الجمهورية العربية السورية

وزارة التعليم العالي

المعهد العالي لإدارة الأعمال

قسم إدارة العمليات والمعلومات

تصميم عروض حسب الشرائح الزمنية باستخدام التنقيب في المعطيات

إعداد الطالبة

نهى حمود

بإشراف

د.كادان الجمعة

مشروع تخرج

العام الدراسي

٢٠٢٠-٢٠٢١

إهداء

إلى من كان فخراً لي أعلو و أسمو به.....

والدي

إلى من آمنت بي وكانت أملي.....

والدتي

إلى تلك الوحيدة وهي الجميع وهي الحياة بأكملها.....

أختي

إلى من أرى فيه المستقبل.....

أخي

إلى جميع الأصدقاء الداعمين الذين جعلوا للحياة طعماً آخر، لا أعلم إن كنت أعطيت ولكن أعلم جيداً أنني أخذت منكم الكثير.....والمخصوصين بالذكر منهم

(محمد، يامن، أية، أجود، نور، ملاك، لين، حسن)

إلى الغائبين الذين لا يراوحون البال.....

(عبدالله، كنان، نيرمين)

إلى من أعطى من ذاته و أنار دربي بالعلم.....الدكاترة الفاضلين والمخصوصين بالذكر منهم

د.كادان الجمعة

د.وائل خنسة

إلى جميع من جعل هذا العمل ممكناً، وإلى جميع ما جعلني ما أنا عليه اليوم.....

الملخص:

يهدف البحث إلى تصميم عروض للزبائن باستخدام المنهج التنبؤي الذي يعتمد عليه علم التنقيب في المعطيات، حيث كانت فترة جمع البيانات ممتدة من تاريخ ٢٠١٣/١١/٢ و حتى ٢٠١٣/١١/٧ وقامت الباحثة بتحليل مجموعة من المعطيات تتكون من سمة واحدة تتعلق هذه السمة بحجم استهلاك الخدمات المقدمة من شركة الاتصالات.

تم شرح محتوى المعطيات وتحليلها من خلال استخدام الأداة google colaboratory، ولغة البرمجة python، واستيراد بعض المكاتب (os– mlxtend– numpy– pandas– seaborn– networkx– warnings– matplotlib– datetime– sklearn)، و أيضاً استيراد بعض المبادئ الإحصائية(association rules– apriori) و بعدها تطبيق خوارزمية kmeans للتعامل مع المعطيات وتم ذلك على المراحل التالية: المرحلة الأولى (مرحلة التحضير للتنقيب على المعطيات) حيث تم استكشاف شكل المعطيات ونوعها وقرائتها باستخدام مكتبة (pandas)، وفي المرحلة الثانية (مرحلة المعالجة المسبقة للمعطيات) تم تحويل نمط المعطيات لنمط مقروء وبعدها تم تجميع المعطيات بحسب الخلية (cell id) وبما أن مدة تجميع المعطيات كانت قصيرة تم إنشاء أعمدة جديدة تحوي (الساعة، الدقيقة، الثانية) لدراسة تفاوت إستهلاك الخدمات بشكل أكثر دقّة ومن ثم تم حذف المعطيات ذات القيمة (null) للحصول على نتائج دقيقة، وفي المرحلة الثالثة (مرحلة معالجة المعطيات) تم إنشاء سمة (TeleV) والمكوّنة من مجموع الخدمات المقدمة من شركة الاتصالات، وبعدها تم استكشاف القيم الشاذّة والتخلص منها وبعدها تم تحويل المعطيات للتوزيع الطبيعي ومن ثم تم تطبيق خوارزمية (Kmeans) مع طريقة (Elbow_curve) لمعرفة عدد العناقيد الأمثل، ولم تكن النتائج دقيقة لذلك تم تالياً تطبيق خوارزمية (Kmeans) مع التابع (silhouette score) لتحديد عدد العناقيد الأمثل بشكل أكثر دقّة، وكان العدد الأمثل للعناقيد (٢) وبعدها تم تطبيق خوارزمية (Kmeans) باستخدام عدد العناقيد السابق للحصول على العنقودين وبعدها تم تحديد العناصر المنتمية للعنقود الأول والعناصر المنتمية للعنقود الثاني ومن بعد الحصول على العنقودين تم تحديد العنقود الذي يحمل هذه السمة بشكل أكبر لتحديد الفئة التي يجب علينا أن نصمم لها عروضاً خاصة.

الفهرس:

الفصل الأول

الإطار التمهيدي

١- المقدمة:

أدّى التقدم العلمي وانتشار استخدام التكنولوجيا في شتى مناحي الحياة اليومية إلى رفع القدرة في توليد وجمع المعطيات بشكل سريع في هذا العصر، وقد ساهم ذلك في حوسبة معظم الأعمال والعلوم والخدمات التي يتم تقديمها يومياً في كل مكان حول العالم، كما أن التقدم التكنولوجي تسبب في ظهور أنواع جديدة من المعطيات كالنصوص والصور والفيديو وأنظمة متعددة المهام بالإضافة إلى شبكة الإنترنت التي تحتوي كميات هائلة من المعطيات بكافة أشكالها. كل ذلك أدى إلى تضخم غير مسبوق في كميات المعطيات التي يتم تخزينها يومياً.

تولّد صناعة الاتصالات وتخزن كمية هائلة من المعطيات. تتضمن هذه المعطيات تفاصيل المكالمات التي تصف المكالمات التي تعبر شبكات الاتصالات وبيانات الشبكة التي تصف حالة مكونات الأجهزة والبرامج في الشبكة ومعطيات العميل التي تصف عملاء الاتصالات، تسمى المعطيات من هذا النوع بالمعطيات الكبيرة. المعطيات الكبيرة هي مصطلح يطلق على مجموعات ذات الحجم الكبيرة والمعقدة بحيث يصبح من الصعب معالجتها باستخدام أدوات معالجة المعطيات التقليدية. يمكن تمييز المعطيات الكبيرة بثلاثة مقاييس: الحجم (كميات كبيرة من المعطيات)، التنوع (يشمل أنواعاً مختلفة من المعطيات)، والسرعة (تراكم معطيات جديدة باستمرار).[١]

Volume	**Variety**
Velocity	**Veracity**

الشكل(١) مقاييس تمييز المعطيات الكبيرة

تصبح المعطيات كبيرة عندما يتجاوز حجمها أو سرعتها أو تنوعها قدرات أنظمة تكنولوجيا المعلومات على تخزينها وتحليلها ومعالجتها. المعطيات الكبيرة هي في الواقع جديدة. يمكن جمع المعطيات الكبيرة ليس فقط من أجهزة الكمبيوتر، ولكن أيضاً من مليارات الهواتف المحمولة

[1] (Abdel hafez, 2016)

ومنشورات الوسائط الاجتماعية وأجهزة الاستشعار المختلفة المثبتة في السيارات وعدادات المرافق والشحن والعديد من المصادر الأخرى. في كثير من الحالات، يتم إنشاء المعطيات بشكل أسرع مما يمكن تحليله مسبقاً.

لقد وعد ظهور تقنية التنقيب عن المعطيات بحلول لهذه المشاكل، ولهذا السبب كانت صناعة الاتصالات السلكية واللاسلكية من أوائل المتبنين لتقنية التنقيب عن المعطيات. تطرح معطيات الاتصالات العديد من القضايا المثيرة للاهتمام للتنقيب عن المعطيات، كالمقياس (حيث أن قواعد معطيات الاتصالات قد تحتوي على بلايين من السجلات وهي من بين الأكبر في العالم)، وأن المعطيات الخام غالباً ما تكون غير مناسبة للتنقيب عن المعطيات (كل من تفاصيل المكالمة ومعطيات الشبكة عبارة عن معطيات سلاسل زمنية تمثل الأحداث الفردية. قبل أن يتم استخراج هذه المعطيات بشكل فعال، يجب تحديد الميزات المفيدة ثم يجب تلخيص المعطيات باستخدام هذه الميزات).

١-١المشكلة:

يعالج البحث بشكل أساسي مشكلة تصميم العروض للزبائن حسب الشرائح الزمنية في شركات الاتصال باستخدام تقنيات التنقيب في المعطيات. وبالتالي يمكن تلخيص مشكلة البحث بالسؤال التالي:

كيف يمكن تطبيق تقنيات التنقيب في المعطيات لتصميم عروض للزبائن بحسب الشرائح الزمنية؟

١-٢ أهداف البحث:

يهدف البحث إلى تصميم عروض لزبائن شركات الاتصال بالاعتماد على الشرائح الزمنية وذلك من خلال الإجابة عن تساؤل البحث المذكور في مشكلة البحث، ويمكن تلخيص الأهداف وفق الآتي:

١- تصميم عروض باستخدام إحدى تقنيات التنقيب في المعطيات.

٢- تحديد ميزات المعطيات لتسهيل تصميم العروض حسب الشرائح الزمنية.

١-٣ منهجية البحث:

المنهجية المستخدمة هي المنهجية التنبؤية والتي يعتمد عليها علم التنقيب في المعطيات في تنبؤاته، وهو ما يعرف بالنموذج التنبؤي (predictive model) و هو يرتكز على نموذج إحصائي. و يتم بناء هذا النموذج بعدة طرق والتي تقوم على نظريات الإحصاء والمعادلات الرياضية ومن أهم هذه الطرق على سبيل المثال لا الحصر: التصنيف، الانحدار،التجميع.

١-٤ حدود البحث:

تتلخص حدود البحث وفق الآتي:

- حدود زمانية: تم إعداد البحث خلال المدة الزمنية الممتدة ما بين ٢٠٢١/٤/١٢ و ٢٠٢١/٦/٣٠.

١-٥ أهمية البحث:

تتجلى أهمية البحث من خلال توضيح ومعرفة العديد من المفاهيم والتعاريف والمصطلحات المتعلقة بالدراسة (التنقيب في المعطيات، أدوات التنقيب في المعطيات، المعطيات الكبيرة وتصميم العروض للزبائن حسب الشرائح الزمنية) ومن خلال النتائج المتوقع الحصول عليها بتطبيق التنقيب في المعطيات للحصول على عروض مصممة بحسب الشرائح الزمنية.

١-٦ معوقات البحث:

١- صعوبة الحصول على دراسات سابقة مفصّلة مشابهة لمنهجية البحث المستخدمة كون المشاريع المشابهة عادة ما تتسم بالحماية الفكرية وبالتالي المعلومات التي تحويها تبقى محدودة.

٢- صعوبة الحصول على المعطيات المستخدمة في البحث.

٣- ضيق الوقت اللازم لإنهاء المشروع.

٢- الدراسات السابقة:

تعرض الدراسات السابقة باللغتين العربية والإنكليزية الأعمال التي نفذت للهدف نفسه.

٢-١ الدراسات العربية:

٢-١-١ دراسة (مراد و الكردي، ٢٠٢٠):

عنوان الدراسة: نظام إحالة لإدارة المعرفة يعتمد على تقنيات الذكاء الصنعي.

يوضح هذا البحث مستقبل نظام إدارة المعرفة وآلية ارتباطه بالذكاء الاصطناعي في المنظمات، وخاصة عندما يتعلق الأمر بتقديم مساعدات الطوارئ الإنسانية والخدمات والرعاية الصحية مثل جائحة فيروس كورونا الحالية سيسمح النظام المقترح للمستفيدين والموظفين والكيانات الرسمية الخارجية بالحصول تلقائياً على اجابات فورية للاستفسارات المتعددة دون الحاجة لتدخل بشري إلا عند الضرورة وما يوافقها من عمليات التخزين والنقل والاسترجاع وتوليد معارف أخرى لإغناء النظام بالمعلومات المحدثة المطلوبة سريعاً، من خلال المرور بثلاث مستويات اعتماداً على

النهج الدلالي وخوارزميات معالجة اللغة الطبيعية وأيضا باستخدام تقنيات الأنطولوجيا بمستوى الأجوبة التلقائية المستخرجة من النظام بالمستوى الأول مروراً بنظام الدردشة التفاعلي مع موظف متوسط الخبرة بالمستوى الثاني وأخيراً هناك خيار إرسال الاستفسار عبر البريد الالكتروني لخبير عالي المستوى بالمستوى الثالث. وقد ثبت صحة الطريقة المتبعة في هذا النظام المقترح التفاعلي المتكامل الذكي. وتم إظهار فعالية هذا النهج عن طريق اختباره بأحد هذه المنظمات وتبين أن النتائج التجريبية كانت مشجعة للغاية لاسيما أن معظم تلك المنظمات تملك نظام معرفة يدوي وليس ألي كما هي الدراسة بالبحث العلمي الحالي.

و قد خلصت هذه الدراسة إلى مجموعة من النتائج أهمها:

توضح ورقة البحث هذه كيف يمكن لأي مستفيد الاستعلام عن المعرفة واستخدامها للإجابة على الاستفسارات المختلفة حسب الحاجة واستثناءات المعلومات المحدثة والمفيدة. الإجابات المتراكمة تغذي النظام بشكل مستمر من خلال إجابات خبراء التدخل الإنساني المخزنة على النظام تلقائياً. إطار عمل متقدم غني لغوياً لأتمتة مراقبة التقاط المعلومات لكل من الموظفين والمستفيدين بين المنظمات وعبرها. يتيح للمستخدم تحديد الخيار الذي يناسب احتياجاته على أفضل وجه، من بين خيارات النظام الممكنة المختلفة التي توفرها المراحل المختلفة. استخراج العناصر الأساسية للمعرفة وتخزينها تلقائياً. الخطوة الأولى نحو بناء نظام تفاوض بشأن السياسة، والذي يمكنه التفاوض مع الأهلية أو التعليمات نيابة عن المستخدم.

٢-١-٢ دراسة (السبيعي ،٢٠٢٠):

عنوان الدراسة: استخدام تقنيات التنقيب في المعطيات للتنبؤ بنتيجة الحملات التسويقية المباشرة.

يهدف البحث من خلال اعتماده النهج التنبؤي لتنبؤ بالحملات التسويقية المباشرة (الهاتفية) لتوقع إيداعات العملاء في بنك (Banco Português de Investimento) البرتغالي/بورتو، وكانت فترة جمع المعطيات ممتدة من (٢٠٠٨) إلى (٢٠١٣) وتحديدا من الشهر (٤) إلى الشهر (١٢) في كل سنة، وقمنا بتحليل مجموعة من المعطيات تتكون من (٢٣) سمة يتعلق أغلبها بعميل البنك وبعض منها يخص جهة اتصال والحملة التسويقية، وتمت إضافة سمات اجتماعية واقتصادية جديدة.

تم شرح محتوى المعطيات وتحليلها من خلال تصوير المعطيات (dv) لفهم المعلومات التي تحملها سمات هذه المعطيات من أجل البحث في تقنيات التنقيب عن المعطيات لاختيار تقنية ملائمة لطبيعة المعطيات والمشكلة المدروسة، وقد تم اختيار تقنية التصنيف رداً لطبيعة المعطيات التي تحمل سمة (attributes) إخراج مكونة من خيار ثنائي (yes/no) وتم اختيار (٥) خوارزميات للتصنيف سيتم بناء المصنفات من خلالهم وهم " الغابات العشوائية (random forest)، شجرة القرار (decision tree j48)، الانحدار اللوجستي (logistic regression)، الجار الأقرب (knn)، و بايز الساذج (naive bayes)"، ثم تم تحضير المعطيات المدروسة لإدخالها على خوارزميات التصنيف وقد تم بمرحلة تحضير المعطيات المتألفة من عدة مراحل بالاستغناء عن عدة سمات (attributes) من خلال تحليل الارتباط الذي يبحث عن الارتباطات المخفية بين هذه السمات، وقد تم حل مشكلة عدم توازن المعطيات باستخدام مبدأ (random oversampling)، و بعد عملية التحضير سيتم تطبيق الخوارزميات السابقة عليها لبناء المصنفات واختيار المصنف الأمثل من خلال المقارنة في ما بينهم بعدة معاير وهم الدقة (accuracy) والإحكام (precision) والحساسية (recall)، وبعدما تمت المقارنة بين نتائج المصنفات الخمس المستخدمة إتضح أن مصنف الغابات العشوائية (random forest) قد حصل تميز بجميع نتائج معاييره على المعايير الأخرى من حيث الدقة بمعدل (٩٥,٣%)، وإضافة نتائج خاصة بسمات معينة للحصول على نتائج عالية للحملات التسويقية جنباً بجنب مع مصنف الغابات العشوائية (random forest).

و قد خلصت هذه الدراسة إلى مجموعة من النتائج أهمها:

- نتائج نظرية:

١- الاستخدامات العديدة التي تقدمها علوم البيان من خلال التنقيب في المعطيات وتقنياتها العديدة في مجال البنوك بشكل عام ومن ثم تطبيق التقنيات على الأقسام العديدة التابعة للبنك (قسم التسويق) بشكل خاص، لما له محاسن من خلال التنبؤ ووصف العميل التابع للبنك بشكل دقيق عبر المعطيات الخاصة بالعميل.

٢- لم يعد للمدراء المقدرة على قياس تلك المعطيات من خلال النظر إليها، ومن جوهر المشكلة تأتي النتيجة باستخدام تقنيات التنقيب في المعطيات لفهم العميل واحتياجاته وتصنيفه والتنبؤ بالحركة القادمة له، مما يعطي مصداقية أكبر للعميل اتجاه الحملات وفي المقابل يحصر

المعطيات اللازمة للحملة من خلال هندسة المعطيات، مما يحدد المعطيات اللازمة معرفتها من العميل دون تطفل وإزعاج لنجاح نتائج الحملات التسويقية على أتم وجه.

٣- تطبيق تقنيات التنقيب في المعطيات (التصنيف) لتوقع إيداعات العملاء يتيح النظر مرة أخرى في المعطيات لهندستها من بعد تحديد قسم التسويق للمعطيات اللازمة من خلال تحضير المعطيات الخاصة بالحملة التسويقية ومعرفة تأثيرات وارتباطات سمات الحملة بين بعضها وحذف السمات التي قد تؤثر على دقة التنبؤ.

٤- استخدام خوارزميات التصنيف في المشكلة المدروسة وعدم استخدام تقنيات أخرى يعطي أفضل نتائج من خلال طبيعة المعطيات الموجودة، وعلى أساسها أصبح ممكناً التوقع لإيداعات العملاء بشكل كبير والنتائج التطبيقية أثبتت ذلك.

٥- إضاقة السمات الاقتصادية والاجتماعية أعطت التنبؤ بنتائج الحملات التسويقية المباشرة دقة كبيرة بتوقع إيداعات العملاء، لأنها تربط بين ثقافة وثقة ورأي مجتمع كامل من وجهة وتغير الأسعار وزيادة وانخفاض أسعار الفائدة على القروض من وجهة أخرى على الصعيد المحلي البرتغالي، مما يربط بين الأموال التي يملكونها العملاء وإمكانية إيداعاتهم من خلال تحليل المؤشرات الصادرة من البنك البرتغالي المركزي.

٦- استخدام تقنيات التنقيب في المعطيات لها أثر كبير على تخفيض تكلفة الحملات التسويقية وإدارة الحملة بالشكل الأمثل، وزيادة إيرادات البنك عن طريق التوقع الصحيح للعملاء الذين سيشتركون بودائع استثمارية، واستهداف شرائح كانت مهمشة وغير مركز عليها.

- نتائج تطبيقية:

١- يعرف بأن محور الحملة التسويقية المباشرة (الهاتفية) هي المكالمة لذلك يمكن من خلال الالتزام بمدة المكالمة التي حددت وهي بين (٩) و (١٤) دقيقة فيمكن الحصول على إيداعات من العملاء بنسبة أكبر وبالتالي تخفيض تكلفة الحملات من خلال المعرفة التقريبية لنسبة مدة الاتصال.

٢- يجب التركيز على الفئات العمرية التي بين (٣٠) و (٤٥) لأنها تملك أكبر شريحة عمرية من العملاء بشكل عام، وبشكل خاص هم أكثر شريحة تشترك بودائع بين العملاء المستهدفين،

الاستفادة من الفئة العمرية ما فوق (٦٠) بسبب حركة تجاوبهم مع الحملة التسويقية لكن بنسب صغيرة.

٣- يمكن للمؤشرين سعر يوربيور (يوربيور٣) وثقة المستهلك (cons.conf.idx) عند معدلات معينة في أيام وأشهر إمكانية الاعتماد عليهم بالنسبة لتكثيف الحملة مرتبطاً بمدة المكالمة الناتجة للحصول على حركة إيداع أكبر من قبل العملاء، وتخفيض الحملة نسبياً بالمعدلات التي أظهر فيها نسب كثيرة لعدم الإيداع، حيث يتأثر المؤشرين بشكل ملحوظ بعملية نتائج الحملة وهذا منطقي من خلال المنطق الاقتصادي بحيث هناك ارتباط بين قدرة العميل على الشراء ورأيه وأمانه على نسبة إيداعه، وكذلك سعر اليوربيور بحيث معرفته بسعر الفائدة على معاملات اليورو ما بين المصارف يؤثر على احتمالية إيداعه.

٤- تم بناء (٥) مصنفات للمعطيات المدروسة واختيار أمثل مصنف واعتماده لتوقع إيداعات العملاء الاستثمارية، وقد كان مصنف الغابات العشوائية هو الأفضل من بين المصنفات المستخدمة متفوقاً عليهم بجميع المعايير بنسبة دقة (٩٥,٣%).

٥- يجب الأخذ بعين الاعتبار جميع النتائج بالإضافة لمدة المكالمة مع للجميع فهي أساس نجاح الحملة إن أتمت على أكمل وجه مع النتائج السابقة.

٣-١-٢ دراسة (الجناعي و الحداد و البار و الزهاري، ٢٠١١):

عنوان الدراسة: استكشاف بعض الأنماط المؤثرة في الأداء الأكاديمي لطلاب جامعة العلوم والتكنولوجيا باستخدام تقنيات التنقيب في البيانات.

يقدم هذا البحث دارسة تطبيقية في مجال استكشاف المعرفة knowledge discovery باستخدام تقنيات التنقيب في المعطيات data mining. الهدف الأساسي من الدارسة هو اكتشاف بعض الأنماط السائدة في المعطيات الأكاديمية للطلاب في جامعة العلوم والتكنولوجيا اليمنية منذ العام ١٩٩٤ و حتى العام ٢٠٠٥م ومن ثم الخروج بمؤشرات عامة حول الأداء الأكاديمي لدعم السياسات التعليمية لدى متخذي القرار في الجامعة لاسيما وأن حجم المعطيات وكذلك البعد الزمني الكبير نسبياً لهذه المعطيات يدعم من نتائج هذا البحث. في هذه الدراسة تم اكتشاف بعض الأنماط patterns السائدة في هذه المعطيات، وقد خلص البحث إلى وجود مجموعة من الأنماط التي يمكن أن تعطي مؤشرات ذات دلالة في الجانب التعليمي. من هذه

الأنماط وجود ارتباط بين مستوى تحصيل الطالب لبعض المواد، وبين معدل الطالب في الثانوية واختيار التخصص في الجامعة، و كذلك علاقة المنح الدراسة بمستوى تحصيل الطالب أكاديمياً. قدم الباحث محاولة لقراءة هذه النتائج وتفسيرها وعرضها والتحقق من مستواها ونوعيتها وذلك باطلاع متخذي القرار بالجامعة عليها. تم اختيار تقنيات التنقيب في المعطيات كونها الأنسب للاستفداة من حجم هذه المعطيات وكذلك لأنها تستخدم خوارزميات استنباطية ذكية تستخدم غالباً لدعم اتخاذ القرار. استخدمت طرائق مختلفة من تقنيات التنقيب في المعطيات لدعم النتائج المكتشفة وهي قواعد الارتباط association rules و التصنيف بأشجار القرار classification using decision trees وذلك بعد عملية المعالجة الأولية preprocessing لقاعدة المعطيات وإعادة هيكلتها على شكل مستودع معطيات منطقي data warehouse logical. و قد استخدمت خوارزميتي apriori, predective apriori في تنقية قواعد الارتباط، و خوارزميتي D3, J48 في تقنية التصنيف بأشجار القرار. هذه الطرائق والخوارزميات تم تطبيقها من خلال الأداة WEKA التي تدعم الكثير من الخوارزميات والطرائق للتنقيب في المعطيات. و في الأخير تم استخلاص الاستنتاجات واقتراح بعض التوصيات التي تهم صانع القرار للاستفادة منها في تحسين الأداء الأكاديمي في الجامعة.

و قد خلصت هذه الدراسة إلى مجموعة من النتائج أهمها:

الحاجة الماسة إلى بناء مستودع معطيات مترابط ومتكامل ونقي وخالي من الأخطاء للجامعة، إضافة إلى نتائج أخرى هامة متعلقة بسجلات الطلاب مثل علاقة معدلات الثانوية العامة وفترة الانقطاع بعد الثانوية باتجاهات الطلاب الأكاديمية ومستوى أدائهم خلال دارستهم الجامعية وهو ما يحتم على الجامعة إعادة النظر في سياسة القبول والتسجيل، أضف إلى ذلك علاقة كثير من المواد الدارسية بتسرب الطلاب وانقطاعهم عن الدارسة، إما لصعوبة المفردات أو لخلل في الخطط الدارسية والمناهج.

٢-٢ الدراسات الأجنبية:

١-٢-٢ دراسة (U.F.Eze, C.J.Onwuegbuchulam, S.Diala, 2017) :

عنوان الدراسة: تطبيقات التنقيب عن المعطيات في صناعة الاتصالات.

طبقت هذه الورقة نموذج استخراج المعطيات في قسم المبيعات والتسويق في صناعة الاتصالات (TI) في نيجيريا. الدافع وراء هذه الورقة هو نتيجة التحديات التنافسية التي تواجه معظم أقسام المبيعات والتسويق في منظمة الشفافية الدولية على مستوى العالم مثل عدم القدرة على الحصول على رؤية دقيقة للمعطيات المستهدفة ، وعدم القدرة على ترجمة وصياغة سؤال العمل بشكل صحيح، ومشكلة معالجة جودة المعطيات. الهدف من هذا العمل البحثي هو تطوير وتنفيذ نموذج يمكن استخدامه للاحتفاظ بالعملاء الحاليين، وجذب عملاء جدد، وإدارة وتخصيص الموارد والسلع والخدمات بشكل فعال في منظمة الشفافية الدولية. كانت تقنيات التنقيب عن المعطيات المستخدمة هي التصنيف والارتباط والاكتشاف المتسلسل والتصور والتنبؤ، الأدوات المستخدمة لتنفيذ النموذج هي PHP و JavaScript و CSS و HTML. مقدمو خدمات الاتصالات (TSP) هم شبكات الهاتف المحمول (MTN) و (GLO) GlobaCom و Airtel وأسواق الاتصالات الناشئة (EMTs) المعروفة أيضًا باسم اتصالات. تم النظر في ثلاثة منتجات عن قسم المبيعات والتسويق في TI مثل Airtime وإعادة الشحن الإلكتروني (up e-top) ومبيعات بطاقة SIM. تتراوح معطيات التدريب المستخدمة في التحليل الاستكشافي النموذجي من ٢٠٠٨ إلى ٢٠١٥ (ثماني سنوات) وتم جمعها من سجلات المبيعات التاريخية لفرق الطوارئ الطبية. تم تنظيف المعطيات وتحويلها. تم تحقيق النظام المعزز من خلال تنفيذ النموذج الذي أثبت أنه أكثر كفاءة من النظام الحالي. كان النموذج الذي تم تنفيذه قادراً على استخراج المعلومات ذات الصلة من قاعدة معطيات TI وجعل توقعات المبيعات للسنة اللاحقة. لذلك ، يوصى باستخدام النظام بواسطة TI لتحسين إنتاجيتها.

في ختام هذه الورقة:

في هذا العمل، تم تطوير نموذج استخراج المعطيات وتنفيذه لتعزيز عمليات أقسام المبيعات والتسويق في TSPs. هذا النموذج قوي جداً ومرن وقابل للفهم ولديه دقة عالية نظراً للحجم الكبير من السجل، مما يجعل تحليل التنقيب عن المعطيات الممل والعمق مهمة بسيطة وسريعة. يسمح النموذج للمستخدمين المصرح لهم بالحصول على رؤية سريعة ودقيقة ورؤى تنبؤية

لمجموعات المعطيات التنظيمية الكبيرة والمعقدة والكشف عن العلاقات والاتجاهات المخفية في المعطيات الجغرافية المكانية. يوفر تطبيق هذا النموذج مجموعة كبيرة من الرسوم البيانية والتقنيات والجداول لوصف سهل للعلاقات في اكتساب المعطيات والمعرفة والتي عالجت الاحتياجات الحالية في تحليل معطيات قسم المبيعات والتسويق في TSP. لذلك، بناءً على الفائدة المذكورة أعلاه لهذا النظام، يوصى باعتماد النظام الجديد بواسطة TSP في نيجيريا لأن الفائدة التي سيتم تحقيقها ستكون رائعة.

٢-٢-٢ دراسة (Lidong Wang, Guanghui Wang, 2015) :

عنوان الدراسة: تطبيقات التنقيب في المعطيات على المعطيات الكبيرة.

التنقيب عن المعطيات هو تقنية لاكتشاف الأنماط المثيرة للاهتمام وكذلك النماذج الوصفية والمفهومة من المعطيات واسعة النطاق. يمكن استخدام التنقيب في المعطيات للعثور على ارتباطات أو أنماط بين عشرات المجالات في قاعدة معطيات علاقاتية كبيرة. التنقيب عن المعطيات هو أيضاً عملية اكتشاف أو العثور على بعض أشكال المعطيات الجديدة والصحيحة والمفهومة والمفيدة. يعد استخراج المعطيات السحابية (CDM) عملية شاقة للغاية تتطلب بنية تحتية خاصة تعتمد على تطبيق تقنيات التخزين الجديدة والمعالجة. المعطيات الكبيرة Hadoop هي أحدث ضجة في مجال معالجة المعطيات. من خلال دمج التحليل المتعمق للمعطيات (استخراج المعطيات) والحوسبة السحابية، ستصبح الحلول التي تصل إلى خدمات استخراج المعطيات في كل مرة وفي كل مكان ومن مختلف المنصات والأجهزة الممكنة. المشكلة الحقيقية الحالية هي أن معظم أساليب التنقيب عن المعطيات لا تعمل بشكل جيد مع المعطيات الكبيرة. هناك الكثير من التحديات عند تطبيق أساليب التنقيب عن المعطيات على تحليلات المعطيات الكبيرة. الهدف من هذه الورقة هو تحديد طرق التنقيب عن المعطيات التي يمكن استخدامها في المعطيات الكبيرة وتقديم التحسينات أو المستجدات لهذه الأساليب من خلال تقديم التقدم التكنولوجي في استخراج المعطيات باستخدام المعطيات الكبيرة. تقدم هذه الورقة التنقيب عن المعطيات، واستخراج المعطيات باستخدام المعطيات الكبيرة، والتحديات والتقدم التكنولوجي في استخراج المعطيات باستخدام المعطيات الكبيرة. تشير التحديات الواردة في هذه الورقة جزئياً

إلى الفجوة / المشكلة من العمل البحثي السابق بالإضافة إلى بعض الأعمال المستقبلية. لذلك، ستعرض هذه الورقة بعض القيمة المهمة لتطبيقات التنقيب عن المعطيات في المعطيات الكبيرة.

في ختام هذه الورقة:

يمكن استخدام التنقيب عن المعطيات لاكتشاف المعرفة المخفية وغير المعروفة ولكنها مفيدة من المعطيات الكبيرة والغامضة والصاخبة وغير الكاملة والعشوائية. لا يمكن استخدام مصنفات KNN لتطبيقات المعطيات الكبيرة. يمكن تطبيق VFDT على تدفق المعطيات، ولكن لديها بعض القيود لتطبيق المعطيات الكبيرة لأن مقاييس الجودة مثل اكتساب المعلومات لتقسيم السمات يتم تقييمها عبر مجموعات المعطيات الفرعية. تتطلب تحليلات المعطيات الكبيرة إجراء التنقيب الموزع لتدفقات المعطيات في الوقت الفعلي. هناك حاجة إلى الكثير من البحث في التحليل العملي والنظري لتوفير طرق جديدة لاستخراج المعطيات الموزعة مع تدفقات المعطيات الكبيرة. تحديات خوارزميات التنقيب عن المعطيات الكبيرة هي: التعلم المحلي ودمج النماذج لمصادر المعلومات المتعددة؛ التنقيب عن معطيات متفرقة وغير مؤكدة وغير كاملة؛ والتنقيب في المعطيات المعقدة والديناميكية. تم إحراز بعض التقدم التكنولوجي مثل طريقة التصنيف للمعطيات الكبيرة ذات السمات الفئوية والرقمية ، والتعلم المختلط غير المتجانس، وخوارزمية اختيار الميزات عبر الإنترنت (OFS)، وما إلى ذلك. أن تكون موضوعات بحثية أخرى.

٣-٢-٢ دراسة (Olga Kurasova, Virginijus Marcinkevi˘cius, Viktor Medvedev, Aurimas Rape˘cka, and Pavel Stefanovic˘, 2014)

عنوان الدراسة: استراتيجيات عنقدة المعطيات الكبيرة.

هناك العديد من تطبيقات المعطيات الكبيرة: الأعمال، والتكنولوجيا، والاتصالات، والطب، والرعاية الصحية، والخدمات، والمعلوماتية الحيوية (علم الوراثة)، والعلوم، والتجارة الإلكترونية، والشؤون المالية، والإنترنت (البحث عن المعلومات، والشبكات الاجتماعية)، إلخ. بعض مصادر المعطيات الكبيرة جديدة بالفعل. يمكن جمع المعطيات الكبيرة ليس فقط من أجهزة الكمبيوتر، ولكن أيضاً من مليارات الهواتف المحمولة ومنشورات الوسائط الاجتماعية وأجهزة الاستشعار المختلفة المثبتة في السيارات وعدادات المرافق والشحن والعديد من المصادر الأخرى. في كثير من الحالات، يتم إنشاء المعطيات بشكل أسرع مما يمكن تهيئتها وتحليلها. يمكن أن تتضمن

المعطيات الكبيرة معطيات منظمة وغير منظمة. المعطيات غير المهيكلة هي المعطيات التي إما لا تحتوي على نموذج معطيات محدد مسبقاً أو غير منظمة بطريقة محددة مسبقاً. المعطيات المنظمة بسيطة نسبياً وسهلة التحليل، لأن المعطيات عادة ما تكون موجودة في قواعد المعطيات في شكل أعمدة وصفوف. التحدي الذي يواجه العلماء هو تطوير أدوات لتحويل المعطيات غير المهيكلة إلى معطيات منظمة. غالباً ما تتكون مجموعة المعطيات المنظمة X من عناصر المعطيات X1، X2، ...، Xm الموصوفة بالميزات x1، x2، ...، xn ، حيث m هو عدد العناصر، n هو عدد الميزات. لذا ، X = {X1، X2، ...، Xm} = {xij، i = 1، ...، m، j = 1، ...، n}، حيث xij هي قيمة الخاصية jth لـ الكائن ith. في حالة المعطيات الكبيرة، تكون m و n كبيرة بما يكفي. إذا كان عدد الميزات n مرتفعاً، فإن المعطيات تسمى المعطيات عالية الأبعاد. يعد تجميع المعطيات عالية الأبعاد مفيداً في حل تقليل الأبعاد بالإضافة إلى مشاكل التصور [٣] ، [٤]. تجلب المعطيات الكبيرة تحديات جديدة للتنقيب عن المعطيات لأنه يجب أخذ الكميات الكبيرة والأصناف المختلفة في الاعتبار. الطرق والأدوات الشائعة لمعالجة المعطيات وتحليلها غير قادرة على إدارة مثل هذه الكميات من المعطيات، حتى لو تم استخدام مجموعات الكمبيوتر القوية. لتحليل المعطيات الكبيرة، تم تطوير العديد من خوارزميات وتقنيات التنقيب عن المعطيات الجديدة والتعلم الآلي. لذلك، لا تنتج المعطيات الكبيرة أنواعاً جديدة من المعطيات وآليات التخزين فحسب، بل تنتج أيضاً طرقاً جديدة للتحليل. عند التعامل مع المعطيات الكبيرة، تعد مشكلة تجميع المعطيات من أهم المشكلات. غالباً ما تتكون مجموعات المعطيات، خاصة مجموعات المعطيات الكبيرة، من بعض المجموعات ومن الضروري العثور على المجموعات. تم تطبيق طرق التجميع على العديد من المشكلات المهمة [٥]، على سبيل المثال، لاكتشاف اتجاهات الرعاية الصحية في سجلات المرضى، والقضاء على الإدخالات المكررة في قوائم العناوين، وتحديد فئات جديدة من النجوم في البيانات الفلكية، وتقسيم المعطيات إلى مجموعات ذات مغزى ومفيدة لتجميع ملايين المستندات أو صفحات الويب. لمعالجة هذه التطبيقات والعديد من التطبيقات الأخرى، تم تطوير مجموعة متنوعة من خوارزميات التجميع. توجد بعض القيود في طرق التجميع الحالية، تتطلب معظم الخوارزميات مسح مجموعة المعطيات عدة مرات، وبالتالي فهي غير مناسبة لتجميع المعطيات الكبيرة. هناك الكثير من التطبيقات التي تحتاج إلى استكشاف مجموعات معطيات كبيرة جداً، ولكنها كبيرة جداً بحيث لا يمكن معالجتها بواسطة طرق التجميع التقليدية. الهدف من هذه الورقة هو استعراض الأساليب

والتقنيات المستخدمة من أجل تجميع المعطيات الكبيرة ووصف استراتيجيات تحليل المعطيات الكبيرة.

في ختام هذه الورقة:

يتم تحليل تحديات ومشاكل تجميع المعطيات الكبيرة في الورقة. تمت مناقشة طرق وتقنيات التجميع. عادةً ما تستخدم أنظمة التنقيب عن المعطيات المعروفة قوة وموارد جهاز كمبيوتر شخصي واحد فقط. في الوقت الحاضر، تولّد الأجهزة الجديدة ووسائل التواصل الاجتماعي والمصادر الأخرى معطيات بأحجام كبيرة. المزيد من التقنيات المبتكرة التي ستكون ضرورية لتحليل المعطيات الكبيرة. في هذه الورقة، تم تقديم استراتيجيات تجميع المعطيات الكبيرة. يعتمد اختيار الاستراتيجية على حجم المعطيات التي يتم تحليلها. عندما نتعامل مع مجموعة كبيرة من المعطيات، عادة ما يتم استخدام أنظمة التنقيب عن المعطيات المعروفة. تتطلب المشاكل المعقدة لتحليل المعطيات استخدام أنظمة وتقنيات قائمة على الحوسبة المتوازية والموزعة. المعطيات الكبيرة تبدأ في تطوير تقنيات جديدة. تعد التقنيات والمكتبات المستندة إلى Hadoop هي الحلول الأكثر شيوعاً لتحليل المعطيات الكبيرة وتجميعها.

الفصل الثاني

الإطار النظري

٣- تمهيد:

مكّن التقدم السريع في تكنولوجيا جمع المعطيات وتخزينها المؤسسات من تجميع كميات هائلة من المعطيات. ومع ذلك، فقد ثبت أن استخراج المعلومات المفيدة أمر صعب للغاية. في كثير من الأحيان ، لا يمكن استخدام أدوات وتقنيات تحليل المعطيات التقليدية بسبب الحجم الهائل لمجموعة المعطيات. في بعض الأحيان، تعني الطبيعة غير التقليدية للمعطيات أنه لا يمكن تطبيق الأساليب التقليدية حتى لو كانت مجموعة المعطيات صغيرة نسبياً. في حالات أخرى، لا يمكن معالجة الأسئلة التي تحتاج إلى إجابة باستخدام تقنيات تحليل المعطيات الحالية، وبالتالي، يجب تطوير طرق جديدة.

من هنا ظهر ما يسمى باستخراج المعطيات (Data Mining) كتقنية تهدف إلى استنتاج المعرفة من كميات هائلة من المعطيات، تعتمد على الخوارزميات الرياضية والتي تعتبر أساس التنقيب عن المعطيات وهي مستمدة من العديد من العلوم مثل علم الإحصاء والرياضيات والمنطق وعلم التعلم، والذكاء الاصطناعي والنظم الخبيرة، وعلم التعرف على الأنماط، وعلم الآلة، والتنقيب عن المعطيات وغيرها من العلوم والتي تعتبر من العلوم الذكية وغير التقليدية.

١-٣ التنقيب في المعطيات (Data Mining):

١-١-٣ مفهوم التنقيب في المعطيات:

هناك الكثير من التعريفات لمفهوم تنقيب المعطيات، وقد تم اختيار التعريفات التالية:

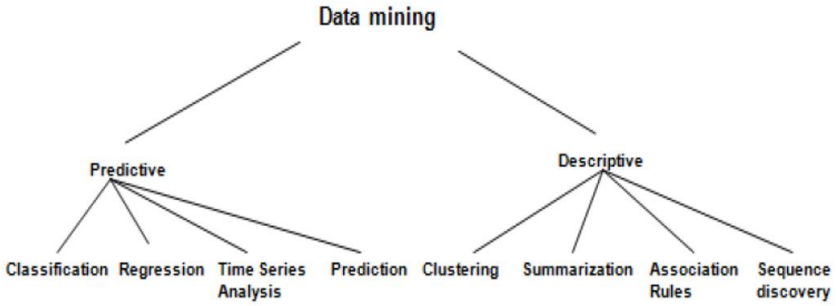
- التنقيب عن المعطيات عبارة عن تقنية تمزج بين طرق تحليل المعطيات التقليدية وخوارزميات معقدة لمعالجة كميات كبيرة من المعطيات.
- التنقيب عن المعطيات هو عملية الاكتشاف التلقائي للمعلومات المفيدة في مستودعات المعطيات الكبيرة.[٢]
- هي النشاط الذي يقوم باستخراج المعلومات المتواجدة في كميات كبيرة من المعطيات ، بهدف البحث عن أنماط معرفية واكتشاف الحقائق الخفية الواردة في قواعد المعطيات.[٣]
- تحليل المعطيات المتواجدة في قواعد المعطيات باستخدام الأدوات التي تبحث عن الاتجاهات أو المعطيات التي لا معنى لها، واستخراج معلومات ضمنية، لم تكن معروفة سابقاً، ويمكن أن تكون مفيدة.
- هي عملية بحث محوسب ويدوي عن معرفة من المعطيات دون فرضيات مسبقة عما يمكن أن تكون هذه المعرفة.
- عملية تحليل كمية معطيات (عادة ما تكون كمية كبيرة)، لإيجاد علاقة منطقية تلخص المعطيات بطريقة جديدة تكون مفهومة ومفيدة لصاحب المعطيات.[٤]

٢-١-٣ أنواع التنقيب في المعطيات:

- التنقيب الاستشرافي: ينتج عنه نموذج عن النظام الذي تصفه المعطيات المستخدمة في التنقيب.
- التنقيب الوصفي: ينتج عنه معلومات جديدة بناءً على المعلومات الموجودة داخل المعطيات المستخدمة في عملية التنقيب.[٥]

---

[2] (Tan, Stein Bach & Kumar, 2006)
[3] (palace, 1996)
[4] (الضاهر,2014)
[5] (محمود و فتوح, 2014)



الشكل (٢): أنواع التنقيب في المعطيات

٣-١-٣ أهمية التنقيب في المعطيات:

للتنقيب في المعطيات أهمية كبيرة في الكثير من التطبيقات الحقيقة حول العالم وأهمها القدرة على التعامل مع المشاكل المعقدة، والقدرة على اكتشاف المعلومات المثيرة للاهتمام والغير متوقعة، واستخراج العديد من المعطيات المخفية، والقدرة على التعلم الذاتي، وإمكانية استخدام التجارب والأخطاء من الماضي لتحسين نوعية النماذج تلقائياً، والتعرف على مسارات المعطيات المخفية وهو ما يؤهله لتقديم العون والمساعدة في بناء التنبؤات المستقبلية، واستكشاف السلوك، والاتجاهات مما يسمح لاتخاذ القرارات الصحيحة وفي الوقت المناسب.[٦]

٣-١-٤ أهداف التنقيب في المعطيات:

١- إن التنقيب في قواعد المعطيات يهدف إلى انتزاع واستخلاص أنماط مفيدة، وهي تكنولوجيا حديثة، أصبحت مهمة في ظل التطور السريع كانتشار استخدام قواعد المعطيات.

٢- استخدامها يوفر للمؤسسات وأجهزة الأمن في جميع المجالات القدرة علي استكشاف، والتركيز على أهم المعلومات في قواعد المعطيات.

٣- تركز تقنيات التنقيب على بناء التنبؤات المستقبلية واستكشاف السلوك والاتجاهات، مما يسمح بتقدير القرارات الصحيحة واتخاذها في الوقت المناسب.

٤- تجنب تقنيات التنقيب علي العديد من الأسئلة، وفي وقت قياسي، بخاصة تلك النوعية من الأسئلة التي يصعب الإجابة عليها، إن لم يكن مستحيلاً، باستخدام تقنيات الإحصاء الكلاسيكية، والتي كانت إن وجدت فإنها تستغرق وقتا طويلاً والعديد من الإجراءات.[٧]

---

[6] (Han, Kamper & Pei, 2012)
[7] (محمود و فتوح, 2014)

٥-١-٣ مراحل عملية التنقيب في المعطيات:

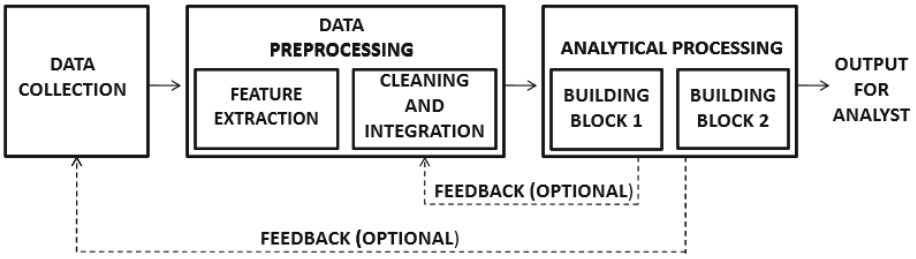
إن عملية استخراج المعطيات عبارة عن خط أنابيب يحتوي على العديد من المراحل مثل تنظيف المعطيات واستخراج الميزات والتصميم الحسابي. يحتوي تدفق العمل الخاص بتطبيق نموذجي للتنقيب عن المعطيات على المراحل التالية:

١- جمع المعطيات: قد يتطلب جمع المعطيات استخدام أجهزة متخصصة. في حين أن هذه المرحلة خاصة بالتطبيق إلى حد كبير وغالباً ما تكون خارج نطاق محلل التنقيب عن المعطيات، إلا أنها مهمة للغاية لأن الخيارات الجيدة في هذه المرحلة قد تؤثر بشكل كبير على عملية استخراج المعطيات. بعد مرحلة التجميع، غالباً ما يتم تخزين المعطيات في قاعدة المعطيات، أو بشكل عام في مستودع معطيات للمعالجة.

٢- استخراج الميزات وتنقية المعطيات: عندما يتم جمع المعطيات، فإنها غالباً لا تكون في شكل مناسب للمعالجة. على سبيل المثال، قد يتم تشفير المعطيات في سجلات معقدة أو مستندات حرة الشكل. في كثير من الحالات، قد يتم خلط أنواع مختلفة من المعطيات بشكل تعسفي معاً في مستند حر الشكل. لجعل المعطيات مناسبة للمعالجة، من الضروري تحويلها إلى تنسيق مناسب لاستخراج المعطيات، مثل تنسيق متعدد الأبعاد أو السلاسل الزمنية أو شبه المركب. التنسيق متعدد الأبعاد هو الأكثر شيوعاً، حيث تتوافق مجالات مختلفة من المعطيات مع الخصائص المقاسة المختلفة التي يشار إليها باسم السمات والصفات والأبعاد. من المهم استخراج الميزات ذات الصلة لعملية التنقيب. غالباً ما يتم تنفيذ مرحلة استخراج الميزات بالتوازي مع تنظيف المعطيات، حيث يتم إما تقدير أو تصحيح الأجزاء المفقودة والخاطئة من المعطيات. في كثير من الحالات، قد يتم استخراج المعطيات من مصادر متعددة وتحتاج إلى دمجها في تنسيق موحد للمعالجة. النتيجة النهائية لهذا الإجراء هي مجموعة معطيات منظمة بشكل جيّد، والتي يمكن استخدامها بشكل فعال. بعد مرحلة استخراج الميزة، يمكن تخزين المعطيات مرة أخرى في قاعدة معطيات للمعالجة.

٣- المعالجة التحليلية والخوارزميات: الجزء النهائي من عملية التنقيب هو تصميم طرق تحليلية فعالة من المعطيات المعالجة. يعتمد هذا الجزء من تصميم الخوارزمية على مهارة المحلل.

[8] (Aggarwal, 2015)



الشكل (٣) مراحل عملية التنقيب في المعطيات

٦-١-٣ مرحلة المعالجة المسبقة للمعطيات:

ربما تكون مرحلة المعالجة المسبقة للمعطيات هي المرحلة الأكثر أهمية في عملية التنقيب عن المعطيات. ومع ذلك، نادراً ما يتم استكشافها بالقدر الذي تستحقه لأن معظم التركيز ينصب على الجوانب التحليلية للتنقيب في المعطيات. تبدأ هذه المرحلة بعد جمع المعطيات ، وتتكوّن من الخطوات التالية:

١- استخراج الميزة: قد يواجه المحلل أحجاماً هائلة من المستندات المسحوبة أو سجلات النظام أو المعاملات التجارية مع القليل من الإرشادات حول كيفية تحويل هذه المعطيات الأولية إلى ميزات قاعدة معطيات مفيدة للمعالجة. تعتمد هذه المرحلة بشكل كبير على المحلل ليكون قادراً على استخلاص الميزات الأكثر صلة بتطبيق معين.

٢- تنظيف المعطيات: قد تحتوي المعطيات المستخرجة على إدخالات خاطئة أو مفقودة. لذلك، قد يلزم حذف بعض السجلات، أو قد يلزم تقدير الإدخالات المفقودة وقد يلزم إزالة التناقضات.

٣- اختيار الخاصية وتحويلها: عندما تكون المعطيات ذات أبعاد عالية جداً، فإن العديد من خوارزميات التنقيب عن المعطيات لا تعمل بشكل فعال. علاوة على ذلك، فإن العديد من الميزات عالية الأبعاد صاخبة وقد تضيف أخطاء إلى عملية استخراج المعطيات. لذلك، يتم استخدام مجموعة متنوعة من الأساليب إما لإزالة الميزات غير ذات الصلة أو تحويل المجموعة الحالية من الميزات إلى مساحة معطيات جديدة أكثر قابلية للتحليل. ومن الجوانب الأخرى ذات الصلة تحويل المعطيات، حيث يمكن تحويل مجموعة معطيات بمجموعة معينة من السمات إلى مجموعة معطيات بمجموعة أخرى من السمات من نفس النوع أو من نوع مختلف.

[9] (Aggarwal, 2015)

٧-١-٣ أنواع مجموعات المعطيات:

هناك العديد من أنواع مجموعات المعطيات، ومع تطور مجال التنقيب عن المعطيات ونضجه، أصبح هناك مجموعة أكبر من المعطيات متاحة للتحليل. في هذا القسم، نصف بعض الأنواع الأكثر شيوعًا، للراحة. تم تجميع أنواع مجموعات المعطيات في ثلاث مجموعات: معطيات السجل، والمعطيات المستندة إلى الرسوم البيانية ، والمعطيات المرتبة. لا تغطي هذه الفئات جميع الاحتمالات ومن المؤكد أن التجمعات الأخرى ممكنة.

❖ الخصائص العامة لمجموعات المعطيات:

قبل تقديم تفاصيل عن أنواع معينة من مجموعات المعطيات، نناقش ثلاث خصائص تنطبق على العديد من مجموعات المعطيات ولها تأثير كبير على تقنيات التنقيب في المعطيات المستخدمة: الأبعاد والتباين والدقة.

١- الأبعاد:

إن أبعاد مجموعة المعطيات هي عدد السمات التي تمتلكها الكائنات في مجموعة المعطيات. تميل المعطيات ذات العدد الصغير من الأبعاد إلى أن تكون مختلفة نوعياً عن المعطيات المتوسطة أو عالية الأبعاد. في الواقع، يشار أحياناً إلى الصعوبات المرتبطة بتحليل المعطيات عالية الأبعاد على أنها لعنة الأبعاد. وبسبب هذا، فإن الدافع المهم في المعالجة المسبقة للمعطيات هو تقليل الأبعاد.

٢- التباين:

بالنسبة لبعض مجموعات المعطيات، مثل تلك التي تحتوي على سمات غير متماثلة، فإن معظم سمات الكائن لها قيم صفرية؛ في كثير من الحالات يكون أقل من ١٪ من الإدخالات غير صفرية. من الناحية العملية، يعد التباين ميزة لأنه عادةً ما يلزم تخزين القيم غير الصفرية فقط ومعالجتها. ينتج عن هذا توفير كبير فيما يتعلق بوقت الحساب والتخزين. علاوة على ذلك، تعمل بعض خوارزميات التنقيب عن المعطيات بشكل جيّد مع المعطيات المتفرقة فقط.

٣- الدقة:

الممكن في كثير من الأحيان الحصول على معطيات على مستويات مختلفة من الدقة، وغالباً ما تختلف خصائص المعطيات باختلاف درجات الدقة. تعتمد الأنماط في المعطيات أيضاً على

مستوى الدقة. إذا كانت الدقة جيدة جداً، فقد لا يكون النمط مرئياً أو قد يكون مدفوناً في الضوضاء؛ إذا كانت الدقة رديئة جداً، فقد يختفي النمط.

١-٧-١-٣ معطيات السجل (record data) :

- ❖ معطيات السجل: الكثير من أعمال التنقيب عن المعطيات تفترض أن مجموعة المعطيات عبارة عن مجموعة من السجلات (كائنات المعطيات)، يتكون كل منها من مجموعة ثابتة من حقول المعطيات (السمات). بالنسبة إلى الشكل الأساسي لمعطيات السجل، لا توجد علاقة صريحة بين السجلات أو حقول المعطيات، وكل سجل (كائن) له نفس مجموعة السمات. عادةً ما يتم تخزين معطيات التسجيل إما في ملفات ثابتة أو في قواعد معطيات علاقاتية. تعد قواعد المعطيات العلاقاتية بالتأكيد أكثر من مجرد مجموعة من السجلات، ولكن التنقيب في المعطيات لا يستخدم في كثير من الأحيان أياً من المعلومات الإضافية المتاحة في قاعدة المعطيات العلاقاتية. بدلاً من ذلك، تعمل قاعدة المعطيات كمكان مناسب للعثور على السجلات.
- ❖ معطيات المعاملات التجارية أو معطيات سلة السوق:

هي نوع خاص من معطيات السجل، حيث يتضمن كل سجل (معاملة) مجموعة من العناصر. معطيات المعاملة عبارة عن مجموعة من مجموعات العناصر، ولكن يمكن عرضها كمجموعة من السجلات التي تعد حقولها سمات غير متماثلة. غالباً ما تكون السمات ثنائية، تشير إلى ما إذا كان قد تم شراء عنصر أم لا، ولكن بشكل عام، يمكن أن تكون السمات منفصلة أو مستمرة، مثل عدد العناصر المشتراة أو المبلغ الذي تم إنفاقه على هذه العناصر.

- ❖ مصفوفة المعطيات:

إذا كانت كائنات المعطيات في مجموعة من المعطيات تحتوي جميعها على نفس المجموعة الثابتة من السمات الرقمية، فيمكن اعتبار كائنات المعطيات كنقاط (متجهات) في مساحة متعددة الأبعاد، حيث يمثل كل بُعد سمة مميزة تصف الكائن. يمكن تفسير مجموعة من كائنات المعطيات هذه على أنها مصفوفة n'L بواسطةn، حيث توجد صفوف rn، واحد لكل كائن، و n عمود، واحد لكل سمة. (التمثيل الذي يحتوي على كائنات معطيات كأعمدة

[1] (Tan, Steinbach & Kumar, 2006)

والسمات كصفوف جيد أيضاً.) تسمى هذه المصفوفة مصفوفة المعطيات أو مصفوفة النمط. مصفوفة المعطيات هي تباين في معطيات السجل، ولكن نظراً لأنها تتكون من سمات رقمية، يمكن تطبيق عملية المصفوفة القياسية لتحويل المعطيات ومعالجتها. لذلك، فإن مصفوفة المعطيات هي تنسيق المعطيات القياسي لمعظم المعطيات الإحصائية.

❖ مصفوفة المعطيات المتفرقة:

هي حالة خاصة لمصفوفة المعطيات تكون فيها السمات من نفس النوع وغير متماثلة؛ على سبيل المثال، القيم غير الصفرية فقط مهمة. معطيات المعاملة هي مثال لمصفوفة معطيات متفرقة تحتوي على (٠، ١) مدخلات فقط. مثال شائع آخر هو معطيات المستند، على وجه الخصوص، إذا تم تجاهل ترتيب المصطلحات (الكلمات) في المستند، فيمكن تمثيل المستند كمتجه للمصطلح، حيث يكون كل مصطلح مكوناً (سمة) للمتجه وقيمة كل مكون هي الرقم من المرات التي يظهر فيها المصطلح المقابل في المستند. غالباً ما يُطلق على هذا التمثيل لمجموعة من المستندات مصفوفة مصطلح المستند.

Tid	Refund	Marital Status	Taxable Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(a) Record data.

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Beer, Soda, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Soda, Diaper, Milk

(b) Transaction data.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

(c) Data matrix.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

(d) Document-term matrix.

الشكل (٤) أمثلة على أنواع مختلفة من معطيات السجل

[1] (Tan, Steinbach & Kumar, 2006)[1]

٢-٧-١-٣ المعطيات المستندة إلى الرسوم البيانية (graph-based data):

يمكن أن يكون الرسم البياني أحياناً تمثيلاً مناسباً وقوياً للمعطيات. نحن نعتبر حالتين محددتين:

١- يلتقط الرسم البياني العلاقات بين كائنات المعطيات.

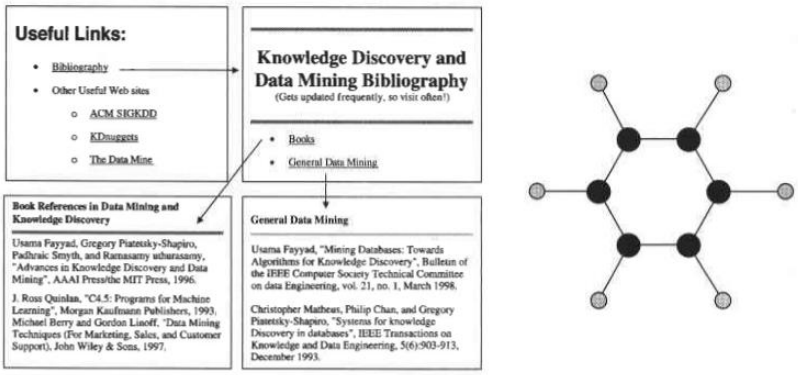
٢- يتم تمثيل كائنات المعطيات نفسها كرسوم بيانية.

❖ المعطيات ذات العلاقات بين الكائنات:

تنقل العلاقات بين الكائنات معلومات مهمة بشكل متكرر. في مثل هذه الحالات، غالباً ما يتم تمثيل المعطيات كرسم بياني. على وجه الخصوص، يتم تعيين كائنات المعطيات إلى عقد الرسم البياني، بينما يتم التقاط العلاقات بين الكائنات بواسطة الروابط بين الكائنات وخصائص الارتباط، مثل الاتجاه والوزن.

❖ المعطيات التي تحتوي على كائنات:

تمثل رسوماً بيانية إذا كانت الكائنات تحتوي على بنية، أي أنه إذا كانت الكائنات تحتوي على كائنات فرعية لها علاقات، فسيتم تمثيل هذه الكائنات غالباً على هيئة رسوم بيانية.



(a) Linked Web pages. (b) Benzene molecule.

الشكل (٥) أمثلة على أنواع المعطيات المستندة إلى رسوم بيانية

[1] (Tan, Steinbach & Kumar, 2006)[2]

٣-٧-١-٣ المعطيات المرتبة:

بالنسبة لبعض أنواع المعطيات، تحتوي السمات على علاقات تتضمن ترتيباً في الزمان أو المكان.

❖ المعطيات المتسلسلة:

يشار إليها أيضاً بالمعطيات المؤقتة، يمكن إعتبارها إمتداداً لمعطيات السجل، حيث يكون لكل سجل وقت مرتبط به. يمكن أيضاً ربط الوقت بكل سمة.

❖ معطيات التسلسل:

تتكوّن من مجموعة معطيات هي سلسلة من الكيانات الفردية، مثل سلسلة من الكلمات أو الحروف. إنها مشابهة تماماً للمعطيات المتسلسلة، باستثناء عدم وجود طوابع زمنية؛ بدلاً من ذلك، هناك مواقف في تسلسل مرتب.

❖ معطيات السلاسل الزمنية:

هي نوع خاص من المعطيات المتسلسلة حيث يكون كل سجل عبارة عن سلسلة زمنية، أي سلسلة من القياسات المأخوذة بمرور الوقت. عند العمل مع المعطيات الزمنية، من المهم النظر في الارتباط التلقائي الزمني؛ على سبيل المثال، إذا كان قياسان قريبان من الوقت، فغالباً ما تكون قيم تلك القياسات متشابهة جداً.

❖ المعطيات المكانية:

تحتوي بعض الكائنات على سمات مكانية، مثل المواضع أو المناطق، بالإضافة إلى أنواع أخرى من السمات. من الأمثلة على المعطيات المكانية معطيات الطقس (هطول الأمطار ودرجة الحرارة والضغط) التي يتم جمعها لمجموعة متنوعة من المواقع الجغرافية. أحد الجوانب المهمة للمعطيات المكانية هو الارتباط الذاتي المكاني؛ على سبيل المثال، تميل الأشياء القريبة جسدياً إلى أن تكون متشابهة بطرق أخرى أيضاً. [٣]

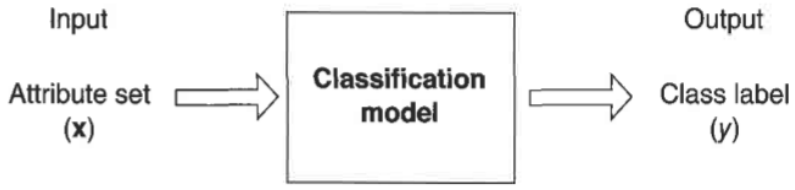
[1] (Tan, Steinbach & Kumar, 2006)[3]

٤ـ تقنيات التنقيب في المعطيات:

٤-١ التصنيف (classification):

التصنيف، هو تقنية للتنبؤ بنتيجة معينة بناءً على مدخلات معينة. تستخرج نماذج تصف بشكل دقيق فئات وتصنيفات المعطيات الهامة، وتتنبأ مثل هذه النماذج، بالتصنيفات الفئوية (المنفصلة و غير المرتبة). تقوم الخوارزمية بتحليل المدخلات وإصدار التنبؤ. تحدد دقة التنبؤ مدى جودة "الخوارزمية".

بيانات الإدخال لمهمة التصنيف عبارة عن مجموعة من السجلات. كل سجل، المعروف أيضاً باسم مثيل أو مثال، يتميز بمجموعة (x,y) حيث x هي مجموعة السمات و y هي سمة خاصة، تم تعيينها على أنها تسمية الفئة (تُعرف أيضًا باسم الفئة أو السمة الهدف).



الشكل (٦) التصنيف كمهمة تعيين سمة الإدخال المحددة x في تصنيف الفئة y

يتكون النهج العام للتصنيف من خطوتين:

١- الخطوة الأولى:

نقوم ببناء نموذج تصنيف بناءً على المعطيات السابقة و تسمى بمرحلة التعلم، و في هذه الخطوة تقوم خوارزميات التصنيف ببناء المصنف. المصنف مبني على مجموعة التدريب (المعطيات السابقة) المكونة من قاعدة معطيات tuples و تصنيفات الفصل المرتبطة (سمة التنبؤ).

٢- الخطوة الثانية:

نحدد ما إذا كانت دقة النموذج مقبولة، وإذا كان الأمر كذلك، فإننا نستخدم النموذج لتصنيف المعطيات الجديدة، في هذه الخطوة يتم استخدام المصنف للتنبؤ بمعطيات غير معروفة، هنا يتم

استخدام معطيات الاختبار لتقدير الدقة قواعد التصنيف. يمكن تطبيق قواعد التصنيف على المعطيات الجديدة، تتم مقارنة هذا المصنف مع المصنف الفعلي.[٤]

١-١-٤ العوامل المؤثرة على دقة التصنيف:

❖ تحويل المعطيات (data transformation):

تحتوي عملية تحويل المعطيات على عمليتين، و ليس بالضرورة استخدامهم معاً، أي أن طبيعة المعطيات هي التي تحدد:

١- التطبيع (normalization):

يتم تحويل المعطيات باستخدام التطبيع، ويشمل التطبيع قياس جميع القيم لسمة معينة لجعلها تقع ضمن نطاق صغير محدد. يستخدم التطبيع عندما يتم استخدام الشبكات العصبية أو الأساليب التي تنطوي على قياسات في خطوة التعلم.

٢- التعميم (generalization):

يمكن أيضاً تحويل المعطيات من خلال تعميمها على المفهوم الأعلى لهذا الغرض يمكننا استخدام التسلسل الهرمي للمفاهيم.

❖ وجود القيم المتطرفة (presence of outliers):

القيم المتطرفة هي نقاط معطيات لا تتوافق مع غالبية المعطيات. القيم المتطرفة هي أيضاً نقاط يصعب تصنيفها بسبب يمكننا أن نقول أنهم لا يملكون خصائص ناقلات ميزة مماثلة مثل غالبية المعطيات. لذا تصنيف القيم المتطرفة مهمة محفوفة بالمخاطر و يمكن ان تؤثر على الدقة بشكل سلبي. طريقة جيدة يمكن استخدامها هنا مقدماً هي إزالة القيم المتطرفة.

❖ إزالة الضوضاء (noise removal):

يمكن تعريف الضوضاء عادةً على أنها خطأ عشوائي أو التباين في متغير مقاس يكون النوعان النموذجيان فيه غير متناسقين قيم الميزات أو الفئات. الضوضاء عادة ما تكون أقلية في مجموعة المعطيات. و يمكن إزالتها باستخدام خوارزميات العنقدة.[٦]

---

[1] (Romero) [4]
[1] (prediction) [5]
[1] (Romero) [6]

❖ تحليل مدى الارتباط (relevance analysis):

قد تحتوي المعطيات في بعض الأحيان على مجموعات قليلة من الصفات التي لا توفر الكثير من المعلومات للمصنف. هذا يعني انه لا تشكل هذه السمات إدخالات مهمة في متجه الميزة. إزالة هذه السمات يمكّن لهذه الصفات تسريع أداء المصنف. يمكن أيضاً القيام بذلك باستخدام (chi-square)، LDA (linear discriminant analysis) تساعد هذه الطرق في تقليل مساحة الميزة بشكل فعّال.[١] [٧]

❖ طرق التقدير الخاطئة (wrong estimation methods):

لا ينبغي أن تكون دقة التصنيف مثالية تقاس في تجربة واحدة. التحقق المتقاطع هو طريقة جيدة جداً لقياس الدقة التي تستخدم نوعاً من الإجراء وتترك الدقة لجميع التكرارات. ولكن الدقة ما تزال تدبير شخصي ولا يمكن تزويدنا بمعلومات كاملة عن أداء المصنف.

٢-١-٤ خوارزميات التصنيف:

١- شجرة القرار (decision tree):

شجرة القرارات وهي رسم تخطيطي، على شكل شجرة متفرعة تستخدم لتحديد مسار العمل أو لتظهر ما هي الاحتمالات الممكنة. يمثل كل فرع من فروع الشجرة قرار محتمل. يتم تنظيم الشجرة لإظهار كيف ولماذا قد يؤدي أحد الخيارت إلى الخيار التالي، مع استخدام الخطوط المتفرعة و التي تشير إلى أنه كل خيار مستقل عن الخيار الآخر، وتعد أشجار القرار أيضاً جزءاً أساسياً من الغابات العشوائية.[١] [٨]

شجرة القرارات هي خوارزمية تعلم الآلة متعددة الوظائف، والتي يمكن أن تؤدي مهام التصنيف والانحدار، وحتى تشمل المهام متعددة المخرجات، و هي خوارزمية قوية جداً و يمكن أن تناسب مجموعات المعطيات المعقدة للغاية.[١] [٩]

مصنف J48 (J48 classifier):

يندرج هذا المصنف ضمن خوارزميات أشجار القرار والتي على اختلاف أنواعها تشابه إلى حد ما خوارزمية التصنيف (Naïve Bayes) من حيث اعتمادها على الاحتمالات الشرطية مع

---

[1] (Prediction) [7]
[1] (Morgan & Techman,1988) [8]
[1] (Han, Kamber & Pei, 2012) [9]

إختلاف رئيسي يكمن في هذه الخوارزمية حيث تقوم بتوليد قواعد (rules) لاستخدامها كجمل شرطية لتحديد السجلات والأحداث الاحتمالية بشكل عبارة شرطية (IF.....THEN).

يستند هذا المصنف إلى هيكلة شجرية مؤلف من عقد رئيسية تدعى الجذر(root) ومجموعة عقد داخلية (nodes) ومجموعة عقد نهائية (terminals)، بحيث يتضمن كل من الجذر والعقد الداخلية قواعد (rules) التي تحدد المسار للفروع المرتبطة بما يسمح في النهاية بالوصول إلى النتيجة النهائية.[٢]

تعتمد هذه الخوارزمية إلى تقسيم مجموعة معطيات التدريب المراد تصنيفها إلى مجالات متقاطعة (mutual exclusive) ذات تسمية أو قيمة أو عملية لتوضيح وشرح المعطيات داخل هذا المجال وذلك بالاعتماد على معيار يستخدم لحساب أو تعين أفضل المعايير لتجزئة هذا المجال من المعطيات التي يتم تدريبها والذي يدعى التابع الإحصائي (information gain) والمعروف بالمعادلة التالية:

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

حيث:

(S) مجموعة معطيات التدريب

(A) مجموعة المعايير

(A) values جميع القيم الممكنة للمعيار A

(Sv) مجموعة جزئية من المعيار S المنتمية للمعيار A ذات القيمة V

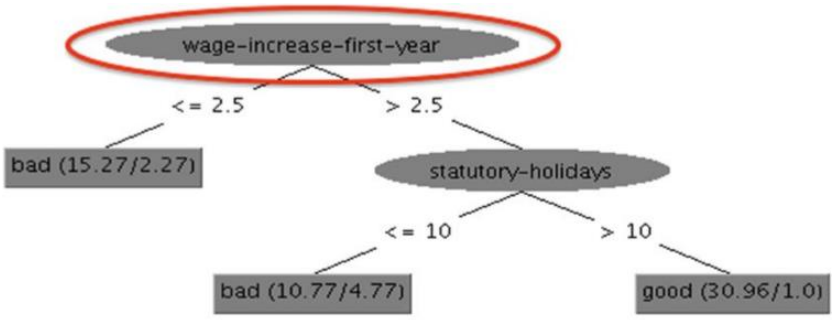
(Entropy) تابع العشوائية و يعبر هذا التابع عن عشوائية المعطيات وتتراوح قيمته بين (0 و 1) ويعبر عنه بالمعادلة التالية:

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

حيث يعبر المتغير (pi) عن احتمالية انتماء مجموعة المعطيات (S) إلى الفئة (i)

[2] (الضاهر, 2014) [0]

تتصف النماذج التي يولّدها هذا النوع من التصنيفات بالدقة العالية والسرعة في بناء النموذج، كما يمكن تطبيقها على المعطيات متعددة الفئات، وبأنها قابلة للتأويل والفهم من خلال تحليل شجرة القرار ومعاينة الرسم البياني المولّد عند بناء النموذج.



الشكل(٧) مثال على شجرة القرار

٢- مصنف بايز الاحتمالي (naïve bayes classifier) :

هي نظرية أساسية في علم الاحتمالات، حيث تحدد الاحتمال الشرطي لحدث ما. هذا الاحتمال الشرطي يعرف بالفرضية، ويتم حساب احتمالية حدوث هذه الفرضية بناءً على معطيات أو دلائل سابقة، باختصار بإمكاننا تعريفها على أنها احتمالية حدوث حدث ما بمعلومية أن حدث آخر قد حدث بالفعل. و الشكل الرياضي للنظرية هو كما يلي:

$$p(H \mid E) = \frac{p(E \mid H)\ p(H)}{p(E)}$$

يتميز مصنف بايز بالسرعة في المعالجة والكفاءة في عمليات التنبؤ. يعتمد هذا الأسلوب على المفهوم الإحصائي لنظرية بايز والذي يحسب احتمالية حدوث نتيجة معينة بتحقق ما هو متاح ومعروف ويسمى ساذج لأنه يعتمد مبدأ Independence Assumption بحيث ينظر للعلاقة بين جميع الخصائص بأنها مستقلّة عن بعضها.[١] [٢]

بمعنى أن النموذج لا يعير اهتماماً للعلاقة بين الخصائص إن وجدت فجميعهم يساهمون في حساب الاحتمال والنتيجة النهائية ستكون رقماً لا يحمل معنى من حيث توضيح اعتماد خاصية

[1] (Han, Kamber & Pei, 2012)

على أخرى أو قيمة للترتيب. كما يتميز نموذج بايز للتصنيف بسهولة البناء والتطوير والقدرة على معالجة المعطيات الكبيرة.[٢] [٢]

٣- خوارزمية الجار الأقرب (K-Nearest Neighbor):

هي إحدى خوارزميات التنقيب في البيانات والتعلم الآلي والتي تعمل بمشرف (موجه) وتعد خوارزمية الجار الأقرب من خوارزميات التصنيف التنبؤية والوصفية ولديها الإمكانيات لتوليد التقديرات المحلية للنقطة، و تكون الطريقة سهلة في أغلب الأحيان عندما يكون حجم المعطيات صغيراً، وتصف المعطيات لمعرفة الجار الأقرب وفعّالة عندما تكون قاعدة المعطيات كبيرة جداً ولها القدرة على التعامل مع البيانات (المشوشة) أي الناقصة، وتعتمد الخوارزمية في عملها على قياس المسافة الإقليدية بين كل نقطة والنقطة الأقرب إليها وعندما تكون البيانات قريبة من بعضها تكون المسافة الاقليدية قليلة جداً بين كل نقطة والنقطة المجاورة لها ولكن كلما تباعدت قيم المعطيات وتبعثرت أصبحت المسافات بين النقاط كبيرة ومنها جاء عنوان الخوارزمية، إذ يشير الحرف K إلى الحالات التي سيتم تصنيفها بناءً على المسافات بينها (أي بين الجيران)، حسب المسافة من المعادلة التالية:[٢] [٣]

$$Euclidean\ Distance = d = \sqrt{\sum_{i=1}^{N} (Xi - Yi)^2}$$

حيث أن:

D: هي المسافة بين أي نقطتين

Xi,Yi نقاط تموضع المعطيات في البحث

ولهذه الخوارزمية محددات مثلما لها مزايا، ومن أهم محدداتها انها تحتاج إلى وقت طويل عند العمل على المعطيات الكبيرة، وكذلك فهي حساسة عند التعامل مع سمات التدريب ذات الخواص غير المترابطة، لأن هذا يضعف التصنيف و لهذا يجب إختيار سمات ذات خواص مترابطة.

[2] (Raschka, 2014) [2]
[2] (Tan & Shi, 2016) [3]

-خطوات عمل الخوارزمية:

حدد عدد الجيران الأقرب و لتكن K

حساب المسافة الإقليدية بين السجل المستكشف و الجار الأقرب من المعادلة الرياضية السابقة.[٤] [٢]

٤-خوارزمية K-Means:

تعتبر خوارزمية K – Means من الخوارزميات غير الخاضعة للإشراف والتي تستخدم لحل مسائل العنقدة/التصنيف على شكل عناقيد، مبدأ عمل هذه الخوارزمية تتبع طريقة بسيطة وسهلة لتصنيف مجموعة معطيات معينة من خلال عدد معين من العناقيد (افترض أن K عدد العناقيد).

تكون مجموعة المعطيات داخل العنقود الواحد متجانسة لكنها غير متجانسة مع مجموعة المعطيات داخل العناقيد الأخرى.

كيف تعمل خوارزمية K – Means على تشكيل العناقيد :

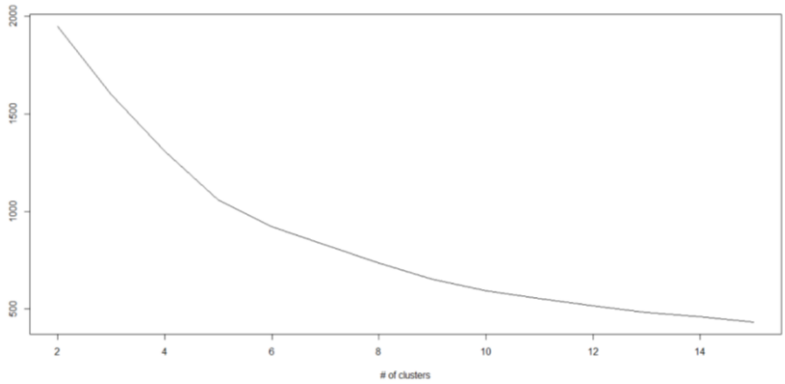
K – Means تختار عدد من النقاط K لكل عنقود(cluster) وتعرف باسم النقاط الوسطى (centroids). تشكل كل نقطة معطيات عنقود به النقاط الوسطى القريبة له، أي K عناقيد (K-clusters). نظراً لظهور نقاط وسطى جديدة، تعمل K – Means على تكرار الخطوتين السابقتين بحيث تبحث عن أقرب مسافة لكل نقطة بيانات من النقاط الوسطى الجديدة وتعمل على ربطها ب K العناقيد الجديدة، تكرّر هذه العملية حتى يحدث التقارب، أي أن النقاط الوسطى لا تتغير.

كيفية تحديد قيمة K في K-means :

لدينا عناقيد (مجموعات) وكل عنقود (مجموعة) لها نقطة وسطى (centroid) خاصة به.
مجموع تربيع الفرق بين النقطة الوسطى ونقاط المعطيات داخل العنقود يشكل مجموع قيمة مربعة لتلك المجموعة أيضاً، عند إضافة مجموع القيم المربعة لجميع العناقيد، يصبح الإجمالي ضمن مجموع القيمة المربعة لحل العنقود.

[2] (Tan & Shi, 2016) [4]

نعلم أنه مع زياده عدد العناقيد، تستمر هذه القيمة في التناقص ولكن إذا قمت برسم النتيجة، فقد ترى أن مجموع المسافة المربعة يتناقص بشكل حاد إلى قيمة معينة من K. ثم ببطء أكثر بعد ذلك، هنا يمكننا إيجاد العدد الأمثل لعدد العناقيد.[٢] [٥]



الشكل (٨) طريقة عمل خوارزمية K-means

الاختبار يتم بشكل يدوي باستخدام مخطط الخسارة مقابل عدد العناقيد للعثور على (k) الأمثل، كما تمت مناقشته فوق : هذه الخوارزمية تكون معتمدة على القيم الأولية للحصول على قيمة صغيرة ل K، يمكنك التخفيف من هذا الاعتماد عن طريق تشغيل k – mean عدة مرات بقيم أولية مختلفة واختيار أفضل نتيجة. مع الزيادة، تحتاج إلى إصدارات متقدمة من K-means لاختيار أفصل قيم للنقاط الوسطى الأولية تسمى (k-means seeding).[٦]

الهدف الرئيسي من خوارزمية K-Means هو تقليل مجموع المسافات بين النقاط والنقط المركزية العنقودية الخاصة بها.

عند عنقدة القيم المتطرفة يجب قص هذه القيم أو إزالتها، لأنه يمكن سحب و إبعاد القيم الوسطى (centroids) بواسطة القيم المتطرفة، أو قد تحصل القيم المتطرفة على مجموعة خاصة بها بدلاً من تجاهلها.

إن خوارزمية K-means سهلة التنفيذ نسبياً، و قابلة للاستخدام مع مجموعات المعطيات الكبيرة كما أنها تعطي نتائج جيدة.

[2] (Aggarwal, 2015)
[2] (Aggarwal, 2015)
[5]
[6]

الفصل الثالث

الإطار العملي

٥- مقدمة:

إن مشكلة البحث تتمحور حول تطبيق تقنيات التنقيب في المعطيات لتصميم عروض للزبائن بحسب الشرائح الزمنية، يهدف هذا الفصل إلى تحليل المعطيات وتجميع الخلايا في مجموعات حسب حجم الاتصالات التي تتم فيها من أجل تصميم عروض للمستخدمين بناءً على الأكثر استخداماً للخدمات، حسب الخلية التي يتصل من خلالها المستخدم.

٥-١ توصيف قاعدة المعطيات:

تتألف قاعدة المعطيات من ستة ملفات بصيغة CSV لشركة اتصالات، وكل ملف يمثل الاتصالات في أحد الأيام، الأيام المتوفرة من تاريخ ٢٠١٣/١١/٢ و حتى ٢٠١٣/١١/٧ ، أنواع الخدمة المدروسة هي: (المكالمة الصوتية، الرسائل النصية، الإنترنت)، كما هو موضح بالشكل التالي:

datetime	CellID	countrycode	smsin	smsout	callin	callout	internet
11/4/2013 0:00	1	0	0.108				
11/4/2013 0:00	1	39	1.0266	0.8069	0.0552	0.2155	50.342
11/4/2013 0:00	2	0	0.1093				
11/4/2013 0:00	2	39	1.043	0.8226	0.0556	0.2195	50.432
11/4/2013 0:00	3	0	0.1106				
11/4/2013 0:00	3	39	1.0604	0.8393	0.056	0.2238	50.5278

الشكل (٩) توضيح لشكل قاعدة المعطيات

يوجد ثمانية أعمدة يعبر كل منها عن:

١- العمود الأول: التاريخ

٢-العمود الثاني: رقم يعبر عن الخلية

٣- العمود الثالث: رقم يعبر عن المقاطعة

٤-العمود الرابع: عدد الرسائل النصية الواردة إلى الخلية

٥- العمود الخامس: عدد الرسائل النصية الصادرة من الخلية

٦- العمود السادس: عدد المكالمات الواردة إلى الخلية

٧- العمود السابع: عدد المكالمات الصادرة من الخلية

٨- العمود الثامن: حجم تراسل الانترنت

بالنسبة للأعمدة الخمسة الأولى التي تعبر عن حجم استخدام الخدمات، تم حساب مجاميعها خلال كل ساعة، فكل سطر يحوي معلومات عن الاتصالات لكل خلية في كل مقاطعة في كل ساعة من ساعات اليوم الواحد.

ومن أجل التخلص من الأرقام الكبيرة تم تقسيم هذه القيم على عشرة مليون والتقريب إلى أقرب أربع أرقام عشرية.

٢-٥ الأداة المستخدمة في التنقيب عن المعطيات:

أداة التنقيب عن المعطيات هي تطبيق برمجي يستخدم لاكتشاف الأنماط والاتجاهات من مجموعات كبيرة من المعطيات وتحويل تلك المعطيات إلى معلومات أكثر دقة. الأداة تساعد على تحديد العلاقات غير المتوقعة بين المعطيات لنمو الأعمال، كما تسمح لك بتحليل المعطيات ومحاكاتها والتخطيط لها والتنبؤ بها باستخدام نظام أساسي واحد.

سنستخدم في هذا البحث أداة Google Colaboratory.

Google Colaboratory : هي خدمة سحابية مجانية، وهي بكلماتها الخاصة، "تتيح لك مشاركة دفاتر Jupyter مع الآخرين دون الحاجة إلى تنزيل أو تثبيت أو تشغيل أي شيء على جهاز الكمبيوتر الخاص بك بخلاف المتصفح".

Jupyter Notebook هو تطبيق ويب مفتوح المصدر يمنحك القدرة على إنشاء ومشاركة المستندات التي تحتوي على كود مباشر ومعادلات وتصورات ونص سردي.

Jupyter Notebook عبارة عن دفتر ملاحظات تفاعلي يحتوي على العديد من التطبيقات. وهو يدعم أكثر من ٤٠ لغة برمجة بما في ذلك Python و R و Julia و Scala وأكثر من ذلك (مما يعني أنه يمكنه تشغيل كود مكتوب بجميع هذه اللغات) بل إنه تم استخدامه للتعلم الآلي.

يمكن لـ Google Colaboratory أيضاً عرض تصورات حيّة، وحفظ نسخة من دفتر ملاحظاتك فيGithub، والسماح لك بتثبيت مكتبات جديدة لاستخدامها في التعليمات البرمجية الموجودة في دفتر ملاحظاتك، وحتى إنشاء نماذج داخل دفتر الملاحظات مباشرةً.

دعم تسريع وقت تشغيلTPI / GPU :

يدعم Google Colaboratory دعم تسريع وقت تشغيلTPU / GPU .

يرمز TPU إلى "Tensor Processing Unit" وهي شريحة مصممة لمكتبة الرياضيات من Google (TensorFlow) والتي تستخدم للتعلم الآلي (مثل الشبكات العصبية).

يرمز GPU إلى وحدة معالجة الرسومات وهي بطاقة يستخدمها جهاز الكمبيوتر الخاص بك للرسومات.

يستخدم تسريع GPU / TPU وحدة معالجة الرسومات (GPU) و وحدة المعالجة المركزية (CPU) في جهاز الكمبيوتر الخاص بك للمهام التي تستهلك قدراً كبيراً من الطاقة، مثل التعلم العميق والتحليلات والتطبيقات الهندسية. يدعم Google Colab استخدام GPU / TPU في التطبيقات الموجودة في دفتر الملاحظات الخاص بك، بحيث يمكنك تشغيل الشبكات العصبية دون مغادرة متصفحك (نظراً لأنك ستستخدم وحدة المعالجة المركزية ووحدة المعالجة المركزية من Google لتشغيل هذه التطبيقات، فهناك حد زمني للمدة يمكنك استخدامها، وهي حوالي ١٢ ساعة) يدعم Google Colab أيضاً الاتصال بوقت تشغيل Jupyter على جهازك المحلي.

يدعم Google Colab أيضاً استخدام Google Drive لتحميل مجموعات المعطيات، وحتى تحميل أو تنزيل النماذج من وإلى أجهزة الكمبيوتر المحمولة التي تستخدمها وجهازك المحلي.

٣-٥ عملية التنقيب عن المعطيات:

سنستعرض في هذا القسم خطوات تطبيق الأداة التي اخترناها (Google Colab) على قاعدة المعطيات المذكورة سابقاً، باستخدام لغة Python و خوارزميات التصنيف (K-means, Apriori, Association-Rules) لتحليل المعطيات و التنقيب فيها من أجل تصميم عروض للمستخدمين.

١-٣-٥ مرحلة التحضير للتنقيب عن المعطيات:

- أولاً تم استيراد مكتبة OS (operating system)

```
import os
os.chdir("d:\\dm\\assoc-rules")
print(os.getcwd())
```

```
d:\dm\assoc-rules
```

- و من ثم تم تنصيب مكتبة Mlxtend (machine learning extensions)

```
!pip install mlxtend
```

```
Requirement already satisfied: mlxtend in c:\programdata\anaconda3\lib\site-packages (0.18.0)
Requirement already satisfied: scipy>=1.2.1 in c:\programdata\anaconda3\lib\site-packages (from mlxtend) (1.5.0)
Requirement already satisfied: numpy>=1.16.2 in c:\programdata\anaconda3\lib\site-packages (from mlxtend) (1.18.5)
Requirement already satisfied: pandas>=0.24.2 in c:\programdata\anaconda3\lib\site-packages (from mlxtend) (1.0.5)
Requirement already satisfied: setuptools in c:\programdata\anaconda3\lib\site-packages (from mlxtend) (49.2.0.post20200714)
Requirement already satisfied: scikit-learn>=0.20.3 in c:\programdata\anaconda3\lib\site-packages (from mlxtend) (0.23.1)
Requirement already satisfied: matplotlib>=3.0.0 in c:\programdata\anaconda3\lib\site-packages (from mlxtend) (3.2.2)
Requirement already satisfied: joblib>=0.13.2 in c:\programdata\anaconda3\lib\site-packages (from mlxtend) (0.16.0)
Requirement already satisfied: python-dateutil>=2.6.1 in c:\programdata\anaconda3\lib\site-packages (from pandas>=0.24.2->mlxte
nd) (2.8.1)
Requirement already satisfied: pytz>=2017.2 in c:\programdata\anaconda3\lib\site-packages (from pandas>=0.24.2->mlxtend) (2020.
1)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\programdata\anaconda3\lib\site-packages (from scikit-learn>=0.20.3->m
lxtend) (2.1.0)
Requirement already satisfied: cycler>=0.10 in c:\programdata\anaconda3\lib\site-packages (from matplotlib>=3.0.0->mlxtend) (0.
10.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\programdata\anaconda3\lib\site-packages (from matplotlib>=3.0.0->mlxten
d) (1.2.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in c:\programdata\anaconda3\lib\site-packages (from mat
plotlib>=3.0.0->mlxtend) (2.4.7)
Requirement already satisfied: six>=1.5 in c:\programdata\anaconda3\lib\site-packages (from python-dateutil>=2.6.1->pandas>=0.2
4.2->mlxtend) (1.15.0)
```

- و بعدها تم استيراد المكاتب التالية:

Numpy للتعامل مع المصفوفات العددية

Pandas لقراءة ملفات الCSV

Matplotlib للتعامل مع الرسوم البيانية

Seaborn للتعامل مع الرسوم الإحصائية

TransactionEncoder لترميز معاملات قاعدة المعطيات في شكل قائمة قوائم

Warnings للتعامل مع التحذيرات و الاستثناءات

– و تم أيضاً استيراد المبادئ الإحصائية التالية:

Apriori, Association rules

```
import numpy as np
import pandas as pd
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
from mlxtend.preprocessing import TransactionEncoder
from mlxtend.frequent_patterns import apriori, association_rules
import networkx as nx
import warnings
warnings.filterwarnings('ignore')
```

– في هذه الخطوة تم قراءة ملفات الCSV بواسطة مكتبة Pandas و استعراض أول خمس أسطر من أول ملف:

```
data1 = pd.read_csv('sms-call-internet-mi-2013-11-02.csv')
data2 = pd.read_csv('sms-call-internet-mi-2013-11-03.csv')
data3 = pd.read_csv('sms-call-internet-mi-2013-11-04.csv')
data4 = pd.read_csv('sms-call-internet-mi-2013-11-05.csv')
data5 = pd.read_csv('sms-call-internet-mi-2013-11-06.csv')
data6 = pd.read_csv('sms-call-internet-mi-2013-11-07.csv')

data1.head()
# data2.head()
# data3.head()
# data4.head()
# data5.head()
# data6.head()
```

و كانت المخرجات كالتالي:

	datetime	CellID	countrycode	smsin	smsout	callin	callout	internet
0	2013-11-02 00:00:00	1	0	0.2445	NaN	NaN	NaN	NaN
1	2013-11-02 00:00:00	1	39	1.4952	1.1213	0.2708	0.3004	46.5094
2	2013-11-02 00:00:00	1	53	0.0018	0.0036	NaN	NaN	NaN
3	2013-11-02 00:00:00	2	0	0.2458	NaN	NaN	NaN	NaN
4	2013-11-02 00:00:00	2	39	1.5028	1.1243	0.2751	0.3023	46.6933

الشكل (١٠) استعراض أول خمسة أسطر من الملف data1

حيث ترمز NaN هنا إلى Not a Number

– ثم تم كتابة هذا الترميز (code) لاستكشاف الملفات:

```
data1.shape
# data2.shape
# data3.shape
# data4.shape
# data5.shape
# data6.shape
```

و كانت المخرجات كالتالي:

```
(1847331, 8)
```

حيث يرمز الرقم الأول لعدد السطور الموجودة في الملف data1

و الرقم الثاني يرمز لعدد الأعمدة الموجودة في الملف data1

– ثم تم استخدام الترميز التالي لضم الملفات كلها إلى ملف واحد ومن ثم استكشاف المعطيات :

```
data = data1
data = data.append(data2)
data = data.append(data3)
data = data.append(data4)
data = data.append(data5)
data = data.append(data6)
data.shape
```

و كانت المخرجات كالتالي:

```
(13197237, 8)
```

حيث يرمز الرقم الأول لعدد الأسطر في الملف data

و يرمز الرقم الثاني لعدد الأعمدة في الملف data

– ثم تم استعراض شكل المعطيات و مواصفاتها من خلال الترميز التالي:

```
data.describe().T
```

و كانت المخرجات كالتالي:

Out[8]:

	count	mean	std	min	25%	50%	75%	max
CellID	13197237.0	5308.762884	2722.913902	1.0	3163.0000	5448.0000	7544.000000	10000.0000
countrycode	13197237.0	363.773914	4386.066813	0.0	33.0000	43.0000	86.000000	97259.0000
smsin	5217770.0	7.837552	27.765199	0.0	0.1008	0.6010	4.696600	2288.7391
smsout	3077028.0	7.436874	25.342487	0.0	0.0917	0.5549	4.449600	2270.6822
callin	3625162.0	7.606378	29.934760	0.0	0.0574	0.2426	2.228300	1328.0880
callout	5629788.0	5.583692	26.575794	0.0	0.0558	0.2080	1.025700	1511.6864
internet	5915110.0	102.303559	429.057813	0.0	0.0579	0.2055	5.522375	31748.6325

الشكل (١١) استعراض شكل المعطيات و مواصفاتها

– ثم تم إدخال الترميز التالي للحصول على معلومات و توصيف لنوع المعطيات:

```
data.info()
```

و كانت المخرجات كالتالي:

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 13197237 entries, 0 to 2407383
Data columns (total 8 columns):
 #   Column       Dtype
---  ------       -----
 0   datetime     object
 1   CellID       int64
 2   countrycode  int64
 3   smsin        float64
 4   smsout       float64
 5   callin       float64
 6   callout      float64
 7   internet     float64
dtypes: float64(5), int64(2), object(1)
memory usage: 906.2+ MB
```

حيث تم وصف كل عامود و نوع المعطيات المتواجدة فيه

و نلاحظ من ذلك أن العامود رقم 0 نوع المعطيات المتواجدة فيه هي object لذلك سنعمل على تحويلها من نمط object إلى نمط date time

٢-٣-٥ مرحلة المعالجة المسبقة للبيانات (preprocessing):

- أولاً تم العمل على تحويل المعطيات ذات النمط object إلى نمط date time باستخدام مكتبة pandas كالتالي:

```
data['datetime'] = pd.to_datetime(data['datetime'])
```

- و من ثم تم استخدام الترميز التالي لعرض الوقت:

```
data['datetime'].dt.time
```

و كانت المخرجات كالتالي:

```
0          00:00:00
1          00:00:00
2          00:00:00
3          00:00:00
4          00:00:00
             ...
2407379    23:00:00
2407380    23:00:00
2407381    23:00:00
2407382    23:00:00
2407383    23:00:00
Name: datetime, Length: 13197237, dtype: object
```

- بعدها تم تجميع المعطيات بحسب الcell id الخاص بها وعرضها حسب حجمها:

```
data.groupby('CellID').size()
```

وكانت المخرجات كالتالي:

```
CellID
1          764
2          764
3          660
4          764
5          787
          ...
9996      1334
9997       893
9998       893
9999      1073
10000     1060
Length: 10000, dtype: int64
```

حيث يرمز الرقم الأول لعدد الcell id و الرقم التاني هو ال id الخاص بالخلية

– ثم تم استخدام الترميز التالي لمعرفة المعطيات المتفردة في الخلية cell id:

```
data.CellID.unique()
```

و كانت المخرجات كالتالي:

```
array([    1,     2,     3, ...,  9998,  9999, 10000], dtype=int64)
```

حيث تم عرض المعطيات المتفردة في الخلية cell id كمصفوفة

– و بعدها تم استخدام الترميز التالي لعرض عدد المعطيات المتفردة في الخلية cell id:

```
data.CellID.nunique()
```

و كانت المخرجات على الشكل التالي:

```
10000
```

– ومن ثم تم استخدام الترميز التالي للتأكد من عدم وجود معطيات ذات قيمة none في الخلية cell id:

```
data[data['CellID']=='NONE'].shape
```

وكانت المخرجات على الشكل التالي:

```
(0, 8)
```

أي أنه لا يوجد أي قيمة ذات النمط none

– ثم تم استخدام الترميز التالي لعدّ القيم داخل الخلية cell id و ترتيبها تنازلياً و عرض أول عشرة قيم:

```
data['CellID'].value_counts().sort_values(ascending=False).head(10)
```

و كانت المخرجات على الشكل التالي:

```
6064    5645
6165    5441
6065    5025
5161    4888
6066    4852
5160    4802
6164    4774
5159    4631
5965    4458
5060    4415
Name: CellID, dtype: int64
```

حيث يرمز الرقم الأول لل id المعرّف للخلية

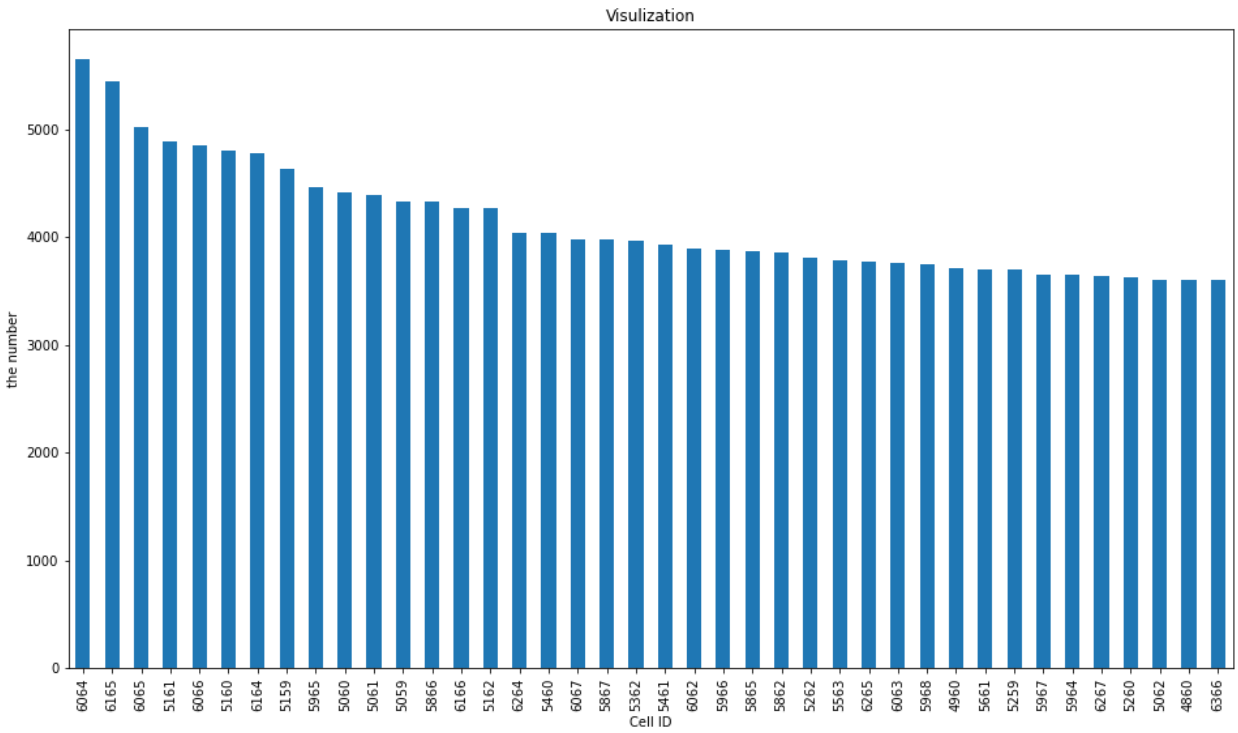
و الرقم الثاني يرمز لعدد التكرارات

– بعدها تم استخدام الترميز التالي لتعريف محاور الرسم البياني الذي سيتم رسمه للقيم من الخلية cell id لأول أربعين قيمة:

```
fig, ax=plt.subplots(figsize=(16,9))
data['CellID'].value_counts().sort_values(ascending=False).head(40).plot(kind='bar')
plt.ylabel('the number')
plt.xlabel('Cell ID')
#ax.get_yaxis().get_major_formatter().set_scientific(False)
plt.title('Visulization')
```

و كانت المخرجات على الشكل التالي:

```
Text(0.5, 1.0, 'Visulization')
```



الشكل (١٢) تعداد القيم من الخلية cell id

– ثم تم التعديل على خلية ال date time و اضافة سمات جديدة للوقت ألا وهي أنه قد تم إضافة عامود جديد إسمه day time وتم تقسيم أوقات اليوم فيه إلى:

الصباح ويبدأ من الساعة الثانية عشر

الظهر ويمتد من الساعة الثانية عشر وحتى الساعة الخامسة

المساء ويمتد من الساعة الخامسة وحتى الساعة التاسعة

الليل ويمتد من الساعة التاسعة وحتى الساعة الحادية عشر وخمسين دقيقة

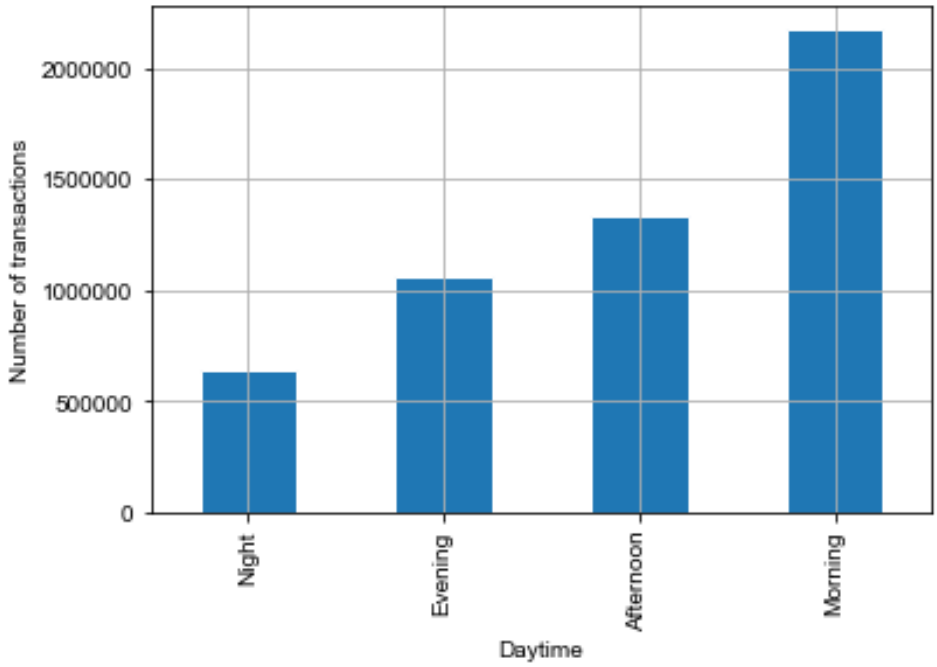
```
from datetime import datetime

data.loc[(data['datetime'].dt.time<datetime.strptime('12::00::00', '%H::%M::%S').time()),'Daytime']='Morning'
data.loc[(data['datetime'].dt.time>=datetime.strptime('12::00::00', '%H::%M::%S').time())&(data['datetime'].dt.time<datetime.str
ptime('17::00::00', '%H::%M::%S').time()),'Daytime']='Afternoon'
data.loc[(data['datetime'].dt.time>=datetime.strptime('17::00::00', '%H::%M::%S').time())&(data['datetime'].dt.time<datetime.str
ptime('21::00::00', '%H::%M::%S').time()),'Daytime']='Evening'
data.loc[(data['datetime'].dt.time>=datetime.strptime('21::00::00', '%H::%M::%S').time())&(data['datetime'].dt.time<datetime.str
ptime('23::50::00', '%H::%M::%S').time()),'Daytime']='Night'
```

– ثم تم استخدام الترميز التالي لتعريف محاور الرسم البياني الذي سيتم رسمه لقيم الخلية sms in مجمعة بحسب ال day time :

```
fig, ax=plt.subplots(figsize=(6,4))
sns.set_style('darkgrid')
data.groupby('Daytime')['smsin'].count().sort_values().plot(kind='bar')
plt.ylabel('Number of transactions')
ax.get_yaxis().get_major_formatter().set_scientific(False)
```

و كانت المخرجات على الشكل التالي:



الشكل (١٣) قيم الخلية sms in مجمعة بحسب الday time

– ثم تم استخدام الترميز التالي لتجميع قيم الخلية cell id بحسب ال day time و ترتيبها تنازلياً:

```
data.groupby('Daytime')['CellID'].count().sort_values(ascending=False)
```

و كانت المخرجات على الشكل التالي:

```
Daytime
Morning      5118988
Afternoon    3774279
Evening      2871974
Night        1431996
Name: CellID, dtype: int64
```

– ثم تم استخدام الترميز التالي لإنشاء ثلاثة أعمدة جديدة (sec, minute, hour) :

```
data['Sec']=data['datetime'].dt.second
data['Minute']=data['datetime'].dt.minute
data['Hour']=data['datetime'].dt.hour
```

- بعدها تم استخدام الترميز التالي لاستعراض أول خمسة أسطر من قاعدة المعطيات:

```
data.head()
```

و كانت المخرجات على الشكل التالي:

	datetime	CellID	countrycode	smsin	smsout	callin	callout	internet	Daytime	Sec	Minute	Hour
0	2013-11-02	1	0	0.2445	NaN	NaN	NaN	NaN	Morning	0	0	0
1	2013-11-02	1	39	1.4952	1.1213	0.2708	0.3004	46.5094	Morning	0	0	0
2	2013-11-02	1	53	0.0018	0.0036	NaN	NaN	NaN	Morning	0	0	0
3	2013-11-02	2	0	0.2458	NaN	NaN	NaN	NaN	Morning	0	0	0
4	2013-11-02	2	39	1.5028	1.1243	0.2751	0.3023	46.6933	Morning	0	0	0

الشكل(١٤) شكل قاعدة المعطيات الجديد

- ثم تم ضم الخليتين (hour, minute) لتشكيل خلية واحدة باستخدام الترميز التالي:

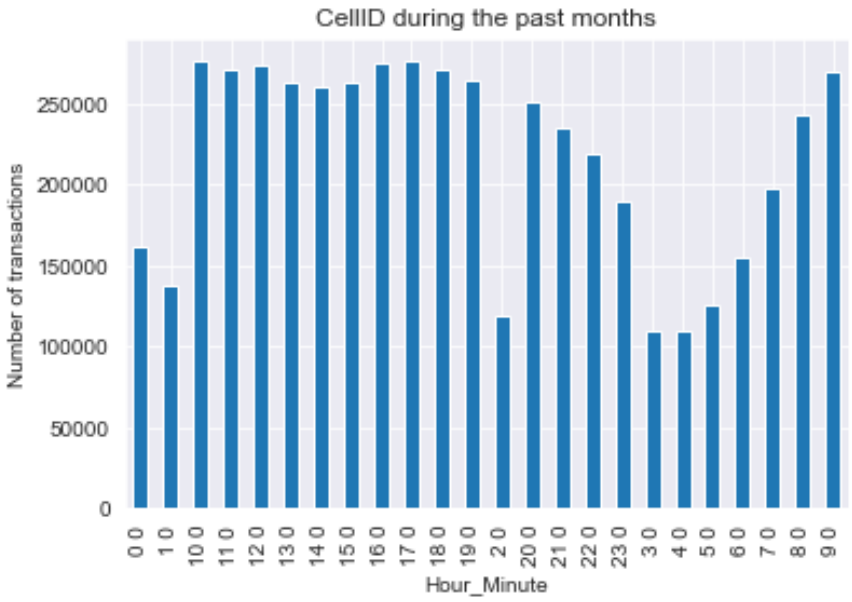
```
data['Hour_Minute']=data['Hour'].apply(str)+' '+data['Minute'].apply(str)
```

- ثم تم استخدام الترميز التالي لتعريف محاور الرسم البياني الذي سيتم رسمه لقيم الخلية sms in مجمعة حسب الhour_minute :

```
data.groupby('Hour_Minute')['smsin'].count().plot(kind='bar')
plt.ylabel('Number of transactions')
plt.title('CellID during the past months')
```

و كانت المخرجات كالتالي:

```
Text(0.5, 1.0, 'CellID during the past months')
```



الشكل (١٥) قيم الخلية sms in مجمعة حسب hour_minute

– و بعدها تم استخدام الترميز التالي لعرض عدد المعطيات المتفردة في الخلية hour_minute في وقت محدد (الساعة الحادية عشر):

```
data.loc[data['Hour_Minute']=='23 0'].nunique()
```

و كانت المخرجات كالتالي:

```
datetime          6
CellID        10000
countrycode     182
smsin         68903
smsout        52521
callin        38172
callout       42403
internet      70851
Daytime           1
Sec               1
Minute            1
Hour              1
Hour_Minute       1
dtype: int64
```

– ثم تم استخدام الترميز التالي للخلية cell id لمعرفة حجم استهلاك الخدمات في الساعة:

```
data['CellID'].resample('60min').count().plot()
plt.ylabel('Number of transactions')
plt.title('Business during the past Hours')
```

و كانت المخرجات على الشكل التالي:

```
Text(0.5, 1.0, 'Business during the past Hours')
```

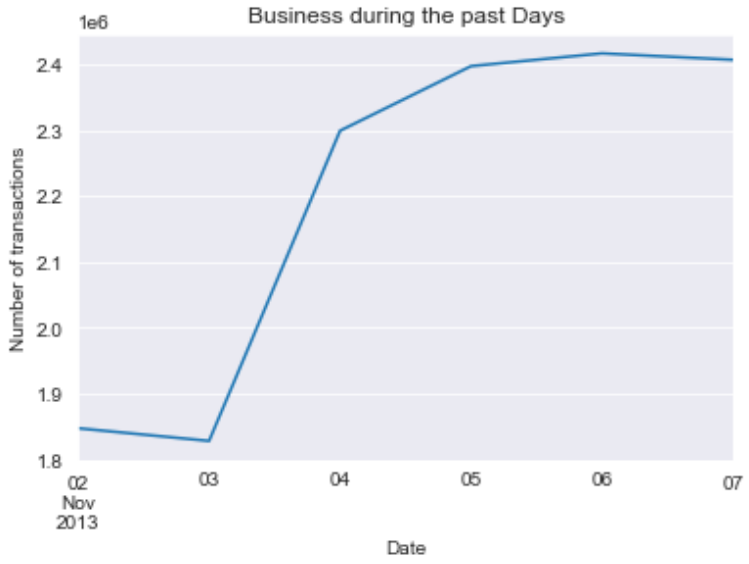
الشكل (١٦) حجم الاستهلاك في الساعة

– ثم تم استخدام الترميز التالي للخلية cell id لمعرفة حجم الاستهلاك اليومي:

```
data['CellID'].resample('D').count().plot()
plt.ylabel('Number of transactions')
plt.title('Business during the past Days')
```

و كانت المخرجات على الشكل التالي:

```
Text(0.5, 1.0, 'Business during the past Days')
```



الشكل (١٧) حجم الاستهلاك في اليوم

– ثم تم إدخال الترميز التالي للحصول على معلومات و توصيف لنوع المعطيات:

```
data.info()
```

و كانت المخرجات على الشكل التالي:

```
<class 'pandas.core.frame.DataFrame'>
DatetimeIndex: 13197237 entries, 2013-11-02 00:00:00 to 2013-11-07 23:00:00
Data columns (total 13 columns):
 #   Column       Dtype
---  ------       -----
 0   datetime     datetime64[ns]
 1   CellID       int64
 2   countrycode  int64
 3   smsin        float64
 4   smsout       float64
 5   callin       float64
 6   callout      float64
 7   internet     float64
 8   Daytime      object
 9   Sec          int64
 10  Minute       int64
 11  Hour         int64
 12  Hour_Minute  object
dtypes: datetime64[ns](1), float64(5), int64(5), object(2)
memory usage: 1.4+ GB
```

– ثم تم استعراض شكل المعطيات و مواصفاتها من خلال الترميز التالي:

```
data.describe()
```

و كانت المخرجات على الشكل التالي:

	CellID	countrycode	smsin	smsout	callin	callout	internet	Sec	Minute	Hour
count	1.319724e+07	1.319724e+07	5.217770e+06	3.077028e+06	3.625162e+06	5.629788e+06	5.915110e+06	13197237.0	13197237.0	1.319724e+07
mean	5.308763e+03	3.637739e+02	7.837552e+00	7.436874e+00	7.606378e+00	5.583692e+00	1.023036e+02	0.0	0.0	1.313569e+01
std	2.722914e+03	4.386067e+03	2.776520e+01	2.534249e+01	2.993476e+01	2.657579e+01	4.290578e+02	0.0	0.0	5.841075e+00
min	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.0	0.0	0.000000e+00
25%	3.163000e+03	3.300000e+01	1.008000e-01	9.170000e-02	5.740000e-02	5.580000e-02	5.790000e-02	0.0	0.0	9.000000e+00
50%	5.448000e+03	4.300000e+01	6.010000e-01	5.549000e-01	2.426000e-01	2.080000e-01	2.055000e-01	0.0	0.0	1.300000e+01
75%	7.544000e+03	8.600000e+01	4.696600e+00	4.449600e+00	2.228300e+00	1.025700e+00	5.522375e+00	0.0	0.0	1.800000e+01
max	1.000000e+04	9.725900e+04	2.288739e+03	2.270682e+03	1.328088e+03	1.511686e+03	3.174863e+04	0.0	0.0	2.300000e+01

الشكل (١٨) شكل المعطيات و مواصفاتها

– بعدها تم كتابة الترميز التالي لمعرفة نسبة المعطيات ذات القيمة null و تقريبها لأقرب رقمين عشريين:

```
df_null = round(100*(data.isnull().sum())/len(data), 2)
df_null
```

و كانت المخرجات على الشكل التالي:

```
datetime        0.00
CellID          0.00
countrycode     0.00
smsin          60.46
smsout         76.68
callin         72.53
callout        57.34
internet       55.18
Daytime         0.00
Sec             0.00
Minute          0.00
Hour            0.00
Hour_Minute     0.00
dtype: float64
```

– بعدها تم استخدام الترميز التالي للتخلص من المعطيات ذات القيمة null واستعراض الشكل الجديد:

```
data = data.dropna()
data.shape
```

و كانت المخرجات على الشكل التالي:

```
(1368838, 13)
```

حيث يرمز الرقم الأول لعدد الأسطر بعد حذف المعطيات ذات القيمة null

و الرقم الثاني يرمز لعدد الأعمدة

– ثم تم استعراض شكل المعطيات و مواصفاتها من خلال الترميز التالي:

```
data.describe()
```

و كانت المخرجات على الشكل التالي:

	CellID	countrycode	smsin	smsout	callin	callout	internet	Sec	Minute	Hour
count	1.368838e+06	1.368838e+06	1.368838e+06	1.368838e+06	1.368838e+06	1.368838e+06	1.368838e+06	1368838.0	1368838.0	1.368838e+06
mean	5.046902e+03	3.956543e+01	2.037100e+01	1.596795e+01	1.973557e+01	2.183965e+01	4.306107e+02	0.0	0.0	1.226297e+01
std	2.847413e+03	1.502273e+01	4.760675e+01	3.593174e+01	4.622084e+01	5.052441e+01	8.069883e+02	0.0	0.0	6.692658e+00
min	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.0	0.0	0.000000e+00
25%	2.627000e+03	3.900000e+01	1.741500e+00	1.601500e+00	1.075000e+00	1.271800e+00	7.159742e+01	0.0	0.0	7.000000e+00
50%	5.097000e+03	3.900000e+01	6.487600e+00	5.339300e+00	5.242100e+00	6.058500e+00	1.905562e+02	0.0	0.0	1.300000e+01
75%	7.484000e+03	3.900000e+01	1.965150e+01	1.534175e+01	1.914340e+01	2.133817e+01	4.420716e+02	0.0	0.0	1.800000e+01
max	1.000000e+04	1.519000e+03	2.288739e+03	2.270682e+03	1.328088e+03	1.511686e+03	3.174863e+04	0.0	0.0	2.300000e+01

الشكل (١٩) شكل المعطيات الجديد و مواصفاتها

٣-٣-٥ مرحلة معالجة المعطيات(processing phase):

أولاً تم إنشاء سمة (feature) للخلية cell id سميت ب TeleV (Telecommunication volume feature) و هي مكونة من مجموع الخدمات المقدمة، واستعراض أول خمسة أسطر باستخدام الترميز التالي:

```
data.loc[:,'TeleV'] = data.loc[:,'smsin']+data.loc[:,'smsout']+data.loc[:,'internet']+data.loc[:,'callin']+data.loc[:,'callout']
cluster_feature = data.groupby('CellID')['TeleV'].sum()
cluster_feature = cluster_feature.reset_index()
cluster_feature.head()
```

و كانت المخرجات كالتالي:

	CellID	TeleV
0	1	9598.4281
1	2	9657.8189
2	3	9559.1936
3	4	9426.4030
4	5	8637.9866

الشكل (٢٠) استعراض لقيم الميزة TeleV مجمعة حسب cell id

- ثم تم كتابة هذا الترميز لاستكشاف الميزة:

```
cluster_feature.shape
```

و كانت المخرجات على الشكل التالي:

```
(9998, 2)
```

- ثم تم استعراض شكل معطيات الميزة و مواصفاتها من خلال الترميز التالي:

```
cluster_feature.describe()
```

و كانت المخرجات على الشكل التالي:

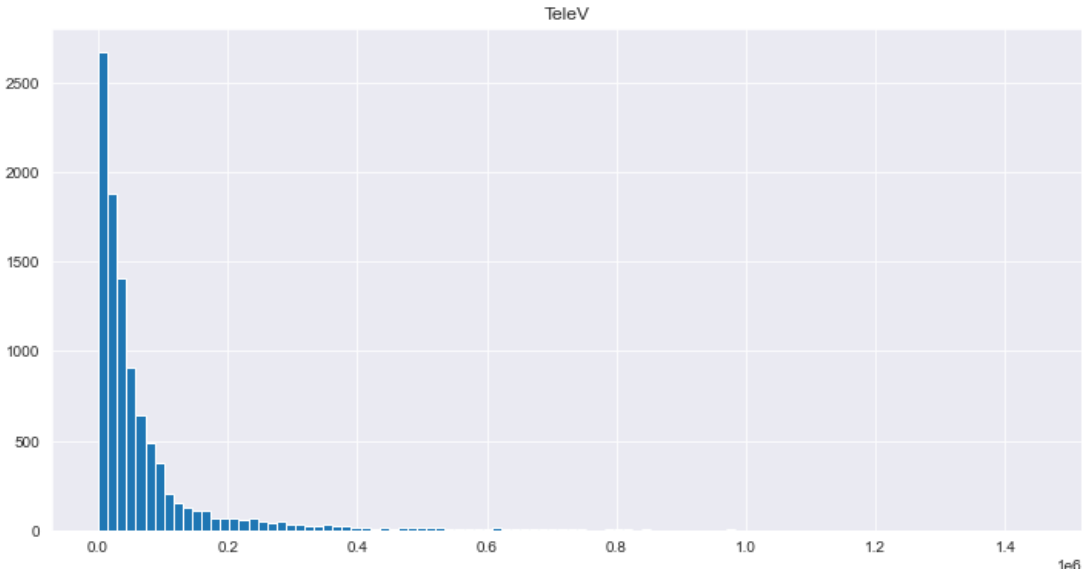
	CellID	TeleV
count	9998.000000	9.998000e+03
mean	5000.442288	6.962274e+04
std	2887.181471	1.122786e+05
min	1.000000	2.893819e+02
25%	2500.250000	1.361706e+04
50%	4999.500000	3.330680e+04
75%	7500.750000	7.249176e+04
max	10000.000000	1.446270e+06

الشكل (٢١) شكل معطيات الميزة و مواصفاتها

– ثم تم استخدام الترميز التالي لرسم حجم الخلية TeleV بيانياً (histogram):

```
cluster_feature.hist(column='TeleV', bins=100, figsize=(12,6))
```

و كانت المخرجات على الشكل التالي:



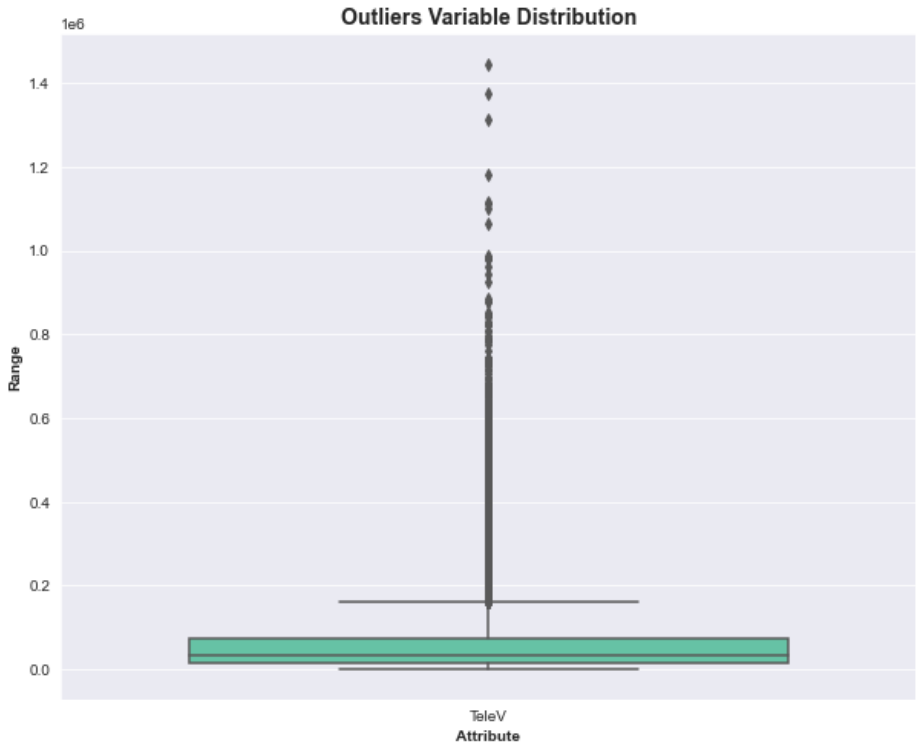
الشكل (٢٢) استعراض حجم الميزة

– ثم تم استخدام الترميز التالي لرسم مخطط الصندوق و الساعدين (box plot) للميزة TeleV لتوضيح توّزع القيم، الربيعات، المتوسط و مكان تواجد القيم الشاذّة:

```
attribute = ['TeleV']
plt.rcParams['figure.figsize'] = [10,8]
sns.boxplot(data = cluster_feature[attribute], orient="v", palette="Set2" ,whis=1.5,saturation=1, width=0.7)
plt.title("Outliers Variable Distribution", fontsize = 14, fontweight = 'bold')
plt.ylabel("Range", fontweight = 'bold')
plt.xlabel("Attribute", fontweight = 'bold')
```

و كانت المخرجات كالتالي:

```
Text(0.5, 0, 'Attribute')
```



الشكل (٢٣) مخطط الصندوق والساعدين (box plot)

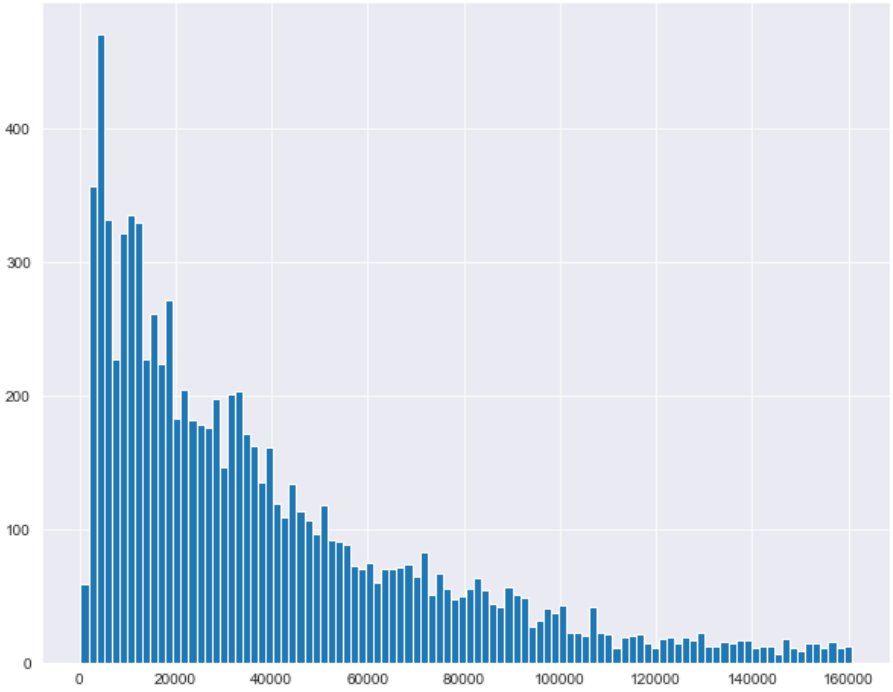
– بعدها تم استخدام الترميز التالي للتخلص من القيم الشاذة:

```
# Removing (statistical) outliers for Amount
Q1 = cluster_feature.TeleV.quantile(0.25)
Q3 = cluster_feature.TeleV.quantile(0.75)
IQR = Q3 - Q1
cluster_feature = cluster_feature[(cluster_feature.TeleV >= Q1 - 1.5*IQR) & (cluster_feature.TeleV <= Q3 + 1.5*IQR)]
```

– ومن ثم تم استخدام الترميز التالي لعرض رسم بياني لقيم الخلية TeleV بعدما تم التخلص من القيم الشاذّة:

```
cluster_feature.TeleV.hist(bins=100)
```

و كانت المخرجات على الشكل التالي:



الشكل (٢٤) قيم الخلية TeleV بعدما تم التخلص من القيم الشاذّة

– بعدها تم استيراد المكاتب التالية:

Numpy للتعامل مع المصفوفات العددية

Pandas لقراءة ملفات الCSV

Matplotlib للتعامل مع الرسوم البيانية

Seaborn للتعامل مع الرسوم الإحصائية

Datetime للتعامل مع تنسيق الوقت

StandardScaler للتوزيع الطبيعي للمعطيات

KMeans خوارزمية تصنيف للتعامل مع العناقيد

silhouette_score للتعامل مع عدد العناقيد

من خلال الترميز التالي:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt

# import required libraries for clustering
import sklearn
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
```

– ثم تم استخدام الترميز التالي لتعريف السمة المستخدمة ألا و هي TeleV و تعريف المتحول Scaler الذي سيحوّل المعطيات للتوزيع الطبيعي، ومن ثم التحويل للتوزيع الطبيعي و عرض النتائج:

```
feature = cluster_feature[['TeleV']]

# Instantiate
scaler = StandardScaler()

# fit_transform
cluster_feature_scaled = scaler.fit_transform(feature)
cluster_feature_scaled.shape
```

و كانت المخرجات على الشكل التالي:

```
(8971, 1)
```

حيث يرمز الرقم الأول لعدد الأسطر

و الرقم الثاني لعدد الأعمدة

– بعدها تم استخدام الترميز التالي لتحويل نمط المعطيات إلى نمط data frame و عرض أول خمسة عناصر:

```
cluster_feature_scaled = pd.DataFrame(cluster_feature_scaled)
cluster_feature_scaled.columns = ['TeleV']
cluster_feature_scaled.head()
```

و كانت المخرجات كالتالي:

	TeleV
0	-0.852475
1	-0.850756
2	-0.853611
3	-0.857455
4	-0.880281

الشكل (٢٥) أول خمس عناصر من سمة العنقود TeleV

– ثم تم استعراض شكل المعطيات و مواصفاتها من خلال الترميز التالي:

```
cluster_feature_scaled.describe()
```

و كانت المخرجات على الشكل التالي:

	TeleV
count	8.971000e+03
mean	7.026297e-16
std	1.000056e+00
min	-1.121985e+00
25%	-7.797759e-01
50%	-2.993352e-01
75%	4.749519e-01
max	3.521989e+00

الشكل (٢٦) شكل المعطيات و مواصفاتها بالنسبة للسمة TeleV

– بعدها تم استخدام الترميز التالي لتطبيق خوارزمية Kmeans بإفتراض أربعة كعدد عناقيد:

```
 kmeans = KMeans(n_clusters=4, max_iter=50)
kmeans.fit(cluster_feature_scaled)
```

– و بعدها تم استخدام الترميز التالي لتحديد العدد الأمثل للعناقيد باستخدام طريقة الساعد (Elbow_curve) و خوارزمية Kmeans:

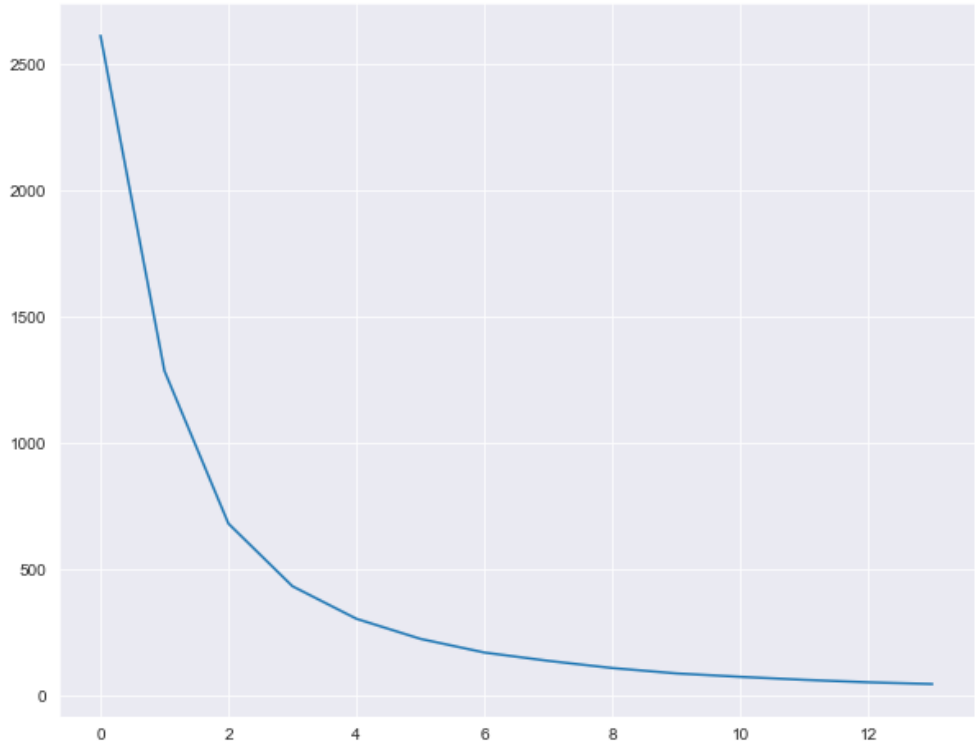
```
# Elbow-curve/SSD

ssd = []
range_n_clusters = [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]
for num_clusters in range_n_clusters:
    kmeans = KMeans(n_clusters=num_clusters, max_iter=50)
    kmeans.fit(cluster_feature_scaled)

    ssd.append(kmeans.inertia_)

# plot the SSDs for each n_clusters
plt.plot(ssd)
```

و كانت المخرجات كالتالي:



الشكل (٢٧) طريقة Elbow_curve لتحديد العدد الأمثل للعناقيد

– و بعدها تم استخدام الترميز التالي لمعرفة عدد العناقيد الأمثل بدقة باستدام التابع silhouette score و خوارزمية Kmeans:

```
for num_clusters in range_n_clusters:

    # intialise kmeans
    kmeans = KMeans(n_clusters=num_clusters, max_iter=50)
    kmeans.fit(cluster_feature_scaled)

    cluster_labels = kmeans.labels_

    # silhouette score
    silhouette_avg = silhouette_score(cluster_feature_scaled, cluster_labels)
    print("For n_clusters={0}, the silhouette score is {1}".format(num_clusters, silhouette_avg))
```

و كانت المخرجات كالتالي:

```
For n_clusters=2, the silhouette score is 0.6704656795623724
For n_clusters=3, the silhouette score is 0.6005597714248295
For n_clusters=4, the silhouette score is 0.5999321569786273
For n_clusters=5, the silhouette score is 0.588131912564738
For n_clusters=6, the silhouette score is 0.5680233883981389
For n_clusters=7, the silhouette score is 0.5607061833883679
For n_clusters=8, the silhouette score is 0.5606423447819456
For n_clusters=9, the silhouette score is 0.542442962646014
For n_clusters=10, the silhouette score is 0.5441607922017695
For n_clusters=11, the silhouette score is 0.5507221685639834
For n_clusters=12, the silhouette score is 0.5515703464761506
For n_clusters=13, the silhouette score is 0.547936663438127
For n_clusters=14, the silhouette score is 0.5540821862954014
For n_clusters=15, the silhouette score is 0.5519398109495917
```

حيث تتراوح قيمة التابع بين الصفر و الواحد، ومع كل عدد مختلف للعناقيد تتغير قيمته وكلما كان أقرب للواحد كان عدد العناقيد المرتبط به هو الأفضل، و نلاحظ من المخرجات أن العدد الأمثل للعناقيد في هذه الحالة هو ٢.

- ثم تم استخدام الترميز التالي لتطبيق خوارزمية Kmeans:

```
# Final model with k=2
kmeans = KMeans(n_clusters=2, max_iter=50)
kmeans.fit(cluster_feature_scaled)
```

- ثم تم استخدام الترميز التالي لتمييز العناصر المنتمية للعنقود الأول والعناصر المنتمية للعنقود الثاني:

```
kmeans.labels_
```

و كانت المخرجات كالتالي:

```
array([1, 1, 1, ..., 1, 1, 1])
```

حيث يرمز الرقم صفر للعناصر التي تنتمي للعنقود الأول

و الرقم واحد للعناصر التي تنتمي للعنقود الثاني

– ثم تم استخدام الترميز التالي لعرض أول خمسة أسطر من الخلايا (TeleV – Kmeans – cell id) :

```
cluster_feature['Cluster_Id'] = kmeans.labels_
cluster_feature.head()
```

و كانت المخرجات على الشكل التالي:

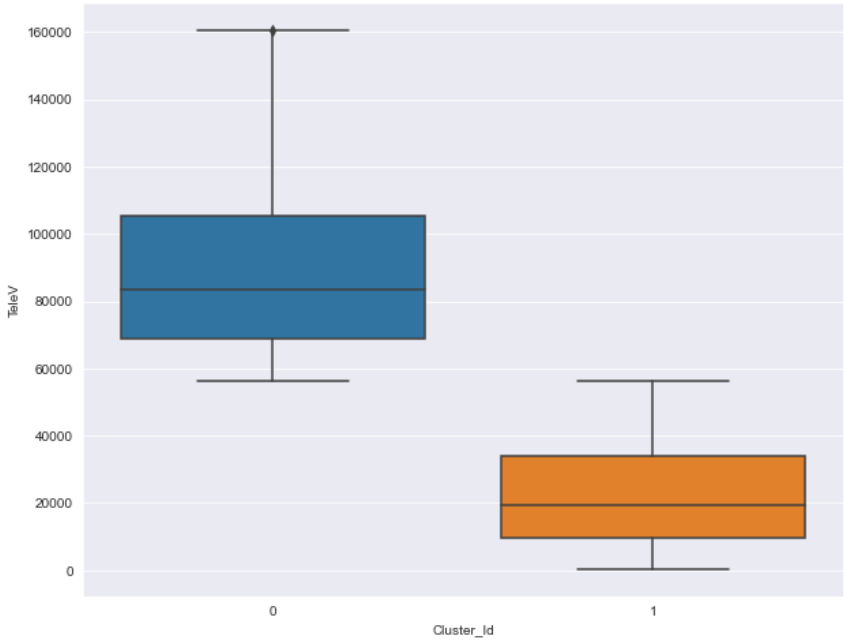
	CellID	TeleV	Cluster_Id
0	1	9598.4281	1
1	2	9657.8189	1
2	3	9559.1936	1
3	4	9426.4030	1
4	5	8637.9866	1

الشكل (٢٨) استعراض لأول خمسة أسطر من الخلايا (TeleV – Kmeans – cell id)

– و أخيراً تم استخدام الترميز التالي لرسم العنقودين بمخطط الصندوق و الساعدين (box plot):

```
sns.boxplot(x='Cluster_Id', y='TeleV', data=cluster_feature)
```

و كانت المخرجات على الشكل التالي:



الشكل (٢٩) مخطط الصندوق و الساعدين للعناقيد

حيث نجد من الشكل أن العنقود الأول ذو الرقم صفر مكوّن من الخلايا التي تستهلك قدراً أكبر من الخدمات المقدمة و بالتالي يتم تصميم عروض خاصة بهذه الفئة على الخدمات التي تم دراستها في البحث.

٦- النتائج والتوصيات:

٦-١ النتائج:

من خلال الدراسة التي أجريت ومن خلال ما تم توضيحه في الشقّين النظري والعملي من البحث يمكن تلخيص النتائج التي توصلت إليها الباحثة من خلال بحثها فيما يلي:

١- يمكن استخدام تقنيات التنقيب في المعطيات لتصميم عروض لزبائن شركات الاتصال.

٢- من خلال القسم النظري تم الحصول على الفئات التي يمكن تصميم عروض خاصّة لها حيث يمكن تصميم عروض خاصّة للفئة المنتمية للعنقود ذو الرقم صفر لأنه كان يحمل السمة (TeleV) بشكل أكبر، ويمكن تصميم عروض خاصة للعنقود ذو الرقم واحد لجذب الزبائن بشكل أكبر لأنه كان يحمل السمة (TeleV) بشكل أقلّ من العنقود السابق.

٦-٢ التوصيات:

❖ توصية خاصة لشركات الاتصالات:

١- تطبيق تقنيات التنقيب في المعطيات على جميع أقسام الشركة وعدم اقتصارها على قسم معيّن لما له من ميزات على المدى الطويل.

٢- التنبؤ المسبق لحاجات العميل ورغباته من أجل إشباعها يكمن في استخدام تقنيات التنقيب عن المعطيات.

٣- فهم المشكلة المراد استخدام تقنيات التنقيب في البيانات عليها لتحديد التقنية المناسبة لها و الحصول على النتائج الأمثل.

❖ دراسات خاصة بالأكاديميين والدراسات المستقبلية:

١- التوسع في هذه الدراسة من خلال تحديد سمات أخرى للمعطيات مثل (sms in & out, call in & out, internet usage).

٢- استخدام أدوات مختلفة عن الذي اعتمدها البحث لتقنيات التنقيب في المعطيات.

٣- أن يكون هذا البحث مرجع غنّي لباقي الدراسات المستقبلية.

٧- المراجع:

٧-١ المراجع العربية:

١- البار علي،الجناعي أواب، الحداد حسين، الزهراوي عمار، استكشاف بعض الأنماط المؤثرة في الأداء الأكاديمي لطلاب جامعة العلوم والتكنولوجيا باستخدام تقنيات التنقيب في البيانات، ٢٠١١.

٢- جلال الضاهر، تصميم نموذج نظام دعم القرار إدارة الموارد البشرية بالإعتماد على تقنيات الذكاء الصنعي، الجامعة الافتراضية السورية، ٢٠١٤.

٣- السبيعي محمد، استخدام تقنيات التنقيب في المعطيات للتنبؤ بنتيجة الحملات التسويقية المباشرة، ٢٠٢٠.

٤- فتوح سيف، محمود الشفيع، التنقيب في البيانات واتخاذ القرارات نموذج تطبيقي لخزان خشم القربة، ٢٠١٤.

٥- الكردي محمد، مراد بيان، نظام إحالة لإدارة المعرفة يعتمد على تقنيات الذكاء الصنعي،٢٠٢٠.

٧-٢ المراجع الأجنبية:

1- Abdel Hafez Hoda, Mining Big Data In Telecommunications Industry: Challenges, Techniques, and revenue opportunity, 2016.

2- Aggarwal Charu, Data Mining The Textbook, 2016.

3- Han Jiawei, Pei Jian, Data Mining Concepts And Techniques, 2012.

4- Kumar Vipin, Ning tan Pang, Steinbach Michael, Book university Of Minnesota And Army High Performance Computing Research, 2006.

5- Kurasova Olga, Marcinkevi˘cius Virginijus, Medvedev Viktor, Rape˘cka Aurimas, and Stefanovi˘ Pavel, StrategiesforBigDataClustering,2014.

6- Morgan S, Teachman, J, Logistic Regression: Description, Examples, And Comparisons, Journal Of Marriage And Family, 1988.

7- Palace B, Data Mining Technology Note Prepared For Management, 1996.

8- Prediction, D.M.-C. (n.d.). Great Islamians, 2015.

9- Raschka S, Naive Bayes And Text Classification, 2014.

10- Romero, Data Mining And Data Classification The Classification Problem, 2017.

11- Shi Yuhui, Tan Ying, Data Mining And Big Data First International Conference, DMBD, 2016.

12- U.F.Eze, C.J.Onwuegbuchulam, S.Diala, Application of data mining in telecommunication industry, 2017.

13- Wang Lidong, Wang Guanghui, Data Mining Applications in Big Data, 2015.