مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(3) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (3) 2022

استخدام تقنيات البيانات الضخمة لتحسين التنبؤ بالحالة المرضية

د. ماهر إبراهيم *

د. ميرنا درغام**

م. حسن محمد وسوف***

(تاريخ الإيداع 3/2 /2022 . قُبِل للنشر في 9 /5 /2022)

□ ملخّص□

أدّى التضخم الهائل في حجم البيانات إلى خلق عجز في الأنظمة التقليدية المسؤولة عن التخزين والمعالجة الأمر الذي جعل تلك الأنظمة غير مجدية ولا سيما أن هذا النمو يتم بشكل متسارع. ظهرت تقنيات البيانات الضخمة لحل هذه المشكلة إذ أنها تمكنت من التعامل مع الكم الهائل من البيانات اعتماداً على النظام الموزع. يمتلك القطاع الصحي أحجاماً كبيرة جداً من المعطيات والتي تُعد مهمة جداً ويمكن الاعتماد عليها في كثير من الأمور منها التشخيص، إذ أنّ السجلات الصحية الخاصة بكل مريض في تزايد مستمر.

يُعد استغلال البيانات المتعلقة بالمريض ليتم الاعتماد عليها في بناء قرار طبي معين، من أهم النقاط المتعلقة بمعالجة البيانات الصحية كونها تسهم في تقليل الأخطاء الطبية وتخفيض الضغوطات على الأطباء. ركز البحث على تحسين المعالجة المسبقة للبيانات الصحية التي تحوي مجموعة من الأمراض مع الأعراض المرتبطة بكل مرض بالاعتماد على تحسين استخدام تقنيات البيانات الضخمة والذي أدى إلى زيادة دقة التنبؤ بالحالة الصحية.

وبناء عليه، قدّم هذا البحث نظام طبي محسن قادر على تشخيص المرض اعتماداً على الأعراض المحددة، مما يعطي القدرة على التعامل مع البيانات الصحية مهما كان حجمها وهذا ما لا نلتمسه في الأنظمة الصحية التقليدية.

الكلمات المفتاحية: البيانات الضخمة، تحليل البيانات الصحية، اتخاذ القرار، تعلم الآلة، الأنظمة الموزعة، خوارزميات تعلم الآلة، التنبؤ

* مدرس في قسم هندسة تكنولوجيا المعلومات –كليّة هندسة تكنولوجيا المعلومات والاتصالات –جامعة طرطوس –سوريا

** مدرس في كلية الهندسة المعلوماتية – جامعة دمشق – سوريا

***طالب ماجستير– قسم هندسة تكنولوجيا المعلومات –كليّة هندسة تكنولوجيا المعلومات والاتصالات – جامعة طرطوس– سوريا

مجلة جامعة طرطوس للبحوث والدراسات العلمية _ سلسلة العلوم الهندسية المجلد (6) العدد(3) 2022
Tartous University Journal for Research and Scientific Studies - engineering Sciences Series Vol. (6) No. (3) 2022

The use of big data technologies to improve the prediction of health status

Dr. Maher Ibrahim*
Dr. Mirna Dargham**
Eng. Hasan Mohammad Wassouf***

(Received 2 / 3/ 2022 . Accepted 9/ 5/ 2022)

□ ABSTRACT □

The massive inflation in the volume of data created a deficit in the traditional systems responsible for storage and processing, which made these systems useless, especially as this growth is accelerating. Big data technologies emerged to solve this problem as they were able to deal with a huge amount of data depending on the distributed system. The health field has very large volumes of data, which are very important and reliable in many matters, including diagnosis, as the health records of each patient are constantly increasing. Exploiting all previous data related to the patient to be relied upon in building a specific medical decision, is one of the most important points related to health data processing as it contributes to reducing medical errors and reducing pressure on Physician. The research focuses on improving the pre-processing of health data that contains a group of diseases with symptoms associated with each disease based on the improvement of the use of big data technologies, which had led to an increase in the accuracy of health prediction. Accordingly, this research presents an improved medical system capable of diagnosing disease based on specific symptoms, which gives the ability to deal with health data regardless of its size, and this is what we do not seek in traditional health systems.

Key words: big data, health data analysis, decision making, machine learning, distributed system, machine learning algorithms, prediction

*Teacher, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria
**Teacher , Informatic Engineering, Damascus University, Syria
***Student Master, Information Technology Engineering Department, Information and communication Technology Engineering, Tartous University, Syria

1- مقدمة:

يسعى الناس في مختلف أنحاء العالم إلى مواكبة التطور الرقمي لما فيه من منفعة تصب في مصلحتهم من خلال زيادة الأرباح وتقليل التكلفة المادية والزمنية، فنجد مثلاً في كثير من الشركات والمؤسسات تم تحويل السجلات الورقية إلى سجلات رقمية منظمة ومخزنة على أجهزة الحواسيب.

نشهد اليوم أحجاماً هائلة من البيانات سببها الاستخدام المتزايد للتكنولوجيا بما فيها الشبكة العالمية والهواتف النقالة والحواسيب.... الخ لدرجة أن تلك البيانات أصبحت عصب الشركات والمؤسسات، حيث أننا نعيش في مجتمع معلوماتي وننتقل باتجاه المجتمع المعتمد على المعرفة بالتالي من أجل الحصول على تلك المعرفة بشكل أفضل نحتاج إلى كميات كبيرة من البيانات.

يقصد بالمجتمع المعلوماتي المجتمع الذي تلعب فيه المعلومات الدور الرئيسي في الاقتصاد والثقافة والسياسة[1] . إن الحجم الضخم من البيانات جعل أنظمة المعالجة ضعيفة وغير قادرة على التعامل مع تلك الضخامة في البيانات. ظهر مصطلح البيانات الضخمة في عام 2005 من قبل Roger Mougalas وهو يشير إلى المجموعات الكبيرة جداً من البيانات التي في الأغلب من المستحيل معالجتها وإدارتها باستخدام الأدوات التقليدية الخاصة بالبيانات [2]، بالتالي مع استخدام تقنيات البيانات الضخمة يمكن التعامل مع ذلك الحجم المتزايد. تُستخدم بيانات المستشفيات في وقتنا الحالي لدعم الرعاية الصحية حيث أن البيانات الضخمة توفر هيكلاً أفضل يمكّننا من ربط البيانات بشكل أكثر كفاءة بالتالي يقود إلى نمو وتحسين الرعاية الصحية. يمكن لتحليل البيانات أن يكشف نمطاً يوصلنا إلى علاج مناسب للأفراد حيث تقوم الشبكة الرقمية بجمع وتبادل المعرفة وتقديم المعلومات مما يؤدي إلى اتخاذ قرارات أكثر كفاءة[3] .

بينت الدراسة[8] بأن تطبيق خوارزميات clusteringعلى البيانات الصحية يعطي دقة تتعلق بالخوارزمية المطبقة، حيث بلغت أعلى دقة تم تسجيلها 87.85% المتعلقة بخوارزمية Make Density Based Clusters وبما أن عدد الأصناف كبير يتعلق بعدد الأمراض الموجودة ضمن البيانات بالتالي استُخدمت خوارزميات التصنيف(Classification) .تعد خوارزمية Support Vector Machines (SVM) وخوارزمية Logistic Regression من أهم الخوارزميات المستخدمة في الرعاية الصحية[9] . بينت الدراسة [10] بأن خوارزميتي Gradient Boosting و Random Forests أعطت أفضل النتائج مع بيانات الرعاية الصحية.

2- هدف البحث:

يكمن الهدف من البحث في بناء نظام قادر على زيادة دقة التنبؤ بالحالة الصحية والذي يساهم في تحسين اتخاذ القرارات الطبية المتعلقة بالمريض وذلك عن طريق تحسين استخدام تقنيات البيانات الضخمة في المعالجة المسبقة للبيانات.

3- مواد وطرق البحث:

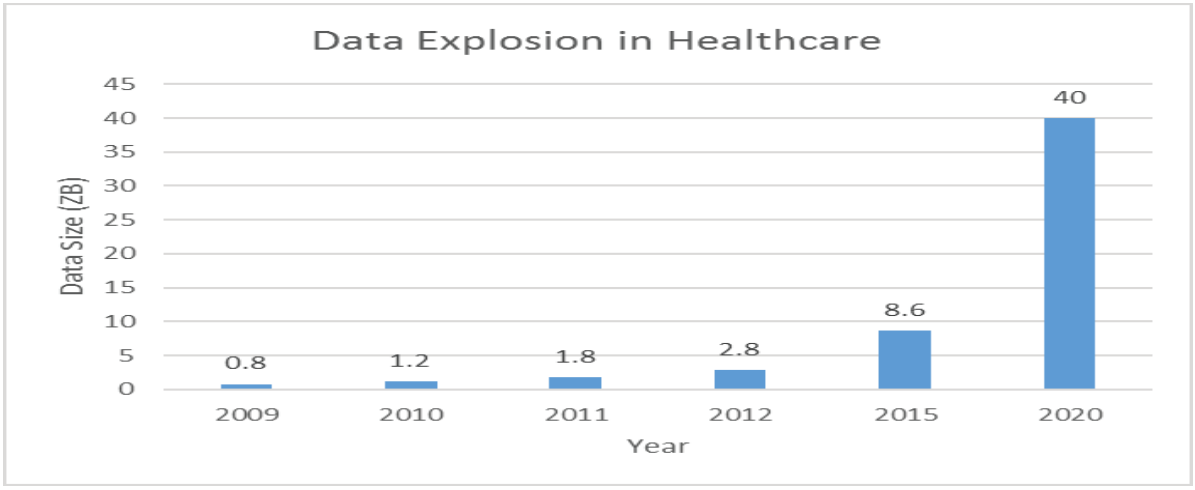
تُعد نقطة البداية في البحث القيام بالدراسة المرجعية عن كيفية استغلال تقنيات البيانات الضخمة في مجال الرعاية الصحية واستخدامها لتحليل البيانات الصحية لتقديم قرار صحي دقيق. استُخدم في هذا البحث إطار العمل Hadoop على نظام التشغيل Centos 6 الخاص بالتعامل مع البيانات الضخمة نظراً لكونه متاح بشكل مجاني على نظام التشغيل آنف الذكر من قبل شركة Cloudera، حيث تم تشغيله على بيئة افتراضية Virtual Box وتم اعتماده لتحليل البيانات الصحية المتاحة ليتم بعدها استخدام خوارزميات تعلم الآلة اعتماداً على لغة البرمجة بايثون (python) لسهولة التعامل معها ولتوفر الكثير من المكاتب التي تسهل عملية البرمجة، والمقارنة بينهما وإظهار النتائج. من الجدير بالذكر القيام بتعديل النموذج البرمجي Map Reduce الموجود في إطار العمل Hadoop باستخدام لغة البرمجة Python الأمر الذي ساهم في تحسين المعالجة المسبقة للبيانات الصحية وزيادة دقة التنبؤ بالحالة الصحية.

3-1- تعريف البيانات الضخمة:

بالرغم من أن مصطلح "البيانات الضخمة" أصبح شائعًا، إلا أنه لا يوجد إجماع عام حول ما يعنيه بالفعل. يشار إلى ذلك المصطلح عادةً بأنه عملية الاستخراج والتحويل والتحميل لمجموعات البيانات الكبيرة. تتميز البيانات الضخمة بثلاث سمات رئيسية هي: الحجم والسرعة والتنوع (عادة ما يرمز لها 3V)[4] .

إن المصادر الرئيسية للبيانات الضخمة هي:

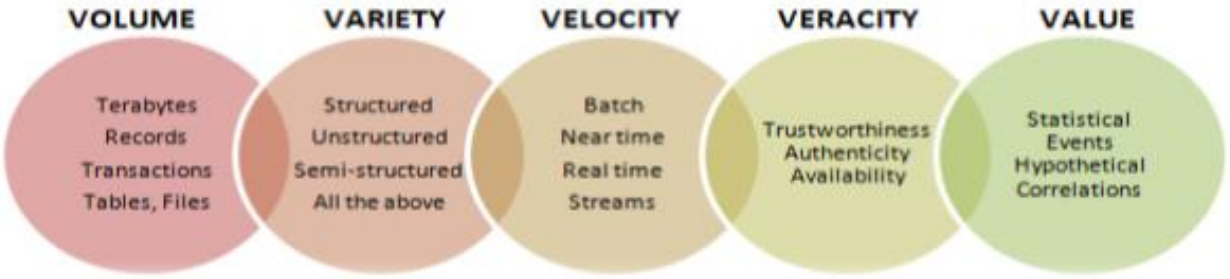
1) الثورة التكنولوجية.
2) انترنيت الأشياء Internet Of Things (IOT).
3) وسائل التواصل الاجتماعي.
4) عوامل أُخرى: يوجد الكثير من العوامل التي أدت إلى ظهور البيانات الضخمة كالأمور المتعلقة بالأسواق المالية والتسوق والتعليم لكن سوف يتم التركيز خلال البحث على البيانات الخاصة بالرعاية الصحية حيث يوضح الشكل (1) الزيادة الكبيرة في حجم البيانات الصحية خلال الأعوام 2009-2020 [5] إذ أن الدراسة[5] تنبأت بالاعتماد على الزيادة الهائلة في الأعوام السابقة بوصول حجم البيانات إلى حوالي 40 ZB في عام 2020.



الشكل (1): التضخم في البيانات الصحية

3-2- خصائص البيانات الضخمة:

تتميز البيانات الضخمة بخمس ميزات أساسية موضحة بالشكل (2)[1].



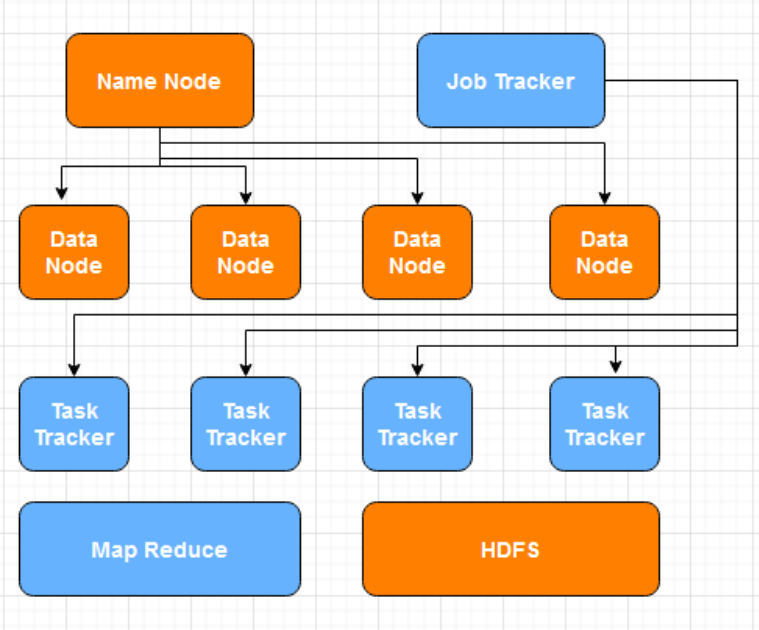
شكل (2): خصائص البيانات الضخمة

- الحجم Volume: يمثل مقدار البيانات التي تم إنشاؤها وتخزينها وتشغيلها داخل النظام مع الحاجة إلى استغلالها لتحقيق أهداف معينة.
- السرعة Velocity: سرعة توليد ووصول البيانات الواجب معالجتها والتي يمكن أن تتطلب معالجة بالزمن الحقيقي.
- التنوع Variety: اختلاف أنواع البيانات سواء كانت مهيكلة، نصف مهيكلة أو غير مهيكلة.
- الثقة Veracity: يدل على مستوى جودة ودقة وتوفر البيانات مع عمليات المصادقة عليها.
- القيمة Value: تشير إلى الفائدة القيّمة التي تم الحصول عليها من البيانات.

3-3- إطار العمل Hadoop:

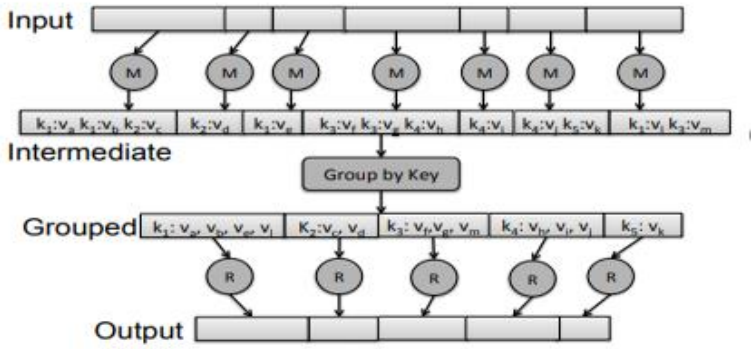
هو عبارة عن إطار عمل مفتوح المصدر مخصص لمعالجة، تخزين وتحليل البيانات. يقوم هذا الإطار بتقسيم البيانات إلى عدة أجزاء وتوزيعها لتتم معالجة وتحليل كل جزء من البيانات الهائلة معاً بدلاً من معالجة البيانات ككتلة واحدة ضخمة.

يتكون إطار العمل من طبقتين أساسيتين هما HDFS (Hadoop Distributed File System) وهو نظام الملفات الموزعة الخاصة بـ Hadoop والنموذج البرمجي Map-Reduce. يقوم HDFS بتخزين البيانات بشكل قطع(blocks) حيث توزع على مجموعة من العقد ونميز نوعين من العقد الأول عقدة الاسم (Name Node) وهي عقدة رئيسة مسؤولة عن تحديد طريقة الوصول إلى البيانات المخزنة. أما الثاني فهو عقدة البيانات (Data Node) وتكون مسؤولة عن تخزين أجزاء البيانات. إن عمليات التحكم بإنشاء، تكرار وحذف البيانات ضمن عقد البيانات تتم بواسطة عقدة الاسم. يوضح الشكل (3) الوظائف الرئيسية لـ Hadoop [6].



الشكل (3): الوظائف الرئيسية لـHadoop

يعد Map-Reduce نموذج برمجي قادر على معالجة البيانات ذات الأحجام الهائلة وعلى التوازي. تحمل المهمة الأولى لهذا النموذج اسم Map وتأخذ شكل مفتاح-قيمة <key-value>. تعمل المهمة الثانية على خرج المهمة الأولى لتعطي النتيجة المرغوبة[6] . يبين الشكل (4) كيفية عمل Map-Reduce [7].



الشكل (4): Map-Reduce

3-4- النماذج المقترحة للعمل:

تكمن الفكرة الرئيسية من البحث في كيفية معالجة البيانات الصحية مهما كان حجمها كبير أم صغير وذلك لعجز الأنظمة التقليدية عن معالجة الكم الكبير من البيانات، بالتالي تم استخدام النموذج البرمجي Map-Reduce وتم تطبيقه على عينة من البيانات من أجل تجريب النظام المقترح حيث تم استخدام بيانات 4921 مريض، بعد ذلك استُخدمت بيانات 43 مريض لأغراض تجريبية ولاختبار النموذج الذي تم بناؤه. إن عدد السجلات الخاصة ببيانات الاختبار هو 43 والسبب في ذلك يعود إلى البيانات المُجمعة والتي تم تحميلها من موقعKaggle وليس هنالك سبب رئيسي لاختيار هذا العدد ولكن تم الالتزام به نظراً لاعتماده كبيانات اختبار من قبل من قام بتجميع تلك البيانات.

3-4-1- البيانات المستخدمة:

اعتُمد في هذا البحث نمط معين من البيانات يجب اعتماده عند استخدام النظام المقترح، إذ أنه يجب أن تكون البيانات مؤلفة من مجموعة من السجلات، كل سجل يمثل مريض سُجلت بياناته سابقاً في المركز الطبي. يُعد كل سجل عن مريض يشكو من مرض معين بالتالي تكون الخصائص التي تصف كل سجل هي عبارة عن الأعراض المرتبطة بمجموعة من الأمراض وأخر عمود عبارة عن اسم المرض الذي يشتكي منه المريض، ويقابل كل عرض من الأعراض قيمة 1 إذا كانت من ضمن أعراض المرض في سجل ما و 0 عكس ذلك والجدول (1) يوضح صيغة البيانات المتبعة.

الجدول (1): صيغة البيانات الصحية المتبعة

Symptoms 1	Symptoms 2	…	Symptoms n	Prognosis
1	0	…	1	disease 1
0	0	…	1	disease 2

يبيّن الجدول (1) كيفية تمثيل الأعراض (Symptoms) والمرض المقابل لها (Prognosis) حيث n عدد الأعراض الأعظمي الموجود ضمن مجموعة البيانات (Dataset) المستخدمة وفي حالة النظام المقترح. حُدد عدد الأعراض المستخدمة بـ 131 أي n=131 وهو عدد الأعراض التي توصّف مجموعة مختلفة من الأمراض. يوضح الجدول (2) مقطع من البيانات المعتمدة في هذا البحث وهي بيانات جاهزة من موقع Kaggle.

الجدول (2): مقطع من البيانات المُستخدمة

skin_peeling	silver_like_dusting	small_dents_in_nails	inflammat	blister	red_sore_around_nose	yellow_crust_ooze	prognosis
1	1	1	1	0	0	0	Psoriasis
0	0	0	0	1	1	1	Impetigo
0	0	0	0	1	1	1	Impetigo
0	0	0	0	1	1	1	Impetigo
0	0	0	0	1	1	1	Impetigo
0	0	0	0	1	1	1	Impetigo
0	0	0	0	0	1	1	Impetigo
0	0	0	0	1	0	1	Impetigo
0	0	0	0	1	1	0	Impetigo
0	0	0	0	1	1	1	Impetigo
0	0	0	0	1	1	1	Impetigo
0	0	0	0	0	0	0	Fungal infection
0	0	0	0	0	0	0	Allergy
0	0	0	0	0	0	0	GERD
0	0	0	0	0	0	0	Chronic cholestasis
0	0	0	0	0	0	0	Drug Reaction
0	0	0	0	0	0	0	Peptic ulcer diseae
0	0	0	0	0	0	0	AIDS
0	0	0	0	0	0	0	Diabetes
0	0	0	0	0	0	0	Gastroenteritis

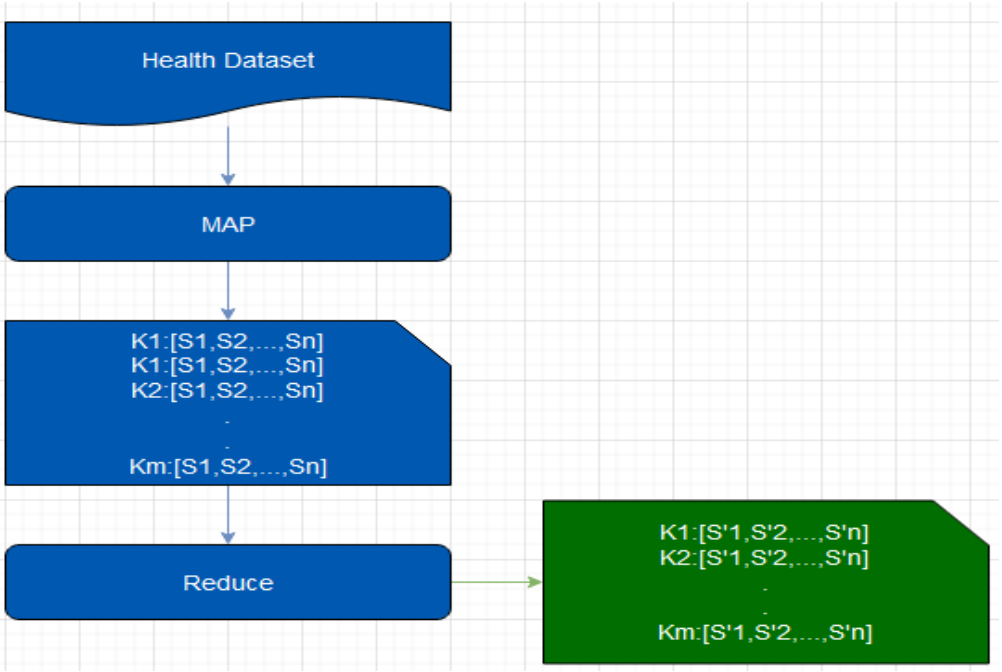
3-4-2- نموذج Map-Reduce المعدل والمطبق على البيانات:

اعتمد النموذج البرمجي Map-Reduce (M-R) على الزوج key-value ، إذ أنه يكون لكل قيمة مفتاح يميزها عن بقية القيم، بالتالي يمكن الحصول على كل قيمة من خلال المفتاح المرتبط بها والذي يكون فريد ومغاير للمفاتيح الباقية.

أُضيف تعديل لهذا المبدأ ليكون هذا النموذج ملائم للبيانات الصحية المتوفرة و مناسب للخصائص المميزة للبيانات والتي هي عبارة عن أعراض الأمراض في حالة الاستخدام المدروسة . تقوم الإضافة بتحويل القيمة المقابلة للمفتاح لتصبح مجموعة من القيم أو بمعنى آخر قائمة قيم (مصفوفة أحادية البعد تحوي مجموعة من القيم كل قيمة تقابل عرض من الأعراض الموجودة ضمن مجموعة البيانات بحيث أنه إذا كانت قيمة منها 0 فإن العَرَض المقابل لهذه القيمة من مجموعة الأعراض غير موجود في المرض المرتبط بهذه المصفوفة والذي هو المفتاح نفسه) بالتالي يصبح الاعتماد على الزوج [list of values] – key.

تتضمن قائمة القيم الأعراض الخاصة بكل سجل أو مريض، بينما يمثل المفتاح اسم المرض وهو اسم فريد لاستحالة تواجد مرضين يحملان اسمين متطابقين.

يوضح الشكل (5) كيفية تطبيق النموذج على البيانات الصحية المتاحة علماً أن تطبيق النموذج يعد مرحلة أولى من مراحل النظام المقترح خلال البحث ليتم استخدام الخرج الذي يأخذ الصيغة الموضحة بالشكل (5) كدخل للمرحلة الثانية.



الشكل (5): تطبيق النموذج M–Rالمعدل على البيانات الصحية

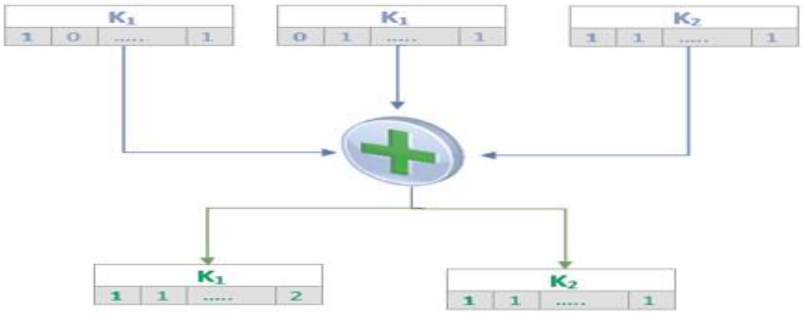
تؤخذ البيانات الصحية لتكون دخلاً للعملية Map، حيث يتم تحويل كل سجل إلى زوج (مفتاح–قائمة من القيم)، وكما ذُكر سابقاً يكون المفتاح اسم المرض والقائمة هي الأعراض، ويكون كل عنصر من عناصر القائمة 1 في حال العرض المقابل لهذا العنصر من أعراض المرض المذكور في المفتاح و0 غير ذلك. يبيّن الشكل (5) خرج العملية Map حيث يمكن لأكثر من سجل أن يدل على نفس المرض لكن ربما بأعراض مختلفة ولذلك فإن تكرار المفتاح أكثر من مرة في هذه العملية وبمصفوفة قيم مختلفة يدلّ على تكرار الحالة المرضية بأعراض مختلفة ضمن مجموعة البيانات. تدل n على عدد الأعراض المتوفرة ضمن البيانات وهي في مجموعة البيانات المستخدمة 131 بينما تدل m على عدد الأمراض الموجودة.

يُرمز للأعراض بالرمز S، وبما أن عدد الأعراض 131، ففي حال كان المرض الأول K_1 متبوعاً بالعرض الأول، الثالث و الخامس من بين كل الأعراض الموجودة فتكون الدلالة على هذا المرض وفق الآتي:

K_1:[1,0,1,0,1,0,0,0,.....,0]، إذ يكون طول القائمة يساوي 131. يجب على المفاتيح الموجودة أن تكون فريدة و غير متكررة، ومن أجل تقليل البيانات الناتجة عن العملية Map، يتم جعل خرجها عبارة عن دخل للعملية Reduce، وينتج لدينا مجموعة من الأمراض المختلفة بعد تجميع النتائج.

يقصد بعملية تجميع النتائج هي تجميع قيم القوائم التابعة لنفس المفتاح لتنتج لدينا قائمة واحدة تابعة لمفتاح واحد وبهذه الطريقة يتم إلغاء تكرار المفاتيح التي تدل على الأمراض كونه تم تجميع المفاتيح المكررة معاً بواسطة العملية Reduce. ينتج عن العملية Reduce قائمة تختلف عن القائمة الناتجة عن العملية Map في القيم التي ضمنها، فقائمة العملية Map تحوي قيماً تساوي 0 أو 1 تدل على تواجد الأعراض أم لا بالنسبة لمرض محدد، بينما قائمة العملية Reduce تحوي قيماً تدل على عدد الحالات التي يكون فيها عرض من الأعراض يتبع مرض معين أي بمعنى أن كل قيمة من القيم الناتجة عن Reduce تدل على عدد مرات ظهور العرض المقابل لهذه القيمة بالنسبة للمرض (أو المفتاح) المرتبط بتلك القائمة. يوضح الشكل (6) كيفية تجميع القوائم للحصول على الخرج النهائي للمرحلة الأولى إذ نلاحظ أن عدد القوائم المرتبطة بالمرض K_1 هو 2 بينما عدد القوائم المرتبط بالمرض K_2 هو 1 بالتالي فأن [1,1,...,2] = [0,1,...,1] + [1,0,...,1] وذلك بجمع كل قيمة من القائمة الأولى مع ما يقابلها من القائمة الثانية.

من الجدير بالذكر أن هذه المرحلة لها أهمية كبيرة في استيعاب البيانات مهما كان حجمها وتعتبر مغزى البحث وخاصة كيفية تحويل البيانات إلى قوائم تابعة لمفاتيح معينة بينما هي في الأصل قيم تتبع مفاتيح، ليتم تجميع القوائم ككل بعد ذلك.



الشكل (6) تجميع القوائم

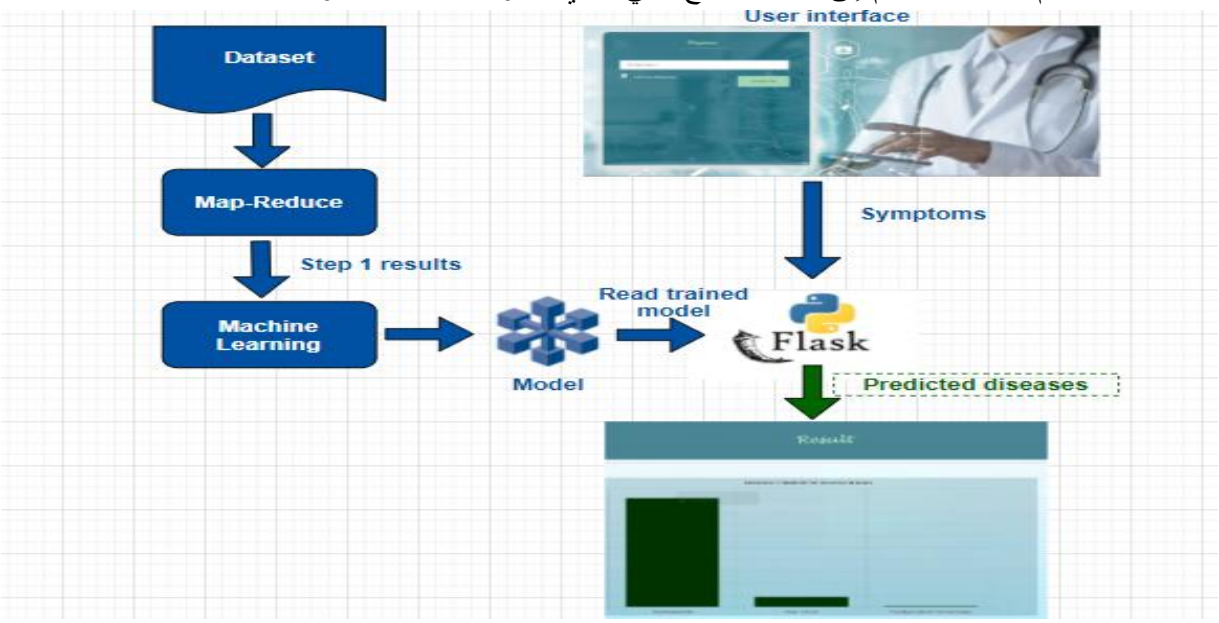
3-4-3- تطبيق خوارزميات تعلم الآلة على نتائج المرحلة الأولى:

طُبقت الخوارزميات الأربعة المومأ إليها سابقاً في المقدمة لمعرفة الخوارزمية التي تعطي الدقة الأفضل بينهم وذلك على نتائج المرحلة الأولى باستخدام مكتبة sklearn المتاحة في لغة البرمجة python، حيث تم القيام بالمعالجة المسبقة للبيانات الوسطية قبل تطبيق الخوارزميات عليها، حيث تم وضع كل القيم ضمن مجال[1-0] لتسهيل عمليات المعالجة.

طُور برنامج خاص بمرحلة المعالجة المسبقة التي تحوي Map-Reduce بعد تعديله ليستوعب كامل الأعراض الموجودة من خلال تحويل القيمة المقابلة للمفتاح وهي قيمة واحدة تدل على كل الأعراض لتصبح قائمة من القيم كل قيمة منها تدل على عرض واحد فقط وبالتالي تسهيل عملية التجميع وزيادة دقة التنبؤ بالحالة المرضية، وبعد ذلك تم تطوير برنامج يطبق خوارزميات تعلم الآلة على خرج المرحلة الأولى وفي النهاية تم تطوير واجهة المستخدم وربطها مع عمليات المعالجة والتنبؤ. كل هذه الأمور تم بناؤها برمجياً في هذا البحث المقترح من البداية وباستخدام لغة البرمجة Python. يوضح الشكل (7) بنية النظام المقترح كاملاً إذ أنه

استُخدم API(Application Programming Interface) اعتماداً على Flask وهي مكتبة ضمن لغة Python تتيح بناء API، والفائدة منه تحقيق الربط بين عمليات المعالجة و واجهة المستخدم.

يقصد بعمليات المعالجة تطبيق خوارزمية تعلم الآلة التي حققت نتائج أفضل على خرج مرحلة M-R ليتم بعدها حفظ النموذج الناتج ليستخدم من قبل API للتنبؤ بالمرض اعتماداً على الأعراض المدخلة. يستطيع المستخدم من خلال الواجهة والتي هي عبارة عن صفحة ويب تتصل بالAPI بواسطة لغة البرمجة PHP و التي بدورها ترسل الأعراض المدخلة إلى API ليتم معالجتها و تحديد المرض المقابل لها، استُخدم Wamp Server من أجل تشغيل صفحة PHP. يتم توجيه المستخدم إلى صفحة النتائج والتي تحوي الأمراض المقابلة للأعراض المدخلة.



الشكل (7) خطوات تصنيف الأمراض

4- النتائج والمناقشة:

اعتُمد في هذا البحث على أدوات القياس Accuracy, Precision, Recall المعرفة وفق المعادلات الآتية:

Tp: True positive

Tn: True negative

Fp: False positive

Fn: False positive

$$Accuracy = \frac{T_p+T_n}{T_p+T_n+F_p+F_n} \quad (1)$$

$$\text{Precision} = \frac{T_p}{T_p+F_p} \quad (2)$$

$$\text{Recall} = \frac{T_p}{T_p+T_n} \quad (3)$$

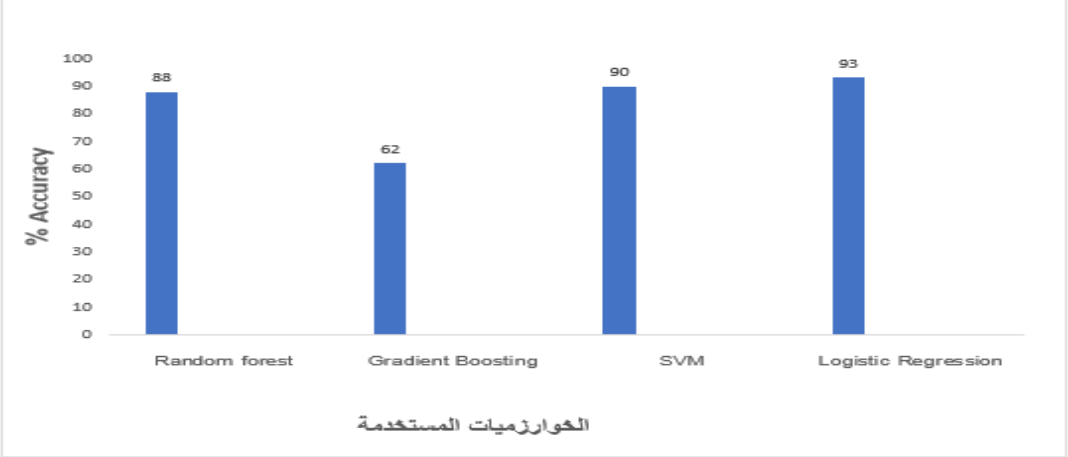
تدل Tp على عدد الحالات التي يتم التنبؤ بها بصنف إيجابي وبالفعل يكون الصنف الصحيح إيجابي بينما تدل Tn إلى عدد الحالات التي يتم التنبؤ بها بصنف سلبي وهي بالحقيقة صنف سلبي.

نميز أيضاً بين Fp و Fn حيث أن Fp هي عدد الحالات التي يتم التنبؤ بها بصنف إيجابي وهي بالحقيقة صنف سلبي بينما Fn هي عدد الحالات التي يتم التنبؤ بها أنها صنف سلبي وهي بالحقيقة صنف إيجابي

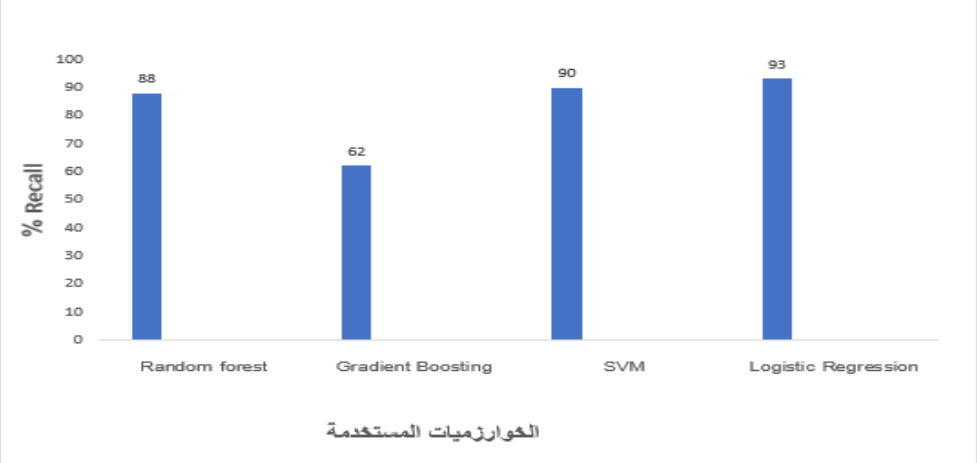
يبين الجدول (3) نتائج مقاييس الآداء التي تم الحصول عليها. يوضح الشكل (8) المقارنة على أساس accuracy بينما يوضح الشكل (9) المقارنة اعتماداً على Recall و أخيراً يوضح الشكل (10) المقارنة اعتماداً على Precision.

الجدول (3): نتائج مقاييس الآداء

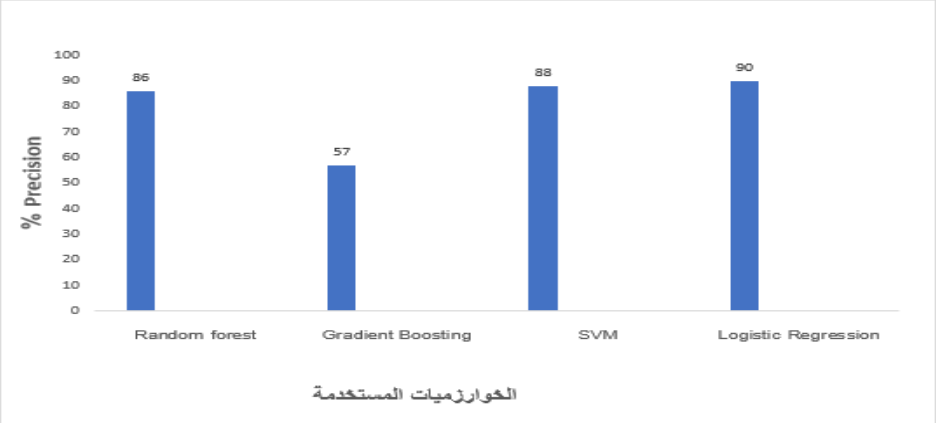
الخوارزمية	مقاييس الآداء		
	Accuracy	Recall	Precision
Random forest	88%	88%	86%
Gradient Boosting	62%	62%	57%
SVM	90%	90%	88%
Logistic Regression	93%	93%	90%



الشكل (8): المقارنة بين الخوارزميات من حيث accuracy

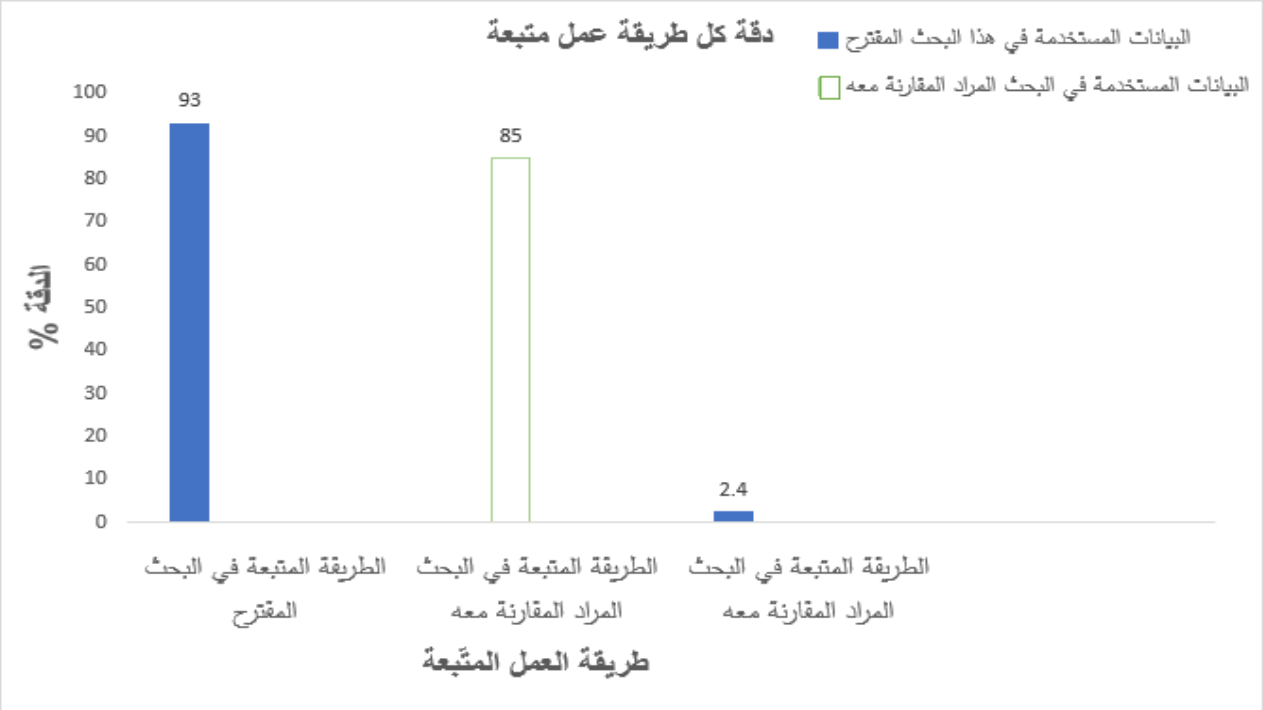


الشكل (9): المقارنة بين الخوارزميات من حيث Recall



الشكل (10): المقارنة بين الخوارزميات من حيث Precision

طُبق مبدأ العمل الذي تم اعتماده في البحث المراد المقارنة معه[6] وكانت الدقة الناتجة بالقيمة التقريبية 2.4%، هذا يعني أن مبدأ العمل الموجود في البحث[6] غير قابل للتطبيق على البيانات المعتمدة في هذا البحث المقترح. إن تطبيق مبدأ العمل[6] على البيانات المعتمدة في [6] أعطى دقة 85%. في هذا البحث تم اتباع صيغة مختلفة للبيانات والتركيز على التشخيص فقط وكانت الدقة الناتجة 93%. يوضح الشكل (11) النتائج النهائية للدقة.



الشكل (11): المقارنة النهائية للدقة

5- الاستنتاجات والتوصيات:

يمكن تلخيص أهمية البحث في النقاط التالية:

- تحليل البيانات الصحية المخزنة والاستفادة منها في تحليل الأعراض مهما كان حجمها.
- إمكانية تحديد المرض اعتماداً على الأعراض المدخلة بعد معالجة البيانات الصحية.
- تحسين المعالجة المسبقة للبيانات الصحية من خلال التعديل على Map Reduce وتطبيقه على البيانات الصحية المتاحة.
- تحسين التنبؤ بالحالة الصحية حيث ازدادت دقة التنبؤ بعد تطبيق التعديل على Map Reduce.

بينت النتائج التطبيقية أنه في مثل هذه الحالة وبعد تطبيق خوارزميات تعلم الآلة آنفة الذكر أن الخوارزمية التي حققت دقة أعلى هي Logistic Regression بالتالي استخدامها في مثل هكذا حالات أفضل.

يمكن تلخيص التوصيات في النقاط التالية:

- تحليل البيانات اعتماداً على Spark من أجل زيادة سرعة المعالجة والقدرة على التعامل مع بيانات تدفقية في الزمن الحقيقي.
- ربط النظام المقترح مع بيانات صحية مخزنة ضمن المستشفيات ليتم معالجتها مباشرة واعتماداً على تلك البيانات يتم اتخاذ القرار الطبي بعد تشخيص الحالة من قبل النظام.
- تطوير النظام ليكون قادراً على تحديد الدواء المناسب للتشخيص الناتج.

6- المراجع:

[1]. Riahi, Y., & Riahi, S. (2018). Big data and big data analytics: Concepts, types and technologies. *International Journal of Research and Engineering*, *5*(9), 524-528.

[2]. Sangeetha, S., & Sreeja, A. (2015). No science no humans, no new technologies no changes "big data a great revolution". *International Journal of Computer Science and Information Technologies*, 6(4), 3269-3274.

[3]. Das, N., Das, L., Rautaray, S. S., & Pandey, M. (2018). Big data analytics for medical applications. *International Journal of Modern Education and Computer Science*, *11*(2), 35.

[4]. Buyya, R., Calheiros, R. N., & Dastjerdi, A. V. (Eds.). (2016). *Big data: principles and paradigms*. Morgan Kaufmann.

[5]. Liang, H., Luo, M., Wang, R., Lu, P., Lu, W., & Long, L. (2018). Big data in health care: Applications and challenges. *Data and Information Management*, *2*(3), 175-197.

[6]. Pasupathi, C., & Kalavakonda, V. (2016, February). Evidence Based health care system using Big Data for disease diagnosis. In *2016 2nd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB)* (pp. 743-747). IEEE.

[7]. Ramírez-Gallego, S., Fernández, A., García, S., Chen, M., & Herrera, F. (2018). Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce. *Information Fusion*, *42*, 51-61.

[8]. Santhi, P., & Bhaskaran, V. M. (2010). Performance of clustering algorithms in healthcare database. *International Journal for Advances in Computer Science*, *2*(1), 26-31.

[9]. 9January.2022. https://analyticsindiamag.com/top-6-ai-algorithms-in-healthcare .

[10]. Kedia, A., Narsaria, M., Goswami, S., & Taparia, J. (2017). Empirical Study to Evaluate the Performance of Classification Algorithms on Healthcare Datasets. *World*, *5*(1), 1-11.