JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

A RESEARCH ON BLOOD GROUP PREDICTION USING MACHINE LEARNING ALGORITHMS

Ms. K. Rajeswari[1], Ms. Mamidi Hanvee Reddy[2], Ms. N Leela Maanasa[3]

[1]Assistant Professor, Department of Computer Science,

[2,3] Students, Department of Computer Science (MSDS)

[1, 2, 3] St. Ann's College for Women, Mehdipatnam, Hyderabad.

Abstract

Blood is a fluid that transports oxygen and nutrients to the cells and carries away carbon dioxide and other waste products. There are various types of blood groups which a person may inherit from the parents. This blood group type is usually determined using the Blood Group Typing test. There is an alternative way wherein a person's blood group is predicted using the parents' blood group. We have developed a machine learning model based on this idea. We used SGDClassifier, RandomForestClassifier, LogisticRegression, KNeighborsClassifier, GaussianNB, Perceptron, LinearSVC and DecisionTreeClassifier algorithms. On analyzing the performance of these algorithms we found that RandomForestClassifier and DecisionTreeClassifier gave us the maximum accuracy.

Keywords: **Blood Group, Random Forest Classifier, Machine Learning, Decision Tree Classifier**

1. INTRODUCTION

Antigens are molecules. They can be either proteins or sugars. The types and features of antigens can vary between individuals, due to small genetic differences. Antigens and antibodies play a role in the immune system's defense mechanism. Blood is classified into groups based on two types of antigens:

i) ABO antigens

ii) Rh antigens

Human blood type is determined by codominant alleles. An allele is one of several different forms of genetic information that is present in our DNA at a specific location on a specific chromosome. There are three different alleles for human blood type, known as IA, IB, and i. For simplicity, we can call these alleles A (for IA), B (for IB), and O (for i).

Each of us has two ABO blood type alleles, because we each inherit one blood type allele from our biological mother and one from our biological father. A description of the pair of alleles in our DNA is called the genotype. Since there are three different alleles, there are a total of six different genotypes at the human ABO genetic locus. The different possible genotypes are AA, AO, BB, BO, AB, and OO.

A blood test is used to determine whether the A and/or B characteristics are present in a blood sample. It is not possible to determine the exact genotype from a blood test result of either type A or type B. If someone has blood type A, they must have at least one copy of the A allele, but they could have two copies. Their genotype is either AA or AO. Similarly, someone who is blood type B could have a genotype of either BB or BO.

Genotype (DNA)	Blood Type
AO or AA	A blood type
AB	AB blood type
BO or BB	B blood type
OO	O blood

Determination of Blood Group

Each biological parent donates one of their two ABO alleles to their child. A mother who is blood type O can only pass an O allele to her son or daughter. A father who is blood type AB could pass either an A or a B allele to his son or daughter. This couple could have children of either blood type A (O from mother and A from father) or blood type B (O from mother and B from father).

Since there are 4 different maternal blood types and 4 different paternal blood types possible, there are 16 different combinations to consider when predicting the blood type of children. In the tables below, all 16 possible combinations are shown. If you know the blood type of the mother and father, the possible blood types for their children can be found.

The Rh factor genetic information is also inherited from our parents, but it is inherited independently of the ABO blood type alleles. There are 2 different alleles for the Rh factor known as Rh+ and Rh-. Someone who is "Rh positive" or "Rh+" has at least one Rh+ allele, but could have two. Their genotype could be either Rh+/Rh+ or Rh+/Rh-. Someone who Rh- has a genotype of Rh-/Rh-.

Just like the ABO alleles, each biological parent donates one of their two Rh alleles to their child. A mother who is Rh- can only pass an Rh- allele to her son or daughter. A father who is Rh+ could pass either an Rh+ or Rh- allele to his son or daughter. This couple could have Rh+ children (Rh- from mother and Rh+ from father) or Rh- children (Rh- from mother and Rh- from father).

PARENT 1		AB	AB	AB	AB	B	A	A	O	O	O
PARENT 2		AB	B	A	O	B	B	A	B	A	O
Possible Blood Type of Child	O					🩸	🩸	🩸	🩸	🩸	🩸
	A	🩸	🩸	🩸	🩸		🩸	🩸		🩸	
	B	🩸	🩸	🩸	🩸	🩸	🩸		🩸		
	AB	🩸	🩸	🩸			🩸				

Figure i. Blood Group Typing

METHODOLOGY

A. Data Collection:

We collected primary data through a Google form and received 534 responses.

```
[7]  data.info()

     <class 'pandas.core.frame.DataFrame'>
     RangeIndex: 534 entries, 0 to 533
     Data columns (total 7 columns):
      #   Column                            Non-Null Count  Dtype
     ---  ------                            --------------  -----
      0   Timestamp                         534 non-null    object
      1   Username                          534 non-null    object
      2   Blood_Group                       534 non-null    object
      3   Mother's_Blood_Group              534 non-null    object
      4   Father's_Blood_Group              534 non-null    object
      5   Do you have any sibling/siblings? 534 non-null    object
      6   Sibling/siblings' Blood Group     499 non-null    object
     dtypes: object(7)
     memory usage: 29.3+ KB
```

Figure ii. Dataset information

B. Data preprocessing:

Visual representation of value counts

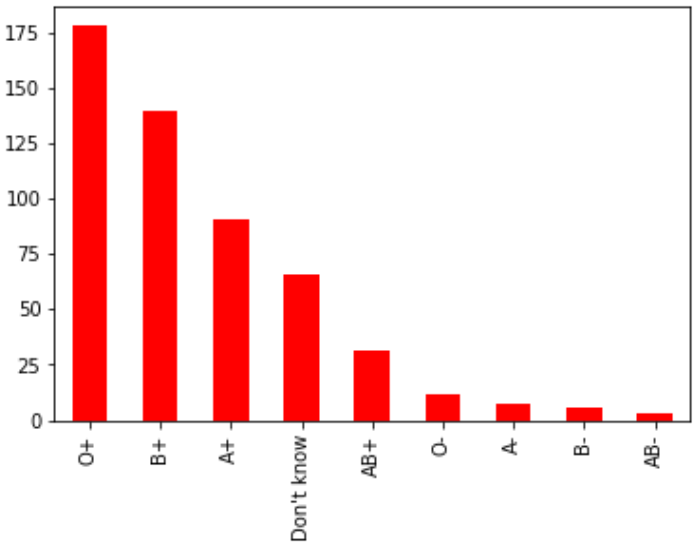


Figure iii. Blood group

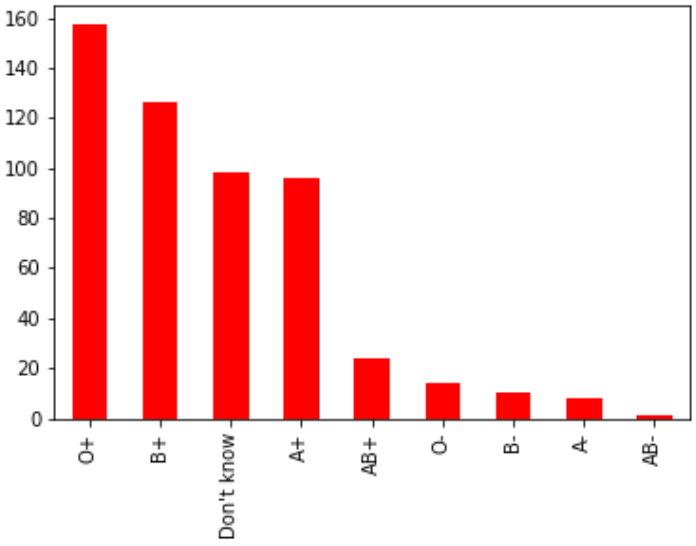


Figure iv. Mother's Blood group

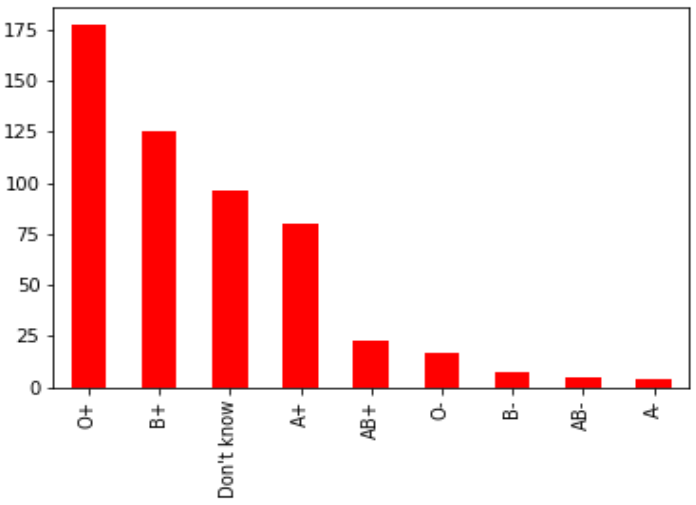


Figure v. Father's Blood group

Step 1: Dropping the unnecessary columns using drop method

Step 2: Removing Null Values using .loc function

Step 3: Encoding the Dataset, i.e, assigning numeric values to categorical data using the LabelEncoder method

```
data.head()
```

	Blood_Group	Mother's_Blood_Group	Father's_Blood_Group
0	4	6	4
2	6	6	4
3	4	5	0
4	6	6	7
6	0	0	0

Figure vi. Preprocessed data

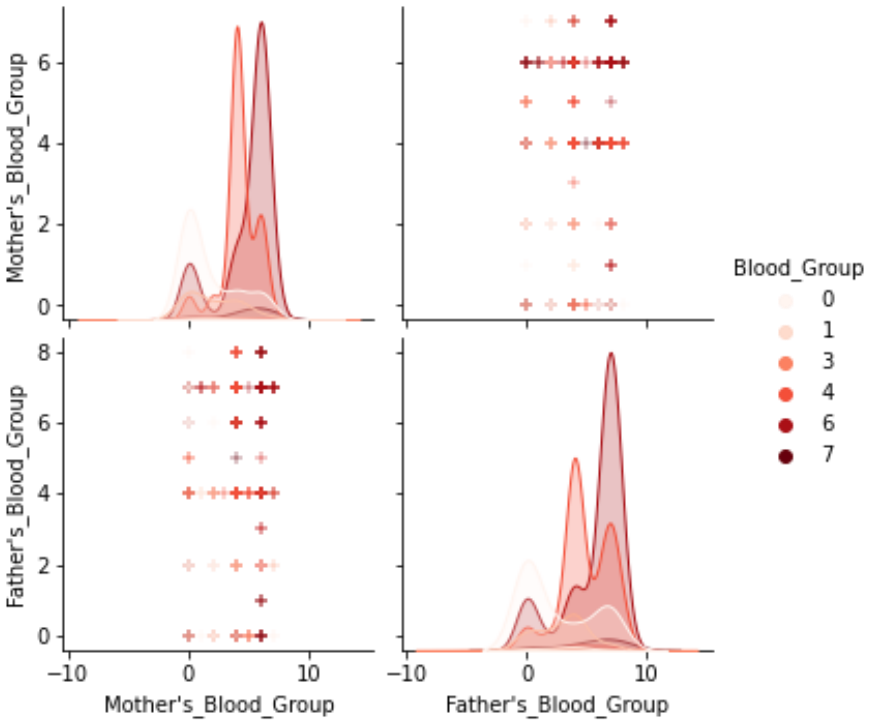


Figure vii. Pairplot showing relationship b/w mother's and father's blood group

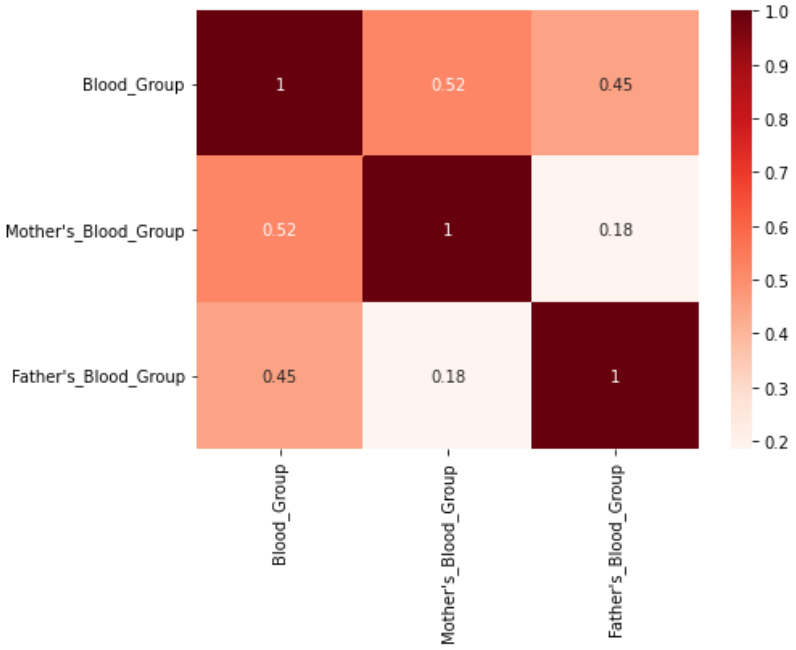


Figure vii. Heatmap showing correlation b/w mothers, fathers and child's blood group

C. Algorithm testing:

Step1: Splitting of data into train and test sets.

Step2: Feeding data into the algorithms.

The algorithms used in for our model:

Stochastic Gradient descent

It is an optimization algorithm often used in machine learning applications to find the model parameters that correspond to the best fit between predicted and actual outputs.

Random Forest

It is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

K Nearest Neighbor

K-NN is one of the most simple and traditional nonparametric techniques to classify samples. It is a type of supervised learning algorithm used for both regression and classification. KNN tries to predict the correct class for the test data by calculating the approximate distance between the different data points on the input vectors and then assigns the unlabelled point to the class of its K-nearest neighbors.

Gaussian Naive Bayes

It is used when each feature is to be distributed according to Gaussian distribution. We can use this formula to compute the probability of likelihoods if our data is continuous.

Perceptron

The Perceptron algorithm learns the weights for the input signals in order to draw a linear decision boundary. This enables you to distinguish between the two linearly separable classes +1 and -1.

Decision Tree

A decision tree classifies a sample through a sequence of decisions in which the current decision helps to make the subsequent decision. Attributes of the samples will be assigned to each node and the branches will hold the corresponding values. Decision Trees are a Non-parametric supervised learning method used for both classification and regression tasks. Tree models where the target variable can take a discrete symbolic class label is called classification tree, whereas a decision tree with a range of continuous numeric values is called a regression tree.

Logistic Regression

It is a Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Support Vector Machine

It is a supervised machine learning algorithm capable of performing classification, regression and even outlier detection. The linear SVM classifier works by drawing a straight line between two classes.

RESULT ANALYSIS

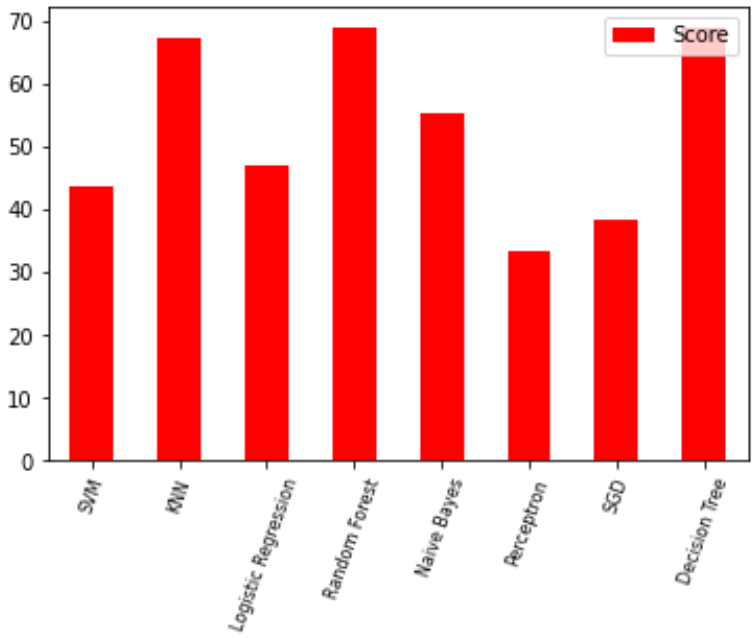


Figure viii. Algorithm accuracy comparison

Score	Model
68.92	Random Forest
68.92	Decision Tree
67.38	KNN
55.38	Naive Bayes
47.08	Logistic Regression
43.69	Support Vector Machines
43.38	Stochastic Gradient Decent
33.23	Perceptron

Figure ix. Algorithm accuracy table

We analysed the performance of these algorithms and found that RandomForestClassifier and DecisionTreeClassifier gave us the maximum accuracy. Since Random Forest gave us the maximum accuracy we calculated the mean, standard deviation and cross validation scores.

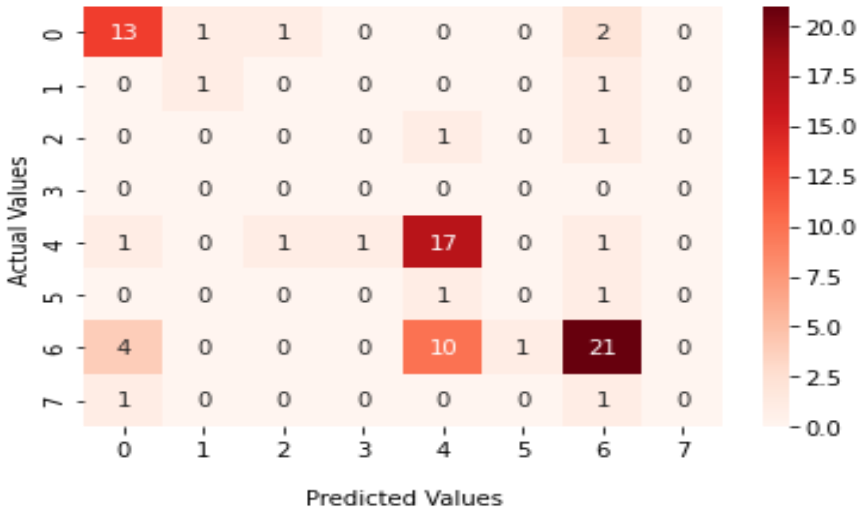
```
Scores: [0.75757576 0.57575758 0.60606061 0.54545455 0.54545455 0.625
 0.71875    0.6875     0.65625    0.65625    ]
Mean: 0.6374053030303031
Standard Deviation: 0.06791489470638822
```

Figure x. Mean, Standard deviation & Cross validation scores

```
y_prediction
```

```
array([6, 1, 6, 4, 4, 1, 4, 6, 4, 6, 0, 2, 6, 4, 6, 4, 0, 6, 6, 2, 6, 4,
       6, 4, 0, 4, 6, 6, 6, 6, 0, 0, 4, 6, 0, 5, 6, 4, 6, 4, 4, 4, 0, 4,
       6, 0, 4, 4, 0, 6, 0, 3, 4, 4, 4, 0, 4, 6, 4, 6, 4, 6, 4, 0, 4, 4,
       4, 4, 6, 4, 4, 6, 6, 6, 6, 0, 4, 4, 0, 6, 0, 0])
```

Figure xi. Array of predicted values



xii. Heatmap showing correlation b/w predicted and actual values

CONCLUSION

This model enables us to predict the blood group of a person using parents' blood group without going through the traditional laboratory testing. Based on our research, RandomForestClassifier and DecisionTreeClassifier gave us the maximum accuracy. Although we had a small dataset which was not very diverse, we got an accuracy of 68.92. Using a larger and more diverse dataset we would get higher accuracy. And enable the model to work better.

REFERENCES:

[1] *https://www.britannica.com/science/blood-biochemistry*

[2] *https://www.nhs.uk/conditions/blood-groups*

[3] *Human Genetics: ABO Blood Group.*