Prediction Model for Risk Factors of Childhood leukemia Based on Data Mining Classification algorithms

Samir Emad Labib, Christina Albert Rayed
Computer and Information System Department, Sadat Academy for Management Science, Cairo, Egypt
Samiremad2000simo@yahoo.com, sams.christina.albert@gmail.com

Abstract

Data Mining aims at discovering knowledge out of data and illustrate it in a form that is easily comprehensible to humans. One of the helpful researches in Egypt is to discover the most Risk Factors of Childhood Leukemia, especially Acute Lymphoblastic Leukemia, which is the most common type of cancer in children.

The Researcher makes use of a Data mining tools for compare some selected machine learning algorithms on data set this system can discover the most Risk Factors of Childhood Leukemia by collecting Knowledge associated with diseases from historical records of the patients, Hospitals and the Biometric scan in Data warehouse then the Data is ready for mining through Risk Factors of Childhood Leukemia classification in terms of impact strength for the chosen sample. The algorithms were compared based on a few parameters including mainly accuracy and training time. The algorithms were applied to the Childhood Leukemia dataset of different fields. for a chosen sample which was divided into eight groups will be processed by R Studio Environment Which Risk Factors for cancer are categorized into a specific number of risks such as genetics, smoking, age, weight, pollution, etc. They are then treated again to reduce these risks into four factors (demographic, social, lifestyle and environmental). In this paper, the researcher's findings of this experiment show that the decision tree algorithm is the most suitable and had the highest accuracy.

Keywords: *Data mining, Data Warehouse, Machine Learning Algorithms, Decision tree, Naïve Bayes and Random forest, DHI (digital health information), DHT (digital health technology), Childhood Leukemia).*

1. Introduction

the data mining method used for diagnosing based on previous data and information has been improving over recent years. The data mining method used currently particularly for disease diagnosis uses various classification techniques which include decision tree and rule classifier. Data mining techniques can not only conclude accurately but also helpful in visualizing patterns inside the dataset itself. And there is no single classifier preferable over the rest, for instance, the classification accuracy depends on the classification method, gene selection method, and datasets.

data mining plays an important role in predicting diseases. Recent advances in microarray technology offer the ability to measure expression levels of thousands of genes simultaneously. Analysis of such data Assists us in identifying different clinical outcomes that are caused by the expression of a few predictive genes. The feature extraction and classification are carried out with a combination of the high accuracy of ensemble-based algorithms, and the comprehensibility of a single decision tree. These allow deriving exact rules by describing gene expression differences among significantly expressed genes in leukemia. It is evident from our results that it is possible to achieve better accuracy in classifying impact factors in childhood Leukemia without reducing the level of comprehensibility. Some of the most important and popular data mining techniques are association rules, classification, clustering, prediction, and sequential patterns[1].

Leukemia disease is a type of cancer that affects the blood and the bone marrow it is characterized by an abnormal proliferation of blood cells. Acute Myelogenous Leukemia (AML), Acute Lymphoblastic Leukemia (ALL), Chronic Myeloid Leukemia (CML) and Chronic Lymphocytic Leukemia (CLL) are categorized as leukemia diseases[2]. In general, leukemia is grouped by how fast it gets worse and what kind of white blood cells it affects [3].

In the era of Data mining driven by the development of technology, different types of Data, including clinical records, imaging, gene information, or even from different areas, can be combined effectively and promptly through the advanced informatics platforms. These platforms can also perform computation and data analysis. This progress provides an unprecedented opportunity for medical research such as outcome prediction of cancer. This research We will briefly review some of the previous Data mining analytics model's Which can be utilized to access the Proposed Data mining analytics Model for prediction of Risk Factors of childhood leukemia[4]. The model can discover the most Risk Factors of Childhood Leukemia Where the researcher initially relied on the identification of Risk Factors causing leukemia for 10000 patients selected in the sample were limited to 18 factors which were divided into four groups Demographic factors, Environmental factors, Live style factors and Social factors that will be processed by Machine learning algorithms To identify the most Risk Factors of Childhood Leukemia in all groups and each group Separately.

The dataset was used in the experiment: A Childhood Leukemia dataset. With the RStudio tool, different algorithms are deployed like Naive Bayes, Decision Tree, and Random forest. Comparisons performed to evaluate various classification algorithms on the same dataset. And comparisons were done to compare machine learning algorithms on the dataset. This study is important because it helps decide which one of the Algorithm classifications is the best fit in this big data. Performance evaluation depends on many standards, but the main important ones are accuracy and training time. Accordingly, this helps in deciding which algorithm is the most suitable for the high accuracy and less consumed time during building the model.

The rest of this paper is organized into five sections: initially, we focus on the introduction of paper in the first section. The second section discusses the objective and model. In the third section, discuss the methodology of the proposed model. Results and dialogues of the experimental works are assumed in the fourth section, whereas the fifth section introduces the conclusion and recommended future work.

2. The Objective of Data mining Analytics Model for Predicting of Risk Factors Childhood Leukemia

The main objective of this paper is to improve the quality of prediction processes in the datasets. This comparative study seeks out a fair judgment and knows which algorithm is best to fit in with the introduced dataset. The study moved on certain phases described in the framework of the experimentation as shown in Fig.1. The framework includes many phases: data collection, data preprocessing and selection, transformation phase, selection of big data tool, selection of programming language, and selection of data mining algorithms depending on the dataset multi-classification.

This research uses data mining techniques for the analysis and evaluation of classification algorithms of impact factors of the leukemia disease dataset. Through R studio environment data mining techniques, the researcher can generate a predictive model for the classification of impact factors of leukemia disease, evaluate accuracies, and performance of several techniques.

Predictive Modeling in health care could help in making better choices, consistent with their long-term goals (cheaper, better health; tailored choices). The Progressive (and other auto insurers) experience is instructive, Progressive (through the use of predictive modeling) identified patients who were classified as high-health risk (assigned to a high-risk pool) and take the necessary measures in advance to prevent the expected diseases.

The experimental design explains the reasons why a phenomenon occurs by making experiments where independent variables are manipulated, extraneous variables are controlled and therefore conclusions are being made which results in actions that the decision-maker should practice. Optimization as a technique suggests balancing the level of a certain variable related to other variables, thus identifying the ideal level of it a recommendation for the decision-maker.

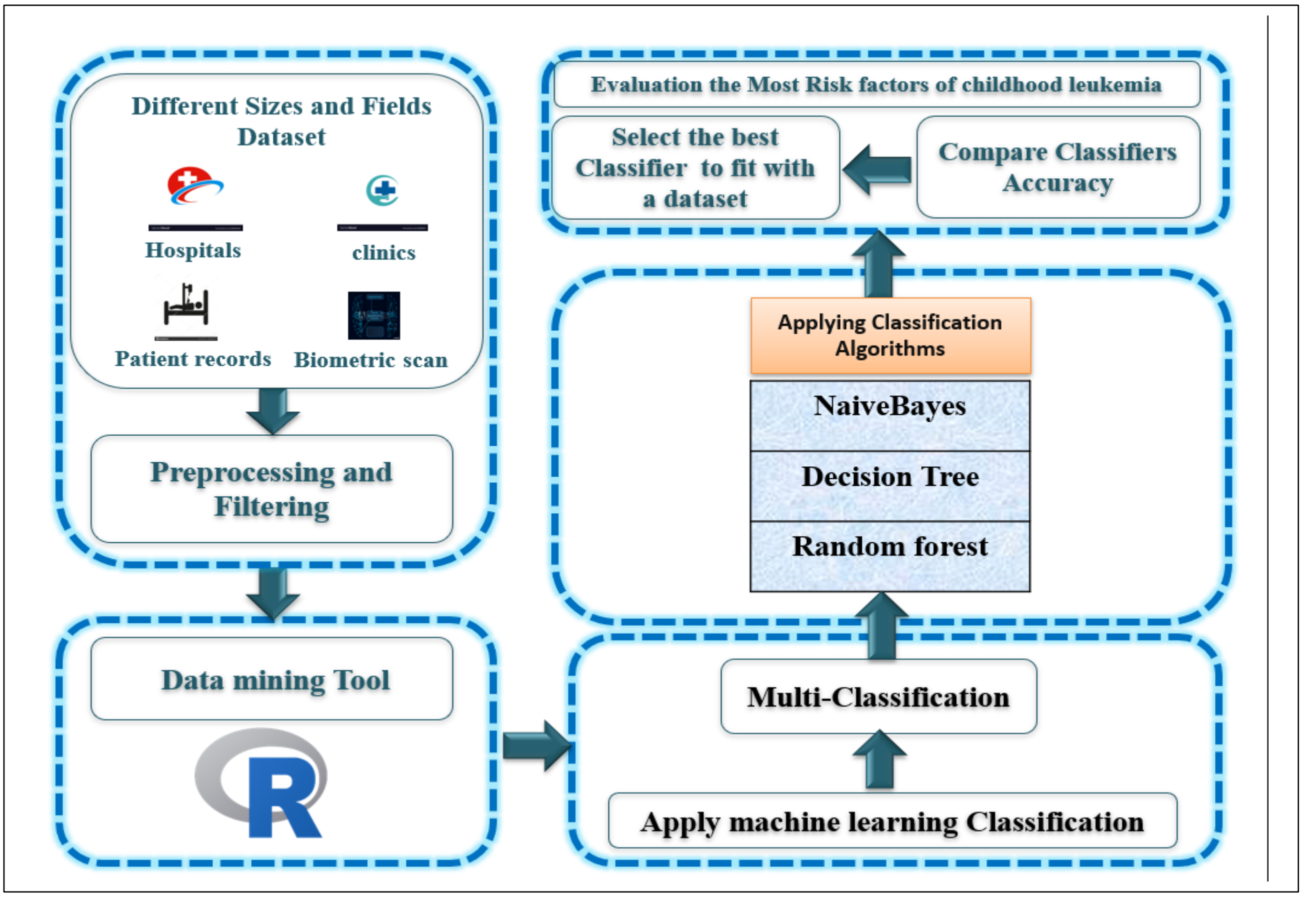
Different Sizes and Fields Dataset
Hospitals
clinics
Patient records
Biometric scan
Preprocessing and Filtering
Data mining Tool
Evaluation the Most Risk factors of childhood leukemia
Select the best Classifier to fit with a dataset
Compare Classifiers Accuracy
Applying Classification Algorithms
NaiveBayes
Decision Tree
Random forest
Multi-Classification
Apply machine learning Classification

Fig. 1. The conceptual Model.

2.1 The Proposed Model

While processing the four phases, this study starts with data collection, data preprocessing and filtering to clean, integrate, and transform the data. Then, the required fields for data mining are selected by programming language. The data was transformed into a certain file format, which is acceptable by the data mining tools. The data mining tools were different data mining algorithms based on the dataset multiclassification which are tested by R programming language.

2.2 Data Collection and Selection Phase

In this paper, there is a Childhood leukemia dataset; the data set is a medium data set obtained from extremely reputational Hospital data set activity which sells only via the hospital. The dataset is collected by a gathering ordering log file for three months. The dataset contains 10000 instances and 25 attributes, three classes, and 18 features. For all the datasets, feature extraction, and feature selection should be done properly. The feature details as shown in table 1.

Table 1. Feature Details

Column Name	Description	Type
Gender	Male/ Female	Plain Text
Age	less than 1 year / From 1 to less than 2 years / From 2 to less than 3 years / From 3 to less than 4 years.	Number
Color skin	Dark skin / Light skin.	Plain Text
Child's weight	From 5 to less than 8 K.G / From 8 to less than 12 K.G / From 12 to less than 14 K.G / From 14 to less than 16 K.G.	Plain Text
Child lived area	urban / rural / Semi-rural / Semi- urban	Plain Text
Type of birth	normal / caesarean /	Date & Time
Q1	Did the mother smoke before pregnancy	Plain Text
Q2	Did the mother smoke during pregnancy	Plain Text
Q3	Did the mother smoke after Birth	Plain Text
Q4	Did the father smoke before mom pregnancy	Plain Text
Q5	Did the father smoke during mom pregnancy	Plain Text
Q6	Did the father smoke after Birth	Plain Text
Q7	Is the child exposed to negative smoking	Plain Text
Q8	Does the child eat unhealthy food	Plain Text
Q9	Did mothers eat unhealthy food or drinks during pregnancy	Plain Text
Q10	Was the fetus physically active during pregnancy	Plain Text
Q11	Was the child physically active during the first four years of his life	Plain Text
Q12	Are there family members suffering from leukemia	Plain Text
Q13	Did the mother undergo regular medical examinations before and during pregnancy	Plain Text
Q14	Have parents suffered from certain diseases before pregnancy?	Plain Text
Q15	Is the Fetus exposed to negative smoke	Plain Text
Q16	Have mothers or babies been exposed to ionizing radiation during pregnancy or breastfeeding	Plain Text
Q17	Has the mother or fetus been exposed to insecticides or other contaminants during pregnancy	Plain Text
Q18	Is the child exposed to pesticides or other contaminants during the first four years of his life	Plain Text

2.3 Data Preprocessing and Transformation Phase

In this step, the dataset goes through more than stages, such as data cleaning, data integration, and data transformation. The gathered data was saved as excel spreadsheets or text documents. The cleaning process is required to analyze the data based on selected classifier algorithms, in which data with missing values are eliminated, inconsistent data is corrected, outliers are identified, and duplicate data is removed. The data was exemplified by numbers and stored in the form of a CSV or txt file so it can be introduced to the data mining tool[5].

2.4 Selection of Data Mining Tool

Analyzing Big Data can be very cumbersome and challenging. There is no particular software that can be used for the analysis. Different enterprises use different tools for Big Data analysis. However, the tool to use depends on the type of data one needs to analyze. The choice of tools can also affect the quality of your data which can have a significant impact on your analysis. In this paper some tools that can be used to analyze both structured and unstructured data (Big Data)[6].

R. is an open-source language that uses data modeling, handling, statistics, prediction, time analysis, and data visualization. The R language uses your computer's RAM, the RAM of your machine is very large, and the larger data you can work for R. We have more than 4000 different packages created by various scholars as per the requirement. The latest version of R will be R 3.0.2, initially, R is not used as a large data analysis language due to its memory limit problems. Gradually, some libraries such as R, Rodbc, rmr2, and Rhdfs were available to handle large data[7].

The R language is well established as the language for doing statistics, data analysis, data-mining algorithm development, Healthcare, credit risk scoring, CRM and all [8] manner of predictive analytics. However, given the deluge of data that must be processed and analyzed today, many organizations have been reticent about deploying R beyond research into health care Applications. The dataset was split into two parts: the first part is a training set which was 80% used to train the model of the real dataset, the second part was a 20%, it was used as a cross-validation to train the model, and it was used as a testing set to train the model.

3. Machine Learning Classification Algorithms

Classification divides data samples into target classes. The classification technique predicts the target class for each data point. The data classification approach is a supervised learning approach having known class categories[9]. The data set is partitioned as a training and testing datasets. Using the training dataset, we trained the classifier. The correctness of the classifier could be tested using the test dataset. Classification is one of the most widely used methods of Data Mining in Healthcare organizations [10]. However, the accuracy of such methods differs according to the classification algorithm used. Identifying the best classification algorithm among all available is a challenging task. The present research proposes a comprehensive analysis of different classification algorithms and the performance of evaluating by applying the leukemia micro-array data set. Hu et al. [10]. used different classification methods such as decision tree, SVM and ensemble approach for analyzing microarray data[10].

Data mining techniques can be segregated by their different model functions, representation, preference criterion, and algorithms. The main function of the model lies in its classification and a brief overview of several classification algorithms that have been used in this study.

A) A decision tree is one of the most popular and efficient techniques in data mining. This technique has been established and well explored by many researchers. However, some decision tree algorithms may produce a large structure of tree size and it is difficult to understand Furthermore, misclassification of data often occurs in the learning process. Therefore, a decision tree algorithm that can produce a simple tree structure with high accuracy in terms of classification rate is a need to work with a huge volume of data. Pruning methods have been introduced to reduce the complexity of tree structure without decrease the accuracy of classification[11].

B) Naïve Bayes is a simple multiclass classification algorithm with the assumption of individuality between every pair of features. It can be used to train professionally, within a single pass of the training data and use it for prediction[11].

C) Random forest is a collection of decision trees. It's one of the most successful machine learning models for classification and regression. It runs well on large datasets, but it is comparatively slower than other algorithms. It can successfully estimate missing values, therefore, it is suitable for handling datasets with a large number of missing values[11].

3.1 Performance Factors Evaluation

The comparison of the different machine learning algorithms mentioned in Section 3 is based on the following measured parameters:

Accuracy: It is the percentage of the correctly classified instances which are the total number of correct predictions.

Precision: It is the element of the identified items that are correct and used to measure how well the proposed algorithm matches the ground truth. It's likewise known as Positive Predictive Value (PPV).

Recall: It is another measure used to compute how the proposed algorithm matches the ground truth. Recall, or sensitivity or consistently True Positive Rate (TPR), which is a measure of the number of true positives relative to the sum of the true positives and the false negatives. It is the element of items that were correctly detected among all the items that should have been detected.

F-measure: It is sensitivity, an overall measure of how well we have been able to classify the ground truth foregrounds and backgrounds.

Training Time: As an accomplished machine learning algorithm are measured. Time taken to build the model is called training time. This varies on the implementations of the algorithms.

From the previously mentioned measured parameters, the researcher will be comparing the accuracies provided by all the algorithms on a dataset. Here, the focus is mainly on comparing major parameters like accuracy and training time to decide which machine learning algorithm is better suited for a selected type of data.

4. Experimental Work

Objectives of the experiment are to compare the performances of the machine learning algorithms when deployed types datasets in terms of the training time, accuracy, true positive rate (TPR), precision, recall and F-measure, this evaluation helps to decide which machine learning algorithm is best suited for a selected of data. The results section has been presented for four experimental results. In all experimental results, the dataset is analyzed using different classifiers. They have been deployed big data mining tool R environment. These experiments were carried out on a Lenovo Laptop using Windows 10 operating system having the following specifications, the number of core processors 1 with Intel® Core ™ i7-5500U CPU, 2.40 GHz Processor, and 16 GB RAM.

4.1 Classification Results Using Decision Tree by Method C5.0

In the Hospital dataset, as shown in table 2, a comparison between 18 Factor, decision tree, to evaluate the performance of the algorithms by several multi-classification evaluation metrics. The table presents the values of accuracy, true positive rate (TPR), precision, recall, F-measure, and time is taken to build model per second, which are classified as to the following dataset. Fig.3 shows that the decision tree algorithm had the highest value of accuracy at 0.9913169% and the decision tree algorithm had the highest value of training time at 8 (Seconds) as shown in Fig.4.

Table 2. Comparison of several different Classes using the Decision Tree Algorithm by C5.0 Method

Classes	TP	FP	TN	FN	Precision	Recall	F-Measure	Accuracy	Time (s)
Q1	0.5145	0.003	0.121	0.015	0.993243	0.9716714	0.9823389	0.9717125	5
Q2	0.564	0	0.121	0.006	1	0.9894737	0.994709	0.9913169	3
Q3	0.568	0.026	0.103	0	0.956229	1	0.9776248	0.962724	4
Q4	0.0135	0.003	0.126	0	0.794117	1	0.8852459	0.9756098	4
Q5	0.0315	0.001	0.090	0	0.969230	1	0.984375	0.9918699	3
Q6	0.0075	0	0.100	0.002	0.758533	1	0.8626887	0.9783496	4
Q7	0.5425	0	0	0.13	0.870869	1	0.9352426	0.8707586	6
Q8	0.0165	0	0	0.016	1	0.9280369	0.976428	0.9894742	3
Q9	0.0780	0	0	0.012	1	0.5	0.6666667	0.5012289	7
Q10	0.7428	0	0.051	0.124	1	0.8564014	0.9226468	0.8644529	8
Q11	0.7643	0	0	0.165	1	0.8223897	0.9025399	0.8223897	5
Q12	0.032	0	0.220	0.001	1	0.9064033	0.9344709	0.9100842	4
Q13	0.7292	0.002	0.018	0.092	0.997262	0.8873325	0.9390912	0.8877005	6

Table 2. Comparison of several different Classes using the Decision Tree Algorithm by C5.0 Method

Classes	TP	FP	TN	FN	Precision	Recall	F-Measure	Accuracy	Time (s)
Q14	0.0475	0	0.004	0.004	1	0.9134615	0.9547739	0.9196429	5
Q15	0.0370	0.001	0.039	0	0.973684	1	0.9866667	0.9870968	4
Q16	0.0380	0.002	0.027	0	0.940123	1	0.9691378	0.9643448	3
Q17	0.011	0	0.220	0.004	1	0.8534088	0.8467203	0.8869314	8
Q18	0.0605	0	0.021	0.006	0.958809	0.9097744	0.9527559	0.9314286	5

Table 2 presents the values of precision, recall, F-Measure, TP, FP, TN, FN and Time taken to build model per second classified as to the following Dataset. The table shows that the decision tree had the highest value of Precision at 0.991334, Decision Tree had the highest value of Recall at 0.9688454 respectively, and Decision Tree had the highest value of F-Measure at 0.9879123, and the highest value of Accuracy at 0.9913169.

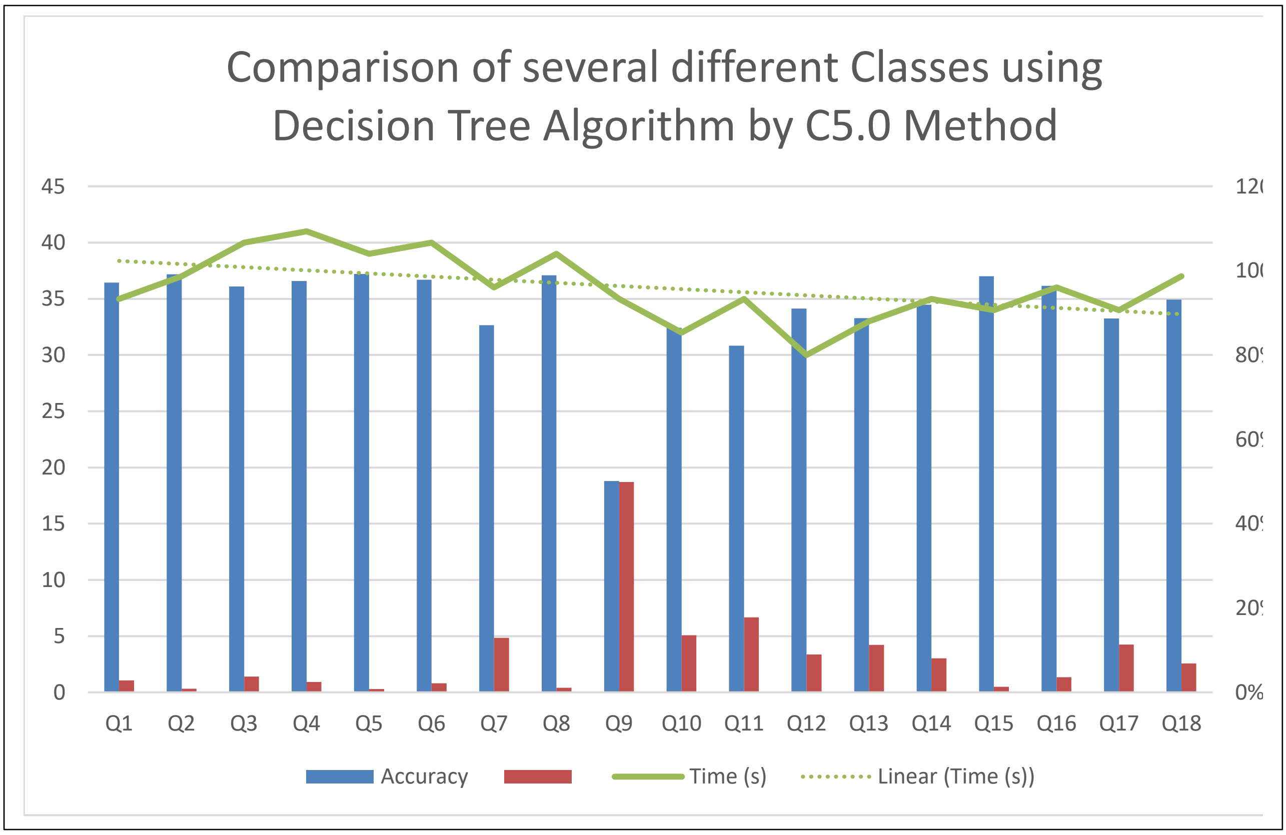

Fig. 2. Comparison between accuracy percentage in different Question's.

As it is shown in Figure 2, the accuracy results for the Decision Tree algorithm tested on our dataset using the R environment mining tool, Decision Tree had the highest value of Questions 1,2,3,4,5,6,8,12,14,16 and 18 which had the highest accuracy more than 90%.

4.2 Classification Results Using Naïve Bayes algorithm

In the Hospital dataset, as shown in table 3, a comparison between 18 Factor, Naive Bayes, to evaluate the performance of the algorithms by several multi-classification evaluation metrics. The table presents the values of accuracy, true positive rate (TPR), precision, recall, F-measure, and time is taken to build model per second, which are classified as to the following dataset. Fig.3 shows that the Naïve Bayes algorithm had the highest value of accuracy at 0.947 % and the decision tree algorithm had the highest value of training time at 4 s as shown in Fig.3.

Table 3 Comparison of several different Classes using the Naïve Bayes Algorithm.

Classes	TP	FP	TN	FN	Precision	Recall	F-Measure	Accuracy	Time (s)
Q1	1103	0	191	39	1	0.9658494	0.9826281	0.9235	3
Q2	1158	0	188	45	1	0.9625935	0.9809403	0.947	2
Q3	1143	24	167	35	0.979434	0.9702886	0.9748401	0.8855	2
Q4	98	0	253	84	1	0.5384615	0.7	0.7194	3
Q5	90	12	143	39	0.882352	0.6976744	0.7792208	0.687	3
Q6	89	0	43	199	1	0.3090278	0.4721485	0.337	4
Q7	48	0	22	112	1	0.3	0.4615385	0.231	3
Q8	22	159	276	11	0.121547	0.6666667	0.2056075	0.161	3
Q9	119	87	119	21	0.577669	0.85	0.6878613	0.5713	4
Q10	1002	115	59	205	0.894642	0.8301574	0.8611947	0.5845	4
Q11	335	0	0	50	1	0.8701299	0.9305556	0.237	3
Q12	15	0	0	0	1	1	1	0.941	4
Q13	1022	183	48	141	0.848132	0.8787618	0.8631757	0.585	2
Q14	74	20	34	34	0.787234	0.6851852	0.7326733	0.6705	3
Q15	40	12	42	15	0.769230	0.7272727	0.7476636	0.7225	4
Q16	23	142	133	0	0.139393	1	0.2446809	0.283	3
Q17	39	143	195	0	0.214285	1	0.3529412	0.3485	3
Q18	69	48	88	35	0.589743	0.6634615	0.6244344	0.648	4

Table 3 presents the values of precision, recall, F-Measure, TP, FP, TN, FN and Time taken to build model per second classified as to the following Dataset. The table shows that Naïve Bayes Algorithm had the highest value of Precision at 0.995566, Naïve Bayes had the highest value of Recall at 0.9783125 respectively, and Naïve Bayes had the highest value of F-Measure at 0.9825443, Naïve Bayes had the highest value of Accuracy at 0.947.

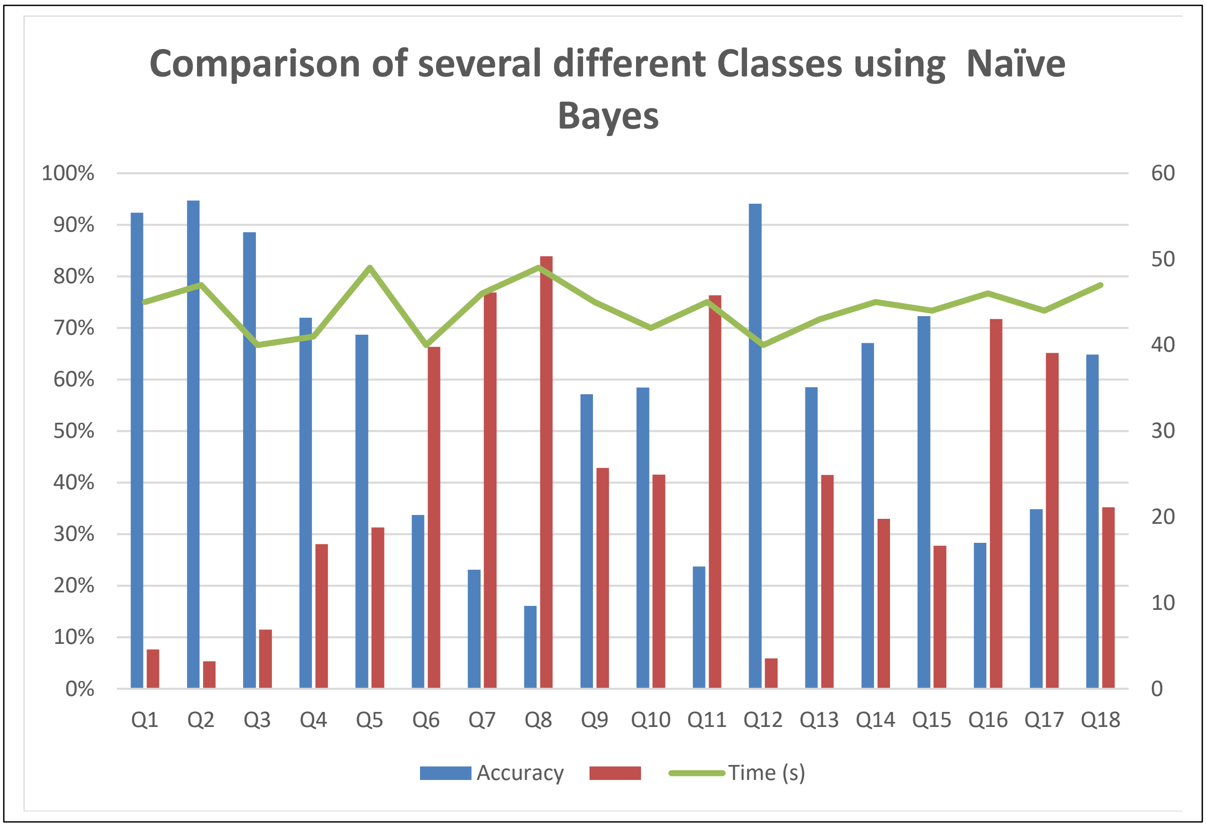

Fig. 3. Comparative study of accuracy percentage between different Questions

As it is shown in Figure 3, the accuracy results for the Naïve Bayes algorithm tested on our dataset using the R environment mining tool, Naïve Bayes had the highest value of Questions 1,2,3,12 and 18 which had the highest accuracy more than 90%.

4.3 Classification Results Using Random forest algorithm

In this study Results, show a Random forest classification algorithms result, First, to measure the model performance, The dataset was split into two parts, a training set which was 80% used to train the model of the actual dataset and 20% was used as a testing set to train the model. To evaluate the performance of the algorithms by several multiclass classification evaluations to measure its accuracy and time taken to build model per second which are defined by equations through a confusion matrix.

Table 4. Comparison of several different Classes using the Random Forest Algorithm.

Classes	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q12	Q13	Q14	Q15	Q16	Q17	Q18
Accuracy	**0.612**	**0.664**	**0.6825**	**0.7944**	**0.741**	**0.809**	**0.8955**	**0.776**	**0.7224**	**0.759**	**0.7745**	**0.7231**	**0.746**	**0.746**	**0.7745**	**0.8895**	**0.8335**	**0.797**	**0.7825**
Time (s)	**6.659**	**4.214**	**3.549**	**2.434**	**4.2**	**2.003**	**2.001**	**3.765**	**3.765**	**4.003**	**4.342**	**3.987**	**5.657**	**3.987**	**3.789**	**2.001**	**2.908**	**3.126**	**3.543**

Table 4 presents the values of the Accuracy build model per second classified as to the following Dataset. The table shows that random forest Algorithm had the highest value of Accuracy at Q6 = 0.809, Q7 = 0.8955, Q15 = 0.8895 and Q16 = 0.8335.

4.4 Classification Results and Performance Evaluation

In this paper Results, show a comparison between three different classification algorithms, which are Naive Bayes, Decision Tree, and Random forest. this paper used a machine-learning algorithm to help achieve high accuracy in different classification algorithms. The results show in Table 5 below.

Table 5. Comparison of several different classifiers

Classes	Accuracy		
	Decision tree	Naive Bayes	Random forest
Q1	**0.9717125**	**0.9235**	0.612
Q2	**0.9913169**	**0.947**	0.664
Q3	**0.962724**	0.8855	0.6825
Q4	**0.9756098**	0.7194	0.7944
Q5	**0.9918699**	0.687	0.741
Q6	**0.9783496**	0.337	0.809
Q7	0.8707586	0.231	**0.8955**
Q8	**0.9894742**	0.161	0.776
Q9	0.5012289	0.5713	0.7224
Q10	0.8644529	0.5845	0.759
Q11	0.8223897	0.237	0.7745
Q12	**0.9100842**	**0.941**	0.7231
Q13	0.8877005	0.585	0.746
Q14	**0.9196429**	0.6705	0.7745
Q15	**0.9870968**	0.7225	**0.8895**
Q16	**0.9643448**	0.283	0.8335
Q17	0.8869314	0.3485	0.797
Q18	**0.9314286**	0.648	0.7825

Table 5 shows the comparison study made between decision tree, naive Bayes and random forest algorithm to help achieve high accuracy in different Questions in the dataset. The result shows that the best one of them was the Decision Tree algorithm, which had the highest accuracy of up to 90%. So, the researcher chose the decision tree algorithm in this case.

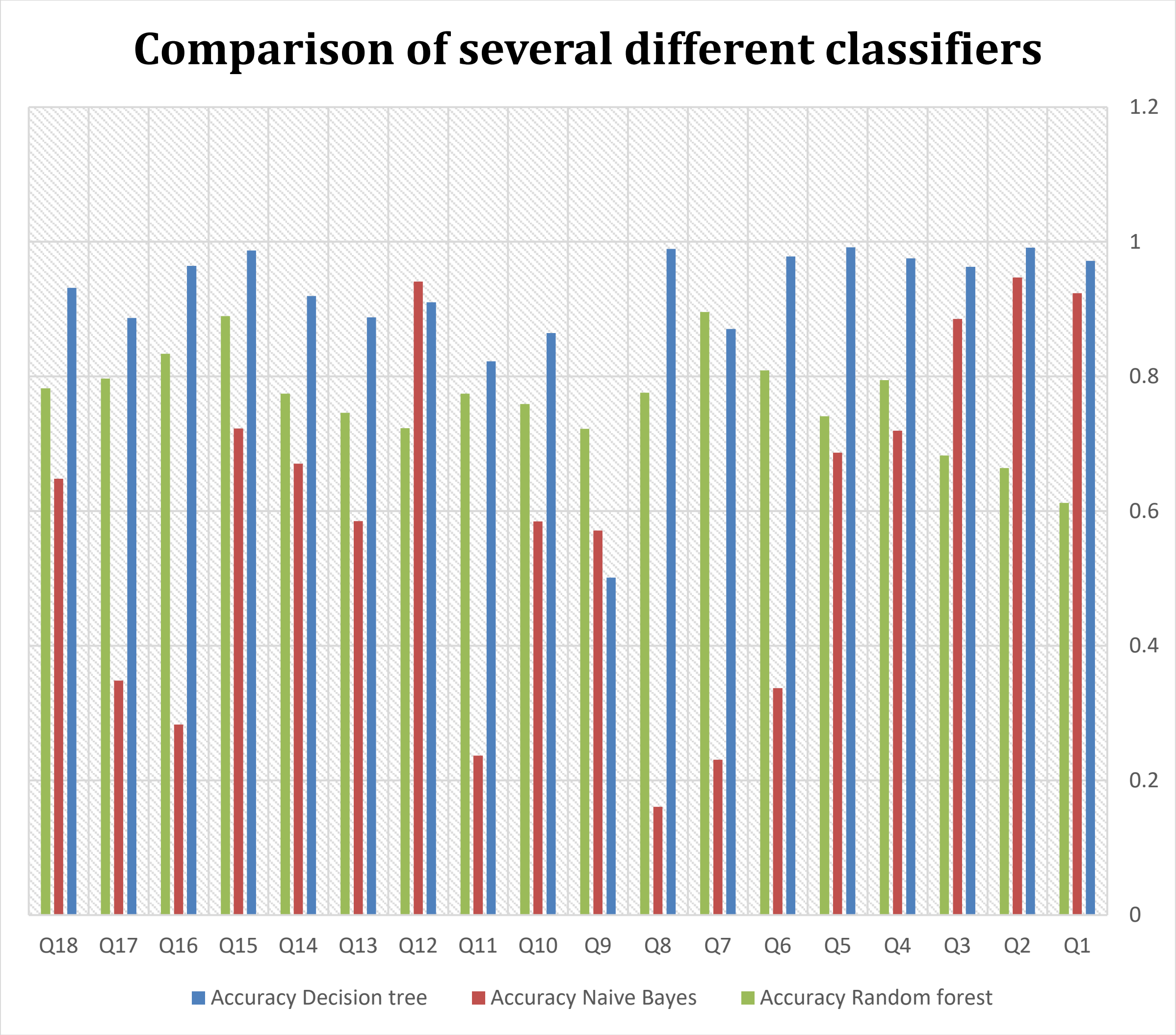

Fig.4. Comparative study of accuracy percentage between different algorithms.

In figure 4 Finally Show, the results of efficiency analysis of each data classification technique showing the best classification algorithm were decision tree algorithm, which had the highest accuracy of 0.9913169 % in 3 seconds followed by a random forest at 0.8955% by taking 2.001 seconds and naïve Bayes had the highest accuracy of 0.947 %. Hence, this result can help Most risk factor datasets by selecting the optimal classification algorithm.

5. Conclusion & Future Work

This paper investigated using different machine learning algorithms on datasets by data mining tool "R languages ". This work focuses on finding the right algorithm for the classification of data that works better on diverse data sets. Algorithms were applied to Hospital datasets for Predicting most risk factors of Childhood Leukemia in Egypt. The finding of this experiment shows that the Decision Tree algorithm is the most suitable. The experiment was implanted locally on specific datasets and needs in future work to be implemented on the cluster and deploying on big data from a different source.

Technological advances in Data mining should be exploited to predict Childhood Leukemia to preserve the health of our children as well as reduce social and economic costs. Reported that 40% of cancers can be avoided and 40% can be easily treated in case of early diagnosis and prediction Consequently. A model has been prepared for the prediction of Childhood Leukemia by applying to a sample that has been examined and preparing its Data for working blood through collecting of all Data on each case. The blood of children and therefore files were made to eight groups according to the division of the study sample and through the use of the application.

These variants and leukemia factors were reduced to four, children Through the statistical analysis of these Data can be reached to the most factors and variables most dangerous to the child who can be infected with leukemia.

Thus, this paper concludes that the child with white skin, who lives in a rural area, was born by cesarean section, in the age group less than a year, weighs between (5-8) kg and at the same time has a family member suffering from leukemia is the most vulnerable child and most likely Exposed to childhood leukemia

References

[1]. B. Rajeswari and A. Rajini, "Survey On Data Mining Algorithms to Predict Leukemia Types," *International Journal for Research in Science Engineering & Technology,* vol. 2, pp. 42-46, 2015.

[2]. S. Dash, B. Patra, and B. Tripathy, "A hybrid data mining technique for improving the classification accuracy of microarray data set," *International Journal of Information Engineering and Electronic Business,* vol. 4, p. 43, 2012.

[3]. M. Madhukar, S. Agaian, and A. T. Chronopoulos, "Deterministic model for acute myelogenous leukemia classification," in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2012, pp. 433-438.

[4]. D. Marcus, J. Harwell, T. Olsen, M. Hodge, M. Glasser, F. Prior, M. Jenkinson, T. Laumann, S. Curtiss, and D. Van Essen, "Informatics and data mining tools and strategies for the human connectome project," *Frontiers in neuroinformatics,* vol. 5, p. 4, 2011.

[5]. M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. Mccauley, M. Franklin, S. Shenker, and I. Stoica, "Fast and interactive analytics over Hadoop data with Spark," *Usenix Login,* vol. 37, pp. 45-51, 2012.

[6]. B. Ratner, *Statistical and Machine-Learning Data Mining:: Techniques for Better Predictive Modeling and Analysis of Big Data*: CRC Press, 2017.

[7]. V. Prajapati, *Big data analytics with R and Hadoop*: Packt Publishing Ltd, 2013.

[8]. R. Nisbet, J. Elder, and G. Miner, *Handbook of statistical analysis and data mining applications*: Academic Press, 2009.

[9]. D. Tomar and S. Agarwal, "A survey on Data Mining approaches for Healthcare," *International Journal of Bio-Science and Bio-Technology,* vol. 5, pp. 241-266, 2013.

[10]. H. Hu, J. Li, A. Plank, H. Wang, and G. Daggard, "A comparative study of classification methods for microarray data analysis," in *Proceedings of the 5th Australasian Data Mining Conference (AusDM 2006): Data Mining and Analytics 2006*, 2006, pp. 33-37.

[11].O. Maimon and L. Rokach, "Data mining and knowledge discovery handbook," 2005.