Predicting Blood Donation Pattern

Computer Science
Georgia State University
Bhavya Induri, Raghavi Ravi, Divya Bhuvanapalli
binduri1@student.gsu.edu ,rravi1@student.gsu.edu ,dbhuvanapalli1@student.gsu.edu

***Abstract*—Blood donation is an essential activity to acquire blood as a raw material into the blood supply chain. It must be managed effectively together with other processes in blood management. In this research, the pattern of blood donors' behaviors based on factors influencing blood donation decision is conducted using online questionnaire. These factors, i.e., altruistic values, knowledge in blood donation, perceived risks, attitudes towards blood donation, and intention to donate blood, are analyzed to find out the possibilities for individuals to become blood donors. The surveyed data are used for machine learning techniques of Artificial Intelligence,Neural Networks ,Random Forest and many other algorithms to predict the possibility of a person donating blood. On the dataset many algorithms and find out accuracy for each algorithm and find out the best algorithm for prediction.**

I. INTRODUCTION

Blood Center is responsible for blood and blood components management. Its main activities are acquiring blood supply from donors, blood screening and processing, blood storage and blood distribution. The Blood Center performs blood collection from donors and distributed blood to various hospitals nationwide via Regional Blood Centers.

Blood collection from voluntary donors is a primary activity to acquire blood as raw material into blood supply chain. Generally, Blood Centers have a standard procedure to screen blood effectively by incorporating collection costs, and amounts of time and units, to obtain blood properly into consideration. The goal is to obtain blood that is safe for use in the subsequent activities. However, the main problem in blood collection is an inability to obtain sufficient blood to meet the patients' needs or a difficulty to balance blood demand and supply in the blood supply chain.

In general, there is no plan to manage potential blood donors. With a good preparation, by classifying potential donors in such a way that intentions of donors to donate blood in the future can be determined, would greatly facilitate the blood acquisition to perform better.Moreover, it will be beneficial for managing blood requisition on emergency requests. Hence, information from blood donor prediction about possibility of donation would suggest the behavior of blood donors whether or not they will donate again in the future. This information would greatly facilitate and improve blood donors recruiting process of the Blood Centers.

II. SYSTEM OVERVIEW

A. Dataset Information

The study of the factors that influence the behavior of blood donors has been conducted extensively due to the significant impact of blood shortages to the survival of patients. We consider five factors leading to the study of blood donor behavior. These factors are 1). Recency - months since last donation 2) Frequency - total number of donation 3)Monetary - total blood donated in c.c. 4) Time - months since first donation 5) Class - a binary variable representing whether a person donated blood or not.For prediction dataset is taken from https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center

Data Set Description:

Data Set Characteristics:	Multivariate	Number of Instances:	748	Area:	Business
Attribute Characteristics:	Real	Number of Attributes:	5	Date Donated	2008-10-03
Associated Tasks:	Classification	Missing Values?	N/A	Number of Web Hits:	162356

B. Data Cleaning

For Data Pre-processing and finding accuracy 'R' language is used.Followed the these steps to make data ready. 1)Converting .data file into .csv file 2) Checking the attributes and predicted values for null values 3) Scaling the datasets.All the attributes are scaled to make their values between 0 and 1. 4) Looking up for categorical attributes and making them categorical if they are not categorical We will make the class variable categorical in our case. 5) Finding out correlations : Look for variables that will contribute to our classification result by observing their correlation with the class variable as well as themselves

III. METHODOLOGY

A.Data Splitting

Now, decide on the number of folds for cross validation. (We have used 5-fold cross validation in our analysis) .And then, split the data into training dataset and test dataset according to n folds.(train_data – 80% and test_data – 20%) For every n fold, we have trained the classifier and tested the model..Calculated the accuracy for each classifier by applying following Machine Learning Models:

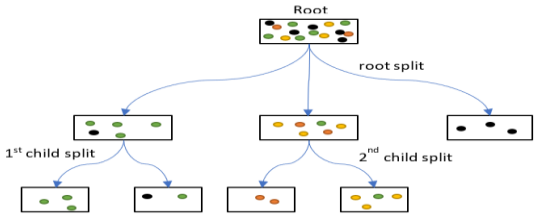
IV.ALGORITHMS

A..Decision Tree :

Decision tree is the most powerful and popular tool for classification and prediction. Decision Tree algorithm belongs to the family of supervised learning algorithm A Decision tree is a flowchart like tree structure,In decision trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with record's attribute. On the basis of comparison,

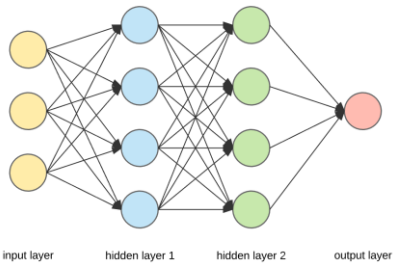
we follow the branch corresponding to that value and jump to the next node.

We continue comparing our record's attribute values with other **internal nodes** of the tree until we reach **a leaf node** with predicted class value. As we know how the modeled decision tree can be used to predict the target class or the value. Now let's understanding how we can create the decision tree model.



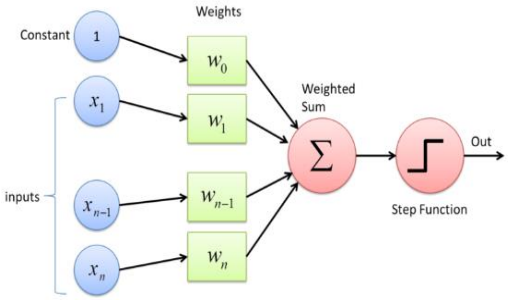
A. *Neural Networks :*

Neural networks represent a brain metaphor for information processing. These models are an exact replica of how the brain actually functions. Neural networks have been shown to be very promising systems in many forecasting applications and business classification applications due to their ability to "learn" from the data , their nonparametric nature(i.e., no rigid assumptions), and their ability to generalize



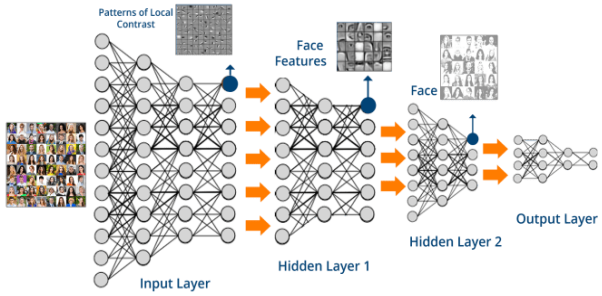
C.Perceptron

Perceptron is a single layer neural network. Perceptron is usually used to classify the data into two parts. Therefore, it is also known as a Linear Binary Classifier. The perceptron model is a more general computational model than McCulloch-Pitts neuron. It takes an input, aggregates it (weighted sum) and returns 1 only if the aggregated sum is more than some threshold else returns 0. Rewriting the threshold as shown above and making it a constant input with a variable weight



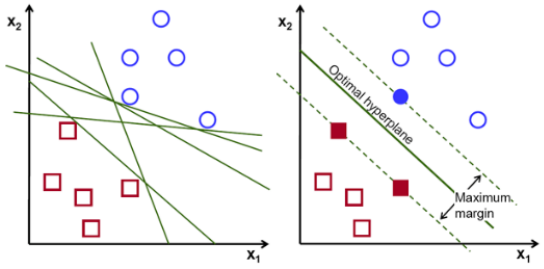
D.Deep Learning

Deep learning is a machine learning technique that teaches computers to do what comes naturally to humans Deep Learning is basically a neural network implementation with large number of layers. Deep learning differs from traditional machine learning techniques in that they can automatically learn representations from data such as images, video or text, without introducing hand-coded rules or human domain knowledge. Their highly flexible architectures can learn directly from raw data and can increase their predictive accuracy when provided with more data



E. SVM

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N—the number of features) that distinctly classifies the data points.



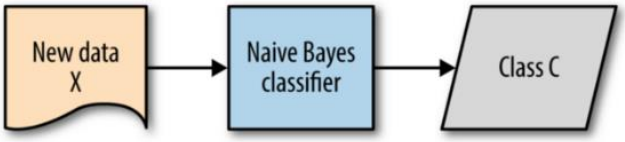
F.Naive Bayes

Statistical method for classification and Supervised Learning Model .Assumes an underlying probabilistic model, the Bayes Theorem.Can solve problems involving both categorical and continuous values attributes It is a probabilistic classifier that makes classifications using the Maximum A Posteriori decision rule in a Bayesian setting. It can also be represented using a very simple Bayesian network.

In order to do this, we use Bayes rule.

$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$

$$P(C_j \mid A_1, A_2, \ldots, A_n) = \frac{\left(\prod_{i=1}^{n} P(A_i \mid C_j)\right) P(C_j)}{P(A_1, A_2, \ldots, A_n)}$$



G. Logistic Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

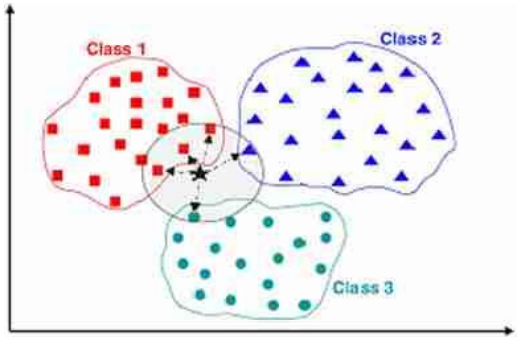
where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

H.KNN

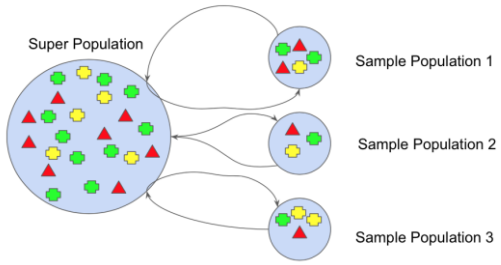
KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. KNN falls in the supervised learning family of algorithms. Informally, this means that we are given a labelled dataset consiting of training observations (x,y)(x,y) and would like to capture the relationship between xx and yy. More formally, our goal is to learn a function h:X→Yh:X→Y so that given an unseen observation xx, h(x)h(x) can confidently predict the corresponding output yy.The KNN classifier is also a non parametric and instance-based learning algorithm.



I.Bagging

Bootstrap Aggregation (or Bagging for short), is a simple and very powerful ensemble method.An ensemble method is a technique that combines the predictions from multiple machine learning algorithms together to make more accurate predictions than any individual model.Bootstrap Aggregation is a general procedure that can be used to reduce the variance for those algorithm that have high variance.

Given a standard training set D of size n, bagging generates m new training sets D_{i}, each of size n′, by sampling from D uniformly and with replacement. By sampling with replacement, some observations may be repeated in each D_{i}. If n′=n, then for large n the set D_{i} is expected to have the fraction (1 - 1/e) (≈63.2%) of the unique examples of D, the rest being duplicates.

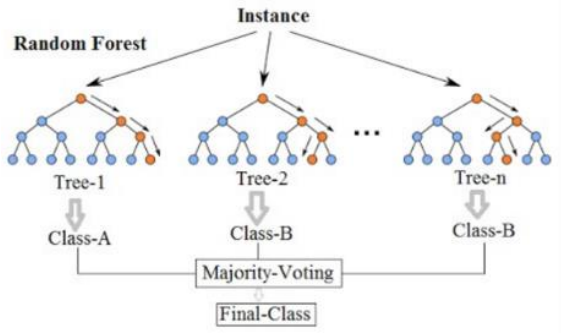


J. Random Forest

Random forest algorithm is a supervised classification algorithm. This algorithm creates the forest with a number of trees.

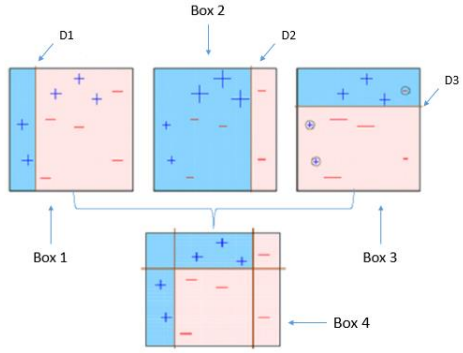
Random Forest pseudocode: Randomly select “k” features from total “m” features.Where k << m Among the “k” features, calculate the node “d” using the best split point.Split the node into daughter nodes using the best split.

Repeat 1 to 3 steps until “l” number of nodes has been reached.Build forest by repeating steps 1 to 4 for “n” number times to create “n” number of trees.



K. Ada Boosting

Ada boosting is nothing but Adaptive Boosting. AdaBoost is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. It retrains the algorithm iteratively by choosing the training set based on accuracy of previous training and the weight-age of each trained classifier at any iteration depends on the accuracy achieved.



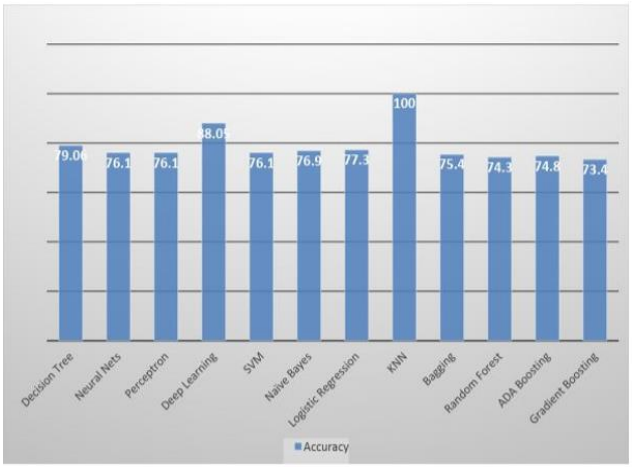
V. MODEL EVALUATION

The blood donor classification model was evaluated based on their accuracy.

Input data consisting of no.records were tested and used to construct network by 10 fold cross-validation. Here we provide the accuracies of the all the algorithms used in predicting the donor.

No.	Algorithm	Accuracy
1	Decision Tree	79.06
2	Neural Networks	76.1
3	Perceptron	76.1
4	Deep Learning	88.05
5	SVM	76.1
6	Naïve Bayes	76.9
7	KNN	100
8	Bagging	75.4
9	Random Forest	74.3
10	Ada Boosting	74.8
11	Logistic Regression	77.3

Bar graph representation of accuracies:



From the above image KNN provides 100% accuracy, followed by SVM.For the given dataset KNN works best.

V. CONCLUSION

Blood donation is an essential activity to import a raw material into the blood supply chain.

Five factors influencing blood donor behaviors suggested are used as a framework to construct a models to predict donors. A sample has yielded a reliability scale in an acceptable significant level of information.

The accuracy test of donor group prediction was done using above mentioned algorithms in order to predict the answer from a series of donors. As a result, it has been found that the models are able to predict the donors willing to donate blood.

Although the sample size chosen in this study is only in the provincial level, the result shows that the KNN model has a relatively high accuracy value. This is an indication that the model is able to learn the pattern of blood donors from questionnaire with satisfactory results. Moreover, this study can be used as a prototype and expand the sample group in order to develop blood donor classification model in ares which will be beneficial for developing blood donor database system in different levels. Furthermore, this classification models will contribute greatly for blood acquisition especially when there are emergency needs for blood for uses in the live saving treatments.

VI. FUTURE WORKS

Other factors affecting blood donation should be explored in order to better manage blood donors' behaviors like location,blood group which adds significant value to the prediction. The work can be extended by enlarging the sample size of the study. Moreover, other machine learning techniques can be used comprehensively in analyzing the model and comparing the results.

REFERENCES

[1] https://www.r-project.org/about.html
[2] https://archive.ics.uci.edu/ml/datasets/Blood+Transfusion+Service+Center
[3] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm
[4] https://en.wikipedia.org/wiki/Decision_tree_learning
[5] https://en.wikipedia.org/wiki/Artificial_neural_network
[6] https://en.wikipedia.org/wiki/Perceptron
[7] https://en.wikipedia.org/wiki/Deep_learning
[8] https://en.wikipedia.org/wiki/Support-vector_machine
[9] https://en.wikipedia.org/wiki/Naive_Bayes_classifier
[10] https://en.wikipedia.org/wiki/Bootstrap_aggregating
[11] https://en.wikipedia.org/wiki/Random_forest
[12] https://en.wikipedia.org/wiki/AdaBoost
[13] https://en.wikipedia.org/wiki/Logistic_regression