الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي

جامعة قاصدي مرباح -ورقلة-
كلية التكنولوجيات الحديثة للمعلومات والاتصال
قسم الاعلام الالي وتكنولوجيا المعلومات

مذكرة

مقدمة لنيل شهادة الليسانس في الاعلام الآلي

العنوان

التنبؤ بأمراض القلب باستخدام الانحدار اللوجيستي

مقدمة من طرف الطلبة:
- محمد لمين بن حبيرش
- عبد الجليل حموية
- عبد الجليل تجيني

تأطير الأستاذ :
- مزوز ميهوب

العام الدراسي: 2022/2021

الفهرس

قائمة الصور

قائمة الجداول

المقدمة

يعتبر القلب من أهم الأعضاء في جسم الانسان، بحيث إذا لم يكن يعمل بشكل صحيح فسيؤثر بشكل مباشر على بقية أجزاء الجسم وقد يؤدي الى الهلاك. فحسب منظمة الصحة العالمية (WHO) حوالي17.9مليون شخص يموت كل عام جرّاءَ أمراض القلب والأوعية الدموية، أي ما يقدر بنحو32 % من جميع الوفيات في جميع أنحاء العالم.

كما نشهد في الوقت الحالي انتشار كبير لأمراض القلب في شتى أنحاء العالم بين مختلف الفئات خاصة الشباب بعدما كان مقتصر فقط على الشيوخ والكهول، وذلك عائد -من بين أسباب أخرى- للاستخدام السلبي المفرط للتكنولوجيا مما أدى لانخفاض الانشطة البدنية والسمنة، وللإدمان على التدخين والافراط في شرب الخمر والمخدرات وغيرها.

مؤخرا أصبحت خوارزميات الذكاء الاصطناعي تستخدم في شتى المجالات نظرا لضخامة البيانات وتعقيدها، ولكن يبقى التحدي الرئيسي للعلماء والباحثين اليوم هو الدقة في التنبؤ وتحسين النتائج مع مرور الوقت.

يهدف عملنا هذا الى تطبيق مجموعة من خوارزميات التعلم الالي للتنبؤ بأمراض القلب لاتخاذ الإجراءات قبل فوات الأوان والتقليل من معدل الوفيات.

تتكون هذه الورقة البحثية من ثلاثة اقسام رئيسية، يتضمن القسم الأول مجموعة من المفاهيم المتعلقة بالتعلم الآلي والتي تقدم فكرة عامة عن التعلم الآلي. في القسم الثاني نشرح بعض خوارزميات التصنيف التي قمنا باستخدامها وفي القسم الثالث نشرح طريقة تهيئة البيانات ومقارنة للنتائج المتحصل عليها. أخيرا ننهي هذه الورقة البحثية بخاتمة وبعض وجهات النظر.

المحور الأول | التعلم الآلي

Machine learning

في هذا القسم سوف نقوم بالتعرف على بعض المفاهيم الأساسية المتعلقة بـالتعلم الالي بدءا بتعريفه وانتهاء ببعض تطبيقاته:

1. تعريف التعلم الآلي

التعلم الآلي هو أحد تطبيقات الذكاء الاصطناعي، وهو عبارة عن تمكين الكمبيوتر من اكتساب قدرته الخاصة في التعلم والتحسن والتطور بناء على مجموعة من بيانات التدريب. وقد عرفه توم ميتشيل (Tom Mitchell)سنة 1997 بقوله: "يقال عن برنامج كمبيوتر أنه تعلم من تجربة (ت) فيما يتعلق بصنف من المهام (م) وبعض مقاييس الأداء (أ) إذا كان أداءها لـصنف المهام (م) يتحسن عن طريق قياس بواسطة (أ) "[1]. وأكثر منه بساطة تعريف آرثر صامويل (Arthur Samuel) حين وصف التعلم الآلي بقوله: "هو ذلك الفرع من الدراسة الذي يعطي الحواسيب القدرة على التعلم دون برمجتها بشكل صريح " [1].

2. تقنيات التعلم الآلي

1.2. التعلم الخاضع للإشراف (Supervised Learning): في التعلم الخاضع للإشراف يتم استخدام مجموعة من أمثلة التدريب (عددها N) على شكل أزواج من المدخلات والمخرجات (x_i, y_i) بحيث i تأخذ قِيَمَها من 1 إلى N. حيث يتم انشاء y_j بواسطة دالة غير معروفة $y = f(x)$ ويكون الهدف هو اكتشاف دالة الفرضية h التي تقترب من الدالة الحقيقية f اعتمادا على تدربها من أمثلة التدريب المقدمة لها. مثال المدخلات عبارة عن صور لمجموعة من الصور للحافلات واشياء اخرى والمخرجات: هذه حافلة، هذه ليست حافلة[2] . هناك العديد من خوارزميات التعلم بإشراف، نذكر منها:

- **الانحدار الخطي (Linear Regression):** يستعمل للتنبؤ بقيم مستمرة. وذلك بتمثيل خط يمر على اغلب نقاط بيانات التدريب [3]. فللتنبؤ بعنصر جديد نحتاج فقط لإسقاط البيانات على هذا الخط لنتحصل على النتيجة وينقسم الى قسمين:
  - ❖ **الانحدار الخطي لمتغير واحد (Univariate Linear Regression):** ويقصد به ان البيانات تحتوي على ميزة واحدة للتعامل معها يستخدم عادة لوصف العلاقة بين متغيرين مثل التنبؤ بسعر المنزل انطلاقا من مساحته.
  - ❖ **الانحدار الخطي لأكثر من متغير (Multuivaeriante Linear Regression):** ويستخدم للتنبؤ بعنصر ما انطلاقا من أكثر من ميزة مثل التنبؤ بدرجة الحرارة في يوم ما باستخدام أحوال الطقس.

- **الانحدار اللوجستي (Logistic Regression):** الانحدار اللوجستي هو أحد الإجراءات الإحصائية الأكثر استخدامًا في البحث. وهو مكون من جميع حزم الإحصائية التجارية للأغراض العامة تقريبًا، إن لم يكن كلها، ويعتبر واحد من أهم أدوات الروتين الإحصائي في العديد من المجالات مثل تحليل الرعاية الصحية، والإحصاءات الطبية، والتصنيف الائتماني، علم البيئة والإحصاءات الاجتماعية والاقتصاد القياسي ومجالات أخرى مماثلة. كما اعتبر العديد من المحللين أن الانحدار اللوجستي هو أهم إجراء في التحليلات التنبؤية[4] ، وسنتطرق للانحدار اللوجستي أكثر في الفصل الثاني.
- **الدعم الآلي المتجه (Support Vector Machines):** الهدف من SVM هو إيجاد الحلول لمشاكل التصنيف (Classification problems) بالبحث عن أفضل حدود القرار (Decision Boundaries) بين مجموعتين مختلفتين من النقاط من صنفين مختلفين. وحدود القرار عبارة عن سطح أو مجال يفصل بيانات التدريب الى مساحتين كل مساحة تعبر عن صنف ما. فلتصنيف عنصر جديد نحتاج فقط ان نعرف في أي جانب من حدود القرار يقع. أفضل حدود القرار هو الذي يأخذ أكبر مسافة بينه وبين أقرب نقطة بيانات من كل صنف[5] .

2.2. التعلم غير الخاضع للإشراف (Unsupervised Learning): يهدف هذا النوع من التعلم إلى إيجاد العلاقات والارتباطات المهمة في مجموعة من أمثلة التدريب دون تقديم أية مخرجات لتصوير البيانات أو ضغطها أو فهم الارتباطات الموجودة بينها وتنظيمها [5]. من أكثر المسائل المعروفة في هذا النوع من التعلم:[1]

- **التحليل العنقودي (Clustering):** من أشهر خوارزمياته K-MEANSوDBSCAN.
- **كشف الشذوذ (Anomaly Detection)**: مثل One-class SVM وIsolation Forest.
- **تقليل الأبعاد (Dimensionality reduction):** مثل PCA وLLE

3. تطبيقات التعلم الآلي

طغت اليوم استخدامات الذكاء الاصطناعي والتعلم الآلي على جميع نواحي حياتنا اليومي، على سبيل المثال نذكر:

- ✓ **التكنولوجيا الحيوية:** أدى التقدم في تكنولوجيا التسلسل والفحص إلى إنشاء مجموعة بيانات ضخمة من العديد من الأنواع المختلفة، مثل تسلسل الحمض النووي، وهياكل البروتين، وتعبير الحمض النووي الريبي .يتم تطبيق تقنيات التعلم الآلي على نطاق واسع على جميع هذه الأنواع من البيانات في محاولة للعثور على أنماط يمكن أن تزيد من فهم العمليات البيولوجية[6] .
- ✓ **كشف الاحتيال المالي:** تبحث شركات بطاقات الائتمان باستمرار عن طرق جديدة للكشف عن العمليات الاحتيالية، تحقيقا لهذه الغاية استخدموا تقنيات التعلم الالي مثل الشبكات العصبية.
- ✓ **رؤية الآلة:** يتم استخدام العديد من تقنيات التعلم الآلي لتحليل الصور والفيديوهات من الكاميرات من أجل الكشف عن المتسللين او تحديد المركبات او التعرف على الوجوه، خاصة تقنيات التعلم غير الخاضعة للإشراف مثل المكون المستقل التحليلي (independent component analysis) الذي له ميزات مثيرة للاهتمام في البيانات الضخمة.

- ✓ **تسويق المنتجات:** يمكن لتقنيات التعلم الآلي أن تعطينا فهما أفضل لانقسامات المستهلكين في الأسواق في ظل وجود القدرة على جمع كم هائل من المعلومات من المستهلكين، وجعلها أفضل عن طريق التنبؤ بالاتجاهات المستقبلية، من بين الخوارزميات التي تستخدم لذلك خوارزمية التجميع.[6]
- ✓ **تحسين سلسلة الاستيراد:** يمكن للمؤسسات الكبيرة توفير ملايين الدولارات من خلال امتلاك سلاسل الاستيراد الخاصة بها التي تمنحها تشغيل فعال ودقيق من خلال التنبؤ بالطلبات على المنتجات في مناطق مختلفة وعدد الطرق التي يمكن من خلالها بناء سلسلة الاستيراد وعدد العوامل التي يمكن أن تؤثر على الطلب.
- ✓ **تحليل سوق الأسهم:** منذ أن كان هناك سوق للأوراق المالية حاول الناس استخدام الرياضيات لكسب المزيد من المال . نظرا لأن المشاركين أصبحوا أكثر تطورا من أي وقت مضى فقد أصبح من الضروري تحليل مجموعات أكبر من البيانات واستخدام التقنيات المتقدمة للكشف عن الأنماط .
- ✓ **الأمن القومي:** يتم جمع كميات هائلة من المعلومات من قبل الوكالات الحكومية في جميع انحاء العالم ويتطلب تحليل هذه البيانات أجهزة الكمبيوتر للكشف عن الأنماط وربطها بالتهديدات المحتملة.[6]

المحور الثاني | مشكلة التصنيف

Classification problem

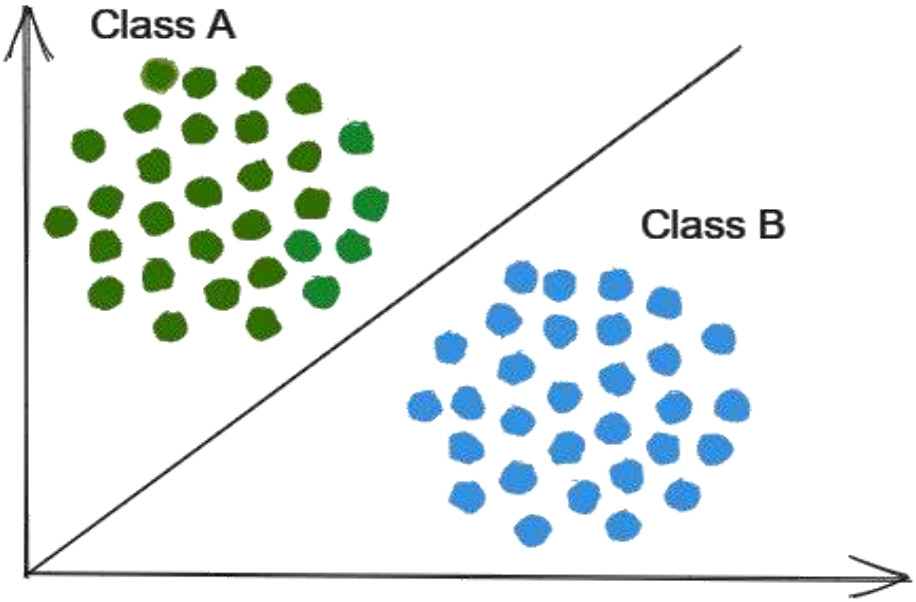
بعد أن تم التعرف على التعلم الآلي في المحور الأول، سنتطرق في هذا المحور أكثر عن التصنيف والخوارزميات الخاصة به.

1. تعريف التصنيف

هو أحد مفاهيم التعلم الخاضع للإشراف كونه يرفق لكل مثال تدريبي النتيجة الفعلية له بحيث يقوم بتصنيف مجموعة من البيانات الى فئات محددة باستخدام مجموعة من الطرق المحددة.

2. أنواع التصنيف

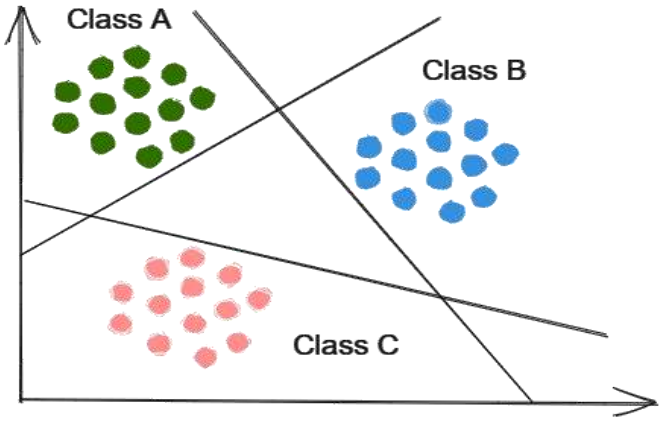
1.2. تصنيف ثنائي (Two-class Classification): ويعتبر النوع الأكثر تطبيقا على نطاق واسع في التعلم الالي، بحيث يتم فيه الفصل بين فئتين بالخط الذي ينتجه النموذج (حدود القرار) والذي يحدد اين يتغير القرار من فئة الى أخرى، وأحد امثلة التصنيف الثنائي تصنيف تقييمات الأفلام إيجابية ام سلبية.



الصورة 1: تصنيف ثنائي لمجموعة بيانات

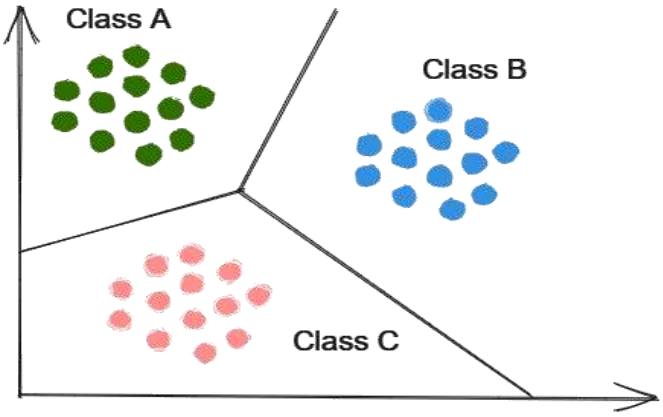
2.2. تصنيف متعدد (Multi-class Classification): وهو تصنيف البيانات الى ثلاث أصناف او أكثر ويوجد طريقتين لتطبيقه هما:

- **الواحد ضد الكل (One-vs-All):** تقوم هذه الطريقة باعتبار أحد الأصناف بانه هو الصنف الصحيح والبقية كلها خاطئة فتقوم بفصله عنها باستخدام التصنيف الثنائي تم تنتقل الى الصنف الموالي وتقوم بفصله هو الاخر وهكذا الى ان تنتهي من كل الأصناف كما هو موضح في الصورة (2).



الصورة 2: التصنيف المتعدد لمجموعة بيانات باستخدام طريقة الواحد ضد الكل

- **الكل ضد الكل (All-vs-All):** وهي تقنيات حديثة يتم فيها فصل كل الأصناف في ان واحد باستعمال خوارزميات جاهزة ذات كفاءة عالية.



الصورة 3: التصنيف المتعدد لمجموعة بيانات باستخدام طريقة الكل ضد الكل

3. خوارزميات التصنيف

يوجد العديد من خوارزميات التصنيف أهمها:

1.3. الانحدار اللوجستي (Logistic Regression):

هي الخوارزمية التي قمنا بالتركيز عليه في عملنا هذا، يقوم الانحدار اللوجستي بأجراء انحدار لكل فئة وتعيين الناتج لإحدى القيم التي تعبر عن النتائج، في التصنيف الثنائي مثلا عادة ما يتم تعيين (1) للتعبير عن القيم الإيجابية و(0) للتعبير عن القيم السلبية ونجد الانحدار اللوجستي مستخدما بكثرة في المجالات الرقمية وعادة ما يستخدم الباحثون والمحللون الانحدار اللوجستي بشكل عام لثلاثة أغراض **[4]** :

- للتنبؤ باحتمالية أن تكون النتيجة أو الاستجابة متغيرة.
- لتصنيف النتائج أو التنبؤات
- للوصول إلى الاحتمالات أو المخاطر المرتبطة بالتنبؤات النموذجية.

لشرح طريقة عمل الانحدار اللوجستي نفترض ان لدينا فئتين فقط، يستخدم الانحدار اللوجستي للتنبؤ بالمخرج والذي لا يمكن تقريبه بدقة بواسطة دالة الفرضية التي تستخدم في الانحدار الخطي:

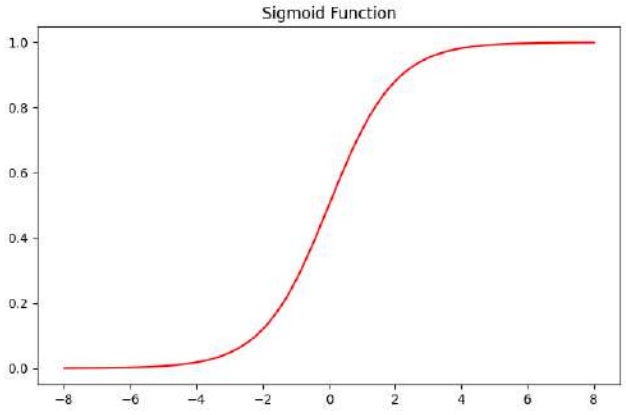
$$h(x_1.x_2 \dots x_n) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

كون نتائجها غير منحصرة على القيمتين (0 و1) بل مجالها مفتوح على الاعداد الحقيقية R.

بحيث x هي عبارة عن الميزات في امثلة التدريب، و θهي عبارة عن الاوزان (والتي سنسعى لإيجادها)

بل يتم استخدام دالة أخرى تدعى بدالة (Sigmoid) والتي تقوم بإخراج قيم محصورة بين الصفر والواحد. [3]:

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



الصورة 4: دالة sigmoid

فتصبح دالة الفرضية كالتالي:

$$h(x_1.x_2 \dots x_n) = \frac{1}{1 + e^{-(\theta_0+\theta_1x_1+\theta_2x_2+\cdots+\theta_nx_n)}}$$

بعد دلك نقوم بالبحث عن قيم اوزان تتناسب جيدا مع بيانات التدريب باستخدام النموذج التالي:

$$J(\theta_{0,}\theta_1.\theta_2 \dots \theta_n) = -\frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \log\left(h_\theta(x^{(i)})\right) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))\right]$$

بحيث y هي قيم المخرجات الحقيقية المقدمة مع امثلة التدريب والتي تأخذ قيمة من القيمتين (0,1)

بدلا من دالة الخطأ التربيعي التي تستخدم في الانحدار الخطي:

$$J(\theta_{0,}\theta_1.\theta_2 \dots \theta_n) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2\right]$$

سعيا لتقليل نسبة الخطأ J.[3]

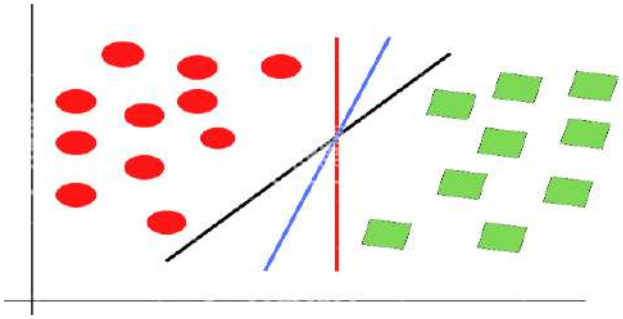
هناك عدة طرق لتقليل قيمة J واحد أكثر الطرق شيوعا طريقة تكرار تغيير قيم الاوزان الى ان تصل قيمة الخطأ الى أدني قيمة ممكنة (global minimum). والتي تصل عادة في عدد قليل من التكرارات. بحيث يتم تغيير قيم الاوزان بالشكل التالي:

$$\theta_j := \theta_j - \alpha \sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)x_j{}^{(i)}$$

في الفئات المتعددة بإمكاننا استخدام نفس الفكرة لكن كل فئة لوحدها وللحصول على الاحتمال المناسب من الضروري اقران النماذج الفردية لكل فئة، ينتج عن دلك مشكلة في التحسين المشترك ولكن هناك عدة طرق فعالة لحلها.[3]

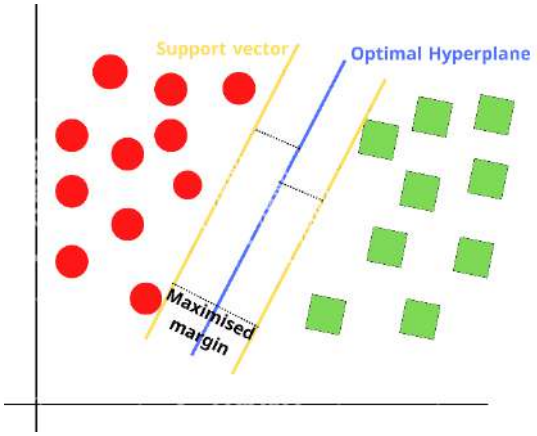
2.3. الدعم الآلي المتجه (Support Vector Machines):

يعتبر الدعم الآلي المتجه أحد أفضل خوارزميات التعلم الآلي في مجال التصنيف كونه يعثر في اغلب الاحيان على الحل الأمثل في تصنيف البيانات **[2].** لشرح الدعم الالي المتجه نضع مثال بسيط لمجموعة من الدوائر والمربعات تمثل مجموعة من البيانات كما هو موضح في الصورة (5).



الصورة 5: مجموعة بيانات مصنفة الى قسمين باستخدام الانحدار اللوجيستي

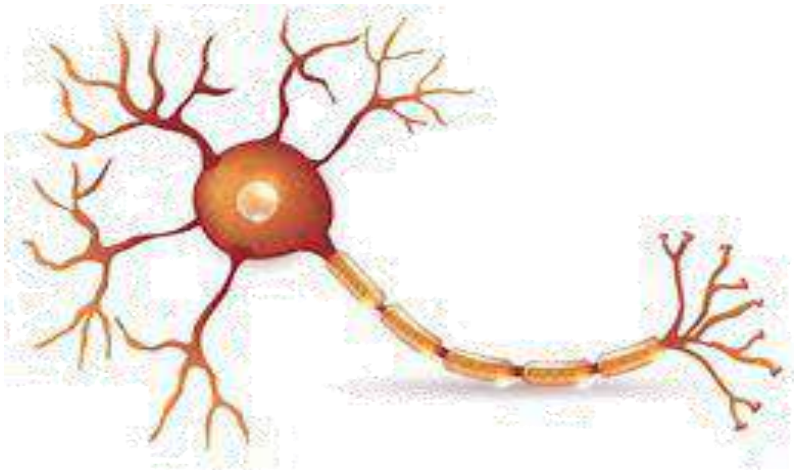
نلاحظ وجود أكثر من خط لفصل الشكلين عن بعضهم البعض لكن أيهم افضل خط، تمكننا خوارزمية الدعم الألي المتجه SVM من العثور على افضل خط والذي يطلق عليه اسم المستوي (hyperplan) بحيث يزيد هامشه إلى اقصي الحدين لفصل الشكلين عن بعضهم البعض كما هو موضح في الصورة (6) , تعرف نهاية كل حد بالمتجهات الداعمة (support vectors) من هذا المنطلق ينحصر الهدف من خوارزمية SVM على التضخيم في طول الهامش للوصول إلى احسن مستوى (optimal hyperplan) يسمي في بعض الاحيان الدعم الألي المتجه بالهامش الأقصى الفاصل (maximum margin separator) .[2](



الصورة 6: مجموعة بيانات مصنفة بخوارزمية الدعم الآلي المتجه SVM

3.3. الشبكات العصبية (Neural Network):

فكرة الشبكات العصبية أتت من الدماغ البشري، حيث ان الدماغ البشري يتكون من الملايين من الخلايا العصبية المربوطة ببعضها البعض. الخلايا العصبية تستقبل عدة رسائل عصبية كهربائية ومن خلالها ترسل رسالة عصبية هي بدورها، حيث نهاية كل خلية عصبية مربوطة بخلية اخرى.[7]

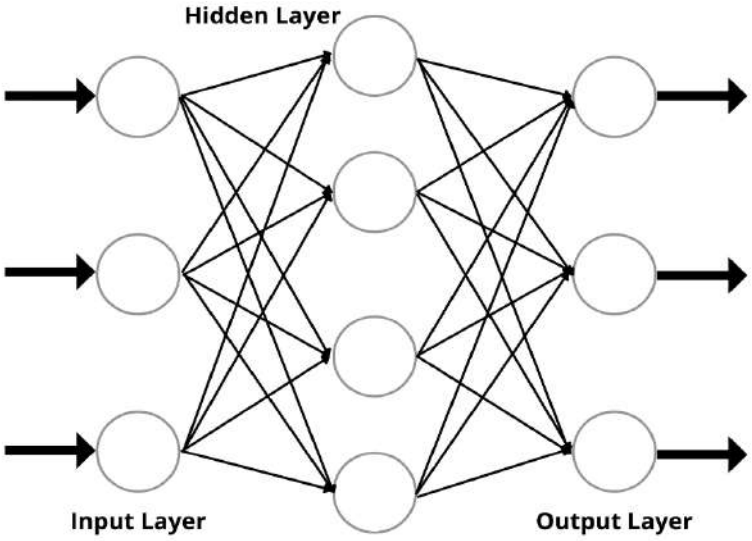


الصورة 7: خلية عصبية

الخطوة الأولى لظهور خوارزمية الشبكات العصبية كانت سنة 1943 وكان ذلك على يد عالم الفسيولوجيا العصبية "والتر بيتس"(Walter Pitts) حيث نشر ورقة بحثية بين فيها كيفية عمل الخلايا العصبية. تم نمذجة اول شبكة عصبية بسيطة مع دوائر كهربائية في عام 1949، كما عزز دونالد هب (Donald Hupp) مفهوم الخلايا العصبية في كتابه تنظيم السلوك.[7]

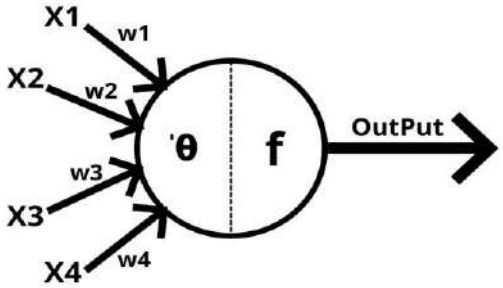
تتمحور فكرة خوارزمية الشبكات العصبية في ثلاث طبقات متصلة وهي طبقة الإدخال، طبقة الإخراج، الطبقات المخفية قد تكون واحدة او أكثر.

تتكون كل طبقة من مجموعة من العقد المتصلة بالعقد في الطبقة التي قبلها وهكذا، الصورة التالية توضح كيفية ترابط خلايا الشبكة العصبية:



الصورة 8: طبقات الشبكة العصبية

كل عقدة مربوطة بكل العقد من الطبقة التي تسبقها بواسطة أوزان (w) تمثل هاته الأوزان قوة الربط بين العقدتين، كما تحمل كل عقدة قيمة بين 0 و1 تمثل قيمة التنشيط.[7]



الصورة 9: عقدة لشبكة عصبية

كما توجد مجموعة من الخوارزميات الأخرى أيضا مثل:

- **الغابة العشوائية (Random Forest):** وهي أداة النمذجة التنبؤية القائمة على التعلم الخاضع للإشراف تعمل على مبدأ التحليل متعدد المتغيرات بتقسيم البيانات الى أساس متعدد والذي يساعد على التنبؤ. حيث تقوم ببناء عدد كبير من اشجار القرار المتخصصة ثم تجميع مخرجاتها.
- **أشجار القرار (Decision Trees):** أشجار القرار هي هياكل تشبه المخططات المخفية تتيح لك تصنيف نقاط بيانات الادخال او توقع قيم الإخراج المعطاة للمدخلات.
- **الجيران الأقربKNN (k-Nearest Neighbors):** تعتبر من أكثر خوارزميات التصنيف انتشارا بحيث تقوم هذه الخوارزمية بالبحث عن أقرب الجيران لنقطة معينة من البيانات ويحدد عدد الجيران على حسب العدد k ويتم تحديد الفئة الغالبة عند الجيران k كنتيجة للنقطة المحددة.
- **Naive Bayes:** هو نوع من مصنفات التعلم الآلي الذي يعتمد على تطبيق نظرية بايز (Bayes):[5]

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

مع افتراض أن الميزات في بيانات الإدخال كلها مستقلة (قوية). وهو أحد أهم الا الأنواع المستخدمة في تحليل البيانات.

المحور الثالث | برمجة النموذج

Model programming

1. الأدوات المستخدمة

استخدمنا خوارزمية الانحدار اللوجستي التي قمنا ببنائها كما استعملنا الخوارزمية الجاهزة في حزمة sklearn وبعض خوارزميات التصنيف الأخرى لمقارنة نتائج نموذجنا بها.

استعملنا أيضا:

- حزمة Python Numpy وذلك للتعامل مع القوائم والمصفوفات بكفاءة وسرعة.
- حزمة Pandas ووحدة CSV ووحدة Json لغرض استيراد وحفظ البيانات.
- قسم pyplotمن Matplotlib بالإضافة الى seaborn لرسم الرسوم البيانية.
- منصة GitHub لحفظ ومراقبة تقدم المشروع.
- موقع Spell كخادم لتنفيذ البرنامج (أثناء مرحلة التعلم) عوضا عن الحاسوب الشخصي كون البرنامج يستغرق وقت طويل خلال التنفيذ.

2. البيانات

1.2. مصدر البيانات

في هذا العمل تم استخدام البيانات من موقع Kaggle [8]. بحيث تحتوي البيانات على 18 سمة وإجمالي عدد أمثلة التدريب 319795 شخص تم استخدام حوالي % 17 منها.

2.2. وصف البيانات

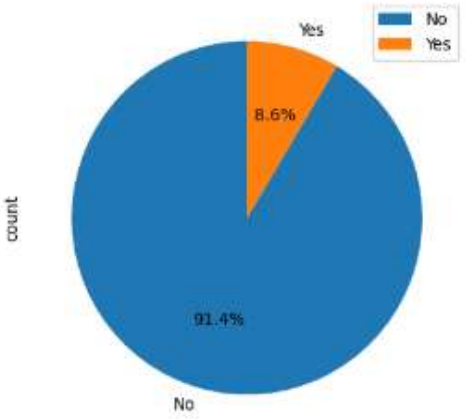
العنصر	الوصف	النوع
Heart Disease	المستجيبون الذين أبلغوا عن إصابتهم بأمراض القلب التاجية (CHD) أو احتشاء عضلة القلب (MI)	Boolean
BMI	مؤشر كتلة الجسم (BMI)	Float
Smoking	هل دخنت ما لا يقل عن 100 سيجارة طوال حياتك؟ [ملاحظة: 5 علب = 100 سيجارة]	Boolean
Alcohol Drinking	من يشربون الخمر (الرجال البالغون يتناولون أكثر من 14 مشروبًا في الأسبوع والنساء البالغات يتناولن أكثر من 7 مشروبات في الأسبوع	Boolean
Stroke	الشخص أصيب بسكتة دماغية من قبل ام لا	Boolean
Physical Health	صحة الشخص الجسدية خلال الأيام الماضية	Float
Mental Health	صحة الشخص العقلية خلال الأيام الماضية	Float
Diff Walking	هل تعاني من صعوبة بالغة في المشي أو صعود السلالم؟	Boolean
Sex	الجنس ذكر ام انثى	String
Age Category	الفئة العمرية التي ينتمي اليها الشخص يوجد أربعة عشر مستوى	String
Race	قيمة العرق / الإثنية المتنازع عليها	String
Diabetic	الشخص أصيب بالسكري ام لا	String
Physical Activity	الشخص يمارس الرياضة ام لا	Boolean
Gen Health	الحالة الصحية	String
Sleep Time	متوسط نوم الشخص في اليوم بالساعات	Float
Asthma	الشخص اصيب بالربو أم لا	Boolean
Kidney Disease	أمراض الكلى لا تشمل حصوات الكلى أو عدوى المثانة أو سلس البول	Boolean
Skin Cancer	الشخص اصيب بمرض السرطان أم لا	Boolean

الجدول 1: وصف عناصر البيانات

3.2. تصوير وملاحظة البيانات (DATA VISUALISATION):

- **قبل التسوية:**
  - Heart Disease

توضح الصورة المقابلة نسبة الأشخاص المصابين والغير مصابين بالمرض في البيانات(المخرجات). حيث نلاحظ وجود تباين كبير بين المخرجات نسبة الأشخاص المصابين 8.6% (27 502 شخص) بينما نسبة الأشخاص غير المصابين 91.4% (292 293 شخص).



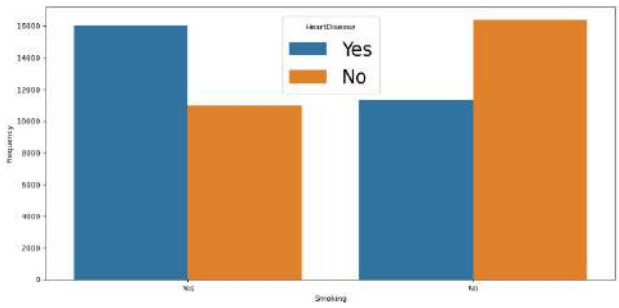
الصورة 10: نسبة المصابين وغير المصابين بأمراض القلب قبل تسوية البيانات

- **بعد التسوية:**
  - Heart Disease

بعد تسوية البيانات نلاحظ أن قيم البيانات أصبحت متساوية تقريبا ما يعني ان النموذج عند بناءه لن ينحاز الى أي فئة.

- Smoking

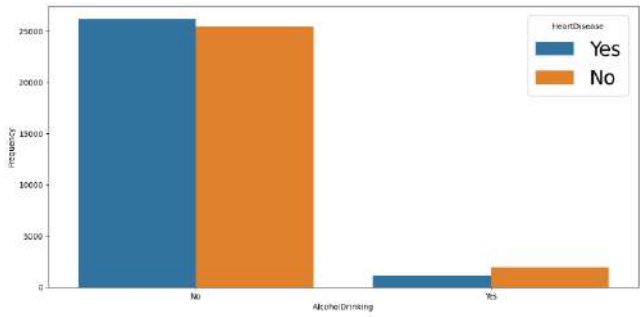
تمثل الصورة المقابلة نسبة حالات الاصابة بالنسبة للأشخاص المدخنين وغير المدخنين حيث نلاحظ على اليساران المدخنين أكثر عرضة للإصابة بالنسبة لغير المدخنين كماهو موضح على الجانب الأيمن من الصورة، مما يدل على أن التدخين يؤثر بشكل كبير على الإصابة بأمراض القلب.



الصورة 11: أمراض القلب حسب التدخين

- AlcoholDrinking

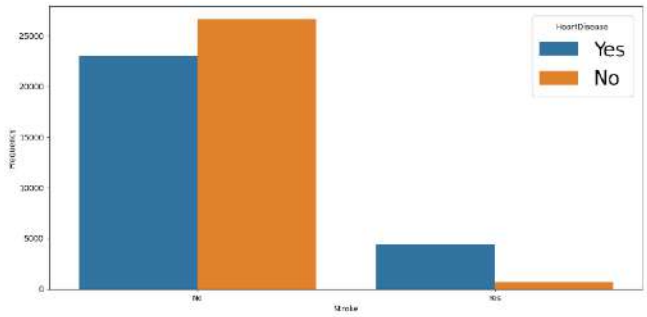
تمثل الصورة المرفقة نسبة إصابة الأشخاص بمرض القلب حسب حالات شرب الكحول حيث نلاحظ على الجانب الأيمن من الصورة أن الأشخاص الدين يشربون الكحول أقل إصابة من الأشخاص الدين لا يشربون الكحول الموضحين على الجزء الأيسر من الصورة مما يدل على أن الكحول له تأثير إيجابي على القلب.



الصورة 12: أمراض القلب حسب شرب الخمر

- Stroke

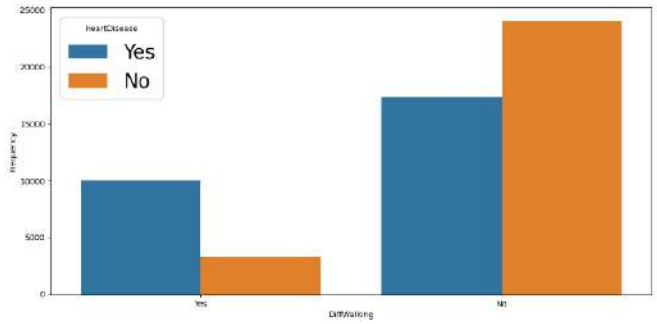
نلاحظ من خلال الصورة التالية أن الأشخاص المصابين بسكتة دماغية هم أكثر الأشخاص عرضة للإصابة بأمراض القلب مقارنة بالأشخاص الاخرين ما يدل على ان السكتة الدماغية لها تأثير كبير على أمراض القلب.



الصورة13: أمراض القلب حسب السكتة الدماغية

- DiffWalking

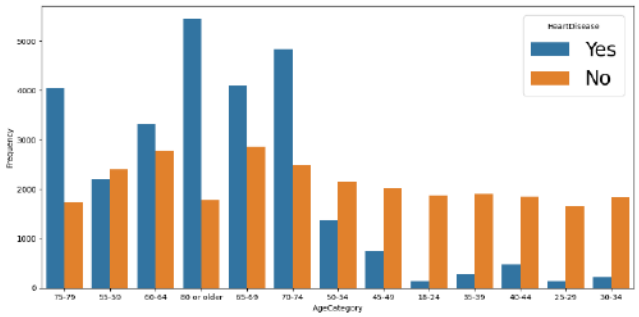
من خلال الصورة المقابلة نلاحظ أنه كلما كان الشخص يواجه صعوبة في المشي كلما زاد احتمال إصابة الشخص بأمراض القلب.



الصورة 14: أمراض القلب حسب الصعوبة في المشي

- AgeCaegory

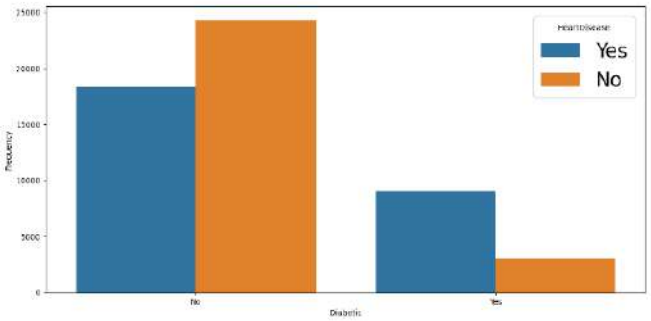
تمثل الصورة نسبة إصابة الأشخاص بأمراض القلب حيث نلاحظ أنه كلما تقدم الشخص في السن كلما زاد احتمال إصابته بأمراض القلب.



الصورة 15: أمراض القلب حسب الفئات العمرية

- Diabetic

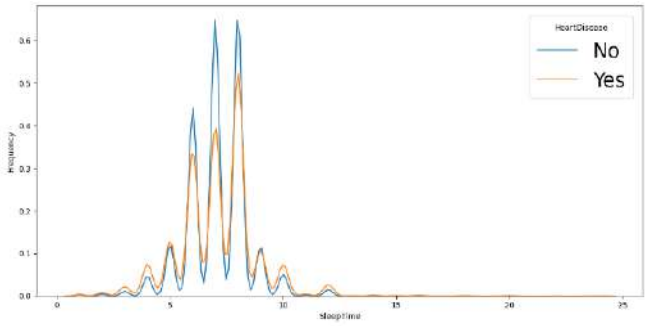
حيث نلاحظ أن الأشخاص المريضين بالسكري أكثر عرضة للإصابة من الأشخاص السليمين من السكر، ما يدل على أن السكري له تأثير مباشر على أمراض القلب.



الصورة 16: أمراض القلب حسب مرض السكري

- SleepTime

نلاحظ من خلال الصورة المرفقة أن الأشخاص الذين ينامون من ستة الى ثماني ساعات في اليوم هم الأقل عرضة للإصابة من أمراض القلب مقارنة بغيرهم من الأشخاص، مما يعني كذلك أن زيادة أوقلة ساعات النوم عن الحجم المفروض له تأثير سلبي على القلب.



الصورة 17: أمراض القلب حسب عدد ساعات النوم في اليوم

نلاحظ أيضا من خلال تحليل البيانات ان السكان البيض وذوي الأصول الإسبانية أكثر عرضة للإصابة من بقية السكان الآخرين وان الرجال أكثر عرضة للإصابة من النساء وان الأشخاص اللذين يمارسون الرياضة واللذين هم في صحة جيدة هم الفئة الأقل عرضة للإصابة. أما بالنسبة لبقية العناصر فهي تؤثر بنسبة قليلة على الإصابة بأمراض القلب.

4.2. تهيئة البيانات:

من خلال صور البيانات الموضحة في العنصر السابق (data visualisation) يمكننا ملاحظة من الصورة (10) أن غالبية امثلة التدريب من الفئة 0 وقليل منها من الفئة 1,إذا استخدمنا هذا التوزيع لتطوير نموذجنا فقد يصبح منحازًا للتنبؤ بفئة الأغلبية 0 نظرًا لعدم وجود بيانات كافية لتعلم أنماط فئة الأقلية 1,فيصبح النموذج يتنبأ بكل الأشخاص على انهم من الفئة 0 (class imbalance)، وهذه مشكلة بالرغم من ان دقة النموذج هنا جيدة جدا (حوالي 91%) الا انه يعتبر نموذج سيء لا فائدة منه كونه لن يتنبأ بفئة الأقلية 1 والتي تعتبر فئة مهمة. لذلك قمنا بتقليل أمثلة فئة الأغلبية لضمان أن كلا الفئتين لهما نفس التوزيع.fifty-fifty

⇦ **ترميز العناصر:** تحتوي البيانات على مجموعة من العناصر ذات قيم غير رقمية، يجب تحويلها الى قيم رقمية ليتم التعامل معها في المعادلات الموجودة في النموذج. يمثل الجدول التالي كيفية قيامنا بتحويل القيم.

العنصر	الترميز	القيمة الأصلية
HeartDisease,Smoking, Alcohol Drinking, Stroke, Diff Walking, Diabetic, Physical Activity, Asthma, Kidney Disease, Skin Cancer	1	Yes
	0	No
Six	1	Male
	2	Female
Race	1	White
	2	Black
	3	American Indian/Alaskan Native
	4	Asian
	5	Hispanic
	6	Other
Gen Health	1	Poor
	2	Fair
	3	Good
	4	Very good
	5	Excellent
Age	(a+b)/2	a-b

الجدول 2: ترميز البيانات

3. النموذج والنتائج:

الخوارزمية	نسبة الدقة على بيانات التدريب	نسبة الدقة على بيانات التقييم
الانحدار اللوجيستي الخاصة بنا	74.18 %	74.22 %
الانحدار اللوجيستي الخاصة ب sklearn	76.19 %	76.13 %
الدعم المتجه الالي	75.11 %	75.19 %
الشبكات العصبية	76.36 %	76.53 %
الغابة العشوائية	99.77 %	74.26 %
أشجار القرار	99.77 %	67.31 %
الجيران الأقرب	78.27 %	71.49 %
المصنف بايز	71.02 %	70.53%

الجدول 3: نتائج خوارزميات التصنيف على البيانات المستخدمة

الخاتمة

قمنا في هذا العمل ببناء نموذج يقوم بالتنبؤ بمرض الشخص بقلبه باستخدام خوارزمية الانحدار اللوجيستي، حيث قمنا ببناء نموذج هذه الخوارزمية ومقارنة نتائجه مع باقي خوارزميات التصنيف الجاهزة في مكتبة sklearn في pythonحيث كانت دقة نموذجنا حوالي 74% وهي نسبة قريبة جدا من نسبة دقة دالة الانحدار اللوجيستي الجاهزة 75%. ما يعني أن نتائج نموذجنا كانت جيدة. لكن المشكلة التي واجهتنا هي طول وقت التنفيذ (25 يوم) وعدم وجود أجهزة قوية لتدريب النموذج، لذلك سنسعى في المستقبل للتقليل في وقت التنفيذ واستخدام التعلم العميق وبعض التقنيات الأخرى التي قد تعطينا دقة أفضل للنموذج.

المراجع

[1] Nicole Tache, editor. Hands-on Machine Learning with Scikit-Learn,Keras, and TensorFlow. Aur´elien G´eron, second edition, June 2019.

[2] Stuart J. Russell and Peter Norvig. Artificial Intelligence a Modern Approach, Third edition, 2010.

[3] Lan h. witten,Eibe Frank,Mark A.Hall, Data Mining Practical Machine Learning Tools and Techniques Third Edition.

[4] Joseph M. Hilbe , Practical Guide to Logistic Regression 2015

[5] Tiffany Taylor, editor. Deep Learning with Python. Francois Chollet, 2018.

[6] Nicole Tache, editor. Hands-on Machine Learning with Scikit-Learn,Keras, and TensorFlow. Aur´elien G´eron, second edition, June 2019.

[7] Abraham, Ajith. "Artificial neural networks." Handbook of measuring system design (2005).

[8] https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease in 20/04/2022

الملخص

تعد أمراض القلب من أكثر الأمراض شيوعا في العالم اليوم وأكثرها رعبا كونها تخلف الكثير من الوفيات. نحن الآن في عصر البيانات حيث يتم جمع وتخزين كميات هائلة من البيانات من مختلف أنحاء العالم في الأنترنت (من الشركات ومواقع التواصل الاجتماعي والعيادات الطبية وغيرها...). وفي العمل البحثي هذا نسعى لبناء نموذج يقوم بالتنبؤ بأمراض القلب لتحسين التشخيص العام المبكر حولها، وذلك باستخدام الانحدار اللوجيستي كنوع من أنواع التصنيف بدون الاستعانة بمجموعة الدوال الجاهزة الخاصة بالتعلم الآلي، وذلك انطلاقا من مجموعة من بيانات تتكون من 319795 مثال يساعد على تدريب النموذج ومقارنة النتائج المتحصل عليها بنتائج خوارزمية sklearn الخاصة بالانحدار اللوجيستي وببعض خوارزميات التصنيف الأخرى. (SVM(SVC), NN, RF, KNN, DT, NBC) بحيث كانت نسبة الدقة في نموذجنا جيدة (74%) وهي نسبة قريبة جدا من بقية خوارزميات sklearn الخاصة بالتصنيف التي كانت تتراوح نسبة الدقة فيهم بين 70 %و 76%.

الكلمات المفتاحية: أمراض القلب، التعلم الالي، التصنيف، الانحدار اللوجيستي، الدعم المتجه الالي، الشبكات العصبية الاصطناعية.

Abstract

Heart disease is one of the most common diseases in the world today and the most terrifying as it causes many deaths. We are now in the data age where huge amounts of data are collected and stored in the internet from different parts of the world (from companies, social networking sites, medical clinics, etc...). In this research work, we seek to build a model that predicts heart disease to improve early general diagnosis around it, that is, using logistic regression as a kind of classification without the help of a set of ready-made functions for machine learning, and that is based on a set of data consisting of 319,795 examples that help train the model and compare the results of Model with sklearn logistic regression algorithm and some other classification algorithms (SVM(SVC), NN, RF, KNN, DT, NBC). where the accuracy rate in our model was good (74%), which is very close to the rest of the sklearn classification algorithms, which were between 70% and 76% accuracy.

Keywords: Heart Disease, Machine Learning, Classification, Logistic Regression, Support Vector Machine Artificial Neural Networks.