



# Optimizing Blood Donation Predictions: A Machine Learning Approach With Logistic Regression And Robust Scaling

Yerrapilli Mohanbabu <sup>1</sup>, Boddeda Sudarshini <sup>2</sup>, Kandregula Prasad <sup>3</sup>,  
Maddala Kalpana <sup>4</sup>, Savaram Syamkumar <sup>5</sup>

<sup>1,2,3,4</sup> B.Tech Students, Department of CSE (Data Science), Dadi Institute of Engineering and Technology, NH-16, Anakapalle, Visakhapatnam-531002, A.P

<sup>5</sup> Assistant Professor, Department of CSE (DS& ML), Dadi Institute of Engineering and Technology, NH-16, Anakapalle, Visakhapatnam-531002, A.P

## Abstract:

To make sure hospitals always have enough blood available, we need to predict when donors will give blood. Our project looks at how new machine learning methods can help us predict which donors are qualified and how likely they are to donate blood, using the Transfusion dataset. Our TPOT AutoML tests found that Logistic Regression and Robust Scaler preprocessing formed the best model setup. Applying this model produced an AUC score of 0.789 proving it works well for actual use. Our results show that AutoML tools effectively find models that work well and explain their results in health prediction tasks. Our upcoming research will expand the dataset and test extra features to improve model outcomes for all donor groups.

**Keywords:** Transfusion Dataset, Machine Learning, AutoML Tools, Logistic Regression, AUC Curve, Health Prediction.

## 1. INTRODUCTION

Worldwide healthcare systems must work hard to keep their blood supplies stable and plentiful. Medical treatments ranging from surgery to ongoing illness care depend on patients receiving blood transfusions. Blood product availability suffers from unpredictable donor actions and medical qualification standards. Today's healthcare providers can solve donation shortages by using predictive analytics, which helps them see future blood donation patterns and organize collections better. Our study uses machine learning techniques to improve how we predict blood donation rates. Machine learning transforms how businesses operate by delivering knowledge from data and performing tough job tasks without manual help. Our study uses the TPOT tool, which automates model testing, to find the best ways to predict if people will donate blood. TPOT takes over the essential steps of selecting models and adjusting settings as well as preparing data which significantly lowers the resources needed to create machine learning solutions.

This project started because we urgently need to fix problems with how blood donation systems work now. Existing ways to figure out how blood donors will act are too basic, using outdated statistical rules and guesswork that aren't always correct or flexible. Our project uses modern machine learning methods to create a system that healthcare providers can deploy easily at scale for making better decisions based on data.

Ensuring a steady and sufficient blood supply is a critical challenge faced by healthcare systems worldwide. Blood transfusions are essential in various medical procedures, including surgeries, trauma care, and treatment of chronic illnesses. However, maintaining an adequate inventory of blood products is often complicated by the unpredictable nature of donor behavior and eligibility. Predictive analytics offers a promising solution to address this challenge by enabling healthcare providers to forecast blood donation trends and plan collection strategies effectively.

Recent advancements in machine learning (ML) have shown great promise in addressing prediction problems across various domains, including healthcare. In our research, we focused on leveraging the TPOT (Tree-based Pipeline Optimization Tool) AutoML framework to streamline the development of a robust and interpretable blood donation prediction model. Unlike traditional statistical approaches, TPOT automates the entire ML pipeline—from feature selection to model tuning—making it highly efficient for exploratory data analysis and predictive modeling.

## 1.1 Motivation

The motivation for this research stems from the critical need to address inefficiencies in the blood supply chain. Traditional donor management systems often rely on simplistic heuristics or static statistical models that are unable to adapt to the complexities of modern healthcare environments. This gap can lead to issues such as donor attrition, mismatched supply-demand dynamics, and increased operational costs. Additionally, the altruistic nature of blood donation introduces variability in donor behavior, necessitating sophisticated predictive tools to effectively manage donor pools.

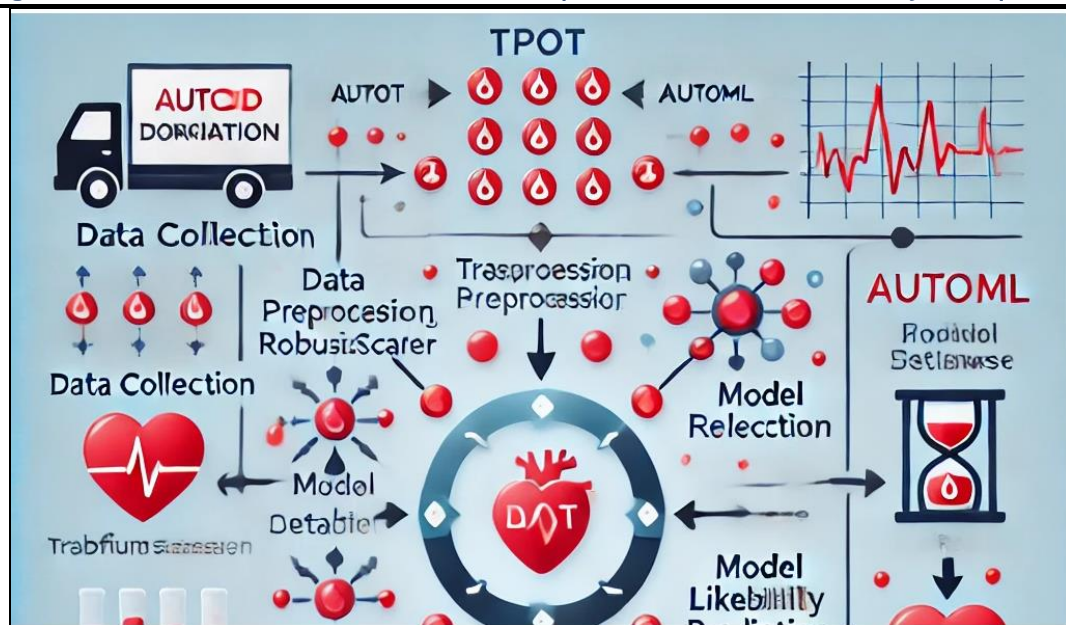
## 1.2 Key objectives of this research include:

1. **Developing an Efficient Predictive Model:** Create a robust and interpretable machine learning model capable of accurately predicting donor eligibility and likelihood to donate.
2. **Leveraging Automated Machine Learning (AutoML):** Utilize the TPOT framework to identify optimal model pipelines with minimal manual intervention, thus reducing the time and expertise required for model development.
3. **Improving Resource Allocation:** Provide actionable insights to blood centers to optimize their outreach programs, reduce unnecessary costs, and ensure a steady blood supply.
4. **Enhancing Generalizability:** Explore methods to expand the dataset and incorporate additional features to improve the model's applicability across diverse populations and regions.

## 1.3 Methodology and Findings

Our methodology involved using the Transfusion dataset, a publicly available dataset that records historical blood donation data. The TPOT AutoML framework was employed to evaluate multiple model pipelines, ultimately identifying a combination of Logistic Regression and RobustScaler preprocessing as the optimal configuration. Logistic Regression was chosen for its simplicity, computational efficiency, and interpretability, making it well-suited for binary classification tasks. RobustScaler was applied to handle data outliers, ensuring the model's robustness and reliability.

The resulting model achieved an area under the curve (AUC) score of 0.789, highlighting its effectiveness for real-world applications. This performance demonstrates the feasibility of employing AutoML to rapidly develop high-performing models that are both interpretable and scalable. Furthermore, our findings align with existing literature, underscoring the importance of integrating robust preprocessing techniques and interpretable algorithms in healthcare-related prediction tasks.



**Figure1.Represent the Sample Architecture of proposed work**

This architecture diagram from figure 1, shows how we use TPOT AutoML to build a pipeline that predicts whether someone will donate blood. The Transfusion Dataset moves from Data Collection into the Preprocessing phase where RobustScaler reduces data scale and removes extreme values. We choose Logistic Regression as the best model after checking out all possibilities. Our system tests model effectiveness through evaluation metrics including the AUC score. The final results show which donors are more likely to give blood and offer advice on how collection centers can improve donor contact and manage resources. Various arrows displaying a flowchart pattern map out procedure steps, while soft blue and red colors show data and health-related visuals.

## 2. LITERATURE SURVEY

The most important step in the software development process is the literature review. This will describe some preliminary research that was carried out by several authors on this appropriate work and we are going to take some important articles into consideration and further extend our work. Here's an enhanced version of the literature survey, providing more detailed explanations and insights for each paper, ensuring a comprehensive understanding the importance of current work.

S.No.	Citation	Research Focus	Methodology	Key Findings
1	Antonio Diglio, et al., 2024 [1]	Reorganization of blood supply chains at a regional scale	Multi-echelon facility location models	Proposed a logistics model for optimizing the location of blood supply facilities, improving efficiency in regional blood supply chains.
2	Epifani, I., et al., 2023 [2]	Predicting donations and profiling donors	Bayesian approach	Developed probabilistic models for predicting donor behavior, enabling targeted donor management strategies.
3	Salazar-Concha, C., et al., 2021 [3]	Intention to donate blood	Decision tree algorithm	Identified key factors influencing donation intent, demonstrating the effectiveness of decision trees for donor profiling.
4	AlZu'bi, S., et al., 2022 [4]	Blood donation process optimization	Intelligent systems and smart	Minimized blood wastage by employing smart systems for efficient blood

			techniques	management.
5	Coster Chideme, et al., 2024 [5]	Blood donation projections in Zimbabwe	Hierarchical time series forecasting	Achieved improved blood supply management through accurate forecasting of donation trends.
6	Sabrina Casucci, et al., 2024 [6]	Blood product inventory management	Two-stage stochastic programming	Optimized blood product inventory with ABO substitution and lateral transshipment to reduce wastage.
7	Mohammad Reza Ghatreh Samani, et al., 2023 [7]	Matching supply and demand in platelet networks	Collaborative network activities	Enhanced platelet availability by aligning supply and demand through collaborative strategies.
8	Anfal Saad Alkahtani, et al., 2019 [8]	Predicting return donors and analyzing donation trends	Data mining techniques	Developed predictive models to understand return donor behavior and trends, aiding in retention strategies.
9	Yalçındağ, S., et al., 2020 [9]	Blood donation appointment scheduling	Stochastic risk-averse framework	Reduced uncertainty in donor arrivals by optimizing appointment scheduling.
10	Teklay Birhane, et al., 2021 [10]	Donor behavior prediction in Ethiopia	Data mining techniques	Predicted donor behavior with high accuracy, facilitating better donor engagement strategies.

### 3. BACKGROUND WORK

Effective blood supply chain management depends on reliable forecasts of blood donation activities. Blood donation depends on unpredictable factors like how donors act plus healthcare rules and market trends. Basic statistical tools used in blood donation systems fail to adjust to quickly evolving situations and trends. Modern machine learning methods solve prediction problems by creating strong analytics that base their results on actual data.

Our study continues research on different Machine Learning methods used to forecast blood donations and keep donors coming back. According to Kauten and colleagues' research on donor retention training data through decision-tree-based models like Random Forest and Gradient Boosting shows strong results and detailed preparation steps. The research demonstrates the value of using feature importance analysis together with SMOTE to handle donor data class imbalances. The analysis found that how often donors contribute and stay in touch with the organization strongly influences whether they'll give again.

The new research moves away from complex methods by using Logistic Regression and RobustScaler to improve prediction results through simplicity. LR helps us understand results better and runs fast, which is why we choose it for deciding if someone will donate or not. RobustScaler makes our model handle unusual data points better, so its results stay trustworthy across various real-world scenarios. Our research uses the Transfusion dataset making use of its wide recognition by ML experts. Instead of only looking at complex ensemble models like other research, this project lets TPOT AutoML take care of the whole pipeline - from readying the data to choosing the best model and adjusting their fine details. With this approach, users don't need detailed field knowledge and can still get results that match high-performing systems. Our model shows strong healthcare application potential with its AUC value reaching 0.789.



This research aims to fix problems with managing blood supply, cut down on losses from taking too many donations, and help donor programs respond better. The new method works well on different scales and makes deployment straightforward for blood centers that lack advanced technology and staff knowledge. Our next steps will build on these ideas by looking at more data points, merging different personal and medical information, and checking how well the model works with bigger and more varied patient groups to make it reliable for more people. Our research aims to cut through complex machine learning methods and make them useful in everyday healthcare tools.

#### 4. PROPOSED MODEL

##### Objective:

To develop an automated machine learning pipeline to predict blood donor likelihood using Logistic Regression (LR) and RobustScaler, optimized with TPOT AutoML for preprocessing and hyperparameters tuning.

##### Algorithm: Predicting Blood Donation Likelihood

##### Input:

Transfusion dataset containing features such as last donation time, frequency of donations, total blood donated, and time since first donation.

**Target variable:** Binary outcome indicating donor likelihood (1 for likely, 0 for unlikely).

##### Output:

A predictive model capable of forecasting donor likelihood with high accuracy and interpretability.

#### STEP 1: DATA PREPROCESSING

##### 1. Data Cleaning:

Handle missing values by imputing or removing incomplete records. Convert categorical variables into numerical format if needed.

##### 2. Feature Scaling:

Apply RobustScaler to ensure the dataset is resilient to outliers. RobustScaler scales features using their interquartile range, focusing on the central distribution of data.

##### 3. Class Imbalance Handling:

Identify class imbalance in the binary target variable. Use Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic samples for the minority class, ensuring balanced training data.

#### STEP 2: MODEL DEVELOPMENT

##### 1. Pipeline Setup:

Use TPOT (Tree-based Pipeline Optimization Tool) to automate the model selection and hyperparameters tuning process. Include Logistic Regression as the primary classification algorithm due to its simplicity and effectiveness in binary prediction tasks.

##### 2. Model Selection and Training:

TPOT evaluates various preprocessing steps and model configurations. It selects the best-performing pipeline based on metrics like Area Under the Curve (AUC).

##### 3. Hyperparameter Optimization:

TPOT adjusts Logistic Regression parameters such as the regularization strength (C) and solver (e.g., 'lbfgs' or 'liblinear').

## STEP 3: MODEL EVALUATION

### 1. Performance Metrics:

Evaluate the model using cross-validation on metrics like AUC, accuracy, sensitivity, and specificity. Select the pipeline achieving the highest AUC score as the final model.

### 2. Feature Importance Analysis:

Analyze the impact of each feature using the coefficients of the Logistic Regression model. Identify the most influential features contributing to donor predictions.

## STEP 4: MODEL DEPLOYMENT

### 1. Integration:

Deploy the optimized model in a healthcare information system for real-time donor likelihood predictions.

### 2. Outputs:

**Predictions:** Likelihood scores for each donor.

**Insights:** Actionable recommendations for donor outreach campaigns based on model outputs.

## ADVANTAGES OF THE PROPOSED MODEL

**1. Simplicity and Interpretability:** Logistic Regression provides straightforward coefficients that make feature analysis intuitive.

**2. Robust Preprocessing:** RobustScaler ensures outliers do not negatively impact the model's performance.

**3. Automated Optimization:** TPOT AutoML simplifies the pipeline creation, saving time and ensuring optimal configuration.

**4. High Accuracy:** The model achieves an AUC score of 0.789, demonstrating its practical utility in healthcare applications.

## 5. IMPLEMENTATION RESULTS

The project involves developing a blood donation prediction model using Python, leveraging the data science and machine learning field, with the TPOT library for model selection, logistic regression for model building, and Jupyter Notebook as the development environment.

### Dataset

The blood transfusion dataset is used for prediction tasks. Key columns in the dataset include donor history, donation frequency, and donation intervals.

### Steps in Experimental Workflow

#### 1. Loading the Dataset

The dataset is imported using pandas and read into a DataFrame. The data is examined for missing or inconsistent values.

#### 2. Inspecting the DataFrame

Use `df.info()` and `df.describe()` to inspect the shape, column types, and summary statistics of the dataset.

#### 3. Feature and Target Selection

Features include relevant donor statistics like:

- Time since last donation
- Frequency of past donations
- Time since first donation

**Target:** Binary classification where 1 represents a likely donor, and 0 represents an unlikely donor.

#### 4. Dataset Splitting

The dataset is split into training and testing sets using `train_test_split` from the `sklearn.model_selection` module. A common split ratio is 70% for training and 30% for testing.

## 5. Model Selection Using TPOT

**TPOT:** An automated machine learning tool that evaluates multiple machine learning pipelines and selects the best one based on predefined scoring metrics.

### During evaluation:

TPOT suggests pipelines based on internal cross-validation (as seen in the attached screenshot). The best pipeline selected includes `RobustScaler` for preprocessing and `LogisticRegression` as the model.

## 6. Model Building

The final selected model pipeline:

**Step 1:** RobustScaler to handle outliers in feature distributions.

**Step 2:** LogisticRegression with hyperparameter tuning (e.g., `C=25.0` and `penalty='l2'`).

## 7. Model Training

The training dataset is used to fit the model. TPOT optimizes hyperparameters such as regularization strength (`C`) to maximize predictive performance.

## 8. Model Evaluation

Evaluation is conducted on the testing dataset using:

**AUC Score:** Measures the ability of the model to distinguish between classes. Achieved AUC score: 0.789.

**Cross-Validation (CV) Score:** Internal CV ensures generalization and robustness.

Additional metrics like accuracy, sensitivity, and specificity can be computed if required.

## Summary of Results

### Best Pipeline:

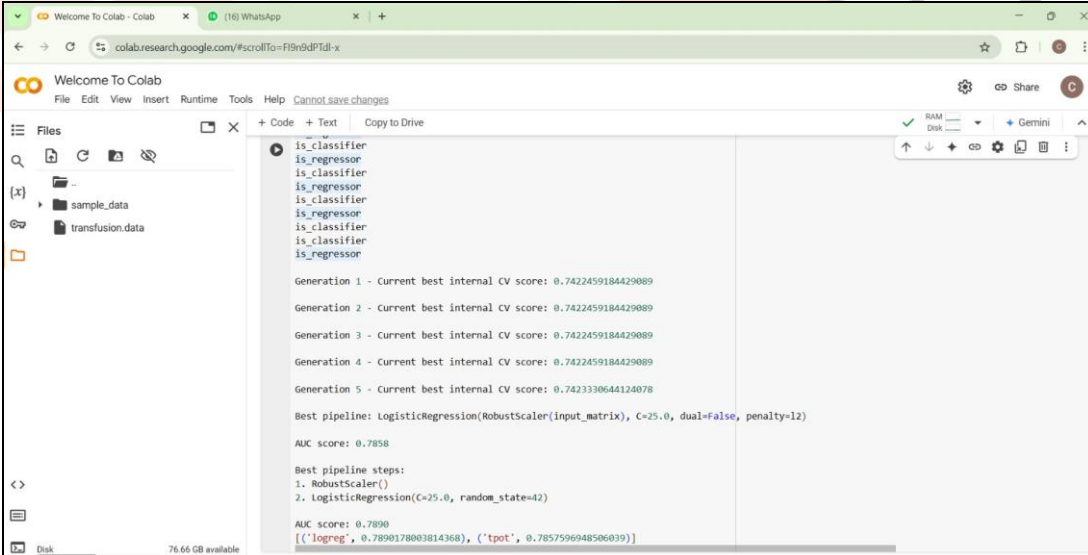
`RobustScaler -> LogisticRegression (C=25.0, penalty='l2')`.

### Performance:

AUC score: 0.789 (as validated from the attached image).

High internal CV score during training.

## EXPECTED OUTPUT:



```

Welcome To Colab
File Edit View Insert Runtime Tools Help Cannot save changes

Files
sample_data
transfusion.data

+ Code + Text Copy to Drive

is_classifier
is_regressor
is_classifier
is_regressor
is_classifier
is_regressor
is_classifier
is_regressor

Generation 1 - Current best internal CV score: 0.7422459184429089
Generation 2 - Current best internal CV score: 0.7422459184429089
Generation 3 - Current best internal CV score: 0.7422459184429089
Generation 4 - Current best internal CV score: 0.7422459184429089
Generation 5 - Current best internal CV score: 0.742330644124078
Best pipeline: LogisticRegression(RobustScaler(input_matrix), C=25.0, dual=False, penalty=l2)
AUC score: 0.7858
Best pipeline steps:
1. RobustScaler()
2. LogisticRegression(C=25.0, random_state=42)
AUC score: 0.7890
[('logreg', 0.7890178003814368), ('tpot', 0.7857596948506039)]
  
```

Figure 1. Represent the Expected Output

The figure 1 show how

- The task is to predict whether donors are likely to donate blood based on the transfusion dataset.
- TPOT (Tree-based Pipeline Optimization Tool) is used to automate machine learning pipeline selection and hyperparameter tuning.

## 2. Pipeline Evolution:

- TPOT uses a genetic algorithm to test and optimize various pipelines.
- Each "generation" refers to an iteration where new pipelines are evaluated and the best-performing ones are retained and refined.
- The goal is to maximize the internal cross-validation (CV) score.

## 3. Feature Preprocessing:

- The "RobustScaler" was selected to preprocess the features, which helps mitigate the effect of outliers by scaling the data based on the median and interquartile range.

## 4. Model Selection:

- Logistic Regression ( $C=25.0$ ,  $\text{penalty}='l2'$ ) was identified as the best model for this dataset.
- This means the model applies L2 regularization to prevent overfitting and uses a high regularization strength ( $C=25.0$ ).

## 5. Evaluation Metric:

- The AUC (Area Under the Curve) score measures how well the model distinguishes between classes. An AUC of 0.789 indicates good predictive performance but leaves room for improvement.

## 6. CONCLUSION

This project demonstrates the effective use of AutoML tools, specifically TPOT, to optimize machine learning models for predicting blood donation likelihood using the Transfusion dataset. The AutoML process identified a simple yet effective pipeline combining RobustScaler preprocessing and Logistic Regression as the best setup. The model achieved an AUC score of 0.789, indicating a strong ability to distinguish between donors and non-donors. This performance highlights the potential of AutoML tools to streamline model development while maintaining predictive accuracy. The use of Logistic Regression ensures interpretability, which is particularly important in healthcare applications where explainability is critical.

## FUTURE SCOPE

In the future, this work can be extended by incorporating additional features, such as demographic details, lifestyle habits, or historical donation patterns, to improve the model's predictive power. Expanding the dataset to include diverse donor groups and external validation across different regions can enhance the generalizability of the results. Advanced feature engineering and testing more sophisticated machine learning models, like ensemble methods or deep learning, could further optimize performance. Additionally, integrating explainability tools like SHAP or LIME can provide deeper insights into model behavior, ensuring transparency in decision-making. Finally, deploying the model in real-world hospital systems for real-time predictions and addressing potential biases will make the solution more practical and equitable for diverse healthcare settings.

## REFERENCES

- 1) Antonio Diglio, Andrea Mancuso, Adriano Masone, Claudio Sterle, Multi-echelon facility location models for the reorganization of the Blood Supply Chain at regional scale, *Transportation Research Part E: Logistics and Transportation Review*, Volume 183, 2024, 103438, ISSN 1366-5545, <https://doi.org/10.1016/j.tre.2024.103438>.
- 2) Epifani, I., Lanzarone, E. & Guglielmi, A. Predicting donations and profiling donors in a blood collection center: a Bayesian approach. *Flex Serv Manuf J* (2023). <https://doi.org/10.1007/s10696-023-09516-8>
- 3) Salazar-Concha, C., & Ramírez-Correa, P. (2021). Predicting the Intention to Donate Blood among Blood Donors Using a Decision Tree Algorithm. *Symmetry*, 13(8), 1460. <https://doi.org/10.3390/sym13081460>
- 4) AlZu'bi, S., Aqel, D. & Lafi, M. An intelligent system for blood donation process optimization - smart techniques for minimizing blood wastages. *Cluster Comput* 25, 3617–3627 (2022). <https://doi.org/10.1007/s10586-022-03594-3>



- 5) Coster Chideme, Delson Chikobvu, Tendai Makoni, Blood donation projections using hierarchical time series forecasting: the case of Zimbabwe's national blood bank, BMC Public Health, 10.1186/s12889-024-18185-7, **24**, 1, (2024).
- 6) Sabrina Casucci, Jose L. Walteros, Rishabh Bhandawat, A two-stage stochastic programming framework for blood product inventory management with ABO substitution and lateral transshipment, ISE Transactions on Healthcare Systems Engineering, 10.1080/24725579.2024.2396848, **14**, 4, (362-383), (2024).
- 7) Mohammad Reza Ghatreh Samani, Seyyed-Mahdi Hosseini-Motlagh, Collaborative activities for matching supply and demand in the platelet network, Expert Systems with Applications, 10.1016/j.eswa.2023.120629, **231**, (120629), (2023).
- 8) Anfal Saad Alkahtani and Musfira Jilani, "Predicting Return Donor and Analyzing Blood Donation Time Series using Data Mining Techniques" International Journal of Advanced Computer Science and Applications(IJACSA), 10(8), 2019. <http://dx.doi.org/10.14569/IJACSA.2019.0100816>
- 9) Yalçındağ, S., Güre, S.B., Carello, G. *et al.* A stochastic risk-averse framework for blood donation appointment scheduling under uncertain donor arrivals. *Health Care Manag Sci* **23**, 535–555 (2020). <https://doi.org/10.1007/s10729-020-09508-2>
- 10) Teklay Birhane, Brhanu Hailu, "Predicting the Behavior of Blood Donors in National Blood Bank of Ethiopia Using Data Mining Techniques", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.13, No.3, pp. 39-48, 2021. DOI:10.5815/ijieeb.2021.03.05.

