Education Visualization: Exploring Data for Academic Success

Ronak Etemadpour, Yongcheng Zhu, Qizhi Zhao, Shirong Zheng, Bohan Chen, Mohammed Sharier, and Yi Hu

Abstract—Academic performances in college carries many importances in life. They determine whether the student is going to graduate, receive scholarship, financial aid, and opportunities more than one can count. The objective of our research is to perform visual analysis with students' track records, and to extract information of how well a student is going to perform in a particular class based on his or her certain characteristics, and so on. Our model allows user to input student's academic dataset from previous years and semesters, and then display the data graphically on a website with many interactive aids.

Index Terms—N/A

1 Introduction

As computer science students, solving practical problems with computer has been our expertise. And now, since we've learned so many useful tools into our kit, it's time to solve a problem of our own—improving our grades. After experiencing the struggle of completing countless credits of courses from freshmen up until now, it's a difficult task to maintain our GPA due to many different factors revolving around the courses. And in attempt to keep up with our studies and GPA, we go from cramming overnight for exams, completing projects and homeworks on time, time management, choosing the right professor, or even the right hours for the classes. Before enrolling into any classes, one trivial action for students to take is to do a research on the professor. And for many of us students, *ratemyprofessor* has been one of the most popular go-to sites to determine whether we should take this professor or not. But the problem with this is that, the sample sizes are often way too little to determine whether this is the right professor to take or not. Hence, we thought it is more reliable to ask the school department for previous students' record with their names hidden, and study upon this dataset. (Although some colleges are strict, and will not allow students to request this without specific purposes.)

We noticed that many college courses does not require student's attendance to be mandatory, so then, does a student's attendance affect their grade overall at the end? Hence, we did some analysis on the student's attendance in respect to the class taken. To do this, we used various machine learning models to attempt to predict the trend of the dataset's behavior, and then we compare this prediction to the real dataset. If the accuracy is high enough, then we use this machine learning model to predict a student that's going to take the class based on his or her characteristics and past records to determine how well he or she is going to perform in this class if absent for this many times.

But, a student's grade may not only depend on his/her past record. We took into consideration of the outside factors, that maybe, their family status also had an impact to their grades. So for this purpose, we did some analysis into their parent's education, and job titles. To do this, we used high dimensional data reduction technique to visualize each student's parental education into two dimension coordinates, then plot them graphically. This visualization allows user to analyze student's gender, parent's educatiaon, parent's job, and even compare average grades with interaction. If their parent's education really matters, then we will see student's whose parent's have similar education grouped together in the plot, and so on.

Getting 'good' grades for courses is important for academic success. But, what's most important is that, are we going to graduate? We are going to study the differences between GPA and graduation rate. As well as doing well on which courses tend to yield higher graduation rate. And then use machine learning to determine how well the student is currently performing, and how likely is this student going to graduate.

Getting good grades on the initial examinations doesn't necessarily merit an overall pass. Similar logic can also be applied towards people who have low grades on the early stages and make it till the end. The grades for students don't only just depend on what they do initially, their grades depend on how active and motivated they are until the end. Since, we did some analysis on the grades of the students and future predictions of their outcomes, machine learning dataset was used to compare with the real data in order to see the accuracy of the predictions. This visualization will allow the users to perceive whether they should see failure as complete fail or an opportunity to learn from mistake and make better of themselves.

We also found that student's gender might also affects their grade performance on specific courses. We think that might related to student's interest to the courses, therefore we think in some courses one gender group will do better than the other. From our dataset, we will calculate the average grade of both males and female student groups, then use zoomable heatmap and dendrogram visualization to display the result to see whether the student's gender and grades shows any sign of correlation.

To summarize, here's a list of our research questions, including in which section they're mentioned in.

- Since many college courses doesn't require attendance to be mandatory, so does a student's attendance affect their resulting grade? (Section 3.2.2 and section 4.1)
- Does a student's parental education and job have any impact on their grades? (Section 3.3.1 and section 4.2)
- Does a student's initial performance reflect heavily on the overall performance? (Section 3.5 and 4.4)
- Does the gender of the student affect his or her grade on different type of course? (Section 3.4 and 4.3)

2 Related Work

Because of privacy issue, it's difficult for data researchers to find ways obtain educational data, therefore there weren't many papers on IEEE Vis that seemed to be relevant to our research. However, we still found some of their methodologies useful for our objective. For example, there were plenty of papers about how instructors can utilize data visualization techniques to aid students in their studies and so on. But luckily, outside of IEEE Vis, we did find some papers that were published in another university that's pretty much relevant to what our research is trying to achieve. We decided to implement some of their ideas into our own research.

2.1 Predicting Student's Graduation Rate

Martinez and Miller [4] have developed a good way to determine how likely a student is going to graduate after four years in college.

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.
Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/

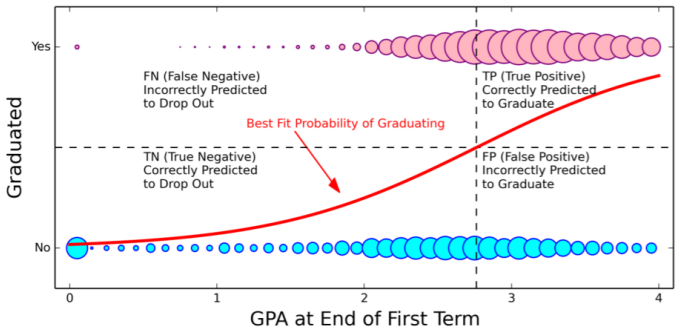

Fig. 1: Logistic Regression

Logistic Regression is a special type of regression where binary response variable is related to a set of explanatory variables, which can be discrete and/or continuous. The important point here to note is that in linear regression, the expected values of the response variable are modeled based on combination of values taken by the predictors. In logistic regression Probability or Odds of the response taking a particular value is modeled based on combination of values taken by the predictors. Like regression we make an explicit distinction between a response variable and one or more predictor variables. We begin with two-way tables, then progress to three-way tables, where all explanatory variables are categorical. Then we introduce binary logistic regression with continuous predictors as well. Logistic regression is applicable, for example, if:

1. we want to model the probabilities of a response variable as a function of some explanatory variables, e.g. "success" of admission as a function of gender.
2. we want to perform descriptive discriminate analyses such as describing the differences between individuals in separate groups as a function of explanatory variables, e.g. student admitted and rejected as a function of gender
3. we want to predict probabilities that individuals fall into two categories of the binary response as a function of some explanatory variables, e.g. what is the probability that a student is admitted given she is a female
4. we want to classify individuals into two categories based on explanatory variables, e.g. classify new students into "admitted" or "rejected" group depending on their gender.

2.2 P-Value Approach

A good way to observe data variation is to use the Kolmogorov-Smirvnov Test which is a method that produce P-value [5]. Following this idea, we are going to calculate a p-value to compare the two samples to determine whether they are produced by the same probability distribution. If the p-value is close to 1 that means there's no significant different, however is it is 0.05 or less, then there's a significant difference. From this article we learn that P-value can be used to verify the credibility of the correlation that we observe between data, therefore, to testify if our dataset produces meaningful data, we will calculate P-value between the two data groups to see if it produces meaningful correlation.

2.3 Design Study Methodology

Following the advices given in [2] we can formulate a plan while carrying out our methodology. The steps are as following,

1. Choose a domain problem that is easy to understand.
2. Work with pre-calculated and "clean" data.
3. Structure and limit the amount of iteration.

And for each step of the methodology, we can further break them up into:

1. Abstract
2. Design
3. Build
4. Evaluate
5. Dissemble

And finally, come to a conclusion from our results. Base on this design study methodology, we try to think of some easy understanding research questions from our education dataset. Then we calculate and re-arrange our dataset into different parts for different research questions. Lastly, we apply these "clean" dataset to our visualizations and analyze them to answer our research questions.

2.4 Principles for interactive visualization

The principles of design weren't created for data gurus to build applications or dashboards; they're the essential building blocks artists use to paint and sculpt. Still, the concepts behind these design fundamentals can apply to any medium, including the data visualization in your dashboard.Two basic principles for interactive visualization of high dimensional data [1]. These are focusing and linking. We going to discover how graphical data analysis methods based on focusing and linking are used in applications including linguistics, geographic information systems, time series analysis, and the analysis of multi-channel images arising. A consequency of focusing is that each view will only convey partial information about the data.Focusing techniques may involve selecting subsets, dimension reduction, or some more general manipulation of the layout of information on the page or screen. Examples of subset selection techiniques are panning and zooming, and slicing. The prncipal mechanism for linking views over time is through smooth change in the position of objects on the screen. Examples for this kind of linking are rotating three-dimensional point clouds and a higher dimensional generalization, the Grand Tour. Any smooth animation can be considered as a set of linked multiple views, spread out over time.The multidimensional point data through simulataneous multiple views that are linked by drawing lines connecting the points in the different views coreesponding to the same observation.

2.5 Course Pair Grade Analysis

Mark Blaise DeCotes has developed a good idea to analyze students' grade performance for different course pairs. Course pair means the a course that's consist of a level one and level two component over two semesters. Here, he is comparing three different course pairs with interesting results.

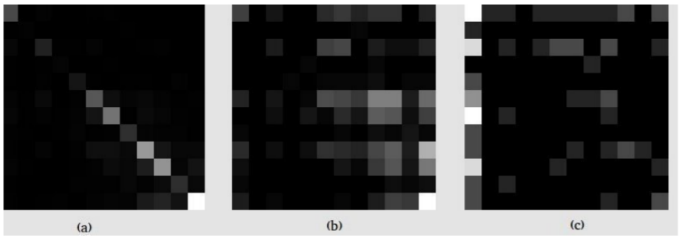

Fig. 2: Heatmaps showing the histogram of three course pairs, where the x-axis is the level 1 course grades and y-axis is the 2 course grades. The grades are scaled from W or F to A. And the color of each box represents the number of students, where black is 0 percent and white is 100 percent.

In these heatmaps, he has shown that different course pairs yield different behaviors. For the first course pair, it appears that student's who did poorly in level one course will receive similar poor grade in the level two course. For the second course pair, grades seems to be a bit more distributed, but we can still tell that if a student got an A in level one then he/she will most likely get an A in level two. The last course pair is the most interesting one, because it appears that students who did good in level one did not transition the good grades over to the level two course, but students who did poorly, however did.

Using similar heatmaps, instead of analyzing course pairs, we used them to analyze student's attendance and grade. Where we have our

x-axis to represent the number of absences, and the y-axis to scale the grades from zero to a hundred. Similarly, the color scale to represent the percentage of students in this category.

2.6 Visualization Module

In this paper [6] we have reviewed different techniques methods and tools which have some shortcomings of their own. We have discussed many paper from which we got a broad idea about a system which is required in today's world for analysis and visualizing the sales data using which the investors and owners of the organization can make proper decision and generate revenue. System also gives a facility of sorting and searching the attribute which may help end user in searching an item faster can make decision, predict the future sales, calculate regional sales, increase the production depending on the demand, to take decisions, generate revenue, and plan promotions. The visualization module is Bar Graph, Pie chart, Line graph, Area graph are used for representing the sales depending upon region, product category, product container, order priority, customer segment, shipping mode, etc. An option of date range is provided using which user can select time duration for which he needs the report. If data is not present for the selected date range then an alert as "data is not available" is given. different module help people easily understand the complex data. The data visualization in my life is primarily in the business-world. At my day job: how do we ensure that people decisions at Google are data-driven? In my presentations and workshops: who is our audience, what do they need to know, and how do we craft a visual and story to do that? But many take data visualization into the personal sphere as well: using visualization to better understand aspects of their world or their life.

2.7 Interactive visualization approach

We can use the way this paper "Gene Slider: sequence logo interactive data-visualization for education and research" [7] creates their data visualization by make the visualization interactive so that users are able to play with it and have better understanding on the main purpose of that visualization. From the figure we can see that it also uses size and color for data representation, the higher score the bigger shape. These element helps the visualization stands out clearly. I learned that our education data set can also use some of these visualizations that the author uses. We can also make more interactive visualizations for our user. For example, the bubble chart can be used to compare the overall grades of each courses between two genders. We also can make it interactive by allow the bubble to change its color and size to show the difference between student gender and the courses.

3 METHODOLOGY

Our method revolves mainly in utilizing and taking advantage of D3.JS' ability to show the movement of data over time, and it's ability to allow user to interact with the visualization to understand and analyze the dataset. We borrowed some ideas from the related works that other researchers have published and modified it to answer our own set of research questions.

3.1 Data Set Preparation

Obtaining the datasets is the hardest part of this research topic, because academic data are generally private and not distributed on the web. But thanks to Dr. Jose Gustav Paiva's help, we managed to obtain our first set of data of student records from the university that he works at. This dataset contains few attribute columns that allow us to analyze how attendance and the specific course that each individual student might affect their grade. And for the second set of data, we were able to obtain it thanks to our team member Qizhi Zhao, he found it after rigorous hardwork of searching all around the web. This data set is more rich in dimensions, which includes the student's gender, semester's grade, family information, and even romantic relationships. We organized these data into .json format to be ready to use for visualization.

3.1.1 Dataset 1

This was the first dataset that we've obtained through our professor's connection with her colleague in another university. This dataset consists of six columns of attributes. The first column is the *Student's ID*, it is a three-digit number that denotes an anonomyous student instead of his or her name. The second column is the student's *Gender*, denoted by 'M' or 'F'. The third column is the *date* in which the student have completed the course, it has data from all the way back in 2009 to 2017. The fourth column is the *course subject* number, which includes Algorithms and Programming I, Algorithms and Programming II, Algorithms and Programming III, Discrete Mathematics I, Discrete Mathematics II, Object Oriented Programming, Data Sctructures, Logic for Computer Science I, Logic for Computer Science II, and Human-computer Interaction. The fifth column is the student's *grade* received for this course, ranging from 0 to 100. And finally, the last column represents how many *absences* each student had for this course in this semester.
For this dataset, there are total of 6 dimensions, consisting of the *Student ID, Gender, Date, Subject, Grade, and Attendance*. Using python's machine learning library *schikitlearn*, we pass in the attributes of each individual student to attempt to predict their grade with different regression models. For this research, we tested the data on a total of 9 different machine learning models, namely, No Regulation Perceptron, Linear Regression, Ridge Regression, Kernel Ridge, Elastic Net, Random Forest Regressor, SVR Linear, SVR RBF, and SVR SIG. We calculated the accuracy of by measuring the percentage of students that exist in the each subject, and divide the correctly predicted percentage of student by the total percentage of predicted students. This gave us a intuition of which model is more accurate than another for specific subjects. We will define each of the machine learning models in a later section.

3.1.2 Dataset 2

This dataset is a bit more complicated, because there are a total of thirty-two attributes in here. But for the sake of clarity, we will not mention any of the attributes that we didn't use. For our research, we only took into the considerations of six attributes. The first is the *age* of the student. The second is the number of *failed classes* that the student has in his or her record, this number is in the range of 0 to 4. The third is the number of *absences* from the beginning to the end of this course (three grading periods). The rest are the *grades* for three separate grading periods in this semester (this is common in highschool), they're numerical values ranged from 0 to 20. However, this dataset is actually seperated into two different files, one for math class and another for portuguese class, so we actually have a seventh attribute called *Subject*.
For this dataset, there are a total of 35 dimenions. However, we attempted to study this dataset by applying the same machine learning techniques, but the results (prediction accuracy) was abysmal. Therefore, we decided to use a different approach suggested by our professor, which is to use isomap to convert all the quantitative dimensions into lower dimensions, namely, two dimensions— the spatial coordinates x-axis and y-axis. We transformed the original dataset into four isomap datasets, by applying the following quantitative dimensions:

1. Subject, Age, Failures, Absences, G1
2. Subject, Age, Failures, Absences, G2
3. Subject, Age, Failures, Absences, G3
4. Subject, Age, Failures, Absences, G1, G2, G3

The spatial coordinates is calculated based on the similarity between the input attributes for each individual students. Therefore, this is a type of clustering that is used to classify different groups of students. Which may also allow us to identify possible outliers for different datasets and filter settings. However, with this approach, we only get to see the relationship between the quantitative dimensions. If we want to see the categorical relationships, we have to visualize with some interesting attributes as we will show in the later sections.

3.2 Attendance Analysis

By attendance analysis, our goal is to figure out by what extent does a student's attendance affect their grade? To answer this question, we thought machine learning is a great approach for predicting a student's overall grade for this particular class based on his/her number of absences. For our research, as we've mentioned in the previous section, we utilized a total of 11 different machine learning models to pass in our data input. And they're No Regulation Perceptron, Linear Regression, Ridge Regression, Kernel Ridge, Elastic Net, Random Forest Regressor, SVR Linear, SVR RBF, and SVR SIG. But to get any further, we need to understand the definition and the underlying behavior for each of these models. After applying machine learning to the dataset, we are going to visualize the results in D3.

3.2.1 Machine Learning Analysis

For this dataset, we create a machine learning model by using python's Scikilearn library for the purpose of learning to predict grade. We fit-in the numerical data to the model and generate the prediction grrade. Therefore, by visualizing the data and the predicted data, we can observe the behavior of our datsaet and conclude some idea.

The perceptron is an artificial neural network invented by Frank Rosenblatt when he worked at the Cornell Aeronautical Laboratory. It can be regarded as one of the simplest forms of feeding forward artificial neural networks and is a binary linear classifier. The MLP (Multilayer Perceptron is a feedforward artificial neural network model that maps multiple input datasets onto a single output dataset via three or more layers of perceptron. And perceptron is an binary linear classifier algorithm for supervised learning of data. Linear regression is a widely used statistical analysis model that uses the regression analysis in mathematical statistics to determine the quantitative relationship between two or more variables. Its expression is $y = w'x + e$, and e is a normal distribution with an error obeying a mean of 0. In the regression analysis, if there is only one independent variable and one dependent variable inclded, then the relationshp between this variable can be represented by a straight line approximation. This regression analysis is called linear regression analysis. Ridge regression is also known as Tikhonov regularization. It is a biased estimation regression method dedicated to collinear data anaylsis. It is essentially an improved least squares estimation method, which loses partial information by abandoning the unbiasedness of the least squares method. Reducing the accuracy at the expense of regression coefficients is more reaslistic and reliable. And for the ill-conditioned data, it is better than the logistic regression as a solution of over-fitting. Kernel Ridge regression (linear least squares with 12-norm regularization) with the kernel trick. ElasticNet is a linear regression model that uses both L1 and L2 penalties of the lasso and ridge methods. This combination is used for sparse models with few non-zero weights. Support Vector Machine (SVM) is a machine learning method based on statistical learning theory developed in the mid 1990s. It seeks to maximize the learning machine's generalization ability by minimizing the structural risk and minimize the empirical risk and confidence range. In the case of a small sample size, the purpose of good statistical behavior can also be obtained. In generally speaking, it is a binary classification mode, and its basic model is defined as the linear classifier with the largest interval in the feature space. That is, the learning strategy of the suppose vector machine is to maximize the interval. Eventually it can be transofrmed into a solution to a convex quadratic programming problem. In the case of linear inseparability, the support vector machine first completes the calculation in low-dimensional space, then maps the input space to the high-dimensional feature space through the kernel function, and finally constructs the optimal separation hyperplane in the high-dimensional feature space, thus Non-linear data that is not well divided on the plane is seperated. LASSO, also known as least absolute shrinkage and selection operator. It is a regression analysis method that simultaneously performs feature selection and regularization to enhance the prediction accuracy and interpretability of statistical models based on Breimans nonnegative garrote. The decision tree is a tree structure (which can be a binary tree or a non-binary tree). Each of its non-leaf nodes represents a test on a feature attribute, each branch representing the output of the feature attribute over a range of values, and each leaf node storing a cateogry. The decision process using the decision tree is to start from the root node, test the correspodning feature attributes in the item to be classified, and select the output branch according to its value unti the leaf node is reached, and the category stored by the leaf node is used as the decision result. The random forst is a type of target estication. It forms a decison tree by using some samples on the dataset, and uses averaging to improve the prediction accuracy and control over-fitting.

3.2.2 Visualization Analysis

To visualize our datasets, we had many options. But, we decided to use D3.js because our professor showed us how elegant this javascript library is, and more importantly, it's easy to learn and understand since we have javascript backgrounds. We came up with two approaches, one is to use heatmap where the user can identify the number of students that is absent for this many times, and received this amount of grade. The amount of students in this category will be scaled in a color range from white to red. Our second approach is to graph this into a spatial domain using the grade and attendance as x-axis and y-axis. This approach allows us to see different clustering in the spatial domain and allow us to identify what which attribute of this student gave him/her this grade.

DeCotes [3] has used a similar heatmap approach to visualize the course grades for three course pairs. But instead of analyzing different course pairs, we are studying the absence in respect to the final grade received. To do this, we let our x-axis represent the amount of absences from 0 to maximum. And y-axis to represent the grade from 0, 10, 20, to 100. The algorithm is very simple, we first created a 2D array in javascript. Then, loop through the dataset to find the maximum number of absences, because this number will be used for our x-axis (our columns for the 2D array.) Then, we loop through the entire array and initialize every entries to zero so that we can increment later to represent the student percentage count. Next, the most important step, is to loop through the dataset, and find the row and column indexes for the array so that we can increment that entry. To do this, we used the following calculations:

$$\begin{aligned} Row =& 10 - (Math.round(data[i]["Grade"]) \\ & - Math.round(data[i]["Grade"] mod 10)) \\ Column =& data[i]["Attendance"] \end{aligned}$$

Now simply index into the array with the row and column variables and increment the entry for every student. But this is still not enough, because we are just counting the number of students, but we need to convert these numbers into percentages based on the total number of students in this class. So, we have to loop through the array again and for each entry, take that number and divide that by the total number of students in this class (by filtering the dataset) and then multiply this number by 100. You can then round this number to the nearest floating point, by we just left it as it is. Next, we defined the color scale for the student percentage, where 0 percent is in white and 100 percent is in red. So, in the heatmap each we will see this color scale for each square entry and we can distinguish the percentage of student that is in this category. For the SVG, we declare the width and height dynamically so that if the maximum number of absences changes based on the selected class, the width of the SVG will change accordingly so that it centers in the middle of the screen. To create the squares, We didn't use conventional method for D3.js we simply created a loop for the 2D array, and append a square for every entry. we scaled the x and y attributes by a factor of 10, so that makes each square 10x10 pixels of width and height. So that basically completes the essential heatmap on the screen. For the details, we added x-scale and y-scale for the user to see. And a tooltip that pops-out upon hovering.

For the machine learning models, we mirrored a heatmap on the bottom for the user to observe and compare the differences between this and the heatmap of the actual dataset. If the heatmaps are somewhat similar, then the prediction is very accurately, and the oppposite otherwise. For better user interactivity, we imple-

mented a sidebar for the user to filter the classes and machine learning techniques. The bottom heatmap which is the machine learning data will change accordingly. For example, let's take a look at the heatmap for the course *Logic in Computer Science I*.

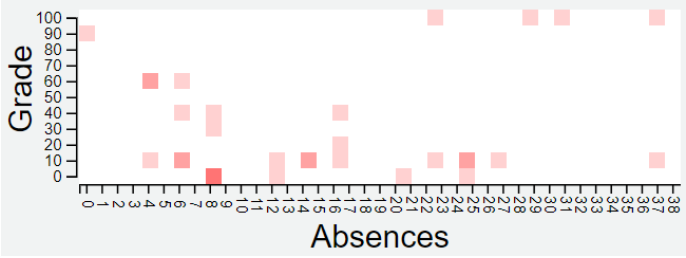

Fig. 3: Heatmap showing the actual dataset for *Logic in Computer Science I*.

Next, we are going to compare this heatmap with the bottom heatmaps, which shows the dataset for our machine learning predictions. On our website, we use the sidebar menu for this purpose, but for convienience in this report we will display all the heatmaps as figures in the following.

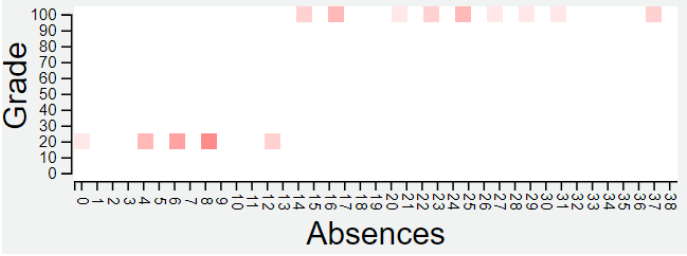

Fig. 4: Heatmap showing No Regulation Perceptron for *Logic in Computer Science I*.

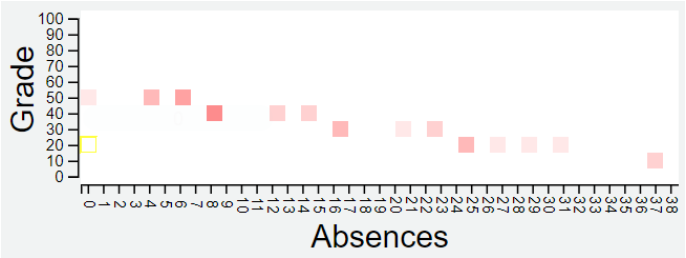

Fig. 5: Heatmap showing Linear Regression for *Logic in Computer Science I*.

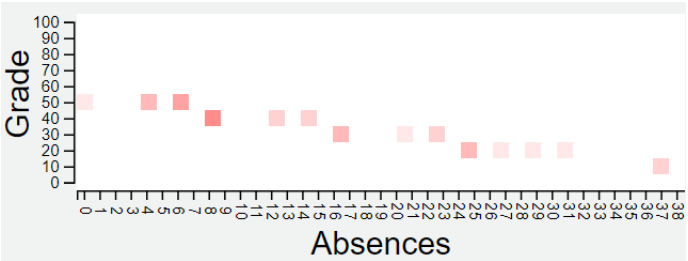

Fig. 6: Heatmap showing Ridge Regression for *Logic in Computer Science I*.

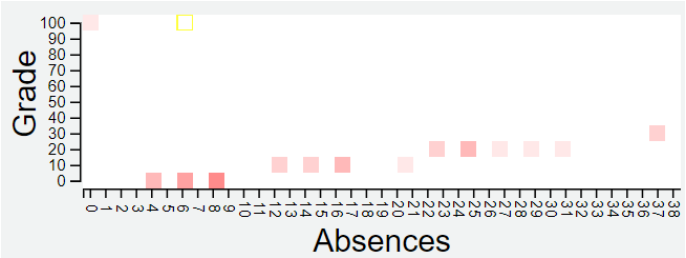

Fig. 7: Heatmap showing Kernel Ridge for *Logic in Computer Science I*.

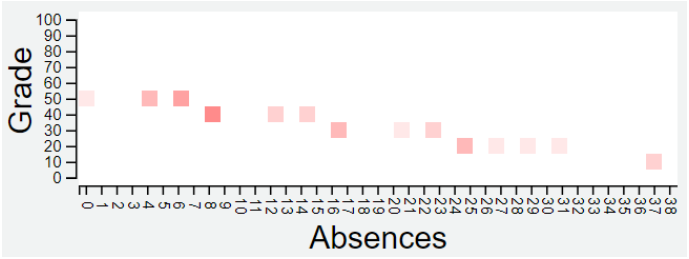

Fig. 8: Heatmap showing ElasticNet for *Logic in Computer Science I*.

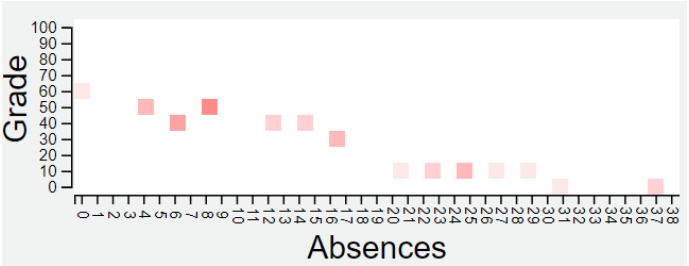

Fig. 9: Heatmap showing Descision Tree for *Logic in Computer Science I*.

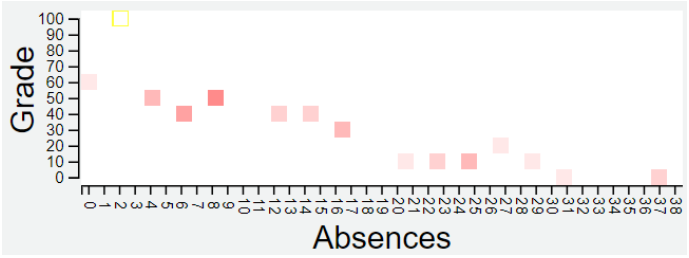

Fig. 10: Heatmap showing Random Forest Regressor for *Logic in Computer Science I*.

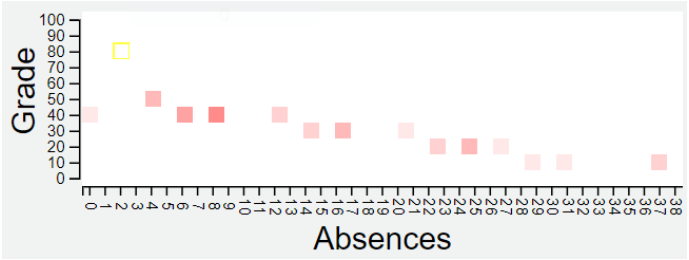

Fig. 11: Heatmap showing MLP Regressor for *Logic in Computer Science I*.

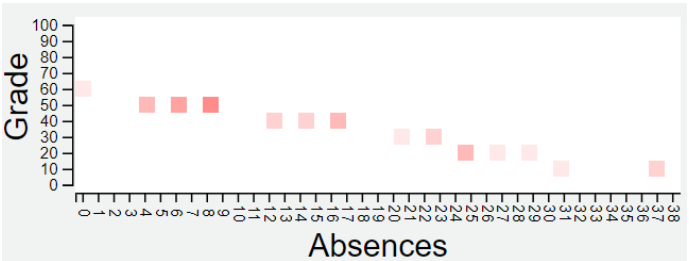

Fig. 12: Heatmap showing SVR Linear for *Logic in Computer Science I*.

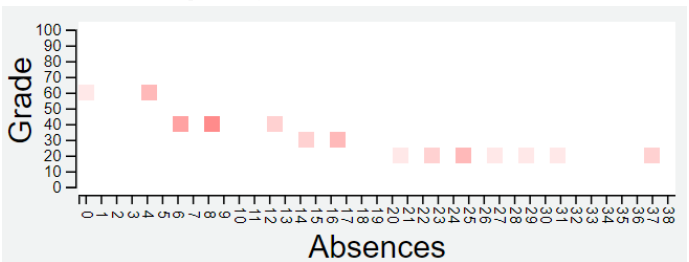

Fig. 13: Heatmap showing SVR RBF for *Logic in Computer Science I*.

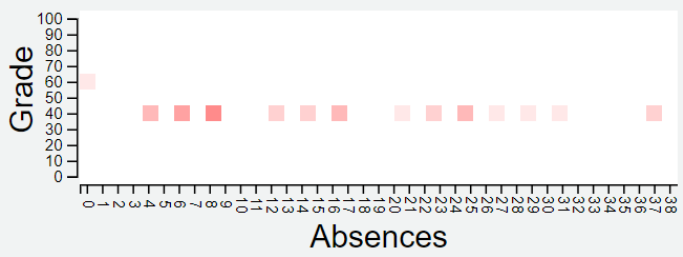

Fig. 14: Heatmap showing SVR SIG for *Logic in Computer Science I*.

As we can see, our predictions are not so accurately. The closest machine model that came closest was around 40% accuracy, which was the Random Forest Regressor. Hence, we cannot conclude any meaningful results from this method, because it would be a bad idea to predict a student's grade with 40 percent accuracy.

Thus, we decided to plot our data into a scatterplot to see if we can find any interesting behaviors. Since we only have two quantitative data dimensions, it's a simple task to visualize this in D3. (But for the latter dataset, in which we have more than two quantitative data dimensions, we had to do something special inorder to plot them.) The x-coordinates in this scatterplot represents the grade received by the student (dot) after completing the course, as grades increases from left to right. The y-coordinates represents the number of absences for each students, as the number of absences increase from top to bottom. Here is a demonstration of our scatterplots from our website.

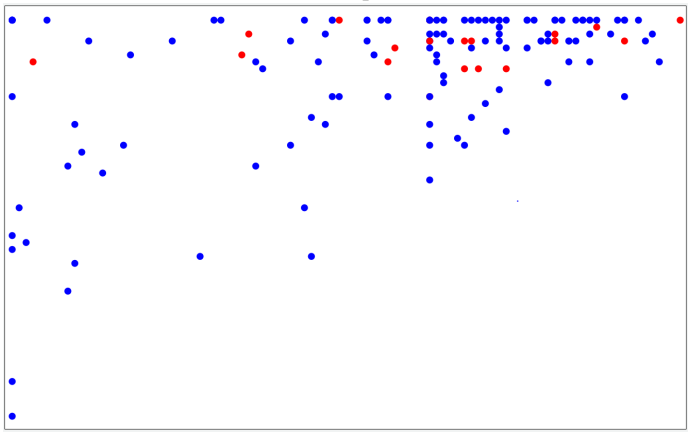

Fig. 15: Scatterplot showing Algorithm and Programming I.

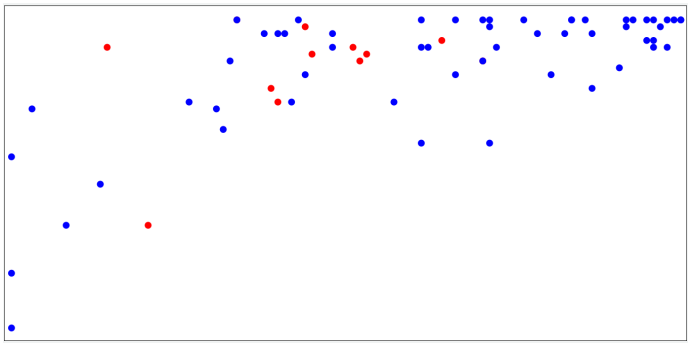

Fig. 16: Scatterplot showing Algorithm and Programming II.

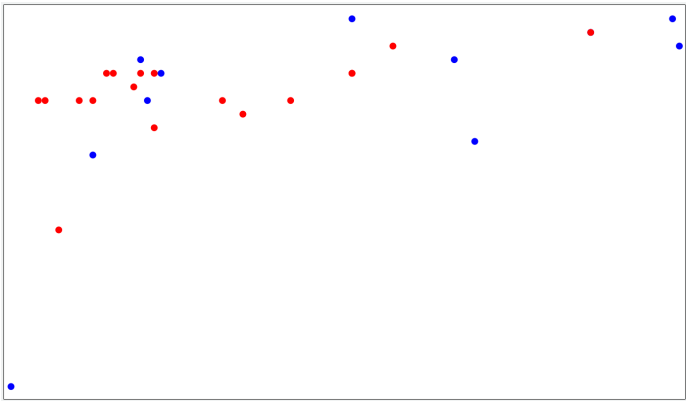

Fig. 17: Scatterplot showing Algorithm and Programming III.

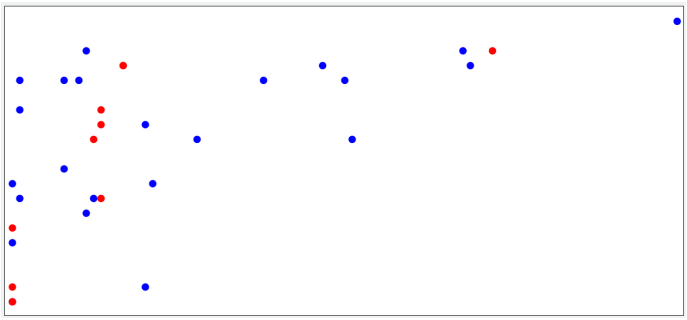

Fig. 18: Scatterplot showing Logic in Computer Science I.

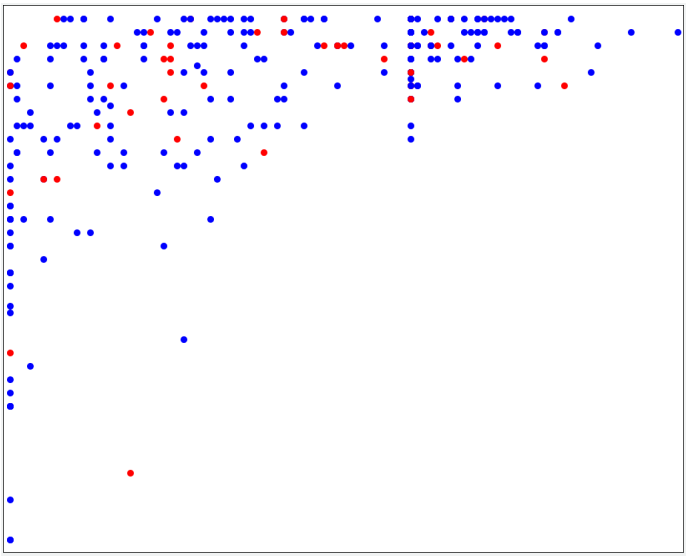

Fig. 19: Scatterplot showing Logic in Computer Science II.

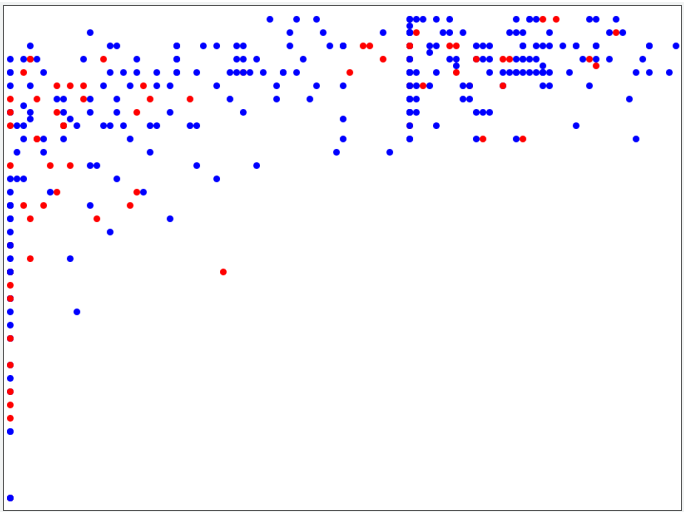

Fig. 20: Scatterplot showing Object Oriented Programming.

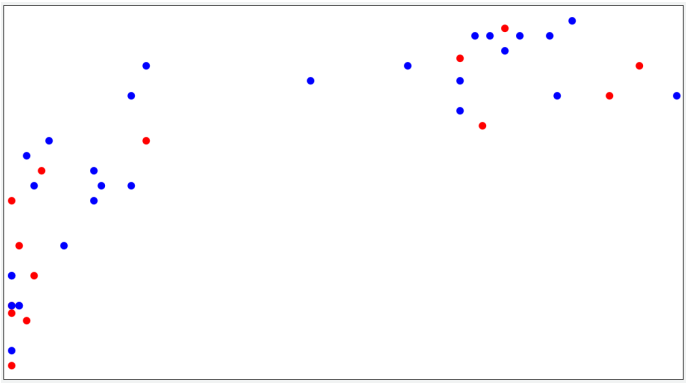

Fig. 21: Scatterplot showing Data Structures.

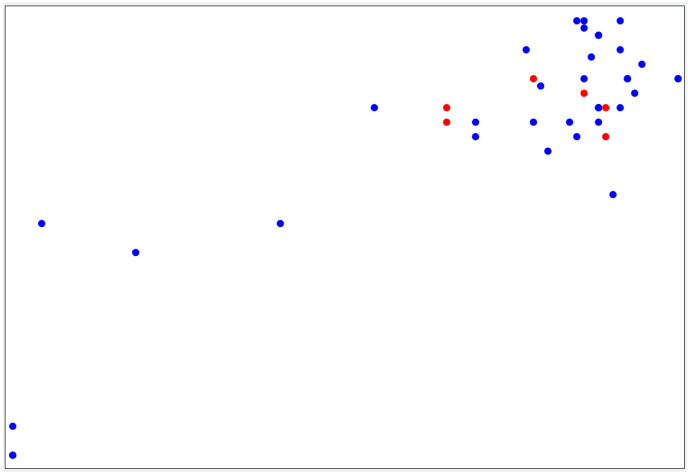

Fig. 22: Scatterplot showing Human-Computer Interaction.

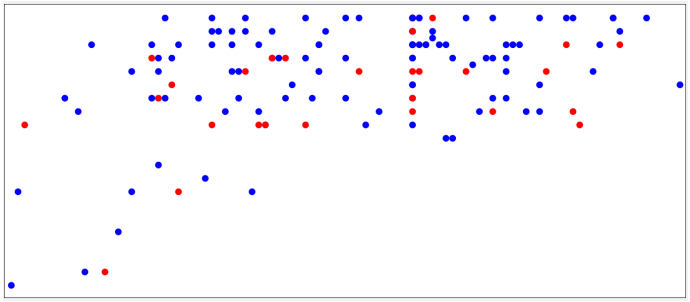

Fig. 23: Scatterplot showing Discrete Math I.

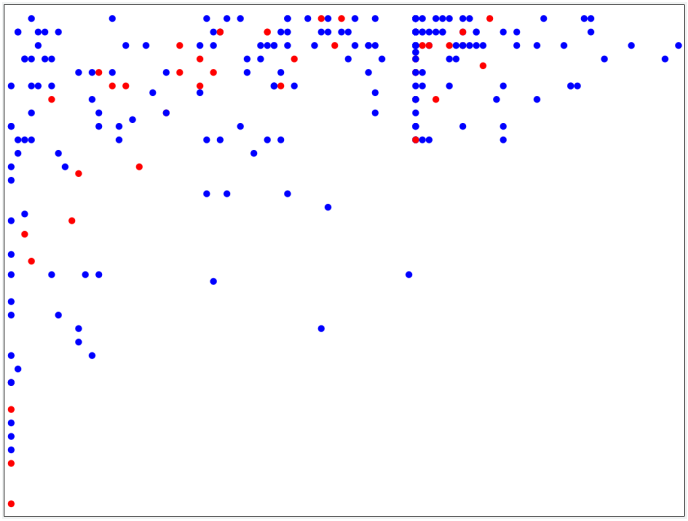

Fig. 24: Scatterplot for Discrete Math II.

As we can see from all of these screenshots, there's a similar behavior in almost all of these classes. That is, the clusters tend to gather at the top-right corner. This result tells us that students who are absent least amount of time, does tend to receive higher grade amongst the class, which does makes sense.

3.3 Parent Education Analysis

This is strictly for our second dataset, where there are total of 35 dimensions. So, it is in our best interest to use isomap to analyze whether a student's parental education has any impact on their grades. As discussed in the previous section, we've mapped the quantitative dimensions into two dimensions, the x and y axis for the dataset. This is done using a python library as well. And again, this will only give us intuition on the quantitative attributes, so we will visualize this in D3.js and add some attributes for our categorical dimensions.

3.3.1 Isomap Visualization

We have decided to use Isomap because Etemadpour et al. [4] has claimed that this high-dimensional reduction technique is effective. To visualize this in D3.js the algorithm is very simple. Since the x-axis and y-axis can be negative, we want to find the min_X, max_X, min_Y, and max_Y to calculate the size of our SVG. Here, we set a condition to check the min_X and min_Y to see if they're less than 0, if they are, we will multiply them by -1 to set them positive. Then, we can calculate the height and width of our SVG using the following formula:

$$width = (min_X + max_X) * 20$$
$$height = (min_Y + max_Y) * 20$$

Notice, we multiply it by a factor of 20 because we want to space out each data point by a factor of 20, because otherwise they're too close to together and we cannot distinguish the clusters. And when we're drawing the circles on the SVG, we simply shift the current x-axis by min_X and y-axis by min_Y, this will satisfy the two dimensional coordinates on the webpage. To represent the father and mother education levels, there are a total of 5 categories, each represented by colors. The father's education will be represented by the circle's stroke color and the mother's education will be represented by the circle's fill color. Here is the layout,

1. Black: No education.
2. Red: Initial education.
3. Orange: Primary education.
4. Green: Secondary education.
5. Blue: Higher education.

Now, we have the essential components of the isomap visualization done. For more visual appeals, I added a sidebar for data filtering, a legend to show the color representations, and a white background for the SVG canvas.

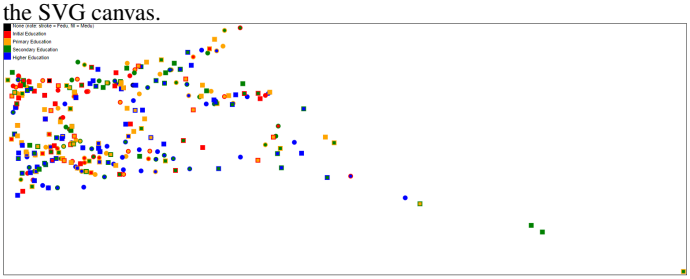

Fig. 25: Isomap for period 1 grades of Math class.

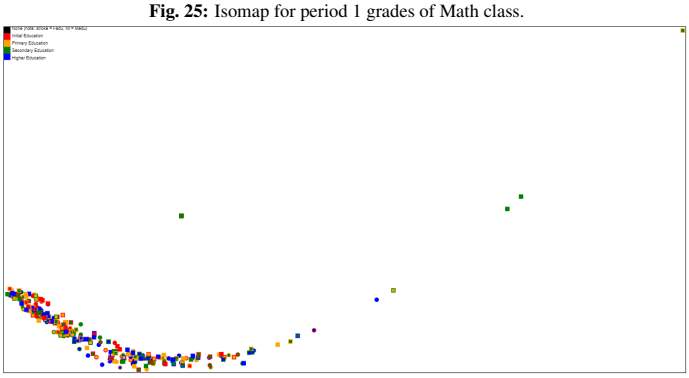

Fig. 26: Isomap for period 2 grades of Math class.

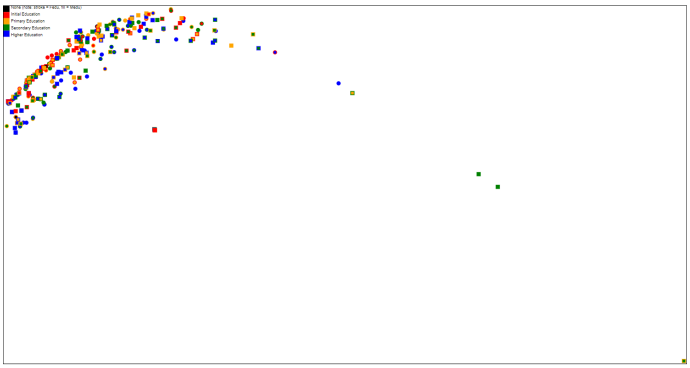

Fig. 27: Isomap for period 3 grades of Math class.

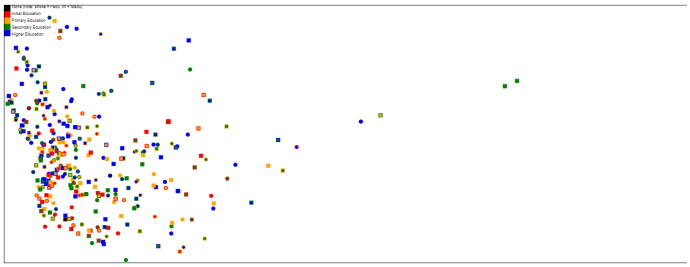

Fig. 28: Isomap for all three periods' grades of Math class.

We have a sidebar menu on this website dedicated for grade period filtering, hence we can see different grading periods for this math class. For interactivity, we've implemented a feature for the user to select any clusters on the screen to display a barchart to compare the grades between the two different genders. This is how it looks like in action:

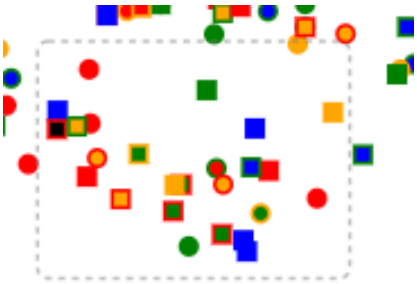

Fig. 29: Selecting cluster on isomap for all three grading periods of math class.

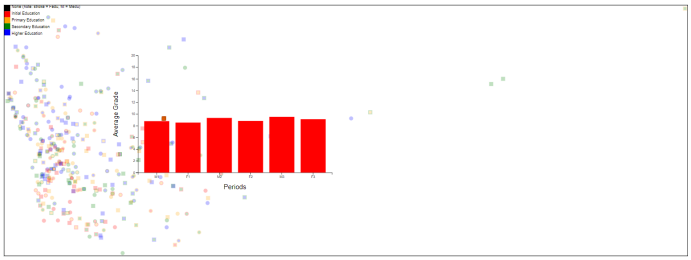

Fig. 30: Barchart displaying data from selected cluster.

As a result, we can compare the gender's grade throughout all three periods of this math class.

3.4 Gender vs Grade Analysis

In gender vs grade analysis, we try to see if there are any correlations between student genders and their grade. We use the following three visualizations to answer our research question: Do male or female tend to do better performance on specific type of courses? Or in other words, do gender affects student's grade on specific course?
For our choise of visualization, the first visualization is a heatmap, second one is a zoomable map, and third one is a dendrogram.

3.4.1 Heatmap

The second research question that we have is the affect of gender on grades. To visualize the possible correlation in the two area we used three heat maps. One of the heat maps will display information for the male, the second one will display the information we had for the female, and the final heat maps will display the significant level of the two datasets, and the value are generated by Z test and double check by the Mann–Whitney U test also know as Wilcoxon signed-rank test.

$$U1 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum_{i=n_1+1}^{n_2} R_i$$

$$U2 = (n_1 * n_2) - U1$$

Where:

1. U=Mann-Whitney U test
2. N1 = sample size one
3. N2= Sample size two
4. Ri = Rank of the sample size

After we calculate the two U value, we will take the smallest of the two value and compare it with the critical value we obtain through search the critical u value table to see if the result obtained through comparing the two group is significant or not. If the value is smaller then the obtain critical value, then the test is significant.

n_2	α	n_1																
		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
3	.05	--	0	0	1	1	2	2	3	3	4	4	5	5	6	6	7	7	8
	.01	--	0	0	0	0	0	0	0	0	1	1	1	2	2	2	2	3	3
4	.05	--	0	1	2	3	4	4	5	6	7	8	9	10	11	11	12	13	14
	.01	--	--	0	0	0	1	1	2	2	3	3	4	5	5	6	6	7	8
5	.05	0	1	2	3	5	6	7	8	9	11	12	13	14	15	17	18	19	20
	.01	--	--	0	1	1	2	3	4	5	6	7	7	8	9	10	11	12	13
6	.05	1	2	3	5	6	8	10	11	13	14	16	17	19	21	22	24	25	27
	.01	--	0	1	2	3	4	5	6	7	9	10	11	12	13	15	16	17	18
7	.05	1	3	5	6	8	10	12	14	16	18	20	22	24	26	28	30	32	34
	.01	--	0	1	3	4	6	7	9	10	12	13	15	16	18	19	21	22	24
8	.05	2	4	6	8	10	13	15	17	19	22	24	26	29	31	34	36	38	41
	.01	--	1	2	4	6	7	9	11	13	15	17	18	20	22	24	26	28	30
9	.05	2	4	7	10	12	15	17	20	23	26	28	31	34	37	39	42	45	48
	.01	0	1	3	5	7	9	11	13	16	18	20	22	24	27	29	31	33	36
10	.05	3	5	8	11	14	17	20	23	26	29	33	36	39	42	45	48	52	55
	.01	0	2	4	6	9	11	13	16	18	21	24	26	29	31	34	37	39	42
11	.05	3	6	9	13	16	19	23	26	30	33	37	40	44	47	51	55	58	62
	.01	0	2	5	7	10	13	16	18	21	24	27	30	33	36	39	42	45	48
12	.05	4	7	11	14	18	22	26	29	33	37	41	45	49	53	57	61	65	69
	.01	1	3	6	9	12	15	18	21	24	27	31	34	37	41	44	47	51	54
13	.05	4	8	12	16	20	24	28	33	37	41	45	50	54	59	63	67	72	76
	.01	1	3	7	10	13	17	20	24	27	31	34	38	42	45	49	53	56	60
14	.05	5	9	13	17	22	26	31	36	40	45	50	55	59	64	67	74	78	83
	.01	1	4	7	11	15	18	22	26	30	34	38	42	46	50	54	58	63	67
15	.05	5	10	14	19	24	29	34	39	44	49	54	59	64	70	75	80	85	90
	.01	2	5	8	12	16	20	24	29	33	37	42	46	51	55	60	64	69	73
16	.05	6	11	15	21	26	31	37	42	47	53	59	64	70	75	81	86	92	98
	.01	2	5	9	13	18	22	27	31	36	41	45	50	55	60	65	70	74	79
17	.05	6	11	17	22	28	34	39	45	51	57	63	67	75	81	87	93	99	105
	.01	2	6	10	15	19	24	29	34	39	44	49	54	60	65	70	75	81	86
18	.05	7	12	18	24	30	36	42	48	55	61	67	74	80	86	93	99	106	112
	.01	2	6	11	16	21	26	31	37	42	47	53	58	64	70	75	81	87	92
19	.05	7	13	19	25	32	38	45	52	58	65	72	78	85	92	99	106	113	119
	.01	3	7	12	17	22	28	33	39	45	51	56	63	69	74	81	87	93	99
20	.05	8	14	20	27	34	41	48	55	62	69	76	83	90	98	105	112	119	127
	.01	3	8	13	18	24	30	36	42	48	54	60	67	73	79	86	92	99	105

Fig. 31: Crtical Value Table U Test

Although the U test alone is enough for confirming if the data groups being compare generate significant result, it is difficult to put in a visualization since the scale is constantly changing with reference to the critical value. Therefore we perform a second test, the Z test, any value greater than 0.05 is insignificant, and any value below 0.05 is significant.

$$StdDev = \sqrt{\frac{(n_1 * n_2) * (n_1 + n_2 + 1)}{12}}$$

$$Z = \frac{U - ((n_1 * n_2)/2)}{(StdDev)}$$

Where:

1. Z= Z score to be search in the z table
2. N1 = sample size one
3. N2= Sample size two
4. StdDev = Standard Deviation

tenths	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	1.00000	0.99202	0.98404	0.97607	0.96809	0.96012	0.95216	0.94419	0.93624	0.92829
0.1	0.92034	0.91241	0.90448	0.89657	0.88866	0.88076	0.87288	0.86501	0.85715	0.84931
0.2	0.84148	0.83367	0.82587	0.81809	0.81033	0.80259	0.79486	0.78716	0.77948	0.77182
0.3	0.76418	0.75656	0.74897	0.74140	0.73386	0.72634	0.71885	0.71138	0.70395	0.69654
0.4	0.68916	0.68181	0.67449	0.66720	0.65994	0.65271	0.64552	0.63836	0.63123	0.62413
0.5	0.61708	0.61005	0.60306	0.59611	0.58920	0.58232	0.57548	0.56868	0.56191	0.55519
0.6	0.54851	0.54186	0.53526	0.52869	0.52217	0.51569	0.50925	0.50286	0.49650	0.49019
0.7	0.48393	0.47770	0.47152	0.46539	0.45930	0.45325	0.44725	0.44130	0.43539	0.42953
0.8	0.42371	0.41794	0.41222	0.40654	0.40091	0.39533	0.38979	0.38430	0.37886	0.37347
0.9	0.36812	0.36282	0.35757	0.35237	0.34722	0.34211	0.33706	0.33205	0.32709	0.32217
1.0	0.31731	0.31250	0.30773	0.30301	0.29834	0.29372	0.28914	0.28462	0.28014	0.27571
1.1	0.27133	0.26700	0.26271	0.25848	0.25429	0.25014	0.24605	0.24200	0.23800	0.23405
1.2	0.23014	0.22628	0.22246	0.21870	0.21498	0.21130	0.20767	0.20408	0.20055	0.19705
1.3	0.19360	0.19020	0.18684	0.18352	0.18025	0.17702	0.17383	0.17069	0.16759	0.16453
1.4	0.16151	0.15854	0.15561	0.15272	0.14987	0.14706	0.14429	0.14156	0.13887	0.13622
1.5	0.13361	0.13104	0.12851	0.12602	0.12356	0.12114	0.11876	0.11642	0.11411	0.11183
1.6	0.10960	0.10740	0.10523	0.10310	0.10101	0.09894	0.09691	0.09492	0.09296	0.09103
1.7	0.08913	0.08727	0.08543	0.08363	0.08186	0.08012	0.07841	0.07673	0.07508	0.07345
1.8	0.07186	0.07030	0.06876	0.06725	0.06577	0.06431	0.06289	0.06148	0.06011	0.05876
1.9	0.05743	0.05613	0.05486	0.05361	0.05238	0.05118	0.05000	0.04884	0.04770	0.04659
2.0	0.04550	0.04443	0.04338	0.04236	0.04135	0.04036	0.03940	0.03845	0.03753	0.03662
2.1	0.03573	0.03486	0.03401	0.03317	0.03235	0.03156	0.03077	0.03001	0.02926	0.02852
2.2	0.02781	0.02711	0.02642	0.02575	0.02509	0.02445	0.02382	0.02321	0.02261	0.02202
2.3	0.02145	0.02089	0.02034	0.01981	0.01928	0.01877	0.01827	0.01779	0.01731	0.01685
2.4	0.01640	0.01595	0.01552	0.01510	0.01469	0.01429	0.01389	0.01351	0.01314	0.01277
2.5	0.01242	0.01207	0.01174	0.01141	0.01109	0.01077	0.01047	0.01017	0.00988	0.00960
2.6	0.00932	0.00905	0.00879	0.00854	0.00829	0.00805	0.00781	0.00759	0.00736	0.00715

Fig. 32: Two tailed Z Table

The Z value obtain through the two function above can be using to search for the p value in the z table, in our case we will be using the two tailed z test table. For example, we have a z value of 1.45, to search for the p-value, you will take the first two value 1.4 and search for it in the vertical axis, then u will look for 0.05 in the horizontal axis that said tenth, the intersection of the two axis will be our p-value.

tenths	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	1.00000	0.99202	0.98404	0.97607	0.96809	0.96012	0.95216	0.94419	0.93624	0.92829
0.1	0.92034	0.91241	0.90448	0.89657	0.88866	0.88076	0.87288	0.86501	0.85715	0.84931
0.2	0.84148	0.83367	0.82587	0.81809	0.81033	0.80259	0.79486	0.78716	0.77948	0.77182
0.3	0.76418	0.75656	0.74897	0.74140	0.73386	0.72634	0.71885	0.71138	0.70395	0.69654
0.4	0.68916	0.68181	0.67449	0.66720	0.65994	0.65271	0.64552	0.63836	0.63123	0.62413
0.5	0.61708	0.61005	0.60306	0.59611	0.58920	0.58232	0.57548	0.56868	0.56191	0.55519
0.6	0.54851	0.54186	0.53526	0.52869	0.52217	0.51569	0.50925	0.50286	0.49650	0.49019
0.7	0.48393	0.47770	0.47152	0.46539	0.45930	0.45325	0.44725	0.44130	0.43539	0.42953
0.8	0.42371	0.41794	0.41222	0.40654	0.40091	0.39533	0.38979	0.38430	0.37886	0.37347
0.9	0.36812	0.36282	0.35757	0.35237	0.34722	0.34211	0.33706	0.33205	0.32709	0.32217
1.0	0.31731	0.31250	0.30773	0.30301	0.29834	0.29372	0.28914	0.28462	0.28014	0.27571
1.1	0.27133	0.26700	0.26271	0.25848	0.25429	0.25014	0.24605	0.24200	0.23800	0.23405
1.2	0.23014	0.22628	0.22246	0.21870	0.21498	0.21130	0.20767	0.20408	0.20055	0.19705
1.3	0.19360	0.19020	0.18684	0.18352	0.18025	0.17702	0.17383	0.17069	0.16759	0.16453
1.4	0.16151	0.15854	0.15561	0.15272	0.14987	0.14706	0.14429	0.14156	0.13887	0.13622
1.5	0.13361	0.13104	0.12851	0.12602	0.12356	0.12114	0.11876	0.11642	0.11411	0.11183
1.6	0.10960	0.10740	0.10523	0.10310	0.10101	0.09894	0.09691	0.09492	0.09296	0.09103
1.7	0.08913	0.08727	0.08543	0.08363	0.08186	0.08012	0.07841	0.07673	0.07508	0.07345
1.8	0.07186	0.07030	0.06876	0.06725	0.06577	0.06431	0.06289	0.06148	0.06011	0.05876
1.9	0.05743	0.05613	0.05486	0.05361	0.05238	0.05118	0.05000	0.04884	0.04770	0.04659
2.0	0.04550	0.04443	0.04338	0.04236	0.04135	0.04036	0.03940	0.03845	0.03753	0.03662
2.1	0.03573	0.03486	0.03401	0.03317	0.03235	0.03156	0.03077	0.03001	0.02926	0.02852
2.2	0.02781	0.02711	0.02642	0.02575	0.02509	0.02445	0.02382	0.02321	0.02261	0.02202
2.3	0.02145	0.02089	0.02034	0.01981	0.01928	0.01877	0.01827	0.01779	0.01731	0.01685
2.4	0.01640	0.01595	0.01552	0.01510	0.01469	0.01429	0.01389	0.01351	0.01314	0.01277
2.5	0.01242	0.01207	0.01174	0.01141	0.01109	0.01077	0.01047	0.01017	0.00988	0.00960
2.6	0.00932	0.00905	0.00879	0.00854	0.00829	0.00805	0.00781	0.00759	0.00736	0.00715

Fig. 33: P Value of 1.45

The following figure is the heatmap visualization for the male population of each course. The color represents the average grade of male students in the course. The darker the color the higher the grade. The male dataset will be default page when you viewer opens our visualization.

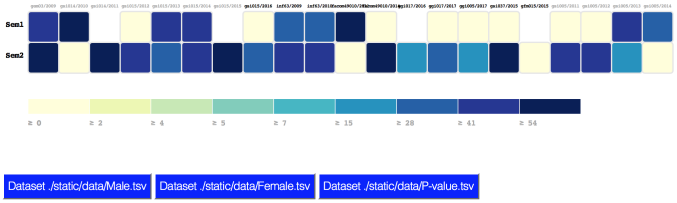

Fig. 34: Heatmap: Male section

The following heatmap visualization for the female population of each courses. The color represents the average grade of female students who took the course. The darker the color the higher the grade. The viewer can access this image by clicking the Female.tsv button under the visualization. To navigate back to the male section of the heat map the viewer can click the Male.tsv, similar for P-value.

Fig. 35: Heatmap: Female

The following heatmap is the p-value heatmap of each course. The white boxes in this visualization represent that the data for this class is not significant, therefore the data of the corresponding class will be ignored when we draw our conclusion. The lighter the color scale in this visualization the more significant the p-value, and the darker the color of the boxes the closer it is to the 0.05 from the left-hand side of the number system. And once the value is over 0.05 it will be replaced with white box to symbolize that selected box is insignificant.

Fig. 36: Heatmap: P-value for each class

The Bar Graph below can be access from clicking any individual square box from the heatmap. Each bar chart contains information of the course that the square box represents. That info includes the individual grade of student who took the course, the gender of that individual student, and a ranking from highest to lowest grade, left most being the highest, rightmost being the lowest. The gender of the student is denoted with F for female and M for male, at the top of each bar. The color scale represent the same as the ranking the darkest color as the highest grade out of 100.

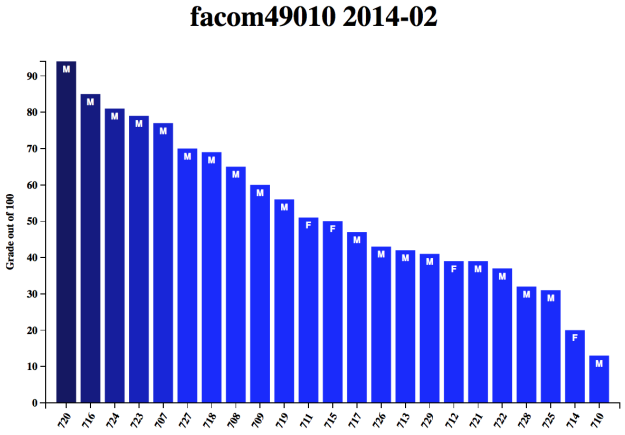

Fig. 37: Heatmap: Bar Chart Pop Out Windows

3.4.2 Zoomable Treemap

We use this zoomable treemap visualization to visualize the class population, the two main point that we will like to outline with this visualization is the difference in population for male and female, and the size of the class. On the first level of the treemap, user can see all the courses' name and each courses' date. When user click on the treemap on a specific course, then the second level of the treemap will display all the dates, number of students in total and gender ratio of that course. When user click on a date of that course, the third level of the treemap will display the number of male and female in that course.

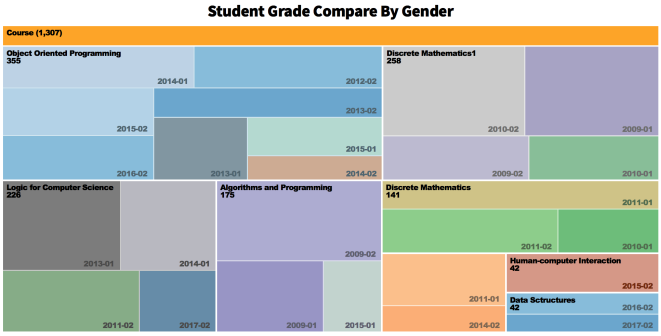

Fig. 38: Zoomable Treemap

Zoomable Treemap walkthrough:

On the first level of treemap, user can see all the course's name and the total number of students took each specific course.

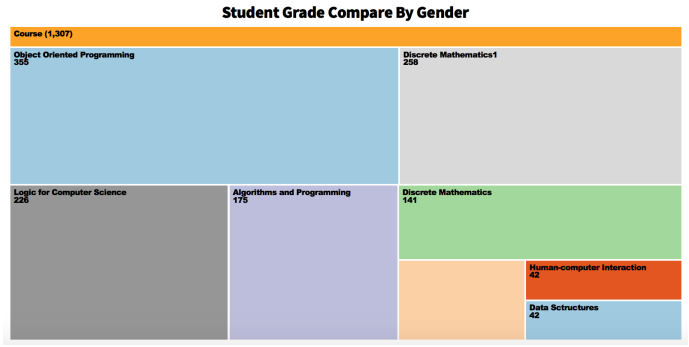

Fig. 39: First level of treemap

For example, if the user wants to see "Object Oriented Programming" course, then he/she can click on the course area to see the next level of this treemap and the information about this course.

Fig. 40: User click on the course "Object Oriented Programming"

Then, the user is able to see different dates/semesters and number of students (the number under date) on each classes of "Object Oriented Programming" course.

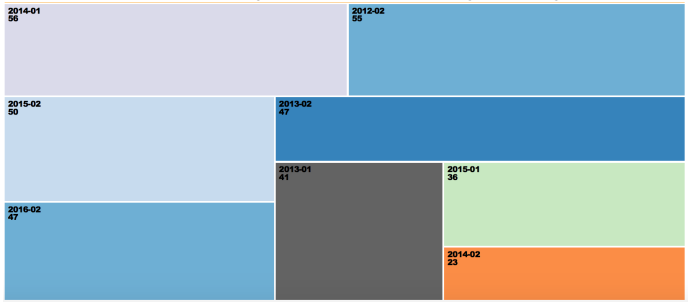

Fig. 41: Second level of treemap

If the user wants to see the information of this course on the first semester of 2014, then the user can just click on that date area to see the next level of this treemap.

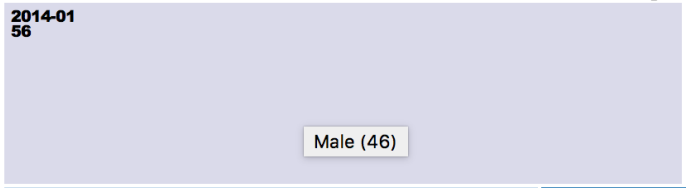

Fig. 42: Click on a date

This is the last level of the treemap, which shows the number of males and females at this class. The size of rectangle represents the ratio between these two genders.

Fig. 43: Third level of treemap

3.4.3 Dendrogram

A dendrogram is a diagram that shows the hierarchical relationship between objects and it is a diagram representing a tree. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters. Why we don't call cluster dendrogram as simple tree? Because dendrogram places the leaves at the same depth, and the tree structure does not have this requirement. We used dendrogram graph as the average course grade of comparison between male and female. Let users could easily to see the advantage and disadvantage for two genders in different subject. The average course grade has been calculated by all combined semester in one subject.

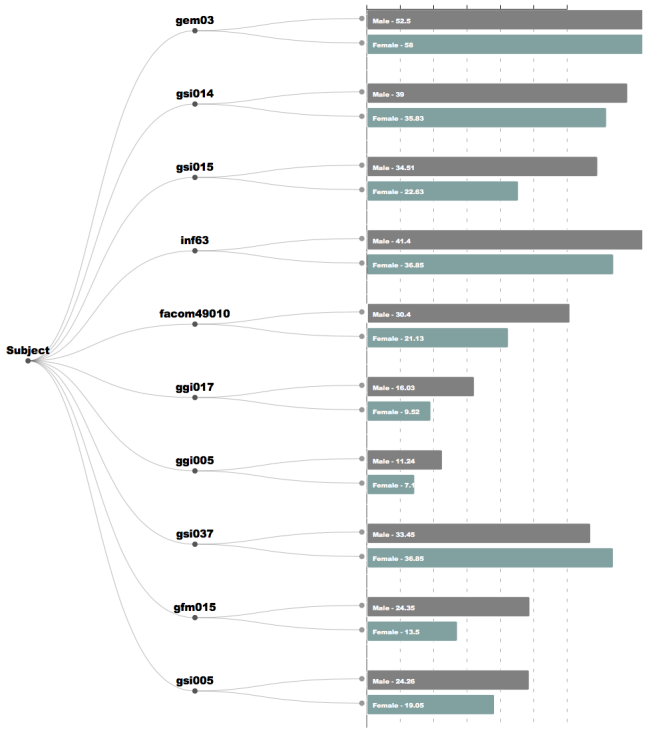

Fig. 44: Dendrogram

3.5 Logistic Regression

Logistic Regression is primarily a data classification technique; it can be used to separate data points into one of two or more classes. In the present study, we used it to identify students as either "passed" or "failed." To perform logistic regression, we first randomly separated the data into two smaller data sets: a "training" set containing 80 precent of the original data, and a "test" set composed of the remaining 20 precent of the original data. It then fits the training set to the explanatory values (e.g., GPA) to find a best fit probability function (e.g., the probability of passing as a function of GPA). Then we assigned a threshold to make predictions, based on the test set, of whether the student will be successful, e.g., we required that the probability of passing must be greater than 50 percent. To evaluate the predictions, the model generates a confusion matrix based on the test set:

$$\begin{bmatrix} FN & TP \\ TN & FP \end{bmatrix}$$

As depicted on the following figure, the confusion matrix is made up of four categories that our predictions can fall into. They can either be: True Positives (TP), False Positives (FP), True Negatives (TN), or False Negatives (FN). Using the confusion matrix, we evaluated the correlation between the explanatory variables and the independent variable. These four categories are plugged into formulas in order to get the Accuracy, Precision, Fallout and Recall of the model. We used the following equations:

$$A = \frac{TP+TN}{TP+TN+FP+FN}$$

$$P = \frac{TP}{TP+FP}$$

$$F = \frac{FP}{TN+FP}$$

$$R = \frac{TP}{TP+FN}$$

Useful predictors can be identified by having high accuracy, recall, and precision alongside with low fallout. Predictors can be good at identifying success but not failure, or they can be good at identifying failure but not success. The best predictor would be able to identify both success and failure simultaneously.

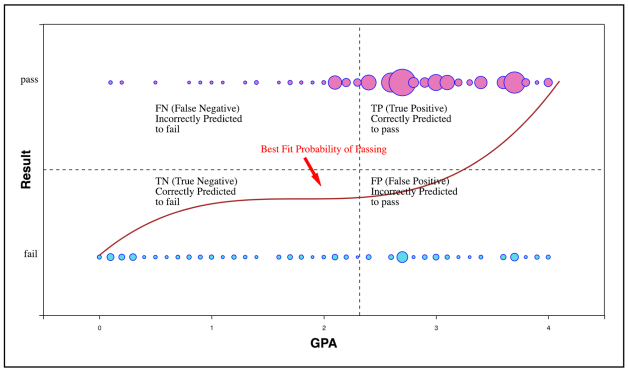

Fig. 45: Logistic Regression Visualization

The following steps were taken when making the visualization above:

- Analyzed the Data Structure.
- Extracted the necessary data fields for the entire data.
- Classified data records by pass and fail as well by the GPA.
- Drew bubbles whose radius stands for total number of students having specific GPA.
- Calculated the start angle and end angle of pie based on the number of male and female.
- Drew bubble at the center of logistic chart area.

A threshold of 2.32 is assigned to make predictions, based on the calculations made on the dataset, of whether or not the student will be successful.

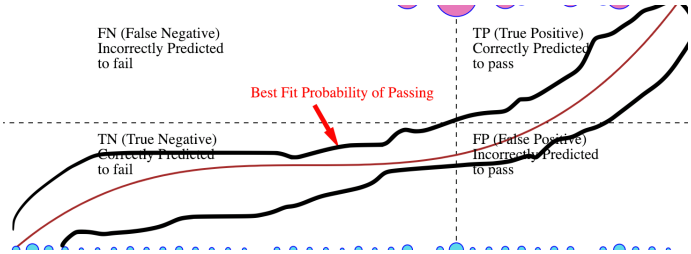

Fig. 46: Best fit probability of passing

In order to map predicted values to probabilities, we use the sigmoid function. The function maps any real value into another value between 0 and 1. In machine learning, we used sigmoid to map predictions to probabilities.

$$S(z) = \frac{1}{1+e^{-z}}$$

1. s(z) = output between 0 and 1 (probability estimate)

2. z = input to the function (your algorithm's prediction)

3. e = base of natural log

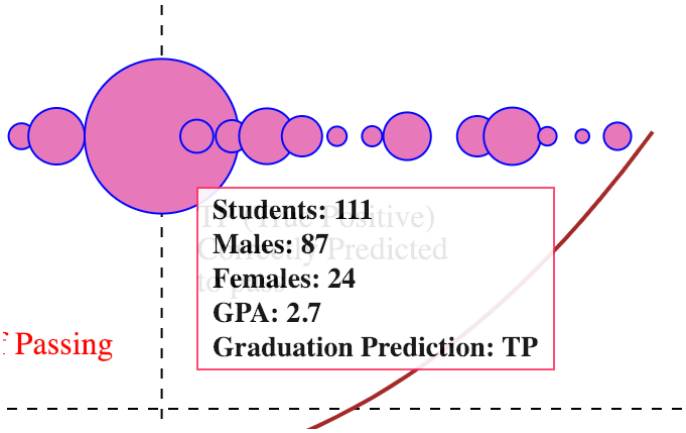

Fig. 47: Tooltip Feature Describing the Bubble

The main purpose of this Logistic Regression Visualization is that it shows the prediction of student's grades based on semesters attempted at college. To implement this visualization, I used the conventional method of JavaScript and D3.js as it was an efficient way to get it done. The Y-axis of the visualization indicates whether a student passed or failed. The X- axis however shows the grade point average of the students based on their performance. The visualization is divided into 4 categories in which confused matrix took place. For pass there is false negative which means that the students were incorrectly predicted to fail and then there is true positive meaning that the students were correctly predicted to pass. For the fail part there is true negative which means the students were correctly predicted to fail and there is false positive in which it states that the students were incorrectly predicted to pass. The line curve represents the probability of passing for the students in which the sigmoid activation method is used. The size of the bubbles indicates the number of students. The bigger means more students and the smaller means less students. In order to make this visualization interactive, zoomed feature is used while touching the bubbles and the bubbles show a tooltip in which all the details are mentioned.

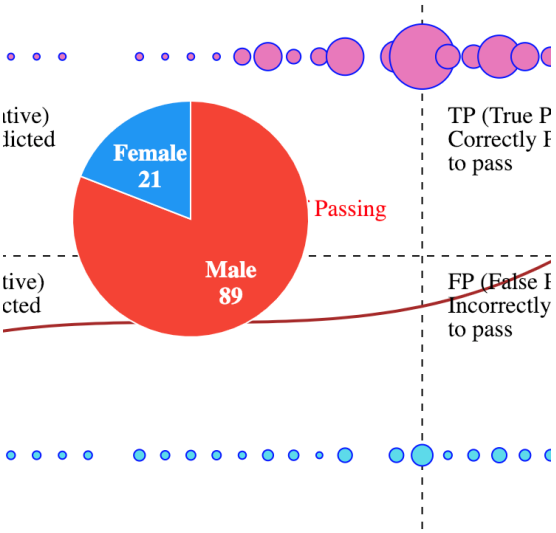

Fig. 48: Popup Dialog Pie Chart

The Popup Dialog feature was implemented, so the pie chart for each bubble can be seen upon clicking on it. The red color indicates male and the blue indicates female. This feature will be useful as it helps to apprehend the number of males and females quickly. In order to use this feature, the user will have to manually click on the bubble and if the user wants to get rid of the pie chart then it can be done by clicking on the showing pie chart.

3.5.1 Pie Chart For Grade Analysis

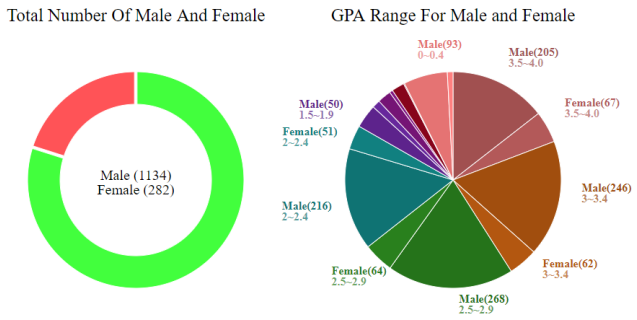

Fig. 49: Pie Chart Visualization

Pie chart is a type of graph in which a circle is divided into sectors that each represent a proportion of the whole. The purpose of this pie chart is to show the total number of male and female students as well as their GPA ranges in order to show a clear indication of their performance. This visualization is enhanced with tooltip feature in which details can be obtained by hovering on the slices of the pie.

4 Results

4.1 Attendance Analysis

From analyzing our machine learning heatmaps, we couldn't find any useful observations because the accuracy was at maximum of 40 percent using random forest regressor and the other models are even lower than that. From our isomaps, we can draw conclusion that for the class Algorithms and Programming I, if a student is never absent, then he or she will most likely receive a grade between 65 and 71. For Algorithms and Programming II, if a student is never absent, then he or she will most like recieve a grade above 90. For Algorithms and Programming III, we cannot conclude anything because the clusters are too small. For Discrete Math I, we cannot conclude anything because the clusters are too scattered. For Discrete Math II, it seems that the attendance doesn't matter, and the students will most likely receive a grade in the 60s. For Logic for Computer Science II, if a student is

never absent, then he or she will receive a grade in the 70s. For OOP class, it seems that if a student is absent for less than 8 times, then he or she will pass with a grade from 60 to 80. For Human-Computer Interaction, attendance doesn't seem to matter, and the student will pass with a high grade regardless.

4.2 Parent Education Analysis

From analyzing the isomap for the second dataset, it is really obvious that students with parents that have higher education has received grades above 10 (highest is 20). While students whose parents have lesser education mostly received grades under 10 with few outliers. And it seems that student's whose parent's jobs are at home or teacher also receive higher grades than other students. This statistic applies to both math and portuguese classes. From our tooltip, we also conclude that parent's with higher education end up becoming a teacher. But there were many parent's whose job is labeled as "other" which is not very helpful for our analysis.

4.3 Gender Statistical Analysis

From our analysis of those the gender vs grade heatmap visualizations, we can see that the student's gender does influence the type of course in question. Thru analyzing the datasets on our heatmap visualization we often see one gender group perform better than the other gender in similar type of course. For example, let's take look at the zoomable heatmap of Student Grade versus Gender. When we click on the heatmap and focus on the Discrete Mathematics 1 for additional information, the heatmap will open a pop window and display the dates that this course was taken, the average grade of male and female for that course and the grades that they receive, and the population of male and female in that course. By observing all the element, we can easily observe for the course Discrete Mathematics1, there is more male student then female student and male student have overall a higher average then the female student in that course.

4.4 Logistic Regression Analysis

From analyzing the logistic regression, it is evident that there is mixture of both correct and incorrect prediction. The students with lower grades initially don't mean that they won't pass. Same logic works for the students with higher grades, as it doesn't guarantee an overall pass. The tooltip used in the visualization shows the prediction status, number of students for a specific GPA criterion. Above all, it can be concluded that the students always have chances to pass if they keep themselves active and motivated.

5 Conclusion and Future Works

Although we've got some interesting results from our analysis, we didn't have get chance to ask a few students to test how effective our analysis are. For our future work, we plan on expanding our knowledge on machine learning to analyze other areas of education to help and motivate students for for academic success.

6 Acknowledgements

We thank our professor Ronak Etemadpour for contacting her colleague, Dr. Jose Gustav Paiva who's a professor in another university to give us the first dataset for our research, as well as giving us many suggestions and improvements along the way.

References

[1] A. B. Bellcore, J. A. Mcdonald, J. Michalak, and W. Stuetzle. Interactive data visualization using focusing and linking. *1991 IEEE*, 1991.

[2] M. A. Borkin, Z. Yan, B. Horn, L. Roe, and B. Berkey. Visualization education through social impact: A service-learning approach for visualization pedagogy. *IEEE VIS 2017*, 2017.

[3] M. B. DeCotes. Mark blaise data analytics of university student records. *Trace: Tennessee Research and Creative Exchange*, 2014.

[4] J. Martinez and A. Miller. Analysis of student success. *Mission College Undergraduate Research Journal (MURJ)*, 2015.

[5] B. E. Shapiro and C. A. Shubin. Does instructor matter? grade variation among math courses at csun, 2005-2014. *Data Science and Mathematical Modelling Research Report*, 2015.

[6] K. N. Singh and R. D. Wajgi. Data analysis and visualization of sales data using data mining techniques. *International Journal of Innovative Research in Computer and Communication Engineer*, 2016.

[7] J. A. P. T. T. W. A. vanWeringh David S. Guttman and N. J. Provart. Gene slider: sequence logo interactive data-visualization for education and research. *University of Toronto*, 2016.