University of Science and Technology

College for Graduate Studies & Academic Advancement

Classification of Student's Examinations Data Set Using Naive Bayes and K-NN Algorithm

Thesis submitted in Partial Fulfillment of the Requirements for the Degree of Master of Information System

Student Name: Jada Salah Mustafa Sid ahmed

Supervisor Name: Dr. Atif Ali Mohamed Ali

January 2014

Abstract

Data is growing rapidly and continuously, the human beings are used in the different technologies to adequate in the society. Every day the human beings are using the vast data and these big data are in the different fields .It may be in the form of documents, graphical formats, videos and records.

This big data has a problems, the most important problem that focused in this research is how to classification these data, where the applied data mining approach to solve this problem.

Due to the importance of extracting knowledge/information from the large data repositories, data mining has become an essential component in various fields of human life.

Data mining have various applications in human life, this research focused about banking, telecommunication, education, Healthcare and manufacturing engineering.

This research apply one of the approaches of data mining which is classification, it's data mining techniques that assigns items in a collection to target categories or classes. Classification in this research applied on the data of student Degrees for examinations office at UST batch 2008-2009 and 2009-2010 for first year by used one of data mining tools named RapidMiner, used for analyzing big data generated by high-throughput instruments though Naïve Bays and K_NN classification algorithm .

The results that were obtained through the K_NN algorithm to classification data of student's Degrees for examinations office batch 2008-2009 and 2009-2010 are (Excellent),(Very Good),(Good),(Pad). The results that were obtained through the Naïve Bayes algorithm is the percentage of students of batch 2008-2009 that obtained (Excellent) is greater than percentage of students of batch 2008-2009, the percentage of students of batch 2009-2010 that obtained (Very

Good) is greater than percentage of students of batch 2008-2009, the percentage of students of batch 2008-2009 that obtained (Good) is greater than the percentage of students of batch 2009-2010, the percentage of students of batch 2009-2010 that obtained (Pad) is greater than percentage of students of batch 2008-2009. Through these results the College Board can identify the performance of students and making appropriate decisions for the benefit of students of the next batches.

المستخلص

تنمو البيانات بصورة سريعة ومستمرة ، يستخدم البشر هذه البيانات في مختلف التكنولوجيات في المجتمع. في كل يوم تستخدم بيانات ضخمة و هذه البيانات في مختلف المجالات ، قد يكون في شكل وثائق ، أشكال رسومية ، الفيديو أو صور وغيرها.

عندما نتعامل مع البيانات الكبيرة هنالك بعض المشاكل التي تواجهنا أهم هذه المشاكل التي نتعامل معها هي كيفية تصنيف البيانات الضخمة ويتم حل هذه المشكلة عن طريق تطبيق مفهوم تنقيب البيانات .

نظراً لأهمية استخراج المعرفة والمعلومات من مستودعات البيانات الكبيرة أصبح تنقيب البيانات عنصراً أساسي في مختلف مجالات الحياة البشرية، وتنقيب البيانات لديه مجموعة من التطبيقات في الحياة البشرية ، هذا البحث وضح تطبيقات تنقيب البيانات في السوق ، البنك ، والطب ، الاتصالات ، التعليم ، الرعاية الصحية وهندسة التصنيع.

هذا البحث طبق واحد من مفاهيم استخراج البيانات هو التصنيف وهو تقنية تنقيب البيانات الذي يضع العناصر في مجموعات لإستهداف الفئات . التصنيف في هذا البحث طبق على بيانات تقدير الطلاب لمكتب الامتحانات في جامعة العلوم والتقانة ،قسم علوم الحاسوب دفعة 2008-2009 ودفعة 2009-2010 باستخدام واحدة من أدوات استخراج البيانات تسمى RapidMiner وهي يستخدم لتحليل البيانات التي يتم إنشاءها بواسطة أدوات الإنتاجية العالية ,عبر خوارزمية Bayes Naïve،وخوارزمية K_NN.

النتائج التي تحصلنا عليها خلال تطبيق خوارزمية K_NN لبيانات نتيجة دفعة 2008-2009 و2009-2010 هي (امتياز) و(جيد جداً) و(جيد) و(ضعيف). والنتائج التي تحصلنا عليها خلال تطبيق خوارزمية Nive Bayes أن نسبة طلاب دفعة 2008-2009 اللذين حصلوا على تقدير (إمتياز) أكبر من نسبة طلاب دفعة 2009-2010 ، ونسبة طلاب دفعة 2009-2010 اللذين حصلوا على تقدير (جيد جداً) أكبر من نسبة طلاب دفعة 2008- 2009، نسبة طلاب دفعة 2008-2009 اللذين حصلوا على تقدير (جيد) اكبر من نسبة طلاب دفعة 2009- 2010 و نسبة طلاب دفعة 2009- 2010 اللذين حصلوا على تقدير (ضعيف) أكبر من نسبة طلاب دفعة 2008- 2009 . خلال هذه النتائج يستطيع مجلس الكلية تحديد أداء الطلاب و إتخاز القرارات المناسبة لمصلحة طلاب الدفع القادمة.

Introduction:

Big data is a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.

Data is being produced at an ever increasing rate. This growth in data production is being driven by individuals and their increased use of media, organizations, the switch from analogue to digital technologies and the proliferation of internet connected devices and systems.

These large amounts of data content challenges in how to classification, so we have to deal with data mining (sometimes called data or knowledge discovery), is process of analyzing data from different perspectives and summarizing it into useful information , it is apply in several application like bank, marketing, medicine, education, telecommunication, healthcare, manufacturing engineering .

Research Problem:

The amount of data in our world is increases every day quickly, this data contain some of the problems one of this problems that focused in this research is classification.

Examination office at UST department of computer sciences has a problems of how to classification data of Student's Examination Degrees that stored in excel files.

This problem cannot be solved by traditional methods, so we use one of data mining tools named RapidMiner.

Research Objective:

The objectives of this research are:

- Apply the classification on big data in Examination Office at UST Department of Computer Science by using RapidMiner software.
- Explain the applications of Data Mining
- Discuss the classification, classification algorithm of big data.
- Discuss the techniques and method of big data.
-

Research Methodology:

This research used experimental methodology , it's a best scientific research methods because this approach depends mainly on scientific experiment, which provides an opportunity to know the facts of the process of enacting laws through these experiences.

So we can say that more research methods are important for the researcher because this approach helped to develop researcher through observation and experimentation , and access to the correct results and learn peaceful ways to deal with the phenomena, so this research using the experimental methodology to experience the classification techniques on the RapidMiner tool for data of Student's Examination Degrees at UST.

Research structure:

Chapter two explain a background for Big Data through the definition, importance, applications, technology, Algorithm, problems and future challenges that related to it. Chapter three discuss data mining concept through task, types, life cycle, technique , stage of data mining application ,data mining applications.

Chapters four discuss classification,types of classification in Data Mining System and classification algorithms. Chapters five description research tool (RapidMiner) and description of case study and discussion of Results. Chapters sex explain conclusion of this research and the recommendations of future work.