4 Application of CART Algorithm in Blood Donors Classification

4.1 Introduction

Blood donorship is a critical aspect of healthcare globally. Section 3 provides some base background in this context. In addition the ability to recruit safe donors and having a pool of regular voluntary donors are critical requirements. In this section we discuss on the classification models of blood donors. Hence the evaluation of potential donors is critical for safety. The donated blood undergoes screening includes testing for diseases that can be transmitted by a blood transfusion, including HIV and viral hepatitis. There is also donor medical history data captured along with a short physical examination to make sure that the donation is not hazardous to his or her health. The following provides a broad overview of the blood donation process a five step process (figure 4.1).

- **Step 1** - The donor registers, Health history and mini physical are completed. About 500 c.c. of blood and several small test tubes are collected from each donor. The bag, test tubes and the donor record are labelled with an identical bar code label to keep track of the donation. The donation is stored in iced coolers until it is transported to a center.

- **Step 2** - The donated blood is scanned into a computer database. Most blood is spun in centrifuges to separate the transfusable components red cells, platelets, and plasma. The primary components like plasma, can be further manufactured
into components such as cryoprecipitate. Red cells are then leuko-reduced. Single donor platelets are leukoreduced and bacterially tested. Test tubes are sent for testing.

- **Step 3** - In general the steps 2 and 3 take place in parallel. The test tubes are received in testing laboratories. An number of tests are performed on each unit of donated blood to establish the blood type and test for infectious diseases. The test results are transferred electronically to the manufacturing facility within 24 hours. If a test result is positive, the unit is discarded and the donor is notified. Test results are confidential and are only shared with the donor, except as may be required by law.

- **Step 4** - Once the test results are received, units suitable for transfusion are labelled and stored. The red cells are stored in refrigerators at 6C for up to 42 days. Platelets are stored at room temperature in agitators for up to five days. Plasma and cryo are frozen and stored in freezers for up to one year.

- **Step 5** - The blood is available to be shipped to hospitals 24 hours a day, 7 days a week.
4.2 Details

A donation is when a donor gives blood for storage at a blood bank for transfusion to an unknown recipient. A donation is known as directed when a person (often a family member) donates blood for transfusion to a specific individual. This event is generally referred to as a blood drive. These have been discussed in section 3.2. It is therefore important to ensure that the donation is utilised within the shelf life of the blood product (and not wasted). Many organizations and governments provide specific regulations to minimize the wastage of blood. Blood [AR13] has the following key components.
• Whole blood (red cells, white cells and platelets 45% of volume) suspended in plasma (55% of volume)

• Red cells (or erythrocytes) carry oxygen from the lungs to your body's tissue and take carbon dioxide (to the lungs to be exhaled)

• Platelets (or thrombocytes) whose main function is to interact with clotting proteins to stop or prevent bleeding

• Plasma is a fluid composed of 92% water and 7% vital proteins (albumin, gamma globulin, antihemophilic factor, and other clotting factors) and 1% (mineral salts, sugars, fats, hormones and vitamins)

• Cryoprecipitated Antihemophilic Factor (Cryo) is a portion of plasma and which is rich in clotting factors

The various blood components are provided graphically (figure 4.2). The details of the shelf life (table 4.1) of some of these components are also important to note.
Table 4.1: Blood Components Shelf Life

<table>
<thead>
<tr>
<th>Component</th>
<th>Shelf Life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Blood</td>
<td>35 Days</td>
</tr>
<tr>
<td>Red Blood Corpuscles</td>
<td>35 Days</td>
</tr>
<tr>
<td>Red Blood Corpuscles (with preservatives)</td>
<td>42 Days</td>
</tr>
<tr>
<td>Platelets</td>
<td>5 Days</td>
</tr>
<tr>
<td>Plasma</td>
<td>1 Year</td>
</tr>
<tr>
<td>Cryoprecipitate</td>
<td>1 Year</td>
</tr>
</tbody>
</table>
Review of Literature

Please note the introduction and section 3 provide some additional references from the overall context of DM. The following are more specific to this specific section. Veldhuizen et al [VF+13] provide insights in the donor cycle and donor segmentation. This research provides new tools for improving blood donor management. The study involved blood establishments (BE’s) from 18 European countries, the Thalassaemia International Federation and a representative from the South-Eastern Europe Health Network joined forces in DOMAINE. A questionnaire assessed blood donor management practices and the composition of the donor population using the newly proposed DOMAINE donor segmentation. The study found the response rate was high (88%). However, only 14 BE’s could deliver data on the composition of their donor population. The data showed large variations and major imbalances in the donor population. In 79% of the countries, inactive donors formed the dominant donor type. Only 21% regular donors were the largest subgroup and in 29%, the proportion of first-time donors was higher than the proportion of regular donors.

Gunanidhi et al [GV+09] have employed Simple Artificial Neural Network (SANN) based classification models. The research used standard datasets [BK13] for data mining (classification) and reviewed the performance of the algorithms. From an India specific context Chaudhary [C09] discusses specific overall governance and controls that have been developed in the area of blood transfusion services. Bharucha [B05] discusses quality management of a BTS
starts with safe donor recruitment and donor care. Vladimir et al [VB+13] have adopted an alternative analytical method to model colorectal cancer (CRC) patients long-term survival by assessing the prognostic value of the Ki-67 protein as a marker of tumour cell proliferation. A score was assigned as the percentage of positive tumour cell staining, denoted as proliferation index (PI), and was used in a multivariate analysis using a recursive partitioning algorithm referred to as classification and regression tree (CART) to characterize the long-term survival after surgery. Malik and Chaudhary [MC13] have compared classification algorithms for decision trees for data analysis. The study covered algorithms such as ID3, SLIQ, SPRINT, PrUning and BuiLDing Integrated in Classification (PUBLIC) and RAINFOREST decision tree classification algorithms. Lomax and Vadera [LV13] identify over 50 algorithms including approaches that are direct adaptations of accuracy-based methods, use genetic algorithms, use anytime methods and utilize boosting and bagging. The survey provides different studies and novel approaches to cost-sensitive decision tree learning.

Mani et al [ME+12] evaluated the performance (diagnostic aspects) of Linear Discriminant Analysis (LDA) and Classification And Regression Tree (CART) algorithms. The research was applied to discriminate glaucoma subjects. Yang and Ross [YR12] have adopted SVM for land-cover characterization using MODIS (MODeRate-resolution Imaging Spectroradiometer) time-series data. Classification performance was examined with respect to training sample size, sample variability, and landscape homogeneity (purity). The results were compared with Multi Layer Perceptron Neural
Networks (MLP) and CART.

Ghao et al [GZ+12] have compared multiple data mining algorithms to the TNM classification of malignant tumours (TNM) staging system using a dataset in which the training and testing data were from different sources. The nine data mining algorithms included artificial neural networks (ANN), Back Propagation network (BP), MLP, Radial Basis Function (RBF) neural network, General Regression Neural Network (GRNN), CART, Logistic Regression (LR), SVM, Adaptive Network-based Fuzzy Inference System (ANFIS), Bayesian Networks (BN) and TNM were compared for colorectal survival analysis. Santhanam and Sundaram [SS10a] have applied classification algorithms to blood donorship data and discusses how the classification can be enhanced with a standard dataset. In addition section 3.2 covers some specific research related to this context.

Strategies towards donor recruitment and retention have been presented from a south East Asian perspective. Another study to understand blood donor behaviour was undertaken by Schlumpf [SG+07]. This study self-administered questionnaire was completed in 2003 by 7905 current donors. With data mining methods, all factors measured by the survey were ranked as possible predictors of actual return within 12 months. Significant factors were analyzed with logistic regression to determine predictors of intention and of actual return.
4.3 Experiments

This research adopted the blood transfusion dataset [BK13]. This is based on donor database of Blood Transfusion Service Center in Hsin-Chu City in Taiwan (more details in section 9.3). The center passes their blood transfusion service bus to one university in Hsin-Chu City to gather blood donated about every three months. It must be noted that the extended nominal class of the Regular Voluntary Donor (RVD) has been extended based on the current dataset, it factors not just the specific blood donation drive in the base data set.

DM Tools

Weka [G95] has been adopted in this research serves as this provides the ability to load, preprocess and visualize data and also perform standard DM algorithms with sufficient parameterization. These algorithms can either be applied directly to a dataset or called from custom Java code.

Decision Tree Methods

Decision Tree (DT) learning [M03] is a common method used in DM. DT induction algorithms function in a recursive fashion. The process starts with an attribute selected as the root node. The most efficient tree (i.e. smallest) requires the root node to effectively split the data. The process continues to split; with each split it pare’s down the set of instances (data) until they all have the same classification. The best split is the one that provides the most information gain.
Information in this context (derived from the concept of entropy from information theory). This concept is applied to each split of the DT to provide an accurate classification. The information gain is computed at each step of the tree (to see if there is a reduction in entropy). One of the benefits of DT is that they can represent diverse types of data (numerical and nominal). It can also factor nominal data (discrete set of symbols) such as weather can be described in either numeric or nominal fashion.

The goal is to create a model that predicts the value of a target variable based on several input variables. Each interior node corresponds to one of the input variables; there are edges to children for each of the possible values of that input variable. Each leaf represents a value of the target variable given the values of the input variables represented by the path from the root to the leaf. If the target variable (response variable) is a nominal or categorical it is referred as classification tree and if continuous (known as regression tree). The key aspects of growing a tree are the following.

- Criteria for selecting a variable for split
- Criteria for split points for a selected variable
- Criteria for decide when to stop the tree growth

The building of a tree is accomplished by an algorithm that examines data from a training sample or created by subject matter experts. Most DT techniques differ on how the tree is created. The complexity of the algorithm is straightforward to analyze. For each tuple in the database, we search the root down to the leaf node. At
each level the complexity depends on the number of levels (branching factor). The following are some salient benefits of decision trees.

- Overall ease of use and efficiency
- Rules can be derived that are easy to interpret
- Decision trees are natural constructs
- Trees are easier to explain to non-statisticians and subject matter experts
- Models are invariant under transformations in the predictor space
- The treatment of missing values is more satisfactory than most other models
- Multi-factor response is also handled by the model

CART Algorithm

In this study the use of CART (Classification and Regression Trees) classification algorithm has been attempted [BF+84]. Classification tree analysis is when the predicted outcome is the class to which the data belongs. Regression tree analysis is when the predicted outcome can be considered a real number. Paul [P09] provides a review and comparison of classification algorithms for medical decision making. Stothers et al [SG+09] have applied DM in the classification of male lower urinary tract symptoms using mathematical modelling and a regression tree algorithm of noninvasive near-infrared spectroscopy parameters.
One of the advantages of using classification trees is their ability to provide easy to understand classification rules. Each node of a classification tree is a rule. The only exception to this would be in cases where the tree is very large and in such cases there may need to be a more specific focus on pruning required to optimize the tree size. Trees are easy off the shelf classifiers that require no variable transformation.

CART builds the tree by recursively splitting the variable space based on the impurity of the variables to determine the split till the termination condition is met. The GINI impurity determines how often a randomly chosen element from the set would be incorrectly labelled if it were randomly labelled according to the distribution of labels in the subset. Using this algorithm for the classification of a regular voluntary donor, the decision tree (figure 3.2) has been generated.

4.4 Results

It should be noted that pruning the tree results in making the tree shorter and simpler and at the same time, one has to guard against the possibility of over fitting of data. Some of the key results are the following.

- In this case the RVD based classification tree has reduced the tree complexity of the number of leaf nodes by 7 in comparison to the donated blood attribute (DB2K7) based tree (figure 4.3).
- The CART decision tree (figure 3.2) shows a simple structure with the root (frequency) and leaf node recency. The other key
observations being that the monetary attribute does not figure in the decision tree.

• This is in conjunction with the study [SG+07] which identified higher prior donation frequency as a predictor. Donor behaviour is dependent on more than one factor alone. The research finding indicated that the following appear to significantly predict donor return.
  * Donation frequency
  * Intention to return
  * Donation experience
  * Convenient location

• The CART RVD derived model provides an optimal classification accuracy (confusion matrix - True Positive, Recall and Precision of 0.999) outlined in the confusion matrix (table 3.4).

In addition the model can be customized to geographic (and other specific attributes). The dashboard (figures 4.4 to 4.7) provides evidence of the capability of the blood donor type concepts discussed. The DT model coupled with these dashboard and representation provide the management of blood banks and health care officials with clear information. This can be extended as well to the existing blood banks systems.

The key benefit of the creating an extended RVD definition based on the donor definition (along with the application of CART) provides a standard model to determine the donor behaviour and provides the capability to build a classification model. This additional nominal class can be easily computed based on the
definition. This demonstrates the ability to create a dashboard to manage the status of blood donors that will help blood transfusion centers manage the blood bank. The next section covers more in details of the blood donorship models.

Table 4.2: Blood Donor Types Definitions

<table>
<thead>
<tr>
<th>Attributes</th>
<th>LVD</th>
<th>RVD</th>
<th>NVD</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recency (Months)</td>
<td>&gt; 6</td>
<td>&lt;=6</td>
<td>&lt;=100</td>
<td>&lt;=100</td>
</tr>
<tr>
<td>Frequency (Months)</td>
<td>&lt;=100</td>
<td>&gt;=4</td>
<td>&gt;=1</td>
<td>&lt;2</td>
</tr>
<tr>
<td>Monetary (c.c.)</td>
<td>&lt;=9500</td>
<td>&gt;=2000</td>
<td>&gt;=250</td>
<td>&lt;=250</td>
</tr>
<tr>
<td>Time (Months)</td>
<td>&gt;80</td>
<td>&gt;24</td>
<td>&lt;=24</td>
<td>&lt;=100</td>
</tr>
</tbody>
</table>

Table 4.3: Blood Donor Base Statistical Measures - Attribute Monetary (in 1000 c.c.)

<table>
<thead>
<tr>
<th>Blood Donor Type</th>
<th>Mean</th>
<th>Median</th>
<th>Variance</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>RVD</td>
<td>51.5</td>
<td>23.75</td>
<td>57.85</td>
<td>360.5</td>
</tr>
<tr>
<td>LVD</td>
<td>10.39</td>
<td>2.75</td>
<td>17.39</td>
<td>114.25</td>
</tr>
<tr>
<td>NVD</td>
<td>8.75</td>
<td>8.75</td>
<td>1.06</td>
<td>17.5</td>
</tr>
<tr>
<td>Others</td>
<td>4.3</td>
<td>3.5</td>
<td>4.03</td>
<td>21.5</td>
</tr>
</tbody>
</table>
Figure 4.3: CART Decision Tree Based on DB2k7
Figure 4.4: Donor Pool Status Dash Board Type 1
Figure 4.5: Donor Pool Status Dash Board Type 2
Figure 4.6: Donor Pool Status Dash Board Type 3
Figure 4.7: Donor Pool Status Dash Board Type 4