*Turkish Journal of Computer and Mathematics Education* *Vol.11 No.03 (2020), 1759-1766*
*Research Article*

# Ensemble Machine Learning Model to Predict Benefaction of an Individual in the Health Sector

**Shashikala B M1*, Pushpalatha M P2, Vijaya B3**

1 Department of Computer Applications, Sri Jayachamarajendra college of Engineering, Mysuru 570006,India
2 Department of Computer Science and Engineering, Sri Jayachamarajendra college of Engineering,Mysuru 570006,India
3 Department of Pathology, JSS Medical College ,Mysuru 570015,India
Corresponding Author Email:shashibmk@gmail.com

## ABSTRACT
Ensemble methodis a machine learning technique that combines several base models in order to produce one optimal predictive model. This work aims to develop blood donor's prediction model, for the management of the blood bank during emergency situations using ensemble method. The proposed model uses two supervised algorithms including multivariate regression and decision tree algorithms. An automated intelligent system is developed that learns from the data presented to the machine learning model to predict blood donors. The system is integrated with score allocation to the blood donors. A network of available ethical blood donors' model has been developed which can be used in case of an emergency for an ailing person. Machine Learning techniques are used to find a perfect matched donor with respect to blood group, medical history and other demographics. A prioritized/ranked donor list based on their medical history, habits and other blood related metrics is generated to benefit the receivers. The ensemble methods used in this intelligent system helps in report generation facilitating medical experts and the society in decision making leading to increased number of donors.
**Keywords:** blood donor,ensemble methods,intelligent system,machine learning

## 1. INTRODUCTION

Over the last couple of decades, the healthcare industry has digitized operational and patient data. Digitization brought instant access to information, made sharing of information easier among healthcare professionals and improved the efficiencies of patient outcomes. Automation has revolutionized the healthcare industry and made it more cost-effective for organizations to run day-to-day operations. Today's automation technologies are capable of far more than human administrators. Tasks such as reduction of administrative workloads, improvement of the consistency of patient care elimination of waste, enhancement of information exchange, analyzation of data, and monitoring of patients can all be streamlined with data automation. Automation of routine tasks can cut the amount of paperwork that healthcare organizations have to deal with. It can also reduce staffing costs and increase operational efficiency. Automation would also result in the efficiencies and synergies across the entire value chain of activities that healthcare organizations provide. The automation of the value chain would especially benefit large hospitals when they integrate the end-to-end activities that they undertake in a single, coherent, and unified set of software. The practice of storing medical records of patients and referring to them on subsequent visits would greatly benefit the doctors and the paramedics who would not only have access to the patients' history but also have details about the patients in terms of medications and allergies to any specific drug.

Healthcare industry requires automated process system to help in its decision making specially in donation services, where in too much of manual process is involved and is time consuming, and increases human errors, hence generation of wrong data and make wrong

decision. Accurate prediction of the number of donors can help medical professionals know the future supply and plan accordingly to entice voluntary donors to meet demand. Hence, an intelligent system that can integrate major operations involved, make efficient decisions, improve communication and introduce significant agility is highly crucial. There are about 1.5 million people require blood regularly. Not all of them get the blood they seek in time. Many people die due to wrong blood donors or lack of donors. The current Blood Donations Systems have failed to verify the authenticity of the data provided by the donors. One cannot be sure if the donor is actually having the described characteristics mentioned in the system and at the end require manual process of verification and tests. The current system also does not involve any automation, and the donors displayed are not analyzed in any way making the best available donor go unnoticed sometimes [1].

So, with an aim of providing an easy-to-use bridge between blood donors and blood seekers a system is developed. The proposed system of automating the selection of donors implements all the features of automation which reduces the manual work and help in accurate decision making and task like analysis of input data, processing the data and final result generation. The system not only help huge institutions but also small and middle scaled clinics due to its cost efficiency and easy to use features and doesn't require any technical knowledge to use the software.

In the current work we use machine learning technique to find the perfect matched donor with respect to blood group, medical history and entire profile. The donors are investigated on regular basis and their profiles are also updated regularly. In this work, two data mining techniques are proposed to predict blood donor. Prediction is performed by using two algorithms: Multivariate regression model and decision tree. Prediction is carried out by combining two algorithms using ensemble method.

## 2. RELATED WORK

Demand for blood increased in hospitals, therefore it is required to maintain an adequate supply of blood that is readily available [2]. To extract the hidden information from databases various data mining techniques have been proposed. When data is massive and heterogeneous data miming and knowledge extraction becomes challenging. To improve the performance of individual algorithm ensemble learning methods used. In distributed data mining ensemble method plays a significant role. Ensemble learning study plays a vital role in real world problems [3].

Intelligent modelling techniques Multilayer Perceptron neural network (MLP) and Probabilistic NeuralNetwork(PNN) are used to study classification abilities of these neural networks and the various dimensions of blood donation behavior by uncovering the patterns in data set [4]. In order to predict the transfusion requirement for patients admitted to the hospital four machine learning algorithms logistic regression, random forest, neural networks and gradient boosting trees used [5].

Sensitivity and specificity analysis demonstrated that Artificial Neural Network (ANN) can accurately predict the transfusion requirement of patients, using the information available at the time of patient entry [6]. To predict whether the individual is donor or not based on questionaries' data Naïve Bayes and K-Nearest Neighbor (KNN) algorithm used. KNN algorithm performed well for the dataset compared with naïve Bayes [7]. Ensemble learning methodologies and techniques gained a major attention in scientific and industry development [8].Ensemble models to improve the accuracy of predicting two complications of blood transfusion, transfusion associated circulatory overload (TACO) and transfusion-related acute lung injury (TRALI) are used. The results are encouraging with Area under curve of 0.84 and sensitivities between 0.82 and 0.92[9]. A model developed using machine learning approach for classification of depression using meta-cognitive neural network (MCNN) classifier with projection-based learning (PBL). To improve the efficiency of model Particle swarm optimization (PSO)

algorithm used to select the best parameter for MCNN-PBL classifier [10]. A mixed ensemble model developed using boosted c5.0 and support vector machine for hospital readmission within a time frame considering admission parameters, the Support vector machine given greater sensitivity value than the c5.0 prediction. Accuracy for the ensemble model ranged from 81%to 85% [11].

To handle imbalance of data a Dynamic Classifier Ensemble method for Imbalance Data proposed(DCEID). The experimental results show the classification performance of DCIED is better than the existing Dynamic Classifier (DCS)and Dynamic Ensemble Selection (DES) [12].

A modified decision tree algorithm based on genetic algorithm proved higher accuracy compared with C4.5 decision tree and Support Vector Machine(SVM) algorithm [13].A ordinal decision-tree based ensemble approach using Adaboost and random forest algorithms developed for classifying the regional daily growth factor of the spread of an epidemic based on several factors[14].

## 3.DATA PREPARATION

Data of blood donors required to train the model had been officially collected directly from the medical laboratories, hospitals and other medical institutions which have verified profiles of the donors, after receiving the donors consent to publish his/her information for Blood donation collected for a period of 2018 to 2019.Table 1 gives the dataset attributesfor blood donor.

### 3.1 Data Preparation

The attributes required for our study is discussed with the medical officer and the columns donor number, marital status, date of collection, date of end, and time were determined to be unnecessary for the study were removed. Cases missing blood types were also removed from the dataset.

### 3.2 Missing data Imputation

Data preprocessing was performed on the dataset. To eliminate the missing values imputation has been performed on the missing vales by considering the mean imputation. In case of Boolean values, missing column are replaced with false value as it fits in our study in order to train the model.

### 3.3 Outlier detection and removal

Clinically accepted ranges for blood transfusion parameters were used to detect outliers. Outliers removed considering the clinical ranges from the data set.

### 3.4 Data Transformation

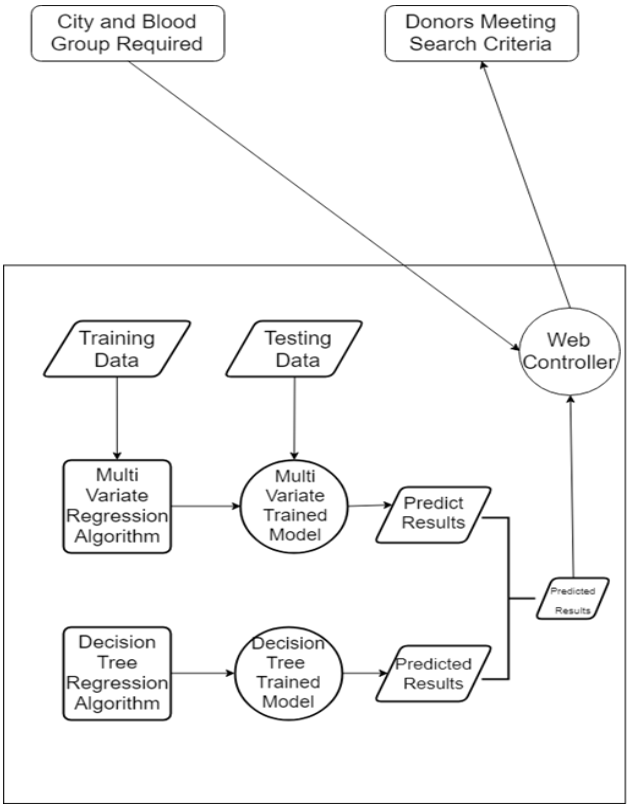
A donor not having any habits and not suffering from diseases is considered as no. otherwise as yes. The attribute is set 0 for the value no,1 for value yes.

**Table 1. Attributes of Blood donor data.**

| Attribute Name | Type |
|---|---|
| Name | Categorical |
| Location | Categorical |
| Gender | Categorical |
| Phone number | Number |
| Blood type | categorical |
| Age | Number |
| Weight | Number |
| Alcohol | Boolean |
| Smoking | Boolean |
| Medications | Boolean |
| Cancer | Boolean |
| HIV/AIDS | Boolean |
| Tattoos | Boolean |
| STD | Boolean |
| Drug Usage | Boolean |
| Haemoglobin | Decimal |
| levels | Decimal |
| Red Blood cell | Decimal |
| count | Boolean |
| WhiteBloodCell | Boolean |
| count | |
| Hepatitis | |
| TB | |

The proposed model is given in Fig 1.

**Figure 1.** The proposed modeldesign




## 4.EXPERIMENTAL SETUP

Multivariate regression, Decision tree regressor and ensemble algorithms are used for implementation. The attribute set used for both multivariate regression, decision tree regressor algorithms are same. The score of each attribute obtained from both the algorithms used for ensemble algorithm. The attributes selection for algorithm is given in Table2. The following gives the proposed model Creation algorithm.

**Step 1**: Select Independent variables and target dependent variable

**Step 2**: Pass Independent variable columns and Target Dependent variablecolumns to multivariate regression algorithm.

**Step 3**: Create Model using the fit() method

**Step 4**: TrainMultivariate Regression Model using training data,with both independent and dependent columns against dependent variable

**Step 5**: Pass Independent variable columns and TargetDependent variable columns to decision tree regressor algorithm

**Step 6:** Create Model using the fit() method

**Step 7:** Train Model using training data with both independent and dependent columns against dependent variable

The following gives the data prediction algorithm.

**Step 1**: Select independent variables and target dependent variable.

**Step 2**: Pass independent variable columns and target dependent variable columns to multivariate regression model

**Step 3**: Predict Score using the Predict() method of Multivariate Model.

**Step 4**: Pass Independent variable columns and target dependent variable columns to decision tree regressor model

**Step 5**: Predict Score using the Predict() method of decision tree regressor model

The following gives the ensemblealgorithm.

**Step 1:** Import Output from multivariate regression model and decision treeregressor model

**Step 2**: Extract score column values from the predicted output of both algorithms.

**Step 3**: Calculate the mean of both scoresand store in array.

**Step 4**: Floor (Round off) the value and replace array value with floored Value

**Step 5**: Assign floored Score to output score column

**Table 2. Attributes selection for algorithm**

| Attribute Name | Type |
|---|---|
| Age | Number |
| Weight | Number |
| Alcohol | Boolean |
| Smoking | Boolean |
| Medications | Boolean |
| Cancer | Boolean |
| HIV/AIDS | Boolean |
| Tattoos | Boolean |
| STD | Boolean |
| Drug Usage | Boolean |
| Haemoglobin | Decimal |
| levels | Decimal |
| Red Blood cell | Decimal |
| count | Boolean |
| White Blood | Boolean |
| Cellcount | |
| Hepatitis | |
| TB | |

## 5. RESULTS AND DISCUSSION

The proposed system which is titled as," Ensemble Machine Learning Model to Predict Benefaction of an Individual in the Health Sector ", is the intelligent automated model for predicting the blood donors based on the score assigned to the users based on their health conditions and other parameters

The score in Multivariate regression model is calculated using the following equation 1 as follows.

$y = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c$ (1)

Here, y is dependent variable, x1, x2, $x_n$ are independent variables, $b_n$'s (n=1,2…n) are the regression coefficients, which represent the value at which the criterion variable changes when the predictor variable changes, C is the intercept. A constant that finds the value of y when x and y are 0.

For simplified explanation here few attributes age, smoking, alcohol and score are considered. Equation calculates Dependent Variable Score value using independent variables Age, Smoking and Alcohol. Table 3 gives few values for equations that determine Score based on other variables.

**Table 3. Score Calculation**

| Age | | Smoking | Alcohol | Score |
|---|---|---|---|---|
| 22 | 0 | 0 | 100 | |
| 22 | 10 | 99 | | |
| 22 | 11 | 98 | | |
| 23 | 00 | 99 | | |
| 23 | 1 | 0 | 97 | |

Score =b+ m1 x Age + m2 x Smoking + m3 x Alcohol

Here, Score=Dependent Variable, m1, m2, m3 = Coefficients, b = Intercept.

Here we create a line such that it stays as close as possible to the predicted values(Score) using Root Mean Square Error equation 2 as given below.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(S_i - O_i)^2}$$

So, when both the LHS and RHS to the algorithm provided it finds out the relationship between the independent variables to the dependent variable and calculates the coefficient and the intercept, when a new dependent variable has to be calculated using independent values it will use the previously available historic coefficients and intercept and try predicting the dependent variable value. Based on the input the equation will be like

score = 133.25 + -1.5 x age + -1.5x smoking + -0.75x alcohol (3)

So now if a new Score is to be predicted we need to multiply his age, smoking, alcohol values with the intercept and coefficients. If the new person's age is 24 and no smoking or drinking habit. The independent variables are age=24, smoking=0, alcohol=0.These values are substituted to equation 3, the score approximately should be 98, since a person with age 23 and no bad habits has a score of 99.The Score obtained is 97.25 which is acceptably correct as we provided very few inputs. Similarly, the algorithm forms the equation for all other attributes against the dependent value score and calculates the score in similar way. The model uses the decision tree algorithm and creates a tree structure of all the attributes considered for assigning the score to theusers. Decision Tress work by creating a tree that will traverse down based on the condition in each node, and the end result will be the value in the leaf node. Table 4 gives the values for calculation of Entropy and Gain.

**Table 4: Values for GAIN and Entropy Calculation**

| Smoking | | HIV | | Alcohol | Eligibility |
|---|---|---|---|---|---|
| 1 | 1 | 1 | I | | |
| 1 | 1 | 0 | I | | |
| 0 | 0 | 1II | | | |
| 1 | 0 | 0II | | | |

Entropy and Gain are calculated for the table 4 values to find the best possible way to split the tree using the following equation 3 and equation 4 as follows [15].

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \qquad (3)$$

Entropy E for a dataset S with *C* classes, where *pi* is the probability of randomly picking an element of class *i* (i.e., the proportion of the dataset made up of class *i).*

The information gain Gain (S, A) of an attribute A relative to a collection of data set S, is calculated as follows.

$$\text{Gain}(S,A)=\text{Entropy}(S)-\sum_{v \in V\ alues} \frac{|S_v|}{|S|} Entropy(S_v) \quad (4$$

Where, Values(A) is the all-possible values for attribute A, and $S_v$is the subset of S for which attribute A has value v.

Split on the feature of smoking Gain = 0. 3112.Similarly Gain is 1 for feature HIV and Gain is 0 for feature alcohol. Information Gain is maximum on HIV, so we split on Feature HIV, the root node best suited is feature HIV, while splitting the dataset on feature HIV, the child contains pure subset of the target variable, so not required to further split the dataset. Generally, if we observe, if HIV is True then the eligibility is No irrespective of the alcohol and smoking habits, and if HIV is False then the eligibility is Yes irrespective of the alcohol and smoking habits. So, by observation, splitting tree based on HIV is best approach. Thetreestructureisusedtoassignthescorefornewarrivingdataandusers.The modeloncecreatedlearnstopredicttheresultsusing thedata andimproves on its own when new data is fed into the model. Here we combined the results of both algorithms which is the score, and get the mean of both algorithms output so that we can get the best of the two results which is much accurate than the individual algorithm result. The donors are allocated with score.

The data which is fed as input iscollectedfromtrustedsourcesandisverifiedbeforeprovidingittothelearning modelandthedetailsofthepatientsanddonorsarescurelymaintainedincloud environment. The model developed manipulates the donors to donate more frequently by utilizing their urge to gain fame on socialmedia and other websites, the donors who donate blood are assigned scores and can be linked to their socialmedia accounts and flaunt their scores to the world not only the model is a hall of fame but also provides further offers when linked with other topsocialmediaapplications,thesefeaturesofthepredictingmodelsencourage the donors to donate more frequently. As the proposed system is an automated system to predict the donors, it reduces the manual work involved, or in present donation system a donor who arrives at the clinic or the hospital to help any patient,themedicalexpertstakethedonorsbloodsampleandcarryoutalonglist of preliminary tests which includes test of blood group, other infections and diseases, ask the donor to fill certain questionnaires which contains their history, habits and other parameters which play a vital role in decision making and donorselection. Theproposedsystemautomatesalltheseprocesses and medical experts are provided with the summary of donor's health report and hence reduces human work and is efficient. Table 5 gives the output for score based on the search of blood group.

**Table 5. Score Result**

| Name | Location | Phone Number | Blood group Score |
|---|---|---|---|
| Dave | Mysuru | 9987896982 | 100 |
| Thomas | Mysuru | 867845632399 | |
| Ravi | Mysuru | 9345645734 | 98 |
| Bharat | Mysuru | 8767788455 | 97 |

## 6.CONCLUSIONS

This paper addresses the implementation of machine learning concepts and migration towards automation in healthcare. The prediction algorithms like decision tree, ensemble methodologies are used in developing the intelligent predicting model. The performance of the hybridschemes is implemented and tested to determine the appropriate algorithms for prediction and decision making. Traditionally, classificationandpredictionproblemsweresolved withhumanintervention.The use of modern computers has expedited large-scale statistical computation, and has also made possible to design decision making models and methods that are impracticaltoachievemanually.Thecomplexnatureofreal-worldclassification applications is solved by the involvement of computationally intelligent paradigms. The machine learning paradigms exhibit an ability to learn/adopt to the given situations and facilitate intelligent behavior with reduced computational burden. The simulation results prove the competency of the decision-making algorithms in terms of classification accuracy, computational time, scalability.

## 7.FUTURE SCOPE

Recent field of technology is growing and data are by nature dynamic. The scope of future work can deal with Incremental learning, which stores the existing model and processes the new incoming data more efficiently. More specifically, the models with incremental learning can be used in categorization process to improve data classification,establishmoresecure,fastandother modesofcommunicationslike messaging,emailfacilitiesandetc. Theconceptsofgeospatialattributesofdonors can be leveraged and prediction parameters can be enhanced based on these parameters too. When these enhancements are incorporated in the classification system, it would help further improve in the performance and be useful for applications meant for the explicit classificationsystem.

## REFERENCES

[1] Selvamani K., Ashok Kumar Rai. (2015). A novel Technique for online blood bank management. ,International Conference on Intelligent Computing,Communication and Convergence, Elsevier,568-573.https://doi.org/10.1016/j.procs.2015.04.137

[2] Oscar Salviano., W.Cezarino., Giselle R.Salviano.(2012).A decision-making tool for demand forecasting of blood components,IFAC Proceedings.14(6):14991504,https://doi.org/10.3182/20120523-3-RO-2023.00201

[3] Prakash chandra Jena., Subhendu Kumar Pani., DebahutiMishra.(2018). A novel approach to ensemble learning in distributed data mining, International Journal of Engineering&Technology,7(2):235238,https://doi.org/10.14419/ijet.v7i2.33.14159

[4] Mohamed M. Mostafa.(2009). Profiling blood donors in Egypt. A neural network analysis. Expert System with Applications.36(2009)5031-5038, https://doi.org/10.1016/j.eswa.2008.06.048

[5] Andreas Mitterecker., Axel Hofmann., Kevin M.Trentino.,AdamLloyd.,MichaelF.Leahy.,KarinSchwarzbauer.,Thomas Tschoellitsch.,Carl Bock.,SeppHochreiter.Jens Meier.(2020)Machine learning-based prediction of transfusion,Original Research,Transfusion.60(9):1977-1986,https://doi.org:10.1111/trf.15935

[6] Steven Walczak.(2005)Artificial neural network medical decision support tool:predicting transfusion requirements of ER patients,IEEE Transaction on Information Technology in Biomedicine ,9(3):468-474,https://doi.org:10.1109/titb.2005.847510

[7] B M Shashikala., M P Pushpalatha., B Vijaya.(2019).Machine learning approaches for potential blood donors prediction,Emerging research in electronics ,computer science and technology,Lecture Notes in Electrical Engineering

545,https://doi.org/10.1007/978-981-13-5802-9_44

[8] Polikar, R. Ensemble learning. In ensemble machinelearning,Springer:boston,MA,USA ,2012;pp.1-34,https://doi.org/10.1007/978-1-4419-9326-7

[9] Dennis Murphree,CheNgurfor,SudhindraUpadhyaya,NagesMadde,LeanneClifford,Daryl J Kor,JyotishmanPathak,Ensemble Learning approaches to predicting complications of blood transfusion,IEEE,2015,doi:10.1109/EMBC.2015.7320058.

[10] Vanishri Arun, Murali Krishna,Arunkumar B V,Padma S K,Shyam v," Exploratory Boosted Feature Selection and Neural Network Framework for Depression Classification, International Journal of Interactive Multimedia and Artificial Intelligence, (2018). http://dx.doi.org/10.9781/ijimai.2018.10.001

[11] Liorturgeman,JerroldH.May,A mixed-ensemble model for hospital readmission,Aritifial Intelligence in Medicine,Elsevier,2016,pp:7282.doi:10.1016/j.artmed.2016.08.005

[12] JinXiao,LingXie,ChanzhengHe,XiaoyiJiang,"Dynamic classifier ensemble model for customer classififcation with imbalanced class distribution",Expert system with applications 39(3):36683675,2012,doi: 10.1016/j.eswa.2011.09.059

[13] Dong-Sheng Liu,Shu-jiang Fan,"A modified decision tree algorithm based genetic algorithm for mobile user classificationproblem",2014,doi:https://doi.org/10.1155/2014/468324

[14] Gonensinger,MatanMarudi,"Ordinal Decision-tree-based ensemble approaches:the case of controlling the daily local growth rate of the COVID-19 epedemic, Application of information theory to industrial and servicesystems,22(8),871,2020,doi:https://doi.org/10.3390/e22080871

[15] HeZang, RunjingZhou,"The analysis and optimization of decision tree based on ID3 algortihm",International conference on modelling, identification and control,IEEE,1012,2017,**doi:** 10.1109/ICMIC.2017.8321588