Implementing different flavours of Classification and Regression Machine Learning Algorithms on different datasets in the US region.

1st Richard Saldanha
Data Mining and Machine Learning 1
National College of Ireland
Dublin, Ireland
x18183034@student.ncirl.ie

***Abstract*—This paper has chosen US as the geographical region for the study. The waterfront view prediction for the different properties is from King County region including Seattle and the results are derived using two machine learning algorithm which are Logistic Regression and Random Forest classifier, the results proved that the percentage of accuracy of prediction is highest for Random Forest Classifier. In case of existence of correlation for the average avocado prices multiple linear regression was applied. SVM and Naïve Bayes Classifier were applied to check for the crowdedness of the gym with SVM giving better prediction results. Confusion Matrix, RSE, ROC are the evaluation parameters used in the project.**

Keywords—Water Front View, Average Haas Avocado Prices, Crowdedness at Gym, Machine Learning, Data mining, Logistic Regression, Random Forest Classifier, Multiple Linear Regression, Support Vector Machine, Naïve Bayes Classifier, R studio.

I. INTRODUCTION

Being highly motivated is uttermost important to achieve your goals in life. In this project the author is highly motivated to apply his budding data analytical skills to carry out different classification and Machine Learning Algorithms on the different datasets acquired from open dataset repository which is Kaggle focusing on the study in the US region[7] and at different area of study like the beautiful properties in Washington DC, Preferred fruits by the "Millennial's[8]" which is Avocado and the other one which is related to health and one of the major requirement of the growing body is fitness acquired by the college/universities students through the source of gym. This motivation has given birth to research questions which this project aims to address and they are What if we could predict that the properties purchased in "King County USA, including Seattle[7]" will have a waterfront view or not, second We would we like to find out if there exist a correlation between the average price of a single avocado with the type of avocado, Total number of avocados with "PLU 4046,4225 and 4770 sold[8]"? and the third one is We would like to find out how many number of people visit the gym during the semester or the question can be framed as is the rate at which the people visiting the gym during the semester is high or low? Addressing these questions would be of a great importance as many individuals prefer having a water front view, enjoy "Avocado Toast [8]" and are fitness enthusiastic persona who prefer going to the gym.

II. RELATED WORK

[1] Rincon-Patino, Juan, et al conveyed a research study about the sales estimation in different cities in the united states in this research worked on the estimation of Haas Avocado using different machine learning algorithms and evaluated the four models like Linear Regression, Multilayer Perceptron, Support Vector Machine for Regression and Multivariate Regression Prediction Model. The study found that there exist a very strong correlation producing the best accuracy with a correlation coefficient of 0.995 and 0.996 and a relative absolute error of around 7.971 and 7.812 which was obtained by applying the model of SVM for regression and Multivariate Regression Model. In another study by researchers [2] Edgar Roa Guerrero and Gustavo Meneses Benavides had proposed a research study on the classification of Haas avocados which is based on image processing technique, they did this by applying a classifying algorithm that would receive fruit images wirelessly from a camera there by transmitting the results is transmitted using IEEE 802.15.4, the correctly matched avocado images for a set of selected avocados was found to be 88 % and the confusion matrix was found to be 82%.Diverting towards the house sales prediction the authors [3]P. A. Viktorovich, P. V. Aleksandrovich, K. I. Leopoldovich and P. I. Vasilevna did a deep research on the study on the predicting of house sales prices using advanced regression technique such as Lasso regression algorithm , Elastic Net regression which is almost similar to elastic net regression , the researchers took a step forward and applied boosting algorithm, they further had used 5-fold cross validation technique.l[4] N. N. Ghosalkar and S. N. Dhage in their study proposed an evaluation of predicting the property prices in Mumbai India using linear regression model with a minimum prediction error obtained at 0.3713. J. Semrl and A. Matei[5] in their research applied a popular machine learning algorithm with an intension to retain their valuable customers of a gym, using two major platforms which are Azure ML and Big ML with the following machine learning algorithms applied such as AzureML, BigML Decision Forests Decision Forest Decision Jungle Boosted Trees Logistic Regression Neural Network which was evaluated using ROC curve method. As heath is an important factor especially for students in their budding years need to have a well-balanced life in order to achieve success as the source available for fitness is through the gym provided by the university researchers at Washington State University Du, Yunshu &

Gebremedhin, Assefaw & Taylor, Matthew. (2018)[6] had done a research to predict the future volumes of students at the fitness center , one of the machine learning model which was used was Random Forest and the performance evaluation methods used were RMSE, MASE and relative RMSE.

III. DATA MINING METHODOLOGY

A. Data Sources

The root of the data sets originates from an open source data repository which is Kaggle which is licensed as open source to use and can be freely downloaded to draw meaningful insights from the data. This project aims to address three datasets emphasized on the US region[7] which is about the property prices in "King County including Seattle[7]", Avocado[8] Fruit Prices and the observation of crowd at the gym.

B. Data Extraction and Loading

The downloaded dataset is in the form of Microsoft csv file format. The csv file downloaded does not have a proper naming convention, so it was renamed appropriately. In this project we will be using R studio environment to apply the different Classification and Regression Machine Learning Algorithms on the three different datasets under observation. The csv files are now loaded i.e. they are read in the R studio environment using **read.csv()** function. Kindly note to set the R studio environment as the current working directory where the csv files are present in order to read the csv files with ease which can be done by navigating to the Files tab ➔ More ➔ set As Working Directory.

C. Data Cleanimg

Once the csv files are read into the environment we then check for any missing values in the dataset using **is.na()** function and produce the sum of the observed null values in the dataset using the function sum(is.na(mydata)).

D. Data Selection

As per the intention to solve the research questions as discussed earlier we would be considering only selected fields under observation from all the three datasets. We drop the unnecessary fields from the data frame using a powerful data mining package which is **dplyr.** We first install the package using the command install.packages("**dplyr**") as it doesn't come preloaded in R Studio environment and next, we have to loaded the downloaded package using **library(dplyr)**.

E. Data Splitting

Once the dataset is free from missing values we split our datasets into two parts one for training our model and the other for testing the accuracy of our model with a ratio of either in the ratio of 7:3 or 8: 2 depending upon the volume of the dataset. This is achieved by installing the package **caTools** which is used for the split function and it is done using the command install.packages("caTools") thereafter we need to import the package by using the command **library(caTools)**. We apply this generic command

split <- sample.split(data, SplitRatio = 0.8)
split
training <- subset(data, split == "TRUE")
testing <- subset(data, split =="FALSE")

Here it will randomly split the main dataset into two datasets for training and testing purpose of the model with 80% True Values assigned to training dataset and 20% False Values assigned to testing dataset. There is another way of splitting the dataset into training and testing dataset which is by using the sample function given as

set.seed(123)
id <- sample(2, nrow(mydata), prob = c(0.8,0.2), replace TRUE)
waterfront_train <- mydata[id==1,]
waterfront_test <- mydata[id==2,]

In this case the training and testing dataset would be randomly split in the ratio of 80:20 where id=1 would be for train dataset and id=2 for test dataset.

F. Data Mining Algorithms

In this project, we would be applying different flavors of machine learning algorithms on the three datasets. Logistic Regression and Random Forest Regression would be applied on the House Sales in King County, USA including Seattle dataset and we would perform a comparative study on them to understand which model gives better results based on the evaluation methods that will be projected in the Evaluation section to predict that the houses sold in King County USA including Seattle will have a waterfront view or not. Multiple Linear Regression would be applied on the Avocado Prices dataset to predict the average price of avocado based on its type and the total number of avocados with PLU 4046,4225 and 4770 sold. The last two classification-based machine learning algorithm which would be applied on the Crowdedness of the gym dataset are SVM which is Support Vector Machine and Naïve Bayes Classifier with an aim to find out whether the gym at "UC Berkeley[9]" is crowded at the beginning of the semester or not. A comparative study would be carried out to determine which would be the best model based on the evaluation parameters which would be discussed in the Evaluation section.

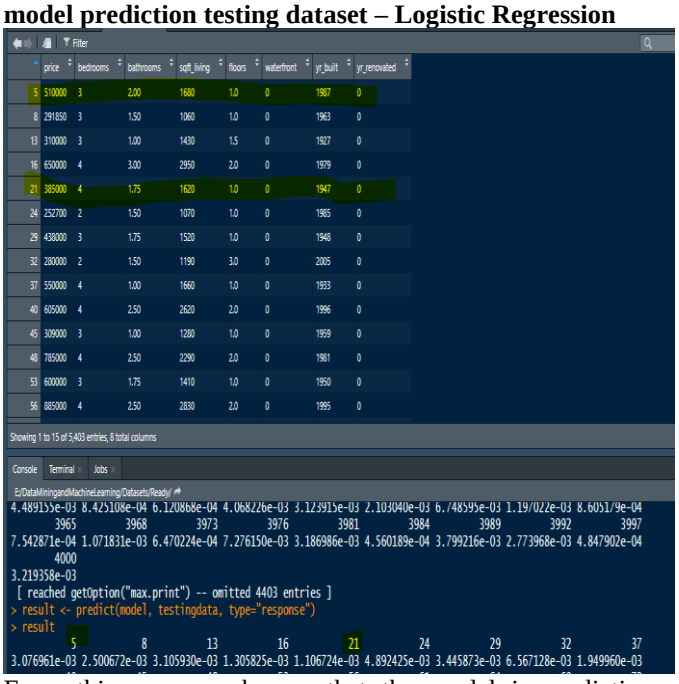
G. Implementation of the Machine Learning Models on the different datasets under observation

1. Logistic Regression

The model is created using glm command in this manner *model<-glm(waterfront~., trainingdata, family = "binomial")* where *waterfront~.* , here waterfront is the dependent variable or the predict variable and ~. Are the independent variables or the predictor variables. The model is applied on the training dataset (*trainingdata)* with a family of binomial means it can take only two values like yes or no, true or false or 0 or 1. The summary of the model applied can be observed using the *summary(model)* method. Observe the figure below **Fig. 1: model summary of Logistic Regression**

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.060e+00  7.119e+00   1.273   0.2031
price        2.997e-06  2.482e-07  12.072  < 2e-16 ***
bedrooms    -9.390e-01  1.455e-01  -6.452 1.10e-10 ***
bathrooms   -1.272e-03  2.011e-01  -0.006   0.9950
sqft_living -4.656e-04  1.955e-04  -2.382   0.0172 *
floors       2.007e-01  2.119e-01   0.947   0.3435
yr_built    -6.526e-03  3.687e-03  -1.770   0.0767 .
yr_renovated 7.148e-04  1.316e-04   5.433 5.55e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1436.08  on 16209  degrees of freedom
Residual deviance:  999.48  on 16202  degrees of freedom
AIC: 1015.5

Number of Fisher Scoring iterations: 9
```

In the above figure we observe the estimate column which specifies the values of the coefficients for the independent variables which are price, bedrooms, bathrooms, sqft_living, floors, yr_built, and yr_renovated. The significant variables are price, bedrooms, sqft_living, yr_built, and yr_renovated with price, bedrooms and yr_renovated being highly significant with level of confidence of 99.9% adding a greater accuracy to the model. From our model it is observed that yr_built and sqft_living is the least significant value. It is observed that bathrooms and floors do not contribute to the model accuracy which can be removed. The next parameter we observe is the Null deviance which is 1436.08 which is the values we get from dataset without inclusion of the independent variables only the intercept is observed. Residual deviance is 999.48 achieved when predictors are taken into consideration for the model. The AIC observed is 1015.5, it should be lower when non-significant predictors are removed from the model. On optimization of the model it was observed that the variable bathrooms can be removed from the model as it had produced a low residual deviance and AIC value when compared with the actual values, On the other hand the variable floors can not be removed from the model as it has some level of significance. We then compare with the Test dataset in order to check if the model is predicting correctly. Observe the below screen shot **Fig. 2 model prediction testing dataset – Logistic Regression**



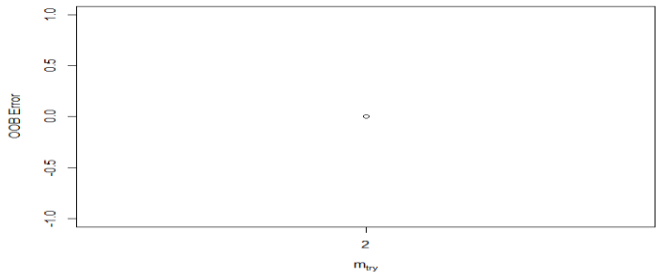
From this we can observe that the model is predicting correctly when we randomly select values to predict means properties with such criteria would not have a waterfront view.

2. Random Forest Classifier

The random forest model is implemented using the random forest method which is done by first installing the packages by using the command *install.packages(randomForest)* and then by importing the package using the package *library(randomForest)* and then we use this command which is *waterfront_forest <- randomForest(waterfront~., data = waterfront_train)* where *randomForest* is the method *waterfront~.* is the dependent variable or the predict variable and ~ . represents the independent variables or the target variables. The summary() function can be used to summarize the model. See below screen shot **Fig.3: Summary-RandomForest**

```
> summary(waterfront_forest)
                Length Class  Mode
call               3   -none- call
type               1   -none- character
predicted      17346   factor numeric
err.rate        1500   -none- numeric
confusion          6   -none- numeric
votes          34692   matrix numeric
oob.times      17346   -none- numeric
classes            2   -none- character
importance         7   -none- numeric
importanceSD       0   -none- NULL
localImportance    0   -none- NULL
proximity          0   -none- NULL
ntree              1   -none- numeric
mtry               1   -none- numeric
forest            14   -none- list
y              17346   factor numeric
test               0   -none- NULL
inbag              0   -none- NULL
terms              3   terms  call
```

We then factorize our categorical variable which is *waterfront* on the processed dataset as well as on the train dataset using *as.factor()* function. In our scenario, there are 7 predictor variables and a target variable, considering the fact that 500 trees are implemented we will be required to make use of the tuneRF function to choose the most optimized random variable the command for it is as given below tuneRF(*waterfront_train,waterfront_train$waterfront,stepFactor = 1.2,improve = 0.01, trace= T, plot = T)*, here step factor indicates the number of iterations to arrive at the number of predicted variable which would be optimized, for every factor increase of 1.2 with an improve of 0.01 to terminate the iterations. See below screenshot **Fig. 4: TuneRF output**

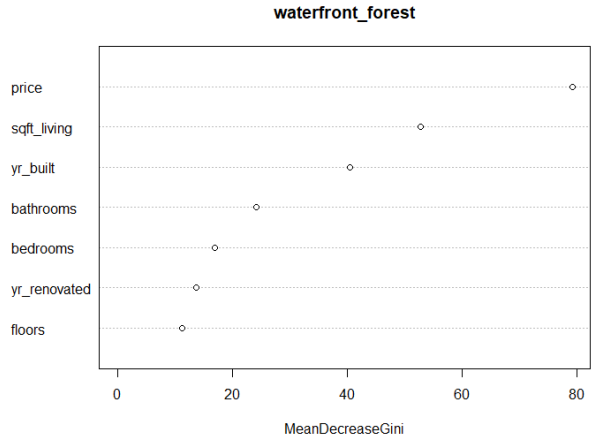


Here we observe that the optimized number of random variables is 2 with respect to OOB which is the prediction error. The significant predicting variables can be found out using the *importance()* function. See below screen shot **Fig.5**

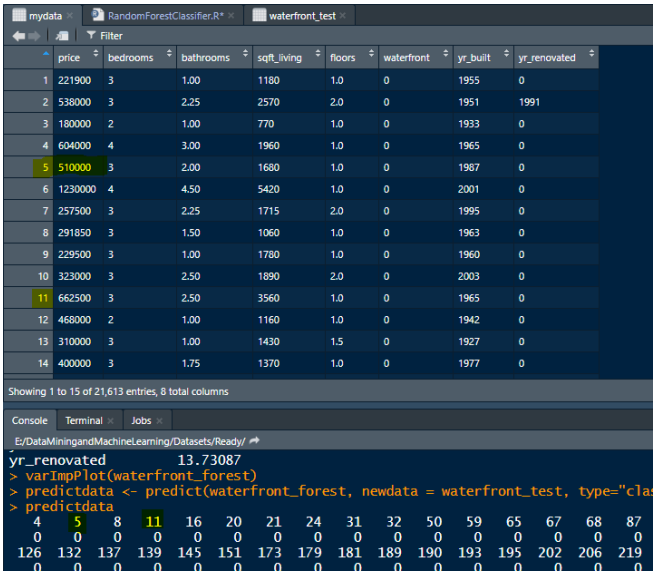
Significance of Predictor Variables

```
> importance(waterfront_forest)
             MeanDecreaseGini
price                79.24149
bedrooms             16.92211
bathrooms            24.15829
sqft_living          52.84512
floors               11.16959
yr_built             40.42402
yr_renovated         13.73087
```

The results shows that price, sqft_living and yr_built being the highest significant variables followed by bathrooms, bedroom, yr_renovated and the least significant independent variable is floors. The visualization of this can be clearly represented in the diagram below. **Fig. 6 – Visualization of Significant level.**



The prediction is done on the test dataset and the observations proved that there were correctly predicted values when cross verified with the test dataset. Observe the below screen shot **Fig. 7: Prediction and test dataset- Random Forest Classifier**



	price	bedrooms	bathrooms	sqft_living	floors	waterfront	yr_built	yr_renovated
1	221900	3	1.00	1180	1.0	0	1955	0
2	538000	3	2.25	2570	2.0	0	1951	1991
3	180000	2	1.00	770	1.0	0	1933	0
4	604000	4	3.00	1960	1.0	0	1965	0
5	510000	3	2.00	1680	1.0	0	1987	0
6	1230000	4	4.50	5420	1.0	0	2001	0
7	257500	3	2.25	1715	2.0	0	1995	0
8	291850	3	1.50	1060	1.0	0	1963	0
9	229500	3	1.00	1780	1.0	0	1960	0
10	323000	3	2.50	1890	2.0	0	2003	0
11	662500	3	2.50	3560	1.0	0	1965	0
12	468000	2	1.00	1160	1.0	0	1942	0
13	310000	3	1.00	1430	1.5	0	1927	0
14	400000	3	1.75	1370	1.0	0	1977	0

3. Multiple Linear Regression

The model is applied on the Average avocado prices dataset is used to find out if the average avocado prices is expressed as a linear combination of total number of avocados with different PLU values and the type of avocado sold. The model is fitted on to the training dataset. The model is fitted on the training dataset using the *lm()* function and it is expressed as *mlregressor = lm(formula = AveragePrice ~ .,*

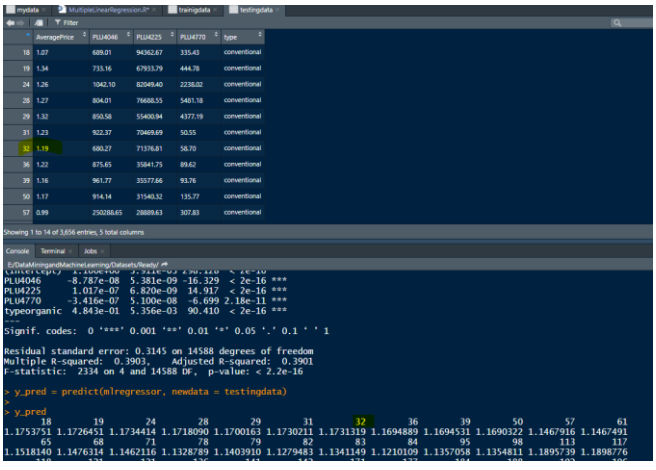
data = trainigdata). Observe the results on applying the model. **Fig.8: Model Summary Multiple Linear Regression**.

```
Call:
lm(formula = AveragePrice ~ ., data = trainigdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.21069 -0.19895 -0.03033  0.18068  1.59884

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.166e+00  3.911e-03 298.128  < 2e-16 ***
PLU4046     -8.787e-08  5.381e-09 -16.329  < 2e-16 ***
PLU4225      1.017e-07  6.820e-09  14.917  < 2e-16 ***
PLU4770     -3.416e-07  5.100e-08  -6.699 2.18e-11 ***
typeorganic  4.843e-01  5.356e-03  90.410  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3145 on 14588 degrees of freedom
Multiple R-squared:  0.3903,    Adjusted R-squared:  0.3901
F-statistic:  2334 on 4 and 14588 DF,  p-value: < 2.2e-16
```

From the above figure we observe the coefficient table for each of the variable, we have from the table is the coefficient in the linear regression equation, standard error, t-value, p-value and the significance level. Valuable information is derived from the P value and the significance level as it tells us about the statistical significance of the independent variable onto the dependent variable. Lower the P value observed higher would be the impact on the dependent variable. We also observe the significant codes and observe that all the predictors are highly significant means that the Average price of the avocado is governed by the factors which are the independent variables under consideration. The prediction is done on the test Results on the testing dataset

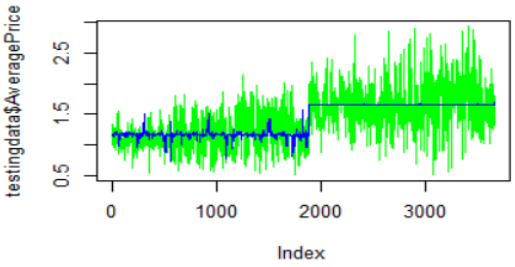
And on observation of record 32 it is found there is no much difference in the predicted and the actual values.

See below screen shot, **Fig. 9 -Prediction with test dataset -Multiple Linear Regression**



We can also plot a graph to show the relationship between the predicted values from the model and the actual values

Fig. 10: Predicted values vs Actual Values



From the above graph there is not much difference in the predicted and actual values of the model.

4. *Naïve Bayes Classifer*

The naïve Bayes classifier is applied on the gym dataset with an intension to find out the crowdedness of the gym at UC Berkey is during the start of the semester or other duration of the semester[8]. The feature of Naïve Bayes machine learning algorithm is such that we can built a model in such a way that it can address either a regression problem or a classification problem but we intend to built it as a classification problem as we are trying to predict the crowdedness at the gym during the beginning of "school semester or otherwise[8]".The factorization of the target variable needs to be prior to applying the naïve Bayes algorithm on the train dataset. The *e1071* package needs to be imported post installation as it contains many functions of machine learning algorithm. The command of execution is *gym_nb<- naiveBayes (is_start_of_semester ~ number_people + day_of_week + is_weekend + is_holiday + temperature + month + hour, data =gymtrain)*
gym_nb. Observe the screen shot of **Fig. 11: Model summary of NAÏVE BAYES Classifier**

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
        0         1
0.92127084 0.07872916

Conditional probabilities:
   number_people
Y       [,1]     [,2]
  0 27.84470 21.77039
  1 43.18126 27.05403

   day_of_week
Y       [,1]     [,2]
  0 3.000970 2.002578
  1 2.921152 1.948226

   is_weekend
Y        [,1]      [,2]
  0 0.2879485 0.4528126
  1 0.2580739 0.4376385

   is_holiday
Y          [,1]       [,2]
  0 0.002660434 0.05151138
  1 0.000000000 0.00000000

   temperature
Y       [,1]     [,2]
  0 58.39937 6.239435
  1 60.60125 6.866916

   month
Y       [,1]     [,2]
  0 7.573137 3.385620
  1 5.834449 3.611471

   hour
Y       [,1]     [,2]
  0 12.23320 6.708122
  1 12.46552 6.654608
```

From the above model we observe the A-priori and Conditional Probabilities values. It is observed that 92 percent of the probability is when the members of the gym do not visit at the start of the semester while roughly around 7 percent of the student population visit the gym at the beginning of the semester as depicted from the A-priori probabilities. The model also produces conditional probabilities for all the predictor variables. On an average the number of students visiting the gym during the semester i.e. at beginning of semester is 43.1818126 percent with prevalence of variability in the data of around 27 percent. Further the prediction with the test dataset and model valuation would be discussed in the Evaluation section.

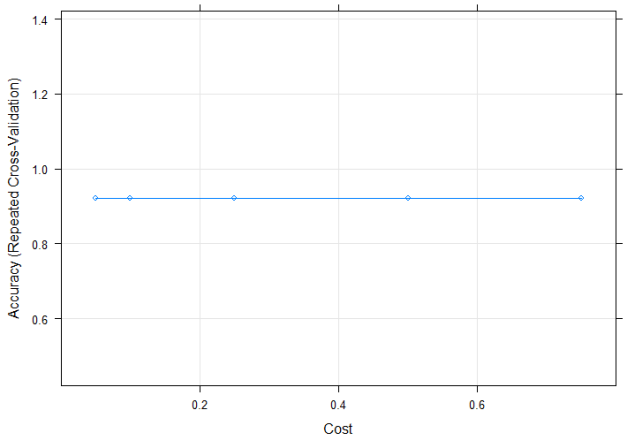
5. *Support Vector Machine*

Like the Naïve Bayes Classifier algorithm, we will also apply Support Vector Machine Algorithm on the campus gym dataset of UC Berkey to find out the when does the gym experiences crowdedness is it at the beginning of the semester or mid-level time of the students' academic year. Prior to applying the model, the dataset is cleaned which means that we check for any null values in them, secondly we need to divide the dataset into training and testing dataset where model will be applied on our train dataset and prediction will be done on our testing dataset. Factorization of the target variable is necessary which can be done using the *factor()* function. The *trainControl()* method made available by the *caret* package for training the dataset. The command is given *as trctrl<- trainControl(method= "repeatedcv", number= 10, repeats = 3)*, here repeatedcv is the cross validation method which will be implemented , followed by number=10 means number of resampling iterations, we set *repeats = 3* indicating repeats for our cross sampling method. The value obtained will be stored in the *trctrl* method which would further be passed to our train method. The model command would look like trctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 3) *svm_Linear <- train(is_start_of_semester~.,data = trainingdata,method = "svmLinear",trControl=trctrl,preProcess=c("center","scale "),tuneLength = 10)* In this, the *trctrl* variable is passed to the trControl method with preprocessing stage where we will be centering and scaling our data, the algorithm is tuned to 10 iterations stored in our *tuneLength* method. Observe the below screenshot of SVM **Fig. 12 SVM Model**

```
Support Vector Machines with Linear Kernel

43528 samples
    7 predictor
    2 classes: '0', '1'

Pre-processing: centered (7), scaled (7)
Resampling: Cross-Validated (10 fold, repeated 3 times)
Summary of sample sizes: 39174, 39175, 39175, 39176, 39175, 39176, ...
Resampling results:

  Accuracy   Kappa
  0.9211772  0

Tuning parameter 'C' was held constant at a value of 1
```

From, the above screen shot we can observe that the train dataset was preprocessed equally, The Resampling results provided an accuracy of 92 % which is a good indicator for our model. The value of C is 1 indicating a linearly separable model. Further, we do the prediction on our testing dataset using the *predict()* function using the *caret* package. The command is as *test_pred <- predict(svm_Linear, newdata = testingdata)* where the svm model is passed to the *predict* function and it is applied on the testing dataset. In Linear

SVM the performance of the model can be enhanced by using *grid()* method to check for different cost value of the model, the command is given as *grid <- expand.grid(C = c(0.05, 0.1, 0.25, 0.5, 0.75)). Observe the figure below*



From the **Fig. 13: Plot for different values of Cost**. We can observe that for different values of cost which are 0.05, 0.10, 0.25, 0.50 and 0.75 the accuracy obtained is same as before which is around 92 %.

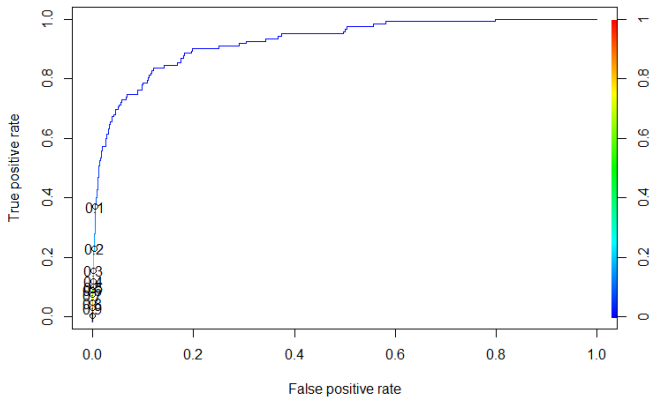
IV. EVALUATION OF MACHINE LEARNING ALGORITHMS

A. Logistic Regression

To predict whether the properties purchased in King County, including Seattle will have a waterfront view or not, the evaluation parameter used is Confusion Matrix as it is efficient way to check for the efficiency of the model. Observe from below screen shot **Fig .14: confusion matrix for Logistic Regression**

```
> confusionmatrix
            Predictedvalue
Actualvalue FALSE TRUE
          0  5357    5
          1    33    8
```

From the above screen shot we observe the actual values derived from our dataset under observation and the predicted values are from the model, here in Actual value 0 means there is no waterfront view and 1 means there is a waterfront view and here it is predicted that when there is no waterfront view for that property the actual predicted value is 5357 whereas it was observed that 5 times the model predicted that there is a waterfront view which is an error occurred around 5 times. On the other hand, when the property had a waterfront view the model predicted that it 33 times that there was no waterfront view while in the case when there is a waterfront view with the property the model predicted it be only 8 times. The accuracy of the model is obtained by adding correct instances to the total number of instances in the model which was found to be 99 % accurate. At varying threshold value, we can check if there is any significant change in the Accuracy of the model by making use of another evaluation method which is Receiver Operating Characteristic Curve (ROC) method. Observe the **Fig. 15 ROC Curve-Logistic Regression**



From the above graph we observe that on our x axis we have our False positive which should be at the lowest rate or they should have a minimum value rate and on the Y axis we have our True positive rate should be high. Here we observe that at 0.1 threshold the true positive rate is high with the lowest false positive rate. Surprisingly at many thresholds values the false positive rate is lowest with high true positive rate observed at threshold 0.1. The intension is to minimize the false predicted value for the condition where the property purchased has a waterfront view as it would be a better for the intended buyers. We now observe the Confusion Matrix and compare it with the New Confusion Matrix at threshold 0.1 which is shown in **Fig. 16: New confusion matrix-Logistic Regression**

```
> confusionmatrix
            Predictedvalue
Actualvalue FALSE TRUE
          0  5357    5
          1    33    8
> newconfusionmatrix
            Predictedvalue
Actualvalue FALSE TRUE
          0  5322   40
          1    23   18
```

From the above figure we can observe that by considering the correct threshold value there is a fall in the false predicted value with an increase in the true positive value as compared to the values obtained in the confusion matrix at threshold 0.5.On the other hand in case where there is no waterfront view the actual value predicted by the model has decreased there to 35 times and false predicted has also increased to 35 times which seems fine. The new Accuracy obtained is around 98% which is pretty good for the model.

B. Random Forest Classifier

Like the Logistic Regression we would be applying the model on the same dataset which is to predict that whether there would be a waterfront view associated with the property purchased in king county including Seattle in USA. The evaluation matrix used would be confusion matrix. See below screenshot **Fig. 17 confusion matrix for Random Forest**

```
> confusionmatrix
Confusion Matrix and Statistics


predictdata    0    1
         0 4233   33
         1    0    1

               Accuracy : 0.9923
                 95% CI : (0.9892, 0.9947)
    No Information Rate : 0.992
    P-Value [Acc > NIR] : 0.4769

                  Kappa : 0.0567

 Mcnemar's Test P-Value : 2.54e-08

            Sensitivity : 1.00000
            Specificity : 0.02941
         Pos Pred Value : 0.99226
         Neg Pred Value : 1.00000
             Prevalence : 0.99203
         Detection Rate : 0.99203
   Detection Prevalence : 0.99977
      Balanced Accuracy : 0.51471

       'Positive' Class : 0
```

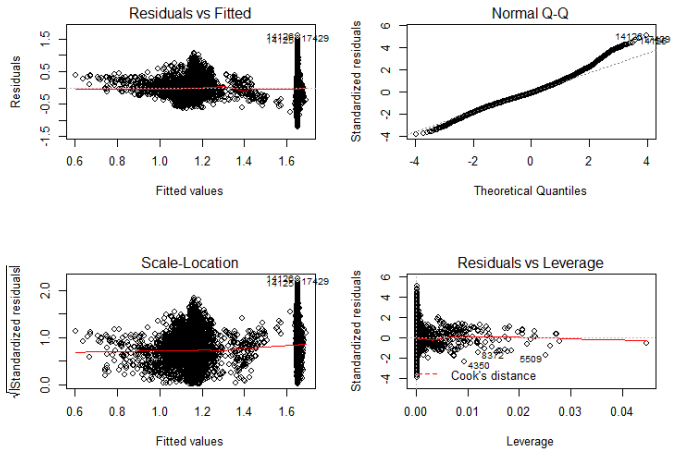
From here we see that the accuracy of the model is quite high around 99 % accurate. From the above confusion matrix, we observe that the actual value and the predicted value for the model when there is no waterfront view is correctly predicted and when there is a waterfront view for the property the actual value predicted is also correct with a very good accuracy of the model at 99 %

C. Multiple Linear Regression

The evaluation is done using multiple R-Squared value, Residual Standard Error. See below screenshot **Fig. 18: Evaluation parameters**

```
Residual standard error: 0.3145 on 14588 degrees of freedom
Multiple R-squared:  0.3903,    Adjusted R-squared:  0.3901
F-statistic:  2334 on 4 and 14588 DF,  p-value: < 2.2e-16
```

From the above screen shot we observe that 39 % of variation in the Avocado Average Prices can be explained by our model. The residual standard error tells us how far the observed average avocado prices are from the predicted fitted average prices. The evaluation of the model can be done by observing the diagnostic plot from below screen shot
Fig. 19: Diagnostic Plot



From the graph we can observe the Residual plot, the linearity assumption is met and their red line is fairly flat, Normal (Q-Q) plot is the ordered observed standardized residuals. There remaining two plots helps us to identify non-linearity, non-constant, variance. Another means of evaluating the model is my observing the variance inflation factor (VIF) as

It tells us about the existence of multicollinearity in the model

```
> vif(mymodel)
  PLU4046    PLU4225    PLU4770       type
7.412642 10.740140  4.835183  1.058280
```

From the above screen shot **Fig.20 VIF values-Multiple Linear Regression** we observe that PLU4046,4225 are highly correlated followed by PLU 4770 moderately and type being the least correlated.

D. Naïve Bayes Classifier

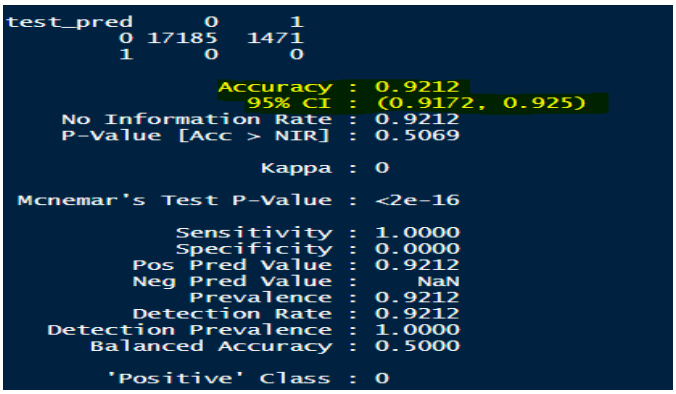
The evaluation of the model is done using the confusion matrix and observation of Kappa Statistics. See below screen shot **Fig. 21 Evaluation- Naïve Bayes**

```
gympredict     0     1
         0  1849    25
         1 15214  1440

               Accuracy : 0.1775
                 95% CI : (0.172, 0.1831)
    No Information Rate : 0.9209
    P-Value [Acc > NIR] : 1

                  Kappa : 0.0159

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.10836
            Specificity : 0.98294
         Pos Pred Value : 0.98666
         Neg Pred Value : 0.08647
             Prevalence : 0.92093
         Detection Rate : 0.09979
   Detection Prevalence : 0.10114
      Balanced Accuracy : 0.54565

       'Positive' Class : 0
```

From the above figure it is evident that the model accuracy is quite low around 17 percent and interpreting the confusion matrix we can observe that the actual value and the predicted value of the students visiting the gym at the mid term of the semester is 1849 which seems to be correct and the falsely predicted is at 25 percent. On observation of visit at beginning of semester it is found that there is a major difference in the actual value and the predicted value. The Kappa statistics observed is also quite low around 0.01 which is the value obtained.

E. Support Vector Machine Algorithm

The model is evaluated using the Confusion Matrix applied after the prediction done on the testing dataset. Observe the below screenshot to check for the accuracy obtained by the model. **Fig. 22: Confusion Matrix of SVM**



At 95% Confidence Interval the scale of (0.9172,0.925) the accuracy obtained by the model is 92 percent which is very good accuracy rate for our model. Further even when the cost values where changed there were no changes in the accuracy value for the model.

V. CONCLUSION AND FUTURE WORK

Dataset	Model	Evaluation	Percentage of Accuracy
"House sales in King county, USA[7]"	Logistic Regression	Confusion Matrix	98%
"House sales in King county, USA[7]"	Random Forest Classifier	Confusion Matrix	99%
"Avocado Prices[8]"	Multiple Linear Regression	Residual Standard Error	39%
"Crowdedness at the campus gym[9]"	Naïve Bayes Classifier	Confusion Matrix	17%
"Crowdedness at the campus gym[9]"	Support Vector Machine	Confusion Matrix	92%

The above table illustrates, the application of 5 machine learning algorithms namely Logistic Regression, Random Forest Classifier, Multiple Linear Regression, Naïve Bayes Classifier and Support Vector Machine across three different datasets focused on the USA region under observation. On a comparative study of Logistic Regression and Random Forest to predict whether the properties purchased would have a waterfront view or not it is evident that both the models applied are good for prediction and with a 1% difference in accuracy level random forest classifier has a better accuracy, thus from the random forest we can say when there is a waterfront view they property prices are quite high the owners would have a waterfront view for lower property prices there are no water front view as evident from the confusion matrix **Fig. 16: New confusion matrix-Logistic Regression.** In the study of Average prices of HAAS avocados where we were interested to find a correlation between the average prices of different PLU sold along with the type whether they are conventional or organic in nature we applied Multiple linear regression model and observed that due to the existence of multicollinearity among the predictor variables see **Fig.20 VIF values-Multiple Linear Regression,** the model is moderately decent to give good results. In case of the dataset where we were trying to predict when the gym is crowded is it when the students just arrive at UC Berkey they visit the gym at the beginning of the semester or is it when they go during the semester or other times, to find out we had applied two Machine Learning Classifier based algorithm which are Naïve Bayes and Support Vector Machine and we observed that SVM is better model for prediction compared to Naïve Bayes Classifier, producing an accuracy of prediction of around 92 percent, thus from this we can say from **Fig. 22: Confusion Matrix of SVM** we can say that university students at UC Berkeley prefer going to gym during their course work of the academic calendar but not during the beginning of the semester. The Future work that I intend is trying to apply different machine learning algorithms to the three different data sets and draw meaningful insights from them, the model where the accuracy obtained was low in case of Multiple Linear Regression and Naïve Bayes, I would try to improve the model accuracy or try another algorithm which would give better accuracy results

REFERENCES

[1] Rincon-Patino, Juan, et al. "Estimating Avocado Sales Using Machine Learning Algorithms and Weather Data." *Sustainability*, vol. 10, no. 10, 2018, p. 3498., doi:10.3390/su10103498.

[2] Edgar Roa Guerrero and Gustavo Meneses Benavides, "Automated system for classifying Hass avocados based on image processing techniques," *2014 IEEE Colombian Conference on Communications and Computing (COLCOM)*, Bogota, 2014, pp. 1-6. doi: 10.1109/ColComCon.2014.6860414

[3] P. A. Viktorovich, P. V. Aleksandrovich, K. I. Leopoldovich and P. I. Vasilevna, "Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning," *2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC)*, Vladivostok, 2018, pp. 1-5. doi: 10.1109/RPC.2018.8482191

[4] N. N. Ghosalkar and S. N. Dhage, "Real Estate Value Prediction Using Linear Regression," *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*, Pune, India, 2018, pp. 1-5. doi: 10.1109/ICCUBEA.2018.8697639

[5] J. Semrl and A. Matei, "Churn prediction model for effective gym customer retention," *2017 International Conference on Behavioral, Economic, Socio-cultural Computing (BESC)*, Krakow, 2017, pp. 1-3. doi: 10.1109/BESC.2017.8256385

[6] Du, Yunshu & Gebremedhin, Assefaw & Taylor, Matthew. (2018). Analysis of University Fitness Center Data Uncovers Interesting Patterns, Enables Prediction. IEEE Transactions on Knowledge and Data Engineering. 10.1109/TKDE.2018.2863705.

[7] "House Sales in King County, USA", *Kaggle.com*, 2019. [Online]. Available: https://www.kaggle.com/harlfoxem/housesalespredictio n. [Accessed: 15- Dec- 2019].

[8] "Avocado Prices", *Kaggle.com*, 2019. [Online]. Available: https://www.kaggle.com/neuromusic/avocado-prices. [Accessed: 15- Dec- 2019].

[9] "Crowdedness at the Campus Gym", *Kaggle.com*, 2019. [Online]. Available: https://www.kaggle.com/nsrose7224/crowdedness-at-the-campus-gym. [Accessed: 15- Dec- 2019].