A COMPARATIVE STUDY ON DATA MINING ALGORITHMS FOR CLASSIFICATION AND REGRESSION

Sai Chethan Singu
School of Computing (Data Analytics)
National College Of Ireland
Dublin, Ireland
x18181937@student.ncirl.ie

Abstract---**This research work contains different machine learning models built for Income classification, House Price Prediction and US accidents severity classification. All the machine learning problems considered in this work play an important role in everyday human life. The classification model built in this work can be able to differentiate an American citizen into either one of two income levels, we applied Decision tree and Logistic regression which more or less gave the same accuracy. The house price prediction is a big problem in the Melbourne city where we built an regression model to predict the House prices based upon various physical and geographical parameters influencing House price where we applied Linear algorithm for the prediction of house price. The final problem set dealt in this work is "US Accidents" which is a very critical and essential problem which requires deployment of classification algorithms to categorize an accident to a severity level based upon various location and environmental factors. To perform this task we applied Random forest and Naïve Bayes algorithms. Out of them Random forest algorithm preformed better delivery in terms of accuracy**.

Keywords—**Decision Tree, Logistic regression, Naïve Bayes, Random Forest, Linear Regression**

I. INTRODUCTION

Human dependence on data and information in society has increased over the past two decades. With the emerging technologies there is a huge demand for machine learning. Machine learning has its applications in all types of the industries. By machine learning, automated decisions will be predicted based on sample data inputs. By machine learning, automated decisions will be predicted based on sample data inputs. In this analysis, problems relating to significant domains of social life, retail sector and public safety have been taken and addressed by using machine learning techniques. These three domains play a very important role in daily human life, bringing machine learning techniques to these domains can bring significant change in the life style of humans. The economic status play an important role in determining the social life of an individual, there is a significant interest in these days from government to standardize these social survey platforms in their country and there is a tremendous scope for machine learning techniques to be implement in these survey to obtain interesting insights on social and economic life of citizens. Retail industry is where machine learning techniques are proved to be lucrative, they are everywhere in this industry right from the e-commerce webistes, to the OOT platforms like Netflix and Prime Video, there is hardly a channel that doesn't use machine learning to provide the best options to buy and watch to its clients. Price optimization, marketing campaigns, advertisement etc. are some of the machine learning applications prominently used by the retail sector. The road safety plays an important in public safety,the transportation industry is associated with high maintenance costs, disasters, accidents, injuries and loss of life. Hundreds of thousands of people across the world are losing their lives to car accidents and road disasters every year. The deployment of smart, connected sensors, combined with machine-learning-powered analytics tools, can enable us to gather information, make predictions and reach decisions that will make our roads safer. The income classifiers gives us with the economic distribution in a society and factors influencing it which can aid governments to make policies. The regressor built in this work answers us with the price that can be associated with a home in a geographical location which will help the buyers, sellers and even governments. The accidents classifier answers us with the severity level which aids to tackle with factors causing accidents of higher severity and even in making quick response in treatment of severity. Knowledge Discovery in Database otherwise known as KDD process was considered for the analysis. KDD is a process of extracting the useful information or knowledge from the data stored in the databases. It is done by using data mining methods or algorithms.

II. RELATED WORK

The inequality of income and wealth in United States is one of the big concerns. To address the income inequality problem, data mining techniques were used by the author [1]. One of the Ensemble Learning algorithm, Gradient Boosting is used with grid search and hyper parameter tuning on the Census data that clocked accuracy of Eighty-eight percentage.

The factors that are most important in predicting the individual income were taken into consideration and performed Random forest classification by the author [2]. The analysis was done on key factors like marital status, capital gain, education, age and work hours. Taking top five important features into account, the model predicted eighty five percentage of accuracy.

Several machine learning algorithms like Naïve Bayes, Decision Tree, KNN, Extreme Gradient Boosting and Support vector machine have been applied along with feature engineering and hyper parameter tuning to classify the adult income [3]. By using various algorithms, accuracy was examined. Out of those algorithms, Random forest and Extreme gradient boosting produced high accuracy.

Supervised methods based on classification and regression provides good generalization along with accuracy according to the author [4]. SVM and PCA algorithms were performed to evaluate the income by taking training time, parameter search, total number of support vectors and accuracy into consideration. With low number of parameters, computational time was reduced by more than half. In terms of accuracy, SVM performed better results. In this research paper, processing time was significantly reduced by using SVM and PCA algorithms.

The income prediction of US citizens was addressed by identifying the various important features that eventually helps to reduce the complexity of different machine algorithms used. Different classification algorithms were performed that produced different accuracy rates [5]. Most useful attributes which were used for prediction includes education, age, sex occupation and hours per week. This suggests that taking only the important attributes will give better accuracy.

By inspecting the two different approaches of traditional statistical model and machine learning [6]. Logistic Regression was used as statistical model and different machine learning techniques used are Neural networks, classification, regression trees and support vector machine. Since majority of the records in the data fall under less than 50,000 $ a year, classifiers were compared using ROC curves.

One of the challenges that many industries are facing is finding out the income facts of the individuals. With the characteristic of the employees like age, work class, education, working hours per week and with other information, salary can be classified using machine learning techniques [7]. Five classifiers have been used to classify the income. From the research it was found that Gradient Boosting algorithm predicted income with high accuracy.

House price prediction or forecasting is very important particularly in Real estate sector. With historical transaction data, Machine learning models were applied to analyze the property prices in Melbourne city [8]. Support vector machine and Step wise regression models that are based on Mean squared error were analyzed. The major finding related to this paper is that Step-wise regression can also be performed to predict the house prices.

To make better decisions by government and developers in the real estate sector on corresponding regions, machine learning models have been performed to predict the prices of houses in Boston. In this paper [9] least square and partial least square and SVM algorithms were adopted to predict prices. It concludes that SVM and Least Squares SVM models performed better than partial least squares as the data is non-linear. The key take away is that if the data is having non linear relation, then using SVM and least squares can perform better predictions.

Change in the real estate prices can affect various investors, bankers and policy makers. Thus evaluating the real estate value is important in terms of economic index. This was addressed by the author [10] by using different prediction models. That includes Logistic Regression, SVM, Lasso Regression and Decision Tree models. These models were compared based on RMSE and accuracy. The important takeaway from this paper is that apart from Lasso Regression other models gives better accuracy.

Prediction of housing prices will lead to the development of Real estate and their policies according to the author [11]. For research, Machine learning methodologies were used for the prediction of housing prices. Model based on Naïve Bayesian and AdaBoost algorithms was built and their accuracy was measured.

There is a close relationship between house prices and the Economy which plays a major role in prediction of house prices stated by the author [12]. It helps in making decisions by sellers and buyers without bias. It was performed by using PCA and Support Vector Regression models. Recursive feature elimination was done to increase accuracy to eighty six percentage. It suggests that reducing the number of features, by selecting most important features boosts the accuracy without overfitting.

The rise and fall of the housing prices and the factors that influences the most were addressed by the author using various Regression techniques [13]. Various features selection techniques have been used to deal with outliers and missing values. The performance of the models was measured by using precision, accuracy, specificity and sensitivity. Random Forest which was used for the analysis produced better accuracy.

It addresses the challenge of predicting house prices by using Multiple Linear Regression analysis. The statistical graphs were used for interpretation of the model results [14]. He used features such as number of bedrooms, floor size, garage size, lot size category and number of bathrooms as influential features and performed Linear Regression techniques. The important finding related to this paper was that all the factors should be analyzed and then taking most important factors into consideration.

Numerous studies have been made on analyzing traffic accidents and predicting traffic severity. In this paper [15] logistic and linear regression algorithms were implemented and evaluated using out of bag error rate, mean square and root mean square error. The variables that have strong correlation were taken into account and performed Random forest algorithm. Random forest was chosen to analyze complex relationships between the variables and to avoid overfitting.

The risk factors associated with motorcycle crash were analyzed by using Random Forest, Instance based learning with parameter K and Decision tree classifier algorithms. Algorithms were validated by using the 10-fold cross validation method in this analysis [16]. From those three algorithms, Random Forest showed more accuracy. The finding from this analysis was Random Forest and eliminating unimportant features with cross validation performed better in terms of accuracy.

Road accidents has become a major problem in many countries. There are many factors that were taken into consideration to find out the intensity of the accidents [17]. This was done using Naïve Bayes, KNN, And AdaBoost methods. Out of which AdaBoost technique predicted better results. The severity was categorized into four and most important eleven factors are used for the analysis. The accuracy achieved was eighty percentage.

Various factors involved in road accidents were analyzed using Random Forest and Decision Tree algorithms to find the patterns in the data by using K-fold Cross Validation [18]. To make sure data is unbiased, cross validation was used. The major finding from the paper was performing Cross validation if the data is biased.

The highway traffic accidents and traffic flow turbulence was addressed by using applying machine learning algorithm on real time traffic data [19]. K-nearest neighbor model was used for the prediction of traffic accidents. The environment factors were taken for analyzing.

II. METHODOLOGY

There are different methods that determine the workflow and processes. The methodology used for this analysis was KDD which is other wise known as Knowledge Discovery in Databases. It refers to the discover of knowledge in the data with the application of specific data mining techniques. KDD includes selection of data, data preprocessing, transformation of data, applying data mining techniques and interpretation (or) evaluation of performance.

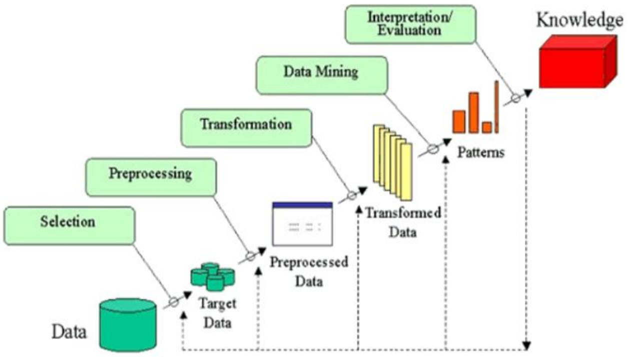


Fig.1. Steps in KDD

Dataset – 1:

(i) Data Selection:

The adult census income data set is to predict whether income exceeds $50k/year in US country. The dataset is extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker that consists of approximately 32k rows and 15 columns. Different independent variables includes work class, age, education, occupation, race and many others. In those, an independent variable named ‘final weight’ is a metric devised by the Population Division at Census Bureau on monthly basis. The weights on Current Population Survey (CPS) files are controlled to independent estimates of the civilian non-institutional population of US. The term estimate refers to population totals derived from CPS by creating "weighted tallies" of any specified socio-economic characteristics of the population. People with similar demographic characteristics should have similar weights.

(ii) Data Pre-processing and Transformation:

This dataset is a collection of independent variables affecting the income and income class as target variable 15 columns (14 Independent + 1 Dependent). The columns of [workclass, occupation, native_country] contain entries in the form of ‘?’ which were replaced by ‘NaN’ values. Of these ‘NaN’ values a few were filled with Mode value of respective columns and a few samples were dropped. The correlation was plotted using heat map for checking correlation between the variables.

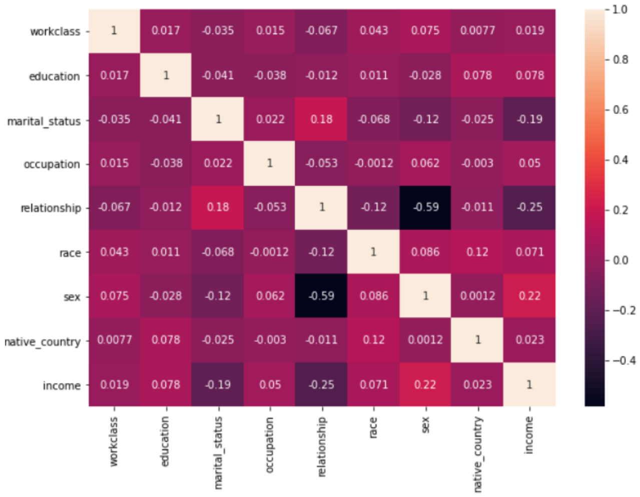
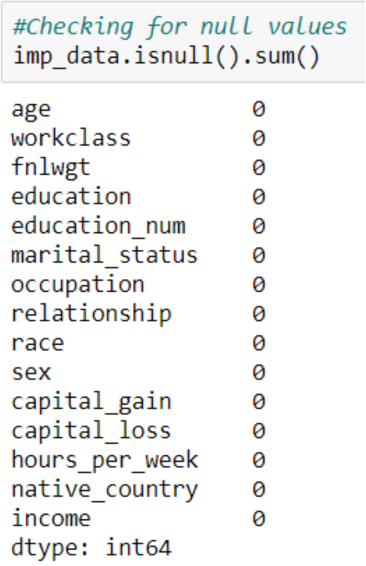


Fig. 2. Correlation Plot for Income Census

As we observe there are various data types across the independent variables of dataset. To maintain uniformity the columns [workclass, education, marital_status, occupation, relationship, race, sex , native_country] were encoded into categorical variables. The target variable ‘income’ is also categorized to form ‘Two’ categories for >$50k/year or <$50k/year. The dataset is made split into train and test datasets with 70:30 ratio and scaling of independent variables is made using standard scaling techniques to obtain data of zero mean and unit variance. The below figure 3 shows there are no null values in the data.



```
#Checking for null values
imp_data.isnull().sum()
```

```
age               0
workclass         0
fnlwgt            0
education         0
education_num     0
marital_status    0
occupation        0
relationship      0
race              0
sex               0
capital_gain      0
capital_loss      0
hours_per_week    0
native_country    0
income            0
dtype: int64
```

Fig. 3. Checking null values

(iii) Data Mining Model:

The logistic regression classifier is built upon the train subset and while classifying the test subset it generates an accuracy of 82%. An alternative approach using Decision Tree Classifier is applied upon the same train and test dataset obtained above. The classifier built on train subset gives an accuracy of 81% on test subset. In figure 4, Decision Tree was plotted by using Sklearn library with Entropy as critirien.

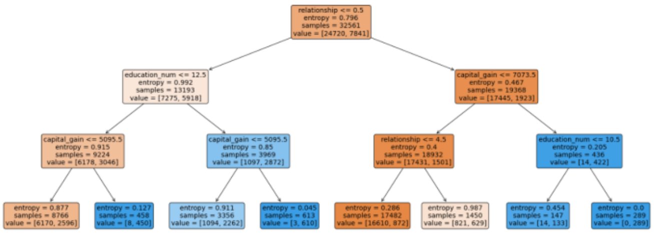


Fig. 4. Decision Tree

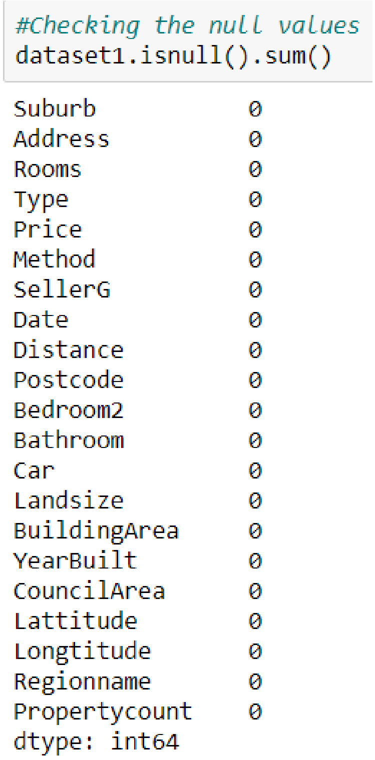
Dataset – 2:

(i) Data Selection:

The “Melbourne Housing Market” dataset is setup in the Melbourne city of Australia which contains history data of house prices in the city and the factors influencing it. Dataset consists of nearly 35k rows with 21 attributes. The data was scraped from publicly available results posted every week from Domain.com.au and made publicly available on dataset repository platforms. The dataset includes Address, Type of Real estate, Suburb, Method of Selling, Rooms, Price, Real Estate Agent, Date of Sale and distance from C.B.D, property size, land size, council area. The machine learning problem here is the regression problem as price prediction is our target variable of interest. The regressor we built can act as assistant in recommending individuals telling them about best suburbs to buy homes, property’s which are value for money, expensive side of town. It can also help government to plan city more efficiently.

(ii) Data Preprocessing and Transformation:

By exploring the dataset gives information that there are 21 columns which includes 20 Independent and one Dependent variable. When correlation map is drawn a high correlation between column “Bedroom2” and “Rooms” variable. considering this high multi-collinearity the “Bedroom2” is dropped which will land us with 20 columns now with a reduction in one independent variable.



```
#Checking the null values
dataset1.isnull().sum()
```

```
Suburb           0
Address          0
Rooms            0
Type             0
Price            0
Method           0
SellerG          0
Date             0
Distance         0
Postcode         0
Bedroom2         0
Bathroom         0
Car              0
Landsize         0
BuildingArea     0
YearBuilt        0
CouncilArea      0
Lattitude        0
Longtitude       0
Regionname       0
Propertycount    0
dtype: int64
```

Fig. 5. Checking null value

The independent variables present in the dataset were categorized to maintain uniformity across the dataset, this will help to preserve the contribution of information from independent variables to outcomes. The effect of outliers on model building is reduced using the z-score function which will characterize a raw score by a value above or below the mean value to identify the outliers. A split is made in dataset to obtain Train and Test subsets with a ratio of 70:30.

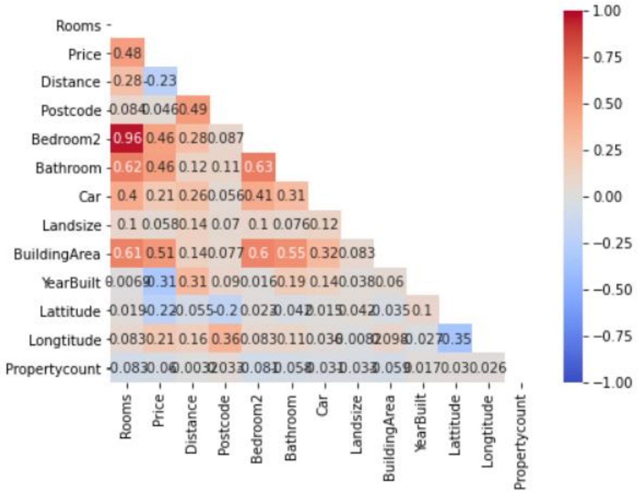


Fig. 6. Correlation Plot for Melbourne Housing prices

(iii) Data Mining Model:

The dataset obtained after Pre-processing and Transformation is one with Train and Test subsets. A feature selection mechanism of Recursive Feature Elimination with Cross Validation (RFECV) is utilized to identify the optimal number of features while improving upon R-squared metric with Random Forest regressor as base line estimator. This mechanism reports a 17 optimal independent which gives best R-squared value among all other possible alternatives. The features including Building area, Date, Year built, Distance, Council area contributes the most in predicting house prices. The linear regression algorithm is built upon train subset containing 17 optimal features obtained above, when deployed on train subset the linear regression algorithm gives an coefficient of determination (R-squared) value of 71.96.

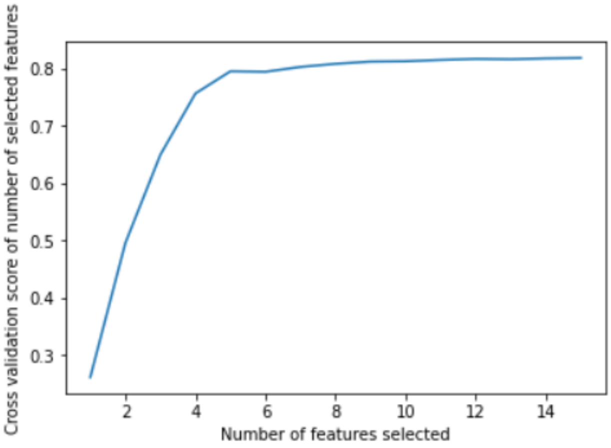


Fig.7. Important features in data

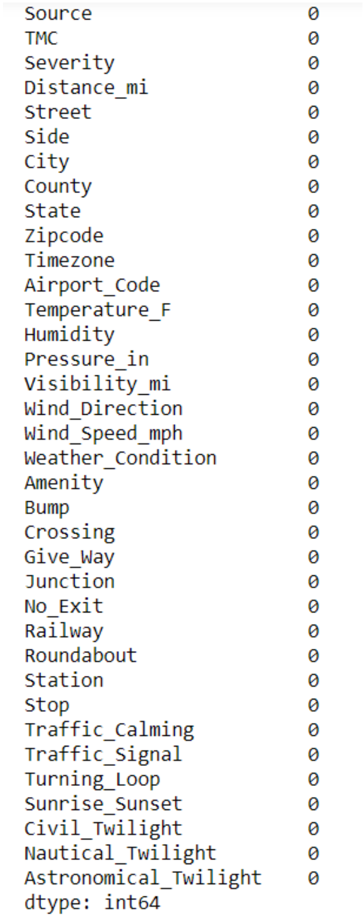
Dataset -3

(i) Data Selection:

The “US-Accidents: A Countrywide Traffic Accident Dataset” is a dataset, which contains accidents information happened across various states of the United States. The data is continuously being collected with the aid of several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. The variant of the dataset choose in this work contains 100 Thousand entries with 49 variables. The dataset was analysed to study accident hotspot locations casualty analysis and extracting cause and effect rules to predict accidents and to study the impact of precipitation or other environmental stimuli on accident occurrence.

(ii) Data Pre-processing:

By exploring the dataset visually and statistically its observed that there are few independent variables which contribute enormous amount of Null-Values when compared to others in dataset. As an advantage the information provided by these variables can also be obtained from their relevant peer variables because of multi-collinearity present among them. Basing upon the insights from correlation map, box-plots and Null-Value detection schemes a few independent variable which contribute less to classifier building were removed. This data pre-processing routine lands us with approximately 75 Thousand entries and 35 variables. A python library for data processing and analysis called ‘Pandas’ is extensively used in this section of work.



```
Source                   0
TMC                      0
Severity                 0
Distance_mi              0
Street                   0
Side                     0
City                     0
County                   0
State                    0
Zipcode                  0
Timezone                 0
Airport_Code             0
Temperature_F            0
Humidity                 0
Pressure_in              0
Visibility_mi            0
Wind_Direction           0
Wind_Speed_mph           0
Weather_Condition        0
Amenity                  0
Bump                     0
Crossing                 0
Give_Way                 0
Junction                 0
No_Exit                  0
Railway                  0
Roundabout               0
Station                  0
Stop                     0
Traffic_Calming          0
Traffic_Signal           0
Turning_Loop             0
Sunrise_Sunset           0
Civil_Twilight           0
Nautical_Twilight        0
Astronomical_Twilight    0
dtype: int64
```

Fig. 8. Checking null values

(iii) Data Transformation:

A glimpse on dataset after pre-processing will disclose that there different data types across the columns of dataset. The target variable here which is “Severity” an indicator of severity of accident is a categorical variable with 4 levels. To maintain uniformity across the dataset all the independent variables were converted to similar data type by converting the string object and Boolean data type independent variables into categorical variables The scaling operation is performed on the independent variables of the dataset using Standard Scaling technique to bring the data around zero mean and unit variance. The dataset is divided into Train and Test (validation) data in a ratio of 70-30, to build the model and evaluate the classifier with relevant metrics. A python based machine learning library “Sklearn” is extensively used in this section for the purpose of encoding and standardizing for Train-Test split.

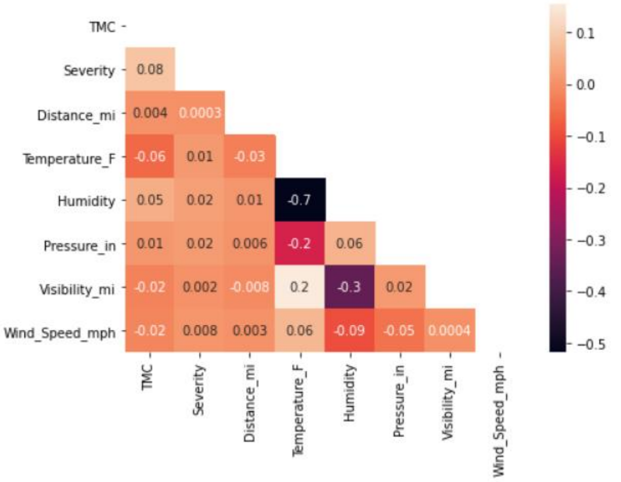


Fig. 9. Correlation Plot for US Accidents

The column ’Source’ a string data type which refers to source of data collection is converted into Two categorical variables with The column ‘TMC’ which contain message codes were categorized to form 21 code categories. The columns [Street, Side, City, County, State, Zipcode, Timezone, Airport_code] which amounts for geographical information of accident is converted into categorical variable to categorize the severity based on the geographical location. The independent variables [Wind_Direction, Weather_condition, 'Sunrise_Sunset','Civil_Twilight', 'Nautical_Twilight', 'Astronomical_Twilight'] which depicts the atmospheric factors influencing the outcome were converted into categorical variables. The columns ['Amenity', 'Bump', 'Crossing', 'Give_Way', 'Junction', 'No_Exit', 'Railway', 'Roundabout', 'Station', 'Stop', 'Traffic_Calming', 'Traffic_Signal', 'Turning_Loop'] which contain presence or absence of respective variables is converted into categorical variables.

(iv) Data Mining Model:

The data set obtained after data transformation is clean and a transformed one into Train and Test subsets with 70:30 ratio. To further reduce the dimension of the machine learning problem a feature selection mechanism is opted by using the Random Forest algorithm’s attribute of feature_importance which calculate the Information Gain of Independent variables with the help of Gini Index. All the 35 Independent variables were ranked as per Feature importance of which top 13 were selected as they contribute 88% of total information for classifier building. With the obtained 13 optimal features a Random Forest Classifier and Naïve Bayes classifier is built on Train subset. A Random Forest Classifier with 500 estimators and Entropy as criterian gives accuracy of 92% on test subset. On the other hand 70% of accuracy is observed on test subset when Naïve Bayes is used for classification.

IV EVALUATION METRICS

Dataset-1:

Two machine learning models were performed on this dataset that includes Logistic and Decision Tree algorithms. The metrics which were used to determine the efficiency includes sensitivity or Recall, specificity and accuracy and also using Receiver operating charecteristics curve. The area under curve value reveals that Logistic Regression performed better than Decision Tree.

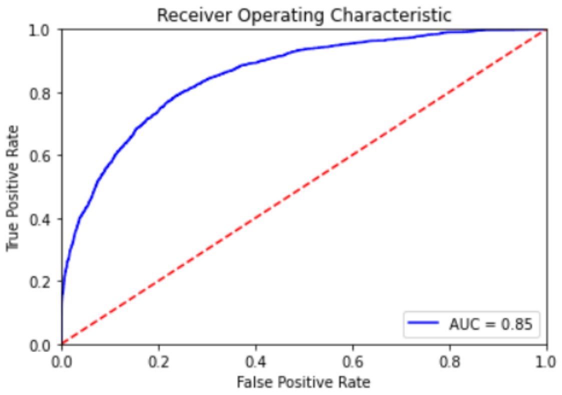


Fig.10. Roc Curve for Logistic Regression

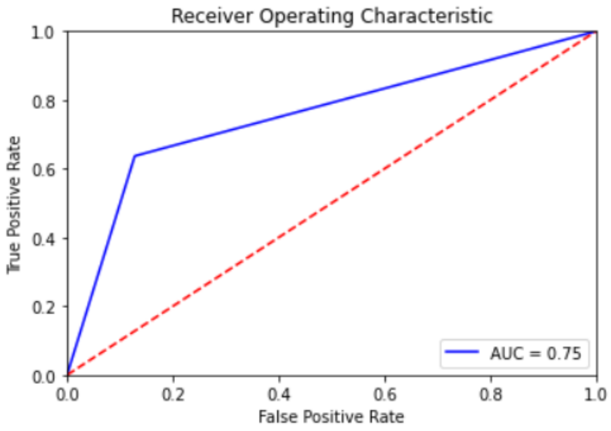


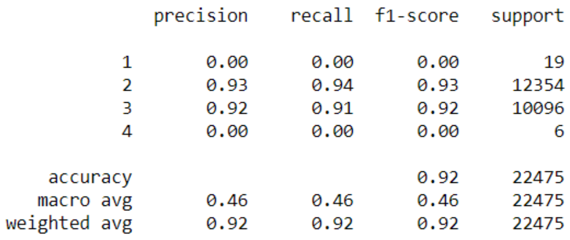
Fig.11. Roc Curve for Decision Tree

Dataset -2:

Linear regression model has been performed on this dataset, Where R-squared values obtained was 72%. R squared value tells how close the data is fitted with respect to the Regression line. In this analysis, Efficiency was improved by considering the most important features by using feature ranking technique. After the prediction the obtained RMSE value is 286835.19.

Dataset – 3:

Two algorithms namely Random Forest and Naïve Bayes have been used for the evaluation of dataset. For evaluating of the performance accuracy was taken into account which shows that Random forest was performed better predictions with accuracy of 92% when compared to Naïve Bayes with accuracy of almost 70%. Kappa was obtained for Naïve Bayes algorithm which is 0.38, which shows that there is a fair classification accuracy. The kappa value for random forest is 0.84 which is close to one, this shows that random forest performs well over Naïve Bayes.



```
              precision    recall  f1-score   support

           1       0.00      0.00      0.00        19
           2       0.93      0.94      0.93     12354
           3       0.92      0.91      0.92     10096
           4       0.00      0.00      0.00         6

    accuracy                           0.92     22475
   macro avg       0.46      0.46      0.46     22475
weighted avg       0.92      0.92      0.92     22475
```

Fig.12. Showing accuracy for Random forest model.

V. CONCLUSION AND FUTURE WORK

The application of logistic regression and decision tree on US income dataset gave almost similar accuracy of eighty two accuracy. The future work on this data could be done by using random forest and advanced ensemble methods like AdaBoost for the improvement of accuracy. Linear regression was performed on second dataset that is on Melbourne housing market for the prediction of house prices. We obtained R-square score of 71.96. Linear regression is applied only when the data is linear but whereas SVM can handle both linear and non linear data. Future work could be done using SVM on the same dataset. Third dataset on US accidents was performed by using Random forest and Naïve Bayes algorithms. Ninety two percent accuracy was obtained by using random forest and almost seventy percentage was given by Naïve bayes. In this dataset, severity levels of 2 and 3 heavily dominates over severity levels of 1 and 4 which causes class imbalance so further work can be done to over come this problem. Since we have nearly fifty independent variables, performing PCA could provide better results by categorizing important features and also tuning hyper parameters by using grid search technique that can eventually lead to better classification.

V. REFERENCES

[1] N. Chakrabarty and S. Biswas: “A Statistical Approach to Adult Census Income Level Prediction.”

[2] Bekena and S Menji: “Using decision tree classifier to predict income levels” 30th July, 2017

[3] M. Topiwalla: ” Machine Learning on UCI Adult data Set Using Various Classifier Algorithms And Scaling Up The Accuracy Using Extreme Gradient Boosting”, University of SP Jain

[4] Alina Lazar:” Income Prediction via Support Vector Machine” ICMLA Dec,2004, USA

[5] C. Lemon, C. Zelazo and K. Mulakaluri:” Predicting if income exceeds 50K$ per year based on 1994 US Census Data with Simple Classification Techniques”, https://cseweb.ucsd.edu/jmcauley/cse190/reports/sp15/048.pdf.

[6] H. Zhu:” Predicting Earning Potential using the Adult Dataset” 5th Dec, 2016

[7] S. Bramesh and B. Puttaswamy:” Comparative Study of Machine Learning Algorithms on Census Income Data Set” in Proceedings – Aug, 2019 International Journal Of Engineering Research And Application,pp 78-81

[8] D. Phan:” Housing Price Prediction using Machine Learning Algorithms: The Case of Melbourne City, Australia” in Proceedings – International Conference on Machine Learning and Data Engineering (ICMLDE), 2018.

[9] Jingyi Mu, Fang Wu and A. Zhang:” Housing Value Forecasting Based on Machine Learning Methods”4th Aug, 2014

[10] N. Shinde and K. Gawande:” VALUATION OF HOUSE PRICES USING PREDICTIVE TECHNIQUES” in Proceedings –“International Journal of Advances in Electronics and Computer Science” Vol-5, Issue-6, 2018

[11] B. Park and J.K Bae:” Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data”. Expert Syst. Appl.

[12] Jiao Yang:” Housing Price prediction Using Support Vector Regression”, San Jose State University.

[13] D. Benerjee and S. Dutta:” Predicting the Housing Price Direction using Machine Learning Techniques” in Proceedings – 2017, IEEE International Conference On Power, Control, Signals and Instrumentation Engineering (ICPCSI).

[14] I. Pardoe:” Modeling Home Prices Using Realtor Data” in Proceedings – Journal Of Statistics Education, Vol-16, 2008

[15] J. Lee, T. Yoon, S. Kwon and Jontae:” Model Evaluation for Forecasting Traffic Accident Severity in Rainy Seasons Using Machine Learning Algorithms: Seoul City Study”, 2019

[16] L. Wahab, H. Jiang:” A comparative study on machine learning based algorithms for prediction of motorcycle crash severity” Published 4th Apr, 2019.

[17] F. Labib, A. Rifat, M. Hossain and F. Nawrine:” Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh” in Proceedings - 7th International Conference on Smart Computing & Communications, 2019

[18] G. Ramani, S. Shanthi:” Classifier Prediction Evaluation in Modeling Road Traffic Accident Data”in Proceedings – IEEE International Conference on Computational Intelligence and Computing Research, 2012.

[19] L. Yisheng, S. Tang:” Real-time Highway Traffic Accident Prediction Based on the k-Nearest Neighbor Method”,in Proceedings – IEEE International Conference on Network and Information Systems for Computers (ICNISC), 2017