الأسلوب البيزي في صنف الانحدار الشجري لتقدير نموذج تجميعي ومقارنته بالنموذج اللوجستي مع التطبيق

أ.م.د. عمر عبد المحسن علي* م.م. سهاد أحمد أحمد**

المستخلص

تم استعمال الاسلوب البيزي وما يتضمنه من ميزات استدلالية لتحليل أنموذج تصنيف الانحدار الشجري للأستفادة من المعلومات السابقة من جانب، ولتجميع الأشجار للمتغيرات التوضيحية كلها معاً وعند كل مرحلة تفرّع من جانب آخر. فضلاً عن استحصال المعلومات اللاحقة عند كل عقدة تفرّع في بناء هذه الشجرة التصنيفية. وعلى الرغم من دقة التقديرات البيزية عموماً، لكن يبدو أن النموذج اللوجستي لازال منافساً جيداً في مجال وصف الاستجابات الثنائية من خلال مرونته وتمثيله الرياضي . ولذا تم استعمال ثلاث طرائق تتم فيها معالجة بيانات البحث، وهي: النموذج اللوجستي، ونموذج مصنف الانحدار الشجري، ونموذج مصنف الانحدار الشجري البيزي. وقد تم في هذا البحث مقارنة هذه الطرائق بصيغة نموذج تجميعي لدالة لامعلمية. وتم اجراء عملية المفاضلة بين هذه النماذج بالأستناد الى معيار دقة التصنيف المتمثل بخطأ التصنيف، ومعيار دقة التقدير المتمثل بجذر متوسط مربعات الخطأ. وتم التطبيق على بيانات مرضى السكري لمن يبلغون من العمر (15) سنة فأقل مأخوذة من عينة بحجم (200) تم سحبها من مستشفى الطفل في الاسكان / بغداد .

الكلمات الرئيسية: التصنيف الشجري، الأسلوب البيزي، النموذج اللوجستي، الطرائق اللامعلمية

Abstract

The use of Bayesian approach has the promise of features indicative of regression analysis model classification tree to take advantage of the above information by, and ensemble trees for explanatory variables are all together and at every stage on the other. In addition to obtaining the subsequent information at each node in the construction of these classification tree. Although bayesian estimates is generally accurate, but it seems that the logistic model is still a good competitor in the field of binary responses through its flexibility and mathematical representation. So is the use of three research methods data processing is carried out, namely: logistic model, and model classification regression tree, and bayesian regression tree model. Having been in this research compare these methods form a model for additive function to some nonparametric function. It was a trade-off between these process models based on the classification accuracy by misclassification error, and estimation accuracy by

* جامعة بغداد / كلية الادارة والاقتصاد .

** جامعة بغداد / كلية التربية ابن الهيثم للعلوم الصرفة .

مقبول للنشر بتأريخ 2015/12/14

مستل من أطروحة دكتوراه

the root of the mean squares error: RMSE. It was the application on patients with diabetes data for those aged 15 years and below are taken from the sample size (200) was withdrawn from the Children Hospital in Al-Eskan / Baghdad.

1- المقدمة وهدف البحث

1-1 المقدمة Introduction

تعدّ مسألة وجود اختلافات جوهرية Essential Variations بين الوحدات التجريبية قيد البحث لظاهرة ما إحدى المسائل المهمة التي يواجهها ويحاول معالجتها الباحثون والعلماء. فيبرز تساؤل في اذهان الباحثين عن معنوية أو عدم معنوية تأثير معين للأسمدة على حقول مختلفة تم زراعتها من المحصول نفسه. أو مدى اختلاف استجابة صفين من الطلبة الجامعيين في احدى الكليات على الرغم من استعمال طرائق التدريس نفسها. فيتم اللجوء الى استعمال صيغ اكثر مرونة واكثر اعماماً تستطيع التخفيف من حدة الاختلافات تلك والتكيّف لشمول جميع الوحدات التجريبية بصورة سهلة وسلسة تضمن مقدار عال من الدقة وجعلها على خط شروع واحد يقل فيها الخطأ التجريبي العائد الى الصدفة.

كما وتعدّ مسألة الكشف عن الاختلافات Variations Detection للأستجابة لظاهرة ما من الأهمية بمكان بحيث تجعل التحليل الأحصائي اكثر سهولة. وهذا الأمر يمكن قياسه ضمن إسلوبين قد متداخلين Nested من أطر علم الأحصاء، الأول: ويدعى بالتصنيف Classification، وهدفه اجراء عملية اختيار لوحدات تجريبية ضمن العينة قيد البحث ضمن أطار مجموعة متشابهة بصفة Characteristic معينة لغرض التقليل من الأختلافات الجوهرية التي قد تسبب في تضخيم الخطأ التجريبي وتزيد من أضطراب تفسير الظاهرة قيد البحث. أما الأسلوب الثاني: فيدعى بالتقدير Estimation، وهدفه ايجاد مقدار Magnitude مايمثل معدل الظاهرة المتنبأ بها لتكون ظاهرة للعيان وتمثل معدل قيم استجابات الوحدات التجريبية بعد ان تصبح أكثر أنسجاماً مع بعضها.

وتبرز جدوى الاسلوبين المذكورين في أعلاه، في موضوع تحليل التباين Analysis of Variance: ANOVA على سبيل المثال، والذي يبين مصادر الاختلافات Source of variations: S.O.V. على أساس إحتساب مجموع مربعات لبين المجموعات Sum of Squares between groups ومجموع مربعات لداخل المجموعات Sum of squares within groups والذي يمثل بدوره مجموع مربعات الخطأ Sum of Squares of Error: SSE والذي يهدف الباحث لجعله أقل مايمكن. ومن آليات تقليل SSE هو إجراء عملية تجميع Grouping الوحدات التجريبية على اساس صفة أو ضمها لقطاعات Blocking معينة للتقليل من أختلافاتها سعياً من الباحث للحصول على تقديرات يتوخى فيها الدقة أكثر من تلك التي يتم التعامل معها بدون تصنيف أو بدون تجميع على أساس صفة/ (صفات).

1-2 مشكلة البحث Research Problem

تتضمن طرائق التصنيف classification أو التمييز discrimination التقليدية المعروفة كالتحليل التمييزي Discriminant Analysis والتحليل العنقودي Cluster Analysis وغيرها، تتضمن بعض القصور والمحدودية في وضعها بعض القيود على طريقة التحليل سواء من حيث افتراضات حول المسافات البينية لاستجابات المتغير التابع أو حول نوع تلك المسافة أو حول افتراضات الصفة التي من أجل تجميع Grouping المشاهدات لمتغير الاستجابة ضمنها. وهو مايقف عائقاً أمام اسلوب مرن يأخذ بنظر الاعتبار التغيرات المضطردة للعالم الجديد اليوم ونحن على اعتاب الالفية الثالثة حيث الاضراب الحاصل في رتابة monotony الظواهر المتعلقة بالنشاط البشري عموماً.

1-3 هدف البحث Research Goal

يهدف البحث الى استعمال اسلوب تصنيف حديث ومرن يستند الى جوارات neighborhood بين مشاهدات متغير الاستجابة متمثلة بالاساليب الاتية:

التحليل اللوجيستي Logistic Analysis
اشجار الانحدار التصنيفية Classification Regression Trees: CART
اشجار الانحدار التصنيفية البيزية Bayesian Classification Regression Trees: BART

ومن ثم اجراء تقدير النموذج التجميعي العام Generalized Additive Model: GAM أزاء كل طريقة تصنيف من الطرق المذكورة في أعلاه. بإعتباره اسلوب لامعلمي حديث ومرن يتغلب على مشكلة البعدية Curse of Dimensionality التي قد تعتري معظم ظواهر العصر الحديث عند تحليل بياناتها.

2- الجانب النظري

1-2 مقدمة

تعدّ طرائق الاستنتاج الشجري tree induction وطرائق الانحدار من التقنيات ذات الطبيعة التكاملية، أي ان احدهما يكمّل الآخر. حيث يظهر في تحيّز عالي وتباين واطئ عند استعمال اسلوب الانحدار، بينما يظهر تحيّز أوطأ ولكن بتباين أعلى عند استعمال اسلوب الاستنتاج الشجري. ولذا يكون الدمج (أو الجمع) بين الاسلوبين وانجازهما ضمن اسلوب موحّد لهما سيكون ذو ميزة استدلالية عالية.

ومن طرائق الاستنتاج الشجري هي النماذج ذات الاساس الشجري tree-based models بسهولتها وكفاءتها عند التعامل مع مجالات domains عدد كبير من المتغيرات والوحدات التجريبية. ويتم الحصول على اشجار الانحدار regression trees باستعمال التقسيم السريع لتلك المجالات باسلوب خوارزمية معينة باستعمال مبدأ التقسيم المتتابع recursive partitioning لجعل مجموعة البيانات المدخلة مقسمة الى مجموعات فرعية اصغر من البيانات. وان استعمال هذه الخوارزمية بهذا الوصف هو سبب كفاءها. وعلى اية حال، لايمكن التعويل على نتائج هذه الخوارزمية بسبب انها ستفرز قرارات ضعيفة في حالة العينات الصغيرة من الوحدات التجريبية. لأن الصفة المهمة في هكذا خوارزميات هو كُبر ونمو الشجرة الابتدائية عبر مراحل التقسيم حتى نهاية التحليل.

وفي هذا السياق، فان النماذج ذات الاساس الشجري تستند على استراتيجيات تُعرف بطرائق التقليم "pruning". ذلك لتلافي الوقوع في حالة فوق التقدير overfitting لهذه النماذج. كما هو الحال في حالة فوق التمهيد oversmoothing التي تصادفنا عند العمل على موضوع موائمة المنحنيات curves fitting في تقدير نموذج الانحدار اللامعلمي. حيث ان الشجرة المكوّنة للنموذج ذو الاساس الشجري ستبدأ بالنمو والاتسّاع الى درجة كبيرة يصعب معها التحليل الدقيق، ولذا تكون عملية التقليم ناجعة في التخلص من الاغصان "branches" التي لايعول عليها وهو مايدعى بالتقليم اللاحق "post pruning". ولذا يكون اختيار اسلوب التقليم الافضل بمثابة اختيار للشجرة المقلّمة الافضل.

وكان اول من استعمل هذه الاستراتيجية هو (Breiman et al., 1984) مع اسلوب اسماه أسلوب CART كحالة خاصة من النماذج ذات الاساس الشجري ضمن مايدعى باشجار القرارات (Decision-Trees) وهي عبارة عن مخططات التي تحمل عرض مجموعة من النتائج المحتملة والقرارات للاحقة التي هي بعد القرار الأولي. حيث يقوم هذا الاسلوب على مرحلتين منفصلتين، حيث نبدأ أبتداءاً بتوليد سلسلة متتابعة من الاشجار المقلّمة ، ومن ثم يصار الى عملية أختيار الشجرة ليتم تطبيقها للحصول على النموذج النهائي.

2-2 التصنيف [10] Classification

يعدّ التصنيف من الآليات الناجعة فيما يخص مسألة تجميع الوحدات التجريبية لعينة البحث. اضافة الى ذلك، فان التصنيف يعدّ من تقنيات التنقيب عن البيانات Data Mining. ويذكر أن هذا الاسلوب ليس الوحيد في هذا المجال فعلى شاكلته يوجد الكثير ولو بدقة متفاوتة مثل استعمال: التحليل التمييزي Discriminant Analysis، التحليل العنقودي Cluster Analysis، وخوارزمية اشجار الانحدار التصنيفية Classification Regression Trees: CART ، وخوارزمية اشجار الانحدار التصنيفية البيزية Bayesian Classification Regression Trees: BART ، وتصنيف الغابات Forest Classification ،...، الخ. وللوصول الى هذا التصنيف سيتم استعمال الاساليب الاحصائية كما في أدناه.

3-2 النموذج التجميعي المعمم [4],[11] Generalized Additive Model

أن دراسة الانحدار المتعدد بشكل عام والانحدار ثنائي المتغيرات بشكل خاص تفرز عنها مشكلة البعدية ، والتي يعاني منها معظم الباحثين، حيث تقيدهم نحو تعميم حالة احادية المتغيرات الى حالة متعدد المتغيرات. ولكن هنا ستبرز مشكلة اخرى في الانحدار المتعدد (P>1) وهي كيفية احتساب حد الجزء، والذي سياخذ عندها جميع فضاءات (مجالات) المتغيرات التوضيحية بنظر الاعتبار مع تفاعلاتها Interactions.

ولذا كانت الاوجه لنماذج من نوع خاص تجاوز هذه المشاكل . فظهرت GAM كحل عملي يقوم على اساس صفة تجميعية Additivity وهي صفة مرغوبة بها في معظم الاستدلالية لانها تساعد في تسهيل وتفسير الظواهر المختلفة.

اذ يتم اخذ ممهدات موضعية Local احادية الابعاد (المتغيرات) بشكل تجميعي ليكون ممهد شامل Global ويعد GAM الحالة اللامعلمية المطورة لـ GLIM ويتم الحصول عليها وذلك من خلال استبدال Linear Predictor:

$$Y_i = \sum_{j=1}^{P} X_j B_j$$

بحد آخر لامعلمي وهو Additive Predictor

$$Y_i = \sum_{j=1}^{P} f_j X_j$$

حيث ان:

$$E(\epsilon) = \sigma \; ; \; var(\epsilon) = \sigma_\epsilon^2$$

ϵ_i : اخطاء مستقلة عن X_{ij} 'S بحيث .

F_j : متجهة دالة مجهولة تعبر عن المتغير التوضيحي X_j.

يقوم GAM في الحقيقة بدمج فكرة التقريب التجميعي Additive Approximation مع فكرة التمهيد smoothing.

اذا النموذج التجميعي سيكون بالشكل التالي :

$$y = \alpha + \sum_{j=1}^{P} f_j(x_j) + \epsilon$$

1-3-2 خوارزمية النموذج التجميعي Algorithm for Additive model

1- **Initialize :**

$$\hat{\alpha} = \frac{1}{N} \sum_1^N y_i \;\; , \;\; \hat{f_j} = 0 \; ,\forall i, j$$

2- **Cycle :**

J=1,2,3,….. , P , ……,1,2,……,P,……

$$\hat{f_j} \leftarrow S_j \left[\{ y_i - \hat{\alpha} - \sum_{k \neq 1} \hat{f_k}(X_{ik}) \}_1^N \right],$$

$$\hat{f_j} \leftarrow \hat{f_j} - \frac{1}{N} \sum_{i=1}^N \hat{f_j}(X_{ij}).$$

Until the function $\hat{f_j}$ change less than a prespecified threshold

وتمتاز الطرائق اللامعلمية أن تسهم الى حد ما في الحصول على قدرة عملية لتلافي مشكلة الابعاد (Curse Of Dimensionality) الموجودة في مثل هذا النوع من البيانات، وبناء على ذلك تعد مسألة تطوير الطرائق اللامعلمية من الامور المطلوبة بشكل كبير، لذا فأن الظروف والاسباب المذكورة أعلاه اجبرت الباحثين على الاتجاه الى اساليب حديثة تتمثل بالطرائق الاحصائية اللامعلمية والشبه معلمية لتحليل البيانات، والتي تزودنا باستدلالات صحيحه في حالة عدم تحقق الشروط أو أن يكون هنالك تركيب خطي للبيانات.

4-2 الانحدار اللوجستي [1],[3],[5],[12] Logistic Regression

يتم بناء نموذج الانحدار اللوجستي على فرض اساسي هو أن المتغير التابع (Y) متغير الاستجابة هو متغير ثنائي التوزيع يتبع توزيع بيرنولي Bernoulli يأخذ القيمة (1) أي تعني بحدوث الاستجابة وبأحتمال (p)، والقيمة (0) أي عدم حدوث الاستجابة وبأحتمال q=(1-p) . وكما هو معلوم فان الانحدار الخطي الذي تكون متغيراته المستقلة والمتغير التابع قيماً مستمرة، وإن الانحدار اللوجستي في حالة البيانات المصنفة لأكثر من مستويين والتي يكون فيها المتغير التابع متعدد الاستجابة (Multiresponse) فإن البيانات تُصنف هنا على شكل فئات، والنموذج الذي يربط تلك المتغيرات هو كما يأتي:

$$Y = b_0 + b_1 x + e$$

إذ إن (Y) هو متغيراً مشاهداً ومستمراً وبفرض أن متوسط قيم (Y) المشاهدة أو الفعلية عند قيمة معينة للمتغير X هي E (Y) وان المتغير e يمثل الخطأ:

$$e = Y - \hat{Y}$$

لذلك يمكن كتابة النموذج على الصيغة الآتية:

$$E(Y/X) = \hat{b}_0 + \hat{b}_1 x$$

ومن المعروف كذلك فان الانحدار في طرفه الايمن لهذه النماذج يأخذ قيماً من (∞-) الى (∞+) ولكن عندما يكون لدينا متغيران احدهما ثنائي (Y) فأن الانحدار الخطي البسيط لا يكون ملائما لأن:

$$E(Y/X) = P(Y-1) = P'$$

لذلك تكون قيمة الطرف الايمن محصورة مابين الرقمين (0 , 1) وبذلك يكون النموذج غير قابل للتطبيق من وجهة نظر تحليل الانحدار، فأن احد الحلول لمثل هذه المشكلة هو ادخال تحويل رياضي مناسب على المتغير التابع (Y)، ومن المعروف كذلك أن:

$$0 \leq P \leq 1$$

لذلك فان نسبة $(\frac{P}{1-P})$ او $(\frac{P}{q})$ هي عبارة مقدار موجب محصور بين (0 , ∞) اي ان $0 \leq \frac{P}{q} \leq \infty$

وياخذ اللوغاريتم الطبيعي لـ $\frac{P}{q}$ ، لذلك فأن مجال قيمة تصبح محصورة $(-\infty \leq \log\left(\frac{P}{q}\right) \leq \infty)$.

وعليه يمكن كتابة نموذج الانحدار في حالة متغير مستقل واحد:

$$\log\left(\frac{P}{q}\right) = b0^{\wedge} + b^{\wedge}X$$

واذا كان لدينا أكثر من متغير مستقل فأن النموذج يصبح:

$$\log\left(\frac{P}{q}\right) = b0 + \sum_{i=1}^{k} b^{\wedge}j\ Xij \quad ,$$

$$i=1,2,3,\ldots\ldots,K \quad , \quad j=1,2,3,\ldots\ldots,N$$

يمكن تحويل المعادلة السابقة الى الصيغة الآتية:

$$P = 1/(1 + \exp[-(\beta_o + \sum b_j X_{ij})])$$

حيث ان : exp هو معكوس اللوغاريتم الطبيعي.

ويسمى هذا النموذج بنموذج الانحدار اللوجيستي وتسمى التحويلة $\log(\frac{P}{q})$ او $\ln(\frac{P}{q})$ بتحويل Logit. وباختصار فأن نموذج الانحدار اللوجيستي هو ببساطة تحويل لوغاريتمي للانحدار الخطي البسيط الذي تم شرحة اعلاه، وبنفس المفهوم للانحدار الخطي المتعدد فأن الانحدار اللوجيستي يأخذ المتغير المعتمد Y متغير ثنائي او اكثر من ثنائي وفي هذه الحالة عندما يكون المتغير Y متغير اكثر من ثنائي ومستمر نقوم بأخذ فئات تكرارية لمتغير الاستجابة (Y) .

لذلك فأن نموذج الانحدار اللوجسيتي للانحدار الخطي المتعدد يأخذ الصيغة الآتية:

$$P = \frac{\exp(B'X)}{1 + \exp(B'X)}$$

حيث X: تمثل المتجه $X = (X1, X2, \ldots\ldots\ldots, Xp)$

B' : تمثل المتجه $B' = (B0, B1, B2, \ldots\ldots\ldots, Bp)$

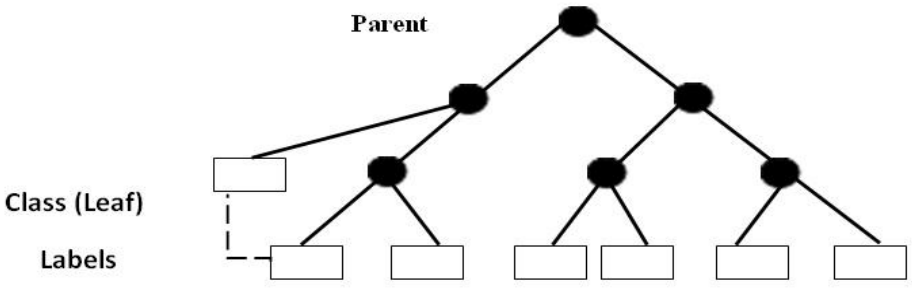
ويتم تقدير معالم النموذج اللوجيستي بطريقة الامكان الاعظم(Maximum Likelihood) ، وهي من اشهر الطرق التقدير في الاحصاء وتقييس دالة الامكان الاعظم (M.L) الاحتمالات المشاهدة لعدد N من المتغيرات المستقلة ولتكن ($P_1, P_2, \ldots P_n$) التي يقع في العينة ويمثل حاصل ضرب هذه الاحتمالات دالة الامكان الاعظم.

$$M.L. = Prob.(p_1, p_2, p_3, \ldots, p_n)$$

5-2 اشجار الانحدار التصنيفي Classification Regression Trees: CART [3],[10]

هو عبارة عن مخطط شجري يقوم بتقسيم مجموعة البيانات Datasets الى مجموعتين او اكثر من المجاميع الفرعية لتحسين تصنيف المتغير الهدف شكل (1) .

والفكرة الاساس هي ان تقسم مجموعة البيانات التنبؤية Predictor الى مجاميع جزئية Subsets كل منها أكثر تجانساً من المجموعة الأصلية .



الشكل (1)
يمثل شكل التصنيف الشجري

وهكذا فأن التصنيف الشجري يعتبر قاعدة تجريبية لتصنيف المتغير التابع من القيم المتغيرات التنبؤية Predictors .

وفيما يأتي مراحل تكوين الاشجار التصنيفية:

1. دمج Merging – تجميع اصناف المتغير التنبؤي predictor غير المعنوية بأصناف المتغير التنبؤي Predictor المعنوية نسبة الى متغير الهدف .
2. انقسام Splitting – تقسيم العقدة بأختيار المتغير التنبؤي .
3. ايقاف Stopping – وضع (حكم) Rules .
4. تقليم Pruning – إزالة وحذف الفروع التي لا تضيف الى عقدة التنبؤية للنموذج .

وتعد طريقة (CART) هي حالة خاصة من اشجار القرارات (Decision Trees). ومن مميزاتها تمتاز ببساطتها ومرونتها وسهلة الفهم فاذا كانت لدينا K من الفئات ($C_1,C_2,...C_K$)، وعينة من البيانات Training Data، يمكن ملاحظة ما يأتي:

1) اذا كانت T تحتوي على واحدة أو اكثر من المشاهدات المنتمية الى فئة واحدة C_j ستكون للشجرة ورقة Leaf مخصصة للفئة C_j .
2) اذا لم تحتوي T على مشاهدات تلك الفئات اذاً لا توجد هناك شجرة لهذه البيانات.
3) اذا احتوت T على خليط من المشاهدات تلك الفئات سيكون هناك اختبار Testing Data مبنياً على الصفات المفردة لتك المشاهدات التي ممكن ان تعطي واحدة او اكثر من النتائج المنفصلة مثنى مثنى ($O_1,O_2,....O_n$) والمجموعة T ستقسم الى مجاميع فرعية ($T_1,T_2,....T_n$) حيث ان T_i تحتوي على جميع المشاهدات التي لها النتائج O_i من الاختبار الذي تم اختيار ويتم تكرار هذه العملية على جميع المجموعات الفرعية من البيانات الاختبار Training data .

ان تصنيف الشجري يتمثل بحاجتة لعدد كبير من البيانات بأفتراض أن البيانات تتكون من متغير الاستجابة (Y) مع متجة من المتغيرات التنبؤية .

$$X_i=(X_1,X_2,X_3,.....X_m)$$

وتكون بشكل مصفوفة ثابتة (M).

X_i تكون أما متغيرات كمية (مستمرة أو متقطعة) أو أن تكون متغيرات وصفية (اسمية أو رتبوية).

وعند كل عقدة (Node) يجب القيام بما يأتي:

1) أختيار كل التقسيمات المسموح بها للمتغيرات التنبؤية ، عادةً التقسيمات الثنائية تولد أسئلة ثنائية.
2) اختيار أفضل تقسيم أن كلمة الافضل في هذه الخطوة تشير الى مصطلح اختيار بعض معايير حسن التقسيم كما هو الحال بمفهوم (حسن المطابقة) وهناك معيارين مشهورين هما (المربعات الصغرى) و (مطلق التباين الصغرى) . فكلاهما تشير الى مقارنة من الناحية التجانس او التقليل من تطبيق القياس عند العقدة (الاب) .
3) يتم التوقف عن التقسيم في العقدة التي لاتحقق فيها الشروط المطلوبة ، لترتيب المتغيرات X_i في السؤال في الخطوه الاولى .
4) هل ان ($X_i > C$) لكل قيم C التي هي تكون ضمن مجى لـ X_i و اي ان X_i تاخذ اعداد محدودة . ($b_0,b_1,b_2,.....b_i$)

ويكون السؤال هنا: هل ان ($Xm \in C$) عندما C هي ضمن المدى للمجموعات الجزئية [$b_0,b_1,b_2,.....b_i$]

هذه الحالات في الشجرة (T) جوابها أما ان يكون (نعم) الذهاب الى يسار العقدة أو يكون جوابها (لا) يتم الذهاب الى يمين العقدة.

وان الطرائق اعلاه تتوقف عند الخطوه الثالثة عندما التطبيق لاينفذ بشكل جيد. الشجرة تكون كبيره جدا عند العقدة عندما يتكون هناك قليل من البيانات في العقدة النهائية المناظره لها .

في كل عقدة هنالك خوارزمية تبحث في المتغيرات واحدة تلو الاخرى تبدأ من X_1 وتستمر حتى تصل الى X_m ولجميع المتغيرات نجد افضل تقسيم ثم مقارنة مع M افضل تقسيم لمتغير مفرد ثم نختاره هو الافضل .

الخطوه الاولى والخطوة الثانية تتكرر لعقدة الابناء حتى نصل الى نهاية الشجرة .

لذلك فان النموذج الاساسي للشجرة Tree هو :

$$f^{\wedge}(X) = \sum_{m=1}^{n} CmI[(X1, X2) \in Rm]$$

حيث أن:

Cm=node means.

$$C^{\wedge}m = \frac{1}{Nm} \sum_{X_i \in Rm} Y_i$$

6-2 خوارزمية أشجار الانحدار التصنيفية البيزية [8],[7]

Bayesian Classification Regression Trees: BART

ازداد الاهتمام بنماذج بيز في السنوات الاخيرة في التطبيقات الاحصائية ويعود ذلك الى سرعة التطور في مجال تطبيقات الحاسوب الالكتروني ناهيك عن تطور اجهزة الحاسوب ذاتها، مما جعل تطبيق طرائق بيز ممكن عملياً وذو قوة استدلالية منافسة لطرائق احصائية اخرى في شتى فروع التطبيقات الاحصائية ولأي نموذج رياضي يخضع لشروط النظرية الاحصائية.

وان السبب في استخدام اساليب بيز يعود إلى سببين: الأول ان نماذج بيز تسمح بعمل استدلال مترابط. والسبب الثاني ، كون نماذج بيز ملائمة وبصورة خاصة حالة دمج المعلومات المسبقة (prior Information) والتي غالبا ما تكون متوقعة لتكون معدل الظاهرة المبحوثة (مثلاً) و هو ما سيمثله توقع التوزيع اللاحق (Posterior Mean) ، وحسب نوع دالة الخسارة المستعملة في التقدير .

يعدّ اسلوب BART من الاساليب البيزية لتقدير دالة لامعلمية باستعمال اشجار الانحدار بصيغة GMA. اذ تعوّل اشجار الانحدار على تقسيم ثنائي متتابع لفضاءات المتغيرات التوضيحية لغرض تقريب دالة f غير معروفة. ويكون بُعد المتغيرات هو p وهو عدد المتغيرات ذاتها. وبذلك يكون للنماذج ذات الاساس الشجري القدرة والامكانية للأستحواذ على التفاعلات interactions والتاثيرات اللاخطية nonlinearities وتمثيلها بتأثيرات تجميعية للدالة f هذا من جانب، ومن جانب آخر فأن هذه النماذج تدمج مجموع اشجار الانحدار لتكون ذات قدرة أكبر من الاشجار المفردة لوحدها. ولذا يعدّ BART بأنه اسلوب تجميعي ensemble لمجموع اشجار ويكون الاعتماد بالتقدير بالاساس على هذا الاسلوب بالاستناد الى النموذج الاحتمالي البيزي بصورة كاملة.

1-6-2 اسلوب بيز لاختيار المتغيرات [6]

ان اختيار المتغيرات في الانحدار الخطي تشكل الاساس لطريقة الانحدار اللامعلمي وبفرض ان Γ عائلة لنماذج الانحدار الخطي التي لها متغير المعتمد نفسه والى $\gamma \in \Gamma$ فان النموذج :

$$y = X_\gamma \beta_\gamma + E$$

إذ أن :

Y : متجه (n *1) للمتغير المعتمد .

X_γ : مصفـــوفة المتغيرات التوضيحـــية ذات اعمدة تامــة الرتبة (full column Rank) ($N * r_\gamma$)، وأن أعمدة X هي وفقا لعناصر التي تساوي واحد حيث $Y_i=1$ إذا العمود X_{ji} هو الانموذج و $Y_i=0$ إذا العمود X_{ji} ليس في الانموذج .

β : متجة (r x 1) للمعالم المجهولة وان β_γ تتضمن كل عناصر β_i بحيث ان $\gamma_i = 1$.

E : متجه (nx1) للاخطاء العشوائية وله متوسط صفر وتباين :

$$\sigma^2 ln$$

γ : متجه (rx1) لدليل المتغيرات، حيث :

$$\gamma_i = \begin{Bmatrix} 0 \;\; if\; \beta_i = 0 \\ 1 \;\; if\; \beta_i \neq 0 \end{Bmatrix} \qquad i = 1, 2, \ldots\ldots, r$$

ولتحديد دالة الكثافة الاحتمالية للمعلمات المسبقة فأن:

$$P\left(\frac{\sigma^2}{\gamma}\right) \alpha\, 1/\sigma^2$$

وان توزيعها الاحتمالي هو توزيع لوغاريتمي منتظم .

وبإعطاء σ^2, y فالاحتمال المسبق الى $\frac{B_y}{\sigma^2, Y}$ يكون

$$P\left(\frac{B}{\gamma}, \sigma^2\right) \sim N\left(\mu_\gamma, c\sigma^2 (X'_\gamma X_\gamma)^{-1}\right)$$

الهدف من طريقة BART هو تزويد خوارزمية (CART) بأدوات الأسلوب البيزية بخصوص تقسيمات العقد، المواقع والاسئلة المستخدمة في العقد غير المعرفة. يمكن التعامل مع هذه المعالم في هذه المشكلة وعمل استدلال حول استخدام هذه البيانات.

أي نموذج يبنى على اساس شجرة ثنائية (Binary Tree) تعرف من خلال تقسيمات العقد، المتغيرات الموجوده في العقد تقسم على اساس قواعد معينة وهذه المتغيرات تعرف بالتتابع .

S_i , S_i^{var} and $S_i$rule (i=1,…,S_{max})

وان جذر العقده الاساسي دائما يختار في التقسيم الاول للعقدة، مـــــوقع هـــــذه العقدة يدعى بـــ (1) $S_i=1$.

بحيث ان اي تقسيم ينحدر من موقع العقده S_i ويكون وحيد يعرف بموقع الاب ويدعى :

S_i^{parent}

وعندما ينفذ يدعى بـــ : $S_i = 2\, S_i^{parent} + 1$

إذا ماذا نعني بـــ S_i^{var} , S_i^{rules}

على سبيل المثال :

$X_3 < 4.2$

$S_i^{var}=3$, $S_i^{rule}=4.2$

ان عدد العقد النهائية في النموذج يدعى بـ K حيث ان $K\ (S_{max}+1)$ للاستدلال نفترض ان النموذج غير معرف جاء تصنيف

$M_1, M_2, \ldots\ldots, M_k$

حيث ان M_k تشير الى نموذج العقد النهائية لـ K التي تاخذ (K-1) لتقسيم العقد المعلمة θ تشير الى اتحاد المجموعات الجزئية المعدودة.

$$\theta = U_1^{\infty}\theta_k$$

حيث ان θ_k مجموعة جزيئة من $R^{n(k)}$ عندما $R^{n(k)}$ تشير الى مصفوفة من المعالم للنموذج :

$$n(k) = 3(k-1)$$

$$S_i, S_i^{var} \quad MK$$

$$\theta^{(k)} = (S_1, S_1^{var}, S_1^{rule}, \ldots\ldots, S_{k-1}^{var}, S_{k-1}^{rule})$$

$$(K, \theta^{(k)}, y) \quad joint\ distribution$$

$$p(k, \theta^{(k)}, y) = p(k)p(\theta^{(k)})/k)p(\frac{y}{k}, \theta^{(k)})$$

احتمالية النموذج = (حاصل ضرب المعالم الاولية * Likelihood)

يستند الاستدلال البيزي حول $K, \theta^{(k)}$ الى:

$$p(K, \theta^{(k)}/y) \quad joint\ posterior$$

$$p(K, \theta^{(k)}/y) = p\left(\frac{k}{y}\right)p(\frac{\theta^{(k)}}{k}, y)$$

ومنها سوف يتم توليد $joint\ posterior$ وفيما يأتي الخوارزمية البيزية لطريقة (CART) .

2-6-2 خوارزمية BART [8] Algorithm - BART

1- تكون البداية من تكوين شجرة بدون عقد تقسيم حالية.
2- وضع K من العدد متساوي من العقد الطرفية (النهائية) للشجرة الحالية.
3- توليد u كمتغير يتوزع توزيع منتظم قياسي (u(0,1 .
4- يتم تحريك أو تفريع الشجرة حسب نوع (u) :
فإذا كانت $u \le b_k$ يتم الذهاب إلى خطوة الولادة،
أما إذا كانت $b_k < u \le b_k + d_k$ يتم الذهاب الى خطوه الوفاة ،
وفيما عدا هذا ، أي عندما : $(b_k + d_k < u < b_k + d_k + v_k)$ يتم الذهاب الى خطوة المتغير ، وعند عدم حصول أي مما سبق يتم الذهاب إلى خطوة قاعدة التوقف .
5- إذا تم انجاز مسألة الانحدار سيتم الحصول على تباين الخطأ وليكن σ_i^2

إعادة الخطوات (2-5) حتى يتم الحصول على تغير صغير جداً (تقارب) في الاحتمال اللاحق لبُنية الشجرة وهي قاعدة التوقف العامة للخوارزمية ككل .

2 -7 معايير المقارنة [13] Comparison Criteria

هناك ثلاثة معايير مستعملة في المقارنة بين الطرائق المستعملة في الفقرات السابقة وهي كم موضحة في أدناه.

1. اختبار خطأ التصنيف (Error of Classification)

حيث تكمن فكرة هذا الاختبار في بناء مصفوفة الخطأ (Cm) والتي تكون مصفوفة مربعة بحجم عدد القنات المصنفة ومن ثم حساب القطر الرئيسي فاذا كان القيمه المصنفة مساويه للعينة الاصلية تعطى قيمه 1 عدا ذلك تعطيى قيمة صفر وبعد ايجاد المصفوفة يتم حساب الاختبار حسب القانون الاتي:

$$Error\ of\ classification = \frac{\left(N - sum(diag(cm))\right)}{N}$$

حيث ان :

$$N = \sum cm$$

2. اختبار دقة التصنيف Test Accuracy of classification

تكمن فكرة هذا الاختبار في حساب المتوسط للبيانات المصنفة صحيحاً وقانون الاختبار كالاتي:

$$Accuracy\ of\ classification = \frac{\sum (y = \hat{y})}{num(y = \hat{y})} \times 100$$

3. جذر متوسط مربعات الخطأ Root mean square error

وهو عبارة عن جذر قانون متوسط مربعات الخطأ ويحسب كالاتي:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y - \hat{y})^2}{N}}$$

3- الجانب التطبيقي

3-1 وصف البيانات Data Description

إن هدف البحث اي اداة تنصيف هو تصنيف المشاهدات الى مجموعتين او اكثر للوصول الى التنبؤ بنتيجة ترتبط بكل مشاهدة ومثال على ذلك: ذكر أو انثى، وراثي أم غير وراثي ... الخ ، وتقنيات التصنيف تزودنا بنماذج تنبؤيه للغرض نفسه.

و تستخدم تقنيات التصنيف بصوره شائعة جداً في الكثير من المجالات التطبيقية. سيما في المجالات الطبية والاقتصادية والتحليلات المالية، تم التناول في هذا البحث المجال الطبي وذلك بإمكانية التنبؤ بنسبة السكر في الدم بالنسبة للأطفال الذين تتراوح اعمارهم (15) فما دون.

علماً أن متغير الاستجابة (Y) هو نسبة السكر في الدم، والمتغيرات التفسيرية هي كما يلي:

1. العمر (Age) تتراوح اعمار المرضى المشمولين ضمن العينة من عمر 15 سنة فما دون.
2. الجنس (Sex) شملت عينة المرضى من كلا الجنسين .
3. هل المريض لدية وراثة (وراثي ، غير وراثي) .
4. الوزن (weight).
5. هل مريض السكر مرض آخر (نعم ، كلا) .
6. هل توجد قرابة بين الام والاب (نعم ، كلا) .

3-2 النتائج Results

3-2-1 الانحدار اللوجستي

بعد تطبيق تقدير نسبة السكر في الدم باستعمال نموذج الانحدار اللوجستي وبحسب الصفات التصنيفية كانت النتائج كما يأتي:

جدول (1)

يوضح تقدير نسبة السكر بالدم باستخدام خوارزمية الانحدار اللوجستي

$\widehat{y}_{LR}$	y	$\widehat{y}_{LR}$	y	$\widehat{y}_{LR}$	y
0	0	2	2	3	3
0	0	2	2	2	3
0	0	3	2	0	3
0	0	2	2	3	3
0	0	1	2	2	3
0	0	1	2	3	3
0	0	2	2	3	3
0	0	2	2	1	3
0	0	2	2	3	3
0	0	2	2	3	3
0	0	2	2	3	3
0	0	2	2	3	3
0	0	2	2	3	3
0	0	2	2	3	3
0	0	3	2	3	3
0	0	2	2	3	3
3	0	0	2	2	3
0	0	3	2	0	3
0	0	2	2	3	3
0	0	3	2	2	3
0	0	3	3	3	3
1	1	1	3	3	3
1	1	3	3		
1	1	3	3		
1	1	3	3		
1	1	3	3		
0	1	3	3		
1	1	3	3		
1	1	3	3		
3	1	3	3		
1	1	3	3		
1	1	3	3		
1	1	3	3		
1	1	3	3		
3	1	3	3		
1	1	3	3		
1	1	3	3		
2	1	1	3		
1	1	3	3		
1	1	0	3		
1	1	0	3		
2	2	2	3		

أما نتائج خطأ التصنيف ودقة التصنيف ومعيار افضليه التقديرللانحدار اللوجستي كانت كالآتي:

جدول (2)

يوضح المعايير المستخدمة في تصنيف وتقدير نسبة السكر بالدم باستخدام خوارزمية الانحدار اللوجستي

Test Set Accuracy	74.75728%
Error of Classification	0.2524
RMSE	0.8923

2-2-3 التصنيف الشجيري CART

أما عند تطبيق اشجار التصنيف على بيانات السكر كانت نتائج التنبؤ بنسبة السكر بالدم حسب الصفات التصنيفية كالآتي :

جدول (3)

يوضح تقدير نسبة السكر بالدم باستخدام خوارزمية التصنيف الشجيري CART

$\hat{y}_{CART}$	y	$\hat{y}_{CART}$	y	$\hat{y}_{CART}$	y
0	0	2	2	3	3
0	0	3	2	3	3
1	0	1	2	3	3
1	0	3	2	0	3
3	0	1	2	3	3
0	0	1	2	3	3
2	0	1	2	3	3
3	0	3	2	0	3
1	0	1	2	0	3
0	0	2	2	3	3
1	0	2	2	2	3
1	0	2	2	3	3
1	0	3	2	2	3
0	0	2	2	2	3
1	0	3	2	3	3
2	0	1	2	3	3
0	0	3	2	3	3
0	0	2	2	3	3
1	0	3	2	0	3
1	0	2	3	3	3
1	1	1	3	3	3
1	1	1	3		
3	1	3	3		
2	1	3	3		
1	1	1	3		
1	1	3	3		
1	1	3	3		
0	1	0	3		
1	1	3	3		
3	1	1	3		
1	1	0	3		
2	1	3	3		
1	1	3	3		
3	1	3	3		
1	1	3	3		
1	1	1	3		
0	1	3	3		
1	1	3	3		
0	1	3	3		
2	1	0	3		
1	2	2	3		

اما نتائج خطأ التصنيف ودقة التصنيف ومعيار افضليه التقدير للتصنيف الشجيري كانت كالاتي:

جدول (4)

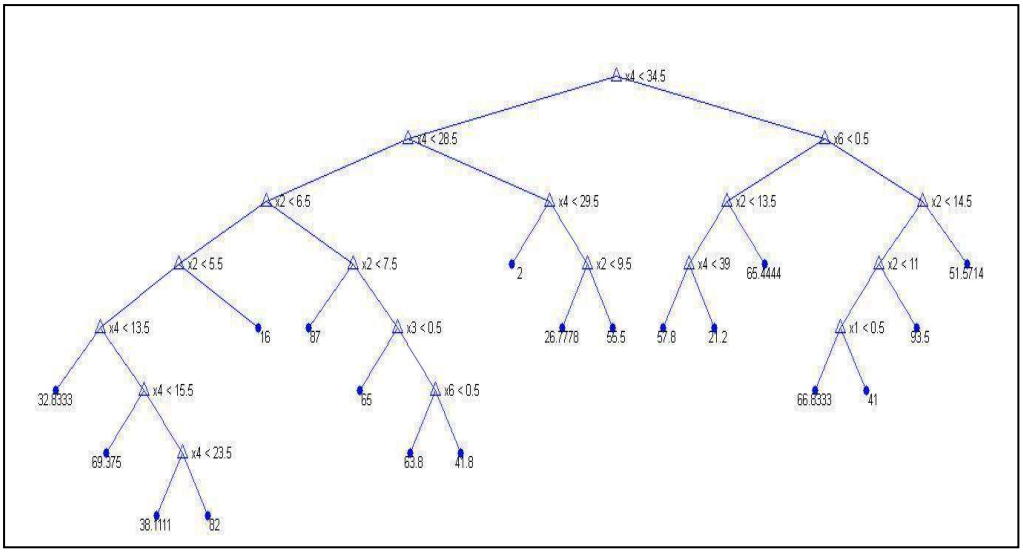
يوضح المعايير المستخدمة في تصنيف وتقدير نسبة السكر بالدم باستخدام خوارزمية التصنيف الشجيري CART

Test Set Accuracy	30.883495%
Error of Classification	0.9612
RMSE	1.2267

وشجرة الانحدار التصنيفية وكما موضحة في ادناه:

الشكل (2)

يوضح شجرة الانحدار التصنيفية CART



3-2-3 الخوارزمية البيزية BART

عند تطبيق الخوارزمية البيزية BART بعد تصنيف البيانات تم الحصول على النتائج كالآتي:

جدول (5)

يوضح تقدير نسبة السكر بالدم باستخدام الخوارزمية البيزية BART

$\hat{y}_{Baysian}$	y	$\hat{y}_{Baysian}$	y	$\hat{y}_{Baysian}$	y
2	2	1	1	0	2
1	1	1	1	0	0
2	2	1	0	0	1
1	1	1	1	1	0
1	2	0	1	1	0
2	2	1	1	1	1
2	0	1	1	0	1
1	1	0	1	0	2
1	1	1	1	0	1
1	1	2	2	0	0
1	1	1	0	0	1
1	1	1	0	0	1
1	2	1	0	1	1
1	1	0	1	0	1
1	1	1	1	0	1
1	1	0	1	0	0
0	1	1	1	1	1
0	2	0	0	1	1

$\hat{y}_{Baysian}$	y	$\hat{y}_{Baysian}$	y	$\hat{y}_{Baysian}$	y
0	2	1	1	0	0
1	2	2	1	1	1
2	1	1	0	0	0
0	2	1	0	0	1
0	0	1	2	0	0
0	0	0	1	0	0
0	1	1	2	0	1
2	1	2	1	0	1
0	1	0	2	1	0
0	1	0	1	0	1
1	1	0	1	2	0
2	0	2	0	1	1
0	1	2	1	1	2
0	1	0	0	1	1
0	1	0	0	2	1
0	1	1	1	1	1
2	1	1	1	2	2
0	1	1	1	1	1
0	0	1	1	1	2
1	1	1	1	1	0
2	1	0	1	2	1
2	2	1	1	1	0
0	0	2	0	2	1

أما نتائج خطأ التصنيف ودقة التصنيف ومعيار أفضلية التقدير للخوارزمية البيزية كانت كالآتي:

جدول (6)

يوضح المعايير المستخدمة في تصنيف وتقدير نسبة السكر بالدم باستخدام الخوارزمية البيزية BART

Test Set Accuracy	45.5285%
Error of Classification	0.5447
RMSE	0.9017

الاستنتاجات Conclusions

1. **كانت نتائج خوارزمية الانحدار اللوجستي قد سجلت أفضل النتائج عن باقي الطرائق ولجميع المعايير المستخدمة في تصنيف وتقدير نسبة السكر في الدم.**
2. **بينما سجلت خوارزمية البيزية BART المرتبة الثانية بعد خوارزمية الانحدار اللوجستي من حيث النتائج ولجميع المعايير المذكورة في أعلاه.**
3. **وأخيراً سجلت خوارزمية التصنيف الشجيري المرتبة الأخيرة عن الطرائق الاخرى من حيث النتائج وأيضاً لجميع المعايير سابقة الذكر.**
4. **ولقد تبين بأن الانحدار اللوجستي قد سجل أفضل النتائج من ناحية تقدير نسبة السكر ويليه في ذلك خوارزمية BART وأخيراً خوارزمية CART.**

المصادر References

1. البلداوي، تنسيم حسن كاظم، (1996)، "مقارنة تحليله بين نماذج اللوجستيك ونماذج الدوال التميزية". أطروحة دكتوراه - جامعة بغداد، كلية العلوم الاقتصادية.
2. ذياب، وسام سرحان، (2006)، "استخدام بعض الطرائق الاحصائية والتصنيف الشجري في التصنيف والتنبؤ بإفلاس الشركات مالياً". رسالة ماجستير، جامعة بغداد، كلية الادارة والاقتصاد.
3. شاهين، حمزة أسماعيل، (2014)، "مقارنة بين بعض طرائق التصنيف الخطية مع تطبيق عملي". مجلة العلوم الاقتصادية والادارية، المجلد 20، العدد 80، الصفحات 393-410.
4. علي، عمر عبد المحسن، (2007)، "مقارنة مقدرات النماذج التجميعية المعممة باستخدام الشرائح التمهيدية عند تحليل الانحدار اللامعلمي وشبة المعلمي". أطروحة دكتوراه، جامعة بغداد، كلية الادارة والاقتصاد.
5. غانم، عدنان، والجوعاني، فريد خليل، (2011) "استخدام تقنية الانحدار اللوجستي ثنائي الاستجابة في دراسة أهم المحددات الاقتصادية والاجتماعية لكفاية دخل الاسرة". مجلة جامعة دمشق للعلوم الاقتصادية والقانونية، العدد (1).
6. يوسف، خلود يوسف خمو، (2004)، "مقارنة اساليب بيز مع طرائق اخرى لتقدير منحى الانحدار اللامعلمي". اطروحة دكتوراه، جامعة بغداد، كلية الادارة والاقتصاد.
7. Adam and Justin, (2014), "BART Machine: Machine Learning with Bayesian Additive Regression Trees". Statistical Learning , non – parametric, R, Java.
8. Banik. Mallick , David G.T. Denison and Adrian F.M. Smith, (1998), "A Bayesian CART Algorithm". Biometrika, Vol. 85, No. 2, (363-377).
9. Breiman , L. , Friedman J.H. , Olshen , R.A. , and Stone , C.J. (1984) . "Classification and Regression Trees". Springer Inc.
10. Francesco, Mola & Raffaele, Miele, (1998), "Evolutionary Algorithms for Classification and Regression Trees". Dipartimento di Economia, Universita di Cagliari, Italy.
11. Huge Chipman , Edward I. George & Robert E. Culloch (2001) , "Bayesian Treed Models". University of Waterloo. Available online at: http://www-stat.wharton.upenn.edu/~edgeorge/Research_papers/treed-models.pdf
12. Imrankurt, Omurlu, Mevlut, True, Merre Katranci, Mustafa Uunbol & Engin Guney, (2014), "Comparing performances of Logistic Regression , Classification & Regression Tree and Artificial Neural Networks for Predicting Albuminuria in Types 2 Diabetes Mellitus".
13. Pratola, M. T., (2014), "Regression Tree Models". Dep. Of Statistics, The Ohio State University

……..……..

……..……

……