Project Report on

ANALYSIS AND GRADE PREDICTION OF CANVAS NETWORK PERSON-COURSE DE-IDENTIFIED OPEN DATASET

State University of New York

Albany

Department of Computer Science

By

RASWITHA NARLA	ID: 001448377
HARSHINI PALAVAI	ID: 001434107
SRILASYA SAMUDRALA	ID: 001448701

CONTENTS:

1. PROBLEM FORMULATION:

One of the biggest challenges faced with online courses is to know what are the factors that help students perform better compared to traditional classroom study. Since the presence of students is not physically seen, it is hard to ascertain what makes the student perform better and have successful online course experience. In our project, we take a Canvas Network Person Course De-Identified Open Dataset, where we have attributes that were involved in online course along with the grade of the student. With the help of that dataset, we analyse the factors and train the data. Then with those data predict the grade. On successfully performing grade prediction, we will be finding the factors that enable good performance for student in online course, so that future students can take note of those factors and perform better.The objective of the project is to predict the accuracy percentages of the grade using three different methods which are Random forest, Bayesian ridge and Linear regression models.

In addition to that we also Analyse the Dataset and do certain visualizations using Tableau Software where we see how certain attributes affect the grades and how each column is related. It also helps in giving the insights for the course creators and students to improve their grade by viewing the results. Often, times people bypass small details in a large dataset, by doing visualizations we are summarizing the complex dataset so that it is easier to understand. The visualizations help future users to get an overview and help in getting accuracy for prediction.

Our project is broadly classified into two parts:

- Analysis
- Prediction

2. MAIN TECHNIQUES USED FOR PREDICTION:

2.1 DATACLEANING AND PREPROCESSING:

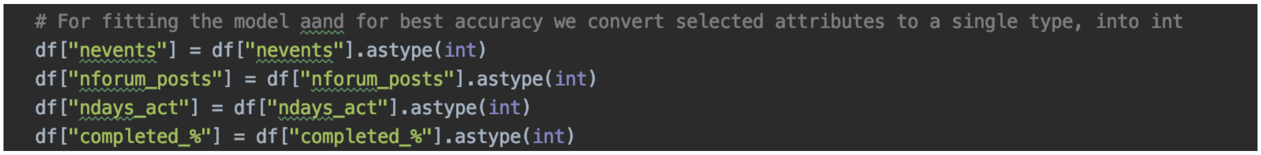
The first step when we choose a dataset is data cleaning. Basically, data cleaning is a process of identifying all the unethical or missing data in the dataset. This is used in datasets where irrelevant or improper data is found and then that data is either removed, modified or replaced. This process is mainly to improve the data quality.

Here in our CNPC dataset, we have performed data cleaning.

- Where we replaced irrelevant data with NAN, then replaced them with 0.
- The given dataset contains text and categorical values, some algorithms can handle categorical values very well but most of them expect numerical values to achieve state-of-art results. For converting these string values to categorical values, we have used Label Encoding.
- First, we have replaced the string values in “LOE_DI” with numbers and then performed Label Encoding. Similarly, for "learner_type", I have reduced the content values to 3, i.e. to Active, Passive, Observer and then performed Label encoding. For “expected_hours_week”, I have directly used Label Encoding to convert string values to integers.

COLUMNS	Values before Label Encoding	Values After Label Encoding
LOE_DI	Master's Degree	1
	Completed 4-year college degree	2
	Some college, but have not finished a degree	3
	Some graduate school	4
	Ph.D., J.D., or M.D. (or equivalent)	5
	Completed 2-year college degree	6
	High School or College Preparatory School	7
learner_type	Active: Active Participant	1
	Passive: Passive Participant	3
	Observer: Drop-in	2
	Missing	0
expected_hours_week	Between 4 and 6 hours	1
	Between 1 and 2 hours	2
	Between 2 and 4 hours	3
	More than 8 hours per week	4
	Less than 1 hour	5
	Between 6 and 8 hours	6

We have normalized some columns using astype, we have converted columns with Dtype float to int for columns nevents, nforum_posts and completed%.



```
# For fitting the model aand for best accuracy we convert selected attributes to a single type, into int
df["nevents"] = df["nevents"].astype(int)
df["nforum_posts"] = df["nforum_posts"].astype(int)
df["ndays_act"] = df["ndays_act"].astype(int)
df["completed_%"] = df["completed_%"].astype(int)
```

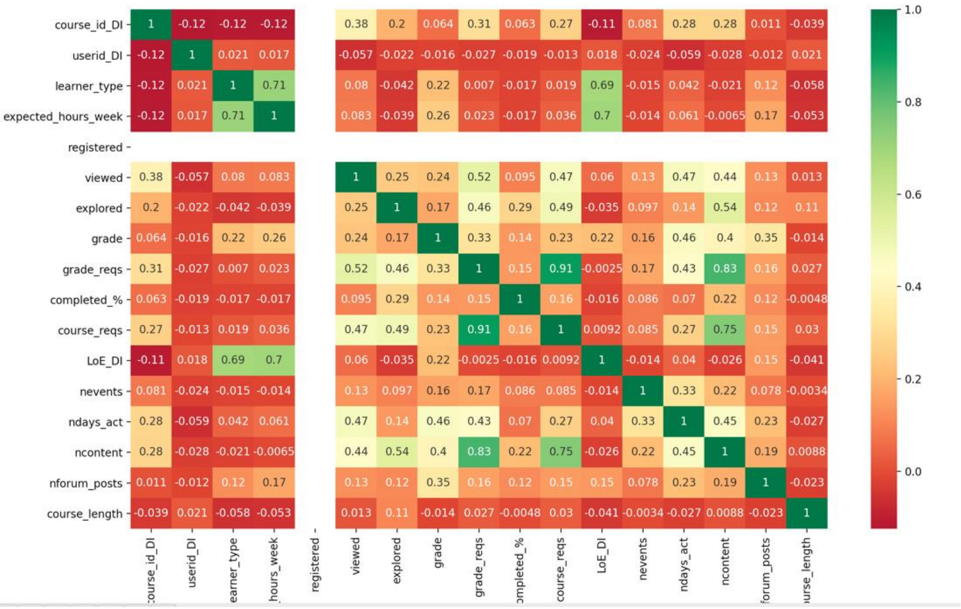
2.2 FEATURE SELECTION:

Feature selection is the process which reduces the number of the input variables and select the subset which consists of relevant features (variables, predictors) for developing a predictive model. The advantage of the feature selection is it helps an algorithm to train faster and reduces the complexity of a model and makes it easier to interpret.

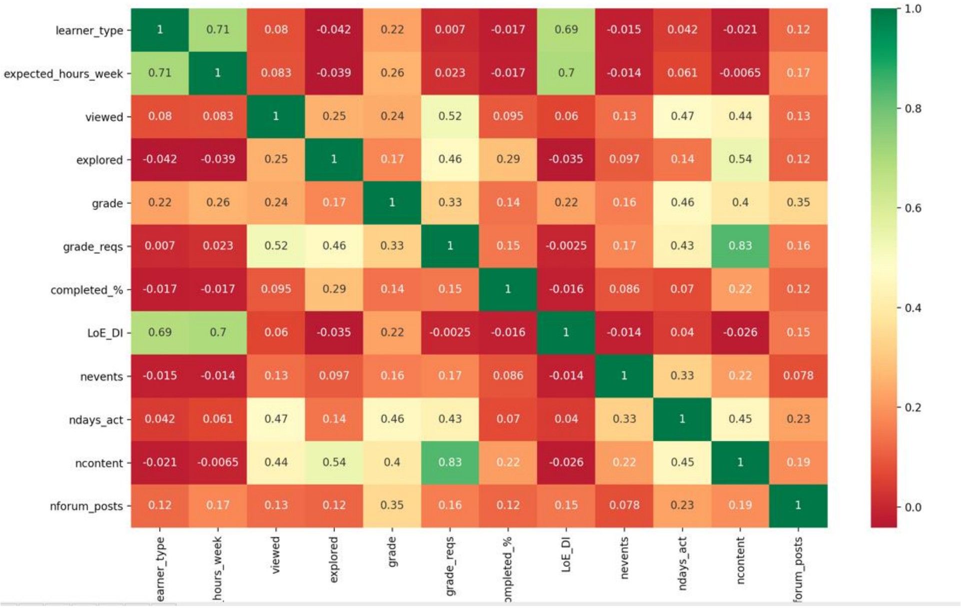
In our feature selection function, we have selected the certain features based on the two feature selection methods which are: correlation using heat map and generic univariate selection.

2.2.1 CORRELATION USING HEAT MAP (PEARSON CORRELATION):

A heat map is a data visualization technique which describes about the magnitude of a phenomenon in two dimensions with different colours. In the heat map which is generated we got values of how much each feature varies with grade. The variation in the colour may be hue or intensity, giving obvious visual cues to the reader about how the phenomenon is clustered or varies over space.



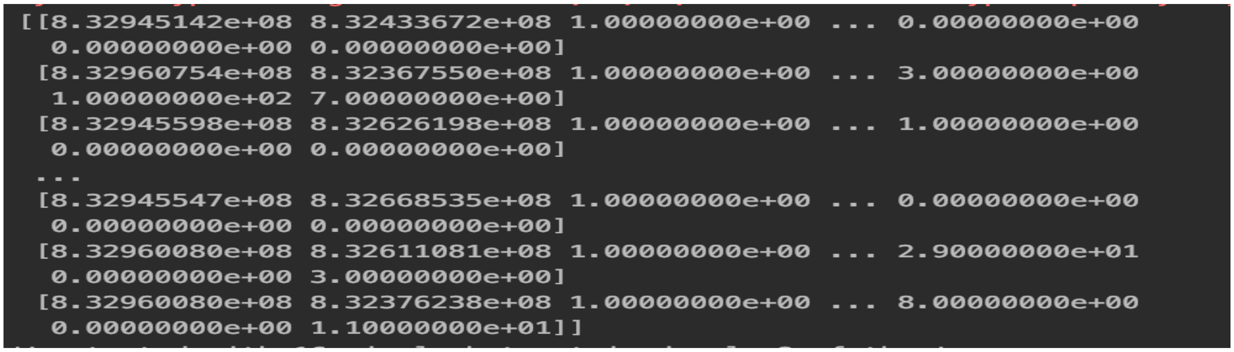
By analysing the values, we have dropped the columns which have negative impact or low impact on the grade and considered the features which have high impact with respect to the grade. For this function, the input provided first is the dataset which contains all the 26 features, we computed and analysed the heat map then we dropped specific columns, showed the plot of the heat map i.e. correlation matrix after dropping specific columns.



2.2.2 GENERIC UNIVARIATE SELECT:

GenericUnivariateSelect is a sklearn.feature_selection tool that allows you to select features from a dataset using a scoring function. It supports selecting columns in one of a few different configurations. Using this function revolves around inputting a function that takes the X and y arrays, performs some kind of statistical test on the values, and then returns the score per feature in X.

We have inputted 16 features and found 8 features which have maximum score by using this function.



```
[[8.32945142e+08 8.32433672e+08 1.00000000e+00 ... 0.00000000e+00
  0.00000000e+00 0.00000000e+00]
 [8.32960754e+08 8.32367550e+08 1.00000000e+00 ... 3.00000000e+00
  1.00000000e+02 7.00000000e+00]
 [8.32945598e+08 8.32626198e+08 1.00000000e+00 ... 1.00000000e+00
  0.00000000e+00 0.00000000e+00]
 ...
 [8.32945547e+08 8.32668535e+08 1.00000000e+00 ... 0.00000000e+00
  0.00000000e+00 0.00000000e+00]
 [8.32960080e+08 8.32611081e+08 1.00000000e+00 ... 2.90000000e+01
  0.00000000e+00 3.00000000e+00]
 [8.32960080e+08 8.32376238e+08 1.00000000e+00 ... 8.00000000e+00
  0.00000000e+00 1.10000000e+01]]
```

Here, we have dropped features that we found were not important after going through all the features in the dataset. Then we have inputted the final dataset to the Generic Univariate Select function so as to know which features vary or are mostly related to the feature “grade” and apply them to compute Accuracy and RMSE values for all the Regressors.

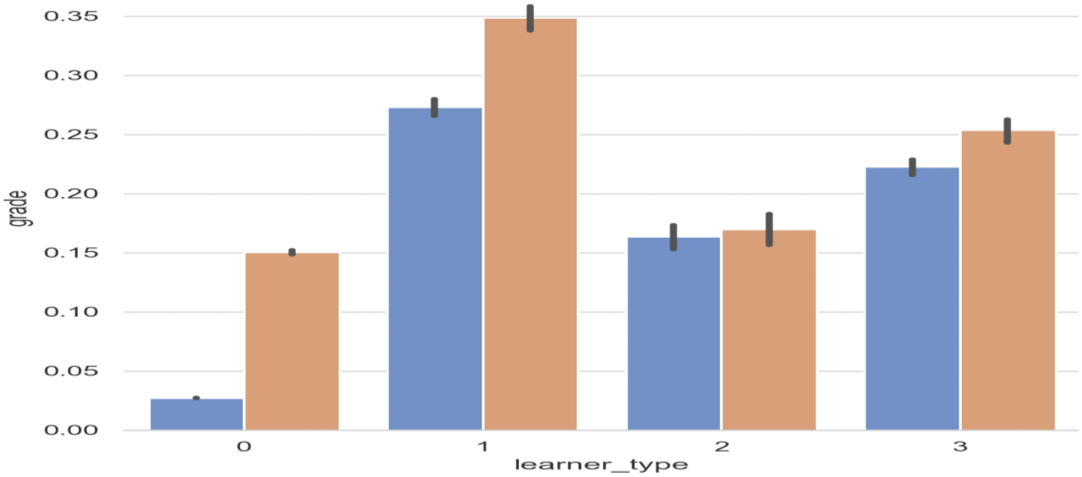
AFTER FEATURE SELECTION METHODS

After applying both of the feature selection methods, we have dropped these particular columns

- 'registered',
- 'primary_reason',
- 'start_time_DI',
- 'last_event_DI',
- 'course_reqs',
- 'final_cc_cname_DI',
- 'age_DI',
- 'gender',
- 'course_start',
- 'course_end',
- 'course_length',
- 'course_id_DI',
- 'discipline',
- 'userid_DI'

2.3 DATA VISUALIZATION GRAPH

This graph predicted how each student scores based on the learner_type and the interactions he/she performed during the course i.e. viewed.



ORANGE - VIEWED (1)
BLUE – VIEWED (0)

This graph detected that a student who is Active and who had interacted during the course scored high (grade high), while a student who is Active and did not interact during the course have a lower grade.
Similarly, for the Observer and Passive student this graph shows how a student scores based on the interactions he/she had made during the course.

2.4 METHODS USED FOR PREDICTION

2.4.1 RANDOM FOREST METHOD:

Random forest is a supervised learning algorithm which is capable of performing both regression and classification tasks. In regression this will have the effect of smoothing the model. In our project we are using random forest regressor which is an meta estimator that fits sub samples of the dataset and uses averaging technique to improve the predicative accuracy. By using random forest, we can handle large data sets with higher dimensionality and estimates missing data and maintains accuracy when large proportion of the data are missing. In the random forest function, we have given two parameters which are X (X represents the features which are considered after the deductions of the other features which has less impact on the grade) and y (which indicates grade).Output of the functions is accuracy percentage for the grade and RMSE value (Root Mean Square Error (RMSE) measures how much error there is between two data sets. In other words, it compares a predicted value).

2.4.2 BAYESIAN RIDGE METHOD:

This method estimates a probabilistic model of the regression. In this we formulate the regression using the probability distributions rather than point estimates and statistical analysis is undertaken with the context of Bayesian inference. In our project the Bayesian ridge method takes two parameters which are X(X represents the features which are considered after the deductions of the other features which has less impact on the grade) and y (which indicates grade).Output of the functions is accuracy percentage for the grade and RMSE value (Root Mean Square Error (RMSE) measures how much error there is between two data sets. In other words, it compares a predicted value). When compared to the RMSE value of the random forest method the RMSE value for the Bayesian ridge method is more and accuracy grade is also high for the Bayesian ridge model.

2.4.3 LINEAR REGRESSION METHOD:

Linear regression method is a linear approach to model the relationship between the scalar variables which is used to predict the relationship between two factors and the factors that are used to predict the value of the dependent variable are called as the independent variable. In our project we will explain the relationship between two features by the observed data and in the linear regression function we took two parameters which are X(X represents the features which are considered after the deductions of the other features which has less impact on the grade) and y (which indicates grade).Output of the functions is accuracy percentage for the grade and RMSE value (Root Mean Square Error (RMSE) measures how much error there is between two data sets. In other words, it compares a predicted value). When compared to the RMSE value of the random forest method the RMSE value for the linear regression method is more.

3. EXPERIMENTAL RESULTS AND EVALUATION

As a result of Feature selection using both the methods specified above, we have considered

X = viewed, grade_reqs, learner_type, nforum_posts, ndays_act, ncontent, LoE_DI, explored, nevents, expected_hours_week.

Y = grade

Here we split the training and test data in the percentage of 70 and 30. We used standard scalar to normalize all the values.

RANDOM FOREST METHOD:

This model takes input arguments as X, Y and then standardize the data. We implemented the Random Forest Regressor and computed the pred_rfc i.e. grade prediction values with respect to X_test and the RMSE value.

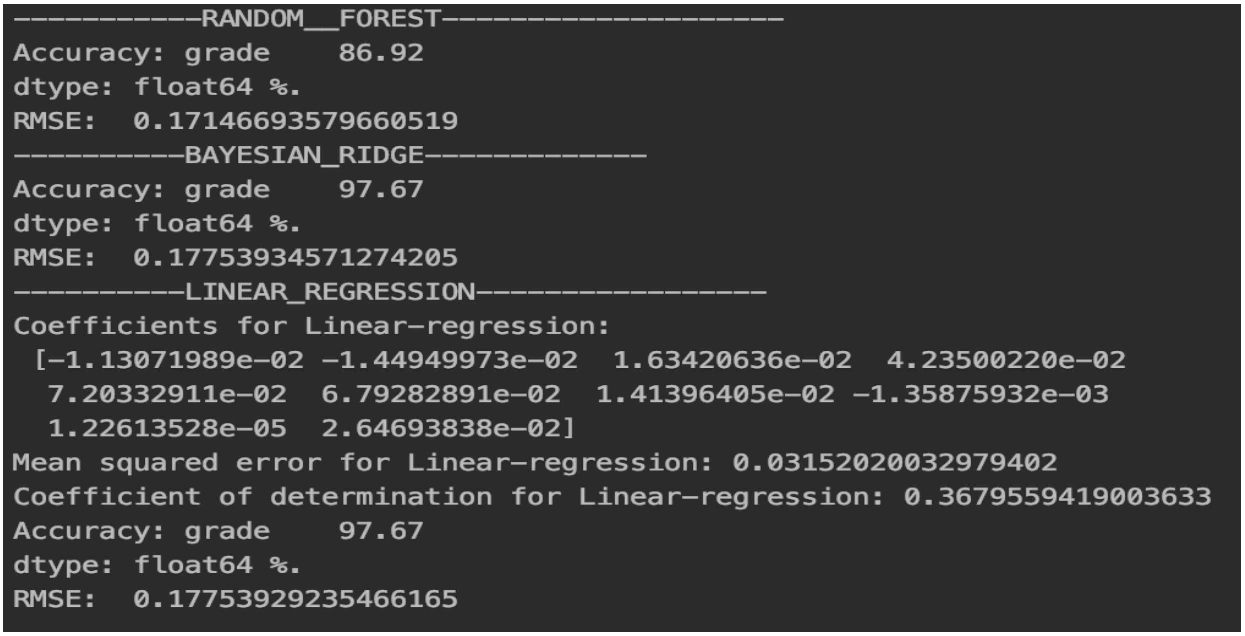
BAYESIAN RIDGE METHOD:

This model takes input arguments as X, Y and then standardize the data. We implemented the Bayesian Ridge and computed the y_pred_gnb i.e. grade prediction values with respect to X_test and the RMSE value.

LINEAR REGRESSION METHOD:

This model takes input arguments as X, Y and then standardize the data. We implemented the Linear Regression Model and computed the y_pred_lr i.e. grade prediction values with respect to X_test, coefficients, mean squared error, coefficient of determination and the RMSE value.

Method Name	Accuracy percentage for the grade	RMSE Value
Random Forest	86.92	0.17190869055032257
Bayesian Ridge	97.67	0.17755474689884654
Linear Regression	97.67	0.17755468853361134



```
-----------RANDOM__FOREST--------------------
Accuracy: grade    86.92
dtype: float64 %.
RMSE:  0.17146693579660519
----------BAYESIAN_RIDGE-------------
Accuracy: grade    97.67
dtype: float64 %.
RMSE:  0.17753934571274205
----------LINEAR_REGRESSION-----------------
Coefficients for Linear-regression:
 [-1.13071989e-02 -1.44949973e-02  1.63420636e-02  4.23500220e-02
  7.20332911e-02  6.79282891e-02  1.41396405e-02 -1.35875932e-03
  1.22613528e-05  2.64693838e-02]
Mean squared error for Linear-regression: 0.03152020032979402
Coefficient of determination for Linear-regression: 0.3679559419003633
Accuracy: grade    97.67
dtype: float64 %.
RMSE:  0.17753929235466165
```

4. VIZUALIZATION TECHNIQUES:

4.1 Tableau Analysis:

Datasets are often having many columns and rows and it is difficult for us to analyze the data through text or csv files. When we have visualizations about how columns are co-related. It becomes easier for us to dissect the data and process them. There are many Visualization tools that are available for us to proceed. One such highly recommended one is Tableau Analytics which is Business Intelligence tool that helps us analyze our Canvas Network Person-Course Open Identified Open Data set. It is usually done in the following steps:

1) **Cleaning Data using Tableau Prep:**

The dataset we have is raw and has a lot of missing values and unprocessed data. Some values in the dataset are incomplete and some duplicate. Having processed dataset is essential for correct Analysis. So, the first step we do is process the data. We use Tableau Prep Software to clean the data. Here we remove the missing values and then perform join operation for duplicate values. Some of the columns like learner type and expected hours are relabeled correctly. Few unnecessary columns like age, gender, and course ID were deleted. Now, we run the flow and output a new output csv file.

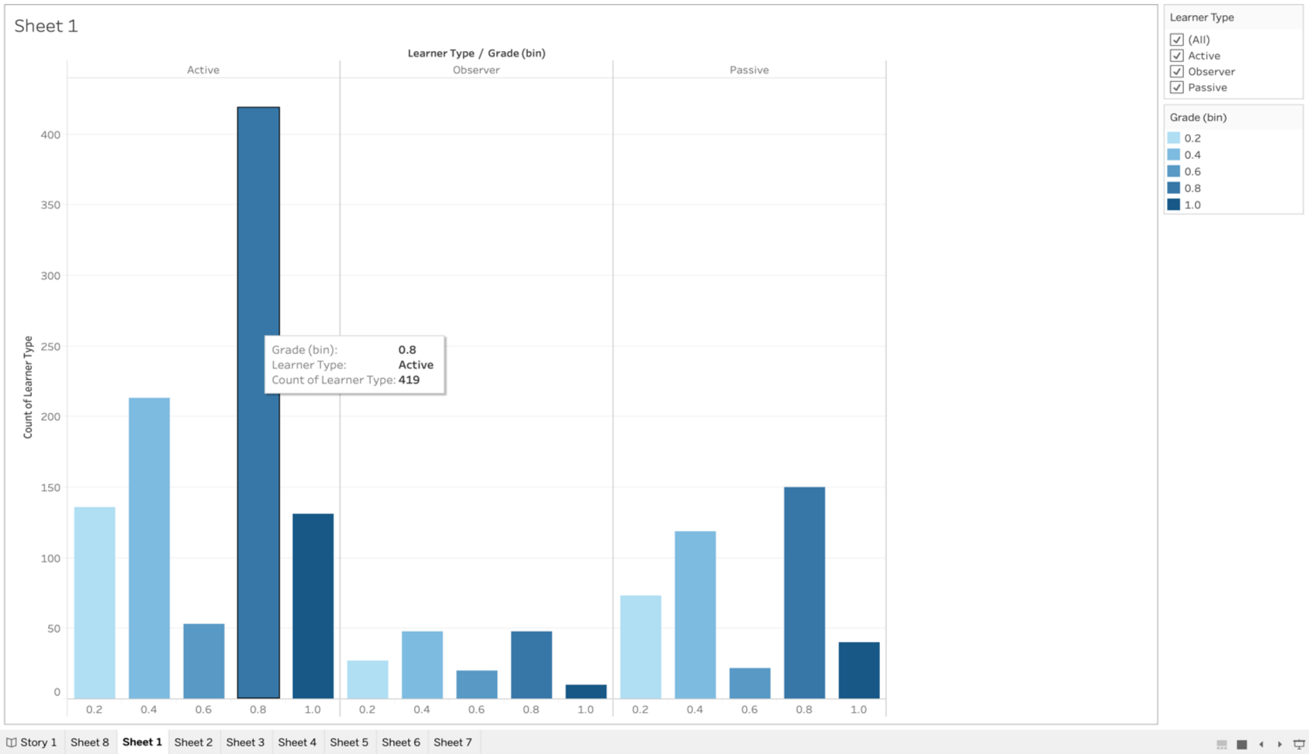
2) **Analysis using Tableau Desktop that we used for our analysis:**

Tableau Desktop is data visualization software It helps us in creating charts so that it is easy to understand data. The following are the resulting visualizations that we obtained:

- **Learner Type/Grade:** Here we divide the grade into 5 bin each with size 0.2 and plot the count of grade scored by various types of learners. We observe that Active Learners have high count of grade for 0.8 grade. Observers have low count of grade for all the bins. This gives us the idea that to get to good grade, we must be active learners.
- **Expected Hours per week/Grade:** Similar to the above one we plot Expected hours against the grade bin and find the count. This is to see if amount of time spent by student reflects the grade they obtain. The result we found were proportional and that to get a good grade range of 0.8, we must spend at least 2-4 hours on the course per week.

- **Views/Average Grade:** If the number of interactions within the course are greater than 1, we assign them 1 else 0. We find the average grade obtained for interactive courses is 0.44 vs non-interactive courses which is 0.27. We can infer from this that we get good grades for interactive courses and that it is best for a course to have interactive section in them so that students pay attention.
- **Explore/Course Length:** If the user interacted with more than 50 % of the course, we give it value 1 else 0. Now we use packed bubbles to find it against the course length against explored and found that the average course length is 67.11, then more exploration is done by students.
- **Discipline/ Expected Hours per week/ Average Grade:** Using Gnatt View, we find the various disciplines of the users and the average grade obtained by them. Moreover, we also check the amount of time spent by the user of a particular discipline and what their corresponding average grade is. We observed that Computer Science students need to spend minimum 8 hours per week to get grade of 0.8 while Mathematics department needs only 2-4 hours.
- **NContent/Nforumposts/Average Grade:** Another major attribute is forum posts which paid a vital part in getting grade. So, the number of posts made in discussion forum for a particular course against average grade and we observe that Nforum post is a vital attribute that affects grade whereas ncontents fluctuates and doesn’t affect the grade.

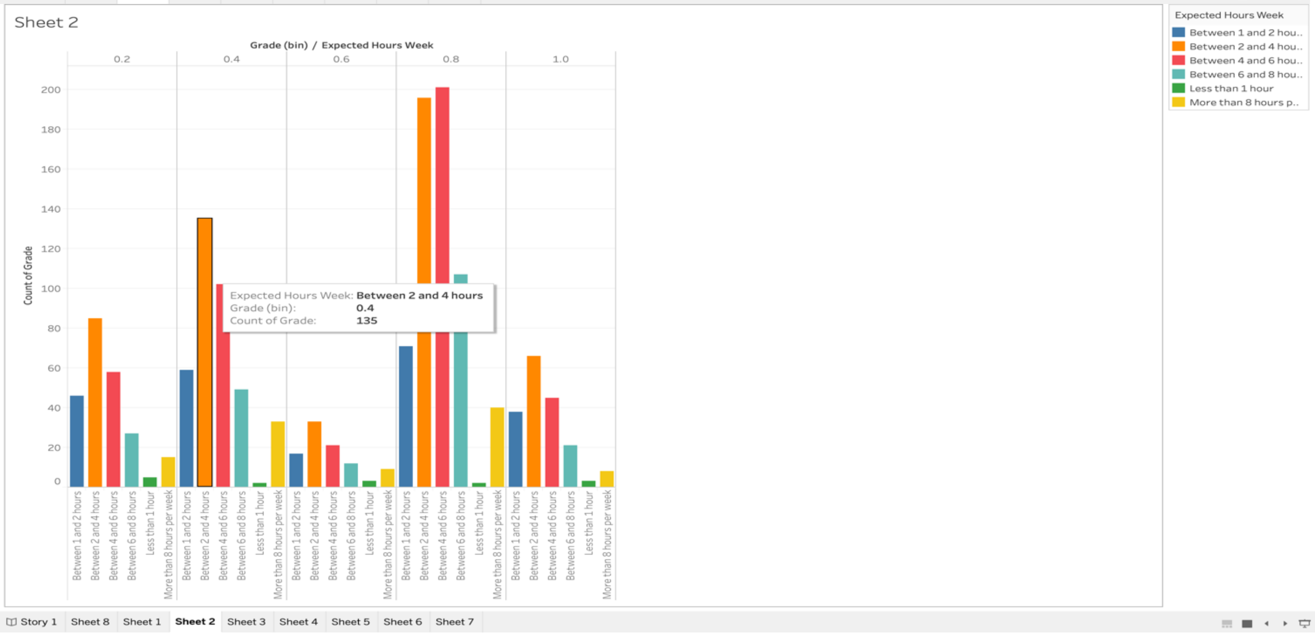
4.2 Output Visualization Screens:



Grade (bin): 0.8
Learner Type: Active
Count of Learner Type: 419

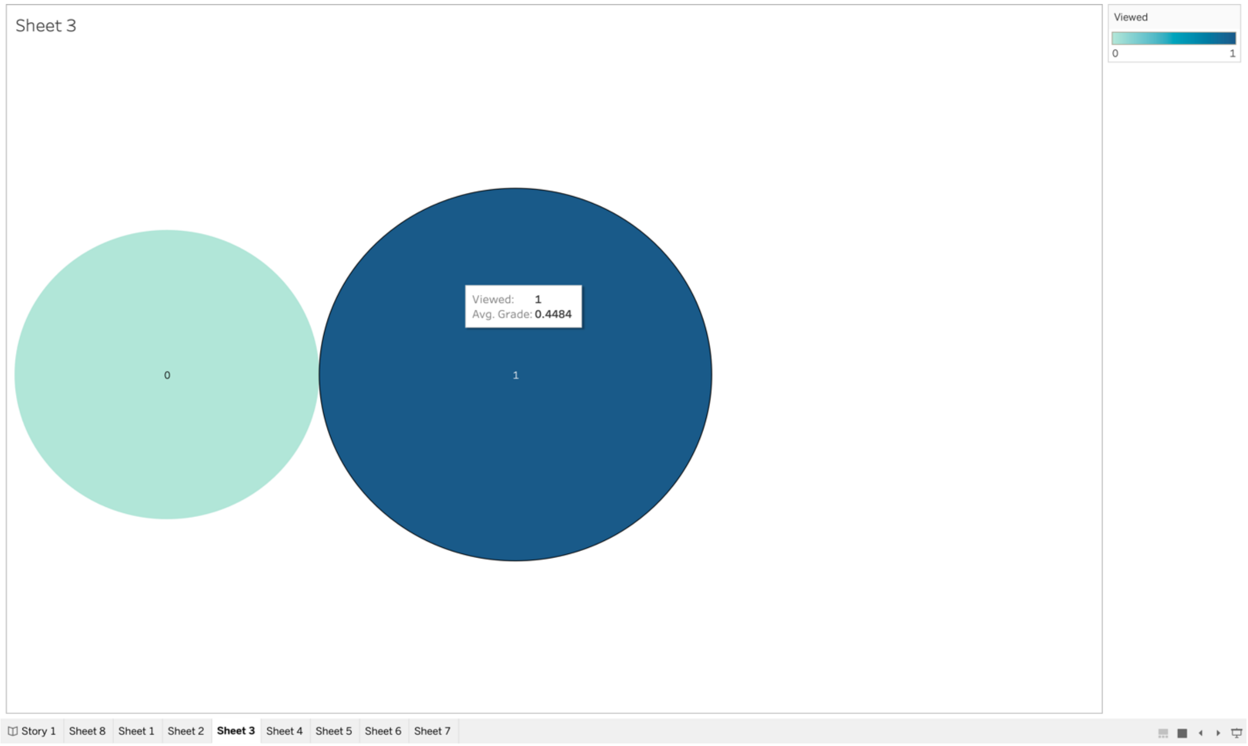
Learner Type
☑ (All)
☑ Active
☑ Observer
☑ Passive

Grade (bin)
■ 0.2
■ 0.4
■ 0.6
■ 0.8
■ 1.0

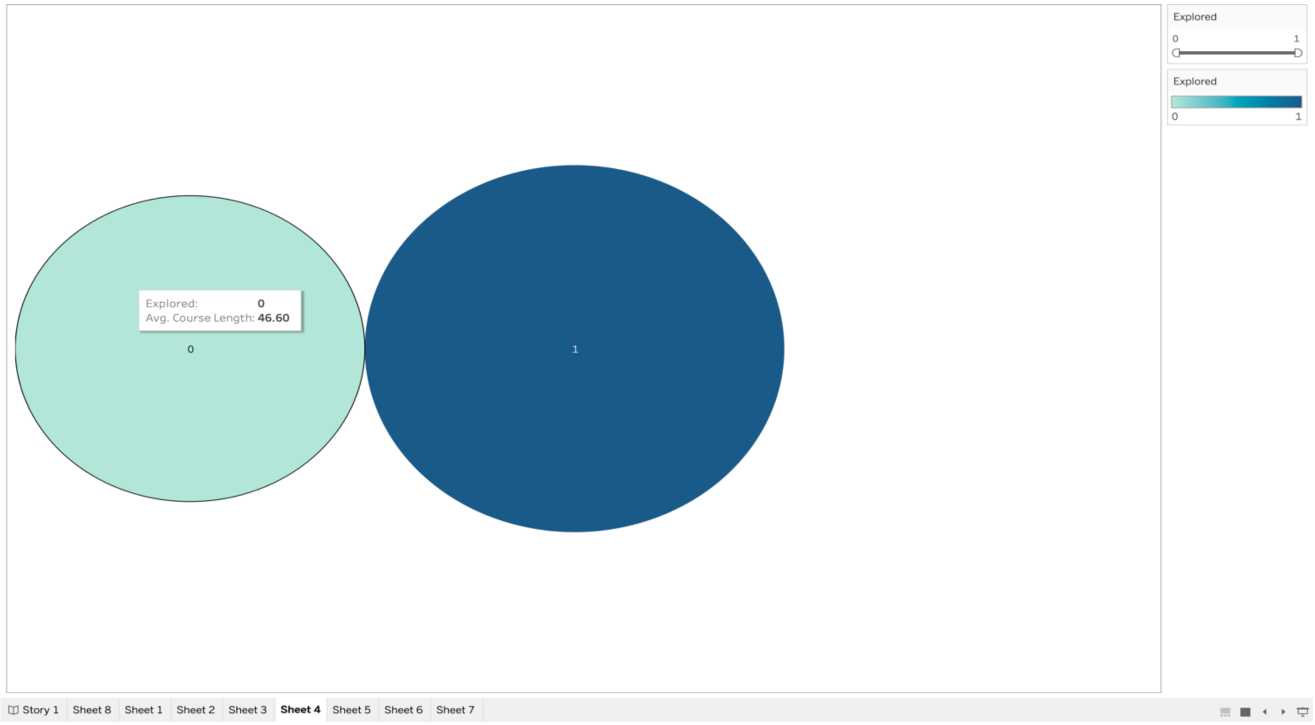


Expected Hours Week: Between 2 and 4 hours
Grade (bin): 0.4
Count of Grade: 135

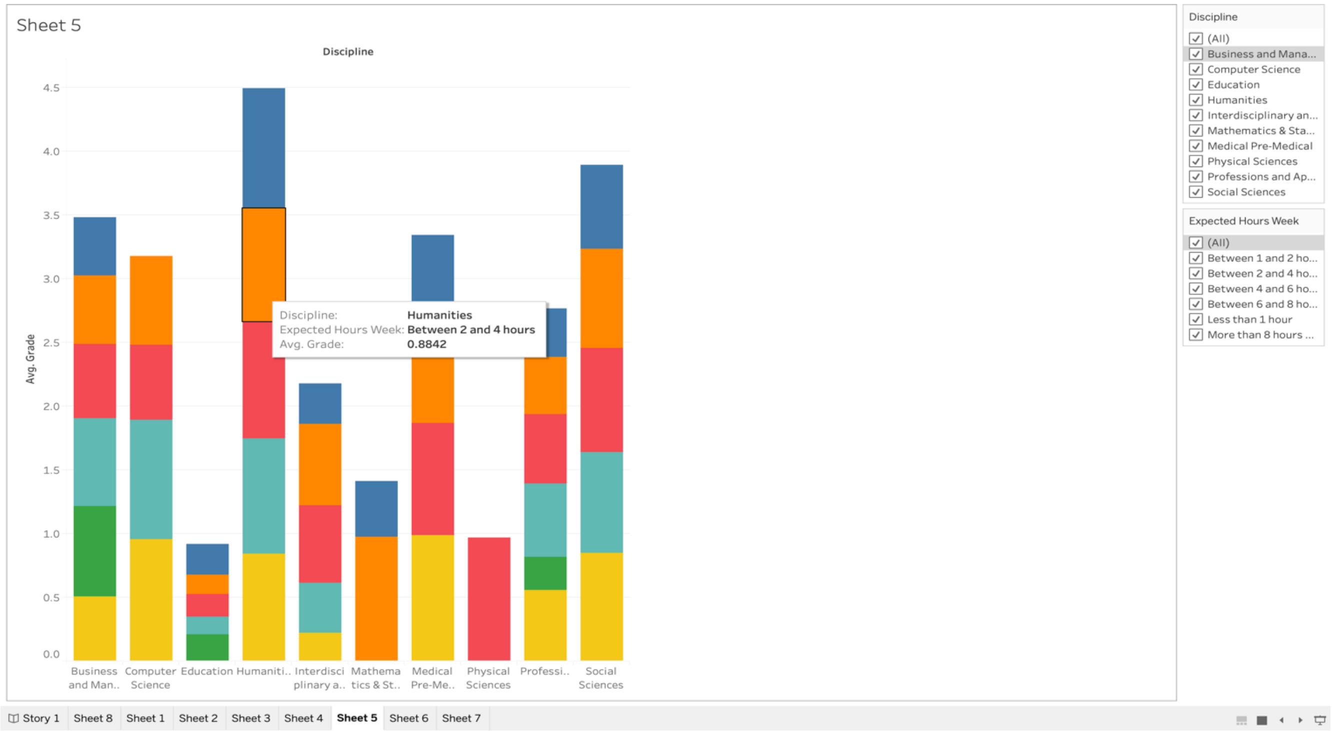
Expected Hours Week
■ Between 1 and 2 hou..
■ Between 2 and 4 hou..
■ Between 4 and 6 hou..
■ Between 6 and 8 hou..
■ Less than 1 hour
■ More than 8 hours p..



Viewed: 1
Avg. Grade: 0.4484



Explored: 0
Avg. Course Length: 46.60

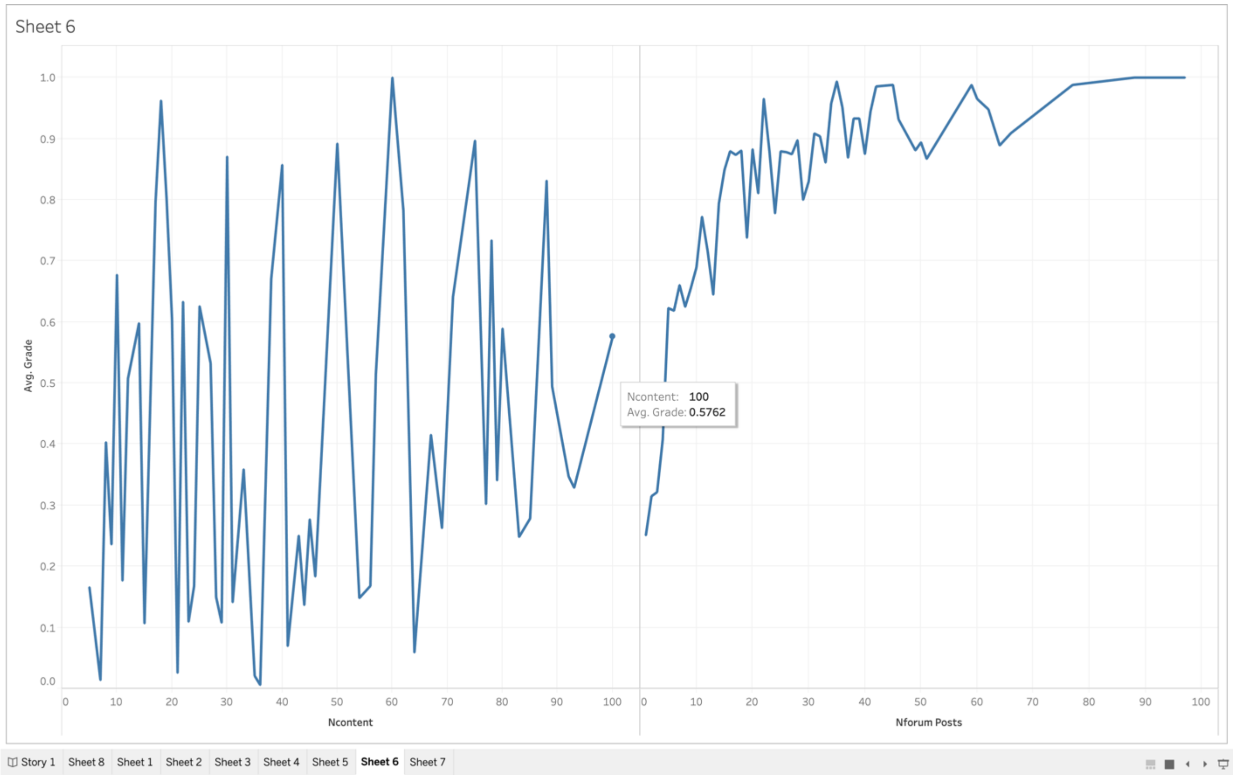


Sheet 5

Discipline: Humanities
Expected Hours Week: Between 2 and 4 hours
Avg. Grade: 0.8842

Discipline
☑ (All)
☑ Business and Mana...
☑ Computer Science
☑ Education
☑ Humanities
☑ Interdisciplinary an...
☑ Mathematics & Sta...
☑ Medical Pre-Medical
☑ Physical Sciences
☑ Professions and Ap...
☑ Social Sciences

Expected Hours Week
☑ (All)
☑ Between 1 and 2 ho...
☑ Between 2 and 4 ho...
☑ Between 4 and 6 ho...
☑ Between 6 and 8 ho...
☑ Less than 1 hour
☑ More than 8 hours ...



Sheet 6

Ncontent: 100
Avg. Grade: 0.5762

5. SUMMARY

In both Accuracy percentage and RMSE value, RMSE value holds more importance. Though the accuracy percentage for the grade of Random Forest is less when compared to Bayesian Ridge and Linear Regression but the RMSE value for the Random Forest method is more when compared to both the Bayesian Ridge and Linear Regression methods.So, by the results we can conclude that Random Forest method is best for predicting the grades.

Based on the Analysis of the dataset we found that Active Learners are more likely to get good grades than observers and passive learners. Another major highlight for the educators to make students interested in course work is to make them interactive and have more content in them. We also found like Disciplines like Computer Science, Medicine and Engineering students must spend more amount in course to get good grades compared to Physical education and Statistical Students. Students must have active participation in discussion forum and discuss course to get high grades.

6. INDIVIDUAL CONTRIBUTION

Member of the Team	**Task done**	**Task Completion Date**
Harshini Palavai	Data cleaning and Pre-processing	March 15, 2020
	Feature Selection method: Generic Univariate Analysis	April 1, 2020
	Data visualization graph	April 3, 2020
	Implemented Random Forest Regression model to predict grades Computed Accuracy and RMSE	April 12, 2020
	Written Report	May 4, 2020
SriLasya Samudrala	Implemented Bayesian-Ridge model to predict grades Computed Accuracy and RMSE	April 12, 2020
	Tableau Prep Builder: Cleaning the data file and producing new output file for analysis of the data	April 6,2020
	Analysis of the Dataset: Using Tableau and creating stories and dashboard sheets.	April 10, 2020
	Written Report	May 4,2020
Narla Raswitha	Feature Selection Method: Correlation using Heatmap	April 1, 2020
	Implemented Linear Regression model to predict grades Computed Accuracy and RMSE	April 12, 2020
	Written Report	May 4, 2020

7. REFERENCES

1) https://www.kaggle.com/residentmario/automated-feature-selection-with-sklearn.

2) https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e.

3) https://scikit-learn.org/stable/.

4) https://www.districtdatalabs.com/how-to-start-your-first-data-science-project.