أساليب التنقيب في البيانات: الطرق المعلمية واللامعلمية

Data Mining Techniques: Parametric and Nonparametric Methods

رزق السيد حامد الوزير* حاتم عبد الواحد سمري**

ملخص:

ظهر مصطلح التنقيب في البيانات لأول مرة في منتصف التسعينات على يد Fayaad et al[13:17]، وكان مرتبطاً وقتها بتقنيين الخطوات التي يجب أن تسير عليها المنشأة لانتهاج تقنية التنقيب في البيانات؛ وهو ما يُعرف اليوم باسم عملية التنقيب في البيانات واكتشاف المعرفة. ثم انهالت البحوث لتساند هذا المدخل بتعديل تلك العملية وتقديم أساليب جديدة وتطويع أساليب قديمة لحل المشاكل وتوفيق النماذج واختبار مصداقيتها في ظل مجموعات البيانات الكبيرة. وقد لاقى ذلك قبولاً واسع الانتشار في المنظمات الكبرى في الغرب لأنها وجدت أن دخولها لهذا العالم يساعد في تحقيق أهدافها وتحسين مراكزها التنافسية بشكل كبير.

ومن أكثر العلوم التي ساهمت في علم التنقيب في البيانات علوم الإحصاء وتعليم الآلة ونظم المعلومات. وتُعد أساليب الشبكات العصبية وشجرة القرارات والعملية التحليلية الهرمية والانحدار اللامعلمي وتحليل التناظر من أكثر الأساليب الحديثة التي تُعد أساليب صريحة للتنقيب في البيانات. كما طُورت و/أو استُخدمت بعض الأساليب الإحصائية التقليدية مثل تحليل المكونات الرئيسة والتحليل العاملي وتحليل التمايز والتحليل العنقودي ونماذج البروبيت واللوجيت وطريقة أقرب الجيران والنماذج الجمعية المعممة والبرمجة الرياضية ودوال الانحدار المقسمة المتوائمة لتكّمل الأساليب الحديثة في عملية تحليل البيانات والتنقيب بها.

وكان من الطبيعي أن يواكب هذا التطور برمجيات جديدة تحتوي على تلك الأساليب الحديثة، ولكن استخدام هذه البرامج وبالتالي هذه الأساليب في البحوث العربية مازال في أضيق الحدود بسبب جدة هذه المواضيع وندرة منشوراتها باللغة العربية وبالتالي صعوبة فهمها. لذا، فقد أحجم معظم الباحثين عن تلك الأساليب في توفيق العلاقة بين المتغير التابع والمتغيرات المستقلة مستعينين بنموذج الانحدار الخطي المتعدد لسهولة فهمه واستخدامه. غير أن التطبيقات الحديثة أثبتت ضعف مصداقية نموذج الانحدار الخطي المتعدد في توفيق معظم المشاكل المعاصرة التي تتسم باللاخطية ووجود تفاعلات بين المتغيرات بفعل مجموعات البيانات الكبيرة. ويهدف هذا البحث إلى التعريف بهذه الأساليب سواء كانت معلمية أو لامعلمية أو نصف معلمية، والتعرف على التطبيقات الحديثة التي استُخدمت فيها تلك الأساليب بنجاح.

الكلمات الدالة: التنقيب في البيانات؛ توفيق النماذج؛ الأساليب المعلمية واللامعلمية والنصف معلمية

Summary:

The term data mining is used for the first time in the mid-nineties by Fayaad et al [13:17], and was associated at the time with the steps that must go by the establishment to pursue technical data mining; which is known today as the data mining and knowledge discovery. Then poured research to support this approach to amend the process and the introduction of new methods and adapting old methods to solve problems and reconcile models and test their credibility in light of large data sets. Has met widespread acceptance in large organizations in the West because they found that entry to this world would help achieve their goals and improve their competitive positions significantly.

The most sciences that contributed to the science of data mining are statistics, machine learning and information systems. The most modern explicit methods for data mining are neural networks, decision trees, the analytic hierarchy process, nonparametric regression and analysis of symmetry. Some traditional statistical methods (like principal components analysis, factor analysis, discriminant analysis, cluster analysis, probit and logit models, nearest neighbors method, generalized additive models and mathematical programming are developed and / or used to complement modern methods of data analysis process within the framework of data mining science.

It was natural to keep pace with this new software development containing these modern methods, but the use of these programs and therefore these methods in research is still in the minimalistic because the modernity of these topics and scarcity of publications in Arabic and therefore difficult to understand. So, it has declined most researchers for those methods to reconcile the relationship between the dependent variable and independent variables with the aid a multiple linear regression model for easy to understand and use. However, the modern applications proved weak credibility of multiple linear regression model in fitting most contemporary problems that characterized nonlinearity and the presence of interactions between variables due to large data sets. The aim of this research is to promote these methods whether parametric, nonparametric or semi-parametric, and meet modern applications which those methods were used successfully.

Keywords: Data mining; Fitting models models; parametric, nonparametric and semiparametric methods

---

* مدرس الإحصاء التطبيقي بكلية التجارة جامعة المنصورة، وأستاذ مساعد الإحصاء بكلية العلوم الإدارية والمالية جامعة الطائف

** مدرس مساعد كلية تجارة جامعة الزقازيق، وطالب دكتوراه معهد الإحصاء جامعة القاهرة.

تاريخ تقديم البحث : 2013/4/29 تاريخ قبول البحث : 2013/5/22

1. الإطار النظري

يتناول هذا البحث بالدراسة المداخل المختلفة للتنقيب في البيانات. ويتعلق ذلك بالنماذج المفترضة لحل المشاكل، وأساليب التقدير المستخدمة، وطرق اختبار صلاحية النماذج والمقارنة بينها لاختيار أكفأها.

1-1 تصميم البحث

يمكن تصنيف طرق التنقيب في البيانات في نوعين أساسيين: الطرق المعلمية، والطرق اللامعلمية بالإضافة للطرق نصف المعلمية. لذلك فسوف يأتي هذا البحث في أربعة فصول: يُخصص الأول منها للجانب النظري؛ ويحتوي على تصميم البحث ومشكلته والدراسات السابقة. ويُخصص الثاني للطرق المعلمية؛ ويحتوي على نموذج الانحدار الخطي، وتحليل المكونات الرئيسة، والتحليل العاملي، وتحليل التمايز، وتحليل التناظر التمييزي، والتحليل العنقودي، ونماذج البروبت واللوجيت. ويُخصص الثالث للطرق اللامعلمية؛ ويحتوي على العملية الهرمية التحليلية، والأنظمة الخبيرة، وطريقة أقرب الجيران، والنماذج الجمعية المعممة، والبرمجة الرياضية، ودوال الانحدار المقسمة المتوائمة، وشجرة القرارت، والشبكات العصبية، والانحدار اللامعلمي. كما خُصص الفصل الرابع وهو الأخير للطرق نصف المعلمية؛ ويحتوي على الانحدار نصف المعلمي.

ويهدف البحث إلى التعريف بأساليب التنقيب في البيانات وأنواعها وخوارزميات تدريبها، كما يهدف إلى استعراض المجالات المختلفة التي استُخدمت فيها تلك الأساليب بنجاح. وقد اعتمد البحث على دراسة العديد من البحوث التي تتضمن أساليب مختلفة للتنقيب في البيانات للتعريف بهذه الأساليب وخوارزميات تقديرها وتطبيقاته وشروط استخدامها بشكل مبسط، ومقارنة بين أفضلية أساليب التنقيب والأساليب التقليدية.

1-2 مشكلة البحث وأهميته

التنقيب في البيانات مجال حديث، وتحليل البيانات إحدى مراحله. وتحتوي برمجياته على أساليب حديثة، ولكن استخدام هذه البرامج وبالتالي هذه الأساليب في البحوث العربية مازال في أضيق الحدود بسبب جدة هذه المواضيع وندرة منشوراتها باللغة العربية وبالتالي صعوبة فهمها. لذا، فقد أحجم معظم الباحثين عن تلك الأساليب في توفيق العلاقة بين المتغير التابع والمتغيرات المستقلة مستعينين بنموذج الانحدار الخطي المتعدد لسهولة فهمه واستخدامه. غير أن التطبيقات الحديثة أثبتت ضعف مصداقية نموذج الانحدار الخطي المتعدد في توفيق معظم المشاكل المعاصرة التي تتسم باللاخطية ووجود تفاعلات بين المتغيرات بفعل مجموعات البيانات الكبيرة.

وتبرز أهمية البحث في أن استخدام الباحثين لهذه الأساليب في تحليل البيانات في ظل الأحجام الكبيرة يجعل نتائج التحليل أكثر مصداقية ويبعد النموذج عن خطأ التوصيف.

1-3 الاستفادة من البحث

يفيد هذا البحث كافة الباحثين الذين يتعاملون مع أحجام بيانات كبيرة، وكافة المنشآت التي تقرر عملية محددة للتنقيب في البيانات. وبمعنى آخر فإن البحث يفيد في التطبيقات المالية مثل تحليل بيانات أسواق الأسهم وتقييم الجدارة الائتمانية لطالبي القروض، وتحليل سلة السوق، وتشخيص الأعطال.

ويحاول البحث جذب انتباه الباحثين –وخاصةً غير الإحصائيين منهم– إلى استخدام تلك الأساليب الجديدة للتنقيب في البيانات، والإشارة إلى التطبيقات الجديدة في هذا المجال.

1-4 الدراسات السابقة

تعاملت بعض دراسات التنقيب في البيانات مع عملية التنقيب في البيانات[13:17] [11]. وبينت الخطوات التي يجب أن تسير عليها المنشأة بغية اكتشاف المعرفة والأنماط الهامة التي لم تكن معروفة من قبل، كما أشار البعض منها لأساليب التنقيب في البيانات ومجالات استخدامها. وقد اهتمت بعض الدراسات[54] بحصر وتقديم تعريفات لهذه الأساليب، وركز البعض الآخر على دراسة الجوانب النظرية[6:10][20:32] والخوارزميات[3:5][35][40][41] التي تستخدمها تلك الأساليب في عملية التقدير، كما اهتم البعض الآخر بتقديم البرمجيات[34][49:52] وأدوات المساعدة لهذه الأساليب. وقد تبين من الدراسات التي اهتمت بتطبيقات أساليب التنقيب في البيانات[21] [23][38][47] أن أهم التطبيقات الناجحة كانت: تحليل سلة السوق، وتقييم الجدارة الائتمانية، وتحليل أسواق الأسهم، واكتشاف الغش والأعطال، وتشخيص الأمراض.

1-5 أسئلة البحث

يجيب البحث عن الأسئلة التالية:

ما هي أساليب التنقيب في البيانات؟

ومتى تُستخدم؟ وما هي خوارزميات تقديرها؟

وما هي التطبيقات الناجحة التي استُخدمت فيها تلك الأساليب؟

وقد تم عمل ذلك لكل من: نموذج الانحدار الخطي، وتحليل المكونات الرئيسة، والتحليل العاملي، وتحليل التمايز، وتحليل التناظر التمييزي، والتحليل العنقودي، ونماذج البروبت واللوجيت، والعملية الهرمية التحليلية، والأنظمة الخبيرة، وطريقة أقرب الجيران، والنماذج الجمعية المعممة، والبرمجة الرياضية، ودوال الانحدار المقسمة المتوائمة، وشجرة القرارت، والشبكات العصبية، والانحدار اللامعلمي، والانحدار نصف المعلمي.

2. الأساليب المعلمية

2-1 الانحدار الخطي المتعدد MLR

يُعد أسلوب الانحدار الخطي المتعدد Multiple Linear Regression أقدم وأشهر الأساليب الإحصائية التي استُخدمت في حقل التنقيب في البيانات. وهو نموذج تنبؤي يلجأ إليه المحللون عند الرغبة في تقييم العلاقة السببية بين أحد المتغيرات الكمية وعدة متغيرات أخرى. ويُطلق على المتغير الذي نريد تفسير التغير فيه أو التنبؤ بقيمه في المستقبل عدة أسماء: المتغير التابع dependent variable، متغير الاستجابة response، أو المتغير المُفَسر explained variable ويأخذ الرمز y_i. كما يُطلق على المتغيرات الأخرى اسم: المتغيرات المستقلة independent vr's، المتغيرات المُفسرة explanatory vr's، المتنبئات predictors، السمات features أو covariates أو benchmarks وتأخذ الرمز x_{ij} ؛ حيث $i = (1,2,...,n)$ وهو ما يشر إلى المشاهدات بينما $j = (1,2,...,p)$ وهو ما يشير إلى المتغيرات.

ويأخذ نموذج الانحدار الخطي المتعدد (بصيغة المصفوفات) الشكل التالي:

$$\underset{n\times 1}{Y} = \underset{n\times p}{X}\ \underset{p\times 1}{B} + \underset{n\times 1}{\ell} \qquad (1)$$

ويتم التوصل لشكله التحليلي بتقدير متجه المعالم B بإحدى طرق التقدير. وتُعتبر طريقة المربعات الصغرى LS أشهر هذه الطرق، حيث يتم اختيار مستوى يصغر مجموع مربعات البواقي ℓ. ويمكن فحص جودة توفيق النموذج من خلال الأدوات التشخيصية برسم البواقي مقابل القيم المقدرة من خط الانحدار $\hat{y}_i$ ثم النظر إلى الشكل الناتج. فإذا كان الانحدار صادقا، فإن قيم المتغير التابع يجب أن تتوزع حول الخط المقدر عشوائياً بدون أن تشكل أي اتجاه عام واضح. كما يمكن فحص جودة توفيق نموذج الانحدار بالاعتماد على مؤشر تلخيصي يُعرف باسم معامل التحديد R^2 الذي يأخذ قيمة تتراوح بين الصفر والواحد، إذ كلما اقتربت قيمته من الواحد، كلما دل ذلك على إمكانية التنبؤ بقيم y_i بشكل أصدق اعتماداً على العلاقة التي تربطها بقيم $x_i's$. وأخيراً، يتم اختبار المعنوية الإجمالية للنموذج باستخدام اختبار F، واختبار المعنوية الجزئية للمتغيرات المستقلة باستخدام اختبار t.

فإذا كان لدينا مجموعة بيانات واحدة (متغير كمي واحد وعدة متغيرات مستقلة) وكانت الأخيرة لا تعتمد على بعضها (بمعنى عدم وجود ازدواج خطي متعدد multicolinearity)، يمكن تطبيق الانحدار الخطي المتعدد بأمان. أما في حالة وجود مجموعتي بيانات (مجموعة للمتغيرات المستقلة ومجموعة للمتغيرات التابعة) أو أكثر (مجموعة للمتغيرات المستقلة وعدة مجموعات للمتغيرات التابعة) أو كان هناك ازدواج خطي في حالة مجموعة البيانات الواحدة، فإن الانحدار الخطي لا يصلح ويمكن تطبيق أحد الأساليب التالية.

2-2 تحليل المكونات الرئيسة PCA

لا بد –عند التعامل إحصائياً مع أي مشكلة– من التعبير عنها بما يسمى بمصطلحات التنقيب في البيانات *بجدول البيانات*. وجدول البيانات data table هو عبارة عن مصفوفة من الدرجة $n\times p$، تشير فيه الصفوف n إلى القياسات التي أخذتها وحدات المعاينة في p من المتغيرات الخاضعة للدراسة.

ويهدف تحليل المكونات الرئيسة Principal Components Analysis إلى ضغط جدول البيانات في ظل القياسات المرتبطة والتعبير عنه بمجموعة جديدة من المتغيرات غير المرتبطة (المتعامدة) وهو ما يُعرف باختزال الأبعاد. وعندئذٍ، يُقال أن المتغيرات الجديدة تعتمد على السياق context أو أنها المكونات الرئيسة principal components أو العوامل factors أو المتجهات المميزة eigenvectors أو المتجهات المنفردة singular vectors أو التحاميل loadings. كما تُمثل أيضاً كل وحدة (صف) بمجموعة من الدرجات scores تناظر تقديرها في المكونات.

ويبدأ تحليل المكونات الرئيسة بمعايرة جميع المتغيرات ثم حساب مصفوفة التغاير S. وتُطبق عملية تكرارية تهدف للتوصل إلى k من المكونات الرئيسة حيث k<p. وتبدأ هذه العملية بالحصول على المكون الرئيس الأول الذي يصف جميع المتغيرات الموجودة، أي الحصول على متجه المعاملات (الأوزان)

$$a_1 = (a_{11}, a_{21}, \ldots . a_{p1})'$$

الناتج عن حل مشكلة تعظيم التباين في Y_1:

$$\max \text{var}(Y_1) = \max(a_1' S a_1)$$

باستخدام مضاعفات لاجرانج في ظل القيد $a_1' a_1 = 1$. ثم الحصول على المكون الثاني، وهكذا حتى الحصول على المكون رقم k، أي الحصول على متجه المعاملات (الأوزان)

$$a_k = (a_{1k}, a_{2k}, ...a_{pk})'$$

الناتج عن حل مشكلة تعظيم التباين في Y_k:

$$\max \text{var}(Y_k) = \max(a_k' S a_k)$$

باستخدام مضاعفات لاجرانج في ظل القيود $a_k' a_k = 1,\ a_2' a_1 = a_3' a_2 = ... = a_k' a_{k-1} = 0$.

ويتم رسم أرقام المكونات الرئيسة على المحور الأفقي مقابل القيم المميزة لها على المحور الرأسي، وهو ما يُعرف برسم الأحجار scree plot، ويُختار المكون ذو أقصى ارتفاع.

2-3 التحليل العاملي FA

يُستخدم التحليل العاملي Factor analysis في اختزال الأبعاد، فهو يختصر المتغيرات من عدد أكبر إلى عدد أقل من العوامل عند نمذجة البيانات. ويختار FA مجموعة فرعية من المتغيرات من مجموعة أكبر استناداً إلى أعلى الارتباطات بين المتغيرات الأصلية مع عوامل المكونات الرئيسة. ويُعد ذلك مدخلاً لعلاج الازدواج الخطي المتعدد عند توفيق نموذج الانحدار المتعدد لأن مجموعة العوامل الناتجة تكون متغيرات غير مرتبطة. لذلك فإن التحليل العاملي يُستخدم في بناء ما يسمى بنماذج المتغيرات المستترة latent vr's، وهي المتغيرات غير المشاهدة التي لا يوجد لها قياسات مسجلة وإنما هي مستنتجة من متغيرات أخرى مشاهدة (من خلال نموذج رياضي) لها قياسات مسجلة. كما يُستخدم أيضاً لاكتشاف الهيكل في العلاقات بين المتغيرات، وهو ما يُعرف باسم تصنيف المتغيرات classify variables.

وينقسم التحليل العاملي إلى نوعين: التحليل العاملي الاستكشافي Exploratory factor analysis والتحليل العاملي التوكيدي Confirmatory factor analysis

- فالتحليل العاملي الاستكشافي EFA: هو الذي يبحث في طبيعة أبنية العلاقات المؤثرة على المتغيرات التابعة (أي هياكل النماذج أو أشكالها البنائية).
- والتحليل العاملي التوكيدي CFA: يختبر أي من هذه الهياكل يؤثر على المتغيرات التابعة عند التنبؤ.

ويتعلق التحليل العاملي بسابقه تحليل المكونات الرئيسة، لكنهما ليسا شيئا واحد. إذ يستخدم FA أساليب نمذجة الانحدار لاختبار حدود الخطأ، في حين أن PCA هو مجرد أسلوب إحصائي وصفي.

2-4 تحليل التمايز DA

يُستخدم تحليل التمايز Discriminant Analysis –في علوم الإحصاء والتعرف على الأنماط pattern recognition وتعليم الآلة machine learning– لإيجاد التوليفة الخطية من المتغيرات المستقلة الكمية التي تميز أو تفصل فئتين أو أكثر من الأحداث (متغير تابع تصنيفي). وبمعني آخر، فإن DA هو طريقة لتصنيف القياسات في مجموعتين أو أكثر. فالغرض الرئيس من DA هو التنبؤ بما يسمى بعضوية المجموعة group membership استناداً إلى توليفة خطية من المتغيرات الكمية. ويبدأ الأسلوب بمجموعة مشاهدات ذات قيم معلومة وذات مجموعات معروفة، وينتهي بنموذج يسمح بالتنبؤ بعضوية المجموعة بمعلومية المتغيرات المستقلة الكمية فقط. والغرض الثاني لتحليل التمايز هو فهم مجموعة البيانات بالفحص

الدقيق لنموذج التنبؤ لأخذ فكرة عن العلاقة بين عضوية المجموعة والمتغيرات المستقلة المستخدمة في التنبؤ بتلك العضوية.

على سبيل المثال، فإن لجنة القبول بالجامعة قد تقسم خريجيها إلى مجموعتين: الطلاب الذين أنهوا البرنامج في خمس سنوات أو أقل، والطلاب خلاف ذلك. ويمكن استخدام DA للتنبؤ بالاستكمال الناجح لبرنامج الدراسة للطلاب الجدد على أساس درجاتهم في اختبار القدرات GRE score ومعدلهم التراكمي في الثانوية undergraduate grade point average. ويعطي فحص نموذج التنبؤ فكرة عن مدى مساهمة كل متغير (بمفرده وبالاشتراك مع المتغيرات الأخرى) في إكمال أو عدم إكمال البرنامج.

ويتشابه DA مع كل من تحليل التباين ANOVA وتحليل الانحدار RA، اللذان يعبران أيضاً عن المتغير التابع بتوليفة من المتغيرات المستقلة. غير أن المتغير التابع في الأسلوبين الأخيرين يشترط أن يكون كمياً، على عكس الحال في DA الذي يكون فيه تصنيفياً. كما يقترب الانحدار اللوجستي logistic regression والانحدار الاحتمالي probit regression أيضاً بشدة من DA، إذ يفسر الكل متغير تصنيفي ما. غير أن الأسلوبين الأولين يفضلا في التطبيقات التي لا تفترض أن المتغيرات المستقلة تتبع التوزيع الطبيعي، وهو الفرض الأساسي الذي بُني عليه DA.

ويتشابه أيضاً DA مع كل من تحليل المكونات الرئيسة PCA والتحليل العاملي FA في أن الكل يبحث عن التوليفات الخطية للمتغيرات التي تعطي أفضل تفسير للبيانات. وإذا كان DA يحاول نمذجة الفرق بين فئات البيانات بصراحة، فإن PCA لا يأخذ في حسبانه أي فرق في الفئات، كما يبني FA توليفات المتغيرات على الفروق بدلاً من التشابه. كما يختلف DA عن FA في أنه ليس أسلوب تداخل interdependence technique: يتم فيه التمييز بين المتغيرات المستقلة والمتغير التابع.

وأخيراً، فإن DA يُطبق عندما تأخذ المتغيرات المستقلة قياسات كمية مستمرة. أما عندما نتعامل مع متغيرات مستقلة تصنيفية، فإن الأسلوب المكافئ يكون تحليل التناظر التمييزي discriminant correspondence analysis.

2-5 تحليل التناظر التمييزي DCA

كما يشير الاسم، فإن تحليل التناظر التمييزي هو امتداد لكل من تحليل التمايز DA وتحليل التناظر CA. ويهدف DCA (مثل DA) إلى تصنيف المشاهدات في مجموعات معرفة مسبقاً (ومثل CA) في أنه يُستخدم مع المتغيرات الاسمية. إن الفكرة الأساسية وراء DCA هي تمثيل كل مجموعة بإجمالي مشاهداتها وإجراء CA بسيط على المجموعات عن طريق مصفوفة المتغيرات. ويتم التنبؤ بالمشاهدات الأصلية وتخصيص كل مشاهدة متوقعة في المجموعة الأقرب. ويمكن استخدام المقارنة بين التصنيفين القبلي والبعدي priori and the a posteriori classifications لتقييم جودة التمايز. كما يمكن تقييم ثبات التحليل باستخدام أساليب التحقق المبدل من الصحة cross-validation techniques.

2-6 التحليل العنقودي Cluster Analysis

يُعد التحليل العنقودي من أشهر الطرق الوصفية (الاستكشافية) للتنقيب في البيانات، وهو منهج لتجميع grouping مجموعة معينة من المشاهدات. فإذا تكونت مصفوفة البيانات من n من المشاهدات (الحالات أو الصفوف) و p من المتغيرات (الحقول أو الأعمدة)، فإن هدف التحليل العنقودي يكون عنقدة أو تصنيف المشاهدات في مجموعات متجانسة (متماسكة) داخلياً internal cohesion وغير متجانسة من

مجموعة إلى أخرى (منفصلة خارجياً external separation). ويُفسَر ذلك على أنه اختزال للأبعاد في الفضاء R^n، ولكن ليس بنفس طريقة المكونات الرئيسة. إذ يقوم التحليل العنقودي بالاختزال الرأسي بتجميع المشاهدات n في g من المجموعات الفرعية (بحيث تكون $g < n$)، بينما يقوم تحليل المكونات الرئيسة بتحويل المتغيرات الأصلية p إلى k من المتغيرات الجديدة (بحيث يكون $k < p$).

ويمكن تكوين التجمعات groupings أو التقسيمات partitions أو العناقيد clusters بنوعين من الطرق:

- الطرق الهرمية hierarchical methods: ويتم فيها تقدير عدد العناقيد بإجراء أسلوب التعاقب succession بدءً من n (وهي الحالة الأبسط التي تُعامل فيها كل مشاهدة على أنها مجموعة منفصلة) حتى 1 (كل المشاهدات تنتمي لمجموعة واحدة).
- الطرق غير الهرمية non-hierarchical methods: ويكون فيها عدد العناقيد معروف مسبقاً.

7-2 نماذج البروبيت واللوجيت

إن نموذج البروبيت أو نموذج الوحدة الاحتمالي probit model هو نوع خاص من الانحدار يكون فيه المتغير التابع من النوع التصنيفي (الثنائي binary) و يأخذ قيمتين فقط، النجاح ويُشار إليه بالرمز 1 والفشل ويُشار إليه بالرمز 0. ومثال ذلك: متزوج وغير متزوج، ناجح وراسب، يفضل ولا يفضل، الإجابة بنعم أو لا، ووجود أو غياب صفة معينة ... إلخ.

ويأخذ نموذج البروبيت الشكل التالي:

$$\Pr(Y = 1 \mid X) = \Phi(X'\beta),$$

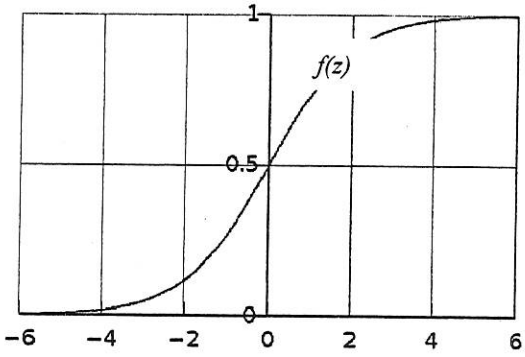
حيث يشير الرمز Pr إلى الاحتمال، والرمز Φ إلى دالة التوزيع المتجمعة للتوزيع المعتاد المعياري، والرمز β إلى المعالم المقدرة باستخدام طريقة الإمكان الأكبر التقليدية.

أما نموذج اللوجيت logit model فهو كسابقه أسلوب أحادي/متعدد المتغيرات يسمح بتقدير احتمال وقوع/عدم وقوع حدث ما لمتغير تابع ثنائي، ولكنه يأخذ الشكل التالي:

$$y=\exp(b_0 + b_1*x_1 + \ldots + b_n*x_n)/\{1+\exp(b_0 + b_1*x_1 + \ldots + b_n*x_n)\}$$

ويُعد نموذج الانحدار اللوجستي LRM مثالاً لهذا النوع من النماذج. وتأخذ فيه الدالة اللوجستية دائماً – مثل الاحتمالات– قيماً تتراوح بين الصفر والواحد، وتُعرّف بالنموذج والشكل التاليين:

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}}$$



شكل (1): مثال للدالة اللوجستية

ويشير المتغير z (الذي يمكن أن يأخذ أي قيم عددية) إلى مدخلات الدالة، بينما تنحصر قيم المخرجات $f(z)$ بين الصفر والواحد. ويمثل المتغير z التعرض لمجموعة ما من المتغيرات المستقلة، بينما

تمثل $f(z)$ احتمال الناتج المقابل في ظل قيم المتغيرات المفسرة. ويقيس المتغير z المساهمة الكلية لجميع المتغيرات المستقلة المستخدمة في النموذج ويُطلق عليه اسم logit، ويُعرّف بالمعادلة:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k,$$

ويُعد الانحدار اللوجستي طريقة مفيدة في وصف العلاقة بين متغير مستقل أو أكثر (مثل العمر، والسن، إلخ) ومتغير استجابة ثنائي يأخذ قيمتين فقط (النجاح أو الفشل).

3. الأساليب اللامعلمية

3-1 العملية الهرمية التحليلية AHP

تُعَرّف العملية الهرمية التحليلية Analytical hierarchy process بأنها أسلوب تفرُع structured technique لتنظيم وتحليل القرارات المعقدة. وهي عبارة عن مزيج من علم الرياضيات وعلم النفس قُدم على يد Thomas L. Saaty في السبعينات للتوصل لأفضل قرار من بين البدائل المتاحة[53][1].

وقد لاقت هذه العملية استحساناً كبيراً من جانب صانعي القرار وقبولاً واسعاً في المجالات الحكومية والتعليم والصحة والصناعة ومنشآت الأعمال. فبدلا من أن تفترض المنشأة بأن قرارها "صحيح"، فإن AHP تساعد على العثور على أفضل قرار يتوافق مع هدفها وفهمها للمشكلة. وتوفر AHP إطارا شاملا وعقلانية لبناء مشكلة القرار، ولتمثيل وقياس عناصرها، ولربط هذه العناصر لتحقيق الأهداف العامة، ولتقييم الحلول البديلة.

وأول ما يفعله مستخدمو AHP هو تفكيك مشكلة القرار بشكل هرمي إلى مشاكل فرعية يمكن فهمها بسهولة أكثر، بحيث يمكن تحليل كل منها بشكل مستقل. ويمكن أن تتعلق عناصر التسلسل الهرمي بأي جانب من جوانب المشكلة، سواء كانت تلك العناصر ملموسة أو غير ملموسة، وسواء قيست بدقة أو قُدرت بشكل تقريبي، وسواء فُهمت جيداً أو بشكل ضعيف.

وبمجرد بناء الهرم، يقوم صانعو القرار بتقييم عناصره المختلفة بمقارنة كل عنصر بالعناصر الأخرى اثنين في كل مرة من حيث تأثيرها على العنصر الذي يعلوها في التقسيم الهرمي. وعند عمل المقارنات، يمكن لصانع القرار أن يستخدم بيانات واقعية عن العناصر، كما يمكنه أيضاً استخدام حكمه عن الأهمية النسبية للعناصر. وهكذا، فإن جوهر AHP يعتمد على استخدام الأحكام الشخصية إلى جانب المعلومات الأساسية في إجراء التقييمات.

وتحول AHP هذه التقييمات إلى قيم عددية[2] يمكن معالجتها ومقارنتها على المدى الكامل للمشكلة. ويُشتق الوزن العددي أو ما يُعرف بالأولوية priority بالنسبة لكل عنصر من التسلسل الهرمي، مما يسمح بالمقارنة مع العناصر المتنوعة وغير القابلة للقياس في كثير من الأحيان مع بعضها البعض بطريقة عقلانية ومتسقة. وتميز هذه القدرة AHP عن غيرها من أساليب صنع القرار.

وفي الخطوة الأخيرة من العملية، يتم حساب الأولويات العددية لكل بديل من بدائل القرار. وتمثل هذه الأرقام القدرة النسبية للبدائل في تحقيق الهدف المقرر.

3-2 الأنظمة الخبيرة ES

يُعَرّف النظام الخبير expert system (في مجال الذكاء الاصطناعي artificial intelligence) بأنه نظام حاسبي يحاكي قدرة الخبرة البشرية في صناعة القرار[25]. وتُصمم الأنظمة الخبيرة لحل المشاكل المعقدة عن طريق المنطق المكتسب من المعرفة، أي بطريقة الخبير وليس بإتباع أسلوب المطوّر كما هو

الحال في البرمجة الاتفاقية[9][38][45]. وقد قُدمت أول الأنظمة الخبيرة في السبعينات ثم انتشرت بعد ذلك في الثمانينات[10].

ويتكون النظام الخبير من قسمين: الأول ثابت مستقل عن النظام هو محرك الاستنتاج the inference engine، والثاني متغير يمثل قاعدة المعرفة the knowledge base. وفي الثمانينات ظهر قسم ثالث يسمح بالاتصال بالمستخدمين هو واجهة الحوار[27].

وقد تم تصميم النظم الخبيرة لتسهيل المهام في مجالات المحاسبة، والقانون، والطب، التحكم في العمليات، والخدمات المالية، والإنتاج، والموارد البشرية. لذلك فقد ساندتها تطبيقات كثيرة في مجالات تشخيص الأعطال، والتشخيص الطبي، ودعم القرارت في الأنظمة المعقدة، والرقابة على العمليات، والبرامج التعليمية، وإدارة المعرفة.

3-3 طريقة أقرب الجيران k-NN

تُعد طريقة أقرب الجيران Nearest Neighbors من الطرق المبنية على الذاكرة[43]، بمعنى أنها لا تتطلب (بمصطلحات التنقيب في البيانات) أي تدريب (توفيق نموذج للبيانات) على خلاف الطرق الإحصائية الأخرى. وتستند k-NN على فكرة بديهية تتلخص في أن المشاهدات القريبة يجب أن تقع في نفس الفئة. فهي أسلوب تصنيف[54] يقرر في أي فئة سنضع الحالة الجديدة بفحص عدد ما (k) في معظم الحالات المشابهة أو الجيران. ويلجأ المحلل لهذه الطريقة عند عمل التحاليل المقارنة باستخدام أساليب اختزال البيانات.ويتطلب تطبيق الطريقة[7] معايرة جميع المتغيرات وحساب المسافات الإقليدية بين كل زوج من المشاهدات. ويمكن تصنيف المشاهدة الجديدة بواحدة من 4 طرق هي: طريقة Papadakis التي تُسمى أحياناً Genesis (وهي مبنية على حساب البواقي ثم استخدام طريقة تحليل التغاير)، وطريقة الارتباط (وهي مبنية على استخدام الارتباط بين كل زوج من المشاهدات من خلال المربعات الصغرى المعممة بشرط معلومية هيكل الارتباط)، وطريقة Wilkinson، وطريقة تمهيد المربعات الصغرى.

ومن أهم تطبيقات k-NN: التعرف على الأنماط، والتصنيف الإحصائي، والتحليل العنقودي، واسترجاع المحتوى المبني على الصور من قواعد البيانات، والتسوق عبر الإنترنت.

3-4 النماذج المعممة المضافة GAMs

إن النماذج المعممة المضافة generalized additive models هي[20] أحد مداخل الانحدار اللامعلمي في حالة تعدد المتغيرات المستقلة. وقد قُدم هذا الأسلوب[23][55] في التسعينات على يد Trevor Hastie and Rob Tibshirani. ويخلط هذا النموذج بين خصائص النماذج الخطية المعممة generalized linear models والنماذج المضافة additive models. وإذا كان النموذج الخطي الجمعي يأخذ الشكل التالي:

$$Y = b_0 + b_1*X_1 + \ldots + b_m*X_m$$

فإن GAM يأخذ الشكل المختلف التالي[51]:

$$g(\mathrm{E}(Y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m).$$

ويُلاحظ بمقارنة النموذجين أن GAMs أخذت من النموذج المتعدد حفاظها على الشكل الجمعي، غير أنها استبدلت الحدود البسيطة في المعادلة الخطية (أي b_i*X_i) بالدوال $f_i(X_i)$ وهي دوال لامعلمية

للمتغيرات المفسرة X_i. وبعبارة أخرى، فإن GAMs تُقدّر دوال لامعلمية غير محددة لكل متغير مستقل بدلاً من المعامل للتوصل لأفضل تنبؤ لقيم المتغير التابع.

ويمكن توفيق الدوال $f_i(X_i)$ باستخدام[28] أحد مُمَهِدات الشكل الانتشاري scatterplot smoother التالية:

1) شرائح التمهيد المكعبة cubic smoothing spline وهي متوفرة في برنامج SAS،
2) أسلوب LOESS وهو أيضاً متاح في البرنامج السابق
3) مُمَهِد النواة Kernel smoother وهو متاح في برنامج STATA
4) الشرائح الرقيقة thin-plate splines التي تسمح بوجود تفاعل بين المتغيرات المستقلة وهو متاح في برنامجي SAS و R.

وأخيراً، تُعد GAMs مفيدة في الحالات التالية[20][39]: 1) إذا كان شكل العلاقة بين المتغيرات شديد التعقيد بشكل يصعب معه توفيق نموذج خطي تقليدي أو أي من النماذج غير الخطية 2) إذا لم يتوفر سبب مسبق لاستخدام نموذج معين 3) إذا كنا نريد أن تقترح البيانات الشكل الدالي المناسب. ويعني ذلك أن تلك النماذج تناسب معظم التطبيقات الحديثة التي تحتوي على عدد كبير من المتغيرات بينها تفاعلات ممكنة في ظل أحجام البيانات الكبيرة مثل أسواق الأسهم.

3-5 البرمجة الرياضية MP

تشير البرمجة الرياضية[30] mathematical programming (في كل من الرياضيات وعلم الإدارة وعلوم الحاسب) إلى الأمثلية optimization؛ أي عملية اختيار أفضل الحلول من بين عدة بدائل متاحة في ظل مجموعة من القيود. وتتكون مشكلة الأمثلية في شكلها البسيط من تعظيم أو تصغير دالة حقيقية باختيار قيم المتغيرات الهامة من بين مجموعة من المتغيرات وحساب قيمة دالة الهدف. ويسمح تعميم مشكلة الأمثلية بوجود تشكيلة متنوعة من دوال الهدف وأنواع مختلفة من النطاقات.

ويتيح زر البرامج الإضافية[31] Add-in في برنامج Excel بناء نماذج البرمجة الرياضية وحلها باستخدام حلاّل المشاكل Solver Add-in. فعندما يتم تحميل البرنامج الإضافي Math Programming add-in، يتم إضافة سطور أوامر لكل من: البرمجة الخطية والكسرية linear and integer programming، والبرمجة غير الخطية nonlinear programming، وشبكات الأعمال network، ومشاكل النقل transportation.

3-6 دوال الانحدار المقسمة المتوائمة متعددة المتغيرات MARS

تُعد دوال الانحدار المقسمة المتوائمة متعددة المتغيرات[18][32] Multivariate Adaptive Regression Splines شكلاً من أشكال تحليل الانحدار. وهي أسلوب انحدار لامعلمي ينمذج اللاخطية والتفاعلات، ويبني النماذج بالشكل التالي:

$$\hat{f}(x) = \sum_{i=1}^{k} c_i B_i(x)$$

وهو عبارة عن مجموع لدوال الأساس $B_i(x)$ المرجحة بالمعاملات الثابتة c_i. وتأخذ كل دالة أساس basis function أحد الأشكال الثلاث التالية: 1) الثابت 1 وهو ما يسمح بظهور حد التقاطع intercept في النموذج 2) دالة مفصلية hinge function على الشكل $\max(0, x - const)$ أو الشكل $\max(0, const - x)$، حيث تختار MARS المتغيرات وقيم العقد knots تلقائياً 3) حاصل ضرب دالتين مفصليتين أو أكثر.

ويمكن توفيق نموذج MARS على مرحلتين؛ بنفس المنهج المستخدم في التقسيم المتكرر recursive partitioning عند توفيق شجرة القرارات:

1) المرور للأمام the forward pass: ويبدأ بنموذج يحتوي على حد التقاطع فقط (متوسط قيم المتغير التابع) ثم إضافة زوج من دوال الأساس إلى النموذج في كل مرة إلى أن نصل لأقصى اختزال في الخطأ المعبر عنه بمجموع مربعات البواقي. غير أن التوفيق الأمامي عادةً ما يبني نموذج ذو جودة توفيق فوقية overfit (نموذج ذو جودة توفيق جيدة بالنسبة للبيانات المستخدمة في بنائه، غير أن أدائه التنبؤي بالنسبة للبيانات الجديدة يكون ضعيف).
2) المرور للخلف the backward pass: وهو استكمال للمرحلة السابقة للتغلب على مشكلة التوفيق الفوقي (لتحسين القدرة التنبؤية) بتقليم prunes النموذج عن طريق حذف حدوده واحداً تلو الآخر، حيث يتم حذف الحد الأقل تأثيراً في كل خطوة إلى أن يتم الوصول إلى أفضل نموذج فرعي. ويُقارن أداء النماذج الفرعية باستخدام طريقة التحقق من الصحة المتقاطعة المعممة Generalized cross validation (GCV) لاختيار أفضل نموذج فرعي؛ حيث تشير القيمة الأقل لـ GCV . لنموذج أفضل. وتُحسب GCV بالصيغة التالية:

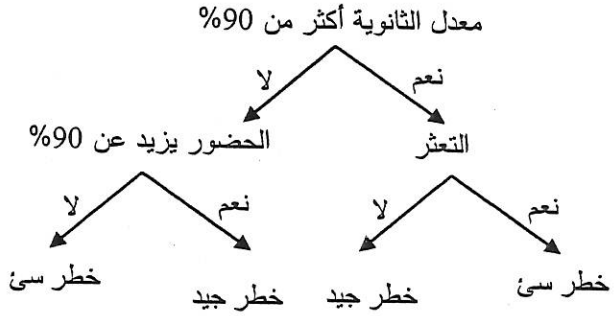
$$GCV = RSS / (N * (1 - \textit{EffectiveNumberOfParameters} / N)^2)$$

7-3 شجرة القرارات DT

تُعرّف شجرة القرارات[20] decision tree ببساطة على أنها طريقة (بيانية و/أو نموذجية) لتمثيل سلسلة من القواعد تقودنا إلى فئة أو قيمة. وهي من أهم النماذج التنبؤية للتنقيب في البيانات، وهي أيضاً من أهم أدوات دعم اتخاذ القرار. وتُعد نماذج الأشجار[21] أسلوب تكراري (مُعاد/يكرر نفسه) recursive procedure يتم فيه تقسيم n من المشاهدات في m من المجموعات بطريقة متتالية حسب قاعدة تقسيم

معينة تهدف إلى تعظيم مقياس الانسجام أو النقاوة homogeneity or purity measure للمتغير التابع في كل مجموعة فرعية.

فعند رغبة الجامعة مثلاً في تصنيف الطلاب الجدد المتقدمين لها من حيث مخاطر التعثر إلى جيد وسيئ، يمكن استخدام شجرة قرارات بسيطة لحل هذه المشكلة على النحو التالي:



شكل (2) شجرة تصنيف بسيطة

مكونات الشجرة:

يوضح شكل (2) أن شجرة القرارات تحتوي على 3 مكونات؛ العُقَد والفروع والأوراق:

1) العُقَد nodes: وهي الجذور التي تناظر العناوين أو المتغيرات (معدل الثانوية، والتعثر، والحضور). وعقدة الجذر root node هي عقدة قمة القرار top decision node الذي تتفرع منه جميع التصنيفات.
2) الفروع branches: ويتفرع من كل عقدة عدد من الفروع، كل فرع يعبر عن أحد الإجابات الممكنة (تقود كل عقدة في المثال التوضيحي إلى فرعين؛ نعم ولا). وتعتمد عدد الفروع الناتجة من كل عقدة على الخوارزمية المستخدمة في تقدير شجرة القرارات. فعلى سبيل المثال، فإن خوارزمية CART تولّد شجرة ذات فرعين من كل عقدة. وعندئذٍ، يُطلق على الشجرة اسم الشجرة ثنائية الفروع binary tree. أما عندما تتفرع العقدة لأكثر من فرعين، فإن الشجرة تصبح متعددة الفروع multiway tree ويمكن تقديرها بخوارزمية[5] CHAID.
3) الأوراق leaves: ويقودنا كل فرع إلى عُقَد أخرى (التعثر والحضور)، أو إلى قاع الشجرة (خطر جيد وخطر سيئ) وهي الأوراق.

قواعد التقسيم:

يتيح علم التنقيب في البيانات فحص البيانات واستنتاج الشجرة وقواعدها التي ستُستخدم في عمل التنبؤات بواحدة من الخوارزميات الأربعة التالية:

1) خوارزمية CHAID: أي اكتشاف التفاعل الذاتي باستخدام χ^2 Chi-squared Automatic Interaction Detection.
2) خوارزمية CART: أي أشجار التصنيف والانحدار Classification And Regression Trees. وتُستخدم أشجار التصنيف مع المتغير التابع المتقطع أو التصنيفي، بينما تُستخدم أشجار الانحدار حين يكون المتغير التابع مستمر.
3) خوارزمية QUEST: أي الشجرة الإحصائية السريعة غير المتحيزة الكفؤة Quick, Unbiased and Efficient Statistical Tree. وهي تشبه خوارزمية CART وتختلف عنها في أنها[56]: أسلوب غير متحيز لاختيار المتغيرات، وتستطيع التعامل بسهولة مع المتغيرات المفسرة التصنيفية متعددة الفئات.
4) خوارزمية C4.5 وخوارزمية C5.0: تُستخدم خوارزمية C4.5 لتوليد شجرة القرارات في حالة التصنيف. أما خوارزمية C5.0 فهي الإصدار الأحدث من الأولى والتي تتميز بأنها: أسرع، وتولد شجرة أصغر، وتدعم عملية التعزيز boosting التي تحسن الشجرة وتعطيها دقة أكثر[4][41].

قواعد التوقف:

لا يمكن ترك الأشجار تنمو بلا حدود لأن ذلك سيتطلب وقتاً أكبر لبنائها كما ستكون غير قابلة للفهم، ولكن الخطورة الأكبر لذلك هي أنها ستُنتج توفيق فوقي للبيانات. فمن الطبيعي أن نتحكم في حجم الشجرة، وهو ما يُعرف بقواعد التوقف stopping rules التي تحدد النمو. ويُعد تحديد العمق الأكبر maximum depth الذي يمكن أن تنمو به الشجرة من أشهر قواعد التحكم. كما أن هناك قاعدة توقف أخرى مبنية على الحد الأدنى لعدد السجلات في العقدة، ولا يُسمح بالتفرع بعد هذا الحد. ويمكن أيضاً أن

يحل التقليم prune بديلاً لقواعد التوقف بالسماح للشجرة بالنمو إلى أقصى حد ممكن، ثم يتم تقليمها من أطرافها (بما لا يتنافى مع الدقة المطلوبة) لتصل إلى الحجم الأصغر.

التقدير:

إذا تم التوصل للتقسيم الأمثل النهائي، فإن شجرة الانحدار تُنتج قيمة مقدرة $\hat{y}_i$ (لكل مشاهدة في المتغير التابع y_i تساوي متوسط قيم المتغير التابع في المجموعة التي تنتمي إليها المشاهدة رقم i:

$$\hat{y}_i = \frac{1}{n_m}\sum_{l=1}^{n_m} y_{lm}$$

حيث تشير m إلى رقم المجموعة التي يُحسب لها القيمة الموفقة، وتشير n_m إلى حجمها.

أما في حالة شجرة التصنيف، فتُحسب تلك القيم بمعلومية الاحتمالات المقدرة لانتساب المشاهدة لمجموعة معينة. ويكون احتمال النجاح في حالة التصنيف الثنائي:

$$\pi_i = \frac{1}{n_m}\sum_{l=1}^{n_m} y_{lm}$$

وتأخذ المشاهدة y_{lm} القيمة 0 أو 1، لذلك فإن الاحتمال المقدر يناظر نسبة النجاح في المجموعة m. ومن الجدير بالذكر أن كلا من $\hat{y}_i$، π_i يُعد ثابتاً لكل المشاهدات.

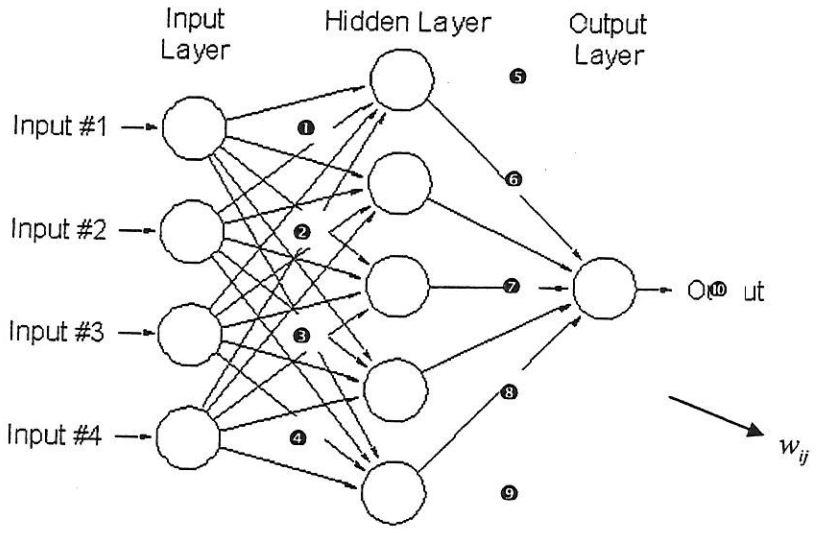
وتُعتبر نماذج الأشجار نماذج تنبؤية لامعلمية لأنها لا تتطلب فروضاً عن التوزيع الاحتمالي للمتغير التابع. وتعني هذه المرونة أن نماذج الأشجار تكون دائماً ممكنة التطبيق مهما كانت طبيعة المتغير التابع والمتغيرات المفسرة. وعلى الجانب الآخر فإن هذه الميزة يمكن أن تنقلب إلى عيوب، إذ يحتاج تطبيق نماذج الأشجار إلى موارد حاسبية عالية. كما أن طبيعتها التسلسلية وتعقّد خوارزمياتها واعتمادها على البيانات المشاهدة من شأنه أن يؤدي أي تغير طفيف في أي منها إلى تغير كبير في هيكل الشجرة.

3-8 الشبكات العصبية NN

تُستخدم الشبكات العصبية Neural Networks لتحقيق العديد من الأغراض الوصفية والتنبؤية عند التنقيب في البيانات[20]. وقد نشأت NN في مجال تعليم الآلة Machine Learning في محاولة لتقليد الوظائف العصبية للمخ البشري من خلال توليفة من العناصر الحاسبية البسيطة (الخلايا العصبية Neurons) في نظام متداخل للغاية. وتتمتع NN بأهمية خاصة[54] لأنها تقدم نمذجة عالية الكفاءة للمشاكل المعقدة (التي تحتوي على مئات المتغيرات المستقلة والعديد من التفاعلات ومتغير تابع أو أكثر) بطريقة لامعلمية من قواعد البيانات الكبيرة. كما يمكن استخدامها في حل مشاكل التصنيف ومشاكل الانحدار سواء كانت البيانات مكتملة أو مبتورة.

مكوناتها: يوضح شكل (1) أن الشبكة العصبية تتكون من مجموعة من الوحدات الحاسبية الأولية (تُعرف باسم الخلايا العصبية متصلة بما يليها من خلال روابط مرجحة. وتُمثل كل خلية بدائرة، وتأخذ رقماً طبيعياً (من 1: 10 في هذا المثال). كما تُمثل الروابط بأسهم وتأخذ الرمز w_{ij}، حيث يشير الدليل i إلى رقم العقدة التي ينطلق منها السهم ويشير الدليل j إلى رقم العقدة التي ينتهي عندها. وتُنظم

هذه الوحدات في طبقات Layers بحيث تتصل كل خلية (في طبقة ما) بجميع خلايا الطبقة السابقة واللاحقة. وتبدأ الشبكة بطبقة المدخلات Input Layer (من 1: 4 في هذا المثال)التي تناظر كل عقدة فيها أحد المتغيرات المستقلة. وتتصل كل عقدة في طبقة المدخلات بجميع عقد الطبقة الخفية (من 5: 9 في هذا المثال)، وربما تتصل عقد الطبقة الخفية بجميع عقد طبقة خفية أخرى (غير موضح على الرسم). وتنتهي الطبقات بطبقة المخرجات Output Layer (رقم 10 في هذا المثال) وهي عقدة (أو أكثر) تمثل المتغير التابع (أو المتغيرات التابعة) وهي التقاء للأسهم الخارجة من آخر طبقة خفية.



شكل (1): نموذج لشبكة عصبية بسيطة

ويُحسب الوزن w_{ij} بمجموع حواصل ضرب الأوزان الداخلة على العقدة التي ينطلق منها في قيم العقد التي تنطلق منها تلك الأوزان. وكمثال، فإن قيمة الوزن الرابط بين الطبقة 7 والطبقة 10 هو:

$$w_{7,10} = w_{17} * value\ of\ node\ 1 + w_{27} * value\ of\ node\ 2 + w_{37} * value\ of\ node\ 3 + w_{47} * value\ of\ node\ 4$$

ويمكن أن يُنظر إلى كل عقدة على أنها متغير مستقل (العقد من 1: 4)، أو على أنها توليفة (تفاعل) من المتغيرات المستقلة (العقد من 5: 10). فالعقدة 10 هي توليفة غير خطية للقيم في العقد من 1: 4 بسبب وجود دالة التنشيط (القيم المجمعة في عقد الطبقة الخفية). وجدير بالذكر أنه إذا كانت دالة التنشيط خطية ولا توجد طبقة خفية، فإن الشبكة العصبية تُختزل إلى الانحدار الخطي. بنما تُختزل الشبكة العصبية إلى الانحدار اللوجيستي في ظل دوال تنشيط غير خطية ذات شكل معين.

الإمكانية Potential:

تعبر الأوزان في الشبكة العصبية (كما في النموذج البيولوجي) عن معاملات قابلة للتعديل استجابة للإشارات التي تسافر في الشبكة بحسب خوارزمية تعلم مناسبة وقيمة فاصلة Threshold (تُعرف أيضاً باسم التحيز Bias) تشبه حد التقاطع في نموذج الانحدار. فالخلية j تأخذ القيمة الفاصلة θ_j وتستلم إشارات داخلة $x = [x_1, ..., x_n]$ من الوحدات (الخلايا/العقد) المتصلة بها من الطبقة السابقة. وتقترن كل إشارة بوزن معين $w_j = [w_{1j}, ..., w_{nj}]$.

وتتم دراسة الإشارات الداخلة وأوزانها والقيمة الفاصلة لكل خلية من خلال ما يسمى بدالة التوليف Combination Function. وتُنتِج دالة التوليف (لكل خلية) قيمة واحدة تسمى الإمكانية (أو الداخل الصافي Net Input). وتقوم دالة التنشيط Activation Function بتحويل الإمكانية إلى إشارة خارجة.

وتكون دالة التوليف عادةً خطية، لذلك فإن الإمكانية p_j تكون مجموع انحرافات قيم الخلايا السابقة x_i المرجحة بالأوزان الخارجة منها w_{ij} عن القيمة الفاصلة θ_j، وهو ما يُعبر عنه رمزياً كالتالي:

$$p_j = \sum_{i=1}^{n}\left(x_i w_{ij} - \theta_j\right) = \sum_{i=0}^{n} x_i w_{ij}$$

حيث $x_0 = 1$، $w_{0j} = -\theta_j$. ويمكن الحصول على الإشارة الخارجة للخلية j (أي y_j) بتطبيق دالة التنشيط على الإمكانية p_j لتعطي:

$$y_j = f(\mathbf{x}, \mathbf{w}_j) = f(\mathbf{p}_j) = f(\sum_{i=0}^{n} x_i w_{ij})$$

أنواع دالة التنشيط:

هناك طرق كثيرة لتنشيط الخلايا في الشبكة العصبية. ومن أشهرها: الطريقة الخطية، والطريقة المجزأة Piecewise، والطريقة الإسية Sigmoidal، وطريقة أقصى تمهيد Softmax:

1) دالة التنشيط الخطية: تُعَرّف دالة التنشيط الخطية بالصيغة التالية:

$$f(p_j) = \alpha + \beta\, p_j$$

حيث تنتمي الإمكانية p_j لمجموعة الأعداد الحقيقية، و α, β ثوابت. وعندما يتطلب النموذج أن يكون مخرج الخلية مساو تماماً لمستوى تنشيطها (الإمكانية)، نضع $\alpha = 0, \beta = 1$ وتتحول الدالة الخطية إلى ما يسمى بدالة الوحدة. يُلاحظ التشابه القوي بين دالة التنشيط الخطية ونموذج الانحدار الخطي البسيط، إذ يمكن النظر للأخير على أنه نوع بسيط من الشبكات العصبية.

2) دالة التنشيط المجزأة: تُعَرّف دالة التنشيط الخطية بالصيغة التالية:

$$f(p_j) = \begin{cases} \alpha & p_j \geq \theta_j \\ \beta & p_j < \theta_j \end{cases}$$

ويتضح أن تأخذ قيمتين فقط بحسب تجاوز الإمكانية للقيمة الفاصلة من عدمه. وعندما تكون $\alpha = 1, \beta = 0, \theta_j = 0$، نكون أمام حالة خاصة من التنشيط المجزأ تُعرف باسم دالة تنشيط الإشارة Sign Activation Function التي تأخذ القيمة 1 إذا كانت الإمكانية موجبة والقيمة 0 بخلاف ذلك.

3) دالة التنشيط الإسية: أي التي تأخذ شكل حرف s، وهي الأكثر استخداماً في التطبيقات العملية. وتُنتج هذه الدالة قيماً موجبة فقط في الفترة [0,1]. ويرجع شيوع استخدامها إلى أنها غير خطية وإلى قابليتها للفهم وللتفاضل بسهولة. وتُعَرّف بالصيغة التالية:

$$f(p_j) = \frac{1}{1 + e^{-\alpha p_j}}$$

حيث تشير α إلى معلمة موجبة تنظم ميل الدالة.

4) دالة أقصى تمهيد: تُستخدم في تطبيع Normalize مخرجات العقد المختلفة التي يوجد بينها علاقة. فإذا كانت الشبكة تحتوي على g من العقد بمخرجات عددها v_j (حيث $j = 1,2,...,g$)، فإن دالة أقصى تمهيد التي تُطبِع v_j (تجعل مجموعها 1) تكون:

$$soft\max(v_j) = \frac{e^{v_j}}{\sum_{j=1}^{g} e^{v_n}}$$

وتُستخدم هذه الدالة في حل مشاكل التصنيف المراقب Supervised Classification Problems عندما يأخذ المتغير التابع عدد g من المستويات.

طرق التدريب[3][46] Training Methods:

يُقصَد بالتدريب (بمفاهيم الشبكات العصبية) تعليم الشبكة كيف تنجز مهمة ما، وهو بلغة الإحصائيين الطريقة المستخدمة في تقدير أوزان الشبكة (المعالم المجهولة). ويمكن تدريب أو تعليم الشبكة بعدة طرق من أشهرها وأوسعها انتشاراً؛ طريقة الإكثار الخلفي Backpropagation التي تبحث في تحديث أوزان الشبكة بتصغير دالة الخطأ في فضاء الأوزان باستخدام عدة خوارزميات، من أشهرها: خوارزمية الهبوط المتدرج[12]، وخوارزمية التدرج المقارن[8][33]، وخوارزمية [40] Quasi-Newton، وخوارزمية Levenberg-Marquardt[35]، والخوارزميات الجينية [19] Genetic Algorithms.

أنواع الشبكات العصبية:

يمكن تصنيف الشبكات العصبية بحسب عدد طبقاتها في نوعين: شبكة الفواهم ذوي الطبقة الواحدة Single-Layer Perceptrons، وشبكة الفواهم متعددة الطبقات Multi-Layer Perceptrons. كما يمكن تصنيف الشبكات العصبية بحسب اتجاه تدفق المعلومات إلى: شبكات التغذية الأمامية Feedforward Networks؛ حيث تتحرك فيها المعلومات من طبقة إلى الطبقة التالية للأمام فقط دون السماح لها بالعودة للخلف، وشبكات التغذية الخلفية Feedback Networks؛ حيث تتحرك فيها المعلومات من طبقة إلى الطبقة التالية للأمام مع السماح لها بالعودة للخلف إلى الطبقات السابقة.

تطبيقات الشبكات العصبية:

تكون الشبكات العصبية قابلة للتطبيق في المشاكل السببية حين توجد علاقة معقدة بين عدة متغيرات مستقلة (مُفسِرة/ مُتنبِئة/ مدخلات) وواحد أو أكثر من المتغيرات التابعة (مُفسَر/ متنبأ به/ مخرجات)، ويصعب التعبير عن تلك العلاقة بالمداخل التقليدية كالارتباط والانحدار والاختلاف بين المجموعات. ومن أمثلة المشاكل التي طُبِقَت فيها الشبكات العصبية بنجاح[6][34][52]: الكشف عن الظواهر الطبية، و التنبؤ بسوق الأسهم، وتقييم الجدارة الائتمانية لطالبي القروض.

9-3 الانحدار اللامعلمي NR

إذا قرر الباحث مثلاً استخدام كثيرة حدود تكعيبية على الشكل[29]:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon$$

في توفيق نموذج الانحدار الذي يربط بين المتغير y والمتغير x، فإن ذلك يشترط صحة الشكل الرياضي المفترض في التعبير عن البيانات. ويُقال أن النموذج معلمي لأنه يعتمد على المعالم $\beta_1, \beta_2, \beta_3$.
أما عندما لا تتوافر معلومات كافية لصنع فرض مثل هذا، أو عند الرغبة في مجرد افتراض أن:

$$y = f(x) + \varepsilon$$

في ظل فروض التمهيد العادية [بأن $f(x), f'(x), f''(x)$ كلها مستمرة] وتقدير $f(x)$ من البيانات، فإننا نستخدم الانحدار اللامعلمي nonparametric regression.

فالانحدار اللامعلمي NR إذن[37] هو شكل من تحليل الانحدار لا يأخذ فيه المتغير المستقل شكل محدد، ولكنه يُبنى من المعلومات المشتقة من البيانات. لذلك فإن NR يتطلب حجم عينة أكبر من الحجم اللازم لحساب الانحدار المعلمي لأن البيانات هي التي تقترح هيكل النموذج وتقديرات المعالم.

ويُقدّر NR من خلال[23][26] دوال الشرائح الممهدة المكعبة cubic smoothing spline التي تأخذ الشكل التالي:

$$RSS(f,\lambda) = \sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f''(t))^2 dt$$

وهو ما يعني تقدير مجموع مربعات البواقي من جميع الدوال الممكنة $f(x)$ في ظل مشتقتين مستمرتين. وتشير λ إلى معلمة التمهيد المثبتة، ويشير الحد الأول من الطرف الأيمن قرب البيانات، ويقيس الحد الثاني مدى انحناء الدالة، وتحدد λ المفاضلة بين الحدين. فإذا كانت $\lambda = 0$، فإن f يمكن أن تكون أي دالة تستكمل البيانات. أما إذا كانت $\lambda = \infty$، يتم توفيق f بخط المربعات الصغرى المستقيم بسبب عدم الاستفادة من المشتقة الثانية.

4. الأساليب نصف المعلمية

4-1 الانحدار نصف المعلمي SPR:

يُعد الانحدار نصف المعلمي[48] Semiparametric Regression توليفة من الانحدارين المعلمي واللامعلمي. وهو يُستخدم إذا كان النموذج اللامعلمي الكامل لا يعبر بشكل جيد عن البيانات و/أو إذا أراد الباحث استخدام نموذج معلمي لكنه لا يعرف بالضبط شكله الدالي بالنسبة لمجموعة فرعية من المتغيرات المفسرة أو إذا كانت كثافة الأخطاء غير معروفة. وحيث أن SPR تحتوي على مركبة معلمية، فإنها تعتمد على فروض معلمية؛ وبالتالي فإنها تكون معرضة لمشكلتين مهمتين: خطأ التحديد misspecified (اختيار شكل رياضي خاطئ و/أو القصور في إدخال المتغيرات المعبرة عن المشكلة)، وعدم الاتساق inconsistent (عدم تمركز توزيع المقدرات بالقرب من القيمة الحقيقية للمعلمة المقدرة) كما في النماذج المعلمية الكاملة.

ويوجد العديد من الطرق لتقدير نماذج SPR، أشهرها:

1] النموذج الخطي الجزئي[42] Partially Linear Model المُعّرف بالشكل التالي:

$$Y_i = X_i'\beta + g(Z_i) + u_i, \quad i = 1, \ldots, n,$$

حيث يشير Y_i إلى المتغير التابع، وكل من X_i , Z_i إلى متجهي المتغيرات المستقلة من الدرجة $p \times 1$ ، و β إلى متجه المعالم من الدرجة $p \times 1$، و $Z_i \in R^q$. ويُعرّف متجه المعالم β الجزء المعلمي من النموذج، بينما تُعرف الدالة المجهولة $g(Z_i)$ الجزء اللامعلمي منه ويتم تقديرها بأي طريقة انحدار لامعلمية مناسبة.

2] نموذج الرقم المفرد[24] single index model والذي يُعرف أيضاً باسم Ichimura's method ويأخذ الشكل التالي:

$$Y = g(X'\beta_0) + u,$$

وتُقدر فيه المعلمه β_0 باستخدام طريقة المربعات الصغرى غير الخطية لتصغير الدالة:

$$\sum_{i=1}(Y_i - g(X_i'\beta))^2.$$

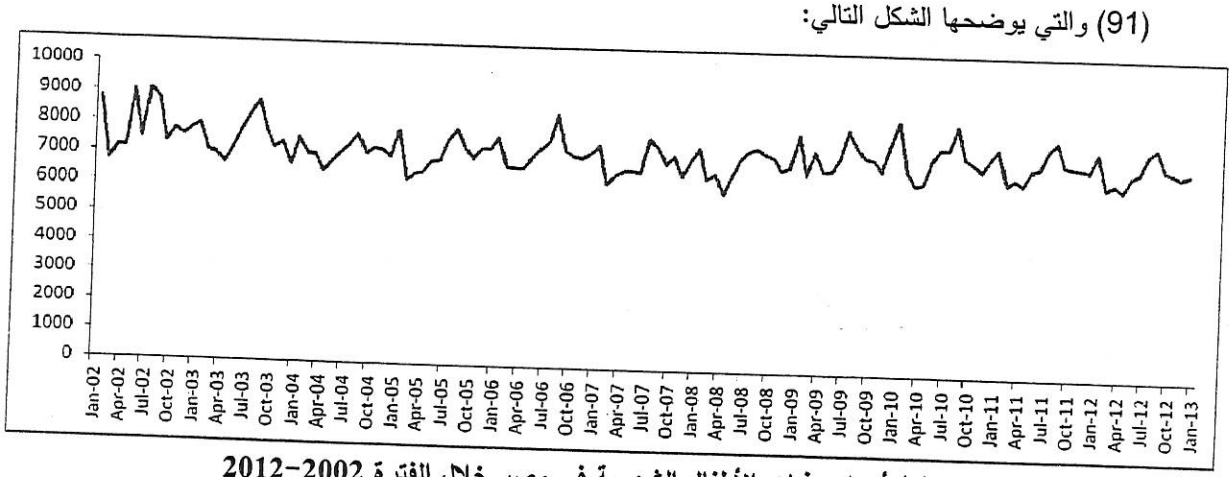
3] نماذج المعامل الممهد أو المتغير[22] Smooth coefficient\varying coefficient models التي تُعرّف بالصيغة التالية:

$$Y_i = \alpha(Z_i) + X_i'\beta(Z_i) + u_i = (1 + X_i')\begin{pmatrix}\alpha(Z_i)\\ \beta(Z_i)\end{pmatrix} + u_i = W_i'\gamma(Z_i) + u_i,$$

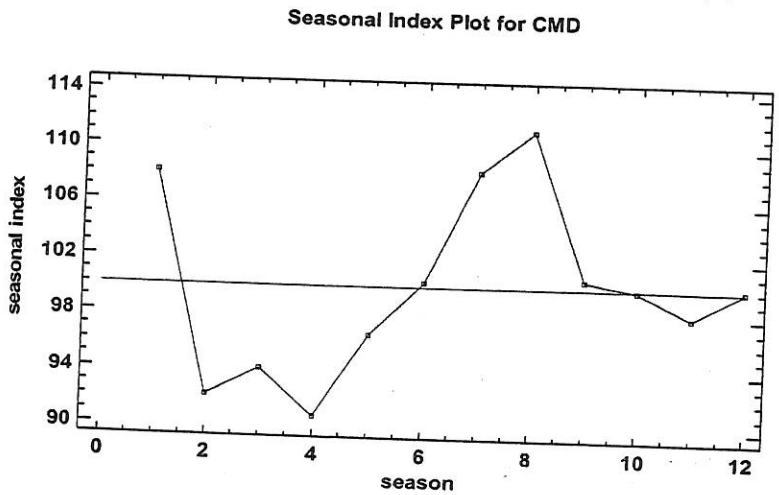
حيث يشير X_i إلى متجه من الدرجة $k \times 1$ و $\beta(z)$ إلى متجه من الدوال الممهدة غير المحددة في z.

5- تطبيق للمقارنة بين النماذج

تم استخدام السلسلة الشهرية لبيانات وفيات الأطفال فى مصر خلال الفترة 2002-2012 (الجهاز المركزي للتعبئة العامة والاحصاء، إدارة الحاسب الآلي) لتوفيق عدة نماذج حسب طبيعة البيانات المستخدمة والتى توضح وجود أثر موسمي والذي بلغ اقصاه شهر اغسطس (111) وأدناه شهر ابريل (91) والتي يوضحها الشكل التالي:



شكل (4) أعداد وفيات الأطفال الشهرية فى مصر خلال الفترة 2002-2012



شكل (5) الدليل الموسمى لوفيات الأطفال الشهرية فى مصر خلال الفترة 2002-2012

في البداية تم توفيق 18 نموذج (تقليدي واريما) موضحة بالجدول التالي، وكان أفضلها وفقاً للمعايير المحددة هو نموذج اريما الموسمي ARIMA(2,1,0)x(1,1,2)12 حيث بلغ معامل الارتباط الذاتي الأول 0.634

جدول (1) مقارنة لمؤشرات الجودة للنماذج التقليدية واريما للوفيات الشهرية للأطفال فى مصر 2002-2012

MAPE	MAE	RMSE	Model
3.71026	272.478	400.013	Random walk
3.71069	272.653	401.594	Random walk with drift = -6.87034
3.48095	258.331	378.073	Constant mean = 7286.31
3.13917	232.489	354.301	Linear trend = 9904.26 + -3.85845 t
3.11279	229.537	329.977	Quadratic trend = 114459. + -312.705 t + 0.227595 t^2
3.11633	231.086	354.103	Exponential trend = exp(9.23494 + -0.000504851 t)
3.10172	229.914	351.385	S-curve trend = exp(8.52728 + 247.206 /t)
3.15207	229.62	346.742	Simple moving average of 2 terms
2.85853	209.337	309.145	Simple exponential smoothing with alpha = 0.1997
2.91387	213.48	315.512	Brown's linear exp. smoothing with alpha = 0.1213
2.92085	214.289	315.573	Holt's linear exp. smoothing with alpha = 0.179 and beta = 0.1848
2.96674	217.416	321.196	Brown's quadratic exp. smoothing with alpha = 0.0924
3.0619	220.712	286.335	Winter's exp. smoothing alpha=0.1984, beta= 0.0308, gamma= 0.197
2.35498	169.704	229.445	ARIMA(0,1,1)x(1,1,2)12
2.32096	167.259	228.754	ARIMA(1,0,1)x(1,1,2)12
2.36162	170.168	231.049	ARIMA(2,1,1)x(1,1,2)12
2.373	170.952	231.534	(ARIMA(1,1,1)x(1,1,2)12
2.30075	165.764	228.397	ARIMA(2,1,0)x(1,1,2)12

وبعد ذلك تم استخدام أفضل نموذج من السابقة ومقارنته مع أربعة نماذج حديثة لعدد 132 شهر، ويظهر جدول (2) مقاييس الجودة للنموذج المقترح والتى تؤكد جودة ملائمة النموذج لتوفيق للبيانات باستخدام المعايير الإحصائية لقياس قدرة النموذج علي التنبؤ واستخدمت المعايير التالية (العباسي، 2011، Wooldridge, 2003):

1- جذر متوسط مربع الخطأ Root Mean Square Error (RMSE)

2- المتوسط النسبي للخطأ المطلق Mean Absolute Percentage Error (MAPE)

3- متوسط القيمة المطلقة للخطأ Mean Absolute Error (MAE)

4- معامل ثيل Theil Coefficient (T.C)

5- معامل التحديد Coefficient of Determination (R^2)

6- مؤشر الدلالة Trade sign (TS)

7- معامل الارتباط الذاتي الأول (ρ_1).

وأثبت المعايير المستخدمة للحكم على أفضلية النموذج، ولتتابع البواقى وعشوائيتها أن نموذج الشبكات العصبية يعد الأفضل والأكثر ملائمة للبيانات المستخدمة خلال الفترة 2002-2012.

جدول (2) مقارنة لمؤشرات الجودة للنماذج الخمس المستخدمة للتوفيق للوفيات الشهرية للأطفال فى مصر 2002-2012

نموذج	جذر متوسط مربع الخطأ (RMSE)	المتوسط النسبى للخطأ المطلق (MAPE)	متوسط القيمة المطلقة للخطأ (MAE)	معامل ثيل Theil	معامل التحديد R^2	مؤشر الدلالة TS	الارتباط الذاتي الأول ρ_1
الانحدار المتعدد	396.752	348.075	4.797	41.83	0.027	-2.756	-0.2493
الانحدار البواسنى	396.690	350.774	4.828	42.36	0.027	-2.855	-0.2482
السلاسل الزمنية (اريما)	301.818	236.752	3.154	79.82	0.021	3.500	-0.0487
المعممة المضافة	225.274	159.047	2.187	79.22	0.015	-5.240	0.0407
الشبكات العصبية	125.981	98.733	1.350	92.64	0.009	-0.043	-0.0293

6. الخلاصة

مما سبق يتضح:

- أن الشبكات العصبية الاصطناعية أكثر دقة وكفاءة في التنبؤ عن الأساليب الإحصائية التقليدية حيث وصلت الشبكات لمعدل مرتفع وعالي من الدقة مع بقاء أفضليتها في التنبؤ للسلاسل الزمنية الطويلة والتى لايوجد بها اثر واضح للموسمية او الارتباط الذاتى.
- إن استخدام نموذج الشبكات العصبية في التنبؤ، ورسم الخطط سواء الطويلة الأجل والقصيرة الأجل لما يتميز به هذا النموذج من سرعة ودقة في البيانات أكثر منه في الأساليب الإحصائية التقليدية.
- من خلال التطبيق لكل من النماذج الإحصائية التقليدية والشبكات العصبية الاصطناعية ANN يتبين لنا أن الشبكات العصبية قد تميزت عن الأساليب الإحصائية التقليدية بأن لديها منهجية في عدم الاعتماد على الخطية في البيانات.
- يجب على كل من يقوم بدراسة يتطلب فيها نظرة مستقبلية أن يقوم باستخدام الشبكات العصبية وأن يتم تحليلها باستخدام الأساليب الإحصائية الحديثة، وذلك لتحقيق الاستفادة القصوى منها حيث أن الشبكات لديها السرعة والدقة.
- وجد أن الشبكات العصبية تتفوق على النماذج التقليدية بدرجة ملحوظة، وبمعنى آخر ونظراً لمنهجية الشبكات العصبية في اعتمادها على غير الخطية فإن أداءها أفضل مقارنة بالنماذج التقليدية ، وينتج أيضاً أنه يمكن تطبيق الشبكات العصبية بنجاح في التنبؤ بالسلاسل الزمنية الشهرية الطويلة والتى تتسم بالموسمية أو الارتباط الذاتى.

فى النهاية نكون قد حققنا هدف البحث وهو التعريف بأساليب التنقيب في البيانات وأنواعها، كما يهدف إلى استعراض المجالات المختلفة التي استُخدمت فيها تلك الأساليب بنجاح. وقد اعتمد البحث على دراسة العديد من البحوث التي تتضمن أساليب مختلفة للتنقيب في البيانات للتعريف بها وتطبيقاته وشروط استخدامها بشكل مبسط، ومقارنة بين أفضلية أساليب التنقيب الحديثة والأساليب التقليدية، وأتضح أن أسلوب الشبكات العصبية يعد أفضل النماذج المستخدمة مقارنة بالنماذج التقليدية لبناء نموذج للوفيات الشهرية للأطفال فى مصر خلال الفترة 2002-2012.

المراجـ ـع

1. العباسى، عبدالحميد محمد (2013)، التنقيب في البيانات Data Mining تطبيقات باستخدام SPSS MODELR ، معهد الدراسات والبحوث الإحصائية– القاهرة.
2. العباسى، عبدالحميد محمد (2012)، قوة العمل الحكومية الكويتية: الواقع والعوامل المؤثرة خلال الفترة 1993-2011، المجلة الإحصائية المصرية، معهد الدراسات والبحوث الإحصائية – القاهرة – مصر، مجلد (56) العدد (2)، ديسمبر 2012 ص (30 – 46).
3. العباسى، عبدالحميد محمد (2010)، التحليل الحديث للسلاسل الزمنية باستخدام Eviwes، معهد الدراسات والبحوث الإحصائية– القاهرة.
4. العباسى، عبدالحميد محمد (2004)،" المقارنة بين استخدام الشبكات العصبية وساريما للتنبؤ بأعداد الوفيات الشهرية الناتجة عن حوادث المرور بالكويت "، المجلة العربية للعلوم الإدارية ، الكويت ، مجلد (3) العدد (11)، ص (333 – 359).

[1] Analytic Hierarchy Process
http://en.wikipedia.org/wiki/Analytic_Hierarchy_Process
[2] Analytic Hierarchy Process (AHP) Tutorial
http://www.cs.toronto.edu/~sme/CSC340F/slides/tutorial-prioritization.pdf
[3] Backpropagation
http://en.wikipedia.org/wiki/Backpropagation
[4] C4.5 algorithm
http://en.wikipedia.org/wiki/C4.5_algorithm
[5] CHAID
http://en.wikipedia.org/wiki/CHAID
[6] Christos Stergiou and Dimitrios Siganos. Neural Network
http://www.doc.ic.ac.uk/~nd/surprise_96/journal/vol4/cs11/report.html#Neural Networks in Practice
[7] CIPFA (2008). Nearest Neighbours Model: Methodology Note and Instructions
http://www.cipfastats.net/default_view.asp?content_ref=2748
[8] Conjugate gradient method
http://en.wikipedia.org/wiki/Conjugate_gradient_method
[9] Conventional programming
http://www.pcmag.com/encyclopedia_term/0,2542,t=conventional+programming&i=40325,00.asp
[10] Cornelius T. Leondes (2002. Expert systems: the technology of knowledge mana-gement and decision making for the 21st century, *Academic Press, pp. 1-22.*
[11] CRISP-DM (2003), CRoss Industry Standard Process for Data Mining
http://www.crisp-dm.org.
[12] Delta rule (gradient descent)
http://en.wikipedia.org/wiki/Delta_rule
[13] Fayyad; U.M., Piatetsky-Shapiro; G., Smyth; P. and Uthurusamy; R (eds) (1996a), Advances in Knowledge Discovery and Data Mining, *AAAI Press*.
[14] Fayyad; U.M., Piatetsky-Shapiro; G., and Smyth; P. (1996b), From Data Mining to Knowledge Discovery: An Overview. In Fayyad; U.M., Piatetsky-Shapiro; G., Smyth; P. and Uthurusamy; R (eds), Advances in Knowledge Discovery and Data Mining , *AI, DDM, AAAI/MIT Press*, pp. 1-34.
[15] Fayyad; U.M., Piatetsky-Shapiro; G., and Smyth; P. (1996c), The KDD process for extracting useful knowledge from volumes of data, Communications of the ACM, 39 (11), pp. 27-34.

[16] Fayyad; U.M., Piatetsky-Shapiro; G., and Smyth; P. (1996d), Knowledge Discovery and Data Mining: Towards a unifying framework, *AI, DDM, AAAI/MIT Press*, pp. 82-88.
[17] Fayyad; U.M., Piatetsky-Shapiro; G., and Smyth; P. (1996e), From data mining to knowledge discovery in databases, *AI Magazine*,17, (3), pp. 37-54.
[18] Friedman, J. H. (1991). "Multivariate Adaptive Regression Splines" *Annals of Statistics*, 19 (1): 1–67. doi:10.1214/aos/1176347963. MR1091842. Zbl 0765.62064.
[19] Genetic algorithm http://en.wikipedia.org/wiki/Genetic_algorithm
[20] Generalized additive model http://en.wikipedia.org/wiki/Generalized_additive_model
[21] Giudici; P. (2003), Applied Data Mining: Statistical Methods for Business and Industry, *John Wiley & Sons Ltd.*
[22] Hastie; T., Tibshirani; R. (1993). "Varying-Coefficient Models" *Journal of the Royal Statistical Society , Series B, 55, pp. 757–796.*
[23] Hastie; T., Tibshirani; R., Friedman; J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, *Springer Series in Statistics.*
[24] Ichimura, H. (1993). "Semiparametric Least Squares (SLS) and Weighted SLS Esti-mation of Single Index Models" *Journal of Econometrics*, 58, *pp. 71–120.* doi:10.1016/0304-4076(93)90114-K.
[25] 1. Jackson, Peter (1998). Introduction to Expert Systems, 3rd ed., *Addison Wesley, p. 2*, ISBN 978-0-201-87686-4.
[26] John Fox (2002). Nonparametric regression http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-nonparametric-regression.pdf
[27] Koch; C. G., Isle; B. A., Butler; A. W. (1988). "Intelligent user interface for expert systems applied to power plant maintenance and troubleshooting" *IEEE Transac-tions on Energy Conversion, PP. 3- 71.*
[28] Mark E. Irwin (2005). Generalized Additive Models, *Harvard University* http://www.markirwin.net/stat135/Lecture/Lecture34.pdf
[29] Mark E. Irwin (2005). Non Parametric Regression, *Harvard University* http://www.markirwin.net/stat135/Lecture/Lecture33.pdf
[30] Mathematical optimization http://en.wikipedia.org/wiki/Mathematical_optimization
[31] Mathematical Programming http://www.me.utexas.edu/~jensen/ORMM/frontpage/pdf/mathprog.pdf
[32] Multivariate adaptive regression splines http://en.wikipedia.org/wiki/Multivariate_adaptive_regression_splines
[33] Multilayer Perceptron Neural Networks http://www.dtreg.com/mlfn.htm
[34] Neural Network Software *For researchers, data mining experts and predictive analysts* http://www.alyuda.com/products/neurointelligence/neural-network-applications.htm
[35] Neural Network Toolbox, Levenberg-Marquardt (trainlm) http://www.caspur.it/risorse/softappl/doc/matlab_help/toolbox/nnet/backpr11.html
[36] Newton's Telecom Dictionary (2010), Harry Newton, CMP Books, http://www.cmpbooks.com.

[37] Nonparametric regression
http://en.wikipedia.org/wiki/Nonparametric_regression

[38] Nwigbo Stella and Agbo Okechuku Chuks (2011). "Expert system: a catalyst in educational development in Nigeria," *Proceedings of the 1stInternational Techn-ology, Education and Environment Conference, (c) African Society for Scientific Research (ASSR)*
http://www.hrmars.com/admin/pics/261.pdf

[39] P.A.Bassett and G. Bishop. Generalized Additive Models
http://www.bioss.ac.uk/smart/unix/mgam/slides/frames.htm

[40] Quasi-Newton method
http://en.wikipedia.org/wiki/Quasi-Newton_method

[41] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning, *Morgan Kaufmann Publishers*.

[42] Racine, J.S.; Qui, L. (2007). "A Partially Linear Kernel Estimator for Categorical Data" *Unpublished Manuscript, Mcmaster University.*

[43] Rajender Parsad and Cini Verghese. NEAREST NEIGHBOURHOOD DESIGNS
http://www.iasri.res.in/iasriwebsite/DESIGNOFEXPAPPLICATION/Electronic-Book/module5/1

[44] Reil, T. (2005), Artificial Neural Network
http://www.google.com.sa/url

[45] Regina Barzilay, Daryl McCullough, Owen Rambow, Jonathan DeCristofaro, Tanya Korelsky, Benoit Lavoie. "A new approach to expert system explanations"
http://www.cogentex.com/papers/explanation-iwnlg98.pdf

[46] Rojas; R. (1996). Neural Networks, *The backpropagation algorithm, Springer-Verlag, Berlin.*
http://page.mi.fu-berlin.de/rojas/neural/chapter/K7.pdf

[47] Rygielski; C., Wang; J. and Yen; C. (2002), Data mining techniques for customer relationship management, *Technology in Society*, 24, pp. 483–502.

[48] Semiparametric regression
http://en.wikipedia.org/wiki/Semiparametric_regression

[49] SPSS (2009), Clementine 16.0, *SPSS*, Inc.
http://www.spss.com/spssbi/clementine/

[50] StatSoft, *STATISTICA*. Data Mining Techniques, *Statsoft Electronic Statistics Textbook*
www.statsoft.com/textbook/data-mining-techniques/.

[51] StatSoft, *STATISTICA*. Generalized Additive Models (GAM)
http://www.statsoft.com/textbook/generalized-additive-models/

[52] StatSoft, *STATISTICA*. Automated Neural Network *(SANN)*
http://www.statsoft.com/textbook/neural-networks/

[53] Thomas L. Saaty (2008). " Decision making with the analytic hierarchy process" *Int. J. Services Sciences, Vol. 1, No. 1, pp. 83-98.*
http://www.colorado.edu/geography/leyk/geog_5113/readings/saaty_2008.pdf.

[54] Two Crows. Introduction to Data Mining and Knowledge Discovery, 3rd ed.
http://www.twocrows.com/intro-dm.pdf

[55] Wood, S. N. (2006). Generalized Additive Models: An Introduction with R. *Chapman & Hall/CRC.*

[56] Yu-Shan Shih (2005). QUEST Classification Tree (version 1.9.2)
http://www.stat.wisc.edu/~loh/quest.html

[57] Wooldridge J. M. (2003) Econometric Analysis of Cross Section and Panel Data, MIT Press.