مذكرة تخرج مقدمة في

جامعة محمد بوضياف – المسيلة

كلية الرياضيات والإعلام الآلي

قسم الإعلام الآلي

من أجل الاستيفاء الجزئي لمتطلبات شهادة

ماستر في الإعلام الآلي

من طرف

ريمة، شيخ

ليلى، بكاي

عنوان المذكرة

التنبؤ بمرض الكلى المزمن باستخدام التعلم الآلي

تحت إشراف الأستاذ

بلقاسم براهيمي

أعضاء لجنة المناقشة

طاهر مهني	جامعة المسيلة	**رئيسا**
بلقاسم براهيمي	جامعة المسيلة	**مقررا**
سعيد حماني	جامعة المسيلة	**ممتحنا**

جوان، 2026

إهداء

بسم الله الرحمن الرحيم

الحمد الله الذي بنعمته تتم الصالحات، الحمد الله ما تم جهد ولا ختم سعى إلا بفضله، وما تخطى العبد من عقبات وصعوبات إلا بتوفيقه ومعونته .اللهم انفعني بما علمتني وانفع بي .فالحمد لله على حسن التمام والختام.

إلى من كلله الله بالهيبة والوقار ..إلى من علمني العطاء بدون انتظار ..إلى من أحمل اسمه بكل افتخار (والدي العزيز)

وإلى ملاكي في الحياة ..إلى معنى الحب وإلى معنى الحنان والتفاني ..إلى بسمة الحياة وسر الوجود إلى من كان دعائها سر نجاحي إلى أغلى الحبايب (أمي الغالية) .

إلى إخوتي وأخواتي ... محبة ووفاء ...أنتم سندي.

إلى القريبين من القلب والداعمين والمساندين في السراء والضراء شكرا لكم ..دمتم لي.

شيخ ريمة

أهدي عملنا هذا إلى عائلتي كنوع من الشكر والتقدير على دعمهم المستمر لي وتعبيرا على إمتناني لوجودهم في حياتي، مع خالص الشكر لجميع جهودهم معي ولو بالكلمة الطيبة، كما أتقدم بخالص الشكر وعظيم الامتنان إلى أصدقائي، الذين كانوا دائماً سنداً ودعماً، وشاركوني بالتشجيع والتحفيز في كل خطوة..

ليلى بكاي

تشكرات

الحمد لله الذي أنشأ وبرى وخلق الماء والثرى الرحمان على العرش استوى والصلاة على الحبيب المصطفى أما بعد

باسم الله الرحمان الرحيم

"قل اعملوا فسيرى الله عملكم ورسوله و المؤمنين"

الأساتذة أعضاء لجنة المناقشة

الأستاذ المشرف الفاضل

الحضور الكرام

السلام عليكم ورحمة الله وبركاته

نحمد الله تعالى على توفيقه وفضله الذي لولاه ما وصلنا إلى هذه المرحلة من مشوارنا العلمي

يسعدنا ويشرفنا أن نناقش معكم اليوم مذكرتنا لنيل شهادة الماجيستر ونتمنى أن تنال رضاكم وتقديركم

كما نتوجه بالشكر الجزيل إلى أعضاء اللجنة

وإلى أستاذنا المشرف على توجيهه

ولعائلتينا الكريمتين فهما السند في كل خطوة

كما نشكر كل من حضر وشاركنا هذه اللحظة المهمة كل باسمه ومقامه

بارك الله فيكم ومرحبا بكم

انتهت الرحلة بعون الله وبفضله

"وآخر دعواهم أن الحمد لله رب العالمين"

الحمد لله عند البدئ وعند الختام

الفهرس

قائمة الصور vii

قائمة الجداول viii

المقدمة العامة 1

الفصل الأول: الذكاء الاصطناعي ومرض القصور الكلوي المزمن 4

1.1 مقدمة 4

2.1 الذكاء الاصطناعي 4

1.2.1 تعريف الذكاء الاصطناعي **5**

1.2.1 أهميته **5**

2.2.1 أنواع الذكاء الاصطناعي **6**

3.2.1 مجالات الذكاء الاصطناعي **7**

4.2.1 تطبيقات الذكاء الاصطناعي في المجال الطبي **7**

5.2.1 مزايا وعيوب الذكاء الاصطناعي في المجال الطبي **8**

1.2 التعلم الآلي 9

1.1.2 تعريف التعلم الآلي **9**

2.1.2 أنواع التعلم الآلي **11**

3.1.2 أهمية فهم البيانات وجودتها في تعلّم الآلي **13**

1.3 مرض القصور الكلوي المزمن 14

1.1.3 تعريف المرض **14**

2.1.3 إحصائيات المرض في الجزائر **15**

3.1.3 الأسباب وعوامل الخطر **15**

4.1.3الأعراض ومراحل المرض.....17

5.1.3 طرق التشخيص 19

1.4 الدراسات السابقة حول التنبؤ بمرض الكلى 20

1.1.4 الخوارزميات المستخدمة 21

2.1.4 مقارنة نتائج الدراسات.....22

3.1.4 الفجوات البحثية 24

1.5 الخاتمة..... 26

الفصل الثاني: المراحل الأساسية المتبعة.....27

1.1مقدمة..... 27

1.2التحليل الاستكشافي للبيانات(Exploratory Data Analysis) 27

1.3المعالجة المسبقة للبيانات(Preprocessing) 27

1.1.3 معالجة القيم المفقودة(Imputation) 28

التعويض بالوسيط(Median Imputation) 28

خوارزمية أقرب الجيران(K-Nearest Neighbors Imputation) 28

التعويض باستخدام الانحدار (Regression Imputation)..... 28

2.1.3 القيم الشاذة (Outliers) 29

1. الكشف عن القيم الشاذة باستخدام المدى الربيعي (IQR) 29

2. معالجة القيم الشاذة (Outliers Handling) 29

التحويل الكمي (Quantile Transformation) 29

تقليص القيم المتطرفة (Winsorization) 29

3.1.3 ترميز المتغيرات الوصفية(Categorical Encoding) 30

4.1.3التقييس باستخدام (Min-Max Normalization) 30

5.1.3اختيار السمات(Feature Selection) 31

اختبار كاي-تربيع(Chi-Square) 31

المعلومات المتبادلة(Mutual Information) 31

خوارزمية الإزالة التكرارية(Recursive Feature Elimination) 31

1.4 تقسيم البيانات (Train/Test Split) 32

1.5 تدريب النماذج(Training) 32

شجرة القرار(Dicision Tree) 32

الغابة العشوائية (Random Forest) 32

الأشجار المتطرفة العشوائية (Extra Trees) 33

تعزيز التدرج المحسّن(XGboost) 33

التكديس(Stacking)..... 33

1.6 الخاتمة 33

الفصل الثالث: الإنجاز 35

1.1 المقدمة 35

1.2 وصف البيانات 35

1.3 بيئة وخطوات سير العمل 36

1.4 التنفيذ(Implémentation) 37

1.1.4 مرحلة التحليل الاستكشافي(EDA) 37

2.1.4. المعالجة المسبقة(Preprocessing) 45

3.1.4 التدريب ومقارنة النتائج 46

1.5 تطوير النموذج النهائي 56

1.1.5 التحسين 56

2.1.5 ضبط المعاملات(Hyperparameter Tuning) 56

3.1.5 قابلية التفسير(Explainability SHAP) 57

4.1.5التقييم النهائي(Final Evaluation) 59

5.1.5 من نموذج الى منصة تفاعلية 60

1.6 عرض المنصة 61

1.7 الخاتمة 65

الخاتمة 66

قائمة المراجع 68

قائمة الصور

الصورة 1: العلاقة بين الذكاء الاصطناعي، وتعلّم الآلة، والتعلّم العميق، وعلم البيانات.....10

الصورة 2: كلية شخص مصاب بفشل كلوي مزمن 14

الصورة 3:المقارنة بين المنهج التقليدي والمنهج المعتمد على الذكاء الاصطناعي في تشخيص مرض الكلى المزمن.....20

الصورة 4: مقارنة بين خوارزميات ML لدراسة النموذج الهجين للتنبؤ ب CKD. 22

الصورة 5: تحميل البيانات المستخدمة وعرض أول الصفوف.....38

الصورة 6:المدرج التكراري لتوزيع المتغيرات العددية.....40

الصورة 7: توزيع المتغيرات الوصفية باستخدام المدرج التكراري. 42

الصورة 8: مصفوفة الارتباط بين المتغيرات(Correlation Heatmap) 43

الصورة 9:الكود البرمجي لخطوات تنظيف البيانات.....44

الصورة 10 : إزالة المسافات وعلامات الجدولة من النصوص الفئوية وتوحيد قيمها.....44

الصورة 11: إزالة العمود id وحفظ البيانات النظيفة.....45

الصورة 12: مخطط متوسط التأثير(قيمة SHAP لكل خاصية) على حجم مخرجات النموذج. 57

الصورة 13 : مخطط تأثير قيم الخصائص المختلفة على المتغير الهدف و اتجاه هذا التأثير. 58

الصورة 14: مخطط الاعتماد للخاصية hemo وتأثيرها على مخرجات النموذج.....59

الصورة 15: مصفوفة الالتباس (Confusion Matrix) للتقييم النهائي على بيانات الاختبار.....60

الصورة 16 : واجهة التطبيق باللغة العربية والوضع الفاتح. 61

الصورة 17: واجهة التطبيق باللغة الانجليزية والوضع الداكن. 62

الصورة 18: واجهة إدخال 5 قيم للمؤشرات الحيوية السريرية الرئيسية. 63

الصورة 19: رسالة تحذير عند تجاهل ملء أحد الحقول.....63

الصورة 20: واجهة تحتوي بيانات ذات قيم طبيعية تُظهر أن المستخدم سليم(لا يوجد CKD) 64

الصورة 21: واجهة تحتوي بيانات ذات قيم غير طبيعية تُظهر أن المستخدم مريض (يُحتمل CKD). 64

قائمة الجداول

الجدول1 : المقارنة بين الأنواع الثلاثة الرئيسية في تعلم الآلة.....12

الكلمات المفتاحية:

AI: الذكاء الاصطناعي(Artificial Intelligence) .

ML : التعلم الآلي(Machine Learning) .

CKD : مرض القصور الكلوي المزمن(Chronic Kidney Disease) .

RFE : خوارزمية الإزالة التكرارية (Recursive Feature Elimination).

MI : المعلومات المتبادلة (Mutual Information).

المقدمة العامة

يعتبر مرض الفشل الكلوي من الأمراض المزمنة الشائعة التي تهدد صحة الإنسان، حيث أن أي خلل يصيب وظائف الكلى يسبب تراكم السموم والسوائل داخل الجسم مما يؤثر بشكل مباشر على حياة المريض. ومن التحديات التي واجهها المختصون في المجال الطبي هو كيفية التشخيص المبكر والتنبؤ بالمرض، وبالتالي تحسين فرص العلاج وتقليل المضاعفات .

فمع التطور الكبير في مجال الذكاء الاصطناعي و التعلم الآلي، أصبح من الممكن تحليل أعداد كبيرة من البيانات الطبية واكتشاف الأنماط الخفية داخلها ، حيث ان الاعتماد على خوارزميات التعلم الآلي يعتبر من التقنيات الحديثة لبناء أنظمة ذكية قادرةعلى التنبؤ بالأمراض بدقة عالية.

يندرج سياق هذه الدراسة في دمج تقنيات التعلم الآلي مع المجال الطبي (التنبؤ بمرض الفشل الكلوي) بالاعتماد على البيانات المدروسة. حيث يهدف هذا العمل إلى دعم الأطباء في الكشف المبكر عن المرض من خلال اتخاذ قرارات طبية أكثر دقة وفعالية، كما يهدف الى مقارنة مجموعة مختلفة من نماذج التعلم الآلي واختيار أفضل نموذج اعتمادا على تحديد الدقة والأداء والكفاءة الزمنية، ويُركز على تقليل عدد السمات المختارة مع الحفاظ على دقة عالية.

يتناول تقرير هذه الدراسة من ثلاث فصول أساسية وهي:

- **الفصل الأول الذكاء الاصطناعي ومرض القصور الكلوي المزمن:** يحتوي مفاهيم الذكاء الاصطناعي والتعلم الآلي وعلاقتهما بالمجال الطبي، إضافة إلى دراسة المرض وعرض أهم الاعمال السابقة المتعلقة بالتنبؤ به ;
- **الفصل الثاني المراحل الأساسية المتبعة:** يعد المنهج المتكامل الذي يهدف إلى عرض المراحل الأساسية المعتمدة في إعداد البيانات وبناء نماذج التعلم الآلي، كما يقدم مفاهيم للتقنيات المستخدمة في كل مرحلة وشرح الخوارزميات المحددة ;
- **الفصل الثالث الإنجاز:** يخصص للجانب التطبيقي والانجاز، حيث يتم تنفيذ مراحل التحليل والمعالجة والتدريب، ثم تحسين النموذج النهائي وعرض النتائج والمنصة التفاعلية المقترحة.

إشكالية هذه الدراسة تتمثل في طريقة بناء نموذج تعلم آلي قادر على التنبؤ بمرض الفشل الكلوي بدقة عالية، مع تقليل عدد السمات المستخدمة من 25 سمة الى 5 سمات فقط بهدف تبسيط النموذج وتحسين سرعة التنفيذ لتسهيل استخدامه عمليا، كما تكمن في اختيار أفضل نموذج تعلم آلي اعتمادا على تحديد أفضل توازن بين المقاييس مثل Accuracyو Precisionو Recallو F1-scoreوROC-AUC، مع المحافظة على استقرار النتائج وكفاءة التنفيذ.

لتحقيق أهداف هذه الدراسة تم اتباع عدة خطوات تمثلت في:

- تقسيم البيانات إلى جزئين (80/20) جزء للتدريب والآخر للاختبار من أجل تجنب مشكلة Data Leakage;
- استكشاف البيانات وتنظيفها من الأخطاء ;
- معالجة البيانات الطبية وتحضيرها للتدريب من خلال:
  - معالجة القيم المفقودة باستخدام عدة تقنيات مثل Median و KNNو Regression Imputation ;
  - معالجة القيم الشاذة باستخدام Winsorization و Quantile Transformation ;
  - ترميز السمات النصيةEncoding ;
- اختيار أهم 5 سمات باستخدام تقنيات Chi-Square و Mutual Information وRFE مع الحفاظ على الأداء العالي ;
- بناء ومقارنة عدة نماذج تعلم آلي مختلفة ;
- تقييم النماذج باستخدام التحقق المتقاطع 10-Fold Cross Validationوعدة مقاييس أداء .

تم الاعتماد في هذ العمل على لغة البرمجة Python ومجموعة من المكتبات الخاصة بتحليل البيانات والتعلم الآلي، كما تم استخدام مجموعة من خوارزميات التعلم الآلي وهي: Random Forest،Decision Tree ، XGBoost Extra Trees، Stacking Ensemble.

ترتكز منهجية هذه الدراسة على منهج تجريبي مقارن، حيث تم استخدام عدة تجارب لنماذج مختلفة مع تطبيق أساليب متعددة للمعالجة المسبقة ولاختيار السمات. بعد ذلك تدريب النماذج واختبارها باستعمال أسلوب التحقق المتقاطع، كما تم مقارنة النتائج بالاعتماد على المقاييس الإحصائية إضافة الى ذلك مقارنة زمن التنفيذ بين النماذج لتحديد النموذج الأفضل من حيث الكفاءة والفعالية في التصنيف.

من خلال نتائج الدراسة تبين أن تقنيات التعلم الالي يمكن أن تحقق دقة عالية للتنبؤ بمرض الفشل الكلوي المزمن بعدد قليل من السمات. كما أظهرت النتائج تفوق بعض نماذج التجميع مثل Stacking Ensemble وRandom Forest مقارنة بباقي النماذج من حيث الأداء العام.

تؤكد هذه الدراسة أهمية استخدام تقنيات الذكاء الاصطناعي في المجال الطبي، لما توفره من سرعة ودقة في تحليل البيانات والتنبؤ بالأمراض، مما قد يساهم مستقبلاً في تطوير أنظمة طبية ذكية تدعم التشخيص المبكر وتحسين جودة الرعاية الصحية.

الفصل الأول: الذكاء الاصطناعي ومرض القصور الكلوي المزمن

1.1مقدمة

يتناول هذا الفصل عرضًا تمهيديًا حول الذكاء الاصطناعي والتعلّم الآلي، لكونهما من أسرع المجالات التقنية تطورًا وأكثرها تأثيرًا في مختلف جوانب الحياة. كما يوضح المفاهيم الأساسية، مدى أهميتهما، تطبيقاتهما ومجالاتهما خاصة في المجال الطبي، كما تم التطرق إلى مناقشة أهم المزايا والتحديات الممكنة، مما يسمح للقارئ اكتساب معلومات أكثر حول هذا المجال واستكشاف فرصه الواسعة.

يعتبر التعلم الآلي أهم فروع الذكاء الاصطناعي، والذي يمكّن الأنظمة من التعلم من البيانات وتحسين أدائها تلقائيا، حيث أصبحت قادرة على تحليل البيانات الطبية والتنبؤ بالأمراض.

في المجال الطبي، يتطلب مرض القصور الكلوي تشخيصا مبكرا ودقيقا، يأتي الذكاء الاصطناعي لدعم اتخاذ القرار الطبي، وتحليل البيانات المعقدة، وتحسين جودة الرعاية الصحية للمصابين بهذا المرض، مما أدى الى الزامية فهم هذه التقنيات لمواكبة التطورات الطبية الحديثة.

2.1 الذكاء الاصطناعي

الذكاء الاصطناعي هو أحد أحدث المجالات في العلوم والهندسة ظهر تقريبًا عام 1950 على يد آلان تورينج ويعني جعل الآلة تفكر وتعمل مثل الإنسان. في بدايته واجه العديد من الصعوبات لأن العلماء حاولوا محاكاة الذكاء والتفكير البشري. رغم كل التقدم الذي وصلت إليه العلوم لم يستطع العلماء تحديد الطريقة التي يفكر بها الإنسان، توصلوا إلى فكرة " أننا كبشر لا يهمنا كيف تعمل الآلة المهم اننا نحصل على نفس النتيجة مثلما يقدمها الانسان ". بعدها بدأ الذكاء الاصطناعي بالتطور تدريجيا حتى دخل لجميع مجالات حياتنا العملية(المجال العسكري، الطبي، الصناعي، التجاري وحتى التعليمي) وأصبح أدائه عاليا أي بدأت الالة تحاكي التفكير البشري، حتى أنها أصبحت تتفوّق عليه في العديد من المجالات. إن كانت الإنترنت هي ثورة السنوات الماضية، فالذكاء الصناعي هو ثورة الحاضر والمستقبل [1].

1.2.1 تعريف الذكاء الاصطناعي

يعد الذكاء الاصطناعي مصطلح صاغه البروفيسور الفخري في جامعة ستانفورد جون مكارثي في عام 1955 لأول مرة ضمن ورشة عمل صيفية تسمى مشروع دار تموث الصيفي للبحوث حول AI على أنه "علم وهندسة صناعة الآلات الذكية" أي أسلوب يُعطي الآلة القدرة على محاكاة الذكاء البشري [1].

ويمكن صياغته أيضا بمفهوم آخر وهو قدرة الحواسيب على محاكاة الوظائف الإدراكية البشرية، مثل التعلم وحل المشكلات. بواسطة AI يستخدم نظام الحاسوب الرياضيات والمنطق لمحاكاة عملية التفكير التي يستخدمها البشر للتعلم من المعلومات الجديدة واتخاذ القرارات [1].

1.2.1 أهميته

تهدف تقنيات AI لخدمة البشرية وتسهيل حياة الأفراد كما يمكن تطبيقه في جميع المجالات للحصول على كفاءة أعلى ومن أهم الحلول التي يقدمها [2] :

تحسين الكفاءة :يعمل على تحسين الكفاءة من خلال أتمتة المهام الروتينية وتقليل الأخطاء البشرية ;

تحليل البيانات :يمكن ل AI تحليل ومعالجة كميات ضخمة من البيانات بسرعة ودقة، مما يتيح للمؤسسات الحصول على نظرة تحليلية لعملياتها وسهولة تسييرها ;

اتخاذ القراروتقديم حلول مبتكرة :يمكنه إتخاذ قرارات أكثر دقة من البشر، وذلك لقدرته على تحليل العلاقات المعقدة وتخزين البيانات الضخمة المتاحة كما يساهم في تطوير حلول جديدة لمشكلات في مجالات متعددة، مثل الرعاية الصحية والطاقة ;

توفير تجربة مستخدم محسنة :عبر تخصيص التوصيات والخيارات، يعزز من رضا العملاء.

2.2.1 أنواع الذكاء الاصطناعي

ينقسم إلى ثلاثة أقسام[3] :

1- الذكاء الاصطناعي الضيق Artificial Narrow Intelligence: يقوم AI الضيق بمهام معينة، مثل التعرف على الوجه، أو التعرف على الكلام أو البحث في الإنترنت أو قيادة السيارات كما يمتاز بكفاءة عالية في أداء المهام التي تمّت برمجته للقيام بها. والآلات التي تعمل به ليس لديها أي قدرة على التفكير، تؤدي فقط الوظائف المطلوبة، اذن هذا النوع لا يحاكي الذكاء البشري ولكن يحاكي فقط السلوك البشري. ومن أمثلة الذكاء الاصطناعي الضيق:

- برامج التعرف على الصور ;
- روبوتات التصنيع والطائرات بدون طيار ;
- لسيارات ذاتية القيادة ;
- مرشحات البريد الإلكتروني العشوائي.

2- الذكاء الاصطناعي العام Artificial General Intelligence : يُعد AI العام (القوي ، العميق)من مراحل تطور AI، وهو الآلات التي تمتلك ذكاء عام يحاكي الذكاء والسلوك البشري ، فتكون قادرة على التفكير واتخاذ القرارات مثل البشر. ويستخدم نظرية العقل، من خلالها تستطيع الآلات تمييز الاحتياجات والعواطف والمعتقدات وعمليات التفكير ;

3- الذكاء الاصطناعي الفائق Artificial Super Intelligence: وهو AI افتراضي، فالآلات التي تستخدمه تصبح مدركة لذاتها وتتجاوز القدرة البشرية. أدرك العلماء أن هذا النوع قد يمثل مصدر تهديد للبشرية عندما يصل لمراحل متقدمة أكثر، لا يفهم المشاعر والتجارب الإنسانية فقط، بل يمكنه إثارة المشاعر والاحتياجات والمعتقدات والرغبات الخاصة به، وستكون للآلات القدرة على اتخاذ القرار وحل المشكلات تتجاوز قدرات البشر.

3.2.1 مجالات الذكاء الاصطناعي

يُعدّ AI علمًا واسعًا يضمّ العديد من المجالات، معظمها بين الجوانب العلمية والهندسية، كل مجال يعالج مواضيع محددة، كما يعتمد أساليب وتقنيات مختلفة لمعالجة المشكلات. ومن أهم المجالات [4]:

- **تعلم الآلة(Machine Learning) :** عرّف آرثر صموئيل تعلّم الالة هو تمكين الحاسوب من اكتساب المعرفة والتعلّم من المشكلات وتحسين أدائه اعتمادًا على الخبرة والبيانات، بدلًا من الاعتماد فقط على التعليمات البرمجية ;
- **الشبكات العصبية الاصطناعية(Artificial Neural Networks) :** هي نماذج حسابية تحاكي طريقة عمل الجهاز العصبي في الدماغ البشري، وتهدف إلى التعرّف على الأنماط، والتصنيف، والتنبؤ، ومعالجة البيانات المعقدة ;
- **التعلم العميق (Deep Learning):** هو أحد فروع تعلّم الآلة يعتمد على شبكات عصبية اصطناعية تتكوّن من عدد كبير من الطبقات المتتالية من وحدات المعالجة، ويستفيد من التطورات الحاسوبية الكبيرة (مثل المعالجات القوية) ومن تقنيات التدريب المحسّنة، لتعلّم الأنماط المعقدة من كميات ضخمة من البيانات.

4.2.1 تطبيقات الذكاء الاصطناعي في المجال الطبي

بفضل تقنيات AI شهد القطاع الصحي تطورًا سريعًا، حيث ساهم في تحسين جودة الخدمات الطبية، ورفع دقة التشخيص، وتعزيز كفاءة الأنظمة الصحية. فمن أهم تطبيقاته في المجال الطبي[5] :

1. **تشخيص الأمراض**

حظيّ الجانب التشخيصي للأمراض باهتمام واسع في تطبيقات AI، إذ تُستخدم الخوارزميات المتقدمة، خاصة تقنيات التعلم العميق، في دعم الأطباء من خلال تحليل بيانات المرضى، والصور الطبية، ونتائج التحاليل المخبرية، مما يساعد في الكشف عن الأورام، والكسور، والاختلالات المرضية وغيرها من المعلومات السريرية، كما يساهم في الاكتشاف المبكر للأمراض الخطيرة بدقة عالية وفي وقت قصير مثل السرطان، أمراض القلب، داء السكري وأمراض الكلى.

ومن أهم التقنيات التي يُساهم AI في تحليل بياناتها ودعم نتائجها: التصوير المقطعي المحوسب، والتصوير بالرنين المغناطيسي، والتصوير بالأشعة السينية، إضافةً إلى مجالي الجينوميات والبروتيوميات[6] ;

2. **العلاج**

يمكن استخدام تقنيات AI لتطوير أنظمة علاج مخصصة وفق للحالة الصحية للمريض ودرجه خطورة المرض، وذلك من خلال تحليل البيانات الطبية الكبيرة. بحيث تساعد في تحديد أفضل الأدوية المناسبة، وضبط الجرعات، واختيار أفضل الاستراتيجيات العلاجية الفعالة ;

3. **الرعاية الصحية**

يعمل AI على تحسين جودة الرعاية الصحية وتجربة المرضى، من خلال استخدام الأنظمة الذكية والروبوتات لتقديم الرعاية الصحية الأساسية للمرضى في المناطق النائية أو البعيدة، التي تعاني من نقص في الخدمات الطبية ;

4. **البحوث الطبية**

AI هو أداة مهمة في تسريع عمليات البحث العلمي والطبي، وتحليل كميات كبيرة من البيانات الطبية، والتحولات الجينية وغيرها من العوامل المرتبطة بالصحة، كما يساهم في اكتشاف العلاقات بين العوامل المختلفة المرتبطة بالصحة، إضافة إلى دعم التنبؤ بالأمراض وتطوير العلاجات المستقبلية الأكثر فاعلية.

5.2.1 مز ايا وعيوب الذكاء الاصطناعي في المجال الطبي

أ- مز اياه: له العديد من المزايا في المجال الطبي أهمها نذكر[5] :

1. **تحسين التشخيص**: يساهم في رفع دقة التشخيص الطبي وتسريعه، مما يسمح بالعلاج المبكر وزيادة فرص الشفاء ;
2. **تحسين الرعاية الصحية**: يمكنه تحسين جودة الرعاية الصحية وتجربة المرضى من خلال تقديم الرعاية الصحية المخصصة لكل مريض وفقًا لحالته ;
3. **توفير الوقت والجهد**: كما يمكنه أن ينقل العديد من المهام المتكررة إلى الأنظمة الذكية، مما يوفر للأطباء والممرضين المزيد من الوقت للتركيز على الرعاية الصحية للمرضى ;
4. **تقليل الأخطاء الطبية**: كما يمكنه تقليل الأخطاء الطبية الشائعة، من خلال تحليل البيانات الطبية وتقدم التشخيص والعلاج الدقيق لكل حالة مريض ;
5. **تطوير البحث الطبي**: ويمكنه أيضا تسريع تحليل البيانات الطبية الضخمة بدقة أكبر، مما يساعد على تطوير الأدوية والعلاجات والتشخيصات الجديدة.

ب- عيوبه: على الرغم من أن AI له العديد من المزايا والفوائد في مجال الطبي، إلا أن له بعض المخاطر التي يمكن أن تنشأ من استخدامه [5]:

1. **عدم الدقة:** قد يكون غير دقيق في بعض الحالات، خاصة إذا كانت البيانات المستخدمة في التدريب غير كافية أو غير دقيقة، وهذا ما يؤدي إلى تشخيص خاطئ أو علاج غير فعال ;
2. **عدم الشفافية:** في بعض الأحيان يكون من الصعب فهم كيفية عمله، ما يجعل القرارات الخاطئة التي يتخذها صعبة الكشف عنها ;
3. **خصوصية المرضى:** يمكن أن يؤثر استخدامه على خصوصية المرضى، خاصة عند استخدام بياناتهم الصحية دون موافقتهم الصريحة، فقد يعرض هذا الأطباء والمؤسسات الطبية للمسائل القانونية ;
4. **الإدمان:** يؤدي الاعتماد الزائد عليه في المجال الطبي إلى الإدمان، فقد يؤدي بالأطباء والممرضين لإضعاف القدرة على اتخاذ القرار الطبي المستقل ;
5. **القضايا الأخلاقية:** يثير استخدام AI في المجال الطبي بعض المسائل الأخلاقية، مثل الأخلاقيات في جمع البيانات الصحية واستخدامها، وتحليلها واتخاذ القرارات المتعلقة بتوجيه الرعاية الصحية.

1.2 التعلم الآلي

1.1.2 تعريف التعلم الآلي

عرَّف البروفسور توم ميتشل- Tom Mitchell وهو عالم حاسوب أمريكي وبروفيسور في جامعة Carnegie Mellon في الولايات المتحدة- تعلم الآلة على انه :

"عندما يتم أمر برنامج الحاسوب للتعلم من الخبرة **E** ليقدر على أداء المهمة **T**، والذي تُحسَب دقته عبر مقياس الدقة**P** ، فإذا استطاع البرنامج أداء المهمة **T**على قدر الدقة المحددة من **P** عن طريق التعلم من الخبرة **E**، فإنّ ذلك هو تعلم الآلة". [7]

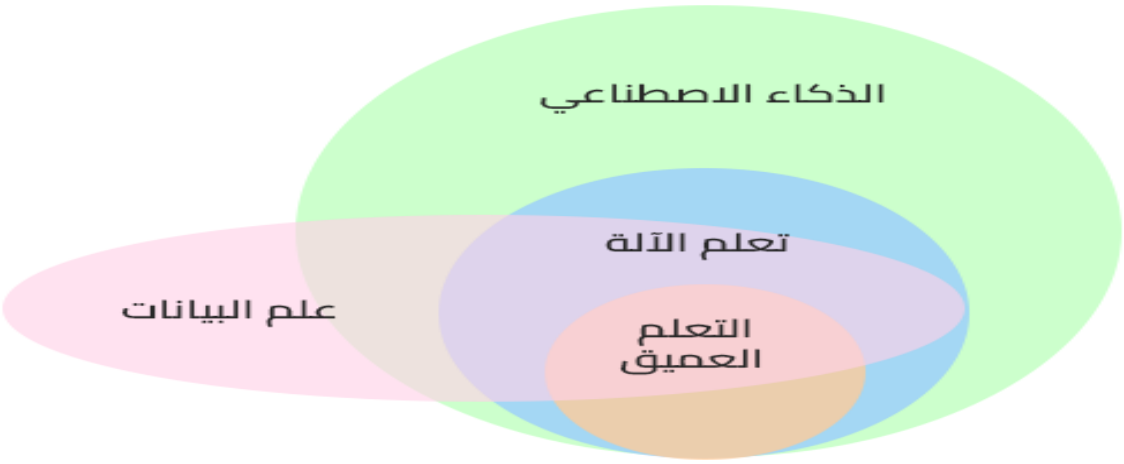
يُعد تعلّم الآلة فرعًا من فروع الذكاء الاصطناعي، حيث شهد تطورا واسعا مع توفر القدرة الحاسوبية والبيانات الضخمة. ومن التخصصات الفرعية المهمة في هذا المجال التعلّم العميق (Deep Learning) الذي يسند قواعده من طريقة عمل

الشبكات العصبية في الدماغ البشري لحل المشكلات المعقدة مثل تطبيقات التعرف على الصور. إضافة الى تطبيقاته العلمية كتحليل تسلسل الحمض النووي، دراسة النجوم والكواكب البعيدة، وتشخيص الأمراض، وتطوير الأدوية.

يعرف التعلم الالي في الأوساط الرياضية باسم التعلم الاحصائي أو التحليلات التنبؤية لأن نماذجه تعتمد على خوارزميات مبنية على الجبر الخطي والتفاضل والتكامل والإحصاء. كما يهدف هذا المجال إلى استخراج الأنماط من البيانات السابقة للتنبؤ بنتائج جديدة بدقة أفضل.

تتعلّم خوارزميات تعلّم الآلة الأنماط من البيانات الضخمة دون الحاجة إلى تعليمات صريحة ومبرمجة مسبقاً، مما يجعلها أكثر كفاءة في المهام المعقدة كتحديد الوجه أو الكائنات الموجودة في الصور. كما يرتبط تعلّم الآلة ارتباطًا وثيقًا بعلم **البيانات(Data Science)** ، إذ تعتمد نماذجه على بيانات مُعالجة ومُهيأة جيدا [7].

ومن خلال هذه الصورة التي توضح العلاقة بين الذكاء الاصطناعي وتعلم الالة والتعلم العميق وعلم البيانات يمكن فهم الترابط والتداخل بين هذه المجالات:



الصورة 1: العلاقة بين الذكاء الاصطناعي، وتعلّم الآلة، والتعلّم العميق، وعلم البيانات.[7]

2.1.2 أنواع التعلم الآلي

تنقسم خوارزميات التعلّم الآلي إلى ثلاثة أنواع رئيسية[8] .

1. **التعلم الآلي الخاضع للاشراف (Supervised Learning):** هو نوع من ML حيث يتم فيه تزويد النموذج ببيانات مُعلَّمة تحتوي على المدخلات والمخرجات الصحيحة المقابلة لها (أي أن متغير الهدف أو النتيجة معروف) ويضم عدة أنواع من الخوارزميات:

- ✓ **خوارزميات التصنيف(Classification) :** هي التنبؤ بمتغيرات المخرجات الفئوية من خلال تصنيف الأمثلة أو العينات بناءا على خصائصها (إسناد البيانات إلى فئات محددة مسبقًا) وتتضمن خوارزميات التصنيف:
  - Support Vector Machines (SVM) ;
  - Decision Trees (DT) ;
  - Logistic Regression ;
  - Random Forest (RF).
- ✓ **الانحدار(Regression) :** يُستخدم الانحدار للتنبؤ بالمخرجات من خلال تحديد العلاقات الخطية بين القيم الحقيقية أو المستمرة (مثل درجة الحرارة، الراتب) ومن أشهر خوارزميات الانحدار:
  - Linear Regression ;
  - Polynomial Regression ;
  - Support Vector Regression (SVR) .

2. **التعلّم غير الخاضع للاشراف (Unsupervised Learning):** يتم في هذا النوع تدريب النموذج على بيانات غير مصنفة (أي أن متغير الهدف أو النتيجة غير معروف مسبقا) فهذا يساعد على اكتشاف الأنماط المخفية في البيانات لتقليل أبعادها والنمذجة التنبؤية وله خوارزميات منها:

✓ ***التجميع(Clustering)***

يُستخدم التجميع لتصنيف البيانات وفقًا لتشابه في الخصائص ومن أشهر خوارزميات التجميع:

- K-Means Clustering ;
- Hierarchical Clustering ;
- Probabilistic Clustering ;
- DBSCAN.

3. **التعلّم المُعزَّز(Reinforcement Learning)** : هو نوع من البرمجة الديناميكية التي تدرب الخوارزميات باستخدام نظام الثواب والعقاب حيث يحصل على مكافآت عند اتخاذ قرارات صحيحة، وعقوبات عند اتخاذ قرارات خاطئة. يُستخدم في تطبيقات مثل الألعاب، الروبوتات، والسيارات ذاتيةالقيادة . من أشهر خوارزميات التعلّم المُعزَّز:

- Q-Learning ;
- Deep Q-Network (DQN) ;
- Proximal Policy Optimization (PPO).

	التعلم الخاضع للإشراف	التعلم غير خاضع للإشراف	التعلم المعزز
نوع البيانات	بيانات مُعلَّمة	بيانات غير مُعلَّمة	تفاع بين بيئة ووكيل
الهدف	التنبؤ بالمخرجات	اكتشاف الأنماط	تعظيم المكافأة
طبيعة المخرجات	فئوية (تصنيف) أو عددية مستمرة (انحدار)	مجموعات أو تمثيل مختزل للبيانات	سياسة (Policy) لاتخاذ القرار
آلية التعلم	مقارنة التنبؤ بالقيمة الحقيقية وتقليل الخطأ	تحليل التشابه أو البنية الداخلية للبيانات	تجربة وخطأ بناءً على المكافآت والعقوبات
أشهر التطبيقات	التنبؤ بالأسعار، التنبؤ بمرض الكلى	تقسيم العملاء، تقليل الأبعاد	الألعاب، الروبوتات، السيارات ذاتية القيادة

الجدول1 : المقارنة بين الأنواع الثلاثة الرئيسية في تعلم الآلة.

3.1.2 أهمية فهم البيانات وجودتها في تعلّم الآلي

يعد فهم البيانات بشكل صحيح والعمل على تحسين جودتها أمرا ضروريا لتحقيق قوة نماذج ML إذ تعد جودة البيانات (المدخلات والمخرجات)أساس نجاح النماذج، كما تساهم في [9]:

1. **الدقة والأداء:**

- **الاتساق والكمال :**تساعد البيانات المتسقة والكاملة نماذج ML على اكتشاف الأنماط الصحيحة من خلال دراسة الأخطاء والقيم الشاذة للبيانات مما تساعد في زيادة دقة النتائج ;
- **الملاءمة:** الخصائص غير الملائمة أو المكررة قد تشتت خوارزميات التعلم، فتؤثر سلبًا على قدرة النموذج التنبؤية. لذلك تعد ملائمة البيانات ضرورية لتدريب نماذج فعّالة.

2. **الموثوقية والثقة:**

- **التحيز والعدالة:** وجود بيانات منحازة يؤدي إلى مخرجات غير عادلة ويقلل من ثقة المستخدمين في النظام. كما أن نزاهة النموذج تعتمد على تمثيل البيانات لجميع الفئات ;
- **الصلابة :**تزيد البيانات ذات الجودة العالية من قدرة النماذج على التعامل مع التغيرات غير المتوقعة والسيناريوهات الواقعية، مما يعزز الاعتماد على النماذج في التطبيقات العملية.

3. **القابلية للتوسع والتطور:**

- **استدامة المستقبل :**تحدد جودة البيانات قدرة النماذج على التكيف مع الوقت والتغيرات. كما تسمح بإعادة تدريب النماذج أو تحديثها بسهولة، ما يضمن استمرار فعاليتها.

1.3 مرض القصور الكلوي المزمن

1.1.3 تعريف المرض

تتمثل وظائف الكلى في تنقية الدم من المخلفات والسوائل الزائدة ويتم التخلص منها في البول كما تنظم توازن الكهارل في الجسم. لكن قد تصيب الكلى بعض الامراض التي تفقدها القدرة على تصفية الدم بكفاءة. ومن الامراض المزمنة الأكثر شيوعا **مرض القصور الكلوي المزمن**. الصورة التالية تمثل كلية شخص مصاب بفشل كلوي مزمن:

الصورة 2: كلية شخص مصاب بفشل كلوي مزمن . [11]

مرض القصور الكلوي المزمن ويُعرف أيضًا بالفشل الكلوي هو اضطراب أو تراجع مستمر لوظائف الكلى على مدى أشهر أو سنوات فعند الإصابة به تفقد الكلى وظائفها، مما يؤدي إلى تراكم السوائل والفضلات والمواد الكيميائية في الجسم إلى مستويات قد تكون خطيرة. ينتج المرض غالبًا عن حالات مزمنة مثل السكري وارتفاع ضغط الدم والتهاب كبيبات الكلى والأمراض الوراثية. ويتم تشخيصه عبر قياس معدل الترشيح الكبيبي وتحليل البول للكشف عن الألبومين.

تظهر القليل من مؤشرات مرض الكلى أو الأعراض في المراحل المتقدمة. وربما لا تكتشف الاصابة به إلا بعد تطوره الى مرحلة خطيرة تجعل الكلى غير قادرة على أداء وظائفها الحيوية بحيث تتطلب هذه المرحلة علاجًا تعويضيًا للوظائف مثل الغسيل الكلوي (الديلزة) أو زراعة الكلى للبقاء على قيد الحياة[10] .

2.1.3 إحصائيات المرض في الجزائر

مرض الكلى المزمن هو أحد أهم المشكلات الصحية في الجزائر، حيث تشير الدراسات إلى تزايد عدد المرضى المصابين به خلال السنوات الأخيرة نتيجة انتشار الأمراض المزمنة مثل السكري وارتفاع ضغط الدم. وتشير بعض التقديرات إلى أن عدد المصابين ب CKD بمختلف مراحله في الجزائر قد يصل إلى **3.5مليون شخص**، بينما يبلغ عدد المرضى الذين وصلوا إلى المرحلة النهائية من الفشل الكلوي حوالي **25ألف مريض** يحتاجون إلى العلاج التعويضي مثل غسيل الكلى أو زراعة الكلى. كما تشير الإحصائيات إلى تسجيل حوالي **4000حالة** جديدة سنويًا من الفشل الكلوي النهائي في الجزائر[12] .

كما أظهرت بعض الدراسات الوبائية أن معدل الإصابة بالمرحلة الخامسة من مرض الكلى المزمن المعالج في بعض مناطق الجزائر بلغ حوالي **75حالة** لكل مليون نسمة سنويًا، بينما يصل المعدل المعياري إلى حوالي **100حالة** لكل مليون نسمة، مع ملاحظة انتشار المرض بشكل أكبر لدى الذكور مقارنة بالإناث[13] .

وتشير الدراسات أيضًا إلى أن مرض السكري وارتفاع ضغط الدم يمثلان أهم الأسباب المؤدية إلى الإصابة بمرض الكلى المزمن في الجزائر، حيث يشكلان نسبة كبيرة من الحالات المسجلة، مما يجعل الوقاية من هذه الأمراض عاملًا أساسيًا في الحد من انتشار المرض [13] .

أما عالميا أصاب مرض الكلى المزمن 753 مليون شخص على مستوى العالم في عام 2016: 417 مليون أنثى و336 مليون ذكر. في عام 2015 تسببت في 1.2 مليون حالة وفاة، ارتفاعًا من 409000 في عام 1990 .الأسباب التي تساهم في أكبر عدد من الوفيات هي ارتفاع ضغط الدم عند 550.000، يليه مرض السكري عند 418.000 والتهاب كبيبات الكلى عند 238.000 [11]

3.1.3 الأسباب وعوامل الخطر

- ✓ تتسبب بعض الامراض او الحالات المرضية في إضعاف وظائف الكلى وقد ينتج عنه CKD ومن أهمها نذكر [14]:
  - **داء السكري الأول والثاني**: يؤدي ارتفاع السكر المزمن إلى تلف الأوعية الدموية الدقيقة في الكلى ;
  - **ارتفاع ضغط الدم:** الضغط المرتفع يجهد الكلى ويؤدي مع الوقت إلى تدهور وظيفتها ;
  - **التهاب كبيبات الكلى:** ومضادات الالتهاب غير الستيرويدية .وهو التهاب يصيب وحدات الترشيح الموجودة بالكلى (الكبيبات) وتؤثر على قدرتها في تنقية الدم ;

- **التهاب الكلية الخلالي**: وهو التهاب في أنابيب الكُلى والتكوينات المحيطة بها ;
- **أمراض الكلى الوراثية:** مثل تكيس الكلى الوراثي ;
- **انسداد مطوّل في المسالك البولية**: بسبب حصى الكلى أو تضخم البروستاتا. وبعض أنواع السرطان ;
- **الجزر المثاني الحالي**: وهو حالة مرضية تتسبّب في ارتجاع البول إلى الكليتين ;
- **التهاب الكلى المتكرر**: ويُسمى أيضًا التهاب الحويضة والكلية ;
- **الاستخدام الطويل لبعض الأدوية:** مثل مضادات الالتهاب غير الستيرويدية (لتخفيف الألم والتهاب وحمى).

✓ يمكن لأي شخص أن يُصاب بمرض الكلى المزمن ومع ذلك، فإن بعض الأشخاص أكثر عرضة للإصابة به من غيرهم.

وتشمل عوامل الخطر الأكثر شيوعًا لمرض الكلى المزمن ما يلي[14] :

- داء السكري ;
- ارتفاع ضغط الدم ;
- مرض القلب والأوعية الدموية ;
- التدخين ;
- السمنة ;
- العوامل الوراثية (إصابة اشخاص من العائلة بمرض الكلى);
- شذوذ بنية الكلى ;
- التقدم في العمر ;
- كثرة تناول الأدوية التي قد تضر الكليتين.

4.1.3 الأعراض ومراحل المرض

أ - الاعراض

في المراحل المبكرة ل CKD قد لا تظهر أي أعراض. وغالبا ما يتم الكشف عنه من خلال الفحوصات المخبرية المتمثلة في ارتفاع مستوى الكرياتينين في الدم أو وجود بروتين في البول. ومع تدهور وظائف الكلى تدريجيًا، تبدأ مجموعة من الاضطرابات السريرية في الظهور نتيجة اختلال التوازن الداخلي للجسم [11]:

- **اضطرابات ضغط الدم والجهاز القلبي الوعائي**

يؤدي احتباس السوائل وزيادة نشاط نظام الرينين وأنجيوتنسين إلى ارتفاع ضغط الدم. الذي يزيد العبء على القلب ويرفع خطر الإصابة بأمراض القلب والأوعية الدموية ;

- **تراكم اليوريا (اليوريميا)**

ينتج عن انخفاض قدرة الكلى على طرح الفضلات النيتروجينية تراكم اليوريا في الدم، مما يؤدي إلى حالة تُعرف باليوريميا (أعراضها التعب والخمول) ، التي تتطور إلى مضاعفات خطيرة مثل التهاب التامور واعتلال الدماغ اليوريمي. وفي الحالات المتقدمة قد تترسب بلورات اليوريا على الجلد فيما يُعرف بـ"الصقيع اليوريمي" ;

- **فرط بوتاسيوم الدم**

عند انخفاض معدل الترشيح الكبيبي إلى مستويات متقدمة تقل قدرة الكلى على إفراز البوتاسيوم مما يؤدي إلى ارتفاعه في الدم. فيسبب ذلك اضطرابات خطيرة في نظم القلب. ويزداد الأمر سوءًا في وجود الحماض الأيضي أو نقص الأنسولين ;

- **احتباس السوائل**

يؤدي ضعف طرح السوائل إلى ظهور وذمات قد تكون خفيفة في الأطراف السفلية، أو شديدة مثل وذمة الرئة التي تمثل حالة طبية طارئة ;

- **اضطراب توازن الفوسفات والكالسيوم**

يؤدي انخفاض طرح الفوسفات إلى فرط فوسفات الدم، مما يساهم في تكلس الأوعية الدموية وزيادة خطر أمراض القلب. كما يحدث نقص في الكالسيوم نتيجة انخفاض إنتاج الكولسيتريول، مما يحفز فرط نشاط الغدة جار الدرقية الثانوي. ومع الوقت قد يتطور الأمر إلى اعتلال عظمي كلوي وتكلسات وعائية ;

- **الحماض الأيضي**

ينتج عن انخفاض قدرة الكلى على التخلص من الأحماض (قلة إنتاج ما يكفي من الامونيا)، مما يؤدي إلى اضطراب في وظيفة الإنزيمات وزيادة استثارة العضلة القلبية والجهاز العصبي ;

- **فقر الدم**

يُعد فقر الدم من المضاعفات الشائعة، ويعود بشكل أساسي إلى نقص إنتاج هرمون الإريثروبويتين من الكلى، وعوامل أخرى مثل الالتهاب المزمن وتأثير السموم اليوريمية على نخاع العظم ;

- **تسارع تصلب الشرايين**

يتعرض مرضى CKD لتطور تصلب الشرايين بشكل أسرع مقارنة بعامة الناس، مما يزيد من معدلات الإصابة بأمراض القلب والأوعية الدموية.

ب- مراحل المرض

يتم تصنيف المرض اعتمادًا على معدل الترشيح الكبيبي(GFR) ، بشرط استمرار الانخفاض أو وجود تلف كلوي لمدة لا تقل عن 3 أشهر بحيث يوجد خمس مراحل أساسية:

1. **المرحلة الاولى**: غالبا لا توجد أعراض، ويُكتشف المرض عبر التحاليل المخبرية:
   - GFR ≥ 90 مل/دقيقة/1.73م² ;
   - وظيفة كلوية طبيعية أو مرتفعة نسبيًا ;
   - وجود دليل على تلف كلوي (مثل بيلة الألبومين المستمرة).
2. **المرحلة الثانية**: أيضا تكون بدون أعراض واضحة:
   - GFRبين 60–89 مل/دقيقة/1.73م²;
   - انخفاض طفيف في الوظيفة الكلوية;
   - وجود تلف كلوي مثبت مخبريًا أو شعاعيًا .
3. **المرحلة الثالثة**: فيها تبدأ المضاعفات بالظهور مثل: فقر الدم واضطرابات ضغط الدم واختلال توازن المعادن:
   - GFRبين 30–59 مل/دقيقة/1.73م² ;

- انخفاض متوسط في الوظيفة الكلوية.

وتُقسم إلى:

- المرحلة 3أ: (59–45).
- المرحلة 3ب: (44–30).

4. **المرحلة الرابعة**: تزداد الأعراض وضوحًا في هذه المرحلة، ويبدأ التحضير للعلاج التعويضي (غسيل الكلى أو الزراعة):
   - GFRبين 15–29 مل/دقيقة/1.73م².
   - انخفاض شديد في وظائف الكلى.

5. **المرحلة الخامسة**: في هذه المرحلة يحتاج المريض الى غسيل الكلى المنتظم او زراعة كلية:
   - GFRأقل من 15 مل/دقيقة/1.73م².
   - فشل كلوي مؤكد.

5.1.3 طرق التشخيص

يتم تشخيص CKD من خلال مجموعة من الفحوصات التي تختبر وظيفة الكليتين ومدى تضررهما، وأهمها [16] ما يلي:

- **فحص الدم:** فحص الدم هو الاختبار الأساسي للكشف عن أمراض الكلى، حيث يقيس مستوى الكرياتينين(هو أحد نواتج الفضلات في الدم) .يقوم الطبيب بحساب ما يُعرف بمعدل الترشيح الكبيبي المقدر(eGFR) بناءً على نتيجة الكرياتينين، مع عمر المريض وجنسه وبعض العوامل الجسدية وهو مؤشر يوضح كمية الدم التي تستطيع الكليتان تصفيتها في الدقيقة.

في الحالة الطبيعية، تكون الكليتان قادرتين على تصفية أكثر من 90 مل/دقيقة. أما إذا انخفض معدل eGFR عن هذا المستوى واستمر الانخفاض، فقد يشير ذلك إلى وجود مرض الكلى المزمن.

- **اختبار البول:** تحليل البول أساسي ومهم لاستكمال الفحوصات، ويهدف إلى:
  - قياس نسبة الألبومين إلى الكرياتينين في البول(ACR) ، والتي تعتبر مؤشر مهم على وجود تلف في الكلى.
  - الكشف عن وجود دم أو بروتين في البول، حيث إن ظهورهما قد يدل على خلل في وظيفة الترشيح الكلوي.

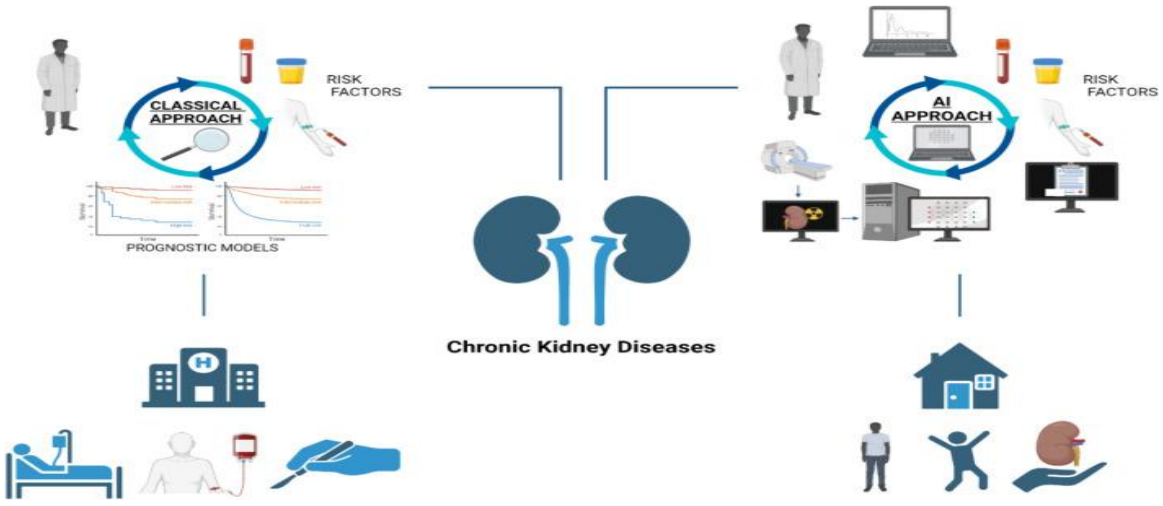
- تساعد اختبارات البول ومعدل الترشيح الكبيبي المقدر **(eGFR)** في إعطاء صورة أكثر دقة عن كفاءة عمل الكليتين ودرجة التلف.

➢ **<u>اختبارات إضافية:</u>** قد يلجأ الطبيب إلى فحوصات أخرى في بعض الحالات لتقييم مستوى الضرر أو معرفة السبب وراء المرض، مثل:

- الفحوصات التصويرية، كالتصوير بالموجات فوق الصوتية أو Computed،Magnetic resonance imaging tomography، وذلك لدراسة شكل الكليتين والكشف عن أي انسدادات أو تشوهات.
- خزعة الكلى وهي اخذ عينة صغيرة من نسيج الكلية وتُفحص مجهريًا للبحث عن علامات الالتهاب أو التلف.

1.4 الدراسات السابقة حول التنبؤ بمرض الكلى

أصبحت تقنيات ML تحظى باهتمام واسع خاصة في المجال الطبي مثل التنبؤ بالأمراض المزمنة كمرض القصور الكلوي المزمن(CKD) . كما أن العديد من الدراسات سعت إلى تطوير نماذج تصنيف تعتمد على البيانات المخبرية والسريرية لتحسين دقة التشخيص المبكر وتقليل المضاعفات. فالكشف المبكر يقلل من الوصول إلى المرحلة النهائية. الشكل التالي يوضح الفرق بين المنهج التقليدي والمنهج المعتمد على تقنيات الذكاء الاصطناعي في تحليل بيانات مرضى الكلى المزمن.



الصورة 3:المقارنة بين المنهج التقليدي والمنهج المعتمد على الذكاء الاصطناعي في تشخيص مرض الكلى المزمن. [17]

1.1.4 الخوارزميات المستخدمة

تناولت دراسات سابقة حول التنبؤ بمرض **CKD** مجموعة متنوعة من خوارزميات ت_آ ، حيث تنوعت أساليب النماذج في العديد من الدراسات، مثل الاستعراض في [18] **المعنون** : "

A Systematic Review of Artificial Intelligence and Machine Learning Approaches" (دراسة مراجعة تناولت 13 دراسة ذات صلة تستخدم خوارزميات متنوعة في مجال الذكاء الاصطناعي) تم تطبيق خوارزميات **Logistic Regression** كأحد الأساليب الأساسية في التصنيف الطبي، إلى جانب **Support Vector Machine (SVM)** و **Random Forest (RF)**، التي تتعتبر متعددة الاستخدامات وفعّالة في التعامل مع البيانات ذات الابعاد المتعددة. كما استخدمت بعض الدراسات الشبكات العصبية الاصطناعية **(Neural Networks)** وتقنيات التعلم العميق القادرة على اكتشاف العلاقات غير الخطية بين المتغيرات، وتم وضع هذه الخوارزميات ضمن الأدوات الشائعة للتنبؤ بتقدّم **CKD** نحو مراحل أكثر تقدّمًا مثل الفشل الكلوي النهائي**(ESRD)** .

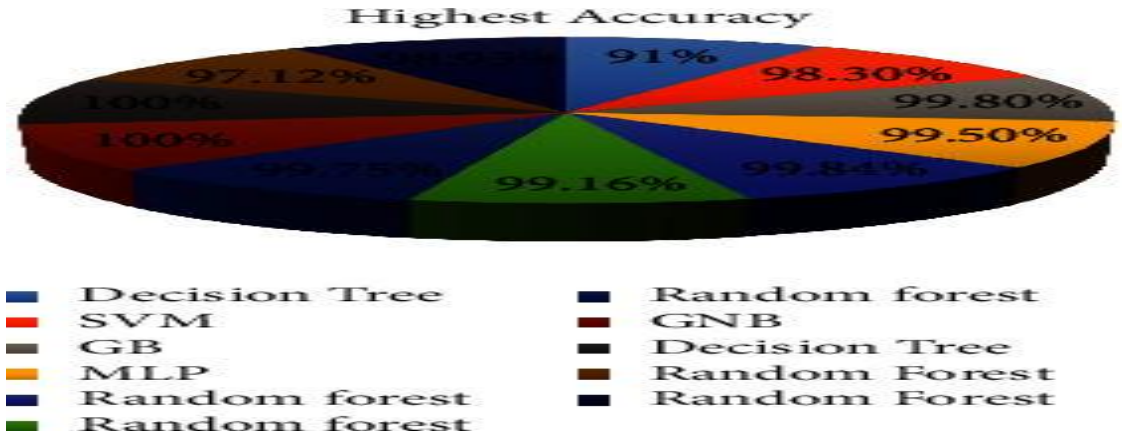
كما أن الباحثين في الدراسة[19] بعنوان **Journal of Big Data** طبقوا بنجاح خوارزميات **Random Forest** و **SVM** و**Decision Tree** على بيانات CKD للتصنيف الثنائي (وجود المرض أم لا) والتصنيف متعدد الفئات لتحديد مراحل المرض، مع طريقتين لاختيار الميزات **RFECV** (Recursive Feature Elimination with Cross-Validation) و **UFS**(Univariate Feature Selection) تحسين أداء النماذج.

وفي الدراسة [20] **باسم "Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease "**، تم استخدام مجموعة من الخوارزميات مع التركيز على بناء **نموذج هجين** يجمع بين عدة مصنفات لتحقيق أعلى دقة وتقليل التلبية الزائدة (Overfitting) حيث تشمل:

- **المصنفات الأساسية**:
  - **Gaussian Naive Bayes (GNB) :** مصنف احتمالي يعتمد على قاعدة بايز ;
  - **Decision Tree (DT) :** مصنف شجري لتقسيم البيانات حسب الميزات ;
  - **Gradient Boosting (GB) :** خوارزمية تعزيز تدريجي لتحسين الأداء.
- **المصنف التلوي**:
  - **Random Forest (RF) :** غابة عشوائية من أشجار القرار لتحسين الدقة العامة للنموذج الهجين.

- **خوارزميات أخرى:** K-Nearest Neighbors (KNN)، Support Vector Machine (SVM)، Linear Discriminant Analysis (LDA)، Neural Networks (NN).
- **النموذج الهجين** (GNB + DT + GB مع RF تلوي) حقق دقة تصل إلى 100٪ على مجموعة بيانات **UC**.

الصورة التالية توضح مقارنة بين خوارزميات ML لدراسة النموذج الهجين للتنبؤ بـCKD للدراسة **" Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease ":**



الصورة 4: مقارنة بين خوارزميات ML لدراسة النموذج الهجين للتنبؤ بـ CKD. [20]

2.1.4 مقارنة نتائج الدراسات

في الأعمال السابقة استخدم الباحثون تقنيات مختلفة من ML لتصنيف مرض الكلى المزمن من خلال استخدام بيانات المرضى في مختلف الأبحاث منها:

✓ **بحث Charleonnan و آخرون [21]**

قاموا بتطبيق هذه النماذج التنبؤية KNN، SVM، Decision Tree وLinear Regression على مجموعة بيانات CKD هندية من أجل اختيار أفضل نموذج للتنبؤ، ثم قارنوا بين النتائج المتحصل عليها من هذه النماذج وتوصلوا إلى أن SVM لديه أعلى دقة تصنيف بلغت 98.3% وأعلى حساسية بلغت 0.99.

✓ **بحث Salekin و Stankovic [22]**

استخدما KNN، Random Forest وANN على 400 حالة من مجموعة بيانات CKD، كما طبقا Wrapper feature selection لاختيار السمات وتم اختيار 5 سمات لإجراء الدراسة. النتائج هي أن أعلى دقة تصنيف كانت 98% باستخدام RF وخطأ جذر متوسط مربعات (RMSE) بلغ 0.11.

✓ **بحث S. Tekale وآخرون [23]**

طبقوا نموذج التنبؤ Decision Tree وSVM على مجموعة بيانات CKD والتي تحتوي على 400 حالة و25 سمة، وهذا بعد أن تمت معالجة البيانات مسبقا وتم تقليل عدد السمات من 25 سمة الى 14 سمة. وقد تم اعتبار SVM نموذجًا أفضل بدقة بلغت 96.75%.

✓ **بحث Xiao وآخرون [24]**

اقترحوا استخدام عدة طرق للتنبؤ بتطور CKD وهي: logistic regression, Elastic Net, lasso regression, ridge regression, support vector machine, random forest, XGBoost, neural network وk-nearest neighbor ثم قارنوا نتائج هذه التقنيات اعتمادا على أدائها. تم تطبيق هذه التقنيات على مجموعة بيانات تاريخية ل 551 مريض مع البروتين في البول و18 سمة، وتم تصنيف النتائج الى خفيف، متوسط وشديد. وتوصلوا إلى أن logistic regression كان الأفضل مع AUC بلغ 0.873، وحساسية وخصوصية بلغت 0.83 و0.82 على التوالي.

✓ **بحث Mohammed و Beshah [25]**

كان بحثهم حول تطوير نظام معرفة ذاتي التعلم لتشخيص وعلاج المراحل الثلاث الأولى من CKD باستخدام ML. استخدموا عدد قليل من البيانات وطوروا نموذج أولي يسمح للمريض من الاستعلام من نظام إدارة المعرفة (KBS)للاطلاع على كيفية تقديم النصائح. كما استخدموا Decision Tree لتوليد القواعد، وقد بلغت دقة هذا النموذج الأولي 91%.

✓ **بحث Almasoud و Ward [25]**

تم فيه اختبار خوارزميات ML لمعرفة مدى قدرة تنبؤها بCKD، وذلك باستخدام مجموعة (Pearson correlation) فرعية من السمات كما تم استخدام معامل ارتباط بيرسون، ANOVA، واختبار Cramer's V وذلك من أجل اختيار السمات التنبؤية . أما بالنسبة لنماذج التدريب فقد استعانوا ب Liear Regression , SVM , Random Forest و Gardient Boosting . النتائج أشارت الى أن Gradient Boosting لديه أعلى دقة مع F-measure بلغت 99.1.

✓ **بحث Vinod [26]**

استخدم سبع خوارزميات ML الخاضعة للاشراف وهي K-Nearest Neighbor, Decision Tree, Support vector Machine (SVM),Random Forest, Neural Network, Naïve Bayes و Logistic Regression والهدف من ذلك هو ايجاد النموذج الأنسب لتنبؤ ب Basic Clinical Dataset (هو مجموعة بيانات سريرية أساسية تم جمعها من المرضى) وذلك اعتمادا على تقييم أداء كل نموذج . توصل إلى أن نموذج KNN هو الأفضل على مجموعة بيانات BCD بدقة بلغت 97%.

✓ في الدراسة [19] باسم" Journal of Big Data "حقق نموذج SVMو Random Forestأداءً ممتازًا في التصنيف الثنائي مع دقة بلغت حوالي 99%، بينما كان أداء بعض الخوارزميات الأخرى مثل XGBoostممتازًا أيضًا في التصنيف متعدد الفئات، مما يشير إلى قوة خوارزميات التعزيز Ensemble Methods في تعلّم الأنماط المتعددة ببيانات CKD كما أبرزت الدراسات الأوسع والمراجعات أن اختيار الميزات (Feature Selection) والتحقق المتقاطع (Cross-Validation) لعبا دورًا مهمًا في تحسين النتائج، بينما بعض الخوارزميات التقليدية قدّم أداءً جيدًا لكنه أقل من نماذج التعزيز أو الشبكات العصبية المتقدمة في بعض حالات مجموعات البيانات الكبيرة.

3.1.4 الفجوات البحثية

على الرغم من توظيف تقنيات ML في التنبؤ ب CKD والوصول لنتائج جيدة، إلا أن الدراسات لا تزال تكتشف العديد من الفجوات البحثية التي تحد من التطبيق السريري الواسع لهذه النماذج والمتمثلة في :

- **محدودية قابلية التفسير للنماذج المتقدمة**

أوضحت بعض الدراسات أن نماذج التعلم المعزز (Gradient Boosting) والشبكات العصبية العميقة تحقق مستويات مرتفعة من الدقة ومنطقة تحت المنحنى(AUC) ، غير أنها تعاني من ضعف في الشفافية وقابلية التفسير[27] .

كما أكدت المراجعة المنهجية المنشورة في Cureusأن غالبية الدراسات ركزت على مؤشرات الأداء الإحصائية دون توفير تفسير سريري واضح يدعم اتخاذ القرار الطبي[28] .

- لذلك من الأفضل دمج تقنيات تفسير النماذج مثل SHAP وLIME لتعزيز الثقة السريرية وتحقيق التكامل بين الذكاء الاصطناعي والممارسة الطبية.

- **محدودية حجم البيانات ونقص التنوع السكاني**

تشير أغلب الدراسات المقارنة في Journal of Big Data إلى أن جزءًا كبيرًا من الأبحاث يعتمد على قواعد بيانات صغيرة أو بيانات مأخوذة من منطقة واحدة (مركز طبي) مما يحد من قدرة النماذج على التعميم [27].

كما بينت بعض المراجعات أن هناك نقص في البيانات الطولية (Longitudinal Data) وغياب التنوع الجغرافي والسكاني في العينات المستخدمة[29] .

  - من الأفضل استخدام قواعد بيانات كبيرة وتمثيلًية للفئات السكانية المختلفة لتعميم نتائج.

- **غياب التحقق الخارجي للنماذج(External Validation)**

أكدت المراجعة أن اغلب الدراسات اكتفت بتقسيم البيانات إلى مجموعتي تدريب واختبار، دون إجراء تحقق خارجي باستخدام بيانات مستقلة من مؤسسات طبية مختلفة[28] .

  - إن الأداء داخل قاعدة بيانات واحدة لا يضمن الكفاءة عند تطبيقه في بيئات مختلفة من البيانات.

- **التركيز على مؤشرات أداء محدودة**

أشارت بعض الدراسات إلى أن هناك أبحاث تعتمد بشكل أساسي على مؤشر الدقة(Accuracy) ، دون الاهتمام الكافي بمؤشرات أخرى ذات أهمية سريرية مثل الحساسية(Sensitivity) ، النوعية(Specificity) ، ومعدلF1 او AUC [27].

  - إن التقييم متعدد المقاييس يضمن تقييم أكثر شمولية وواقعية وتفادي النتائج المضللة.

- **محدودية المقارنة بين النماذج التقليدية والحديثة**

بعض المراجعات تشير إلى أن اغلب الدراسات اقتصرت على مقارنة نماذج تقليدية مثل الانحدار اللوجستي وSVM، دون إدراج نماذج حديثة قائمة على Boosting مثل XGBoost وCatBoost ، أو نماذج التعلم العميق[29] .

  - المقارنة الشاملة تحدد النموذج الأمثل في سياقات سريرية مختلفة.

- **قلة الدراسات المتعلقة بالتنبؤ المبكر**

تشير الدراسات الحديثة إلى أن معظم النماذج تركز على التنبؤ بوجود المرض بعد ظهور مؤشرات سريرية واضحة، في حين لا تزال الأبحاث المتعلقة بالتنبؤ بالمراحل المبكرة محدودة نسبيًا[29] .

  - يساهم الكشف المبكر في الحد من تطور المرض إلى الفشل الكلوي

1.5 الخاتمة

في هذا الفصل تناولنا الجزء النظري لشرح بعض المفاهيم الأساسية التي تخدم موضوع دراستنا والمتمثلة في مفاهيم الذكاء الاصطناعي والتعلم الآلي ومرض القصور الكلوي. تطرقنا في عنوان الذكاء الاصطناعي لتعريفه وذكر أهميته وأنواعه ومجالاته المختلفة مع التركيز على تطبيقاته في المجال الطبي وتم توضيح أبرز مزاياه وتحدياته. أما في عنوان التعلم الآلي استعرضنا مفهومه وأنواعه وبينا أهمية جودة البيانات وفهمها في تحسين أداء النماذج التنبؤية خاصة في التطبيقات الطبية الحساسة التي تتطلب دقة وموثوقية عاليتين . وفي العنوان الموالي المتمثل في مرض القصور الكلوي قمنا بتعريف المرض وأسبابه وعوامل الخطر وأيضا الأعراض وطرق التشخيص حيث أن هذا العنوان يوضح الفهم الطبي الدقيق لمشكلة دراستنا.

وأخيرًا، قمنا بمراجعة الدراسات السابقة المتعلقة بالتنبؤ بمرض الكلى، مع تحليل الخوارزميات المستخدمة، ومقارنة نتائجها، وتسليط الضوء على الفجوات البحثية التي ما تزال بحاجة إلى معالجة.

يشكّل هذا الفصل النظري تمهيد للفصول التطبيقية اللاحقة، حيث سيتم في الفصل الثاني دراسة قاعدة البيانات المستخدمة والمعالجة المسبقة لها وفي الختام سنقوم بتوظيف تقنيات تعلّم الآلة لبناء نموذج تنبؤي يساهم في الكشف المبكر عن مرض القصور الكلوي المزمن يهدف إلى دعم القطاع الصحي.

الفصل الثاني: المراحل الأساسية المتبعة

1.1 مقدمة

لبناء نماذج التنبؤ الفعالة في مجال التعلم الالي يجب فهم البيانات المستخدمة وتحليلها، خاصة في التطبيقات الطبية التي تتطلب دقة عالية وموثوقية في النتائج. فنجاح أي نموذج تنبؤي لا يعتمد فقط على الخوارزميات المستخدمة، بل يرتبط بشكل كبير بجودة البيانات ومدى معالجتها بشكل صحيح.

يتناول هذا الفصل المراحل الأساسية المتبعة في إعداد البيانات المستخدمة وبناء نماذج التنبؤ وفقا للمنهجية العلمية، حيث يبدأ الفصل بتحليل قاعدة البيانات وفهم خصائصها من خلال التحليل الاستكشافي، ثم ينتقل إلى مرحلة المعالجة المسبقة التي تهدف إلى معالجة القيم المفقودة والقيم الشاذة، إضافةً إلى تحويل البيانات إلى صيغة مناسبة للنمذجة. كما يتم التطرق لتقنيات اختيار السمات الأكثر أهمية، وصولًا إلى مرحلة تقسيم البيانات وتدريب مجموعة من نماذج التعلم الآلي الذي تسمح بمقارنة النتائج لاحقا.

1.2 التحليل الاستكشافي للبيانات(Exploratory Data Analysis)

التحليل الاستكشافي للبيانات يُعد مرحلة أساسية في بناء نماذج التعلم الآلي، حيث يهدف إلى فهم خصائص البيانات واكتشاف الأنماط والعلاقات بينها. كما يساعد في تحديد المشاكل المحتملة مثل القيم المفقودة، القيم الشاذة، وعدم توازن البيانات، فهو يساهم في اختيار تقنيات المعالجة المسبقة المناسبة وتحسين أداء النماذج التنبؤية.

وقد تم في هذه الدراسة إجراء التحليل اعتمادًا على مجموعة من الأدوات الإحصائية باستخدام لغةPython ، وذلك لفهم بنية البيانات وجودتها قبل مرحلة النمذجة.

1.3 المعالجة المسبقة للبيانات(Preprocessing)

تُعد مرحلة المعالجة المسبقة من أهم مراحل إعداد البيانات في مشاريع التعلم الآلي، حيث تهدف إلى تحسين جودة البيانات ومعالجة المشكلات التي قد تؤثر على أداء النماذج، مثل القيم المفقودة والقيم الشاذة، بالإضافة إلى تحويل البيانات إلى صيغة عددية مناسبة للتعلم.

وقد تم في هذه الدراسة تطبيق مجموعة من التقنيات لمعالجة البيانات، شملت تعويض القيم المفقودة، والكشف عن القيم الشاذة ومعالجتها، وترميز المتغيرات الوصفية، إضافةً إلى تقييس البيانات واختيار أهم السمات.

1.1.3 معالجة القيم المفقودة(Imputation)

نظرًا لاحتواء قاعدة البيانات على قيم مفقودة(NaN) ، تم تطبيق عدة طرق لتعويض هذه القيم، بهدف الحفاظ على أكبر قدر ممكن من البيانات وتحسين جودة التحليل وتشمل هذه الطرق:

1 التعويض بالوسيط(Median Imputation)

يُعرّف التعويض بالوسيط بأنه تقنية إحصائية تُستخدم لاستبدال القيم المفقودة بالقيمة الوسيطية للمتغير، حيث يُمثل الوسيط القيمة التي تقع في منتصف الترتيب التصاعدي للبيانات. وتُعد هذه الطريقة مناسبة بشكل خاص للبيانات ذات التوزيع غير الطبيعي أو التي تحتوي على قيم شاذة، نظرًا لكون الوسيط مقياسًا قويًا (Robust Measure) لا يتأثر بالقيم المتطرفة[33] .

2 خوارزمية أقرب الجيران(K-Nearest Neighbors Imputation)

يُعرّف التعويض باستخدام خوارزمية KNN بأنه أسلوب يعتمد على إيجاد أقرب k جيران مشابهة للسجل الذي يحتوي على قيمة مفقودة، وذلك باستخدام مقاييس المسافة مثل المسافة الإقليدية. ويتم بعد ذلك تقدير القيمة المفقودة اعتمادًا على قيم هذه الجيران، مما يسمح بالحفاظ على البنية المحلية للبيانات واستغلال العلاقات بين المتغيرات[34] .

3 التعويض باستخدام الانحدار (Regression Imputation)

يُعرّف التعويض باستخدام الانحدار بأنه تقنية تعتمد على بناء نموذج إحصائي (مثل الانحدار الخطي) للتنبؤ بالقيم المفقودة، من خلال استخدام المتغيرات الأخرى كمتغيرات مستقلة. ويُتيح هذا الأسلوب استغلال العلاقات الوظيفية بين المتغيرات، مما يؤدي إلى تقديرات أكثر دقة مقارنة بالطرق البسيطة، خاصة في حال وجود ارتباط قوي بين الخصائص [33].

2.1.3 القيم الشاذة (Outliers)

تُعرّف القيم الشاذة بأنها القيم التي تنحرف بشكل كبير عن بقية البيانات، وقد تؤدي إلى تشويه نتائج التحليل.

1. الكشف عن القيم الشاذة باستخدام المدى الربيعي (Interquartile Range- IQR)

قد تم استخدام طريقة المدى الربيعي للكشف عن هذه القيم، حيث يتم حساب:

- الربع الأول (Q1)
- الربع الثالث(Q3)
- المدى الربيعي: IQR=Q3−Q1

وتُعتبر القيم التي تقع خارج المجال: $[Q1 - 1.5 \times IQR\ ,\ Q3 + 1.5 \times IQR]$ قيمًا شاذة.[34]

2. معالجة القيم الشاذة (Outliers Handling)

بعد تحديد القيم الشاذة، تم تطبيق تقنيتين للتقليل من تأثير هذه القيم على البيانات والنتائج النهائية.

- **التحويل الكمي(Quantile Transformation)**

يُعرّف بأنه تقنية إحصائية تهدف إلى إعادة توزيع البيانات من خلال تحويلها إلى توزيع محدد مثل التوزيع المنتظم (Uniform)أو التوزيع الطبيعي(Normal) ، وذلك بالاعتماد على ترتيب القيم(Ranking) تقوم هذه الطريقة بتكييف القيم الأصلية مع القيم المقابلة لها في التوزيع المستهدف، مما يؤدي إلى تقليل تأثير القيم الشاذة وضغط القيم المتطرفة ضمن نطاق أكثر توازنًا، دون حذفها.

وتُستخدم هذه التقنية بشكل خاص عندما تكون البيانات غير موزعة بشكل طبيعي، حيث تساعد في تحسين استقرار النماذج وزيادة كفاءتها. [33]

- **تقليص القيم المتطرفة(Winsorization)**

تُعرّف بأنها أسلوب إحصائي يُستخدم للحد من تأثير القيم الشاذة من خلال استبدال القيم المتطرفة بقيم حدية (Boundary Values)ضمن نطاق معين، بدلاً من حذفها .حيث يتم تحديد حدود دنيا وعليا (مثل النسب المئوية 3% و97%)، ثم يتم تعويض القيم التي تتجاوز هذه الحدود بأقرب قيمة داخل النطاق المحدد. وتُعد هذه الطريقة مفيدة في

الحفاظ على حجم البيانات وتقليل تأثير القيم المتطرفة، مع الحد من الخلل الذي قد تسببه هذه القيم في التحليل الإحصائي والنماذج التنبؤية. [35]

3.1.3 ترميز المتغيرات الوصفية(Categorical Encoding)

تحتوي قاعدة البيانات على متغيرات وصفية، لذلك يجب تحويلها إلى صيغة عددية باستخدام تقنيات الترميز، التي تُعرّف بأنها تحويل البيانات غير العددية إلى قيم رقمية لأن معظم خوارزميات التعلم الآلي لا يمكنها التعامل مع البيانات النصية بشكل مباشر، بل تتطلب مدخلات عددية لإجراء العمليات الحسابية.

وقد تم في هذه الدراسة تطبيق عملية الترميز (Encoding) على المتغيرات الوصفية، خاصة في حالة المتغيرات الثنائية (Binary Variables)مثل (Yes/No) إلى (0 و1)، مع الحفاظ على الترتيب في المتغيرات الترتيبية.

4.1.3 التقييس باستخدام (Min-Max Normalization)

تُعد عملية التقييس (Scaling) من الخطوات الأساسية في مرحلة المعالجة المسبقة للبيانات، حيث تهدف إلى توحيد نطاق القيم بين المتغيرات المختلفة، مما يساهم في تحسين أداء نماذج التعلم الآلي، خاصة تلك التي تعتمد على حساب المسافات بين البيانات.

وفي هذه الدراسة، تم استخدام تقنية **Min-Max Normalization** لتقييس المتغيرات العددية، حيث تعمل على تحويل القيم الأصلية إلى مجال محدد، غالبًا بين **[0, 1]**، دون تغيير شكل توزيع البيانات.

- ✓ **Min-Max Normalization** : هي عملية تحويل خطي (Linear Transformation) يتم من خلالها إعادة ضبط قيم المتغير بحيث تقع ضمن نطاق محدد، اعتمادًا على القيم الدنيا والعليا لذلك المتغير.
- ✓ يتم حساب القيمة وفق العلاقة التالية[34] :

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

حيث: Xالقيمة الأصلية / **Xmin**أصغر قيمة في المتغير / **Xmax**أكبر قيمة في المتغير.

5.1.3اختيار السمات(Feature Selection)

عملية اختيار السمات هي من المراحل الاختيارية التي تسبق بناء نماذج التعلم الآلي، حيث تهدف إلى اختيار أهم المتغيرات المؤثرة في المتغير الهدف، مع حذف المتغيرات غير المهمة أو المكررة، مما يساهم في تحسين دقة النماذج وتقليل التعقيد الحسابي .وقد تم في هذه الدراسة استخدام عدة تقنيات لاختيار السمات والمتمثلة في:

- ✓ **اختباركاي-تربيع(Chi-Square)**

يُعد اختبار كاي-تربيع من الأساليب الإحصائية المستخدمة لقياس درجة الاعتماد بين المتغيرات الوصفية والمتغير الهدف، حيث يعتمد على مقارنة التكرارات الفعلية بالتكرارات المتوقعة في حالة الاستقلالية. ويُستخدم هذا الاختبار لتحديد أهمية المتغيرات في عملية التنبؤ، حيث تُشير القيم المرتفعة لإحصائية كاي-تربيع إلى وجود علاقة قوية بين المتغير والمتغير الهدف، مما يجعله أكثر تأثيرًا في نماذج التصنيف[34] .

- ✓ **المعلومات المتبادلة (Mutual Information)**

تُعرّف المعلومات المتبادلة بأنها مقياس إحصائي مستمد من نظرية المعلومات، يُستخدم لقياس مقدار المعلومات التي يوفرها متغير معين حول متغير آخر، حيث تعكس درجة الاعتماد الإحصائي بينهما، سواء كانت العلاقة خطية أو غير خطية. وتُستخدم في تقييم أهمية المتغيرات، حيث تسمح باكتشاف العلاقات المعقدة بين الخصائص والمتغير الهدف، دون افتراض شكل محدد للعلاقة. كما تُعد قيمة المعلومات المتبادلة مؤشرًا على قوة العلاقة، إذ كلما زادت هذه القيمة دلّ ذلك على وجود ارتباط أقوى بين المتغير والمتغير الهدف، مما يجعله أكثر أهمية في عملية التنبؤ والتصنيف [33].

- ✓ **خوارزمية الإزالة التكرارية (Recursive Feature Elimination)**

خوارزمية الإزالة التكرارية للسمات تُعد من تقنيات الاختيار المعتمدة على النماذج، حيث تعتمد على تدريب نموذج تعلم آلي على مجموعة البيانات، ثم تقوم بإزالة السمات الأقل أهمية بشكل تدريجي اعتمادًا على أوزان أو معاملات النموذج. وتتم هذه العملية بشكل تكراري، حيث يتم في كل خطوة إعادة تدريب النموذج بعد حذف أضعف سمة، إلى أن يتم الوصول إلى مجموعة السمات الأكثر أهمية. وتُساهم هذه الطريقة في تحسين أداء النموذج من خلال تقليل عدد المتغيرات غير المهمة، كما تساعد في تقليل التعقيد الحسابي وزيادة قابلية تفسير النتائج[33] .

1.4 تقسيم البيانات (Train/Test Split)

تُعد عملية تقسيم البيانات من الخطوات الأساسية في تقييم نماذج التعلم الآلي، حيث يتم تقسيم البيانات إلى مجموعتين رئيسيتين:

- **مجموعة التدريب (Training Set):** التي تُستخدم في بناء النماذج وتعلّم الأنماط والعلاقات الموجودة في البيانات.
- **مجموعة الاختبار (Test Set):** والتي تُستخدم لتقييم أداء النماذج على بيانات لم يتم رؤيتها مسبقًا، ما يسمح بقياس قدرتها على التعميم.

1.5 تدريب النماذج(Training)

هي عملية يتم من خلالها تدريب نماذج التعليم الالي باستخدام بيانات مجموعة التدريب (80% مع تقنية 10-FCV) حيث تتمكن النماذج من اكتشاف العلاقات والانماط بين المتغيرات للتنبؤ بالنتائج على بيانات جديدة حيث تم استخدام ما يلي:

✓ **شجرة القرار(Dicision Tree)**

شجرة القرار هي نموذج تعلم آلي تستخدم في التصنيف، تعتمد على تقسيم البيانات بشكل تكراري إلى مجموعات فرعية متجانسة. ويتم هذا التقسيم من خلال اختيار السمات التي تحقق أفضل معيار للتقسيم مثل كسب المعلومات (Information Gain) أو مؤشر جيني (Gini Index).

هذا النموذج يتم تمثيله كبنية شجرية حقيقة يتكون من عقد داخلية تمثل اختبارات على قيم الخصائص، وفروع تمثل نتائج هذه الاختبارات، وأوراق نهائية تمثل التنبؤات (التصنيف) [36].

✓ **الغابة العشوائية (Random Forest)**

الغابة العشوائية هي أسلوب تجميعي يعتمد على إنشاء مجموعة من أشجار القرار باستخدام تقنية Bootstrap **Bagging** (Aggregatio). حيث يتم تدريب كل شجرة على عينة عشوائية من البيانات مع التكرار، ثم في كل عقدة في الشجرة يتم اختيار مجموعة فرعية عشوائية من السمات (تجربة عدة تقسيمات واختيار أفضل نقطة تقسيم). ويتم دمج نتائج الأشجار من خلال التصويت(Voting) في حالة التصنيف أو المتوسط في حالة الانحدار، مما يؤدي إلى تحسين الدقة وتقليل التباين وزيادة قدرة النموذج على التعميم[37] .

- ✓ **الأشجار المتطرفة العشوائية (Extra Trees)**

الأشجار العشوائية المتطرفة هي تحسين لنموذج Radom Forest، حيث تزيد من العشوائية أثناء بناء الأشجار من خلال اختيار نقاط تقسيم عشوائية بدلا من اختيار أفضل تقسيم (الغابة العشوائية). تسمح هذه الطريقة ب تحسين سرعة التدريب وتقليل التباين، اما في البيانات المتعددة الابعاد تحافظ على الأداء الأقوى[38] .

- ✓ **تعزيز التدرج المحسّن(XGboost)**

تعزيز التدرج المحسّن (XGBoost) هو من النماذج المتقدمة في التعلم التجميعي (دمج عدة نماذج للحصول على نموذج قوي ودقيق)، يعتمد على خوارزمية Gradient Boosting لبناء سلسلة من النماذج الضعيفة بشكل متتابع.

كل نموذج جديد لا يبدأ من الصفر بل يحاول تصحيح أخطاء النماذج السابقة من خلال تقليل دالة الخطأ. كما يتضمن XGBoost تقنيات تنظيم (Regularization) للحد من الافراط والتحكم في تعقيد النموذج مما يجعله فعالا في معالجة البيانات المعقدة والكبيرة[39] .

- ✓ **التكديس(Stacking)**

التكديس هو من تقنيات التعلم التجميعي التي تحسن الأداء من خلال تدريب نماذج مختلفة وجمع التنبؤات (أي اعتبار مخرجات كل النماذج كمدخلات جديدة). باستخدام نموذج ثانوي يعرف ب(**Meta-Learner**) يتم دمج التنبؤات، وهذا ما يسمح بالاستفادة من نقاط القوة لكل نموذج وتحقيق أداء تنبؤي قوي "عدة عقول تفكر ثم عقل أخير يقرر النتيجة الأفضل" [40].

1.6 الخاتمة

يمثل هذا الفصل المنهج المتكامل الذي يهدف إلى عرض المراحل الأساسية المعتمدة في إعداد البيانات وبناء نماذج التعلم الآلي بطريقة متسلسلة، حيث تم فيه التطرق إلى التحليل الاستكشافي للبيانات باعتباره خطوة أولية لفهم قاعدة البيانات واكتشافها. يليه عرض مرحلة المعالجة المسبقة، وتشمل معالجة القيم المفقودة والقيم الشاذة، إلى جانب عمليات ترميز المتغيرات الوصفية وتقييس البيانات، بما يضمن تهيئة البيانات بالشكل المناسب لمرحلة النمذجة. وتم كذلك استعراض أهم تقنيات اختيار السمات، بالإضافة إلى تقسيم البيانات إلى مجموعتي التدريب والاختبار وفق منهجية معتمدة تضمن التقييم الموضوعي للنماذج.

وفي مرحلة لاحقة، تم تقديم مجموعة من نماذج التعلم الآلي المعتمدة على الأشجار وتقنيات التعلم التجميعي، دون التطرق إلى النتائج التطبيقية، وذلك بهدف توضيح الأساس النظري والمنهجي الذي يقوم عليه العمل.

وبناءً على ما سبق، يُعد هذا الفصل قاعدة منهجية أساسية سيتم الاعتماد عليها في الفصل التطبيقي من أجل اختبار فعالية هذه المراحل وتقييم أداء النماذج المقترحة في سياق عملي.

الفصل الثالث: الإنجاز

1.1 المقدمة

الانجاز يعتبر من أهم الفصول في هذه الدراسة، حيث يم فيه عرض العمل التطبيقي وشرح المنهجية المعتمدة لبناء نموذج تنبؤي لمرض القصور الكلوي باستخدام تقنيات التعلم الالي.

في البداية يتم تقديم وصف شامل لمجموعة البيانات المستخدمة، ثم عرض بيئة العمل التي تم اعتمادها. كما يتم شرح مراحل التنفيذ الأساسية للدراسة بداية بمرحلة التحليل الاستكشافي التي يمكن من خلالها فهم بنية البيانات، ثم الانتقال الى مرحلة المعالجة المسبقة للبيانات التي تشمل معالجة القيم المفقودة والشاذة، ترميز المتغيرات، تقييس البيانات إضافة الى اختيار الميزات الأكثر تأثير على النموذج. وفي اخر مرحلة يتم تدريب النماذج المختلفة وتقييمها باستخدام منهجية دقيقة ومقاييس أداء مناسبة من اجل اجراء مقارنة موضوعية بين النماذج واختيار النموذج النهائي وتحسينه. وفي نهاية هذا العمل يُحفظ النموذج ويُحول الى منصة تفاعلية ويُشرح كيف يتم استخدامها للتنبؤ بالمرض.

1.2 وصف البيانات

✓ مصدر البيانات

في هذه الدراسة تم استخدام قاعدة بيانات مرض القصور الكلوي المزمن المتوفرة عبر منصة Kaggle [30]، تحت عنوان *CKD Disease Dataset*، والتي تُعد من قواعد البيانات الشائعة في أبحاث التعلم الآلي والتطبيقات الطبية. فهذه القاعدة ليست مستقلة تمامًا، بل تُعد نسخة مُعاد نشرها من مجموعة بيانات أصلية متوفرة في UCI Machine Learning Repository [31] والتي تم جمعها من سجلات سريرية حقيقية لمرضى في مستشفى بالهند. وقد تم اعتماد هذه البيانات في العديد من الدراسات العلمية المتعلقة بالتنبؤ ب CKD باستخدام تقنيات التعلم الآلي، حيث تم استخدامها لتقييم أداء نماذج التصنيف المختلفة، مما يعكس موثوقيتها واستخدامها كمرجع معياري (Benchmark Dataset)في هذا المجال[32] .

✓ **حجم البيانات**

تتكون قاعدة بيانات هذه الدراسة من **400عينة**(Samples) ، بحيث تمثل كل عينة مريضًا واحدًا، وتتضمن مجموعة من المؤشرات السريرية والمخبرية المرتبطة بوظائف الكلى. وتشتمل هذه القاعدة على **25متغيرًا**(Attributes) ، بالإضافة إلى متغير الهدف (Target Variable) الذي يتم استخدامه لتصنيف المرضى إلى مصابين وغير مصابين بمرض القصور الكلوي المزمن[30] .

1.3 بيئة وخطوات سير العمل

في هذه الدراسة تم استخدام بيئة تطوير متكاملة ترتكز على منصة **Anaconda** لإدارة المكتبات والبيئات الخاصة بلغة **Python**، كما تم الاعتماد على بيئة **Jupyter Notebook** لتنفيذ مراحل العمل التي يتم شرحها في جزء التنفيذ.

لتطوير المنصة وتنظيمها تم استخدام **Visual Studio Code** مع اعتماد إيطار عمل **Flask** لدمج النموذج النهائي داخل تطبيق ويب تفاعلي.

✓ **بايثون 3 (Python 3):** هي لغة برمجية حديثة من النسخة الشهيرة بايثون، عالية المستوى تتميز بسهولة التعلم وبساطة الإنشاء هذا ما يجعلها أكثر استخداما في مجالات علم البيانات والتعلم الآلي، تم إطلاقها لأول مرة في 3 ديسمبر 2008 (المعروفة بPy3K) لتصحيح أخطاء تصميمية أساسية في الإصدارات السابقة (Python 2). [41] حيث توفر مكتبات قوية تم استخدامها في هذه الدراسة :

Numpyو Pandas (لمعالجة البيانات وتحليلها) ، Matplotlibو Seaborn (لتمثيل البيانات بصريا) ، Scikt-Learn(لتطبيق خوارزميات التعلم الالي والتقييم) ، XGBoost (لبناء نموذج عالي الاداء) ، أما الأدوات الإضافية تتمثل في : Pickle (لحفظ النموذج النهائي) ،وJSON (لتبادل البيانات بين الواجهة والخادم).

✓ **أناكوندا (Anaconda):** هي منصة متكاملة وتوزيع مجاني ومفتوح المصدر للغات البرمجية python، مصممة لتبسيط ادارة الحزم والبيئات الخاصة بمشاريع علوم البيانات والتعلم الالي، حيث تتضمن أدوات مدمجة مثل Jupyter Notebook، كما تسهل تثبيت الحزم اللازمة اعتمادا على مدير الحزم Conda.[42]

- ✓ **Jupyter Notebook** : هو موقع ويب تفاعلي يستخدم لكتابة دفاتر تحتوي على أكواد برمجية ونصوص توضيحية لها، معادلات رياضية، تمثيلات بيانية. يسمح الموقع بتنفيذ خلايا الكود(Cells) بشكل منفصل وعرض النتائج تحت كل خلية مما يجعله أداة مثالية للتعلم والتجربة. كما يستخدم لتحليل البيانات ومعالجتها. [43]

- ✓ **Flask** : هو إطار عمل ويب (Web Framework) مصغر و مرن مفتوح المصدر، مكتوب بلغة Python يسمح بإنشاء تطبيقات ويب مع التحكم في مكوناتها حيث يُصنف كـ "Micro-framework" لأنه خفيف الوزن[44] . في هذه الدراسة تم استخدام Flask لتحويل النموذج النهائي المدرب الى تطبيق ويب تفاعلي.

- ✓ **Visual Studio Code** : هو محرر أكواد تم تطويره من طرف Microsoft، يدعم العديد من اللغات من بينها Pythonو يعتبر بيئة تطوير قوية وشائعة بين المبرمجين لأنه يحتوي أدوات مدمجة كما يتميز بواجهة سهلة الاستخدام وإمكانية إضافة امتدادات تدعم تحليل البيانات وتطوير تطبيقات ويب. تم استخدامه في هذه الدراسة لإدارة الكود وتنظيم الملفات إضافة الى تطوير تطبيق Flask.[45]

1.4 التنفيذ(Implémentation)

تم تنفيذ أربعة مراحل أساسية في هذه الدراسة وهي:

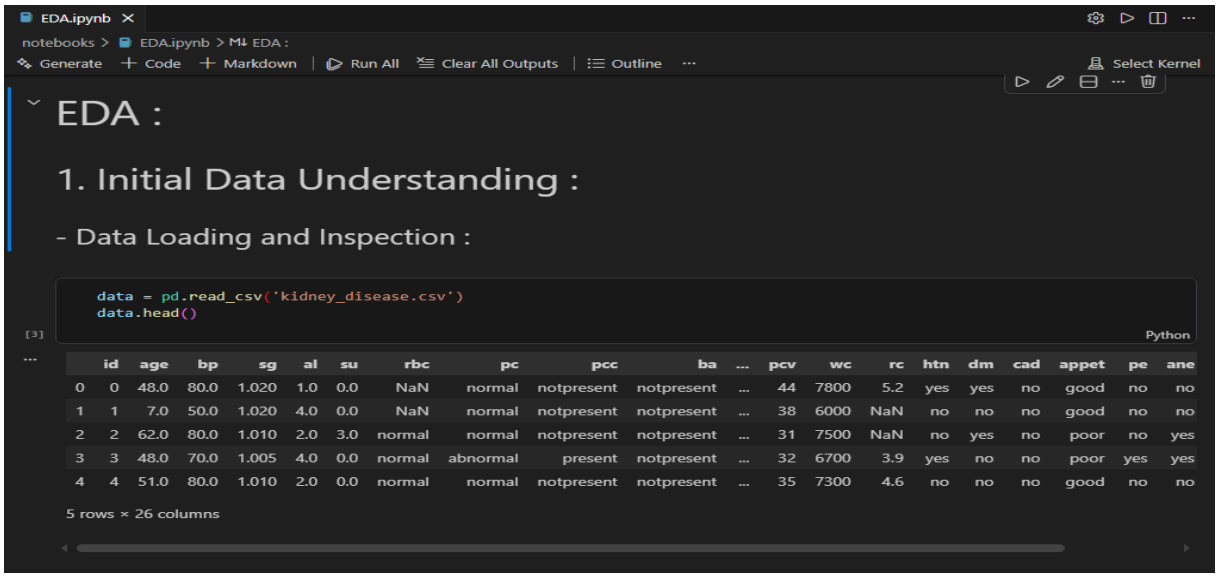
1.1.4 مرحلة التحليل الاستكشافي(EDA)

تم في هذه المرحلة استكشاف البيانات من خلال استخدام التعليمات البرمجية الأساسية لتحميل البيانات ودراسة نوع المتغيرات والعلاقات فيما بينها والتحليل الكامل لها.

✓ تحميل البيانات

تم تحميل البيانات باستخدام مكتبة **Pandas** بصيغة **CSV**، حيث تم تحويلها إلى جدول منظم من البيانات وهوما يُسمى **DataFrame** لتسهيل معالجتها. الفحص الاولي للبيانات يتم باستخدام دوال مثل**()read_csv** و**()head** لقراءتها وعرض

أول 5صفوف، إضافة الى الدالة **info ()** لفهم البيانات من خلال إعطاء نظرة شاملة لها (اكتشاف القيم الناقصة ونوع كل عمود). في الصورة المقابلة يتم عرض 5 صفوف الأولى للبيانات المستخدمة:



	id	age	bp	sg	al	su	rbc	pc	pcc	ba	...	pcv	wc	rc	htn	dm	cad	appet	pe	ane
0	0	48.0	80.0	1.020	1.0	0.0	NaN	normal	notpresent	notpresent	...	44	7800	5.2	yes	yes	no	good	no	no
1	1	7.0	50.0	1.020	4.0	0.0	NaN	normal	notpresent	notpresent	...	38	6000	NaN	no	no	no	good	no	no
2	2	62.0	80.0	1.010	2.0	3.0	normal	normal	notpresent	notpresent	...	31	7500	NaN	no	yes	no	poor	no	yes
3	3	48.0	70.0	1.005	4.0	0.0	normal	abnormal	present	notpresent	...	32	6700	3.9	yes	no	no	poor	yes	yes
4	4	51.0	80.0	1.010	2.0	0.0	normal	normal	notpresent	notpresent	...	35	7300	4.6	no	no	no	good	no	no

الصورة 5: تحميل البيانات المستخدمة وعرض أول الصفوف.

✓ المتغيرات(Features/Targets)

تتكون قاعدة البيانات من مجموعة المتغيرات الطبية التي تمثل الخصائص السريرية والمخبرية للمرضى (Features) بالإضافة إلى المتغير الهدف (Target) الذي يحدد حالة المريض من حيث الإصابة بالمرض.

1. **المتغيرات المستقلة(Features)** تتضمن:

✓ **<u>المتغيرات الرقمية (Numerical Features)</u>** :تشمل الخصائص التي تم قياسها بوحدات رقمية، مثل:

- **ID(المعرف)** : وهو رقم فريد خاص بكل مريض ;
- **Age (العمر)** : يمثل عمر المريض بالسنوات ;
- **BP(ضغط الدم)** : ضغط الدم الانقباضي والانبساطي بوحدة mmHg ;
- **BGR (مستوى السكر في الدم)** : تركيز الغلوكوز في الدمmg/Dl ;
- **BU (اليوريا في الدم)**: تركيز اليوريا في الدم mg/dL ;
- **SC (الكرياتينين في الدم)** : تركيز الكرياتينين في الدمmg/Dl ;

- **SOD (الصوديوم)و POT (البوتاسيوم) :** تركيز المعادن ;
- **HEMO (الهيموغلوبين) :** تركيز الهيموغلوبينg/dL ;
- **PCV (حجم كريات الدم الحمراء) :** نسبة الهيماتوكريت %;
- **WC (عدد كريات الدم البيضاء) :** عدد خلايا الدم البيضاء لكل وحدة حجم الدم ;
- **RC (عدد كريات الدم الحمراء) :** عدد خلايا الدم الحمراء لكل وحدة حجم الدم .

✓ **المتغيرات الوصفية(Categorical Features) :** تمثل الخصائص التي تُصنف في فئات غير رقمية، مثل:

- **SG (الكثافة النوعية للبول) :** قيم مرتبة بين 1.005 و1.025 ;
- **AL(الألبومين في البول) :** قيم مرتبة 0 – 5 ;
- **SU (السكر في البول) :** قيم مرتبة 0 – 5 ;
- **RBC (حالة كريات الدم الحمراء) :** طبيعي/ غير طبيعي ;
- **PC (وجود خلايا القيح في البول) :** طبيعي/ غير طبيعي ;
- **PCC (تجمع خلايا القيح في البول) :**موجود / غير موجود ;
- **BA (وجود البكتيريا في البول):** موجود / غير موجود ;
- **HTN (ارتفاع ضغط الدم) :** نعم / لا ;
- **DM (مرض السكري):**نعم / لا ;
- **CAD (أمراض القلب) :** نعم / لا ;
- **PE (وذمة) :** نعم / لا ;
- **ANE (فقر الدم) :** نعم / لا ;
- **APPET (الشهية) :** جيدة / ضعيفة.

2. **المتغير الهدف(Target Variable)**

- **للتصنيف(Classification) :**يحدد هل المريض مصاب بمرض القصور الكلوي المزمن أو غير مصاب، ويعتبر متغير ثنائي(Binary) فهو مناسب لتطبيق خوارزميات التصنيف في التعلم الآلي.

✓ تحليل توزيع المتغيرات

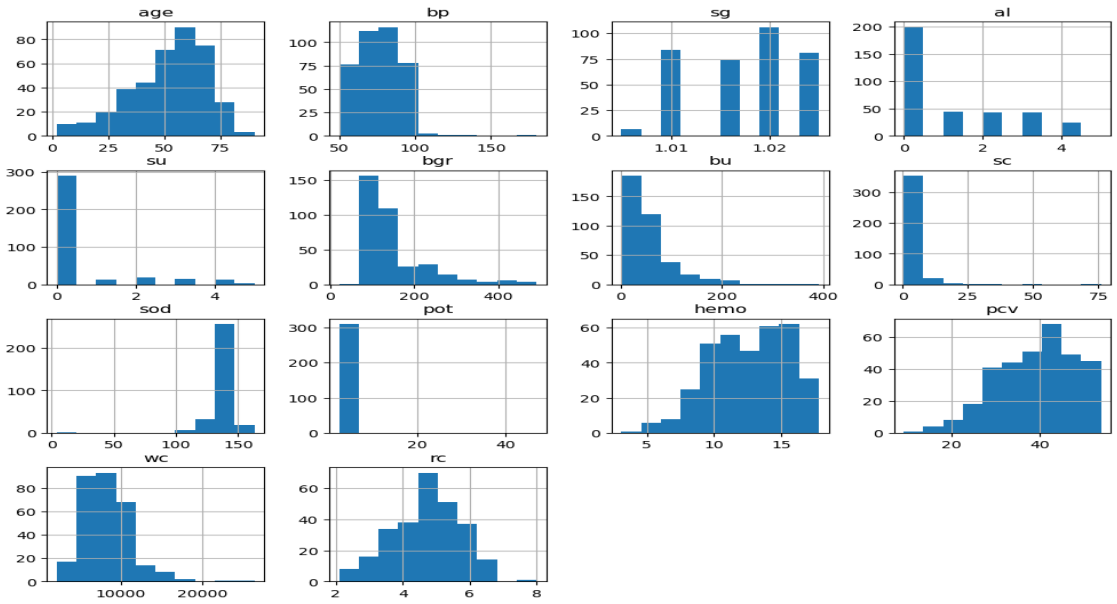
تم تحليل توزيع المتغيرات من خلال دراسة انتشار القيم داخل كل متغير، وذلك باستخدام أدوات برمجية تحليلية مثل:

- الخاصية **shape ()** :لفحص أبعاد البيانات.
- الدالة**() value_counts :** لدراسة توزيع القيم داخل المتغيرات (تحسب تكرار كل قيمة في عمود معين).

كما تم استخدام العرض البصري لتوزيع المتغيرات الرقمية من خلال التمثيل البياني المتمثل في **المدرج التكراري**، حيث يساعد على:

- **تحديد شكل التوزيع**: طبيعي: القيم مركزة في الوسط(جرس) / منحرف: القيم تميل لجهة معينة ;
- **الكشف عن القيم الشاذة**: تظهر كأعمدة صغيرة بعيدة عن البقية ;
- **فهم مدى تركز القيم و انتشارها** :إذا كانت الأعمدة قريبة فالقيم متقاربة وإذا الأعمدة منتشرة فالقيم متباعدة.

توضح الصورة المقابلة مدرج تكراري لتوزيع المتغيرات العددية بحيث من خلاله يمكن تحديد شكل توزيع القيم والكشف عن الشاذة منها وفهم تركزها وإنتشارها:



الصورة 6:المدرج التكراري لتوزيع المتغيرات العددية.

يوضح مخطط المدرج التكراري توزيع قيم السمات في مجموعة البيانات. على سبيل المثال، **سمة العمر (Age)** تُظهر أن أغلب المرضى تتراوح أعمارهم بين **40و70 سنة**، مع انخفاض عدد الحالات في الأعمار الصغيرة، مما يشير إلى أن مرض الفشل الكلوي المزمن أكثر انتشارًا لدى الفئات الأكبر سنًا. كما تظهر بعض السمات مثل **الكرياتينين (SC)** توزيعات منحرفة بسبب وجود قيم مرتفعة وشاذة، مما يستدعي معالجتها قبل التدريب.

✓ تحليل احصائي وصفي

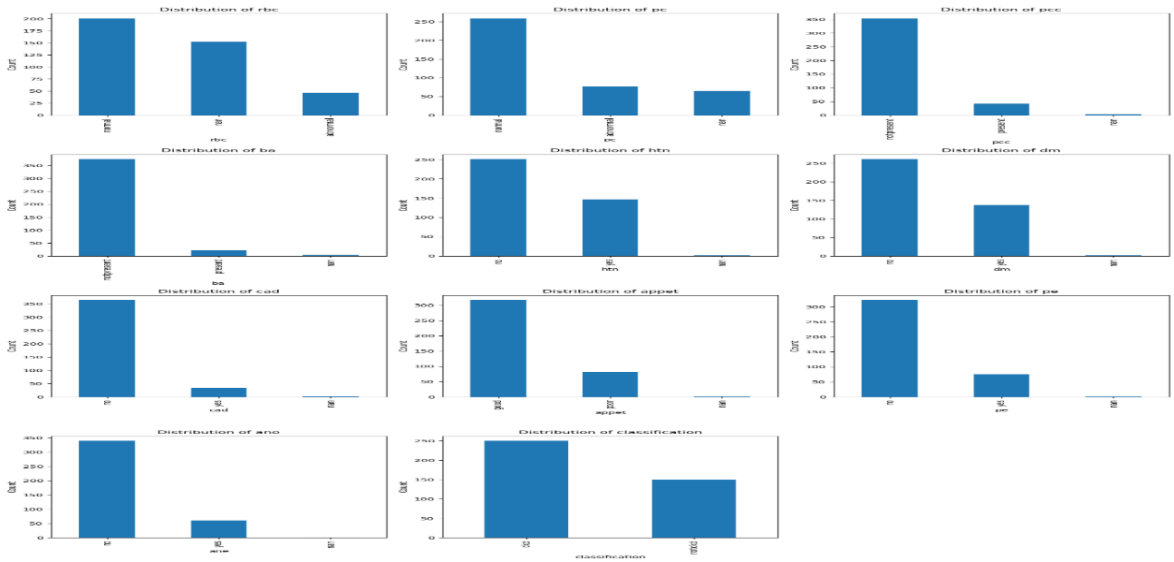
تم التحليل الوصفي للمتغيرات الرقمية باستخدام الدالة **() describe**، حيث استُخرجت مجموعة المؤشرات الإحصائية المهمة مثل:

- **المتوسط الحسابي(Mean)**: يحدد القيمة المركزية ;
- **الانحراف المعياري (Standard Deviation)**: يقيس درجة التشتت ;
- **القيم الدنيا والعليا(Min/Max)**: لتحديد حدود البيانات ;
- **الربعيات(Quartiles)**: تستعمل لفهم توزيع القيم وتقسيمها لأجزاء متساوية.

كما تم تحليل المتغيرات الوصفية باستخدام التعليمة **describe(include='object')** حيث سمح ب:

- معرفة القيم المتكررة التي تظهر أكثر في المتغير ;
- تحديد الفئات المختلفة من خلال عدد القيم المختلفة في المتغير والقيم الفارغة ;
- فهم توزيع الفئات داخل المتغيرات من خلال توزيع القيم (البيانات متوازنة ام لا).

كما تم استخدام المدرج التكراري لعرض تكرار القيم داخل كل متغير وصفي وهو ما توضحه الصورة التالية حيث لوحظ وجود عدم توازن في توزيع الفئات داخل المتغيرات.



الصورة 7: توزيع المتغيرات الوصفية باستخدام المدرج التكراري.

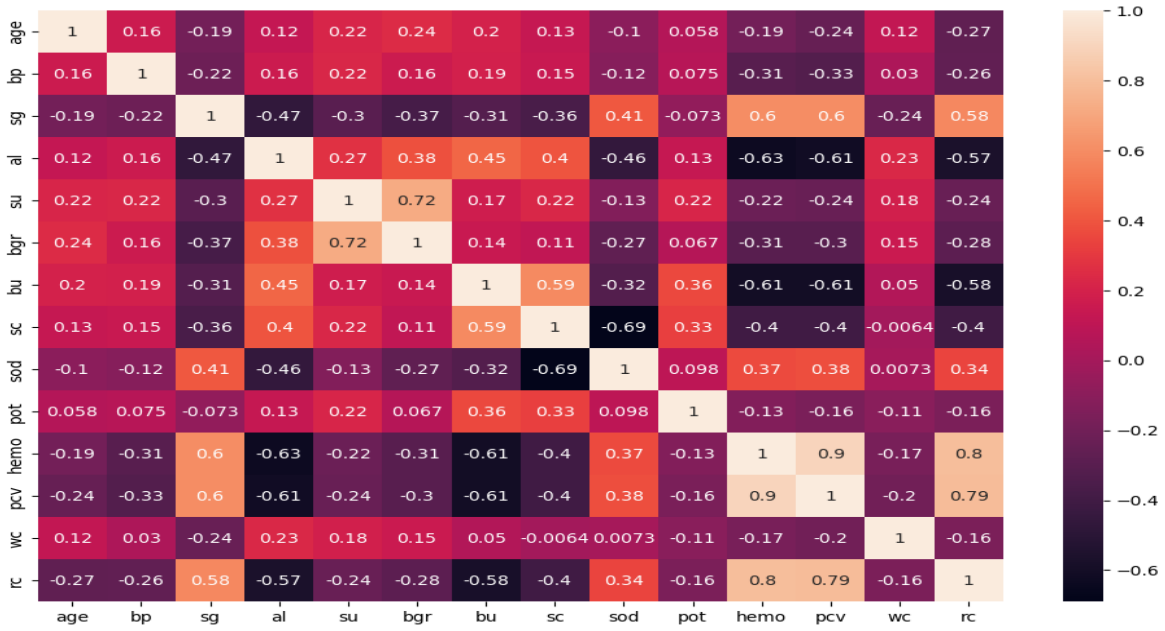
✓ العلاقات بين المتغيرات

تمت دراسة العلاقات بين المتغيرات بشكل أولي من خلال تحليل خصائص البيانات وفهم طبيعتها لتحديد قوة الارتباط بينها، مما يساعد في اختيار المتغيرات الأكثر تأثيرا في التنبؤ بالمرض. كما تم فحص جودة البيانات من خلال التعليمتين:

insa().sum() {الكشف عن القيم المفقودة}، **duplicated().sum()** {التحقق من وجود بيانات مكررة}.

كانت النتيجة وجود قيم مفقودة في بعض الأعمدة وعدم وجود تكرار في البيانات.

إضافة إلى ذلك تم استخدام **مصفوفة الارتباط (Correlation Heatmap)** لتمثيل العلاقات بين المتغيرات بشكل بصري بحيث كل ما كان اللون افتح كانت العلاقة اقوى هذا ما توضحه الصورة المقابلة:



الصورة 8: مصفوفة الارتباط بين المتغيرات(Correlation Heatmap) .

يبين مخطط مصفوفة الارتباط قوة العلاقة بين السمات. أقوى الارتباطات الموجبة كانت بين **الهيموغلوبين (Hemo) وحجم خلايا الدم المضغوطة (PCV)** بقيمة تقارب **0.90**، وبين **سكر الدم (BGR) وسكر البول (SU)** بقيمة **0.72**. أما أقوى الارتباطات السالبة فكانت بين **الصوديوم (SOD) والكرياتينين (SC)** بقيمة **0.69-**، وبين **الألبومين (AL) والهيموغلوبين (Hemo)** بقيمة **0.63-**، مما يدل على وجود علاقات قوية بين المؤشرات المرتبطة بوظائف الكلى والدم.

✓ تعديل البيانات والحفظ

تم إجراء بعض التعديلات على البيانات المستخدمة لجعل النماذج تتعلم بسرعة وتعطي نتائج أدق. حيث تم تنظيف البيانات وذلك بإزالة المسافات وعلامات الجدولة من النصوص الفئوية وتوحيد قيمها، كما تم تحويل بعض الميزات الفئوية إلى ميزات رقمية مثل (pcv، wc، rc)لأنها ذات قيم عددية.

إضافة إلى حذف العمود id نظرا لعدم تأثيره على عملية التنبؤ، وحفظ التغييرات على البيانات واعتمادها في باقي المراحل والتدريب. الصور التالية توضح التعليمات البرمجية لتنظيف البيانات وحفظ التغييرات:

EDA.ipynb

notebooks > EDA.ipynb > ...

Generate | Code | Markdown | Run All | Clear All Outputs | Outline

- Data Cleaning:

- Remove spaces and tabs from category texts :

Generate | Code | Markdown

Start Chat to Generate Code (Ctrl+I)

```
for col in data.select_dtypes(include='object'):
    data[col] = data[col].str.strip()
```

[11]

- Converting some categorical features into numerical categories :

```
for col in ['pcv','wc','rc']:
    data[col] = pd.to_numeric(data[col], errors='coerce')
```

[12]

- Conversion verification :

```
data[['pcv','wc','rc']].dtypes
```

[13]

[13]

```
pcv      float64
wc       float64
rc       float64
dtype: object
```

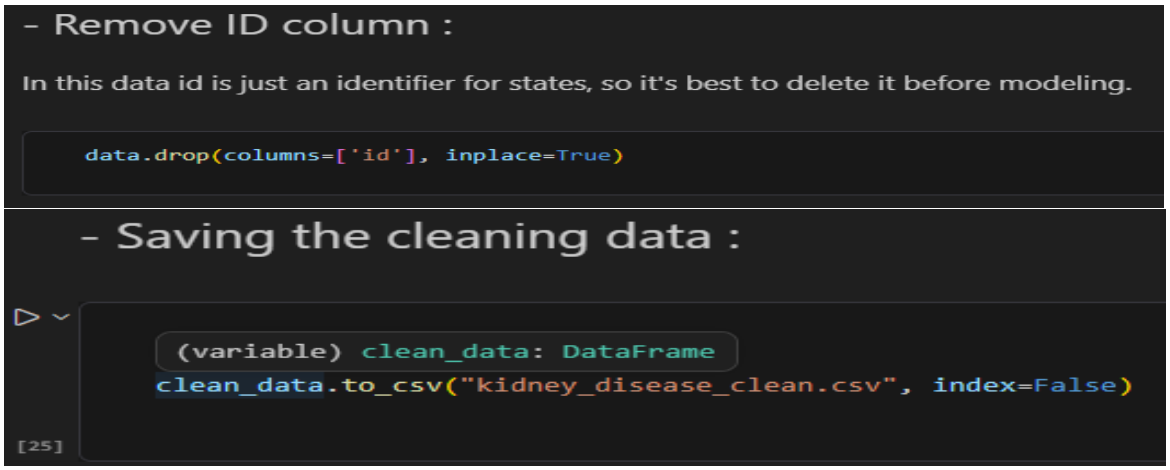
الصورة 9:الكود البرمجي لخطوات تنظيف البيانات.

```
# 1. Clean up text: Remove spaces and tabs + convert to lowercase
for col in data.select_dtypes(include='object'):
    data[col] = data[col].str.strip().str.lower()
# 2. Dictionary for standardizing category values
replace_map = {
    'yes':'yes', 'no':'no',
    'present':'present', 'notpresent':'notpresent',
    'normal':'normal', 'abnormal':'abnormal',
    'good':'good', 'poor':'poor',
    'ckd':'ckd', 'notckd':'notckd'}
# 3. Apply standardization to all text columns
for col in data.select_dtypes(include='object'):
    data[col] = data[col].replace(replace_map)
# 5. Verify unique values after cleaning
for col in data.select_dtypes(include='object'):
    print(col, data[col].unique())
```

[14]

```
rbc ['nan' 'normal' 'abnormal']
pc ['normal' 'abnormal' nan]
pcc ['notpresent' 'present' nan]
ba ['notpresent' 'present' nan]
htn ['yes' 'no' nan]
dm ['yes' 'no' nan]
cad ['no' 'yes' nan]
appet ['good' 'poor' nan]
pe ['no' 'yes' nan]
ane ['no' 'yes' nan]
classification ['ckd' 'notckd']
```

الصورة 10 : إزالة المسافات وعلامات الجدولة من النصوص الفئوية وتوحيد قيمها.



الصورة 11: إزالة العمود id وحفظ البيانات النظيفة.

2.1.4. المعالجة المسبقة(Preprocessing)

في هذه المرحلة تمت معالجة البيانات تحضيرًا للتعلم الآلي من خلال:

1. تقسيم البيانات إلى **80%**مجموعة تدريب و**20%** مجموعة اختبار ;
2. ملأ القيم المفقودة بثلاث طرق« **Knn**، **Median**، **Regression**» وإنشاء ملفات **Imputation** الخاصة بكل طريقة ;
3. الكشف عن القيم الشاذة بواسطة« **IQR**»ومعالجتها بطريقتين **Winsorization**و**QuantileTransformer** وتطبيقهما على الملفات الثلاثة وإنشاء الملفات الخاصة بها (طريقة للملأ_طريقة للمعالجة لكل حالة) ;
4. وبما أن نماذج التعلم الآلي لا تفهم إلا الأرقام تم ترميز (**Encoding**) المتغيرات النصية والمتغير الهدف باستخدام **One-Hot Encoding** (التعليمة: **drop_first=True**)لتفادي مشكلة التعدد الخطي(**Multicollinearity**) وحفظ الملفات الخاصة بالترميز ;
5. كما تم تقييس (**Scaling**)المتغيرات العددية من خلال تحويل قيمها إلى نطاق موحد باستخدام تقنية **"MinMaxScaling** " لتحسين أداء النماذج خاصة الحساسة لمقياس البيانات ;
6. تتميز هذه الدراسة بإضافة خطوة اختيار السمات (**Feature Selection**) لتعيين أفضل الميزات المؤثرة في عملية التنبؤ حيث استُخدمت ثلاث تقنيات **"rfe**، **mi**، **chi2**" وحُدد عددها المتمثل في **5ميزات**.

3.1.4 التدريب ومقارنة النتائج

✓ منهجية التقييم

لتقييم أداء النماذج بطريقة علمية ومنهجية، تم اعتماد تقييم يرتكز على اختبار النماذج باستخدام بيانات لم تُستخدم أثناء مرحلة التدريب، بهدف قياس قدرتها على التعميم وتجنب مشكلة فرط التعلم حيث:

- تم تقسيم البيانات الى مجموعتين رئيسيتين: مجموعة تدريب (Training Set) لبناء النموذج واستخلاص الأنماط، مجموعة اختبار (Test Set) لتقييم أدائه على بيانات لم يتم رؤيتها. كما تم تطبيق أسلوب التقسيم الطبقي (Stratified Sampling) للحفاظ على التوزيع النسبي للفئات داخل المجموعتين، مما يعطي تقييم أكثر دقة ويمنع من تحيز البيانات.
- تم إجراء مقارنة منهجية بين عدة نماذج من خلال تدريبها باستخدام إعدادات مختلفة، شملت تقنيات متنوعة في المعالجة المسبقة وعدد الميزات المختارة، حيث تهدف المقارنة إلى دراسة تأثير هذه العوامل على أداء النماذج. كما تم استخدام عدة مقاييس إحصائية دون الاعتماد فقط على مقياس الدقة لتقييم كل نموذج.
- بالإضافة إلى ذلك تم اعتماد تقنية **Fold Cross Validation-10** على مجموعة التدريب حيث يتم فيها تقسيم البيانات إلى **10**أجزاء متساوية، ففي كل تكرار يتم استخدام **9** أجزاء للتدريب وجزء للتحقق (Validation) مع تكرار العملية عشر مرات.

✓ المقاييس المعتمدة

لتقييم أداء النماذج بشكل دقيق تم اعتماد المقاييس الإحصائية الشائعة في مجال التصنيف، والتي تسمح بتحليل الأداء من زوايا متعددة، والمتمثلة في[47] :

1. **الدقة (Accuracy):** هي نسبة الحلات التي تم تصنيفها بشكل صحيح من إجمالي الحالات وتُستخدم كمؤشر عام على أداء النموذج، إلا أنه قد يكون غير كافي في حالة عدم توازن البيانات ;
2. **الضبط (Precision):** قدرة النموذج على عدم تصنيف عينة سلبية على أنها إيجابية ;
3. **الاستدعاء (Recall): يعكس** قدرة النموذج على اكتشاف جميع الحالات الموجبة، وهو من المقاييس الأساسية في التطبيقات الطبية لان تجاهل الحالات المصابة يؤدي لنتائج خطيرة ;

4. **مقياسF1 (F1-Score):** المتوسط الهندسي بين الدقة والاستدعاء، وهو مفيد عند وجود توازن بين الفئات ;
5. **مصفوفة الارتباك (Confusion Matrix):** تُستخدم لتحليل نتائج التصنيف من خلال عرض جدول يوضح تفاصيل التصنيفات الصحيحة والخاطئة وهذا يساعد في فهم الأخطاء التي يرتكبها النموذج ;
6. **مساحة تحت منحىROC (ROC-AUC):** يُستخدم لتقييم قدرة النموذج على التمييز بين الفئات المختلفة، حيث تعكس القيم المرتفعة أداء أفضل للنموذج في الفصل بين الحالات.

✓ مقارنة أداء النماذج

بعد مرحلة المعالجة المسبقة للبيانات، تأتي المرحلة الأساسية وهي تدريب عدة نماذج تعلم آلي شملت:Decision Tree ، Random Forest، Extra Trees، XGBoost وStacking Ensemble ومقارنتها بالاعتماد على المقاييس F1،Accuracy ، Recall وزمن التنفيذ لتحديد النموذج الأفضل في الأداء للتنبؤ بالمرض.

كما تم استخدام تقنيات مختلفة لتحديد السمات المهمة للحفاظ على الدقة وتقليل التعقيد، وتم اختيار ثلاث قيم للسمات وهي 10،5،3 والمقارنة فيما بينها لهدف تأثير عدد السمات على أداء النماذج.

أوضحت النتائج أن المعالجة المسبقة المختلفة للبيانات المتمثلة في طرق تعويض القيم الناقصة ومعالجة القيم الشاذة تؤثر على أداء النماذج، وقد كانت النتائج متقاربة نوعا ما كما هو مبين في الجداول التالية التي تعرض 5أفضل نتائج لكل نموذج مع قيم مختلفة للسمات (10،5،3) .

1. **<u>نتائج النماذج باستخدام 3سمات</u>**

- **شجرة القرار(Decision Tree)(وقت التنفيذ : 15.81 s)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
Regression	quantile	rfe	0.9531	0.9429	0.9833
Regression	winsor	rfe	0.9406	0.9263	0.9596
Regression	winsor	ch2	0.9219	0.9113	0.9923
Regression	quantile	ch2	0.9219	0.9113	0.9923
knn	quantile	mi	0.9187	0.9017	0.9673

- **الغابة العشوائية (Random Forest)(وقت التنفيذ :s 229.87)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
Regression	quantile	rfe	0.9594	0.9491	0.975
Regression	winsor	rfe	0.9563	0.9454	0.975
median	winsor	rfe	0.9406	0.9252	0.9429
knn	quantile	rfe	0.9281	0.9136	0.975
Regression	quantile	ch2	0.9219	0.9113	0.9923

- **الأشجار المتطرفة العشوائية(Extra trees) (وقت التنفيذ :s 172.81)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
Regression	winsor	rfe	0.9406	0.9235	0.934
Regression	winsor	ch2	0.9219	0.9113	0.9923
knn	winsor	ch2	0.9187	0.908	0.917
knn	quantile	ch2	0.9156	0.9026	0.9756
Regression	quantile	ch2	0.9156	0.9026	0.9756

- **تعزيز التدرج المحسّن (XGBoost) (وقت التنفيذ :s 262.73)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
Regression	winsor	rfe	0.9625	0.9522	0.9667
Regression	quantile	rfe	0.9563	0.9439	0.959
median	winsor	rfe	0.925	0.9075	0.9429
Regression	winsor	ch2	0.9187	0.9073	0.9923
Regression	quantile	ch2	0.9127	0.9073	0.9923

- **التكديس (Staking) (وقت التنفيذ :s 3564)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
Regression	winsor	rfe	0.9625	0.9522	0.9661
Regression	quantile	rfe	0.9531	0.9389	0.9506
median	winsor	rfe	0.9375	0.9226	0.9513
Regression	winsor	ch2	0.9219	0.9113	0.9923
Regression	quantile	ch2	0.9219	0.9113	0.9923

1.1 نتائج النماذج باستخدام 3 سمات

أظهرت النتائج أنه رغم الاعتماد على عدد محدود من السمات المتمثل في 3 من أصل 24، إلا أن النماذج المدروسة حققت أداء جيد مما يدل على فعالية اختيار السمات، خاصة تقنية RFE و CH2التي ظهرتا ضمن أفضل النتائج لأداء النماذج وقدرتها على اختيار السمات الأكثر تأثير.

حقق النموذج XGBoost أداء أفضل، بلغ أعلى دقة تمثلت في 96.25%وقيم مرتفعة للمقياسين F1 و Recallوهذا ما يدل على كفاءته مع عدد سمات أقل. كما ان النموذج Random Forest حقق نتائج جيدة أما النموذج Extra Trees حقق أداء أقل نسبيا من النماذج. في حين تميز النموذج Decision Tree عن البقية بسرعة التنفيذ وسهولة التدريب لكن نتائجه أقل استقرار. من جهة أخرى حقق النموذجStacking Ensemble أيضا أفضل أداء لكن زمن التنفيذ مرتفع نتيجة اعتماده على دمج عدة نماذج تعلم آلي.

كما بينت نتائج3سمات أن استخدام Winsorizationوquantile لمعالجة القيم الشاذة، وطرق التعويض Regression، كانت لها أكبر تأثير في تحسين أداء النماذج واستقرارها.

تشير النتائج الى أنه يمكن تحقيق أداء مرتفع نسبيا باستخدام 3 سمات فقط، وخاصة مع نماذج Stackingو XGBoost.

2. **نتائج أداء النماذج باستخدام 5سمات**

- **شجرة القرار(Decision Tree) (وقت التنفيذ :14.59s)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
Regression	winsor	rfe	0.9875	0.984	0.9917
Regression	quantile	rfe	0.9875	0.984	0.9917
Regression	quantile	mi	0.9844	0.98	0.984
median	winsor	mi	0.9812	0.9756	0.9833
median	winsor	rfe	0.9812	0.9756	0.9833

- **الغابة العشوائية (Random Forest) (وقت التنفيذ :252.87 s)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
Regression	winsor	rfe	0.9938	0.992	1.00
Regression	quantile	rfe	0.9938	0.992	1.00
Regression	quantile	mi	0.9938	0.9923	1.00
Regression	winsor	mi	0.9844	0.9793	0.9833
median	winsor	mi	0.9844	0.9793	0.975

- **الأشجار المتطرفة العشوائية (Extra trees) (وقت التنفيذ :224.99 s)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
Regression	winsor	rfe	0.9812	0.9766	1.00
Regression	quantile	mi	0.9781	0.9737	1.00
Regression	winsor	mi	0.9781	0.9729	1.00
median	winsor	mi	0.9781	0.9713	0.9673
median	winsor	rfe	0.9781	0.9713	0.9673

- **تعزيز التدرج المحسّن (XGBoost) (وقت التنفيذ :s 345.65)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
Regression	winsor	rfe	0.9906	0.9879	0.9917
Regression	quantile	rfe	0.9906	0.9879	0.9917
Regression	quantile	mi	0.9875	0.984	0.984
median	winsor	mi	0.9812	0.9756	0.9673
median	winsor	rfe	0.9812	0.9756	0.9673

- **التكديس (Staking) (وقت التنفيذ :s 3866.19)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
Regression	winsor	rfe	0.9938	0.9923	1.00
Regression	quantile	rfe	0.9938	0.9923	1.00
Regression	quantile	mi	0.9906	0.9883	0.9923
Regression	winsor	mi	0.9875	0.9833	0.9833
median	quantile	mi	0.9844	0.9793	0.9673

1.2 نتائج أداء النماذج باستخدام 5 سمات

أظهرت نتائج أداء النماذج باستخدام 5سمات تحسن ملحوظ مقارنة بعدد سمات أقل (3)، وهذا يدل على أن زيادة عدد السمات يساهم في توفير معلومات إضافية تحسن من قدرة النماذج على التمييز بين الحالات المصابة وغير المصابة. في حين حقق النموذج Random Forest دقة بلغت 99.38%مع Recall مثالي، كما حقق Stacking Ensemble نتائج قوية من حيث الدقة وباقي المقاييس لكن زمن تنفيذ أكبر، أما النموذجXGBoost حافظ على أداء قوي ومستقر.

من جهة استمر النموذج Decision Tree في تقديم نتائج جيدة مع زمن تنفيذ منخفض لكن أقل كفاءة من النماذج الجماعية، ومن جهة أخرى حقق النموذج Extra Trees أداء جيد لكنه أقل استقرار مقارنة بـRandom Forest وXGBoost.

بينت النتائج في هذا الإعداد أن أفضلها تحققت باستخدام Regressionلتعويض القيم المفقودة، Winsorization لمعالجة القيم الشاذة، وتقنية RFE لاختيار أهم السمات .كما أشارت النتائج إلى أن استخدام 5سمات فقط وفر توازن بين الأداء وتقليل تعقيد النموذج.

3. نتائج أداء النماذج باستخدام 10سمات

- **شجرة القرار(Decision Tree) (وقت التنفيذ :s 9.57)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
median	winsor	mi	0.9875	0.9826	0.975
median	quantile	rfe	0.9875	0.983	0.9833
median	winsor	rfe	0.9875	0.983	0.9833
median	quantile	mi	0.9844	0.9786	0.975
Regression	winsor	mi	0.9844	0.9796	0.9833

- **الغابة العشوائية (Random Forest) (وقت التنفيذ :s 253.5)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
Regression	winsor	mi	0.9938	0.9917	0.9917
median	quantile	mi	0.9938	0.9909	0.9833
median	quantile	rfe	0.9938	0.9909	0.9833
median	winsor	ch2	0.9906	0.988	1.0
Regression	winsor	rfe	0.9906	0.9877	0.9917

- **الأشجار المتطرفة العشوائية (Extra trees) (وقت التنفيذ :s 173.48)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
median	winsor	mi	0.9969	0.996	1.00
median	winsor	ch2	0.9906	0.988	1.00
median	winsor	rfe	0.9906	0.988	1.00
median	quantile	mi	0.9906	0.9869	0.9833
Regression	winsor	mi	0.9844	0.9803	1.00

- **تعزيز التدرج المحسّن(XGBoost) (وقت التنفيذ :s 253.1)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
median	quantile	mi	0.9906	0.987	0.9833
Regression	winsor	mi	0.9875	0.9833	0.9833
Regression	winsor	rfe	0.9875	0.9833	0.9833
median	winsor	mi	0.9875	0.983	0.9833
median	winsor	rfe	0.9844	0.9786	0.975

- **التكديس(Staking) (وقت التنفيذ :s 3917.6)**

تعويض القيم المفقودة	القيم الشاذة	اختيار السمات	Accuracy	F1	Recall
Regression	winsor	mi	0.9969	0.9957	0.9917
Regression	winsor	rfe	0.9938	0.9913	0.9833
median	winsor	mi	0.9938	0.9909	0.9833
median	quantile	mi	0.9938	0.9909	0.9933
median	quantile	rfe	0.9938	0.9909	0.9933

1.3 نتائج أداء النماذج باستخدام 10 سمات

أظهرت النتائج باستخدام 10سمات تسجيل أعلى قيم أداء بين التجربتين السابقتين، حيث حقق النموذج Stacking Ensemble أعلى دقة بلغت 99.69% مع مقياسين F1 وRecall مثاليين نسبيًا.

في حين حققا النموذجين Random ForestوXGBoost دقة متقاربة للأداء بلغت 99 %، أما Decision Tree بقي أقل كفاءة ولكن حافظ على سرعة التنفيذ. في المقابل أبدى النموذج Extra Trees تحسنا واضحا مقارنة بعدد اقل من السمات كما حقق دقة عالية للأداء بلغت 99.69% وRecallمثالي .إضافة الى ذلك أوضحت النتائج ان استخدام Winsorization لمعالجة القيم الشاذة وmedian لتعويض القيم الناقصة، وتقنيات اختيار السمات خاصة RFE وMI ساعد في تحسين أداء واستقرار النماذج.

وبالرغم من تحقيق أفضل النتائج باستخدام 10سمات، إلا أن زيادة عددها يؤدي إلى ارتفاع تعقيد النماذج وزمن التنفيذ خاصة في النماذج المركبة (Stacking Ensemble).

4. مقارنة عامة بين النماذج

هذه الدراسة تمثل مقارنة أداء النماذج المختارة (Staking/ XGboost/ Extra / Random Forest/Decision Tree) باستخدام تقنيات مختلفة للمعالجة المسبقة تشمل ثلاث طرق لملأ الفراغ (Regression/ knn /median) ،تقنيتين لمعالجة القيم الشاذة (quantile/ winsor) ،ثلاث طرق لاختيار أهم السمات (mi/ref /chi2) واختيار ثلاث قيم لأهم السمات (3/5/10)نتج عنها مقارنات مختلفة تتكون من 270 (3×3×2×3×5)حالة ،تم أخذ 75 حالة تتمثل في 5أفضل نتائج لكل نموذج بعدد سمات مختلف، حيث أنه من خلالها تبين أن نماذج التعلم الجماعي حققت أداء أفضل مقارنة بالنماذج التقليدية من حيث الدقة و الاستقرار.

- سجلت كل من Random ForestوXGBoostو Stacking Ensembleأعلى قيم RecallوF1 وAccuracy في مختلف التجارب وهذا يدل على كفاءتها العالية في التنبؤ بالمرض.
- لوحظ أنه عند زيادة عدد السمات يتحسن أداء النماذج، إلا أن التحسن باستخدام 5سمات كان محدودا مقارنة بزيادة التعقيد وزمن التنفيذ. لذلك يُعد اختيار 5سمات قرار متوازن بين الكفاءة والدقة.

- كما أن تقنيات المعالجة المسبقة في هذه الدراسة أظهرت تأثير واضح على تحسين النتائج، حيث حققت طرق التعويض Regression وMedian أفضل أداء مقارنة ب Knn، كما ان تقنية Winsoration عالجت تأثير القيم الشاذة وحسنت استقرار النماذج بطريقة أفضل منQuantile . أيضا أثبتت الطريقة Min-Max المختارة للتقييس فعاليتها في تحسين أداء خوارزميات التصنيف. أما تقنيات اختيار السمات أظهرت النتائج تفوق تقنية RFE ثم MIعلى التقنية CH2 حيث ساهمتا في تحسين جودة التنبؤ واختيار السمات الأكثر تأثيرا.
- إضافة إلى ذلك حقق النموذج Stacking Ensemble أفضل أداء من حيث الدقة وباقي المقاييس، لكنه استغرق زمن تنفيذ مرتفع مقارنة بالنماذج الأخرى، في المقابل قدم Random Forest أداء قريب من النموذج الأفضل مع كفاءة زمنية واستقرار مرتفع، مما يجعله الأنسب للتطبيق العملي.

✓ اختيارالنموذج النهائي

من خلال نتائج هذه الدراسة تم تحليل النماذج المستخدمة اعتمادا على المقاييس الإحصائية التالية Accuracy و Precisionو Recallو F1-Scoreو ROC-AUC، وزمن التنفيذ بهدف تحديد النموذج النهائي الذي سيتم تحسينه وتحويله لتطبيق خاص بالتنبؤ بمرض الكلى.

بالاعتماد على المقارنة حسب عدد السمات أوضحت النتائج أن استخدام 3 سمات حقق نتائج جيدة ثم تحسن الأداء عند استخدام 5سمات، أما عند استخدام 10سمات تم تسجيل أعلى النتائج لكن يُلاحظ زيادة في تعقيد النموذج وزمن التنفيذ. ومن بين النماذج التجميعية التي حققت نتائج أفضل يأتي Stacking Ensemble كأعلى دقة بين جميع النماذج، خاصة باستخدام 10سمات ولكن بزمن تنفيذ مرتفع، ثم يأتي النموذج Random Forest الذي قدم ثاني أفضل أداء بدقة تجاوزت 99% مع زمن تنفيذ أقل واستقرار مرتفع وهذا ما يجعله أفضل اختيار للتطبيق العملي.

وبناءً على المقارنات المتحصل عليها تم اختيار Random Forest كنموذج نهائي مع الاعدادات التالية:

5 سمات، تقنيةRFE لاختيار أهم السمات، تعويض القيم المفقودة باستخدامRegression ،معالجة القيم الشاذة باستخدامWinsorization ، وتطبيق Min-Max Scaling لتقييس البيانات. هذا الاختيار راجع الى أن النموذج وُفق بين الدقة العالية، الاستقرار الاحصائي، سرعة التنفيذ وتقليل التعقيد.

1.5 تطوير النموذج النهائي

بعد اختيار النموذج Random Forest الذي يعتبر من أكثر النماذج استقرارا وكفاءة في التعامل مع البيانات الطبية إضافة إلى تحقيق نتائج دقيقة، تم الانتقال إلى مرحلة التطوير لتحسين الأداء وزيادة القدرة على التعميم وتوفير نموذج أكثر موثوقية وقابلية للاستخدام العملي للتنبؤ بالمرض.

1.1.5 التحسين

تم تنفيذ تحسين النموذج النهائي بهدف إيجاد أفضل الاعدادات الممكنة له، وتحسين جودة التنبؤ بحيث اعتمدت هذه المرحلة على تقنيات ضبط المعاملات واستخدام التحقق المتقاطع للحصول على نموذج قابل للتعميم.

2.1.5 ضبط المعاملات(Hyperparameter Tuning)

لاختيار أفضل المعاملات للنموذج Random Forest تم استخدام تقنية **Grid Search**، بحيث تعتمد على تجربة عدة تركيبات مختلفة من المعاملات المحددة مسبقاً، ثم تقييم أداء كل تركيبة باستخدام 10-Fold Cross Validation واختيار الأفضل بينها اعتمادًا على مقاييس التقييم التالية:

Accuracy ،F1-Score،Reall كمقاييس أساسية أثناء التحسين نظرًا لأهميتها في البيانات الطبية.

في عملية التحسين تم ضبط أفضل قيم للمعاملات المهمة التالية:

- عدد الأشجار(n_estimators)= 100;
- العمق الأقصى للشجرة(max_depth) = 3;
- أقل عدد عينات لتقسيم العقدة(min_samples_split) = 2;
- أقل عدد عينات في الأوراق النهائية(min_samples_leaf) = 1;
- العدد الأقصى للسمات (max_Feature)= 'sqrt'.

والتي تم تدريب النموذج عليها فنتج عنها القيم التالية:

- Accuracy = 0.990625;
- F1-Score = 0.98795;

- Reall =1.0.

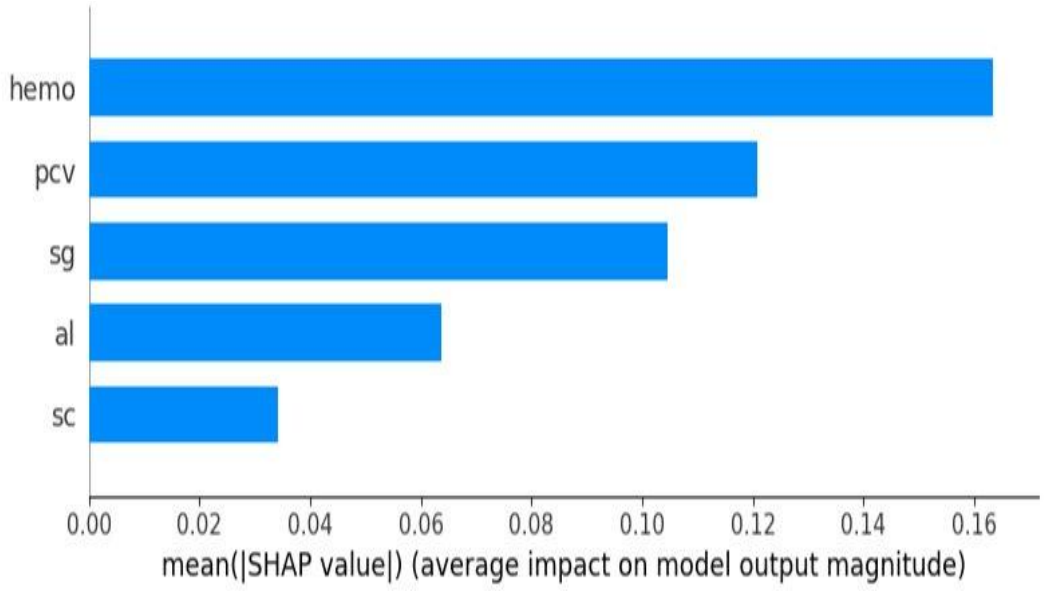
3.1.5 قابلية التفسير (Explainability SHAP)

بعد تدريب النموذج Random Forest على أفضل قيم للمعاملات، تم التركيز على تفسير قراراته وتحليل مساهمة كل خاصية في عملية التصنيف باستخدام تقنية SHAP (SHapley Additive exPlanations) التي تعتبر من أشهر تقنيات التعلم الالي، بحيث تعتمد على حساب قيم SHAP لتوضح مدى تأثير كل متغير على التنبؤ النهائي للنموذج.

في هذه الدراسة ساعدت قابلية التفسير في تفسير السمات الخمسة المختارة وعلاقتها بالمتغير الهدف والمتمثلة في: الهيموغلوبين (HEMO)، الثقل النوعي للبول(SG) ، حجم الخلايا المكدسة(PCV)، الألبومين(AL)، الكرياتينين في المصل(SC). باستخدام ثلاث مخططات:

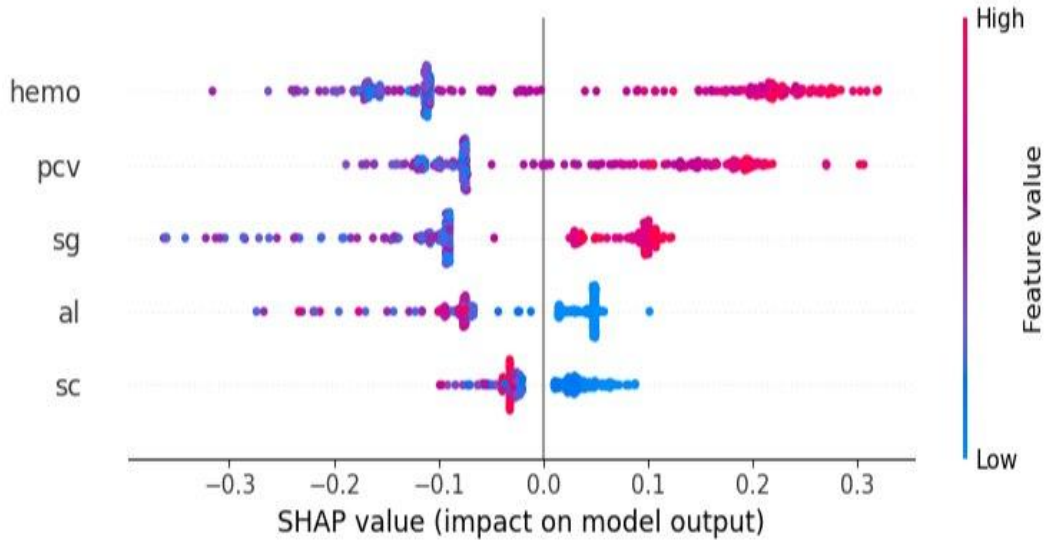
1. **مخطط أهمية السمات (Feature Importance Plot):** لتحديد الخصائص الأكثر تأثيرًا، وهذا ما يوضحه مخطط متوسط التأثير (قيمة SHAP لكل خاصية) على حجم مخرجات النموذج. من خلاله تبين أن: مستوى الهيموغلوبين من أهم العوامل التي يعتمد عليها النموذج، لأن الخاصية hemo كانت أكثر تأثيرا في التنبؤ. كما تليها الخاصيتين pcv،sg المهمتين في تحسين دقة التصنيف. أما sc،al كانت أقل تأثير مقارنة بالخصائص الثلاثة.

كما ان المخطط يدل على ان النموذج يعتمد أكثر على خصائص تحاليل الدم ووظائف الكلى أثناء اتخاذ القرار.



الصورة 12: مخطط متوسط التأثير (قيمة SHAP لكل خاصية) على حجم مخرجات النموذج.

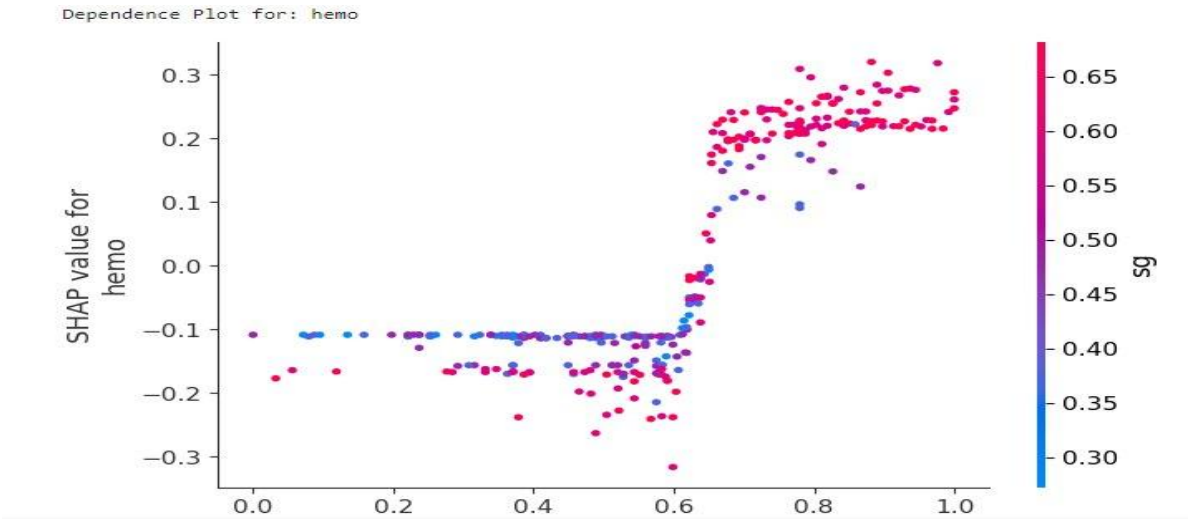
2. **المخطط التلخيصي (Summary Plot):** وهو ما تعرضه الصورة التالية التي تمثل تأثير قيم الخصائص المختلفة على المتغير الهدف واتجاه هذا التأثير. حيث تشير النقاط الحمراء الى القيم المرتفعة للسمة، أما النقاط الزرقاء تشير للقيم المنخفضة لها. كما أن القيم الموجبة ل SHAP تدفع النموذج نحو التنبؤ بالفئة الإيجابية (not ckd) بينما السالبة تقلل احتمال التنبؤ بها. إضافة الى ذلك نلاحظ: ارتفاع قيم hemo يرتبط بزيادة تأثيرها الإيجابي على المتغير الهدف، كما ان باقي الخصائص pcv،sg ،sc كان لها تأثيرا متفاوت في عملية التصنيف.



الصورة 13 : مخطط تأثير قيم الخصائص المختلفة على المتغير الهدف و اتجاه هذا التأثير.

3. **مخطط الاعتماد (Dependence Plot):** حيث تم اختيار الخاصية hemoمن بين الخصائص المدروسة كما يوضحه مخطط العلاقة بين قيمها وتأثيرها على مخرجات النموذج، بالإضافة إلى دراسة تفاعلها مع الخاصيةsg .

أظهرت النتائج أن القيم المنخفضة لـ hemoترتبط بقيم SHAP سالبة، بينما يؤدي ارتفاعها إلى زيادة التأثير الإيجابي على قرار النموذج. كما يظهر وجود تفاعل بين الخاصيتين hemoوsg، حيث يؤثر تغير قيم sgعلى قوة تأثي hemoفي بعض الحالات. يدل ذلك على أن النموذج يأخذ بعين الاعتبار العلاقات والتفاعلات بين السمات أثناء عملية التنبؤ.



الصورة 14: مخطط الاعتماد للخاصية hemo وتأثيرها على مخرجات النموذج.

4.1.5 التقييم النهائي(Final Evaluation)

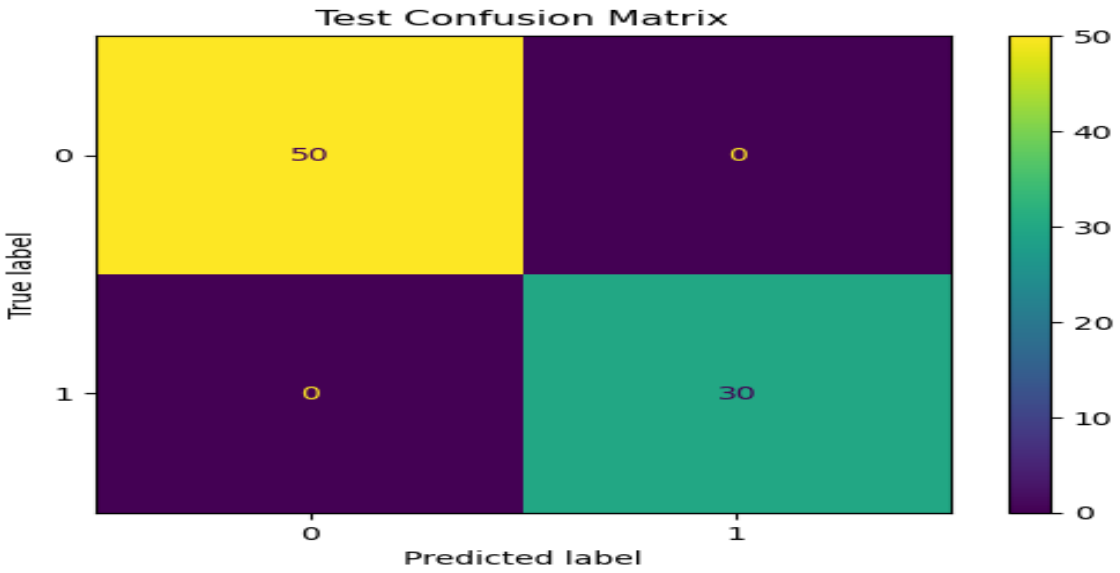
بعد تحسين النموذج Random Forest من خلال ضبط أفضل قيم للمعاملات ثم تدريبه عليها، تم إجراء التقييم النهائي باستخدام بيانات الاختبار التي لم تُرى أثناء التدريب أو التحسين للحصول على تقييم دقيق ونموذج قادر على التعميم. حيث ان بيانات الاختبار تم معالجتها سابقا بنفس سلسلة المعالجة للنموذج (rfe + Regression + winsor + minmax).

نتج عن التقييم النهائي القيم الإحصائية التالية:

ROC-AUC=1 / F1-score= 1 / Recall= 1 / Precision= 1 / Accuracy= 1.

التي اشارت أنه تم تصنيف جميع العينات بشكل صحيح دون أخطاء على مجموعة الاختبار. كما تُظهر أن النموذج حقق أداءً مرتفعًا ودقة عالية في عملية التنبؤ.

كما تم استخدام مصفوفة الالتباس(Confusion Matrix) التالية: التي بينت أنه تم التعرف على جميع الحالات الإيجابية (30حالة غير مصابة) والسلبية (50حالة مصابة) بدقة كاملة دون تسجيل أي أخطاء تصنيف.



الصورة 15: مصفوفة الالتباس (Confusion Matrix) للتقييم النهائي على بيانات الاختبار.

بلاعتماد على النتائج يمكن الاستعانة بالنموذج **Random Forest** للكشف عن مرض الكلى المزمن لبيانات هذه الدراسة، كما أشارت إلى أن اختبار النموذج على بيانات خارجية مستقبلًا قد يساهم في التحقق بشكل أكبر من قدرته على التعميم.

5.1.5 من نموذج الى منصة تفاعلية

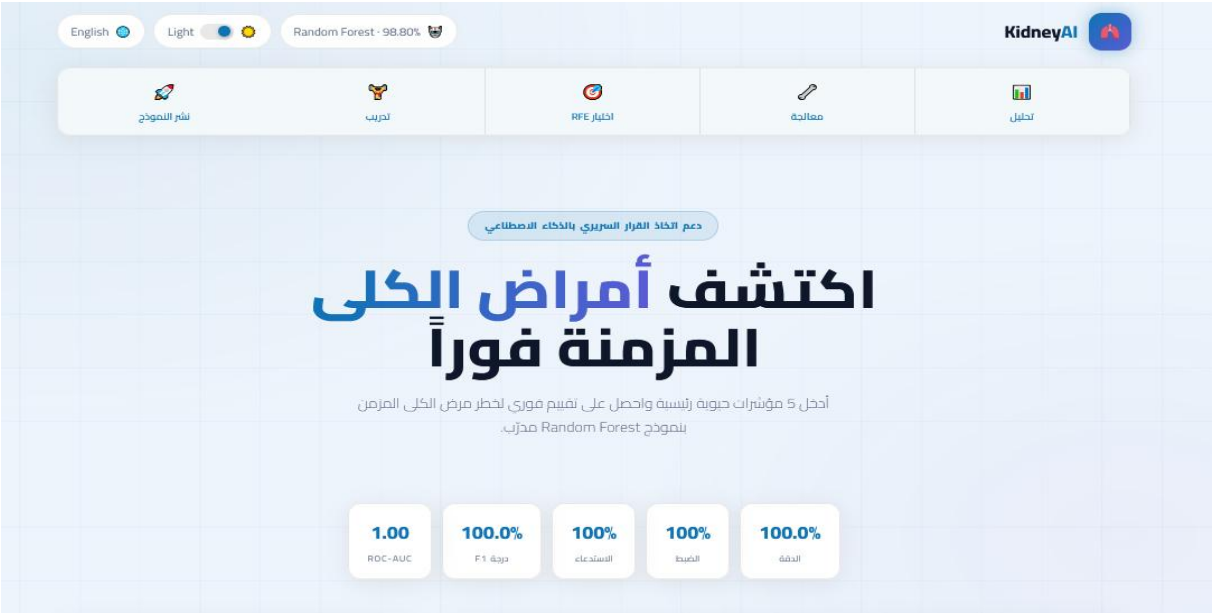
بعد بناء النموذج النهائي وتحسينه، تم الانتقال الى مرحلة تطوير منصة تفاعلية تهدف إلى تحويل Random Forest من نموذج تعلم آلي إلى نظام تطبيقي قابل للاستخدام في التطبيقات الطبية الخاصة بالكشف المبكر عن مرض الكلى المزمن. تم تطوير المنصة بالاعتماد على إطار عمل Flaskبلغة البايثون، مع دمج النموذج النهائ Random Forest داخل تطبيق ويب قابل للاستخدام حيث يتيح للمستخدم إدخال البيانات الطبية المختارة للحصول على نتائج التنبؤ بالمرض بطريقة سهلة وسريعة. يعتمد النظام في بنيته على استقبال المدخلات المتمثلة في التحاليل المخبرية المحددة عبر واجهة ويب، ثم يتم تمريرها إلى النموذج بعد تطبيق نفس سلسة المعالجة المسبقة خلال تدريبه، وهذا يضمن اتساق للبيانات واستقرار النتائج. كما تم حفظ النموذج النهائي بصيغة (PKL) مما يتيح إعادة استخدامه داخل التطبيق دون تدريبه مرة أخرى عند كل تشغيل للنظام.

في واجهة الويب تم الاعتماد على مقاييس التقييم النهائي لمجموعة بيانات الاختبار لعرض أداء النموذج داخل المنصة. حيث ان هذه المقاييس تُثبت الأداء الممتاز للنموذج على هذه البيانات غير المرئية أثناء التدريب كما تشير إلى قدرته على التعميم. كما تم توفير عدة خصائص وظيفية لتحسين تجربة المستخدم التي تتمثل في واجهة سهلة وبسيطة بتصميم احترافي يمكن من خلاله تغيير اللغة ودعم الوضع الداكن.

1.6 عرض المنصة

يتم في هذا الجزء عرض تفصيلي للمنصة تحت عنوان"**KidneyAI**" لاكتشاف مرض الكلى المزمن، والتي تتيح للمستخدم إدخال المؤشرات الطبية الأساسية للمرض، بهدف دعم اتخاذ القرار السريري بالذكاء الاصطناعي

الصور التالية تمثل واجهة النظام التي يتم فيها عرض قيمة المقياس **F1** للنموذج النهائي **Random Forest** بعد التحسين والتي تقدر ب **98.8%**، كما يتم عرض **مقاييس التقييم النهائي** للنموذج على مجموعة بيانات الاختبار. إضافة إلى ميزات تفاعلية تشمل تغيير اللغة بين **العربية** و**الإنجليزية** وتبديل الوضع بين **الداكن** و**الفاتح**، مما يعزز سهولة الاستخدام ومرونة التفاعل مع النظام:



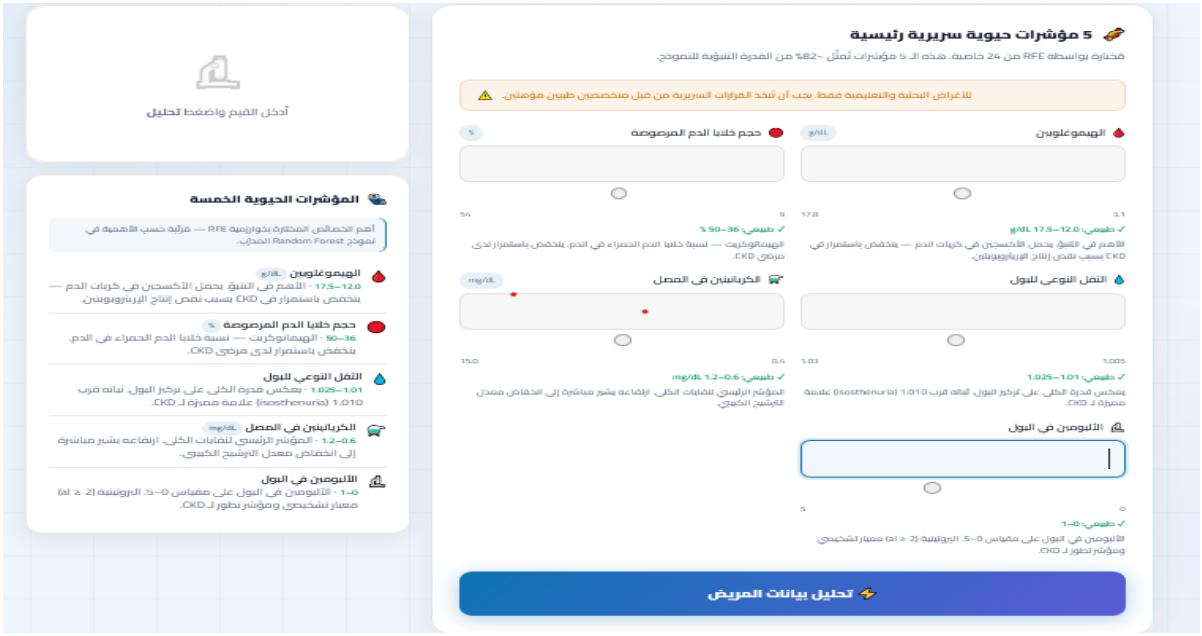
الصورة 16 : واجهة التطبيق باللغة العربية والوضع الفاتح.

الصورة 17: واجهة التطبيق باللغة الانجليزية والوضع الداكن.

تمكن المنصة للمستخدم من إدخال 5مؤشرات حيوية سريرية رئيسية وهي الهيموغلوبين، حجم خلايا الدم المرصوصة، الثقل النوعي للبول، الكرياتينين في المصل، الالبومين في البول، بحيث يوجد تحت كل حقل خاص بالسمة ملاحظتين واحدة تعبر عن مجال القيمة والاخرى تُعطي معلومة للسمة.

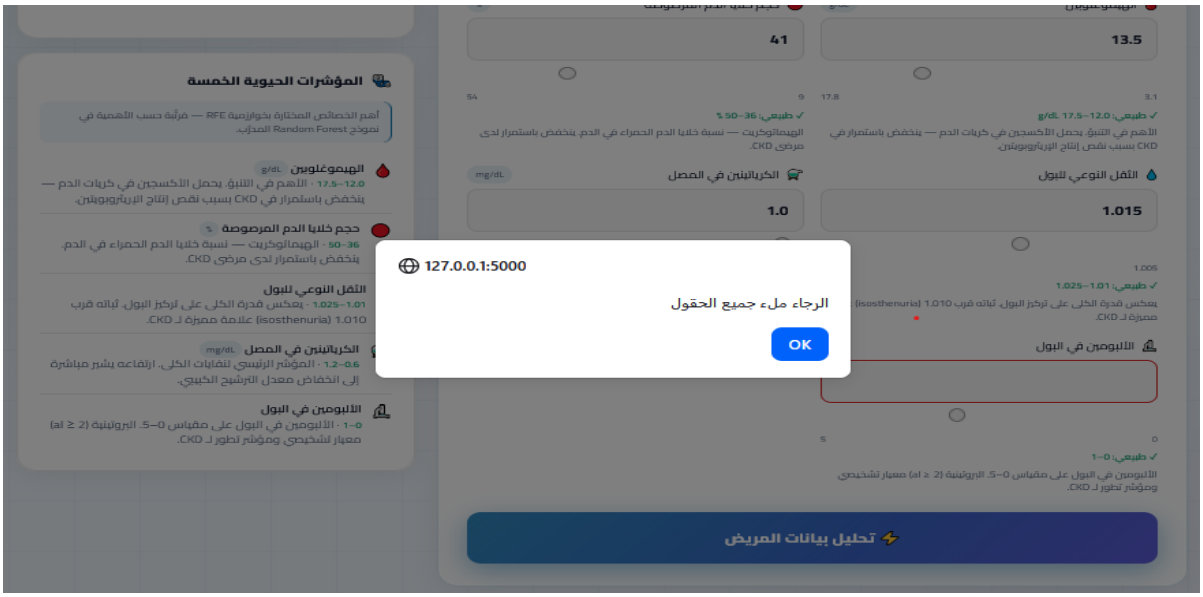
عند ادخال جميع الحقول يتم الضغط على زر "**تحليل بيانات المريض**" بهدف عرض نتيجة التنبؤ (احتمال الاصابة/احتمال عدم الإصابة) مع نسبة ثقة (Confidence Score) التي تعكس درجة يقين النموذج.

والصورة التالية توضح ذلك:



الصورة 18: واجهة إدخال 5 قيم للمؤشرات الحيوية السريرية الرئيسية.

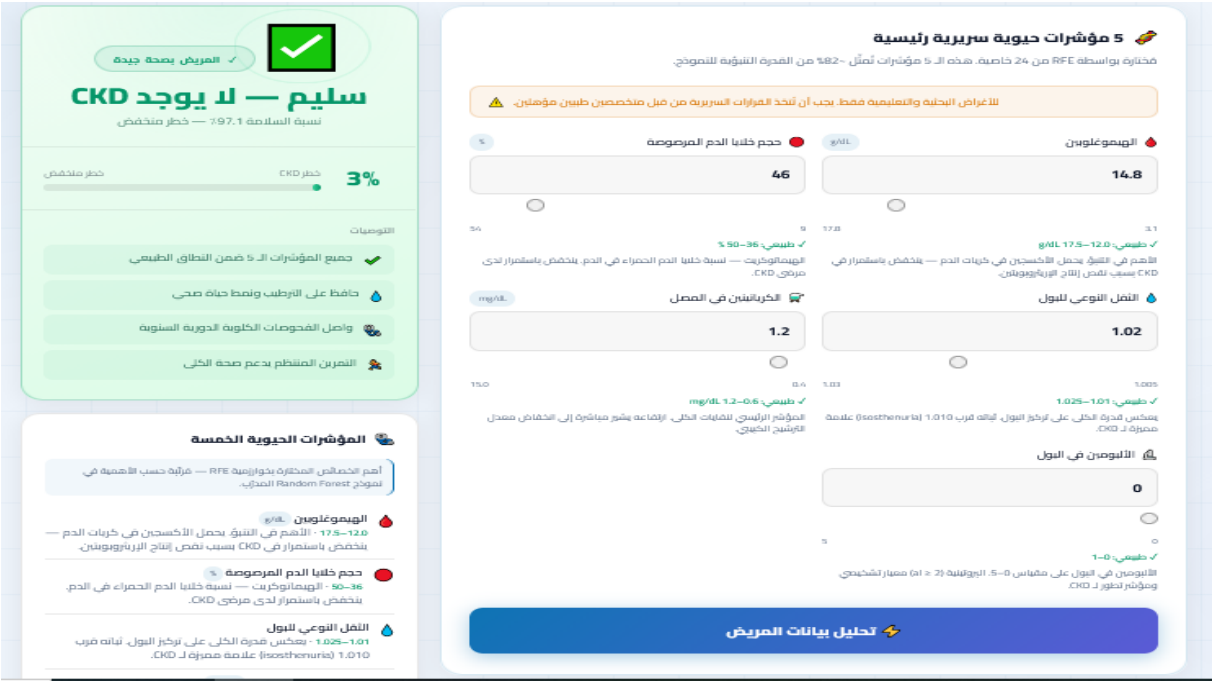
يجب ادخال جميع قيم المؤشرات فبمجرد تجاهل أحد هذه الحقول تظهر رسالة تحذير، والتي تمنع ظهور نتيجة التنبؤ(لانه عند تجاهل حقل يعني تم الاعتماد على أقل من 5مؤشرات)، هذا ماتُظهره الصورة أدناه :



الصورة 19: رسالة تحذير عند تجاهل ملء أحد الحقول.

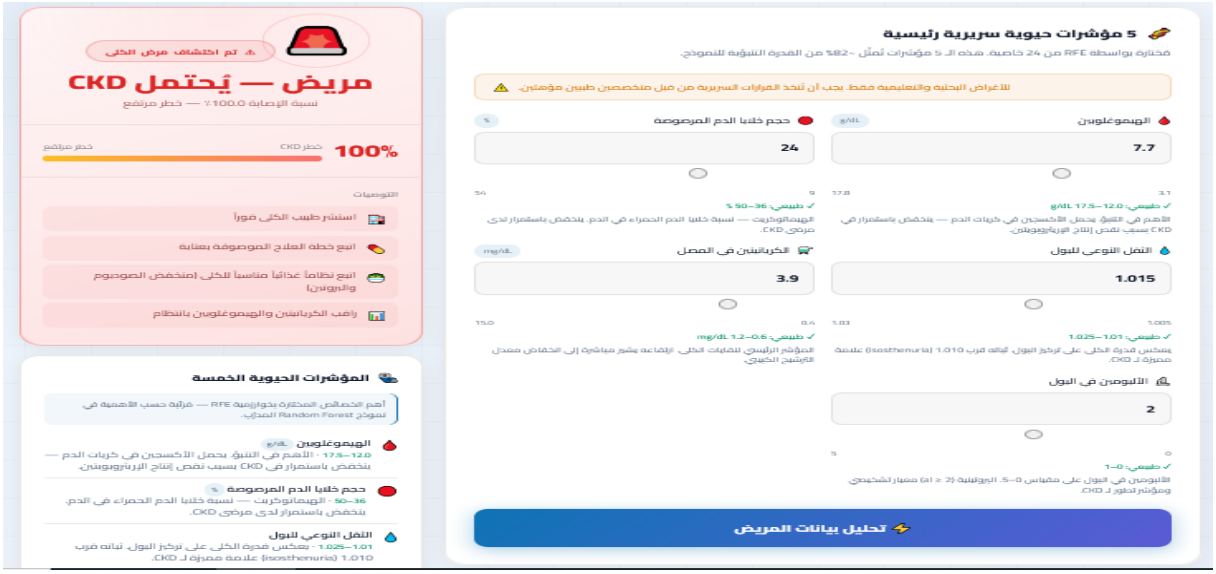
بعد إدخال جميع قيم المؤشرات السريرية الرئيسية، يقوم المستخدم بتفعيل عملية التنبؤ حيث يتم إرسال القيم إلى نموذج **Random Forest** المدرب، ومعالجتها وفق نفس خطوات المعالجة المسبقة، ثم عرض النتيجة مباشرة في الواجهة.

كما يتم مقارنة البيانات بالقيم الطبيعية فعندما تكون في المجال الصحي تظهر نتيجة إيجابية أي المستخدم سليم (لا يوجد CKD) مع نسبة الثقة (Confidence Score) وبعض التوصيات وهذا ما تمثله الصورة التالية:



الصورة 20: واجهة تحتوي بيانات ذات قيم طبيعية تُظهر أن المستخدم سليم(لا يوجد CKD) .

وعندما تكون قيم البيانات غير طبيعية تظهر نتيجة سلبية أي المستخدم مريض (يُحتمل CKD) مع نسبة الثقة (Confidence Score) وبعض التوصيات وهذا ما تمثله الصورة التالية:



الصورة 21: واجهة تحتوي بيانات ذات قيم غير طبيعية تُظهر أن المستخدم مريض (يُحتمل CKD).

1.7 الخاتمة

في هذا الفصل تم التطرق إلى عرض العمل التطبيقي وشرح الخطوات المعتمدة لبناء نموذج تنبؤي وذلك بالاستعانة بتقنيات التعلم الالي وتحويله لمنصة قابلة للاستخدام الخاصة بمرض القصور الكلوي.

تم فيه عرض مراحل الدراسة التي بدأت بوصف مجموعة البيانات المستخدمة ثم عرض بيئة العمل المعتمدة كما تم شرح مراحل التنفيذ الأساسية للدراسة التي تتكون من مرحلة تحليل البيانات تليها المعالجة المسبقة التي تم فيها معالجة القيم المفقودة والشاذة ترميز المتغيرات وتقييس البيانات وأيضا اختيار 5مؤشرات الأكثر تأثيرا في المرض من خلال تقنيات معتمدة.

وفي المرحلة الأخيرة تم تدريب النماذج المحددة وتقييمها اعتمادا على مقاييس مناسبة للبيانات الطبية بهدف إجراء المقارنة الموضوعية ومن خلالها تم اختيار النموذج النهائي ذو الأداء العالي والكفاءة الزمنية وتحسينه عن طريق ضبط المعاملات.

وأخيرا تم إجراء تقييم النموذج النهائي على مجموعة بيانات الاختبار حيث أظهر هذا التقييم الأداء المثالي للنموذج. والذي تم الاعتماد عليه لبناء نظام تفاعلي وعرض واجهاته وشرح كيفية استعماله للمستخدمين للتنبؤ بمرض الكلى المزمن.

الخاتمة

في ختام هذه الدراسة، تم بناء نظام قادر على التنبؤ بمرض الفشل الكلوي بالاعتماد على تقنيات التعلم الآلي، من أجل تحسين دقة التشخيص المبكر والمساهمة في دعم اتخاذ القرار الطبي. وهدف من هذه الدراسة هو الحفاظ على أحسن أداء ممكن مع تقليل عدد السمات الى خمس اساسية، لتبسيط النموذج وتقليل التعقيد الحسابي.

تناولت هذه الدراسة في بدايتها مفاهيم حول الذكاء الاصطناعي والتعلم الآلي وتطبيقاته في المجال الطبي، بالإضافة الى مفهوم مرض الفشل الكلوي وأسبابه وطرق تشخيصه، ثم تم شرح التقنيات المستخدمة في معالجة البيانات واختيار السمات، كما تم التطرق الى الجانب التطبيقي من خلال تطبيق معالجة البيانات باستخدام عدة تقنيات لملء القيم الفارغة ومعالجة القيم الشاذة وتقييس البيانات ثم تطبيق تقنيات اختيار السمات Chi-Squareو Mutual Information و RFEمن أجل تحديد أهم 5 سمات تساعد على التنبؤ بالمرض.

ومن جانب تطبيق خوارزميات التعلم الآلي فقد تم استخدام 5 نماذج لتدريب البيانات وهي : Decision Tree و Random Forestو Extra Treesو XGBoostوStacking Ensemble، مع تقييم أدائها باستخدام عدة مقاييس مثل Accuracy و Precisionو Recallو F1-Scoreو ROC-AUC مع أسلوب 10-Fold Cross Validation .

وقد أظهرت نتائج المقارنة أن نماذج التجميع، خاصة Random Forest وStacking Ensemble، حققت أفضل أداء من حيث الدقة والاستقرار، مما يؤكد فعالية تقنيات التعلم الآلي في التنبؤ المبكر بمرض الفشل الكلوي حتى مع استخدام عدد محدود من السمات. كما ساهمت الدراسة في إبراز أهمية تقنيات اختيار السمات في تحسين الأداء وتقليل زمن المعالجة.

مع ما تم إنجازه لحد الآن إلا أنه لا تزال بعض النقائص، مثل محدودية حجم البيانات (حيث تم البحث على بيانات حقيقية مخبرية بمساعدة مخبر لتحاليل الطبية ولكنها لم تستوفي الشروط المطلوبة فتم الاستغناء عنها والاستعانة بهذه البيانات المحدودة)، وارتفاع زمن التنفيذ في بعض النماذج بسبب التعقيد، إضافة إلى ضرورة اختبار النموذج على بيانات طبية حقيقية ومتنوعة لضمان قدرة أكبر على التعميم.

أما بالنسبة للآفاق المستقبلية المتعلقة بهذا العمل، يمكن استخدام تقنيات التعلم العميق والشبكات العصبية على بيانات أكبر من أجل تحسين دقة التنبؤ وبالتالي تطوير نظام طبي قادر على مساعدة الأطباء في التشخيص المبكر واتخاذ القرارات الطبية بشكل أسرع وأكثر فعالية.

وفي الأخير، قدمت هذه الدراسة مساهمة عملية في تطبيق الذكاء الاصطناعي في المجال الطبي والرعاية الصحية، وذلك باستغلال تقنيات التعلم الآلي من أجل بناء نظام فعال ودقيق وقابل للتطوير، يساعد في عملية التنبؤ المبكر بالأمراض المزمنة وبالتالي تحسين جودة الخدمات الطبية.

قائمة المراجع

[1] H. Jbour, " أكاديمية حسوب، "الذكاء الاصطناعي : دليلك الشامل, Jan. 26, 2023. [Online]. Available: https://academy.hsoub.com/programming/artificial-intelligence/الذكاء-الاصطناعي/.[Accessed: Mar. 10, 2026].

[2]"أهمية وتأثير الذكاء الاصطناعي في جميع المجالات وكيف أصبح ثورة العصر الحديث," Ghanaem. [Online]. Available: https://ghanaem.com/blogs/الذكاء-الاصطناعي/. [Accessed: Mar. 10, 2026].

[3] Bakkah Learning, "أنواع الذكاء الاصطناعي," Bakkah Knowledge Center. [Online]. Available: https://bakkah.com/ar/knowledge-center/أنواع-الذكاء-الاصطناعي. [Accessed: Mar. 10, 2026].

[4] H. Jabbour, " أكاديمية حسوب، "مجالات الذكاء الاصطناعي, Jan. 30, 2023. [Online]. Available: https://academy.hsoub.com/programming/artificial-intelligence/مجالات-الذكاء-الاصطناعي/. [Accessed: Mar. 10, 2026].

[5] A. Belouahem, "Using Artificial Intelligence in the Field of Medicine," University of 8 May 1945 Guelma, Algeria, 2023. [Online]. Available: https://dspace.univ-guelma.dz/jspui/bitstream/123456789/17642/1/في%20مجال%20الطب%20د%20بلواهم%20عبد%20الحليم.doc. [Accessed: Mar. 10, 2026].

[6] N. Al-Dabbas, "الذكاء الاصطناعي في الطب," WebTeb, Oct. 31, 2022. [Online]. Available: https://www.webteb.com/articles/36654_الذكاء-الاصطناعي-في-الطب. [Accessed: Mar. 10, 2026].

[7] A. E. El Naggar, " أكاديمية حسوب، "نظرة سريعة على مجال تعلم الآلة, Mar. 19, 2023. [Online]. Available: https://academy.hsoub.com/programming/artificial-intelligence/نظرة-سريعة-على-مجال-تعلم-الآلة-r1933/. [Accessed: Mar. 10, 2026].

[8] C. R. China, "Types of machine learning," IBM Think, IBM, [Online]. Available: https://www.ibm.com/sa-ar/think/topics/machine-learning-types. [Accessed: Mar. 10, 2026].

[9] "The Critical Role of High Quality Data in Machine Learning," Kotwel. [Online]. Available: https://kotwel.com/the-critical-role-of-high-quality-data-in-machine-learning/. [Accessed: Mar. 10, 2026].

[10] Mayo Clinic, "Chronic kidney disease (CKD) – Symptoms and causes," *mayoclinic.org*, 2026. [Online]. Available: https://www.mayoclinic.org/ar/diseases-conditions/chronic-kidney-disease/symptoms-causes/syc-20354521. *[Accessed: Mar. 10, 2026]*.

[11] "مرض الكلى المزمن — العلامات والأعراض"، *الموسوعة الحرة – ويكيبيديا*، 2026. [Online] Available: https://ar.wikipedia.org/wiki/مرض_الكلى_المزمن#العلامات_والأعراض . *[Accessed: Mar. 10, 2026]*.

[12] K. Berkache *et al.*, "End-stage kidney disease in Sidi Bel Abbes, Algeria," *Clinical Epidemiology and Global Health*, vol. 12, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2213398421001160. [Accessed: Mar. 11, 2026].

[13] N. Meguellati *et al.*, "Incidence of treated end-stage renal disease in south-eastern Algeria," *La Tunisie Médicale*, vol. 101, no. 6, 2023. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/38982968/. [Accessed: Mar. 11, 2026].

[14] Mayo Clinic, "Chronic kidney disease (CKD) – Symptoms and causes," *mayoclinic.org*, 2026. [Online]. Available: https://www.mayoclinic.org/ar/diseases-conditions/chronic-kidney-disease/symptoms-causes/syc-20354521. *[Accessed: Mar. 10, 2026]*.

[15] Cleveland Clinic, "Kidney Failure," *my.clevelandclinic.org*, 2026. [Online]. Available: https://my.clevelandclinic.org/health/diseases/17689-kidney-failure . *[Accessed: Mar. 10, 2026]*.

[16] NHS, "Kidney disease – Diagnosis," *NHS.uk*, 2024. [Online]. Available: https://www.nhs.uk/conditions/kidney-disease/diagnosis/ . *[Accessed: Mar. 10, 2026]*.

[17] A. Simeri, G. Pezzi, R. Arena, et al., "Artificial intelligence in chronic kidney diseases: methodology and potential applications," *International Urology and Nephrology*, 2024. https://pubmed.ncbi.nlm.nih.gov/39052168/.*[Accessed: Mar. 10, 2026]*.

[18] Khalid F, Alsadoun L, Khilji F, Mushtaq M, Eze-Odurukwe A, Mushtaq MM, Ali H, Farman RO, Ali SM, Fatima R, Bokhari SFH. Predicting the Progression of Chronic Kidney Disease: A Systematic Review of Artificial Intelligence and Machine Learning Approaches. Cureus. 2024 May 12;16(5):e60145. doi: 10.7759/cureus.60145. PMID: 38864072; PMCID: PMC11166249. https://pubmed.ncbi.nlm.nih.gov/38864072/. *[Accessed: Mar. 10, 2026]*.

[19] D. A. Debal and T. M. Sitote, "Chronic kidney disease prediction using machine learning techniques," *Journal of Big Data*, vol. 9, art. no. 109, Nov. 2022, doi: 10.1186/s40537-022-00657-5. https://link.springer.com/article/10.1186/s40537-022-00657-5 . *[Accessed: Mar. 10, 2026]*.

[20] H. Khalid, A. Khan, M. Z. Khan, G. Mehmood and M. S. Qureshi, "Machine Learning Hybrid Model for the Prediction of Chronic Kidney Disease," *Computational Intelligence and Neuroscience*, vol. 2023, Article ID 9266889, Mar. 14, 2023. doi: 10.1155/2023/9266889. https://pmc.ncbi.nlm.nih.gov/articles/PMC10030216/ . *[Accessed: Mar. 10, 2026]*.

[21] A. Charleonnan, T. Fufaung, T. Niyomwong, W. Chokchueypattanakit, S. Suwannawach, and N. Nintachai, "Predictive analytics for chronic kidney disease using machine learning techniques," in Proc. Management and Innovation Technology International Conference (MITiCON), 2016, pp. 80–83.

[22] A. Salekin and J. Stankovic, "Detection of chronic kidney disease and selecting important predictive attributes," in Proc. 2016 IEEE International Conference on Healthcare Informatics (ICHI), 2016, pp. 262–270.

[23] S. Tekale, P. Shingavi, S. Wandhekar, and A. Chatorikar, "Prediction of chronic kidney disease using machine learning algorithm," International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), vol. 7, no. 10, pp. 92–96, 2018.

[24] J. Xiao et al., "Comparison and development of machine learning tools in the prediction of chronic kidney disease progression," Journal of Translational Medicine, vol. 17, no. 1, pp. 1–13, 2019.

[25] M. Almansour and T. E. G. Al-Juaid, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," International Journal of Advanced Computer Science and Applications, vol. 10, no. 8, pp. 89–96, 2019.

[26] V. Kumar, "Evaluation of computationally intelligent techniques for breast cancer diagnosis," Neural Computing and Applications, vol. 33, no. 8, pp. 3195–3208, 2021.

[27] M. U. Uddin, S. A. Ali, M. A. Rahman, and A. K. M. Masum, "A comparative study of machine learning algorithms for chronic kidney disease prediction," *Journal of Big Data*, vol. 8, no. 1, pp. 1–21, 2021.

[28] S. N. Almansour, F. M. Alharthi, and M. A. Alzahrani, "Machine learning approaches for the prediction of chronic kidney disease: A systematic review," *Cureus*, vol. 15, no. 3, pp. 1–12, 2023.

[29] R. K. Singh and P. Kumar, "Machine learning approaches for chronic kidney disease prediction: A review," *International Journal of Medical Informatics*, vol. 146, pp. 104–118, 2021.

[30] M. Daku, "CKD Disease Dataset," Kaggle, 2019. [Online]. Available: https://www.kaggle.com/datasets/mansoordaku/ckdisease

[31] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2017. [Online]. Available: https://archive.ics.uci.edu/ml

[32] S. Vijayarani and S. Dhayanand, "Kidney Disease Prediction Using SVM and ANN Algorithms," International Journal of Computing and Business Research, 2015.

[33] A. Géron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow," 2nd ed., O'Reilly Media, 2019.

[34] J. Han, M. Kamber, and J. Pei, "Data Mining : Concepts and Techniques," 3rd ed., Morgan Kaufmann, 2011.

[35] J. Tukey, "Exploratory Data Analysis," Addison-Wesley, 1977.

[36] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA : Springer, 2009.

[37] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[38] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine Learning, vol. 63, no. 1, pp. 3–42, 2006.

[39] T. Chen and C. Guestrin, "XGBoost : A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2016, pp. 785–794.

[40] D. H. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.

[41] Python Software Foundation, "What's New in Python 3.0," *Python Documentation*, Dec. 3, 2008. [Online]. Available: https://docs.python.org/3/whatsnew/3.0.html. Accessed: Apr. 27, 2026.

[42] Anaconda, Inc., "Anaconda Distribution," *Anaconda*, [Online]. Available: https://www.anaconda.com/. Accessed: Apr. 27, 2026.

[43] Project Jupyter, "Project Jupyter Documentation," *Jupyter*, [Online]. Available: https://docs.jupyter.org/. Accessed: Apr. 27, 2026.

[44] O. Saleh, "تعرف على إطار العمل فلاسك Flask," *أكاديمية حسوب*, 31 ديسمبر 2023. [Online]. Available: https://academy.hsoub.com/programming/python/flask/%D8%AA%D8%B9%D8%B1%D9%81-%D8%B9%D9%84%D9%89-%D8%A5%D8%B7%D8%A7%D8%B1-%D8%A7%D9%84%D8%B9%D9%85%D9%84-%D9%81%D9%84%D8%A7%D8%B3%D9%83-flask-r2201/. [Accessed: 28 أبريل 2026].

[45] "Blog Details 10," *Repoteq*, [Online]. Available: https://newsite.repoteq.com/blogs/Details/10. [Accessed: 28-Apr-2026].

[46] Université Mohamed Khider Biskra, "أدوات القياس وجمع البيانات (المحاور 23-30)," Plateforme d'enseignement à distance (Moodle), Algérie, [Online]. Available: https: //elearning.univ-biskra.dz/moodle2024/pluginfile.php/784141/mod_resource/content/1/23-30-%D8%A7%D9%84%D9%85%D8%AD%D8%A7%D9%88%D8%B1.pdf. [Accessed: May 5, 2026].

الملخص

الهدف من هذا العمل هو تطوير نظام ذكي للتنبؤ بمرض الكلى المزمن بالاعتماد على تقنيات التعلم الآلي، من خلال دراسة تأثير مراحل المعالجة المسبقة واختيار السمات على أداء النماذج المختلفة. تم تطبيق تقنيات متعددة لمعالجة البيانات شملت تعويض القيم المفقودة، معالجة القيم الشاذة، وتقييس البيانات، إضافة إلى استخدام تقنيات اختيار السمات مثل RFE و MIوChi-Square. كما تمت مقارنة عدة خوارزميات، وهي Decision Tree و Random Forestو Extra Treesو XGBoostوStacking Ensemble، باستخدام مقاييس تقييم مختلفة إضافةً إلى زمن التنفيذ. أظهرت النتائج أن نماذج التعلم الجماعي حققت أداءً أفضل مقارنة بالنماذج التقليدية، حيث سجل نموذج Stacking Ensembleأعلى دقة، بينما حقق نموذج Random Forest توازنًا فعالًا بين الدقة العالية وسرعة التنفيذ. كما بينت الدراسة أن استخدام خمس سمات مختارة ساهم في تحسين الأداء وتقليل تعقيد النموذج. وبناءً على النتائج المتحصل عليها، تم اعتماد نموذج Random Forest كنموذج نهائي ودمجه داخل تطبيق ويب تفاعلي باستخدامFlask ، مما يسمح بالحصول على نتائج التنبؤ بشكل مباشر وسريع عبر واجهة استخدام سهلة وفعالة.

Abstract:

The objective of this work is to develop an intelligent system for predicting chronic kidney disease based on machine learning techniques, by studying the impact of preprocessing stages and feature selection on the performance of different models. Various data preprocessing techniques were applied, including missing value imputation, outlier handling, and data scaling, in addition to feature selection techniques such as RFE, MI, and Chi-Square. Several algorithms were also compared, namely Decision Tree, Random Forest, Extra Trees, XGBoost, and Stacking Ensemble, using different evaluation metrics as well as execution time. The results showed that ensemble learning models achieved better performance compared to traditional models, where the Stacking Ensemble model recorded the highest accuracy, while the Random Forest model achieved an effective balance between high accuracy and execution speed. The study also showed that using five selected features contributed to improving performance and reducing model complexity. Based on the obtained results, the Random Forest model was adopted as the final model and integrated into an interactive web application using Flask, allowing direct and fast prediction results through an easy and efficient user interface.