جامعة الفيوم
كلية الحاسبات والمعلومات
قسم علوم الحاسب

تحليل البيانات المتدفقة باستخدام التعلم الآلى على الأنظمة الكبيرة

مقدمة من:

فوزية رمضان سيد حسان

مدرس مساعد، قسم علوم الحاسب،

كلية الحاسبات والمعلومات، جامعة الفيوم

رسالة مقدمة الى كلية الحاسبات والمعلومات، جامعة الفيوم

كجزء من متطلبات الحصول على درجة دكتوراة الفلسفة في الحاسبات و المعلومات

أ.د/ عبدالمجيد أمين على

قسم علوم الحاسب
كلية الحاسبات والمعلومات
جامعة المنيا

د/ مسعود إسماعيل مسعود شاهين

قسمعلومالحاسب
كلية الحاسبات والمعلومات
جامعة الفيوم

تخصص علومالحاسب

إشراف:

كلية الحاسبات و المعلومات
جامعة الفيوم

2020

ملخص الرسالة

مع الزيادة المستمرة فى تدفق البيانات وسرعتها الهائلة فى الأنظمة الكبيرة ومع الاحتياجات المتزايدة لتحليل البيانات وخصوصا فى مجال الرعاية الصحية. لقد أصبح تحليل تلك البيانات والتبؤ بالأمراض المزمنه فى الوقت الحالي هو أمر ضرورى للغاية ، وذلك لأن من الصعب استيعاب ومعالجة وتحليل مثل هذه البيانات الضخمة لأتخاذ إجراء فى الوقت الحالي فى حالات الطوارىء باستخدام الطرق التقليدية. . لذلك ، فإن العمل في هذه الأطروحة يتعلق بكيفية بناء نظام يمكنه تحليل ومعالجة تدفق البيانات فى الوقت الحالى باستخدام بيانات وسائل التواصل الأجتماعي المستندة إلى الصحة أو أجهزة الاستشعار الطبية التي يمكن ارتداؤها والتنبؤ بالحالة الحالية لصحة المريض. لذلك فأن العمل فى هذه الاطروحة هدفه تحسين تحليل البيانات المتدفقة فى الوقت الحالى، وقد تم ذلك من خلال مساهمتين مختلفتين .

المساهمة الأولى وهي تقديم نظام للتنبؤ لمرض السكري عند وصول البيانات من وسائل التواصل الاجتماعي (تويتر). وقد تم تطوير هذا النظام لأستقبال البيانات المتدفقة المستندة إلى الصحة وذلك للتبؤ بالحالة الصحية للمريص يهدف النظام المقترح إلى إيجاد نموذج التعلم الآلي الأكثر دقة التي لديها أعلى دقة من التنبؤ بمرض السكري. وقد حددت النتائج التجريبية أن نموذج Random Forest حقق أعلي دقة بين النماذج الآخرى عند 84.11 %. للتنبؤ عبر الإنترنت من خلال وسائل التواصل الاجتماعي ، لقد تم تنفيذ نظام مقترح للتعامل مع بيانات تويتر المتدفقة الخاصة بصحة المرضى. فلقد تم دمج تدفق Kafka و Spark في الواجهة الخلفية للنظام المقترح. ومن ثم، يتم استخدام Random Forest للتنبؤ بالحالة الصحية الحالية للمريض في الوقت الحالي.

المساهمة الثانية وهي عبارة عن نظام Online Prediction System الذى قد تم تطويره فى هذة الأطروحة لحل مشكلة تحليل البيانات المتدفقة فى الوقت الحالي من خلال تطبيق ال Streaming machine learning على البيانات الصحية المتدفقة التى يتم تطبيقها على Spark Streaming من خلال Kafka. ولقد تم أجراء النتائج التجريبية على مجموعات مختلفة للبيانات الطيبية التاريخيه وتحويلها لبيانات متدفقة وايضا بيانات أجهزة الاستشعار الطبية القابلة للارتداء. وقد أثبتت النتائج التجريبية أن Online Prediction System قادر على التعلم وتحديث النموذج وفقا لوصول البيانات الجديدة وحجم النافذة.