|  |  |  |
| --- | --- | --- |
| **Syrian Arab Republic** |  | **الجمهورية العربية السورية** |
| **Ministry of Higher Education** | **وزارة التعـليــم العـالـــــــــي** |
| **Syrian Virtual University** | **الجامعة الافتراضية السورية** |
| **MCS** |  | **ماستر علوم الحاسب** |

**عنوان المشروع:**

**التنبؤ بمرض القلب بالاعتماد على خوارزميات التنقيب عن البيانات**

**Heart Disease Prediction**

**إعداد الطالب:**

**زين العابدين موريس موسى**

[**Zain\_alabideen\_242399@svuonline.org**](mailto:Zain_alabideen_242399@svuonline.org)

**إشراف الدكتور:**

**عصام سلمان**

[**T\_isalman@svuonline.org**](mailto:T_isalman@svuonline.org)

**جدول المحتويات**

[المقدمة: 3](#_Toc203306029)

[الكلمات المفتاحية: 3](#_Toc203306030)

[أولا: فكرة موسعة عن المشروع: 4](#_Toc203306031)

[ثانيا: اهداف المشروع: 5](#_Toc203306032)

[ثالثا: المتطلبات الوظيفية: 5](#_Toc203306033)

[رابعا: المتطلبات غير الوظيفية: 6](#_Toc203306034)

[خامسا: آفاق مستقبلية للمشروع: 8](#_Toc203306035)

[سادسا: مراحل عمل المشروع: 8](#_Toc203306036)

[سابعا: الادوات البرمجية المستخدمة في عمل المشروع: 9](#_Toc203306037)

[ثامنا: الدراسات المرجعية: 9](#_Toc203306038)

[الفصل الاول: تحليل النظام والمنهجية المقترحة: 12](#_Toc203306039)

[**1.1.** **النموذج المقترح:** 12](#_Toc203306041)

[**1.2.** **الأدوات البرمجية:** 13](#_Toc203306042)

[**1.3.** **منهجية جمع البيانات ومعالجتها:** 14](#_Toc203306043)

[**1.3.1.** **جمع البيانات Collect data:** 14](#_Toc203306044)

[**1.3.2.** **مجموعات البيانات قبل المعالجة:** 14](#_Toc203306045)

[**1.3.3.** **استخراج المعرفة knowledge extraction:** 14](#_Toc203306046)

[**1.3.4.** **تنظيف البيانات data cleaning:** 15](#_Toc203306047)

[**1.4.** **النماذج تعلم الالة المستخدمة:** 16](#_Toc203306048)

[**1.4.1.** **نموذج آلة المتجه الداعم :Support Vector Machine:** 16](#_Toc203306049)

[**1.4.2.** **نموذج شجرة القرار Decision Trees model:** 18](#_Toc203306050)

[**1.4.3.** **نموذج الجار الأقرب KNN model:** 18](#_Toc203306051)

[**1.4.4.** **نموذج الانحدار الخطي Linear Regression model:** 19](#_Toc203306052)

[**1.5.** **التقييم:** 20](#_Toc203306053)

[**1.6.** **التصنيف والتنبؤ:** 23](#_Toc203306054)

# المقدمة:

الطب هو مجال علاجي يشمل الممرضين والأطباء بمجالات مختلفة، ويغطي التشخيص، والعلاج، والوقاية من الأمراض، والبحوث الطبية، والعديد من الجوانب الأخرى للصحة، ويهدف إلى تعزيز الصحة والحفاظ عليها، وهي تنطوي عن استخدام العقاقير أو الجراحة، وغالبًا ما تدعمها تدابير المشورة ونمط الحياة، وتشمل الأنواع البديلة والتكميلية من الطب الوخز بالإبر، والمعالجة المثلية، والأدوية العشبية، والعلاج بالفن، والطب الصيني التقليدي، وغيرها الكثير

إنّ التطور السريع الآخذ بالتزايد والحاجة الملحة إلى استخدام الحاسوب في حياتنا اليومية لتسهيل الكثير من المهام التي تطلبت فيما مضى الكثير من الوقت والجهد وصل إلى مجال الطب، الأمر الذي جعل المجال الطبي يعتمد بشكل ملح على أجهزة الكمبيوتر لتقديم أكثر أشكال التشخيص العلاج تقدمًا للمرضى، فقد أصبح من الممكن مشاهدة تأثير الحاسوب في العديد من المجالات الطبية، مثل أنظمة التصوير والمختبرات السريرية وتخزين السجلات الطبية وعمليات المستشفيات وغيرها من المجالات الطبية الأخرى.

# 

# الكلمات المفتاحية:

|  |
| --- |
| * **Bayes algorithms**   هي عائلة من المصنفات الاحتمالية البسيطة على أساس تطبيق نظرية بايز مع افتراضات استقلالية (ساذجة) قوية بين الميزات.   * **KNN algorithms**   تقوم خوارزمية KNN بعملية تصنيف نقطة الاختبار test point اعتمادا على نقاط التدريب المحيطة بها، أي الجيران الأقرب لنقطة الاختبار لذلك سميت هذه الخوارزمية neighbors nearest   * **data mining**   تنقيب البيانات هو مجال يجمع عدة تقنيات من عدة تخصصات مثل تعليم الآلة، الإحصاء، التعرف على الأنماط، قواعد البيانات والإظهار المرئي، يهدف هذا العلم إلى استخراج المعلومات من البيانات الضخمة المخزنة والمجمعة في المستودعات |

# أولا: فكرة موسعة عن المشروع:

عمل الباحثون منذ سنوات عديدة للبحث عن مسببات الأمراض وكيفية النجاة منها، وقد برز مؤخراً اثر العلوم المعلوماتية على المجال الطبي، وخاصة الدور الذي لعبته تقانات الذكاء الصنعي في حل المشاكل الطبية ومن بينها خوارزميات التنقيب عن البيانات في العلوم الطبية.

وقد تمكنت التكنولوجيا من توفير الات جديدة وأدوية وعلاجات ساعدت في حماية العديد من الأشخاص وحسنت من فرص علاجهم، ليس هذا وحسب، بل استطاعت التكنولوجيا تحسين الأبحاث والدراسات العلمية لجعل الرعاية الصحية والتشخيص أكثر كفاءة.

وبالتالي فان السبب الرئيسي الذي يجعل استخراج المعلومات هو محور الاهتمام في العلوم الطبية ومن بينها أمراض القلب هي توافر كميات ضخمة من البيانات والحاجة الملحة لاستخراج المعلومات والمعرفة منها. واستخراج البيانات هو تكييف أو استخراج المعرفة من مجموعة من البيانات ، ويمكن أن يعزى استخراج البيانات إلى التطور الطبيعي لتكنولوجيا المعلومات ، والذي هو نتيجة تطور آلية بناء قواعد البيانات ، مثل: عمليات جمع البيانات وإنشاء قاعدة البيانات.

وان مرض القلب يعتبر وصف لنطاق من الأمراض التي تصيب القلب. وتشمل الأمراض التي تندرج تحت طائلة مرض القلب أمراض الأوعية الدموية، مثل مرض الشريان التاجي، والمشاكل في نظم القلب (اضطراب نظم القلب) وعيوب القلب التي يولد المرء بها (عيوب القلب الخلقية) من بين أمراض أخرى.

وتعتمد الفحوص الواجب إجرائها لتشخيص مرض القلب على الحالة التي يعتقد الطبيب أن الشخص مصاب بها. بصرف النظر عن مرض القلب الذي قد يكون مُصابًا به، من المرجح أن يجرى الطبيب فحصًا جسديًا وأن يسأل عن التاريخ الطبي للمريض قبل إجراء أي فحوصات. بجانب اختبارات الدم والأشعة السينية على الصدر.

ونظراً لأهمية الكشف المبكر عن وجود مشكلة صحية في الجسم وتحديد ماهيتها والذي يلعب دوراً هاماً في سرعة العلاج وعاملاً مؤثر في الشفاء وهنا تكمن أهمية الحاجة إلى التشخيص الدقيق والسريع لوجود المرض وتحديد درجة الحالة المرضية.

وبالإضافة الى الحاجة عن الكشف المبكر هنالك حاجة الى تشخيص المرض بشكل صحيح وسريع اعتمادا على بيانات المريض واعطائه نتائج واضحة في هذا المجال وانطلاقا من هذه الاسباب نسعى لبناء نظام قادر ذكي باستخدام خوارزميات التنقيب عن البيانات قادر على تشخيص الاصابة بأمراض القلب بالاعتماد على مجموعة كبيرة من البيانات التي سوف نعمل على جمعها من خلال الجولات على بعض المراكز الصحية والمستشفيات ذات العلاقة وبالتالي تحديد الحالة المرضية للمريض

# ثانيا: اهداف المشروع:

تأتي الأمراض القلبية الوعائية في صدارة أسباب الوفيات في جميع أنحاء العالم، ذلك أنّ عدد الوفيات الناجمة عن هذه الأمراض يفوق 17.7 مليون نسمة، ممّا يمثّل 17 % من مجموع وفيات العالم. كان لابد من الاهتمام بهذا الجانب للمساهمة في تشخيص وعلاج هذا النوع من الأمراض

يهدف عملنا على جمع عدد كبير من البيانات الأولية المبنية على التحاليل والتقارير الطبية المُوصفة لحالة المرضى ثم العمل على إعدادها وترتيبها لنعمل على دراستها بالاعتماد على خوارزميات التنقيب عن البيانات لاستخراج المعلومات ومقارنة النتائج فيما بينها والهدف تحسين تشخيص أمراض القلب والتنبؤ بحالة المريض بشكل سليم وبالتالي تحسين عملية بناء القرار .

ويمكن تلخيص أهداف البحث بما يلي:

* تشخيص مرض القلب بالاعتماد على خوارزميات التنقيب عن البيانات
* تطوير خوارزميات التعلم الآلي في مجال أمراض القلب ومعالجتها
* المساهمة في التشخيص الطبي السليم باستخدام خوارزميات التنقيب عن البيانات
* بناء نظام ذكي بالاعتماد على قادر على تشخيص حالة المريض من خلال المعرفة المكتشفة
* تصميم وبناء تطبيق ويب يعمل وفق خوارزميات التنقيب عن البيانات في التعامل مع مرضى القلب

# ثالثا: المتطلبات الوظيفية:

ان المتطلبات الوظيفية لهذا المشروع يمكن إنجازها من خلال شيفرة برمجية للتنبؤ بمرض القلب عرض نتائج الفحص والتشخيص على شكل مخططات بيانية تظهر احتمالات الاصابة وخطأ التصنيف ودراسة نظرية شاملة لأدوات التنقيب عن البيانات والخوارزميات المراد استخدامها ضمن التطبيق كخوارزمية الاحتمال الشرطي BAYES وخوارزمية الجار الاقرب KNN.

وسوف يتم الاعتماد في عملية التنقيب على مجموعة من بيانات التدريب لحالات من المصابين بمرض القلب حيث تحتوي عينة التدريب والتنبؤ على سجلات بمقدار 200 سجل بحالتين مختلفتين تتراوح قيمها بين النتيجة الايجابية والنتيجة السلبية وتحتوي هذه العينة على بارامترات محددة تمثل هذه البارامترات السمات العامة للمرض

# رابعا: المتطلبات غير الوظيفية:

بالنسبة لمستخدمين هذا المشروع يجب أن تتوفر لديهم المتطلبات التالية:

# خامسا: آفاق مستقبلية للمشروع:

* هناك العديد من الإضافات التي يمكن إضافتها إلى هذا المشروع والتي قد تسمح له بأن يصبح أكثر احترافية وعملية. انطلاقًا من الحاجة إلى البيانات الإضافية اللازمة لتدريب النماذج بشكل أفضل.
* من الممكن إضافة خاصية وضع صفحة الويب على الاستضافة الإلكترونية وربطها بالنظام الصحي الإلكتروني ليتمكن الشخص من استخدام التطبيق بشكل تشاركي وسريع.
* وهناك ميزة اخرى ايضا وهي توسيع عمل التطبيق بحيث نقوم ببناء موقع كامل لاحد المستشفيات وامكانية تقديم الطلبات وادارتها من خلال هذا الموقع ومن ثم استخدام النموذج الافضل الذي اخترناه في عملية الكشف عن مرض القلب بشكل آلي ومباشر دون الرجوع للأطباء المسؤولين مما يساهم في توفير الوقت والجهد في عملية دراسة الحالة الصحية المقدمة من المريض.

# سادسا: مراحل عمل المشروع:

1. **اجراء دراسة مرجعية عن خوارزميات التنقيب المقترحة**
2. **دراسة العينات المقترحة وتصديرها الى قاعدة البينات**
3. **تحديد حالات الاستخدام للموقع وتجهيز الدارسة التحليلية والمخططات اللازمة**
4. **اجراء الدراسة التصميمية لقاعدة البينات ورسم المخططات اللازمة**
5. **تصميم الواجهات**
6. **البدء ببرمجة الموقع وكتابة الأكواد الخاصة بالخوارزميات المطلوبة**

# 

# سابعا : الادوات البرمجية المستخدمة في عمل المشروع:

# ثامنا: الدراسات المرجعية:

توجه العديد من الباحثين في الآونة الأخيرة لاستخدام تقانات الذكاء الصنعي وأدواتها في المجال الطبي ومن العلوم الطبية التي شغلت الكثيرين من العلماء أمراض القلب الوعائية

* قام الباحث Ameer H. Ali وآخرون عام 2018 ببناء شبكة عصبونية ذكية الهدف منها تشخيص الأمراض السرطانية بالإعتماد على ست ميّزات هامة تم استخلاصها من الصور الشعاعية المأخوذة للمريض بمرض السرطان ومعالجتها بعمليات حسابية لتحويلها لدخل جيد للشبكة العصبونية المصممة. الهدف من الدراسة كان الكشف عن مرض السرطان وفقا للمعايير التي وضعتها الكلية الأمريكية لأمراض الروماتيزم. استخدم الباحثون شبكة عصبونية حيث كان عدد خلايا طبقة الدخل فيها ستة وكان خطأ التصنيف يساوي 3.8968\*10^-10. عمل الباحث على التحكم بعدد الميزات وبالتالي عدد خلايا طبقة الدخل للشبكة العصبونية فاستخدم خمسة ثم ثمانية فكان خطأ دقة التصنيف 0.0041 و1.0611\*10^-10. وهنا أمكننا القول أن جميع النتائج مقبولة فالخطأ في التصنيف صغير جدا [[1]](#footnote-1)
* أما في عام 2011، قام الباحث Jyoti Soni وآخرون بتقديم مسح للتقنيات الحالية لاكتشاف المعرفة من قواعد البيانات الضخمة باستخدام تقنيات استخراج المعلومات حيث استخدم الباحثون ثلاث عشرة ميزة هامة تصف حالة وأعراض مريض القلب, وقد قام الباحثون بعدد من التجارب باستخدام خوارزمية شجرة القرار فيها فكان معدل التصنيف 89% وقد تفوقت هذه الخوارزمية في أدائها على خوارزمية تصنيف بايز والتي كانت دقة التصنيف فيها 86.53 %. بينما كانت دقة التصنيف في خوارزمية الجار الأقرب 85.53 %[[2]](#footnote-2).
* وفي دراسة للباحث Dr. G. Karraz   واخرون عام 2006، بعنوان "التصنيف التلقائي لنبضات القلب باستخدام تصنيف الشبكة العصبية بناءً على إطار بايزيان ".حيث قام الباحث بتصميم خوارزمية للمعالجة التلقائية لإشارة مخطط كهربائية القلب (ECG) والهدف هو تصنيف دقات القلب . جمع الباحث البيانات من قاعدة البيانات MIT-BIH حيث عدد السجلات المدروسة هو / 48 / حالة مرضية تعاني من عدم انتظام ضربات القلب. درس الباحث خمسة أنواع من التصنيفات لعدم انتظام ضربات القلب وفق الخوارزمية المصممة، حيث كان النوع الأول هو ضربات التوصيل البطيني المبكرة (PVC) ، والنوع الثاني هو ضربات التوصيل الأذيني المبكر (APC) ، والنوع الثالث هو ضربات كتلة الفرع الأيمن (RBBB) ، أما النوع الرابع فهو ضربات كتلة الفرع الأيسر (LBBB) ، وآخيرا النوع الخامس هو ضربات إيقاع السرعة ( PRB) ،مضافاً الى ذلك النوع النظامي هو ضربات القلب النظامية . استخدم الباحث في مرحلة التدريب بيانات خمس سجلات وهي على التوالي (124 و214 و111 و100 و107).  تم قام بتعيين الميزات الهامة بالاعتماد على مورفولوجيا ECG وعلى الفترات الزمنية. بعد القيام بالعديد من العمليات اللازمة لتكون مناسبة لاستخدامها كبيانات تدريب الشبكة العصبونية . تم تصنيف نوع ضربات القلب بشكل يدوي بالاعتماد على الخبرة المهنية لأخصائي أمراض القلب وعلى دليل قاعدة بيانات MIT-BIH . كانت الدراسة حول تصنيف نوع المرض وهو خرج للشبكة العصبونية المصممة. توصل الباحث وفق النظام المصمم الى الحد الأدنى من الحساسية وبمعدل تصنيف 86٪ والحد الأدنى من الخصوصية بمعدل تعرف 90٪.[[3]](#footnote-3)
* وفي عام 2013، عمل الباحث Ishtake S.H على استخدام تقنيات التنقيب عن البيانات لتشخيص مرض القلب باستخدام 15 ميزة هامة استخدم الباحث في الدراسة خوارزمية بايز فكان معدل التصنيف 95% استخدم الباحث 909 حالات مرضية . كما درس 106 حالة مرضية باستخدام شجرة القرار فكان معدل التصنيف 94.93% بينما عندما استخدم شبكة عصبونية لدراسة 432 حالة مرضية فكان معدل التصنيف 93.54%.[[4]](#footnote-4)
* قدم الباحث Nidhi Bhatla وآخرون عام, 2012 دراسة تهتم بمقارنة نتائج تقنيات تحليل البيانات التي تم تطويرها في السنوات الأخيرة لاستخراج المعلومات المختلفة والتنبؤ بأمراض القلب, استخدم الباحثون كمدخلات للشبكات العصبونية 15 ميزة هامة حيث بلغت دقة التنصيف في حالة الشبكات العصبونية 100% وكانت هذه النبة الأعلى في التصنيف اي كانت الأفضل بين الدراسات مقارنة بخوارزمية Naive Bayes والتي بلغت دقة التصنيف فيها 90.74% وخوارزمية Decision Treesحيث كانت دقة التنصيف فيها 99.62% [[5]](#footnote-5).

وأخيراً، توضح الدراسات السابقة أهمية دمج المعلومات الناتجة عن الفحوصات السريرية مع بيانات سجلات المرضى المجمعة من التقارير الطبية السابقة والحالية والصور الشعاعية والتحاليل الطبية لبناء القرارات الطبية السليمة مبنية على المعرفة التراكمية وهذا الأسلوب يمكن أن يقلل عدد الأخطاء الطبية ، وتعزيز سلامة المرضى ، وتقليل تباين الممارسات غير المرغوب فيها أو القرارات الخاطئة. فقد أظهرت الدراسات السابقة أن استخدام تقنيات التنقيب عن البيانات أعطت نتائج واعدة في التنبؤ أو تصنيف الأمراض بشكل عام وخاصة أمراض القلب الوعائية.

# الفصل الاول: تحليل النظام والمنهجية المقترحة:

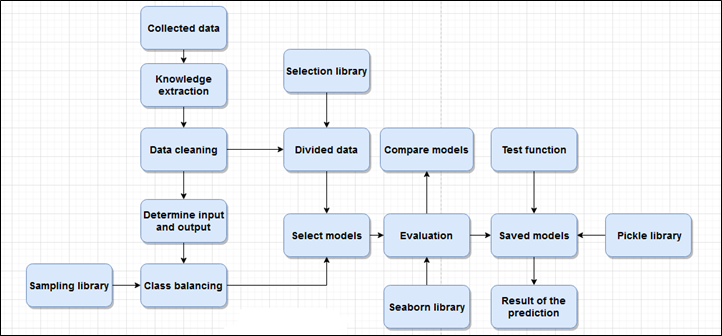
1. 1. **النموذج المقترح:**

**في هذا المشروع ، يمكن اعتماد منهجية خاصة للوصول إلى الهدف المنشود. في البداية ، سنعمل على جمع البيانات لعملية التصنيف والتنبؤ ، وبعد ذلك سنعمل على معالجة البيانات الأولية يدويًا ، للانتقال إلى المرحلة التالية ، وهي إجراء مجموعة من العمليات لتنظيف البيانات ، وملء القيم الفارغة ، وتحويل القيم النصية إلى قيم رقمية.**

**بعد ذلك تتم عملية اختيار حزمة من نماذج التصنيف والتي ستمر بدورها بمرحلتين: التدريب ثم اختبار وقياس القيم الناتجة عن طريق تطبيق معايير التقييم المعتمدة مثل الدقة وعدد مرات الإجابة الصحيحة وعدد مرات الإجابة الخاطئة.**

**حتى نصل أخيرًا إلى إنشاء مجموعة من المصنفات القادرة على إجراء عملية التنبؤ بدقة تامة ويمكن استخدامها لاحقًا عن طريق إدخال المعلومات وإعطائنا النتيجة المطلوبة وفقًا لعملية الإدخال.**

**تتم هذه المراحل كما ذكرنا سابقًا من خلال عدة خطوات ، ويوضح الشكل (1-4) مخطط تدفق عملية التنبؤ والمنهجية المستخدمة في هذا المشروع:**



**الشكل رقم 1-1 نموذج التنبؤ المقترح**

* 1. **الأدوات البرمجية:**

**عند تطوير التطبيق ، تم اختيار Python لتكون لغة البرمجة الأساسية ، وتحديداً Python الإصدار 3.6 لأنه الإصدار الأكثر استقرارًا وفقًا لتوثيق المكتبات الفردية التي يُراد استخدامها (TensorFlow ، 2020).**

**يرجع السبب وراء استخدام Python إلى النطاق الواسع من الدعم لمكتبات التعلم الآلي وأطر العمل الضرورية للتعلم الآلي ، مثل مكتبة "Sklearn" وSeaborn وPickle.**

* 1. **منهجية جمع البيانات ومعالجتها:**
     1. **جمع البيانات Collect data:**

تُعرف عملية جمع البيانات الكمية والنوعية حول متغيرات معينة بهدف تقييم النتائج أو اكتساب نظرة ثاقبة يمكن تنفيذها باسم جمع البيانات. إجراء واضح المعالم.

ومع ذلك ، قد يكون إنشاء هذا الإجراء صعبًا ، فهو يتضمن تقييم الأهداف ، وتحديد متطلبات البيانات ، واختيار طريقة لجمع البيانات ، وتنظيم خطة جمع البيانات التي تجمع بين أهم مكونات البرنامج.

* + 1. **مجموعات البيانات قبل المعالجة:**

يجب أن تحتوي مجموعة البيانات المناسبة لهذا المشروع على عدد من الميزات القوية المفيدة في عملية التنبؤ والتصنيف التي تعكس بيانات الإدخال ، ويجب أن تحتوي هذه الميزات على مجموعة من القيم وفقًا لمعايير محددة مثل العمر والجنس وعدد من معلومات الحالة الصحية مثل معدل التدخين ومعدل التلوث البيئي ومعدل تعاطي الكحول .. إلخ. تحتوي معظم الميزات على نطاق من القيم التي تعبر عن القيمة المتوسطة

في الوقت نفسه ، يوجد عمود يمثل الناتج المرغوب من خلال عملية التنبؤ ، على سبيل المثال ، حدوث المرض. قيم هذا العمود في نطاق مجموعة البيانات بين نعم أو لا.

* + 1. **استخراج المعرفة knowledge extraction:**

تتضمن مرحلة جمع البيانات لاستخراج المعرفة عملية البحث داخل كميات كبيرة من البيانات للكشف عن العلاقات التي لم يتم كشفها سابقا بين عناصر البيانات؛ والمعروفة أيضا باسم اكتشاف المعرفة في قواعد البيانات (KDD) وبالتالي هي النشاط الذي يقوم باستخراج المعلومات المتواجدة في كميات كبيرة من البيانات، بهدف البحث عن أنماط معرفية واكتشاف الحقائق الخفية الواردة في قواعد البيانات.

كما تتضمن عملية جمع البيانات ايضا تحليل البيانات المتواجدة في قواعد البيانات باستخدام الأدوات التي تبحث عن الاتجاهات أو البيانات التي لا معنى لها، واستخراج معلومات ضمنية، لم تكن معروفة سابقا، ويمكن أن تكون مفيدة.

وتتم عملية استخراج أنماط معرفية من مجموعات البيانات الكبيرة من خلال الجمع بين الأساليب من الإحصاءات والذكاء الاصطناعي مع إدارة قواعد البيانات وصولا للهدف المراد وهو اكتشاف المعرفة داخل قواعد البيانات (KDD) حيث تستخدم أساليب التحليل مثل: الشبكات العصبية Neural Networks أو الخوارزميات الجينية Genetic Algorithms أو شبكة القرارات Decision Trees والتقنية والنماذج الهجينة Hybrid Models لتحديد الأنماط والعلاقات في مجموعات البيانات.

وتتم عملية اكتشاف المعرفة في قواعد البيانات عن طريق تحديد الأنماط والاتجاهات في البيانات التي تم جمعها باستخدام طرق مختلفة مثل: التصنيف Classification أو التحليل التسلسلي Sequential analysis أو العنقدة (التجميع) clustering أو قواعد الارتباط Association Rule.

* + 1. **تنظيف البيانات data cleaning:**

يُعد تنظيف البيانات عملية أساسية في تجهيز البيانات الأولية لتطبيقات تعلّم الآلة وتطبيقات ذكاء الأعمال (BI). قد تحتوي البيانات الأولية على العديد من الأخطاء، ما قد يؤثر في دقة نماذج تعلّم الآلة ويؤدي إلى تنبؤات غير صحيحة ويؤثر سلبًا في الأعمال.

تتضمن الخطوات الأساسية لتنظيف البيانات تعديل حقول البيانات غير الصحيحة وغير المكتملة وحذفها، والتعرُّف على المعلومات المكررة والبيانات بعيدة الصلة وحذفها، وتصحيح أخطاء التنسيق والقيم المفقودة والأخطاء الإملائية.

تستلزم عملية تنظيف البيانات عدة خطوات لتحديد إدخالات المشكلة وإصلاحها. الخطوة الأولى هي تحليل البيانات لتحديد الأخطاء. قد يتضمن ذلك استخدام أدوات التحليل النوعي التي تعتمد على القواعد والأنماط والقيود لتحديد القيم غير الصالحة. الخطوة التالية هي إزالة الأخطاء أو تصحيحها.

**تتضمن الخطوات الشائعة لتنظيف البيانات علاج كل من:**

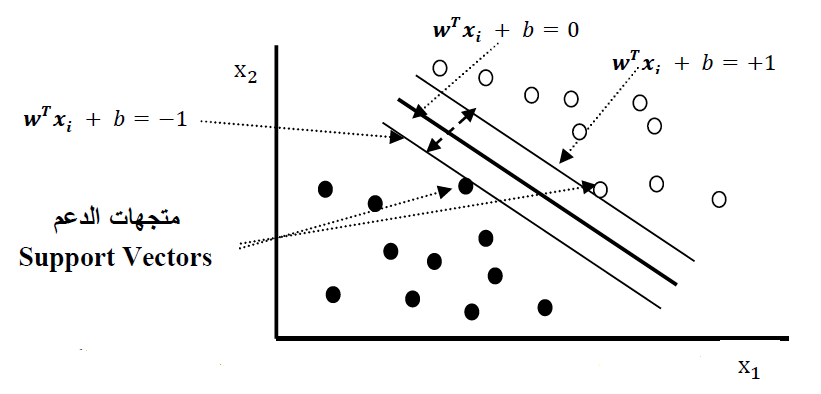
* البيانات المكررة: إسقاط المعلومات المكررة.
* البيانات غير ذات الصلة: تحديد الحقول المهمة لإجراء تحليل معين وإسقاط البيانات غير ذات الصلة من التحليل.
* القيم المتطرفة: يمكن أن تؤثر القيم المتطرفة جذريًا في أداء النموذج، لذا تُحدد القيم المتطرفة والإجراء المناسب.
* البيانات المفقودة: تمييز البيانات المفقودة بعلامة وإسقاطها أو تنسيبها.
* الأخطاء الهيكلية: تصحيح الأخطاء المطبعية وغيرها من التناقضات، وجعل البيانات متوافقة مع نمط أو اصطلاح شائع.
  1. **النماذج تعلم الالة المستخدمة:**

هناك العديد من نماذج التعلم الآلي التي يمكن استخدامها في هذا المشروع ، وقد اخترنا مجموعة من النماذج التي أثبتت نجاحها في مجال التنبؤ الطبي والصحي.

* + 1. **نموذج آلة المتجه الداعم :Support Vector Machine:**

يعتمد نموذج أدلة المتجه الداعم على عدد من المعلمات الأساسية مثل المستوى الفاصل Hyperplane ومضاريب لاكرانج Lagrange multipliers التي تؤثر على نحو كبير على دقة عملية التصنيف اذ ان البيانات في فضاء الادخال Input Space تصنف على وفق الانموذج الرياضي الاتي:

W: متجهة الاوزان Weights, X تمثل متجهة الادخال Input Vector, b تمثل قيمة التحيز Bias وd تمثل قيمة الإخراج كما يمكن ملاحظة معادلات الحدود في المستوى من الشكل التالي:



**الشكل رقم 2-1 معادلات الحدود في المستوى**

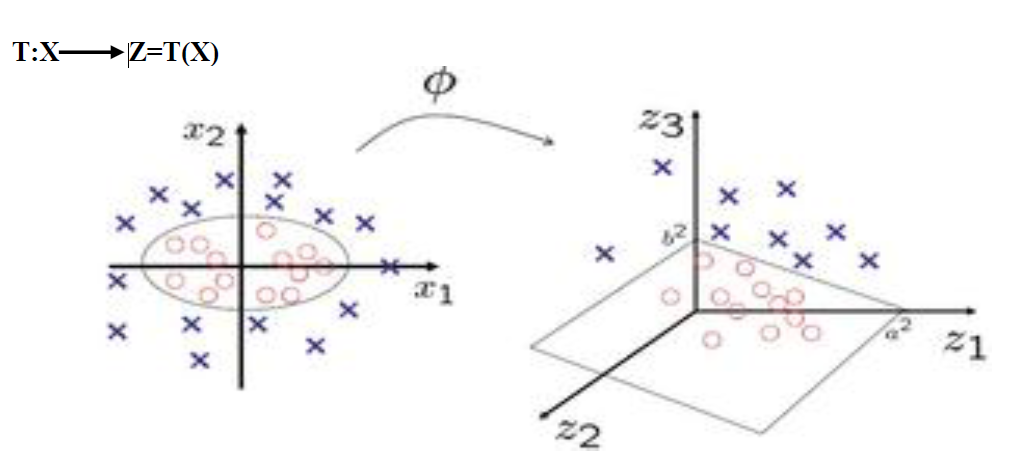
من الشكل السابق نلاحظ ان معادلة المستوي الفاصل Hyper plane تكتب بالشكل الاتي:

ان البيانات القريبة أو التي تقع على حدود الحد الفاصل تسمى متجهات الدعم أو المساندة Support vectors كما يمكن حساب المسافة بين النقاط في المستوى ومعادلة المستوى الفاصل من خلال العلاقة التالية:

وبعد عدد من الإجراءات والتحويلات الرياضية، يتم إيجاد قيم كل من متجه الاوزان المثالي والتحيز المثالي ومن بعدها تحسب دالة التنصيف الاتية:

اذن ان تمثل الوزن المثالي , تمثل قيمة التحيز المثالية وsgn تمثل القرار النهائي لانتماء (x) لاحد الأصناف Classes

في بعض الحالات لا يمكن فصل البيانات بوساطة الحد الفاصل Hyperplane لذلك يفضل استخدام تقنية Kernel trick ، التي تقوم بتحويل بيانات المدخلات من الخطية إلى بيانات ذات مساحة ابعاد (بيانات غير خطية) ويكون التحويل الجديد قابلاً لمفصل الخطي بوساطة الحد الفاصل Hyperplane ويبين الشكل التالي كيفية تحويل البيانات من الخطية الى غير خطية باستخدام Kernel Trick



**الشكل رقم 3-1 تقنية Kernel trick**

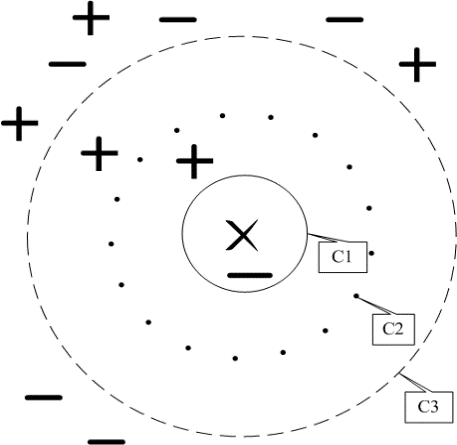
* + 1. **نموذج شجرة القرار Decision Trees model:**

يمكن استخدام نموذج شجرة القرار Decision Trees model للتنبؤ بمرض القلب ، وذلك عن طريق تحليل العوامل المختلفة التي تؤثر على احتمالية الإصابة بهذا المرض، مثل التدخين والتعرض للملوثات البيئية والوراثة والعمر والجنس والتاريخ الطبي للشخص. يتم تحليل هذه العوامل وتصنيفها في شجرة قرارية تساعد على تحديد المعطيات الأكثر أهمية في التنبؤ بالمرض ، مما يساعد على اتخاذ قرارات أكثر دقة في التشخيص والعلاج.

* + 1. **نموذج الجار الأقرب KNN model:**

على خوارزمية الجار الأقرب (K-Nearest Neighbor) التي تعتبر من التقنيات التنبؤية التي تهدف للتنبؤ بقيمة الفئة عن طريق مقارنة السجلات الشبيهة بالسجل المراد التنبؤ بقيمته وتقدير القيمة المجهولة لهذا السجل بناء على مقدار تلك السجلات. يعتمد عمل هذه الخوارزمية بشكل أساسي على عدد الحالات الأقرب (K) ووحدة القياس (metric) حيث يمثل الرمز (K) عدد الحالات المماثلة أي إيجاد مجموعة جزئية مؤلفة من (K) عنصر الأكثر تشابهاً مع الحالة المراد التنبؤ بقيمتها, وفي حال تماثل بعض أعضاء هذه المجموعة الجزئية يتم الاعتماد على وحدة القياس بغية تحديد العنصر الأكثر تشابهاً.

الشكل (4-4) يوضح آلية عمل هذه الخوارزمية حيث تظهر النقطة المجاورة الأقرب لإحدى نقاط البيانات المراد تصنيفها (X) ضمن الحد الفاصل (C1) بينما يظهر ضمن الحد الفاصل (C2) النقطتين المجاورتين للنقطة (X) وضمن الحد الفاصل (C3) النقاط الثلاثة المجاورة للنقطة (X).



**الشكل رقم 4-1 توزع البيانات ضمن المصنف KNN**

* + 1. **نموذج الانحدار الخطي Linear Regression model:**

معادلة النموذج للخوارزمية والتي ستحدد لنا كيفية التنبؤ للبيانات المستقبلية بالشكل التالي:

حيث:

* (h (xi: الفرضية hypothesis التي تستند عليها الخوارزمية.
* B: معامل الانحدار Regression Coefficient.
* Xi: البيانات التي لدينا.

لأننا نحتاج أن يكون مُخرج output ألخوارزمية على شكل احتمالية، فهذا يعني يجب أن تكون القيم محصورة بين (0,1)، لذلك سنستعين بدالة Sigmoid والتي ستكون صيغتها الرياضية بالشكل التالي:

لذلك سيكون تعريف الاحتمالية في النموذج الذي لدينا والذي سيوفر احتمالية التصنيف للفئتين (01،) بالشكل التالي:

ويتم اختصار هاتين المعادلتين بمعادلة رياضية واحدة لتكون بالشكل التالي:

والشيء المهم في هذه المرحلة هي أيجاد دالة الخطأ Error Function أو تسمى في بعض الأحيان Cost Function حيث نستخدم مع خوارزمية Logistic Regression نوع من دوال الخطأ تسمى أقصى تقدير للاحتمال maximum likelihood estimation (MLE) والتي تكون صيغتها الرياضية بالشكل التالي:

وهدفنا من استخدام دالة الخطأ هو حساب مقدار الخطأ الذي ينتج عن التنبؤ الخاطئ أثناء عملية التصنيف Classification، فكلما كان مقدار الخطأ قليل أثناء عملية التدريب training كلما كانت عملية التصنيف لدينا جيدة وحيث تمثل قيمة B معامل الانحدار Regression Coefficient الذي يمكن التحكم فيه لمحاولة تقليل مقدار الخطأ لدينا.

* 1. **التقييم:**

في التصنيف الثنائي، هناك عدد من المصطلحات التي ستتكرر عند التحدث عن تجريب النموذج الذكي، منها:

**True Positives (TP):** عند توقع النموذج ل 1, ووجود 1 فعلي حقيقي للبيانات

**True Negatives (TN):** عند توقع النموذج ل 0, ووجود 0-حقيقي للبيانات

**False Positives (FP):** 0عند توقع النموذج ل 1, لكن القيمة الحقيقية للبيانات تكون

**False Negatives (FN):** 1**ع**ند توقع النموذج ل 0, لكن القيمة الحقيقية للبيانات تكون

* **Precision score:**

يعرف على أنه نسبة الملاحظات الإيجابية المتوقعة بشكل صحيح **(true positives/tp)** إلى إجمالي الملاحظات الإيجابية المتوقعة.

* **Recall score**

هو النسبة **tp / tp + fn** حيث **tp هو** عدد الإيجابيات الحقيقية وfn عدد السلبيات الخاطئة.

* **F1 score**

هو مقياس لدقة النموذج في مجموعة البيانات. ... تعتبر أيضًا طريقة للجمع بين الاحكام **(Precision)** والاسترجاع **(Recall)** النموذج، ويتم تعريفها على أنها الوسط التوافقي لدقة النموذج واسترجاعه ويعطى بالعلاقة:

* **Accuracy score**

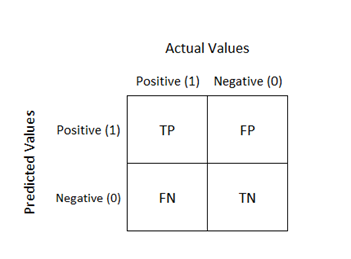
الدقة هي مقياس واحد لتقييم نماذج التصنيف. يتم شرحه أيضًا على أنه جزء من التنبؤات التي حصل عليها نموذجنا بشكل صحيح رياضيا، الدقة لها التعريف التالي: **الدقة = عدد التنبؤات الصحيحة/ العدد الإجمالي للتنبؤات.**

بالنسبة للتصنيف الثنائي، يمكن أيضًا حساب الدقة من حيث الإيجابيات والسلبيات على النحو التالي:

* **Confusion Matrix**

هي مصفوفة ثنائية الأبعاد nxn، حيث أن n هو عدد الأصناف المتوقعة، المصفوفة مهمتها قياس أداء لمشكلة تصنيف التعلم الآلي حيث يمكن أن يكون الناتج فصلين أو أكثر، جدول يحتوي على 4 مجموعات مختلفة من القيم المتوقعة والفعلية:

تستخدم Confusion Matrix لقياس Precision، accuracy، recall بطريقة مصورة وسهلة الفهم، ويوضح الشكل (6-4) ذلك:



**الشكل رقم 5-1 مصفوفة الارتباك**

يمثل كل صف من المصفوفة نقاط البيانات في فئة متوقعة، بينما يمثل كل عمود الحالات الموجودة في فئة فعلية (أو العكس). ينبع الاسم من حقيقة أنها تجعل من السهل معرفة ما إذا كان النظام يخلط بين فئتين أو صنفين.

* 1. **التصنيف والتنبؤ:**

يمكن استخدام طرق التعلم الآلي المذكورة أعلاه للتنبؤ بمرض القلب ، فعلى سبيل المثال:

* + طريقة الخوارزميات الجيدة لإنشاء نماذج تعلم الآلة التي تستخدم في التنبؤ بمرض القلب، وذلك عن طريق تحليل البيانات الطبية المتعلقة بالأشخاص المصابين والأشخاص الأصحاء.
  + طريقة الشبكات العصبونية التكرارية لتحليل التسلسلات الزمنية المتعلقة بالأشخاص المصابين ، مثل نتائج فحوصات الأشعة السينية والتحاليل الطبية.
  + طريقة الخوارزميات الجينومية لتحليل البيانات الجينومية المتعلقة بالأشخاص المصابين ، وذلك لتحديد الجينات المسؤولة عن هذا المرض والعلاقات بينها.
  + طريقة التعلم العميق لتحليل الصور الطبية المتعلقة بالأشخاص المصابين ، وذلك لتحديد المعالم المهمة في الصور وتشخيص المرض بدقة أكبر.
  + طريقة الخوارزميات الجيوماتيكية لتحليل الصور الجوية المتعلقة بالمناطق التي تشهد ارتفاعًا في معدلات إصابة السكان بهذا المرض، وذلك لتحديد العوامل البيئية والجغرافية التي تؤثر على انتشار المرض.

1. Ameer H. Ali, Mohannad A. M. Al-Ja'afari, Saif Abdulwahed," Rheumatoid Arthritis Diagnosis Based on Intelligent System" , Journal of University of Babylon for Pure and Applied Sciences ,Vol.(26), No.(7): 2018 [↑](#footnote-ref-1)
2. [Jyoti Soni , Ujma Ansari, Dipesh Sharma, Sunita Soni," Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 –8887), Volume 17–No.8, March 2011 [↑](#footnote-ref-2)
3. Dr. G. Karraz," Automatic Classification of Heartbeats using Neural Network Classifier based on a Bayesian Framework" ,International Conference of the IEEE Engineering in Medicine and Biology Society,2006 [↑](#footnote-ref-3)
4. Ishtake S.H , Sanap S.A., "Intelligent Heart Disease Prediction System Using Data Mining Techniques", International J. of Healthcare & Biomedical Research, Volume: 1, Issue: 3, Pages 94-101, April 2013 [↑](#footnote-ref-4)
5. Nidhi Bhatla, Kiran Jyoti," An Analysis of Heart Disease Prediction using Different Data Mining Techniques" International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 1 Issue 8, October - 2012 [↑](#footnote-ref-5)