LITERATURE SURVEY

We examined the academic literature and grouped what we found into a couple different categories. First, blood banks often will survey donor volunteers to try and understand the factors that led them to donate. For example Godin, Conner et al. (2007) found that the important factors that lead to repeated blood donation among experienced donors were intention, perceived control, anticipated regret, moral norm, age, and past donation frequency. Moreover, the factors leading to repeated blood donation among new donors were only intention and age. Others have designed studies to understand one’s motives for donating blood. Sojka and Sojka (2008) surveyed over five hundred donators and found that the most commonly reported motivator among their participants was friend influence (47.2%), followed by media requests (23.5%). Lastly, they found that altruism (40.3%), social responsibility (19.7%), and friend influence (17.9%) were the primary drivers for blood donors to continue to be blood donors in the future. As stated previously, only around 5% of eligible donor population actually donate (Katsaliaki 2008). The reasons for this are regularly reviewed by social and behavior scientists to help improve population participation (Ferguson, France et al. 2007). The studies just discussed are outlined below in Table 1 are just a fraction of the many studies being performed to better understand the social and behavioral aspects of why people donate blood. We have found that many studies are trying to extend what is known as a theory of planned behavior (TPB), which continues to be developed in this area. This theory predicts the occurrence of certain behavior given that it is intentional and under volitional control (Veldhuizen, Ferguson et al. 2011). A systematic review and meta-analysis was performed in this area by Bednall, Bove et al. (2013).

|  |  |  |  |
| --- | --- | --- | --- |
| Authors | Methods | Data | Drivers |
| (Godin, Conner et al. 2007) | Logistic Regression | Survey (2070 experience donors, 161 new donors) | Experienced donors: intention, perceived control, anticipated regret, moral norm, age, and past donation frequency. New donors: intention and age |
| (Sojka and Sojka 2008) | Descriptive statistics | Survey (531 participants) | General motivators: friend influence (47.2%), media requests (23.5%). Continued donations: altruism (40.3%), social responsibility (19.7%), friend influence |
| (Masser, White et al. 2009). | Structural equation modeling | Survey 1 (263 participants); Follow-up survey (182 donors) | Moral norm, donation anxiety, and donor identity |

Table 1: Social and psychological studies investigating drivers of blood donations .

The focus of our study is to understand the performance that using traditional machine learning techniques can have at predicting future blood donation. Table 2 outlines what we believe is an exhaustive list of all published studies in this domain, the data set used, methods employed, and results achieved. The “-” symbol indicates that nothing is reported in their paper in this table field.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Authors | Methods | Data | RESULTS | |
| (Mostafa 2009) | ANN (MLP), ANN (PNN), LDA | Survey (430 records, 8 features) | ANN (MLP): Test accuracy (98%) ANN (PNN): Test accuracy (100%) LDA: Test accuracy (83.3%) | |
| (Santhanam and Sundaram 2010) (Sundaram 2011) | CART CART vs. DB2K7 | UCI ML blood transfusion data2 (748 donors, 5 features) | Precision/PPV (99%), Recall/Sensitivity (94%) | |
| (Darwiche, Feuilloy et al.2010) | PCA for feature reduction ANN (MLP) vs SVM (RBF) | UCI ML blood transfusion data (748 donors, 5 features) | SVM (RBF) using PCA: Test Sensitivity (65.8%); Test Specificity (78.2%); AUC (77.5%) MLP with features recency & monetary: Test Sensitivity (68.4%); Test Specificity (70.0%); AUC (72.5%) | |
| (Ramachandran, Girija et al. 2011) | J48 algorithm in Weka (aka C4.5) | Indian Red Cross Society (IRCS) Blood Bank Hospital (2387 records, 5 features) | Recall/Sensitivity (95.2%), Precision/PPV (58.9%), Specificity (4.3%) | |
| (Lee and Cheng 2011) | k-Means clustering, J48, Naïve Bayes, Naïve Bayes Tree, Bagged ensembles of (CART, NB, NBT) | Blood transfusion service center data set (748 records/donors, 5 features) | Bagged (50 times) Naïve Bayes: Accuracy (77.1%), Sensitivity (59.5%), Specificity (78.1%), AUC (72.2%) \* model had best AUC among competing models | |
| (Zabihi, Ramezan et al. 2011) | Fuzzy sequential pattern mining | Blood transfusion service center data set (748 records/donors, 5 features) | Precision/PPV (Frequency feature 88%, Recency feature 72%, Time feature 94%) | |
| (Boonyanusith and Jittamai 2012) | Artificial Neural Network (ANN), J48 algorithm (aka C4.5) | Survey (400 records, 5 features) | ANN: Accuracy (76.3%); Recall/Sensitivity (81.7%); Precision/PPV (87.9%); Specificity (53.8%) J48: Recall/Sensitivity (81.2%); Precision/PPV (87.3%); Specificity (52.5%) | |
| (Ashoori, Alizade et al. 2015) | C5.0, CART, CHAID, QUEST | Blood transfusion center in Birjand City in North East Iran (9231 donors, 6 features) | Model accuracy (train/test): C5.0 (57.49/56.4%), CART (55.9/56.4%), CHAID (55.56/55.61%), QUEST (55.34/56.11%) |
| (Ashoori, Mohammadi et al. 2017) | Two-step clustering, C5.0, CART, CHAID, QUEST | Census survey from a blood transfusion centers from Birjand, Khordad, & Shahrivar (1392 participants) | Important features: Blood pressure level, blood donation status, temperature Model accuracy: C5.0 (99.98%), CART (99.60%), CHAID (99.30%), QUEST (89.13%) |

Table 2: Predicting blood donation with a focus on data mining/machine learning techniques.

The first published study we found investigating machine learning classification techniques to identify donors versus non-donors was by Mostafa (2009). They show that it is possible to identify factors of blood donation behavior using machine learning techniques. They train and test two artificial neural network (ANN) variants; one using a multi-layer perceptron (MLP); the other a probabilistic neural network (PNN). They then compare these non-linear models to a linear discriminant analysis (LDA) model. They conclude that the ANN models both perform very well compared to LDA due the nonlinearities that exist in their data. Santhanam and Sundaram (2010) used the Classification and Regression Tree (CART) from the University of California – Irvine Machine Learning repository. They showed on this data set that this algorithm has the ability to classify future blood donors accurately that had donated previously (i.e. recall/sensitivity of 94%). We found a very similar study published by one of the original authors the following year with a comparison of what they call a Regular Voluntary Donor (RVD) versus a DB2K7 (Donated Blood in 2007), which led to slightly better recall and precision (Sundaram 2011). Their key contribution was that the RVD model realized better accuracy than DB2K7. Darwiche, Feuilloy et al. (2010) extend this investigation of this data set by testing ANN with a radial basis function (RBF) as well as investigate performance using Support Vector Machines (SVMs). Even though the feature space is limited they also build and evaluate these models using principal components analysis (PCA) as feature inputs instead of the raw feature inputs. The SVM (RBF) model performed best using PCA as inputs because this model achieved the highest area under the curve (AUC) on the test set (i.e. 77.5%). The ANN model achieved the best AUC of 72.5% using only the features recency and monetary value. Lastly, we found the study design of (Darwiche, Feuilloy et al. 2010) better than (Santhanam and Sundaram 2010) and (Sundaram 2011) because their models are assessed on a test (i.e. holdout) set, which provides more realistic performance on future observations. Furthermore, this design allows one to identify if a model has overfit to the data by comparing the testing set statistics to the training set statistics. Zabihi, Ramezan et al. (2011) investigate the use of fuzzy sequential pattern mining to try and predict future blood donating behavior. The features investigated in this study were (1) months since last donation, (2) total number of donations, (3) time (in months) since first donation, and (4) a binary feature indicating whether blood was donated in March 2007 or not. These features are similar in nature to those we investigated in our study. Ramachandran, Girija et al. (2011) investigated the performance of the J48 algorithm provided in Weka3 . The J48 algorithm is an implementation of the C4.5 decision tree written in Java (Wikipedia , Quinlan 1993). They found this methodology to also perform well at predicting blood donors whom had donated before having a sensitivity of 95.2%, but performed poorly at future non-donors. Sharma and Gupta (2012) also used the J48 algorithm in Weka on a different blood donation data set obtained from a blood bank in Kota, Rajasthan, India. While they were attempting to predict the “number” of donors through their age and blood group, they actually performed a classification of donors versus non-donors which raised concerns over the validity of this study. Boonyanusith and Jittamai (2012) performed a blood donation survey in Thailand. Like previous studies they used the J48 decision tree, but also tried an artificial neural network. Both models yielded similar performance with sensitivity (81.7% vs 81.2%) and specificity (53.8% vs. 52.5%).

Bhardwaj, Sharma et al. (2012) provided a very limited review of data mining in blood donation and do not actually train and test any models. They propose to do this in the future research. Likewise Khalid, Syuhada et al. (2013) provide a slightly more extensive review of the literature, but also do not perform any modeling or analysis. Testik, Ozkaya et al. (2012) use the idea of trying to group similar donors based on arrival patterns using Two-Step clustering (SPSS 2001). Then once clusters are formed, CART was implemented on the individual clusters to try to improve predictive accuracy. This approach has been tested in other domains and is an approach we investigate in our study. However, instead of Two-Step clustering we implement models based on more widely known k-Means clustering algorithm. The authors do not report the predictive accuracy of their approach, nor provide a comparison of using Two-Step clustering-CART versus using CART alone. Their primary contribution is the formulation of a serial queuing network model that could be used in the case of blood center operations where arrival patterns could be estimated and used to support workforce size utilization. Ashoori, Alizade et al. (2015) collected census data collected from a blood transfusion center located in Birjand City, North East Iran. This data set consisted of 9,231 donors and measured six features. They tried to predict future blood donors using four types of decision trees (C5.0, CART, CHAID, and QUEST). Their cross-validated models all yielded poor performance ranging from 55 to 57 percent accuracy. One interesting aspect of their results was that the best performing model, the C5.0 tree, had 41 rules compared to only 13 (CHAID), 8 (CART), and 5 (QUEST). With trees the more rules (or splits) used often will lead to overfitting to the data, but can also lead to more distinct probability values in the prediction. Ashoori, Mohammadi et al. (2017) extend research into the performance of these techniques by first using two-step clustering before employing the same decision tree algorithms used in their previous study. They conclude that this approach helped them predict faster and more precisely compared to their previous study.