Analyzing Housing Prices,Suicide Rates and Likelihood of Rain Using Data Mining Techniques

Raksha Muddegowdana Koppalu Prakasha
School of Computing
MSc in Data Analytics
National College of Ireland
Dublin, Ireland
x19193181@student.ncirl.ie

***Abstract*—Real Estate is a booming industry across the world and the analysis of past price trends plays a trivial role in predicting the selling price of a real estate property in the present. Through this project we will predict the housing prices based on past price trends and multiple other factors, to help a seller successfully define the price of the house he wants to sell. Suicide Rates in countries across the globe has been increasing rapidly in the last few years, and World Health Organization has defined multiple indicators that might play a role in the number of Suicides. Through this project, we will predict the factors/indicators that are correlated to increased suicide rates and their extent of involvement, along with predicting the suicide numbers from the discovered correlated factors. Forecasting of unpredictable weather is an imminent trend in the field of weather analysis. In this project, we will look at predicting the occurrence of rainfall based on the historical weather data of the previous day.**

I. Introduction

Real Estate is booming across the globe and housing price analysis could be a step towards making a seller and every stakeholder involved, become more aware of the market trend of a place, and a step towards understanding the trends in real estate pricing. For a Seller, it is important to define the right price for the house he/she wants to sell. If the price defined is too low, potential profits might be missed out on and if the price is high, the possibility of the home getting sold decreases. This project will focus on an unbiased evaluation of the real estate pricing trends, keeping into consideration the extensive research and statistical data collected over the years to successfully help a seller predict the right price of the house he/she wants to sell. The dataset used for the prediction is created by scraping publicly available results posted to Domain.com.au by Melbourne county. The data-set includes Type of Real estate, Address, Suburb, Methods of Selling, No of Rooms, Worth, Real Estate Agent, Date of Sale, and distance from City Centre.

Death by suicide is one of the most painful matters in question to thousands of people around the globe. Every death is a tragedy, but suicide beats all by a margin because of the emotional trauma it invokes and the shock it is to people. Every minute on an average 2 people commits suicide around the world. The survey by the World Health Organization has compiled a list of multiple indicators such as GDP of a country, age of a person, population of a country, etc, in the study of suicide rates in countries across the globe. In this project, we focus on predicting the factors/indicators that are correlated to increased suicide rates and their extent of involvement through a set of research-intensive evaluations on the statistically collected data. Along with it we will also predict the suicide nos from the discovered correlated factors. The dataset used for the prediction is formed by fusing four different data-sets by WHO linked by place and time and built to find correlated factors affecting the overall suicide rates in the countries of the world. It is a dataset which compares the socio-economic information with suicide rates by year and country.

Everyone individual talks about weather always being unpredictable. A sunny, beautiful day may suddenly turn into a rainy mess and people in most cases are not prepared for the abrupt change in the weather. Forecasting the unpredictable weather is an imminent trend in the field of weather analysis. Through this project, we will be evaluating the numerous observations of historical weather data to predict the the occurrence of rainfall, the next day. The dataset used contains daily weather observations from numerous weather stations in Australia. The observations are obtained from http://www.bom.gov.au/climate/data.

II. Related Work

A. Housing Price Analysis

Accurate price prediction in real estate is significant to various stakeholders [1] such as real estate agents, house owners, house buyers, investors, etc. The paper by author by Eduard Hromada [2] talks about a novel solution involving software for analyzing real estate advertisements published online in the country of the Czech Republic. The solution involves the systematic collection, analysis, and assessment of data to depict the changes in the real estate market. The advertisements from the net are continuously collected and rigorously analyzed for their reliability. And as a unique value, the solution deals with an unbiased data set computed from intense research.

Authors, Alejandro Baldominos, Iván Blanco, Antonio José Moreno, Rubén Iturrarte, Óscar Bernárdezand Carlos Afonso in their paper [3], have explored the application of different machine learning techniques, to identify the trends of real estate opportunities for investment. They have modeled the analysis as a regression problem and built learning models with four different machine learning techniques of k-nearest neighbors, decision trees, support vector machines, and multilayer perceptrons. And the results of the models proved the regression trees to be more outperforming than the rest.

Similar to the above, another paper [4] compared the performances of machine learning techniques of Neural Network, Random Forest, and Support Vector Machine to predict housing prices. From the results of the analysis, the Random forest was seen to perform better than the other models.

The researchers have also demonstrated that one of the important components of prediction modeling is feature selection [5].

G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu, recommend a cost prediction model [6] to support house vendors and real estate agents to gain better insight on the house value.

There is also a demand and proposal [7] for high accuracy, intelligent automatic real estate evaluation system based on machine learning. And a paper [8] talks about the application of machine learning algorithms to explore the accuracy of various methods such as random forest regressions, XGboost, and the stacked regression, with results showing a significant reduction in the forecasting variance which confirms that the artificial intelligence understands real estate prices much deeper.

B. Suicide Rate Analysis

K. Hammond [9] exploits the suicidal behavior among the US Veterans and active-duty troops. They take into consideration the fact that the prediction of suicide and its attempt is hard and they project a high false-positive rate. The paper talks about the utilization of clinical data extracted from electronic health records and a study of suicidal behavior in them. Additionally, their approach offers structured data, text search, and classification techniques. The results of their paper indicated that a model with a large data set can be instrumental and be an additional approach to predictions of suicidal behavior.

According to Khon Kaen Rajanagarindra [10],there are numerous factors playing a role in the cause of suicide attempts. These factors include poverty, loss, disappointment, alcohol addiction, addiction, chronic illness, economic problems, financial problems, depression, life changes, etc. And they also point to the cases of unsuccessful suicide and repeated suicide attempts. The main focus of the report was to predict the traits of individuals who get repeated suicidal thoughts and attempt suicide multiple times, using data mining techniques.

A research conducted [11] has analyzed the pattern of suicide rates and has predicted the causes of future suicides using machine learning techniques, Artificial Neural Networks, and Support Vector Machines.

Social Network is one of the platforms where people tend to put out their thoughts, a thought which also includes suicidal ones. A research was conducted to analyze suicidal trend analysis on the social network platform of Twitter [12], wherein machine learning techniques, Support Vector Machine, and Neural Network were used. The research was able to attain an accuracy of 95.2 percent using SVM and 97.6 percent using Neural Network. Similarly, another paper addresses [13] detection of suicide ideation through deep learning and machine learning-based approaches applied to Reddit social media. It employs a combination of classification models and LSTM-CNN as a way to achieve the best relevant results.

The research applies data mining techniques to analyze and predict suicidal behaviours [14] in individuals. Prediction is done on the basis of an analysis of various risk factors such as Anxiety, Stress, Depression, etc which are accounted for using various psychological measures. The model, classification via Regression resulted in the highest accuracy in the prediction.

Artificial Intelligence is also used in optimizing suicide risk prediction and behavior management [15]

C. Rain Prediction Analysis

In India, rainfall prediction is an important aspect of the economy and helps to prevent natural disasters. Agriculture being the primary occupation, people depend a lot on rainfall. Linear and Nonlinear models of machine learning are used to predict rainfall [16].

A paper is proposed which focuses on comprehending the significance of changes in parameters such as precipitation, temperature, humidity, etc. Daily observations of weather parameters were considered for the prediction, and accuracy of the same was assessed using validation of results with ground truth. The model concludes that for forecasting, ARIMA and Neural Network works the best, and random forest gives the best result of the classification methods [17]. Another work talks about using, Linear Regression algorithms for classification [18] to forecast the rainfall of any location, wherein the classifier is trained with past weather information.

A comparison analysis was done between linear correlation and average mutual information to discover optimum input technique [19]. A multi-model method is used and compared to model a novel hybrid model for the forecast. The models included artificial neural networks, regression splines, k- nearest neighbor, and vector regression. These models are applied in addition to pre-processing techniques of moving average and principal component analysis.

Heuristic prediction of rainfall using machine learning techniques is carried out. Various categories of weather data are analyzed by linear regression methodologies [20].

A paper by R. K. Grace and B. Suganya proposes a prediction model using Multiple Linear Regression [21]. The model is evaluated using parameters such as Mean Square Error, Accuracy, and Correlation.

The paper [22] by R. S Kumar and C Ramesh from India talks about analyzing the crop productivity and rainfall prediction necessary for agriculture. It talks about the application

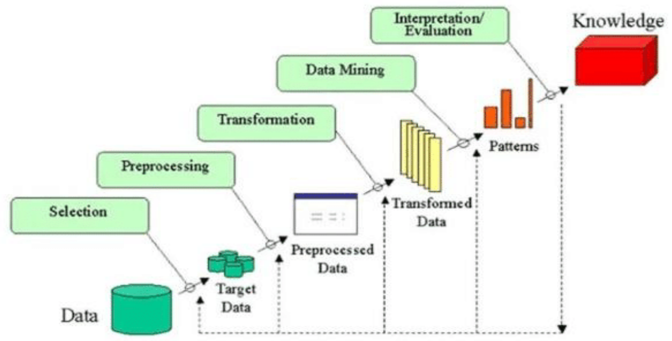


Fig. 1. KDD Methodology

of technology to predict the state of the weather and aptly determine the rainfall for use efficient use of water resources, crop productivity, and planning water stores. The paper focuses on the use of data mining techniques to estimate rainfall numerically, associated with the accuracy of the prediction

III. Data Mining Methodology

For analyzing the dataset, I have used the methodology KDD -Knowledge Discovery in Databases, which is a nontrivial extraction of unknown and potentially useful information from the dataset. KDD is an iterative process as seen in 'Fig. 1', where evaluation measures are enhanced, mining methodologies are refined, data are integrated and transformed to get appropriate efficient results.

A. Step 1: Data Selection

For housing price analysis, a large dataset with more than 10,000 rows was accessed from Kaggle [23]. The dataset consists of 21 columns with attributes such as Price, Rooms, Suburb, YearBuilt, BuildingArea, etc.

For Suicide Rate analysis, a large dataset with more than 10,000 rows was used [24]. The dataset contains 12 columns with attributes like suicide nos, gdp of a country, gdp per capita, age, etc.

The third dataset for Weather Prediction was accessed from Kaggle [25]. The dataset consisted of 24 columns with attributes like Min Temperature, Max Temperature, Humidity, Rain Tomorrow, etc.

B. Step 2: Data Pre-processing

Pre-processing of data involves Data cleaning and Data Integration. It is necessary for the data to be normalized and cleaned without having to contain any missing values or outliers to build an efficient model. These imbalanced values can affect the results of predictions. For all the datasets used in the project, Missing values were checked and appropriately imputed and dropped without ruining the dataset or incurring any data loss. The datasets were also checked for outliers, and any outliers computed were removed and dealt with appropriately without incurring any data loss.

C. Step 3: Data Transformation

In the data transformation step, variables in the data were transformed into appropriate forms required in the mining process. In the case of the housing dataset, the category levels of variable 'CouncilArea' was converted into character type. In the case of the suicide dataset, the variables gdp per capita, and gdp for year were normalized, and the variable names for the same were changed appropriately. In the case of weather dataset, correlated predictors were checked and predictors with little variance were removed. And for all the three datasets, dummy variables for different levels of the categorical variables were created, to get individual coefficients for individual states. And also to a single regression equation for multiple groups.

D. Step 4: Data Mining

Post transformation, the data was analyzed using different data mining techniques applied to each dataset. Initially, in order to understand the data, spreads of data based on parameters like mean, median, mode, 1st Quartile, standard deviations, etc.were checked, and also at this stage, we made sure that there were no missing values in the data being analyzed.

For all the datasets, in order to apply the machine learning algorithm, random samples of data were divided into the ratio of 75% training set 25% test set. The model was trained on the training set and Evaluated for performance on the remaining test set.

For the housing price analysis dataset, the objective was to predict the price for a house based on the historical data of real estate pricing, and other related variables. Plots were drawn to project the trends of pricing with respect to a year, area of housing, housing type, etc. The variable 'YearBuilt' was not used in the prediction because it had too many missing values.

Data Mining Methods used for Prediction: Regression Trees(rpart) and Random forest.

For the suicide analysis dataset, the objective was to find the correlated factors for the increased suicide rate and their extent of the effect. Also, predict the suicide nos from the discovered correlated factors. Plots are drawn to visualize the effect of variables such as gdp per year, gdp per capita, age, etc. on the suicide numbers. Regression Analysis was done to find the cause(variable) of that suicide rate. Stepwise Regression analysis was also carried out to find the factors causing an increase in suicide numbers. GDP per year was found to be the most influential variable to suicide numbers, and once this was found data mining methods were applied to predict the suicide nos.

Data Mining Methods used for Prediction: Multiple Linear Regression, PLS Regression, and Random Forest.

For the Weather Data, the objective was to predict the occurrence of rain the next day in a particular City, based on the weather data of the past. Based on the objective stated the response variable is categorical (Yes/No) in nature. The columns RISKMM, WindGustDir, WindDir9am, WindDir3pm, RainToday, Date were dropped from the dataset because more than 90% of the data were missing. Hence an overall of 17 variables was considered for analysis out of 24. Correlation plots are visualized.

Data Mining Methods used for Prediction: Logistic Regression, Decision Trees, and k-Nearest Neighbor Regression.

E. Step 5: Data Evaluation

Various Evaluation methods such as Confusion Matrix, Accuracy, ROC Curve, Area Under Curve, R Squared, Root Mean Squared Error are used to evaluate the results of prediction from different data mining methodologies. The next section will explain in detail the various evaluation methods used on each dataset and the results interpreted by them.

IV. Evaluation

In this section we will discuss the performance evaluation of different data mining methods applied to our datasets, using various evaluation metrics.

A. Housing Dataset

For the housing dataset, we have used Data mining techniques of Regression Trees and Random Forests for predicting the housing prices. To evaluate the performance of the model, in this case, we have considered evaluation methods Root Mean Squared Error and R Squared.

Root Mean Squared Error: Root Mean Squared Error or RMSE is the measure of average deviation of the output from the observations [27]. This gives us a palpable idea of how efficiently an algorithm is working.

$$RMSE = 1 - \frac{SS_{regression}}{SS_{total}} \tag{1}$$

Root Mean Squared Error gives an estimate of how far the predictions have deviated from the actual output value. The model is said to be well fit if the RMSE for the test set and training set are very similar and low as possible.

For our dataset, the Regression tree resulted in an RMSE value of 0.3245 for test data and a value of 0.3166 for training data. And in case of Random Forest, the RMSE value was 0.1861 for test data and 0.0971 for the train set. And from the values we can infer that RMSE values for test and train are similar in both the data mining methods, but take the RMSE value into consideration, Random Forest method is seen to perform better and the best fit model compared to regression tree.

R Squared: R Squared which is also known as the coefficient of determination gives us a "goodness of fit" metric for the output to the observations [27]. It is valued between 0 and 1 for no-fit and perfect fit respectively.

$$Rsquared = \frac{ExplainedVariation}{TotalVariation} \tag{2}$$

R Squared gives a measure of how close the data were to the fitted regression line. It gives the goodness of fit of the model and the best model has the value closest to 1 and similar for both test and train set.

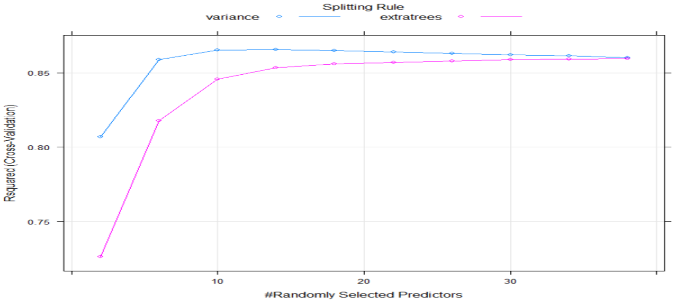


Fig. 2. R Squared

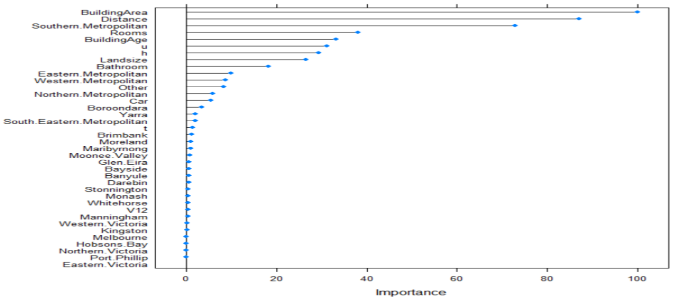


Fig. 3. Variable Importance

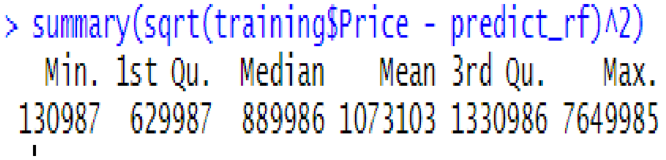
For our dataset, the Regression tree resulted in an R Squared value of 0.6426 for test data and a value of 0.6578 for training data. And in the case of Random Forest, the R Squared value was 0.8839 for test data and 0.9701 for the train set. And from the values we can infer that R Squared values for test and train are similar in both the data mining methods, but taken the R Squared values into consideration, Random Forest Method is seen to perform better and the best fit model compared to regression tree.

Insights Drawn:

- From the above evaluation done, we were able to conclude that random forest is a best fit model compared to regression trees to predict the housing price, explaining about 80% of variance in price.
- 15 predictors were used to model the Price, and as seen in 'Fig. 2', R Squared value begins to fall with additional predictors, using splitting rule. Hence we can say 15 predictors were sufficient to predict the Price.
- The variable 'CouncilArea' was not important to the Prediction of Housing Price 'Fig. 3'
- From 'Fig. 4', we can see that on average, our predictions of housing prices are within 980,000 with an interquartile range between 630,000 and 1,330,000, wherein 50% of the data falls between these Quartile and 75% of Prices is below Third Quartile. The highest Price Prediction is off by 7,640,000.

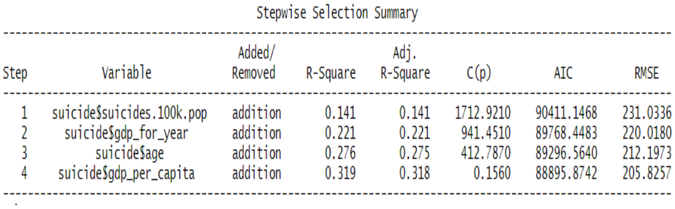
B. Suicide Dataset

For the Suicide dataset, we have used Data mining techniques of Multiple Linear Regression, PLS Regression, and Random Forests for predicting the Suicide numbers. To evaluate the performance of the model, in this case, we have considered evaluation methods Accuracy and R Squared.



```
> summary(sqrt(training$Price - predict_rf)^2)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 130987  629987  889986 1073103 1330986 7649985
```

Fig. 4. Price Prediction



```
                               Stepwise Selection Summary
-------------------------------------------------------------------------------------------------
                                     Added/                    Adj.
Step           Variable             Removed     R-Square     R-Square     C(p)         AIC          RMSE
-------------------------------------------------------------------------------------------------
   1    suicide$suicides.100k.pop   addition       0.141        0.141    1712.9210    90411.1468    231.0336
   2      suicide$gdp_for_year      addition       0.221        0.221     941.4510    89768.4483    220.0180
   3          suicide$age           addition       0.276        0.275     412.7870    89296.5640    212.1973
   4     suicide$gdp_per_capita     addition       0.319        0.318       0.1560    88895.8742    205.8257
-------------------------------------------------------------------------------------------------
```

Fig. 5. Stepwise Regression

Accuracy: Accuracy is the rate/proportion of the correctly classified occurrence out of all the occurrences [27].

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictionsMade} \tag{3}$$

Accuracy is measured as an indicator of the performance of the model and its value should be near to 1, which refers to 100% accuracy.

For our dataset, the Multiple Linear Regression resulted in an Accuracy of 38.1%. In the case of PLS Regression, we got a negative accuracy and hence the model was discarded. And in the case of Random Forests, we got an accuracy of 64.69%. From the values of accuracy, we could infer that Random Forest Method is seen to perform better and the best fit model compared to Multiple Linear Regression.

R Squared: R Squared Method is used to evaluate the performance of the model to predict suicide numbers. (2)

For our suicide dataset, the Multiple Linear Regression resulted in an R Squared value of 0.3810 for test data and a value of 0.2964 for training data. And in the case of Random Forest, the R Squared value was 0.7627 for test data and 0.7468 for the train set. And from the values we can infer that R Squared values for test and train are similar in both the data mining methods, but taken the R Squared values into consideration, Random Forest Method is seen to perform better and the best fit model compared to Multiple Linear Regression.

Insights Drawn:

- From Regression Analysis and Stepwise Regression, we found out that the variables “gdp_for_year“, “gdp_per_capita“ and “suicides.100k.population“, is the most valuable variable in the prediction of suicide numbers. We could conclude that as gdp per year for a country increases suicide numbers increases too ’Fig. 5.
- The Random Forests model resulted the high value of R Squared of 0.7627 and is considered to be the best model for predicting suicide numbers.
- Mean squared error plot for Random Forest falls very quickly for the first 100 iterations, but after this the error rate barely changes. This is as close the RF model can get to irreducible error ’Fig. 6.

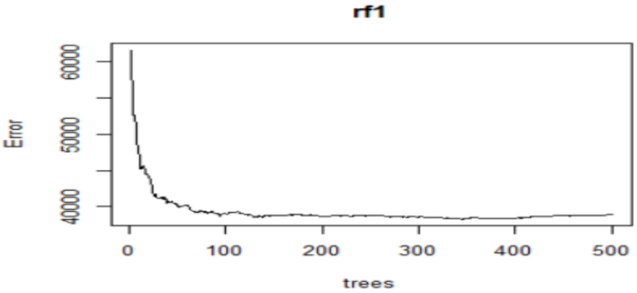
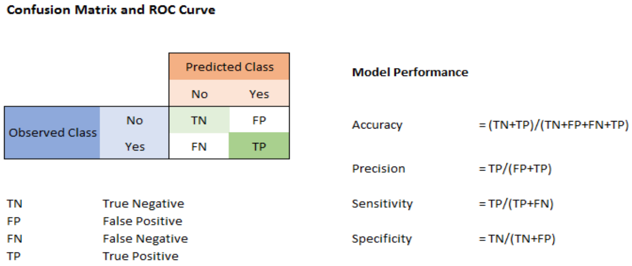


Fig. 6. Mean Square Error Plot for Random Forest



Confusion Matrix and ROC Curve

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN True Negative
FP False Positive
FN False Negative
TP True Positive

Model Performance

Accuracy	= (TN+TP)/(TN+FP+FN+TP)
Precision	= TP/(FP+TP)
Sensitivity	= TP/(TP+FN)
Specificity	= TN/(TN+FP)

Fig. 7. Structure of Confusion Matrix

C. Weather Dataset

For the Weather dataset, we have used Data mining techniques of Logistic Regression, Decision Trees, and k-Nearest Neighbors for predicting the occurrence of rainfall the next day. To evaluate the performance of the model, in this case, we have considered the evaluation methods Confusion matrix, ROC, and Area Under Curve(AUC).

Confusion Matrix: Confusion Matrix is a measure of performance of a classification model [26], measured on a set of test data for which the true values are known. Calculating the confusion matrix for our classification models gives us a better idea of the values our model is getting right and the types of errors it is making. The confusion matrix table is divided into 4 quadrants and is used to tabulate Predicted vs actual values ’Fig. 7’. Additionally, the Confusion Matrix also gives information on factors such as Accuracy, Sensitivity, and Specificity. For an ideal model, the accuracy should be as near to 100%, Specificity, and Sensitivity value near to 1.

For our weather dataset, the resulting confusion matrix for Logistic Regression has an accuracy value of 79.24%, Specificity of 0.7850, and Sensitivity of 0.8189. In the case of Decision Trees, the confusion matrix resulted in an accuracy of 83.65%, Specificity of 0.9692, and Sensitivity of 0.3664. And in the case of kNN Regression, the confusion matrix resulted in an accuracy of 91.09%, Specificity of 0.3034 and Sensitivity of 0.9214. Even though kNN Model has high accuracy, its Specificity value is low. And when considered Specificity, Sensitivity, and Accuracy, Logistic Regression performs the best.

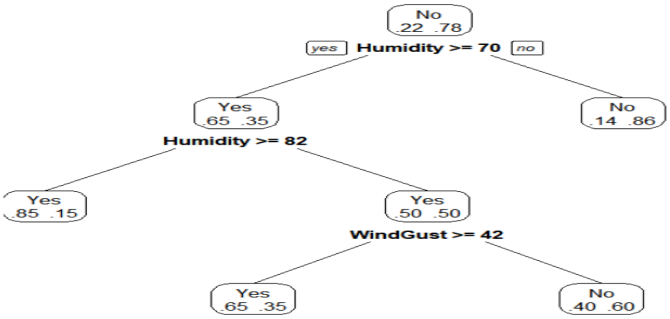


Fig. 8. Decision Tree for rainfall Prediction

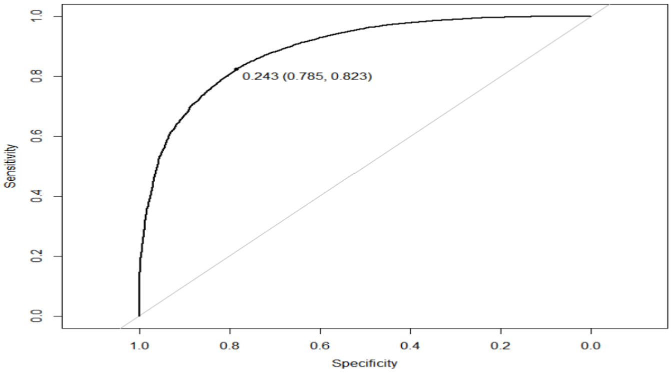


Fig. 9. ROC Curve for Logistic Regression

Hence we will consider Cohen Kappa's value to determine the best model.

Cohen's Kappa: Kappa or Cohen's Kappa is a type of classification accuracy metric, which is normalized by the randomness on your data set [27]. It is very efficient in evaluating imbalanced classes.

$$k = \frac{po - pe}{1 - pe} \tag{4}$$

While Accuracy indicates the performance of the model, it falls behind in the case of imbalanced classes. Kappa can be used to evaluate our model for its accuracy and performance even in the case of imbalanced classes. The best kappa value is one near to 1.

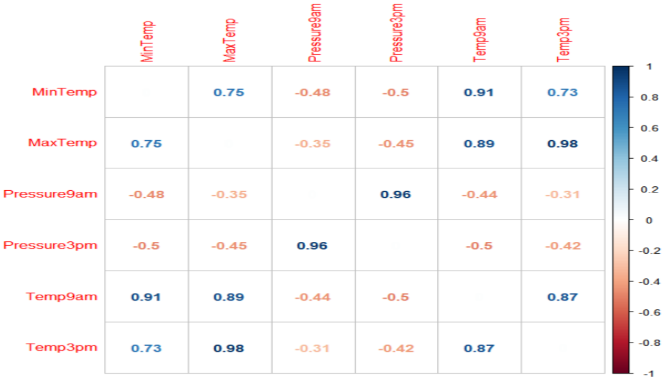
For our dataset, Logistic Regression results in a Kappa value of 0.4998. The decision tree results in a Kappa value of 0.4135 while kNN Model results in a Kappa value of 0.0784. Considering the Kappa Value we can say that the Logistic Regression Model performs the best in the prediction of rainfall.

ROC Curve: ROC(Receiver Operator Characteristic Curve) [26] helps select the best threshold value, which is generated by plotting the True Positive Rate against False Positive Rate from Confusion Matrix. ROC can be used on the models to evaluate the best threshold, and the area under the ROC is called Area Under the Curve(AUC) will give the rate of successful classification by the logistic model.

For our dataset, Logistic regression results an AUC value of 0.8883, in the case of a Decision tree it results in a value of 0.2909. Hence considering the AUC value and ROC Curve 'Fig. 9', we can say that Logistic Regression is the best fit model for predicting the occurrence of rainfall. Decision Tree doesn't generate any significant results hence the model is discarded 'Fig. 8'.

Insights Drawn:

- We were able to predict the rainfall correctly 80% of the time using Logistic Regression
- Variables related to pressure and temperature are highly correlated in the dataset 'Fig. 10'.



	MinTemp	MaxTemp	Pressure9am	Pressure3pm	Temp9am	Temp3pm
MinTemp		0.75	-0.48	-0.5	0.91	0.73
MaxTemp	0.75		-0.35	-0.45	0.89	0.98
Pressure9am	-0.48	-0.35		0.96	-0.44	-0.31
Pressure3pm	-0.5	-0.45	0.96		-0.5	-0.42
Temp9am	0.91	0.89	-0.44	-0.5		0.87
Temp3pm	0.73	0.98	-0.31	-0.42	0.87	

Fig. 10. Correlation Plot for weather dataset

V. Conclusion and Future Work

In this project, various data mining methods were applied to different datasets to analyze the prediction of housing prices, suicide numbers, and rainfall. And from the analysis, we could conclude that,

- In case of housing price analysis, Random Forest Model performed the best compared to Regression tree with an accuracy as high as 97%.
- 75% of the predicted house prices lie in the range of 630,000 to 1,330,000.
- In case of suicide number prediction, the variable “gdp_for_year“ was found to play a significant role. And the data mining technique of random trees performed better than other models in predicting the suicide numbers.
- In case of Rainfall prediction, Logistic Regression outperformed the other models with an accuracy as high as 80%, along with a high value for specificity and sensitivity.

As part of future work, I would like to apply Principle Component Analysis, to the existing models to reduce their dimensions in the dataset and thereby increase the efficiency and model performance.

References

[1] Gao, Guangliang Bao, Zhifeng Cao, Jie Qin, A. Sellis, Timos Wu, Zhiang,“Location-Centered House Price Prediction: A Multi-Task Learning Approach“,2019.

[2] Hromada, E, “Mapping of Real Estate Prices Using Data Mining Techniques“, Procedia Engineering, vol. 123, pp 233-240,2015.

[3] Alejandro Baldominos, Iván Blanco, Antonio José Moreno, Rubén Iturrarte,Óscar Bernárdez and Carlos Afonso,“Identifying Real Estate Opportunities Using Machine Learning“, .17 November 2018.

[4] Masías, Víctor Hugo and Valle, Mauricio and Crespo, Fernando and Crespo, Ricardo and Vargas Schüler, Augusto and Laengle, Sigifredo. “Property Valuation using Machine Learning Algorithms: A Study in a Metropolitan-Area of Chile“, 2016.

[5] Thuraiya Mohd, Suraya Masrom, Noraini Johari,“Machine Learning Housing Price Prediction in Petaling Jaya, Selangor, Malaysia“,International Journal of Recent Technology and Engineering (IJRTE),vol. 8, September 2019.

[6] G. Naga Satish, Ch. V. Raghavendran, M.D.Sugnana Rao, Ch.Srinivasulu,“House Price Prediction Using Machine Learning“, International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 8, July 2019.

[7] Jiafei Niu, Peiqing Niu,“An intelligent automatic valuation system for real estate based on machine learning“, vol 12, pp 1–6, December 2019[AIIPCC ’19: Proceedings of the International Conference on Artificial Intelligence, Information Processing and Cloud Computing]

[8] Marcelo Cajias, 2019.“Can a machine understand real estate pricing? – Evaluating machine learning approaches with big data“, ERES eres2019-232, European Real Estate Society (ERES).

[9] A. Ben-Ari and K. Hammond, “Text Mining the EMR for Modeling and Predicting Suicidal Behavior among US Veterans“ ,2015 48th Hawaii International Conference on System Sciences, Kauai, HI, 2015, pp. 3168-3175

[10] K. Boonkwang, S. Kasemvilas, S. Kaewhao and O. Youdkang, “A Comparison of Data Mining Techniques for Suicide Attempt Characteristics Mapping and Prediction“,2018 International Seminar on Application for Technology of Information and Communication, Semarang, 2018, pp. 488-49

[11] Syed, Sobia & Amin, Imran,“Prediction of Suicide Causes in India using Machine Learning“, Journal of Independent Studies and Research - Computing,2017.

[12] Shahreen, Nabia & Subhani, Mahfuze & Rahaman,“Suicidal Trend Analysis of Twitter Using Machine Learning and Neural Network“,pp 1-5,2018.

[13] Michael Mesfin Tadesse, Hongfei Lin, Bo Xu, Liang Yang, “Detection of Suicide Ideation in Social Media Forums Using Deep Learning“, 24 December 2019.

[14] Alina Joseph,Ramamurthy B, “Suicidal Behavior Predicting Using Data Mining Techniques“,vol. 9, Issue 4, pp 293-301,April 2018.

[15] Fonseka, Trehani M.,“The Utility of Artificial Intelligence in Suicide Risk Prediction and the Management of Suicidal Behaviors“, Australian & New Zealand Journal of Psychiatry, vol. 53, no. 10, pp. 954–964,October 2019.

[16] G. Bala Sai Tarun, J.V. Sriram, K. Sairam, K. Teja Sreenivas, M.V.B.T. Santhi, “Rainfall prediction using Machine Learning Techniques“, vol.8, Issue 7, May 2019.

[17] U. Shah, S. Garg, N. Sisodiya, N. Dube and S. Sharma, “Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques“, 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), Solan Himachal Pradesh, India, 2018, pp. 776-782.

[18] J.Refonaa, M. Lakshmi, Raza Abbas, Mohammad Raziullha,“Rainfall Prediction using Regression Model“, vol. 8, Issue 2S3, July 2019.

[19] S. Monira Sumi, M. Faisaal Zaman, Hideo Hirose,“A Rainfall Forecasting method using Machine Learning Models and its application to the Fukuoka City case“,Int. J. Appl. Math. Comput. Sci., Vol. 22, No. 4,pp 841–854, 2012.

[20] C. Thirumalai, K. S. Harsha, M. L. Deepak and K. C. Krishna, “Heuristic prediction of rainfall using machine learning techniques“, 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, 2017, pp. 1114-1117.

[21] R. K. Grace and B. Suganya, “Machine Learning based Rainfall Prediction“, 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 227-229.

[22] R. S. Kumar and C. Ramesh, “A study on prediction of rainfall using data mining technique“, 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1-9.

[23] Dans Becker, “Melbourne-Housing-Snapshot Housing Data—Kaggle“, 2018. [Online]. Available: https://www.kaggle.com/dansbecker/melbourne-housing-snapshot [Accessed on: May. 3, 2020]

[24] Rusty, “Suicide Rates Overview 1985 to 2016 Suicide Data—Kaggle“, 2019. [Online]. Available:https://www.kaggle.com/russellyates88/suicide-ratesoverview-1985-to-2016 [Accessed on: May. 3, 2020]

[25] Joe Young, “Rain in Australia Weather Data—Kaggle“, 2019. [Online]. Available:https://www.kaggle.com/jsphyg/weather-datasetrattle-package [Accessed on: May. 3, 2020]

[26] “Medium - Machine Learning Classifier: Basics and Evaluation.“ [Online]. Available: https://medium.com/cracking-the-data-science-interview/machine-learning-classifier-basics-and-evaluation-44dd760fea50. [Accessed: 03-May-2020].

[27] P. S. Kumar and S. Pranavi, “Performance analysis of machine learning algorithms on diabetes dataset using big data analytics,“ 2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions) (ICTUS), Dubai, 2017, pp. 508-513.