An approach with Data Mining and Machine Learning Methodologies and Evaluation Methods

Gudivada Manikanta Dinesh (x18191851)
School of Computing
National College of Ireland
Dublin, Ireland
x18191851@student.ncirl.ie

***Abstract*—Data Mining is one of the processes in acquiring the appropriate patters from the huge datasets using the methods of machine learning and statistics. Three main objectives are discussing in this paper. The First dataset describes about diabetes. Coming to the dataset we are going to predict the taking of diabetes medications by the people based on previous procedures, tests and Inpatient or Outpatient history using k-Nearest Neighbor and Logistic regression models. The Second objective is predicting room type from the dataset Airbnb's New York houses which are taken from the source Kaggle. This dataset gives information about the hosts, areas, availability, reviews, price and neighborhood counting. For this dataset, we can perform Decision tree and Random Forest classification models where we must classify the room type whether to rent a private room or a shared room. These are the superlative models in Machine Leaning which proffers the satisfactory results. The third dataset demonstrates the Critical Temperature in a superconductor based on the atomic mass, atomic radius, entropy value and Fusion heat. For this dataset, we can perform Multiple regression for prognosticating temperature variables. This paper furnishes statistics of Data Mining Machine learning algorithms that are utilized for predicting the output dependent variables in all three different datasets. This whole paper is entirely based on the scientific study of implementing statistical algorithms using R programming.**

I. Introduction

This paper is organized by performing Data Mining and Machine Learning algorithms using R programming. R programming is used for predicting the patterns using statistical analysis on datasets. I took three datasets are related to three different sectors whereas, The First dataset is based on the diabetics wherein US high number of cases were registered in adults facing diabetes problems. I took a dataset related to the medications which need to be given for the people of the US in 130-US hospitals where the suitable algorithms that can be performed are the KNN classification and Logistic Regression models. In the same way, the second dataset is related to the Airbnb New York houses. In this dataset, we are having many cases which can be predicted whereas, the hosts and guests can be predicted. Based on the price, location and availability we can predict the people's interest in choosing a private/shared/Entire home. In this way, I have used the Decision Tree and Random Forest for predicting the room type in which many of the people were fascinated. The final dataset is related to the critical temperature in superconductivity where the critical temperature is a quantitative value which is depending on atomic mass, entropy valence, fusion heat and thermal conductivity. For this, the good algorithm suitable to use is Multiple Regression where this is used for predicting the quantity values based on the multiple independent variables. These five algorithms were performed on the three datasets for predicting their independent values by giving training and predicting testing data.

II. Related Work

From the paper Analysis of Airbnb in New York city the author examines the guests, hosts and room availability in the New York city. Airbnb is a one of the best websites for house renting, lodging and tourism. From the year 2008, the tenants were using Airbnb to revamp all the facilities like travelling and housing for fabricating the superior experience of the world. There are many tenants who are advertising their houses for renting [1]. In the same way in New York city there are many guests and hosts who want to rent their houses for renting and hosts for purchasing [2]. Furthermore, in recent years economy growth of Airbnb has increases enormously where the whole market of Airbnb has gradually swelled from low to very high in experiencing the astounding results. The main objective of choosing related to Airbnb dataset is for predicting astonishing results using the locations, prices and reviews. Etc. and by collecting this data we even can give estimate result of busiest hosts in the world. In addition to this we also can come to know about hosts and areas of the renting houses. This paper discusses key issues of the Airbnb whereas the believing and trusting the online reviews. There are also many raising issues in all over the countries, but the major impact of revenue is generated only in the New York city of US [3][4].

By considering other paper they also mentioned that the Airbnb have autonomously increased their business in New York city based on the revenue generated in last few years and also the short term rentals are being increased the problems also been increased in many ways the three main problems that discussed in this paper are violating the country laws of New York, supply of houses and the impact on effective prices in the New York[5]. These problems can be solved by collecting the data from different sources and creating a database. After the creation of the database we need to perform the correlation matrix, regression analysis and finally we need to find the key results of the Airbnb [6].

In addition to this the recently performed studies were informing that the trust on Airbnb is reduced by consumers. To improve the growth of the business we need to bring back the trust of the customers in all possible ways. So, based on all these studies relating my work I used Decision Tree and Random Forest for predicting the values of US Airbnb where I got accuracy of 82 percent in decision Tree and 76 percent in the Random forest [7].

In the same way one more paper discussed the evolution of Airbnb in Australia using logistic regression we can also develop the consumption patterns across the world. So finally based on the data I extracted from the Kaggle and performing the regression and classification models to evaluate the type of room in which the people are much interested in based on the price, neighborhood, host listings count and availability of the house [8].

The author examines that the diabetics the occurs due to the increase of recent changes in the diagnosis of diabetics and these also changes the enhance case detection. Main objective of choosing dataset related to diabetics is this is an out going fast spreaded diseases so in my dataset I am predicting the frequency of admission of a person for the treatment based on the different types of tests performed. Basing this objective I choose this paper where the data for last 5 years and examined the trends in the adults of US whose age is between 18 to 79 using logistic regression. By observing the results in the year 1997 to 2003 there is hike in diabetics of 41 percent. The diabetics has increased both in men and women [9]. There also adults who are facing diabetics problem are 100 million till now. This more significance of diabetics is not same in all the areas and also it differs from the non-Hispanic whites and the non-Hispanic blacks. The symptoms of the undiagnosed diabetics is they will lead to serious illness with chronic kidney disease[10]. So for this problem US government came up with a solution of calling through landline or cell to all the individuals of the country and knowing whether they are facing any severe issues with diabetic. After the survey they combined all the results of the years 2011, 2013, 2015 and 2017 for obtaining the sample size for performing analysis. In the final result of screening process, they came to know the results of past three years whether a person is having the high blood sugar or diabetics [11][12].

In different paper the author is explaining about the organic pollutants and gestational diabetics in US women. Firstly, the insulin is also one of the factors that effecting the organic pollutants across the US. In this paper they also discussed about how diabetics occur in humans and it also occurs more in the women due to the destruction of cells in the pancreas that causes the insulin. There are also two types of diabetics which are type 1 and type 2. Type 1 occurs because of fighting infection in the immune system which is caused by the low blood and more sugar in the body [13].

According the paper US Monogenic organic registry the general symptoms of the diabetics are loss of weight, huge hungry and thirsty, improper vision having more fatigue are the major factors that effecting the diabetics. In US there are 217 families with the Monogenic diabetics in the last one decade. In the same the way the variations in the blood glucose levels will also increase the state of diabetics. So for decreasing this diabetics most important treatment is we need to take a good body diet and decrease the weight and need to perform exercises daily. The people with Type 2 diabetics should consume the medication which controls the blood sugar and so that they can easily cure the diabetics. There are other more medications which may suggest by doctors are using pills and insulin together. All these are some of the medications to cure the diabetics [14].

By taking reference as self-management in the 2nd diabetics we can see that the relationship is more important in utilizing the Type 2 medications in different ethnic groups in the US. Comparatively the Americans in China has very low foot care and healthy eating habits. The people who are taking care of people should use approaches which are patient centered which means there should be an analysis on the patient in a way of how he is responding to the treatment [15].

The final objective is based on the super conductivity where I am calculating the critical temperature based on the factors like atomic mass, thermal conductivity, valence, entropy. Etc. Basing the paper critical temperature of a super conductor, the author examines that the prediction which I am doing for the critical temperature is for predicting the super conductors which is not including the prediction of materials of super conductors. In this paper they also discussed about the super conductivity which is a bundle of physical properties where we can observe in many materials. Super conductivity can exhibit zero resistance flow in many materials for the effective flow of current and neglects the magnetic fields from the transition temperature. Irradiation in the particles have a high range of temperature in superconductors for creating the flux pinning centers where it can enhance the super conductors. In the there will cooper pairs which are very small, and the energy gap is carried by the surface of the conductivity. If the gap is increased, then there will be more implications occurred in the free electrons. By the analysis of whole paper, it is given whole energy is irradiated and shows an effect on the super conducting materials [16].

In other the author discussed about the mechanism of flow of conductivity in the super conductors which is done with the strength of the aerogel and xonotlite the atomic mass is based on the flow of electrons between the conductors. Super conductivity is also the having the metallic an nonmetallic metals and also is it noted that super conductors will not be found in any of the ceramic metals. And in the same way the unique ness is also having different in all the metals of the physical phenomena and all other phenomena is having a true relation with the periodic table where such as, atomic number and Ionization potential will also be there in the normal conductivity [17]. We also know that there is relation between the super conductors and the normal conductors. From the super conductivity there we can say that we also have the magnetic flux inside the electrons which are revolving in the circles where the negative and positive poles are opposite to each other where there is also a connection of where the atoms will be revolving around the electrons and provide super conductors [18].

The other author is describing he conductivity in different ways such as the entropy and the valence is not having equal and informing about the BCS theory of the super conductivity where this is known as quantum of theory. Finally to calculate the critical temperature all the external factors were included in the scenario of which each and every one of the factor has their own importance like mean density, fusion heat, thermal conductivity, atomic radius all these will be major effecting factors for identifying the critical temperature. There are also having many ways for calculating the critical temperature which is Meissner Effect. In this effect the voltage bias is been applied using the electron temperature of the detector which will give the non-zero resistance based on this the critical temperature varies. Based on these all the factors the super conductivity and the electricity will be passing through all the cables connected and the insulators will also play a crucial role in evaluating a temperature of the detector [19][20].

So, based on these scenarios the relevant data sources were taken and performed some of the machine learning algorithms.

III. Data Mining Methodology

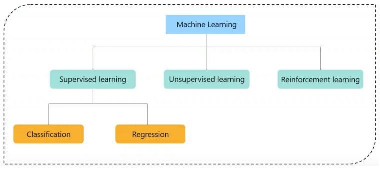


Fig. 1. Machine Leaning Techniques

A. K- Nearest Neighbors

- There are two types of learning which are classified as Supervised learning and Unsupervised learning. The KNN algorithm come under the supervised learning. KNN algorithm is one of the best and simple algorithms which is used to perform the both classification and regression models. The classification and regression models both come under supervised learning. Where the classification is used for predicting categorizing values and the regression, we will be used for predicting the quantity[Fig 1].

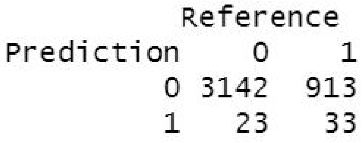
Data cleaning: The first for performing the model is to clean the data. I have cleaned the data using R programming and then we need to transform the original data to other data frame. After transformation of data the KNN model is to be performed.

Data Selection: For performing the KNN model we need to choose a dataset. I choose the dataset related to the diabetes in 130-US hospitals. In this dataset we are going to predict the Diabetes Med based on the previous history of the patient, procedures or tests performed on the person and based on inpatient or outpatient and previous medications all these factors will help us to predict the medications to be used.

Data Transformation: For performing the KNN algorithm we need to train the data and give for testing. In this process we should divide the dataset based on the model priority and then we should predict the values based on the training values we have to predict the test values.

Data Mining: So here, I am performing the KNN Classification model based on the dependent variable as Diabetes Med which is a binary value and all other variables as independent variables. KNN is mainly known as “Recommended system”. What exactly the recommended system is if for instance a person is looking for some item in amazon based on his search, he will be get some recommendations for purchasing of item. This happens because of KNN algorithm. Then calculate the Euclidean distance and Manhattan distance and divide them into two groups. Based on the K value the dependent variable will compare with the Euclidian or Manhattan distance values and check for the majority K value and shift to that group. So, based on this we will find each value of the dataset. Here in my dataset diabetes Med is the dependent variable where whenever a new patient enters hospital, we should give medications for that we need to check all the tests performed previously and medications using previously. In this dataset the values of lab procedures, emergency, medications, outpatient and diagnoses will calculate their Euclidean distance and Manhattan distance and then I have given different values of k = 10,20,30,200 based on this the prediction of values also varies like whether we need to give the medications to that person or not necessary [21].

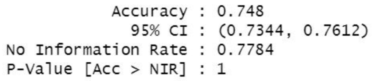
Knowledge Representation:



```
          Reference
Prediction    0    1
         0 3142  913
         1   23   33
```

Fig. 2. Confusion Matrix

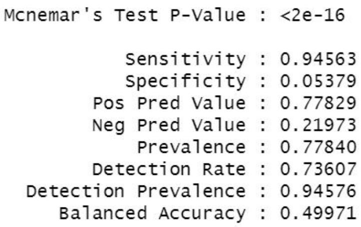
- Here from the result the correctly predicted values are 3142 and 913 in correctly predicted values [Fig 2].



```
             Accuracy : 0.748
               95% CI : (0.7344, 0.7612)
No Information Rate : 0.7784
P-Value [Acc > NIR] : 1
```

Fig. 3. Accuracy of the KNN

•The above figure is confusion matrix performed for knowing the accuracy. The accuracy value is 74 percent[Fig 3].



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

                Sensitivity : 0.94563
                Specificity : 0.05379
             Pos Pred Value : 0.77829
             Neg Pred Value : 0.21973
                 Prevalence : 0.77840
             Detection Rate : 0.73607
       Detection Prevalence : 0.94576
          Balanced Accuracy : 0.49971
```

Fig. 4. Specificity of the Model

• In this Fig 3 we can check the values which were giving the sensitivity, specificity and Pos Pred value, Neg Pred value, Prevalance, Detection Rate, Detection Relavance and Balanced Accuracy [Fig 4].

B. Logistic Regression:

• Logistic regression is coming under supervised learning. In this model we will consider one dependent variable and remaining as independent variables. In this the dependent variable should be the categorical variable. Using these variables, we can define a sigmoid curve by mapping both dependent and independent variables[22].

• **Data Cleaning:** For performing this Logistic model, firstly we need to check for the null values and if there are any null values we need to clean the dataset and then we should store the dataset in a data frame. we need a dataset containing categorical values as dependent values and all other values should full fill all the requirements of the model.

• **Data Selection:** This is very important process where we need to select the dataset which is relevant to the logistic regression. Here the dataset which we are choosing is diabetics 130-US hospitals. Where this is a dataset relevant to the diabetes patients whether they are having the diabetics or not using some procedures we can predict the diabetics Med whether the medications is to be given or not based on the test results. In this model it will predict the outcome variable in the form of Yes or No type. This also finds the probability of event occurring like yes is occurring how many times and No is occurring how many times. In my case the diabetics Med is the dependent variable where it should predict the value of taking medications in the format of Yes or No. So here we can predict the probability of people who need to take the medication and people who no need to take the medication.

• **Data Transformation:** For predicting the probability of the variables we need to train the 70 percent of data and remaining 30 percent need to be tested.

• **Data Mining:** Here we can predict the values using the GLM function. In addition to this for making sure of the output value we use confusion matrix where we can find the accuracy of the model.

Knowledge Representation:

• This explains the relationship between the variables[Fig 5]

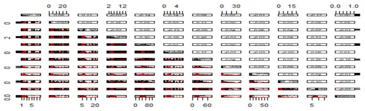


Fig. 5. Scatter Plot

C. Decision Tree

• Decision Tree can be used for both classification and regression. The decision tree is used as a decision tree it predicts the values in the form a tree where the dependent variable which is to be predicted will be considered as root node and the remaining values will be considered as the child nodes.

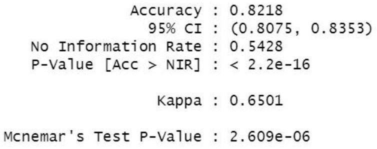
• **Data Cleaning:** As I mentioned in every process in this step we are cleaning the unnecessary data. If we do not delete the null values, output will not be predicted accurately.

• **Data Selection:** For decision tree classification the dependent variable should be binary value. I considered the Airbnb as my dataset where in this I am going to predict the room type based on the attributes price, availability, last review, minimum nights and reviews per month all these will be useful for predicting the room type. In this decision tree the algorithm works in a way that every independent and dependent variable use to build a tree. In detail scenario of this model I am having room type as a dependent value and there are price of the room, no of nights staying in the room, number of reviews, reviews per month, calculated host listings count and availability. Here the algorithm performs in a way that it compares the values with price and room type it considers the room type as root node and then price as child node where there are two categories in the room type which are shared room and private room and then it calculates the values of how many people with more price were interested in choosing shared room or private room based on this value the tree gets divided and also calculates for all other values and final leaf nodes were produced based on these nodes we can predict the output

• **Data Transformation:** Here for performing the decision tree we need to split the data for training and testing then we apply decision tree algorithm. This gives the plot of tree where we can check the leaf values and based on which leaf value, we are getting our output and so that we can decide how many no of people were interested in choosing which type of room. In this format the training data will be more than test data where we give 70 percent to the training and 30 percent to the testing and perform the algorithm. In this step we also can do post pruning or pre pruning where the tree gets divided for the better understanding.

• **Data Mining:** in this step we will perform the decision tree algorithm and predict the value based on the diagram obtained in the form of a tree. Where this tree is a classifier it classifies the values of room type as that is my dependent variable. For this model we use Decision Tree function for performing the decision tree. Once every thing is done we will calculate the accuracy of how much accurate the model is based on the confusion matrix.

Knowledge Representation:



```
                Accuracy : 0.8218
                  95% CI : (0.8075, 0.8353)
     No Information Rate : 0.5428
     P-Value [Acc > NIR] : < 2.2e-16

                   Kappa : 0.6501

  Mcnemar's Test P-Value : 2.609e-06
```

Fig. 6. Confusion Matrix of Decision Tree

• In this we can observe that the model is the best fit for this dataset where the accuracy is very high which is 0.8218.

D. Random Forest

This is used for both classification and regression. In my case I am using classification where it should be classified based on the preference of room type. The random forest is mainly works very better for the missing data.

• **Data Cleaning:** In this process as we already mentioned there is cleaning of the data process takes place where there should be no null values in the dataset before performing a model. If we have null values the model cannot predict the correct output and the whole process accuracy will be dropped down.

• **Data selection:** This step is the main step to perform any analysis because based on the algorithm the model need to be performed because some models can be performed only by using binary values and some models can be performed by continues values. In this case the prediction output variable should be categorical where we are performing the random forest classifier. I choose the dataset based on the Airbnb New York city where it explains about the hosts and guests who were renting the house and buying the house. In this case I am going to predict the room type based on the price of the room, no of nights staying in the room, number of reviews, reviews per month, calculated host listings count and availability. Based on these factors how the people previously choose their house we can predict the people are much interested in choosing which kind of house.

• **Data transformation:** In this step as we are performing the random forest we need to split the dataset for training and testing. So based on the trained values we can predict the testing values. For this I have approximately 12000 rows and 10 columns so I splitted the data where 70 percent for training and 30 percent for testing. Based on this algorithm we can easily predict the accuracy of how accurate the prediction is using confusion matrix.

• **Data Mining:** In this model the random forest is useful better than decision tree where it gives more accurate values where in random forest the whole model is taken as one input and divides the whole tree at a time and then it gives the results. But in random forest there will be a random state where we can give that random state based on this the algorithm. For example if we give random state to 10 it randomly picks the 10 values from the dataset and perform the analysis if we give random state to 20 it randomly picks 20 values for analysis. So here the values were very less at the time of prediction but in decision tree the whole values will be picked randomly so the accuracy will be less to the decision tree compared to the random forest where it considers less no of values and predicts perfect outcome. This is also a kind of tree analysis based on the random state this performs the tree analysis and provides the best result.

E. Multiple Regression

This is a regression model where this is used for the predicting an outcome variable. Where this is used as a extension for linear regression where in multiple regression the outcome value will be dependent on the multiple independent values and based on these values we can predict the dependent value variable.

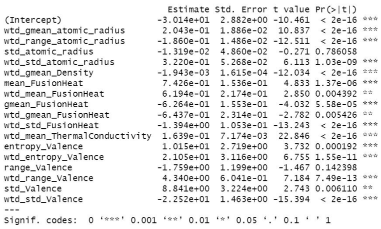
• **Data Selection:** In this process we need to select the dataset which can be able to perform the multiple regression model. According to my dataset which I have chosen is super conductivity. In this I am calculating the critical temperature based on the atomic mass, fusion heat, entropy valence, thermal conductivity, atomic radius and some more factors relating to the critical temperature. Using these factors we can only find the critical temperature but we cannot say about the material whether it is a super conductor or not.

• **Data Cleaning:** This is first foremost step where we need to clear the null values and perform the algorithm and this algorithm is used as the regression model for various datasets.

• **Data Tranformation:** In this process the data is divided into training and testing where the training data is large in size because the value based on the training data only we are going to predict the testing data. As I am giving the same ratio of division in this dataset also I am taking 70 percent as training data and 30 percent as testing data. So based on the training data the model predicts the values of the critical temperature.

• **Data Mining:** In this model we are going to find the line of best fit. Here in my case the values are dependent on each other where we cannot calculate the critical temperature if even one value is been modified. Once if we calculate the values then the dots will be plotted based on that we will draw a best fit line if the points were close to the best fit then we can say that the model is predicting correctly.

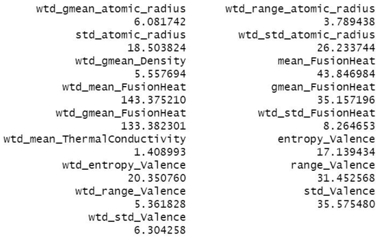
knowledge Represenatation:



```
                               Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  -3.014e+01  2.882e+00 -10.461  < 2e-16 ***
wtd_gmean_atomic_radius       2.043e-01  1.886e-02  10.837  < 2e-16 ***
wtd_range_atomic_radius      -1.860e-01  1.486e-02 -12.511  < 2e-16 ***
std_atomic_radius            -1.319e-02  4.860e-02  -0.271 0.786058    
wtd_std_atomic_radius         3.220e-01  5.268e-02   6.113 1.03e-09 ***
wtd_gmean_Density            -1.943e-03  1.615e-04 -12.034  < 2e-16 ***
mean_FusionHeat               7.426e-01  1.536e-01   4.833 1.37e-06 ***
wtd_mean_FusionHeat           6.194e-01  2.174e-01   2.850 0.004392 ** 
gmean_FusionHeat             -6.264e-01  1.553e-01  -4.032 5.58e-05 ***
wtd_gmean_FusionHeat         -6.437e-01  2.314e-01  -2.782 0.005426 ** 
wtd_std_FusionHeat           -1.394e+00  1.053e-01 -13.243  < 2e-16 ***
wtd_mean_ThermalConductivity  1.639e-01  7.174e-03  22.846  < 2e-16 ***
entropy_Valence               1.015e+01  2.719e+00   3.732 0.000192 ***
wtd_entropy_Valence           2.105e+01  3.116e+00   6.755 1.55e-11 ***
range_Valence                -1.759e+00  1.199e+00  -1.467 0.142398    
wtd_range_Valence             4.340e+00  6.041e-01   7.184 7.49e-13 ***
std_Valence                   8.841e+00  3.224e+00   2.743 0.006110 ** 
wtd_std_Valence              -2.252e+01  1.463e+00 -15.394  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 7. Significant Statistic Variables

• In the Figures [7],[8] we can observe the best accuracy values to be considered which are representing with the stars



```
          wtd_gmean_atomic_radius      wtd_range_atomic_radius
                         6.081742                     3.789438
                 std_atomic_radius        wtd_std_atomic_radius
                        18.503824                    26.233744
                 wtd_gmean_Density              mean_FusionHeat
                         5.557694                    43.846984
              wtd_mean_FusionHeat             gmean_FusionHeat
                       143.375210                    35.157196
             wtd_gmean_FusionHeat           wtd_std_FusionHeat
                       133.382301                     8.264653
  wtd_mean_ThermalConductivity               entropy_Valence
                         1.408993                    17.139434
              wtd_entropy_Valence                range_Valence
                        20.350760                    31.452568
                wtd_range_Valence                  std_Valence
                         5.361828                    35.575480
                  wtd_std_Valence
                         6.304258
```

Fig. 8. Significant values Table

will gives the more accuracy for the model. This diagrams gives us the values conclusion to be which values need to be considered for performing the models.

IV. Evaluation Methods

The evaluation methods are for evaluating the results which we have obtained from the datasets. We have executed some code based on training and testing of the data. So, after training the data we tested the data and predicted the output variables. So, to confirm the results whether the output variables were correct or not we perform Evaluation techniques. For instance, in my second dataset I have used Airbnb New York dataset where the prediction variable is room type so by performing the algorithms, we have predicted the room type. So in this module I am discussing about how accurately the values are predicted.

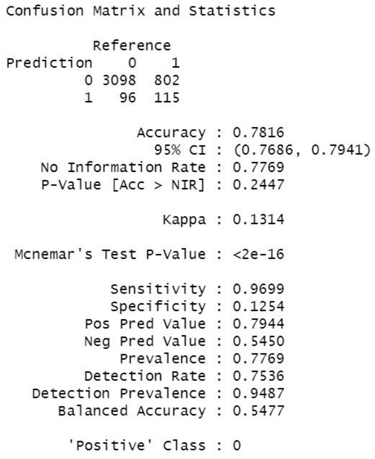
Dataset 1:

In this dataset we have performed the K-Nearest Neighbour and Logistic Regression models for predicting the diabetes Med based on the independent variables. In this dataset the diabetes Med is the effecting where it is categorical variable. The categorical variable is nothing but categorizing the values in yes or no format. In this case a person is facing diabetes issue he needs to visit hospital and check for the treatment. So here doctors checks whether he is an Inpatient or outpatient based on that if he is inpatient his details will already been registered so they will go with all these factors and perform procedures and check for the previous medications for how many years he is using those medications and after checking all back ground doctors will confirm whether a patient is having diabetes or not. If he is facing issue with diabetes, then they will suggest the Diabetes Medications. This is how the real time process work.

• Now using the machine learning models we can store all the data in a dataset and give data for training and testing in the end we can directly predict whether a person need to get the medications or nor. Like this we have predicted the Diabetics Med value using KNN and Logistic models.

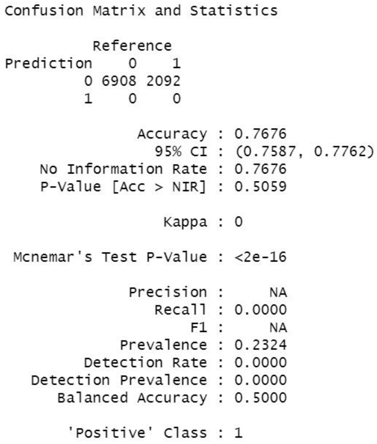
• From the above diagrams we are calculating the accuracy. As we can observe that the accuracy in the KNN is 0.7816 and in Logistic also the accuracy is 0.7676 so comparing both the models we can say that KNN performing better on the dataset where it is giving 78 percent of accuracy.

• The specificity value is 96 percent in KNN and sensitivity is 12 percent Fig [9][10].



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 3098  802
         1   96  115

               Accuracy : 0.7816
                 95% CI : (0.7686, 0.7941)
    No Information Rate : 0.7769
    P-Value [Acc > NIR] : 0.2447

                  Kappa : 0.1314

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

            Sensitivity : 0.9699
            Specificity : 0.1254
         Pos Pred Value : 0.7944
         Neg Pred Value : 0.5450
             Prevalence : 0.7769
         Detection Rate : 0.7536
   Detection Prevalence : 0.9487
      Balanced Accuracy : 0.5477

       'Positive' Class : 0
```

Fig. 9. Confusion Matrix of KNN



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 6908 2092
         1    0    0

               Accuracy : 0.7676
                 95% CI : (0.7587, 0.7762)
    No Information Rate : 0.7676
    P-Value [Acc > NIR] : 0.5059

                  Kappa : 0

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

              Precision :     NA
                 Recall : 0.0000
                     F1 :     NA
             Prevalence : 0.2324
         Detection Rate : 0.0000
   Detection Prevalence : 0.0000
      Balanced Accuracy : 0.5000

       'Positive' Class : 1
```

Fig. 10. Confusion Matrix of Logistic

Dataset 2:

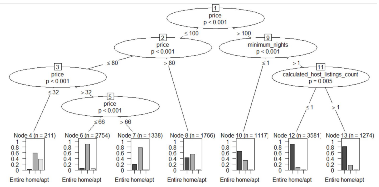
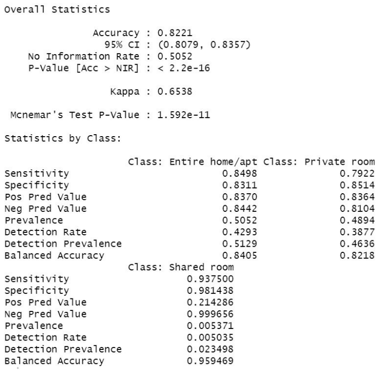


Fig. 11. Decision Tree

• As we can see that the figure it is showing that there are leaf nodes where it is predicting the values in one by one order and gives the accuracy values [Fig 11].

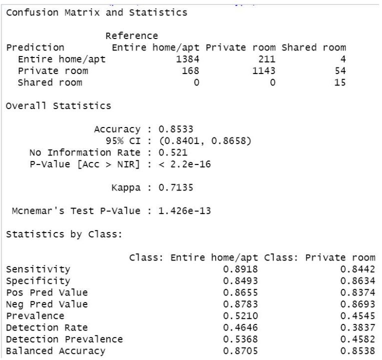
• In this dataset we are predicting the values for the room type using both Decision Tree and Random Forest where the room type is evaluated based on the availability and the no of neighbourhood counting value is the value of representing how many houses do the tenant have for renting the house and price, reviews all these are the basic factors gives the result of room type. Choosing of Decision tree and Random Forest algorithms I can answer my objective where the room type is



```
Overall Statistics

               Accuracy : 0.8221
                 95% CI : (0.8079, 0.8357)
    No Information Rate : 0.5052
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.6538

 Mcnemar's Test P-Value : 1.592e-11

Statistics by Class:

                     Class: Entire home/apt Class: Private room
Sensitivity                          0.8498              0.7922
Specificity                          0.8311              0.8514
Pos Pred Value                       0.8370              0.8364
Neg Pred Value                       0.8442              0.8104
Prevalence                           0.5052              0.4894
Detection Rate                       0.4293              0.3877
Detection Prevalence                 0.5129              0.4636
Balanced Accuracy                    0.8405              0.8218
                     Class: Shared room
Sensitivity                    0.937500
Specificity                    0.981438
Pos Pred Value                 0.214286
Neg Pred Value                 0.999656
Prevalence                     0.005371
Detection Rate                 0.005035
Detection Prevalence           0.023498
Balanced Accuracy              0.959469
```

Fig. 12. Confusion Matrix of Decision Tree

having values like shared room, Entire room/ apartment and private room. So for classifying which kind of room the people were more interested in based on the tree diagram where it classifies the nodes based on the dependent and independent variables. Now in our case the root node is room type where as the price reviews and all other factors will act as child nodes based on these values it classifies the people's choice of choosing room.



```
Confusion Matrix and Statistics

                Reference
Prediction        Entire home/apt Private room Shared room
  Entire home/apt            1384          211           4
  Private room                168         1143          54
  Shared room                   0            0          15

Overall Statistics

               Accuracy : 0.8533
                 95% CI : (0.8401, 0.8658)
    No Information Rate : 0.521
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7135

 Mcnemar's Test P-Value : 1.426e-13

Statistics by Class:

                     Class: Entire home/apt Class: Private room
Sensitivity                          0.8918              0.8442
Specificity                          0.8493              0.8634
Pos Pred Value                       0.8655              0.8374
Neg Pred Value                       0.8783              0.8693
Prevalence                           0.5210              0.4545
Detection Rate                       0.4646              0.3837
Detection Prevalence                 0.5368              0.4582
Balanced Accuracy                    0.8705              0.8538
```

Fig. 13. Confusion Matrix of Random Forest

- Here we can observe that the accuracy value in decision tree is 82 percent. Where this is predicting 81 percent accurately. In addition to this we also can observe that the specificity value for entire room is 81 percent and private room is 81 percent and for the shared room is 71 percent. Furthermore, we can observe the specificity value for entire room is 86 percent for private room is 82 percent and for the shared room is 98 percent. Where we can say that this is having very best result and predicting very much accurately Fig [12][13].
- Now coming to the random forest this is used for the same values which are categorical values in a way that the predicted variable should be in the form of classification where we can classify easily based on the independent variables. Both models are equal but in the random function we can give the random state where it can randomly takes only few values and predicts but in decision tree the whole tree is considered and the output is predicted. So, Random Forest gives better result compared to the Decision Tree.
- Comparing both the trees we can identify that the Random Forest is producing the more accuracy result with percent accuracy where as Decision Tree is providing only 82 percent of accuracy but the random forest is giving more than that of about 85 percent accuracy based on this we can say that the Random Forest is better compared to the Decision Tree[Fig 12].

Dataset 3:

- This dataset is completely based on calculating the critical temperature in Super conductors where the multiple regression model is best suitable for evaluating the values and the values are also in the continuous format. So that we can predict the quantity using regressor models in the same way we also can predict the critical temperature values. But after performing the Multiple Regression model it is not providing much good accuracy values.

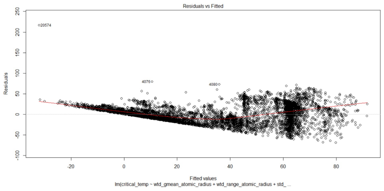


Fig. 14. Residuals Vs Fitted

- The vertical distance between the data point and regression line is Residuals. Each and every residual is formed by the data points. One data point will have one residual. If the points occurs above the line is positive below the regression line is negative and if it lies on the line then the residual at that point is Zero [Fig 14].

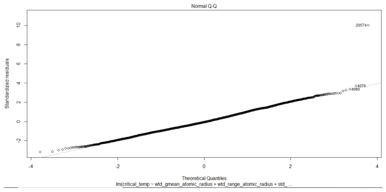


Fig. 15. Normal Q-Q Plot

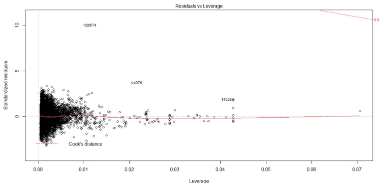
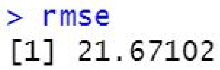


Fig. 16. Residuals Vs Leverage

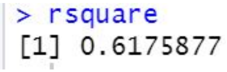
- In this regressor the evaluation methods which have been used are R-Squared and RMSE where these techniques provide the better understanding. If the R-Squared value is good, we can say that this model predicted better output.

- As my dataset is not that good because of more no of values which is not having good collinearity the independent variables are not well participated in calculating the Critical temperature values of super conductor. Based on my R-Squared value which is having only 61 percent and RMSE value which is 21 we can come to know that it is giving less accuracy which means Multiple regression for this algorithm for this dataset is not that good maybe we can try SVM are Neural networks or clustering to solve this problem [Fig 17,18].



```
> rmse
[1] 21.67102
```

Fig. 17. RMSE Value



```
> rsquare
[1] 0.6175877
```

Fig. 18. R-Squared

V. Conclusions and Future Work

According to the research, we are performing the Data Mining and Machine Learning algorithms which are famously renowned as Decision Tree, Random Forest, K Nearest Neighbour, Logistic Regression and Multiple Regression models on three datasets for predicting Diabetics Med in the diabetes dataset using K Nearest Neighbour and Logistic Regression models and also to predict room type using Decision Tree and Random Forest algorithms and to predict Critical Temperature in Super Conductors using Multiple Regression models. But from the final study of the paper, we can see that the accuracy is very low for the Multiple Regression Model.This model is not predicting good accurate values in the future studies I may use the other algorithms for revamping the accuracy of the model. In the final study, we can see that the other two datasets where we predicted the values of diabetics Medications and room type have predicted with a good result of accuracy. So these algorithms can be used for the prediction of outcomes. In the same way, I have applied the R programming for performing all the five algorithms which is the best coding language where it is having the good number of libraries and functions where we can phenomenally fabricate our way [23].

References

[1] What do Airbnb users care about? An analysis of online review comments Mingming ChengaXin Jinb

[2] The effects of Airbnb's price positioning on hotel performance LinchiKwokb

[3] A socio-economic analysis of Airbnb in New York City Lajos Boros Gábor Dudás

[4] Do airbnb host listing attributes influence room pricing homogenously? Manojit ChattopadhyayaSubrata Kumar Mitrab

[5] Enabling people with impairments to use Airbnb MelanieRandleaSaraDolnicarb

[6] Sources of distrust: Airbnb guests' perspectives Erose Sthapita,∗, Peter Björkb

[7] The evolution of 'Airbnb-tourism': Demand-side dynamics around international use of peer-to-peer accommodation in Australia. Michael Volggerab1 RossTaplinc1 ChristofPforr

[8] Airbnb Superhosts' talk in commercial homes Madalyn A.Scerria RajkaPresburyb

[9] Changes in Incidence of Diabetes in U.S. Adults, 1997–2003.Linda S.GeissMALipingPanMD, MPHBetsy Cadwell MSPHEdward W.GreggPhDStephanie M.BenjaminPhDMichael M.EngelgauMD, MPH

[10] A cross-sectional analysis of racial disparities in US diabetes screening at the national, regional, and state level

[11] Persistent organic pollutants and gestational diabetes: A multicenter prospective cohort study of healthy US women. Mohammad L.Rahmana1CuilinZhangaMelissaM.SmarrbSunmiLeecMasatoHondacKurunthachalamK AyeleaGermaine M.Buck Louisd

[12] GCK-MODY in the US Monogenic Diabetes Registry: Description of 27 unpublished variants MaySanyouraaLisaLetourneauaAmy E.Knight JohnsonbDanieladel GaudiobSiri Atma W.GreeleyaLouis H.PhilipsonaRochelle N.Naylora

[13] US ethnic group differences in self-management in the 2nd diabetes attitudes, wishes and needs (DAWN2) study MarkPeyrotaLeonard E.EgedebMartha M.FunnellcWilliam C.HsudLaurieRuggieroeLinda M.SimineriofHeather L.Stuckeyg

[14] Hypoglycemic and uricosuric properties of acetohexamide and hydroxyhexamide Ts'ai-fanYu1LawrenceBerger2Alexander B.Gutman3

[15] Diabetic kidney disease: New clinical and therapeutic issues. Joint position statement of the Italian Diabetes Society and the Italian Society of Nephrology on "The natural history of diabetic kidney disease and treatment of hyperglycemia in patients with type 2 diabetes and impaired renal function GiuseppePuglieseaGiuseppePennobAndreaNatalicFedericaBaruttadSalvatoreDi PaoloeGianpaoloReboldifLoretoGesualdogLucaDe NicolahItalian Diabetes Society and the Italian Society of Nephrology

[16] Superconducting Parameters of Cuprates Due to Microwave Irradiation In The Framework Of The Variational Theory Munasia E Rapando B.W Ndinya B.

[17] Mechanism of low thermal conductivity of xonotlite-silica aerogel nanoporous super insulation material Hailong YangWen NiDeping ChenGuoqiang XuTaoLiangLiXu

[18] An anisotropic enhanced thermal conductivity approach for modelling laser melt pools for Ni-base super alloys ShakeelSafdaraAndrew J.PinkertonbLinLibMohammed A.SheikhbPhilip J.Withersc

[19] Most Downloaded Physica C: Superconductivity and its Applications Articles J.E. Hirsch F. Marsiglio

[20] A data-driven statistical model for predicting the critical temperature of a superconductor Kam Hamidieh

[21] https://www.r-bloggers.com/machine-learning-in-r-for-beginners/

[22] https://www.youtube.com/watch?v=ypO1DPEKYFo

[23] https://www.superdatascience.com/pages/machine-learning