متعددة. (Pencina D'Agostino and Steyerberg, 2011: 12)

والصيغة التقديرية لمؤشر (NRI) هي على النحو الاتي:

$$\widehat{S_j} = \sum_{m=1}^{M} \frac{\omega_m}{n_m} \sum_{i=1}^{n} I\{\widehat{p}_{mi}(\mathcal{M}_2) = max\widehat{p}_i(\mathcal{M}_2), \widehat{p}_{mi}(\mathcal{M}_1) \neq max\widehat{p}_i(\mathcal{M}_1),\ Y_i = m\} \cdots\cdots (24)$$

اذ ان:

$\widehat{p}_i(\mathcal{M}_j) = \left(\widehat{p}_{1i}(\mathcal{M}_j), \widehat{p}_{2i}(\mathcal{M}_j), \ldots\ldots, \widehat{p}_{Mi}(\mathcal{M}_j)\right)$ تمثل الاحتمالات المقدرة

ويعتبر أفضل مقياس دقة تصنيف الذي يمتلك أكبر قيمة لمؤشر أعادة التصنيف الصافي (NRI).

الجانب التجريبي

٤. **مفهوم المحاكاة:** عبارة عن اسلوب يتضمن استخدام نموذج رياضي نظري ومشابهته بالأنموذج الحقيقي الذي يمثل المشكلة المدروسة أو كما عرفه (Naylor) انه اسلوب يستخدم للتحكم بالتجربة المراد دراستها ومعرفة صفاتها وخصائصها باستخدام القابلية الرياضية والمنطقية المتوفرة في الحاسبة الالكترونية أو كما عرفه آخرون اسلوب يتم من خلاله ايجاد نموذج بديل مماثل للنموذج الحقيقي من دون محاولة الحصول على النموذج الحقيقي نفسه. (بتال، ٢٠٠٨: ١٧٥)

٤-١. مراحل بناء تجارب المحاكاة: لتحليل البيانات باستعمال نماذج المحاكاة لا بد من تحديد مراحل بناء المحاكاة والتي تكون كما يلي:

٤-١-١. توليد المتغيرات التوضيحية: يتم توليد المتغيرات التوضيحية وفق التوزيع الطبيعي متعدد المتغيرات وفق الصيغة الاتية:

$$X_1, X_2, \dots, X_p \sim \mathbf{MN}(\mathbf{0}, \boldsymbol{\Sigma})$$

اذ ان:

$$\boldsymbol{\Sigma} = (\sigma_{ij})_{p\times p} \quad , \quad \sigma_{ij} = 0.5^{|i-j|} \quad , 1 \le \mathbf{i}, \mathbf{j} \le \mathbf{p}$$

اما بالنسبة لمتغير الاساسي يتم توليد من التوزيع الطبيعي القياسي وكما يأتي:

$$U \sim \mathbf{N}(\mathbf{0}, \mathbf{1})$$

٤-١-٢. تحديد المعلمات الاولية: تم وصف القيم الافتراضية لتجارب المحاكاة وفقاً للجدول التالي:

الجدول (١): القيم الافتراضية للمعلمات الاولية

$(Case\ one)$	عدد المعلمات $= p$	$(parameter)$ $\boldsymbol{\beta} = (\beta_0, \beta_u, \beta_1, \dots, \beta_p)$
1	$p = 75$	$\boldsymbol{\beta} = (1.5, 1.5, 1.5, 0.5.0.5, 0.5), \beta_j = 0, j = 7, \dots, 75$
2	$p = 150$	$\boldsymbol{\beta} = (1.5, 1.5, 1.5, 0.5.0.5, 0.5), \beta_j = 0, j = 7, \dots, 150$
3	$p = 350$	$\boldsymbol{\beta} = (1.5, 1.5, 1.5, 0.5.0.5, 0.5), \beta_j = 0, j = 7, \dots, 350$
4	$p = 550$	$\boldsymbol{\beta} = (1.5, 1.5, 1.5, 0.5.0.5, 0.5), \beta_j = 0, j = 7, \dots, 550$

٤-١-٣. حساب متغير الاستجابة (Response Variable): يتم حساب متغير الاستجابة الاحتمالي بالاعتماد على أنموذج الانحدار اللوجستي وعلى النحو الاتي:

$$y = \log \frac{p}{1-p} \quad \dots\dots\dots\dots (25)$$

$$\log \frac{p}{1-p} = \beta_0 + \beta_u U + \beta_1 X_1 + \dots + \beta_p X_p \quad \dots\dots (26)$$

٤-١-٤. توليد مؤشر اعادة التصنيف الصافي: بعد ان تم توليد الانموذج الاحتمالي في الفقرة (٤-١-٣) يتم توظيف قيمه الاحتمالية لحساب مؤشر اعادة التصنيف الصافي الذي يمثل الخلط بين نموذجين لوجستيين احتماليين أحدهما يتضمن المتغير الاساسي ومتغير الاستجابة الاحتمالي الذي يرمز له $\hat{p}(\mathcal{M}_1)$ والذي يأخذ الصيغة الاتية:

$$\hat{p}_i(\mathcal{M}_1) = \log\frac{p}{1-p} \quad \cdots\cdots \quad (27)$$

$$\log\frac{p}{1-p} = \beta_0 + \beta_u U \quad \cdots\cdots \quad (28)$$

والاخر يتضمن المتغير الاساسي بالإضافة الى المتغيرات التوضيحية ومتغير الاستجابة الذي يرمز له $\hat{p}(\mathcal{M}_{2j})$ الذي يأخذ الصيغة الاتية:

$$\hat{p}_i(\mathcal{M}_{2j}) = \log\frac{p}{1-p} \quad \cdots\cdots \quad (29)$$

$$\log\frac{p}{1-p} = \beta_0 + \beta_u U + \beta_1 X_1 + \cdots + \beta_p X_p \quad \cdots\cdots \quad (30)$$

٤-١-٥. مناقشة تجارب المحاكاة: تم استعمال برنامج لغة البرمجة الإحصائية (R) حيث تم الاعتماد على معيار متوسط اكبر خطا مطلق (MME) لمؤشر إعادة التصنيف الصافي (NRI) وكانت النتائج كالاتي:

الجدول (٢): مقارنة طرائق تحديد حجم العينة

NRI	$(P = 75)$ $\beta = (1.5, 1.5, 1.5, 0.5, 0.5, 0.5)\,, \beta_o = 0.5, \beta_u = 0.5$		
Methods		Bonf. Inq.	Berns. Inq.
$(\varepsilon = 0.1)$	n^*	1158	1655
	$MAE\,\hat{\varepsilon}$	0.006074626	0.005433601
	(MME)	0.02074634	0.02327051
$(\varepsilon = 0.05)$	n^*	4632	6512
	$MAE\,\hat{\varepsilon}$	0.002983268	0.001549753
	(MME)	0.01916864	0.00746901

الجدول (٣): مقارنة طرائق تحديد حجم العينة

NRI	$(P = 150)$ $\beta = (1.5, 1.5, 1.5, 0.5, 0.5, 0.5)\,, \beta_o = 0.5, \beta_u = 0.5$		
Methods		Bonf. Inq.	Berns. Inq.
$(\varepsilon = 0.1)$	n^*	1287	1798
	$MAE\,\hat{\varepsilon}$	0.00872644	0.001934587
	(MME)	0.03047043	0.03465371
$(\varepsilon = 0.05)$	n^*	5149	7076
	$MAE\,\hat{\varepsilon}$	0.00209387	0.001082647
	(MME)	0.01151935	0.006489755

الجدول (٤): مقارنة طرائق تحديد حجم العينة

NRI	$(P=350)$ $\beta=(1.5,1.5,1.5,0.5,0.5,0.5)\,,\beta_o=0.5,\beta_u=0.5$		
Methods		Bonf. Inq.	Berns. Inq.
$(\varepsilon=0.1)$	n^*	1446	1973
	$MAE\,\hat{\varepsilon}$	0.006740543	0.002581634
	(MME)	0.0293	0.02326463
$(\varepsilon=0.05)$	n^*	5786	7765
	$MAE\,\hat{\varepsilon}$	0.001964876	0.001197568
	(MME)	0.01093737	0.01370418

الجدول (٥): مقارنة طرائق تحديد حجم العينة

NRI	$(P=550)$ $\beta=(1.5,1.5,1.5,0.5,0.5,0.5)\,,\beta_o=0.5,\beta_u=0.5$		
Methods		Bonf. Inq.	Berns. Inq.
$(\varepsilon=0.1)$	n^*	1532	2065
	$MAE\,\hat{\varepsilon}$	0.004702194	0.003010647
	(MME)	0.0508137	0.02102493
$(\varepsilon=0.05)$	n^*	6127	8131
	$MAE\,\hat{\varepsilon}$	0.002059838	0.00110434
	(MME)	0.01076622	0.005256992

نتائج المحاكاة: نلاحظ من نتائج الجدول (٢) و (٣) عندما تكون عدد المعلمات (P =75)، (P =150) وقيم المعلمات الافتراضية اعلاه في الجدول بان افضل طريقة لتحديد حجم العينة عندما تكون قيمة هامش الخطأ $(\varepsilon=0.1)$ هي طريقة (Berns. Inq.) كونها تملك اعلى متوسط اكبر خطا مطلق (MME)، وتعتبر طريقة (Bonf. Inq.) افضل طريقة لتحديد حجم العينة عندما تكون قيمة هامش الخطأ $(\varepsilon=0.05)$ كونها تملك اعلى متوسط اكبر خطا مطلق (MME).

ونلاحظ من نتائج الجدول رقم (٤) عندما تكون عدد المعلمات (P =350) وقيم المعلمات الافتراضية اعلاه في الجدول بان أفضل طريقة لتحديد حجم العينة عندما تكون قيمة هامش الخطأ $(\varepsilon=0.1)$ هي طريقة (Bonf. Inq.) كونها تملك اعلى متوسط اكبر خطا مطلق (MME)، وتعتبر طريقة (Berns. Inq.) افضل طريقة لتحديد حجم العينة عندما تكون قيمة هامش الخطأ $(\varepsilon=0.05)$ كونها تملك اعلى متوسط اكبر خطا مطلق (MME). ونلاحظ من نتائج جدول رقم (٥) عندما تكون عدد المعلمات (P =550) وقيم المعلمات الافتراضية اعلاه في الجدول بان أفضل طريقة لتحديد حجم العينة عندما تكون قيمة هامش الخطأ $(\varepsilon=0.1,\varepsilon=0.05)$ هي طريقة (Bonf. Inq.) كونها تملك اعلى متوسط أكبر خطا مطلق (MME).

٥. الاستنتاجات والتوصيات

٥-١. الاستنتاجات: من خلال تجارب المحاكاة وما تم عرضه من نتائج استنتج الباحث ما يأتي:

١. من خلال الجانب التجريبي اثبتت النتائج أن أفضل طريقة تحديد حجم العينة هي طريقة (Bonf. Inq.) وعند قيمة هامش الخطأ $(\varepsilon = 0.1, 0.05)$.

٢. هناك علاقة عكسية بين حجم العينة وهامش الخطأ اذ نلاحظ انه كلما قل هامش الخطأ زادت حجم العينة والعكس صحيح.

٣. تم الحصول على أفضل تقدير للمعلمات الافتراضية للبيانات ذات الابعاد العالية من خلال طريقة (Bonf. Inq.) لكونها تمتلك اعلى متوسط أكبر خطأ مطلق (MME).

٥-٢. التوصيات: على ضوء الاستنتاجات التي توصلنا اليها من خلال البحث يمكن ادراج التوصيات الاتية:

١. نوصي باستعمال طريقة (Berns. Inq.) لتحديد حجم العينة في حالة البيانات ذات الابعاد العالية وللدراسات وخصوصاً في الجانب الطبي وفي الطب التشخيصي.

٢. نوصي باستعمال مؤشر اعادة التصنيف الصافي (NRI) للدراسات الطبية الخاصة بالطب التشخيصي للأمراض.

٣. استعمال اساليب مختلفة غير الشبكات العصبية الاصطناعية لتقدير انموذج الانحدار اللوجستي الثنائي وبالتالي تقدير مؤشر اعادة التصنيف الصافي (NRI).

المصادر

أولاً. المصادر العربية:

١. بتال، احمد حسين، احمد، عصام كامل، خضير، البراء عبد الوهاب (٢٠٠٨)، استخدام المحاكاة في تدريس الانحدار الخطي البسيط، مجلة جامعة الانبار للعلوم الصرفة، العدد الثالث، المجلد الثاني.

ثانياً. المصادر الأجنبية:

1. Haykin, S, (2008), Redes Neurals.2rd Edition ,Principiose pratica.
2. Habibnia, A., & Maasoumi, E., (2019), Forecasting in Big Data Environments: an Adaptable and Automated Shrinkage Estimation of Neural Networks (AAShNet) p.1-26
3. Jiang B., & Li J., (2018), Sample size determination for high dimensional parameter estimation with application to biomarker identification. Computational Statistics and Data Analysis, 118, 54–65.
4. Li, J., Jiang, B., and Fine, J., (2013), Multicategory reclassification statistics for assessing improvements in diagnostic accuracy. Biostatistics 14, 382–394.
5. Kleinbaum, D. G., & Klein, M., (2002), Logistic Regression A Self-Learning Text. 3d ,Springer.
6. Pencina, M., D'Agostino, R., and Vasan, R., (2008), evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Statistics in Medicine 27, pp- 157–172
7. Pencina, M., D'Agostino, R., and Steyerberg, E., (2011), Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Statistics in Medicine 30, pp- 11–21.

8. R. J. Simes, (1986), An Improved Bonferroni procedure for multiple tests of significance. Biometrika, Vol. 73, No. 3, pp. 751-754.
9. Riple, B. D., (1996), Pattern Recognition and Neural Network . Cambride University.
10. Shen, J., & Gao, S., (2008), A Solution to Separation and Multicollinearity in Multiple Logistic Regression. Journal of Data Science : JDS, 6(4), PP-515-531.
11. Scott, M., (2002), Applied Logistic Regression Analysis. sage.
12. Van Der Vaart, A. and Wellner, J. A., (1996), Weak Convergence and Empirical Processes. New York: Springer.
13. Van der Laan, M., and Bryan, J., (2001), Gene expression analysis with the parametric bootstrap. Biostatistics 2, PP- 445-461.