Implementing Decision Trees With Metaheuristic Optimization Search Algorithm For Learner's Performance Prediction

Jayanth H. N.[1], Sharath Kumar D.A.[2]
[1, 2] Dept of Computer Science and Engineering
[1, 2] New Horizon College of Engineering, Bangalore, Karnataka, INDIA

Abstract- *The main objective of education institution these days is to refine their student's performance in overall grind. One way to achieve the zenith for these is to identify factors including curricular and extracurricular activities. The specific objective of the proposed work is to find out if there are any patterns in the available data that could be useful for predicting a learner's performance such as number of absences, parent's job and education. We have used CART algorithm for the construction of our initial population of decision trees. The trees are taken as both combined as well as five different trees for next step. Genetic algorithm is a term used for survival of the fitter off-spring or generation. However, its performance might get deeply affected by the choice of algorithm used in its each step. A thorough comparison of different techniques of crossovers has been carried out. These steps yield a tree with the highest and the most efficient system. The use of an appropriate algorithm is essential and hence the results of these techniques have been compared to find the most optimum system.*

I. INTRODUCTION

Rightful evaluation of students is a classic problem that has been around in the industry for years. And several attempts have been made to curb the evaluation process without a lot of success. But with the ample amount of data available these days we are trying to leverage machine learning and genetic algorithms to detect patterns as well as create a healthy relationship. Although the problem of handling student performance has been around for a very long time, there isn't a definite system in place to address the issue. The automated systems present use a single algorithm for prediction. The proposed model automates the filtering process and produces a more accurate prediction system that will enable institutions to focus their resources on students who are valuable and are at a higher risk of failing. It will also help them prioritize students based on the classes. The system also arms the institution with possible reasons why a student might be failing. The proposed system uses ensemble methods(CART Algorithm) for prediction. The model is trained on the labeled data and is stored; accuracy is further improved by encoding the trees and using crossover methods on them.

The proposed framework predicting the performance of a learner is considered one of the most important topics these days. These topics are used by schools or universities since it helps to find any hidden pattern to improve the student's academic result. We are proposing an algorithm for predicting a learner's performance using decision trees and genetic algorithms. The genetic algorithm is applied to the initial population of decision trees created. An appropriate algorithm in each step is chosen, two crossovers are implemented to compare the final results. These steps yield a tree with the highest and the most efficient system. The main objective is to provide its students with quality education. One way to achieve the highest quality standard is to recognize factors that influence academic performance and then try to overcome the deficiency of those factors. The basic reason for the proposed algorithm is to figure out if there are trends in the available data. The proposed model covered the future scope and the demand for this product in the market. Also, how this product can help large institutions monitor student's performance. It even marked the main features of the device which helps in overcoming the drawbacks of the existing system. We review previous approaches, compare them with our model and search for any further potential.

II. METHODOLGOY

The proposed method as follows

2.1. Software Requirements Specifications

The requirement specification is the movement of interpreting the data assembled amid an investigation into a prerequisite report. Software requirements specifications are the detailed enlisting of all necessary requirements that arise in the project. The aim of having these requirements is to gain an idea of how the project is to be implemented and what is to be expected as a result of the project.

Operating Environment

This section gives a brief about the hardware and software prerequisites for the project.

Hardware Requirements

- Processor: 1.6GHz or faster processor
- RAM: 1 GB(32 bit) or 2GB(64 bit)
- Storage: 250GB of available hard disk space
- Other general hardware such as a mouse and keyboard for inputs and a monitor for display.

Software Requirements

- Operating system: Windows 10
- Programming languages: Python,JavaScript, HTML, CSS
- Documentation: Overleaf

2.1.2Functional Requirements

Functional requirements are a formal way of expressing the expected services of a project. We have identified the functional requirements for our project as follows:

- The system should be able to gather data from the user.
- The system should be able to check for the correctness of data entered by the user.
- The system should be able to predict correctly if the employee is at risk of leaving or
- not.
- The system should be able to have the capacity to decide the contribution of each attribute towards the decision made by the predictor.

2.1.3 Non-Functional Requirements

Non-functional requirements are the various capabilities offered by the system. These have nothing to do with the expected results, but focus on how well the results like Usability, Reliability, Security, Performance, Portability and Re-usability.

2.1.4 User Characteristics

There are basically three types of users like School Admin Staff, Students and Advisors associated with the system.The model could be trained on data for different sectors and used to predict student behavior among different classes. This tool can be used to plan better strategies and guide the student which in turn leads to better productivity.

2.2. High Level Design

This section mainly covers the design technique of the entire system which involves the implementation of 2 modules which are as follows:

- Training the machine learning model and saving it
- Using the trained model as a web service for prediction

2.2.1 Design Approach

Here are two methodologies for software designing:

Top-down Design: It takes the entire programming framework as one entity and after that disintegrates it to accomplish in excess of one subsystem or some components based on a few attributes. Bottom-up Design: The model begins with the most particular and essential components. It accedes with making a more elevated amount out of subsystems by utilizing essential or lower level components. We used a bottom-up design strategy in this product design phase as we start designing the basic components in each module and finally we interlink both the modules to get the final product.

2.2.1 System Architecture

The system architecture diagram is a design whose primary function is to outline the whole framework by distinguishing the primary segments that would be created for the objects along with its interfaces. The proposed algorithm shows us how the process of natural selection is done where the fittest individuals are selected for generation of new population by producing offspring of the next generation.

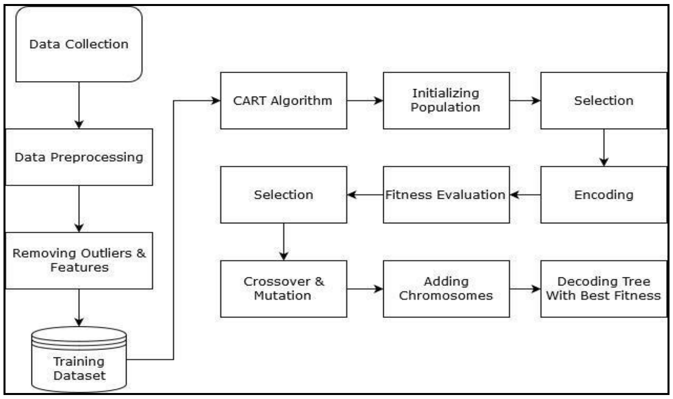


Figure1. System Architecture

2.2.2 Data Flow Diagram

A data-flow diagram is a type of representing a process of data flow or a system. Along with the representation part it also conveys information about the outputs and inputs of each entity and the process itself.

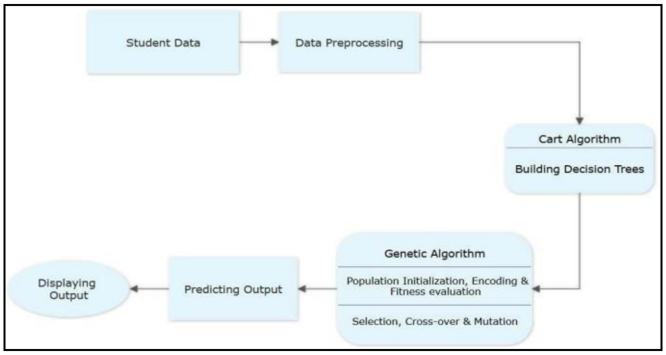


Figure 2 Data Flow Diagram

2.2.3 Sequence Diagram

A sequence diagram is the depiction of interactions of components among each other in order and also interactions between the objects in the system with the time order that particular interaction takes place.

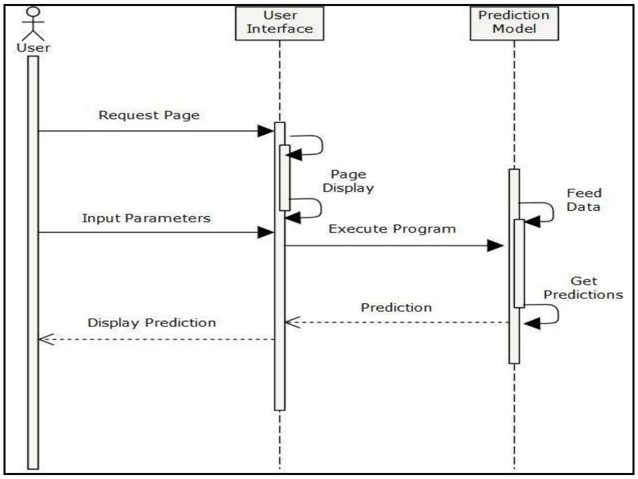


Figure 3 Sequence Diagram

2.3 Detailed Design

The purpose of the detailed design is to plan our system to meet the requirements specified at the start. In the detailed design we see what the input data for each model is, how the model implementation is carried out and how the output is interpreted. The basic purpose of the project is to design a HR tool that can predict the chances of a particular person leaving the company.

2.3.1 Module 1: Data Pre-processing

Consider the following flowchart

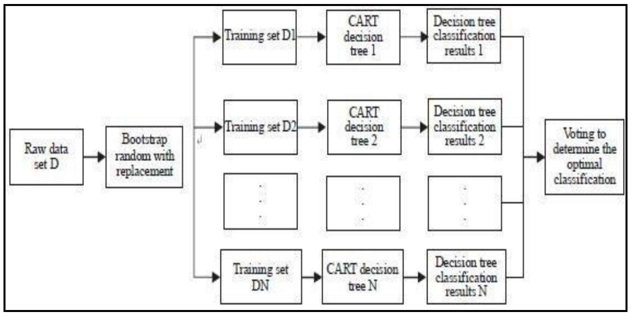

Raw data set D
Bootstrap random with replacement
Training set D1
CART decision tree 1
Decision tree classification results 1
Training set D2
CART decision tree 2
Decision tree classification results 2
Training set DN
CART decision tree N
Decision tree classification results N
Voting to determine the optimal classification

Figure 4. Detailed Design

2.3.2 Classification

Classification is the method of collecting entities with the similar attribute under one class label. This is essential because it validates our hypothesis on the efficiency of our model. Classification in our case is a multi classification showing the overall attrition of the company by predicting

2.3.3 Student Dataset

After data pre-processing the following 5 tables are constructed: Study Patterns, Family Background, Previous academic performance, Extracurricular Activities and General factors

2.3.4 CART Algorithm

The CART algorithm consists of a set of if-else sentences. The main elements of CART algorithm are:

- Data is split at a node based on one variable;
- When a branch reaches terminal node, stopping criteria is applied;
- Finally, target variable is obtained through prediction.

2.3.5 Encoding

Each individual in the tree is traversed in the breadth-first order, which is then stored in two arrays. One of the arrays consists of the integer as well as null values. The other consists of threshold values corresponding to each node.

2.3.6 Fitness Evaluation

The fitness function is defined as a model taking input and producing corresponding output. Input parameters from the candidates are taken and solution of how “good” or how “fit” the current generation is provided. Fitness values of different crossovers are compared for the selection of the fittest.

$$f(x)= \frac{\text{Number of samples correctly predicted}}{\text{Total number of samples}} \quad {*100}$$

In our project, fitness function returns the accuracy of the chromosome with which it predicts the output for test data.

2.3.7 Selection

This phase of the genetic algorithm selects individual solutions from the current population to produce a new generation of the population by applying the crossover between the selected individuals. In our project, we have used elitism selection strategy.

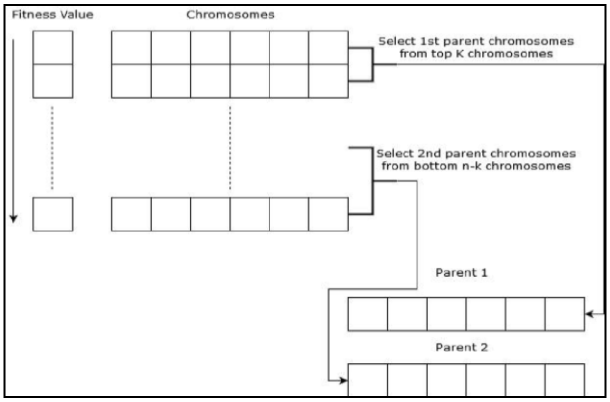

Fitness Value
Chromosomes
Select 1st parent chromosomes from top K chromosomes
Select 2nd parent chromosomes from bottom n-k chromosomes
Parent 1
Parent 2

Figure 5 Elitism

2.3.8 Crossover

Crossover in the genetic algorithm performs a similar functions that crossover performs in a biological process. It takes two chromosomes and produces new chromosomes. In our project we have implemented two types of crossover: Single point crossover and Dominance based crossover.

2.3.8 Mutation

We have implemented mutation in our project by randomly selecting an equal number of attribute and class nodes and alter their value. For attributes nodes, we have either increased or decreased the threshold value for that particular node by 0.5 which is selected randomly. Similarly for class nodes the one of the possible class value which is also selected randomly

2.4 Implementation

Implementation is the phase of the undertaking when the hypothetical plan is transformed out into a working framework. This proposal frame work dealt with the various techniques used in the development of the project, starting with the language and platform selection to finally explain the entire process of implementation steps.

2.5 Testing

Software testing is a process to show the customers the quality of the product. Unit testing is a method used to test individual modules. Integration testing aggregates all modules that are unit tested and tests them using Integration testing methods. System testing basically tests the system to check whether the system meets all the specified requirements.

III. RESULTS

Graphical User Interface:

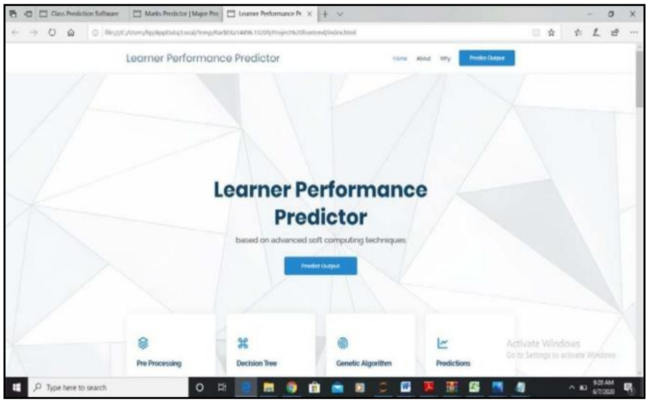
A web based user interface was implemented for this project. The Web UI is a HTML-based application used to design and deal with the server apparatus from a remote client. The Website provides a clear description of all the work done in the project. Users will feed the data to the trained model and the results of the prediction are presented on the webpage. Following are the technologies used for implementing the UI: Bootstrap, HTML and CSS, Java script and Flask.

Bootstrap is used for front-end development, it contains HTML and CSS based plan layouts for typography, frames, catches, route and other interface components, as well as Java script extensions. Following are the advantages of using Bootstrap: It already has predefined design templates and classes, which saves a lot of time, all bootstrap components share a consistent design throughout and it is easy to use and compatible with all browsers.

Flask is a miniaturized scale web structure written in Python and in light of the Werkzeug toolbox and jinja2 layout engine. Flask is considered and used for the following advantages: It's easy to set up, it's well documented, it's very simple and minimalist, and doesn't include anything you won't use and it's flexible enough that you add extensions,if you need more functionality.

Jinja is a layout engine for the python programming dialect and is authorized under a BSD License. It gives Python like articulations while guaranteeing that the layouts are assessed in a sandbox. It is content based format dialect and along these lines can be utilized to produce any increase and in addition source Code.

As the below figure 6 is the Landing Page of our Project.


Class Prediction Software
Marks Predictor | Major Pro
Learner Performance Pr
Learner Performance Predictor
Home
About
Why
Predict Output
Activate Windows
Go to Settings to activate Windows
Type here to search
9:20 AM
6/7/2020

Learner Performance Predictor

based on advanced soft computing techniques

Predict Output

Pre Processing

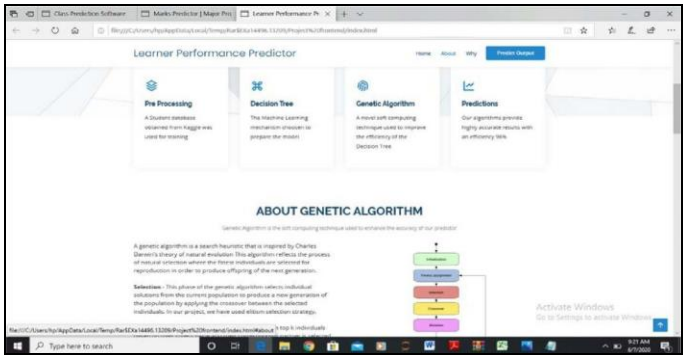
Decision Tree

Genetic Algorithm

Predictions

Figure 6. Landing Page

Figure 7 below displays the details about our system.


Class Prediction Software
Marks Predictor | Major Pro
Learner Performance Pr
Learner Performance Predictor
Home
About
Why
Predict Output
Activate Windows
Go to Settings to activate Windows
Type here to search
9:21 AM
6/7/2020

Pre Processing

A Student database obtained from Kaggle was used for training

Decision Tree

The Machine Learning mechanism choosen to prepare the model

Genetic Algorithm

A novel soft computing technique used to improve the efficiency of the Decision Tree

Predictions

Our algorithms provide highly accurate results with an efficiency 96%

ABOUT GENETIC ALGORITHM

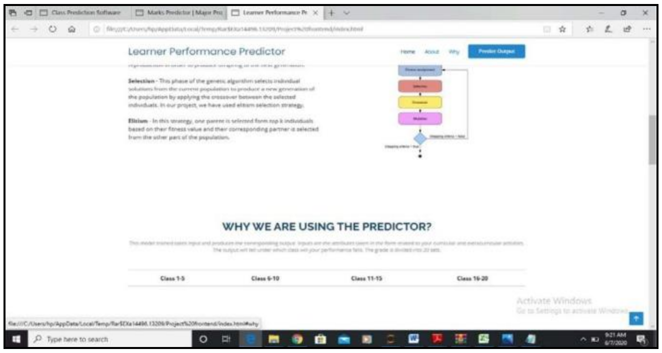
Genetic Algorithm is the soft computing technique used to enhance the accuracy of our predictor

A genetic algorithm is a search heuristic that is inspired by Charles Darwin's theory of natural evolution This algorithm reflects the process of natural selection where the fittest individuals are selected for reproduction in order to produce offspring of the next generation.

Selection - This phase of the genetic algorithm selects individual solutions from the current population to produce a new generation of the population by applying the crossover between the selected individuals. In our project, we have used elitism selection strategy.

Figure 7. About us Page

Figure 8 tells why we should use this system.


Class Prediction Software
Marks Predictor | Major Pro
Learner Performance Pr
Learner Performance Predictor
Home
About
Why
Predict Output
Activate Windows
Go to Settings to activate Windows
Type here to search
9:21 AM
6/7/2020

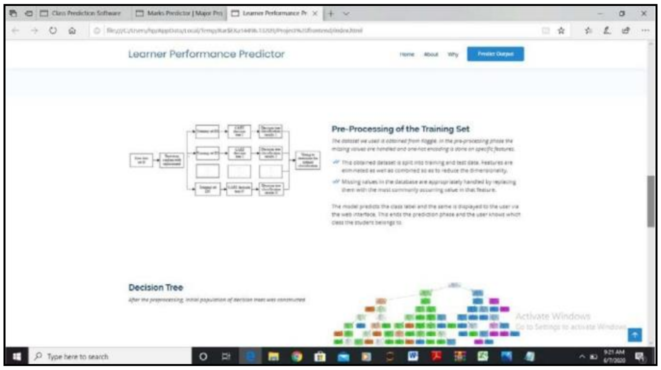
Selection - This phase of the genetic algorithm selects individual solutions from the current population to produce a new generation of the population by applying the crossover between the selected individuals. In our project, we have used elitism selection strategy.

Elitism - In this strategy, one parent is selected form top k individuals based on their fitness value and their corresponding partner is selected from the other part of the population.

WHY WE ARE USING THE PREDICTOR?

Class 1-5	Class 6-10	Class 11-15	Class 16-20

Figure 8. Why to use the predictor


Class Prediction Software
Marks Predictor | Major Pro
Learner Performance Pr
Learner Performance Predictor
Home
About
Why
Predict Output
Activate Windows
Go to Settings to activate Windows
Type here to search
9:21 AM
6/7/2020

Pre-Processing of the Training Set

The dataset we used is obtained from Kaggle. In the pre-processing phase the missing values are handled and one-hot encoding is done on specific features.

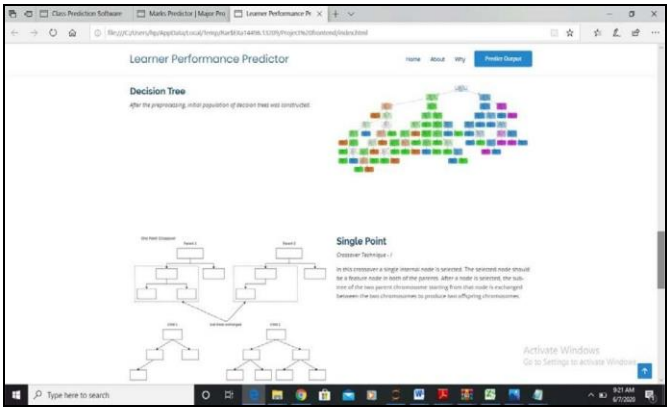
- This obtained dataset is split into training and test data. Features are eliminated as well as combined so as to reduce the dimensionality.
- Missing values in the database are appropriately handled by replacing them with the most commonly occurring value in that feature.

The model predicts the class label and the same is displayed to the user via the web interface. This ends the prediction phase and the user knows which class the student belongs to.

Decision Tree

After the preprocessing, initial population of decision trees was constructed.

Figure 9. Page1



Decision Tree

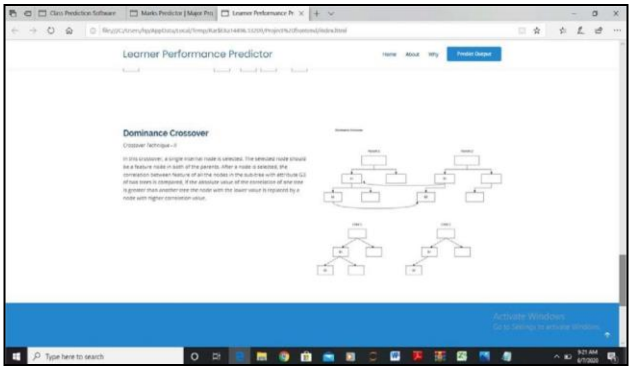
After the preprocessing, initial population of decision trees was constructed.

Single Point

Crossover Technique - I

In this crossover a single internal node is selected. The selected node should be a feature node in both of the parents. After a node is selected, the sub-tree of the two parent chromosome starting from that node is exchanged between the two chromosomes to produce two offspring chromosomes.

Figure 10. Page1



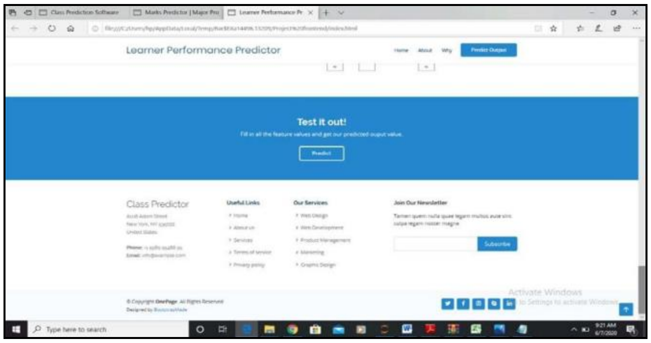
Learner Performance Predictor

Dominance Crossover

Crossover Technique - II

In this crossover, a single internal node is selected. The selected node should be a feature node in both of the parents. After a node is selected, the correlation between feature of all the nodes in the sub-tree with attribute G3 of two trees is compared. If the absolute value of the correlation of one tree is greater than another tree the node with the lower value is replaced by a node with higher correlation value.

Figure 11. Page1



Learner Performance Predictor

Test it out!

Fill in all the feature values and get our predicted ouput value.

Class Predictor

Useful Links
- Home
- About us
- Services
- Terms of service
- Privacy policy

Our Services
- Web Design
- Web Development
- Product Management
- Marketing
- Graphic Design

Join Our Newsletter

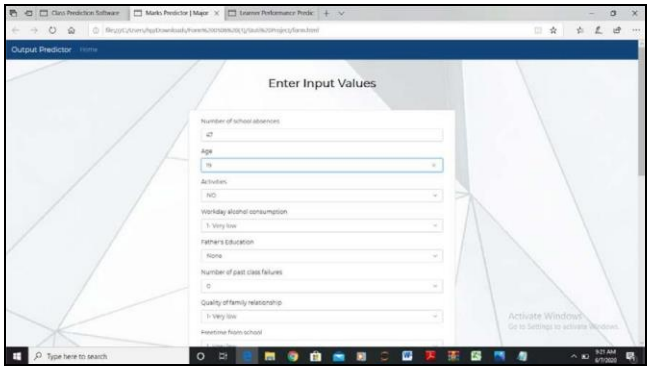
Tamen quem nulla quae legam multos aute sint culpa legam noster magna

© Copyright OnePage. All Rights Reserved
Designed by BootstrapMade

Figure 12. Page1

.

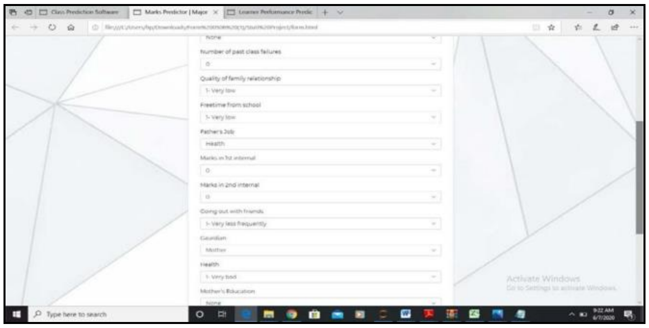
Figures 9,10 and 11 are the input form page. This data is accessed from the users.



Enter Input Values

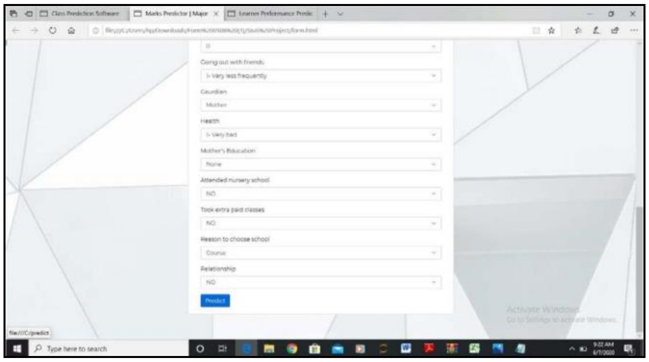
Number of school absences: 47
Age: 15
Activities: NO
Workday alcohol consumption: 1- Very low
Father's Education: None
Number of past class failures: 0
Quality of family relationship: 1- Very low
Freetime from school

Figure 13. Forms Page1



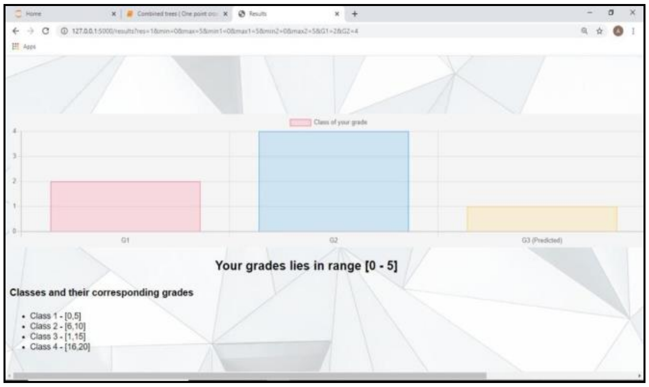
None
Number of past class failures: 0
Quality of family relationship: 1- Very low
Freetime from school: 1- Very low
Father's Job: Health
Marks in 1st internal: 0
Marks in 2nd internal: 0
Going out with friends: 1- Very less frequently
Guardian: Mother
Health: 1- Very bad
Mother's Education: None

Figure 14. Forms Page2



0
Going out with friends: 1- Very less frequently
Guardian: Mother
Health: 1- Very bad
Mother's Education: None
Attended nursery school: NO
Took extra paid classes: NO
Reason to choose school: Course
Relationship: NO

Figure 15. Form Page3

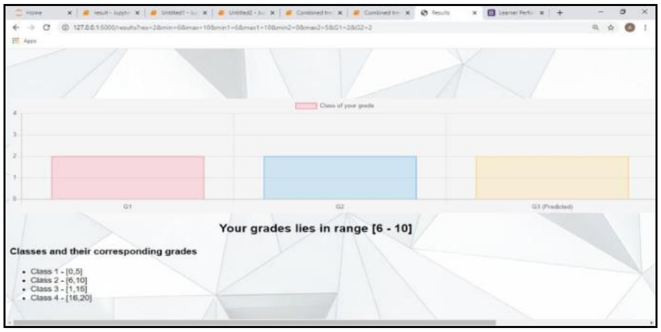

127.0.0.1:5000/results?res=1&min=0&max=5&min1=0&max1=5&min2=0&max2=5&G1=2&G2=4
Class of your grade
0
1
2
3
4
G1
G2
G3 (Predicted)

Your grades lies in range [0 - 5]

Classes and their corresponding grades

- Class 1 - [0,5]
- Class 2 - [6,10]
- Class 3 - [1,15]
- Class 4 - [16,20]

Figure 16. Result Page


127.0.0.1:5000/results?res=2&min=6&max=10&min1=6&max1=10&min2=0&max2=5&G1=2&G2=2
Class of your grade
0
1
2
3
4
G1
G2
G3 (Predicted)

Your grades lies in range [6 - 10]

Classes and their corresponding grades

- Class 1 - [0,5]
- Class 2 - [6,10]
- Class 3 - [1,15]
- Class 4 - [16,20]

Figure 17.Result with different outputs

Figures 16 and 17 as shown above include the student prediction according to the details entered by the user and it also includes the contributions of each feature that leads to that particular prediction this score changes according the details entered by the user.

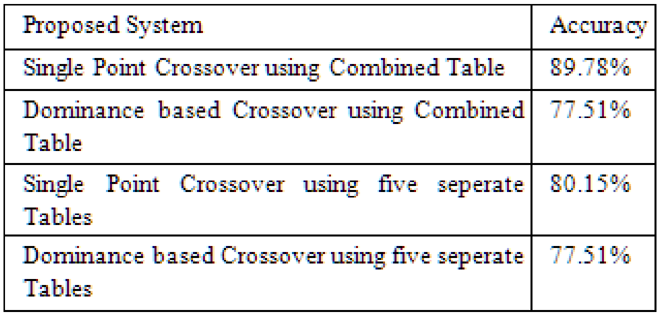
IV. DISCUSSION AND CONCLUSION

As a part of the prediction stage in our system, we initially split the dataset obtained from the Kaggle repository to five tables depending on their traits and habits. This data was then fed into the model by encoding them and forming an array. The data was passed through the steps of the genetic algorithm to obtain a better accuracy of. Also, we have applied different types of crossovers to compare the obtained accuracy.

- The aim was to obtain a set of decision trees using a genetic algorithm which can correctly classify our data.
- Building an interface that takes attributes as inputs from the user and correctly classifies the data given.
- In our project we are mainly using two types of crossovers to do the same, whilefour different types of methods are processed for comparing the end results.

The accuracy of our proposed system is as follows:

Table 8.1: Accuracy of crossovers



Proposed System	Accuracy
Single Point Crossover using Combined Table	89.78%
Dominance based Crossover using Combined Table	77.51%
Single Point Crossover using five seperate Tables	80.15%
Dominance based Crossover using five seperate Tables	77.51%

The size of the dataset may not be sufficiently large. The clarity of the results depends a lot upon the dataset, and only with sufficient data can we have good results. The data needs to be available exactly in the required form. For that, we need a good amount of data cleaning. Both the models are very far from being versatile. Future researchers can examine the same correlations by carrying out a longitudinal research study i.e, doing research on data that is gathered over a long span of time. Future work can also consider including more factors from the database that could have more effect on deciding representative classes.

REFERENCES

[1] Suman Khatwani and Arti Arya , *A Novel Framework for Envisaging a Learner's Performance using Decision Trees and GeneticAlgorithms* . 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 09 – 11, 2013, Coimbatore, INDIA

[2] Paulo Cortez and Alice Silva , *Using Data Mining to Predict Secondary School Student Performance*. "Proceedings of 5th Annual Future Business Technology Conference, Porto, 2008". [S.l. : EUROSIS, 2008]. ISBN 978-9077381-39-7. p. 5-12.

[3] Sung-Hyuk Cha , *Constructing Binary Decision Trees using Genetic Algorithms*, . In GEM, pp. 49-54. 2008.

[4] Cristóbal Romero, Sebastián Ventura, Pedro G. Espejo and César Hervá , *Data Mining Algorithms to Classify Students* . Educational data mining 2008.

[5] Brijesh Kumar Baradwaj and Saurabh Pal , *Data Mining: A prediction for performance improvement using classification*. (IJCSIS) International Journal of Computer Science and Information Security, Vol. 9, No. 4, April 2011.