Data Mining & Machine Learning Project

Karan Mahendra Singh
School of Computing
National College of Ireland
Dublin, Ireland
karanmsingh10@gmail.com

***Abstract*—This is the report of a machine learning project where five learning models were applied on three large sized datasets. The three datasets were from three different domains and hence, not related. The first dataset contained data about landslides, while the second contained data about air pollution and the third – contained Google Play Store data. The KDD methodology of discovering knowledge was used throughout the project. Three research questions were devised to form the basis of analysis. Random forest and Multiple Regression were applied on the first dataset to predict the number of fatalities with the location accuracy of rainfall triggered landslides. Thereafter, Simple Linear regression was applied to predict the levels of PM2.5 pollutant with NO_2 levels as a predictor variable. Finally, on the third dataset, Logistic Regression and Naïve Bayes were applied to classify if an app was free or paid based on the number of installs it attracted. The project thus attempts to compare the working and results of the above-mentioned machine learning models to check which one fits better.**

Keywords—Machine learning, KDD, Random Forest, Multiple Regression, Simple Linear Regression, Logistic Regression, Naïve Bayes

I. INTRODUCTION

Broadcasting information about probable disasters takes a lot of efforts. Sometimes it is just the media of the spread of news that takes time while sometimes it is the detection. In the past, the spread of information among people was relatively more difficult and time consuming, for there was lack of media. Furthermore, even the disaster detection systems were not efficient enough. However, that is not the case now.

The number of smartphone users is ever increasing and therefore, the spread of news is not a concern anymore. The remaining: early recognition of disasters too is now gaining pace. The usage of wireless sensor networks (WSN) for detecting any shift in the usual state of natural objects such as the mountains: for landslides, the ocean: for tides and so on is now becoming prevalent so that we could identify and prepare for any upcoming natural calamity.

There has been constant upgrade in this domain in the recent couple of years, enabling us to prepare and mitigate the loss of lives and/or property wherever possible.

Landslides occur due to a variety of reasons. They occur when the shear pressure on slopes of hills is high [5]. They also initiate when there is high precipitation in the area and large quantities of water percolate inside the soil, loosening it up from the inside – so much that it is not able to stand on its own and collapses. Furthermore, geographical aspects such as the under structure of the land in that area also plays a major role. If the structure is loose with no rocks or only rocks, it is unstable.

There is no doubt that precipitation plays a key role in landslides and therefore regional landslide recognition uses a predefined limit for detection of the movement of debris [1].

Thus, in its first analysis, this project attempts to find the following:

RO 1: Predict loss of life with the accuracy of rain fall location and the size of the landslide.

The second analysis relates to air pollution. It is a major problem in a wide number of cities across the world. Many larger cities are facing a variety of problems due to air pollution and are seeking solutions to it.

Smog, a mixture of smoke and fog causes a thick layer leading to visibility and breathing problems. Exposure to such heavy amounts of pollution causes compulsive pulmonary disorders [9]. The composition of other pollutants such as SO_2, CO, PM 2.5, PM 10 and so on causes them to react with themselves in the air and cause further airborne hazard.

Numerous measures are being implemented by those who are aware of air pollution. These include carpooling, using public transport, not burning garbage in the air, moving industries that excrete smoke to parts outside the city and so on. However, air pollution levels remain about the same and the air we constantly inhale remains hazardous.

Since PM 2.5 is the most harmful, in its second analysis, this project has the following aim:

RO 2: Predict the PM 2.5 levels with the levels of NO_2 present in the air.

The third and final analysis pertains to smartphone apps. With the high degree of dependence of almost all individuals on their smartphones, an analysis on something related to them was necessary.

The hardware sophistication being used in smartphones is getting higher by the day. All the mobile phones in the past could was call some one and provide some basic functionality. However, as years passed and development gained pace, they started offering more and more features.

Not only features, but also the kind of hardware being used to manufacture a phone kept getting better. Today an average smartphone has a 2 GB RAM as opposed to a 5MB hard drive in the past. Furthermore, even the battery backup of phones is improving constantly. Power chargers that charge a battery as big as 4000 mah in 1hr or lesser are being used to charge a battery in lesser time and give longer power back up.

With the elegance in hardware and software came the elegance of features. Smart phones today can do many complex tasks which required a computer in the past. Just sending and receiving mails and other data is too trivial an application. With the plethora of apps available on app stores, users can select the app that performs the tasks they require to be done and then even delete it. Such is the kind of feature dynamism smartphones have achieved in the date.

Developers develop apps that have features pertaining to every field: media, video, gaming, science, e-commerce and so on. In the process, they gain profits by selling these apps on app stores or integrating multiple libraries of ads in their free versions to facilitate income to them.

Many users prefer using free versions of aps rather than buying them and therefore, do not mind when ads pop up on their screens. However, sighting certain ads could lead to the app getting low user ratings on the store [14].

Therefore, the third objective of this project is:

RO 3: Predict from the number of downloads, whether an app is paid or free.

A brief literature review of related work done in the above three domains is presented under Related Work, followed by a description of the methodology used in the project work is given under the Data Mining Methodology section. Thereafter, a brief evaluation of the models used to meet the objectives are under Evaluation. Furthermore, the section of Conclusions and future work describes conclusions with respect to the objectives and possible future work related to the project. Finally, references to future work, datasets and supporting material is presented in the References section.

II. RELATED WORK

A. *Literature review related to the data set about landslides [21] is presented here.*

1. Many landslides are set off due to rainfall. They are predicted with predefined intensity-duration (ID) limits in those regions. Usually, precipitation incidents are plotted with occurrence or non-occurrence of landslides. However, it this approach often suffers from false positives as the events of landslides could be due to both meteorological as well as hydrological reasons. Thus, Boggard and Grecco [1] have studied the ID limits for landslides and from a comprehensive view by considering both the aforementioned reasons. Further, propose a theoretical for landslides. In their methodology, they have compared the ID thresholds and IDF (intensity-duration-frequency) to see which one is better at predicting landslides [1].

2. Currently, most of the mudslide vulnerability evaluation do not assess the exact points of landslide and potentially dangerous regions. [2] have proposed a site-based assessment apparatus uses vector calculations to estimate the risk of landslides in a particular region. Considering the area of study – Schenzhen, variables such as the slope, the probable distance of the slide and the height of the slope when plot together form a semi-ellipse. When the above process is applied, precise information of the landslide and its effect can be derived.

3. Using multi-mode data for detecting distortion in the shape and size of an area with a probable landslide is a complex task because there is constant variation in the landscapes, water-content of the soil and anthropogenic activities. Thus, an SVR-based prediction model is presented by Miao et al. [3]. They have also verified better performance of their model as opposed to that done typically.

4. Lian et al. [4] has used a probabilistic approach to predict landslides. Their model gains knowledge of static and volatile points with k-means. ELM is then applied to obtain predictions.
   However, if there is misclassification, the probabilistic outcome could lead to false positives.

5. A model switching strategy which can select different predictors is proposed here [5]. It can estimate landslides with the trends in predictor variables.

6. The objective of this paper is to build an early warn ing system using a rain gauge and an intelligent acc elerometer to detect potential landslide events. The system measures precipitation and accelerometer for angular movement and earth movement for surveillance and estimation [6].

7. This paper has integrated seismic activity alerting system into Android systems since smartphones have a low-cost apparatus to detect movements. In spite of being low on preciseness, the dispersion of smartphones over large areas cover more ground. Using the data from the crowd, may help reduce false positives triggered by usual sensors [7].

B. *The following are works related to the dataset that contains data about Air pollution [23].*

1. Aggregated the past 4 years data of PM 2.5 from Beijing's observing locations to gain an overall understanding of the air quality. They were able to statistically prove that there had been a considerable amount of fall in the PM2.5 levels in the air [8].

2. Being exposed to smoke may lead to prolonged respiratory diseases by damaging the inner air passage; the world's almost half of the inhabitants are exposed to it [9].

3. Body cells were exposed to PM to examine and study the kind of the biological changes that occur due to it. It was observed that it induces stress and cell division as well as inflammation was sited [10].

4. A new method of reducing air pollution levels is put forward in this paper. They have proposed the use of distilled water for purification of air. Air is passed through the water and the pollutants are stuck in the water, yielding pure air [11].

5. Put forth a way of envisioning air pollution by further adding more variables, sources of air pollution can be realized. It can be implemented on a sensing network [12].

6. Studied the reduction of air pollution in a starch factory. The effects they studied were: tapioca yielded better, drying of starch process improved. Hot air initiators performed better [13].

C. *Works pertaining to the data set containing App Types are introduced here [24].*

1. Sarro et al. [14] put forth a study that the features an provides, may be used to predict the user ratings it will draw. Their study was based on the existing ratings and features information on the App Store and yield highly accurate predictions. They used the CBR (Case Based Reasoning) with the ANGEL tool

and implied the Euclidian Distance to calculate app similarity and the average rating of *k* nearest analogies. They used app data of apps present on both Samsung Android and Blackberry App Store at the time of the study and thus cannot be generalized to all app stores.

2. Developers often integrate ads into free apps to derive money from it. Various advertising companies provides libraries to developers for including ads in them. However, ads not necessarily received with every request, sometimes the demand exceeds supply and no ads are retrieved. Therefore, many developers incorporate more than one ad libraries in their apps and yet not study has examined if the app's ratings are affected by including several ad libraries. The authors studied numerous Android apps and found that there was no effect of the presence of several ad libraries on the overall ratings of the app. Ruiz et al. [15] limited their study to ad-supported version of apps. Nevertheless, it was found that there is some correlation between certain kinds of app libraries and user ratings, if the users do not like the ads, they rate the apps down.
3. When developers release their apps and users demand for either fixing of existing features or addition of new features, it is cumbersome to elicit the demands of users from a large number of requests. They realized that users either provided information that was not actionable at all or highlighted just the exact feature that required to be included in the app [16].
4. Algorithms of rankings of apps are studied. Further they have studied the relationship of app reputation and number of installations it achieves and whether it is profitable for attackers to exploit this information [17].
5. A prediction model that considers various aspects that influences the use of apps. The factors being: user preferences and environment and user actions [18].
6. False rankings on app stores enable apps to quickly escalate in the popularity list. Three ways of preventing fraudulent ratings of apps are proposed: authorizations established on ratings, position established ratings and review-built scores [19].
7. Apps are often deployed across multiple platforms on smartphones. This is a common practice to achieve a larger user base. A study of how users recognize the same app on different platforms a comparison-based research is presented here. They found many differences with respect to user ratings, star ratings and so on [20].

III. DATA MINING METHODOLOGY

The KDD methodology of discovering knowledge was used throughout the project to meet the set objectives. An in-depth report of the entire process is present here in this section. The datasets contained a lot of columns which were not in the scope of the objectives of this project. Hence, they were removed. [25] was referred for the entire KDD process.

A. Data Cleaning and Processing

All datasets had a lot of NaN, NA and missing data. Therefore, they were dropped in order to have all complete datasets. Thereafter, a good quantity of data processing was done which is described in the presentation submitted along with this report. The values of data sets are shown as follows:

1. *Data set for Landslides*

   The data set contained several columns which were not directly related to the objectives. Details are shown in Fig. 1.

```
> colnames(slides)
 [1] "i..source_name"          "source_link"           "event_id"              "event_date"
 [5] "event_time"              "event_title"           "event_description"     "location_description"
 [9] "location_accuracy"       "landslide_category"    "landslide_trigger"     "landslide_size"
[13] "landslide_setting"       "fatality_count"        "injury_count"          "storm_name"
[17] "photo_link"              "notes"                 "event_import_source"   "event_import_id"
[21] "country_name"            "country_code"          "admin_division_name"   "admin_division_population"
[25] "gazeteer_closest_point"  "gazeteer_distance"     "submitted_date"        "created_date"
[29] "last_edited_date"        "longitude"             "latitude"
```

Fig. 1 – Landslides

2. *Data set for Air Pollution*

   This data set is as shown in Fig. 2.

```
> colnames(df)
 [1] "No"    "year"    "month"   "day"     "hour"    "PM2.5"   "PM10"    "SO2"     "NO2"     "CO"      "O3"
[12] "TEMP"  "PRES"    "DEWP"    "RAIN"    "wd"      "WSPM"    "station"
>
```

Fig. 2 – Air pollution

3. *Dataset for App Types*

   The data set is Illustrated in Fig. 3.

Fig. 3 – App Types

```
> colnames(df)
 [1] "App"          "Category"      "Rating"        "Reviews"       "Size"          "Installs"
 [7] "Type"         "Price"         "Content.Rating" "Genres"       "Last.Updated"  "Current.Ver"
[13] "Android.Ver"
>
```

B. Data Integration

This step was not necessary as columns from additional datasets were not required to meet the objectives.

C. Data Selection

1. *Data set for Landslides*

   Three columns: one dependent: fatality count and two predictor variables: *location_accuracy* and *landslide_size* were selected. Please refer Fig. 4.

```
> summary(df)
slides.location_accuracy slides.Landslide_size slides.fatality_count
5      :3174             2:  3                 Min.   :  0.000
1      :2182             3: 750                1st Qu.:  0.000
0      :1928             4:6551                Median :  0.000
25     :1470             5:3618                Mean   :  2.013
10     :1434             7: 102                3rd Qu.:  0.000
50     : 793                                   Max.   :491.000
(Other):  43
```

Fig. 4 – Landslides

2. *Data set for Air Pollution*

   In this data set, two columns: one dependent: PM2.5 and one independent variable: NO2 was used. Shown in Fig. 5.

```
> summary(df2)
    df.PM2.5          df.NO2
 Min.   :  0.00   Min.   :  0.00
 1st Qu.: 20.00   1st Qu.: 28.00
 Median : 56.00   Median : 52.00
 Mean   : 80.59   Mean   : 57.58
 3rd Qu.:112.00   3rd Qu.: 81.00
 Max.   :898.00   Max.   :290.00
> |
```

Fig. 5 – Air Pollution

3. *Data set for App Types*

   Here again, two columns: one dependent: PM2.5 and one independent variable: NO2 was used. Shown in Fig. 6.

```
> summary(inst)
      df.Installs   df.Type
 1,000,000+ :1579   0   :    1
 10,000,000+:1252   Free:10039
 100,000+   :1169   NaN :    1
 10,000+    :1054   Paid:  800
 1,000+     : 907
 5,000,000+ : 752
 (Other)    :4128
> |
```

Fig. 6 – App Types

D. Data Transformation

All those columns that were factors were converted into continuous numeric (integer) or character form before splitting into train and test data.

E. Data Mining

The data was split into two parts one for training and the other for testing. It was split in the ratio of 70: 30. All models were first applied on the training dataset and subsequently on the testing dataset to check how much the machine had learned. The machine gained knowledge from the training dataset and applied it in real time to the testing dataset to arrive on conclusions.

1. *Random Forest*

   Application of tuneRF() for selecting the dependent and predictor variables in the dataset. These will be used to apply the model to extract knowledge. Please refer Fig. 7.

```
#Splitting the data
set.seed(2)
id <- sample(2, nrow(df), prob = c(0.7,0.3), replace = TRUE)

train <- df[id ==1,]
test <- df[id==2,]

head(df)
summary(df)


library(randomForest)


bestmtry <- tuneRF(train, train$slides.fatality_count, stepFactor = 1.2, improve = 0.01, trace = T, plot = T)

land_slides <- randomForest(slides.fatality_count ~., data = df )
land_slides

importance(land_slides)

varImpPlot(land_slides)
```

Fig. 7 – Model on the dataset for Landslides

2. *Multiple Regression*

   The lm() function is used to define and feed the dependent variable and predictor variables in the dataset to the multiple regression model. Please see Fig. 8.

```
set.seed(2)
library(caTools)
split <- sample.split(df, SplitRatio = 0.7)
split
train <- subset(df, split ="TRUE")
test <- subset(df, split = "FALSE")
train

#create the model
Model <- lm(slides.fatality_count ~., data =train )
summary(Model)

#Prediction
pred <- predict(Model, test)
pred
```

Fig. 8 – Model on dataset for Landslides

3. *Simple Regression*

   The lm() function is used for performing simple regression as well. With the predictor and dependent variables set, the model is run to train and the machine. Shown in Fig. 9.

```
#Splitting the data
set.seed(123)
library(caTools)
split <- sample.split(df2$df.PM2.5, SplitRatio = 2/3)
split

train <- subset(df2, split ="TRUE")
test <- subset(df2, split = "FALSE")
train

#creating the model
Model <- lm(formula = df.PM2.5 ~ df.NO2, data = train )
summary(Model)

#Prediction
pred <- predict(Model, test)
pred
```

Fig. 9 – Model on the dataset for Air Pollution

4. *Logistic Regression*

   The data of apps being free (1) or paid (0) was uneven and was thus down-sampled and

subsequently up-sampled to have equal data for training. The original binary data was as follows:

0	1
560	7028

Please refer to Fig. 10 and Fig. 11

```
######Logistic
table(inst$df.Type)

#Sampling
library(caret)
'%ni%' <- Negate('%in%')  # define 'not in' func
options(scipen=999)  # prevents printing scientific notation

#Training and Test data
set.seed(100)
trainDataIndex <- createDataPartition(inst$df.Type , p=0.7, list = F)  # 70% training data
trainData <- inst[trainDataIndex, ]
testData <- inst[-trainDataIndex, ]

table(trainData$df.Type)

#Need to downsample "Free" apps
set.seed(100)
down_train <- downSample(x = trainData[, colnames(trainData) %ni% "Type"],
                         y = trainData$df.Type)

summary(down_train)

table(down_train$df.Type)
```

Fig. 10 – Model on dataset for App Types

```
## Up Sample
set.seed(100)
up_train <- upSample(x = trainData[, colnames(trainData) %ni% "Type"],
                     y = trainData$df.Type)
table(up_train$df.Type)
up_train <- up_train[,-1]
summary(up_train)


# Build Logistic Model
logitmod <- glm(df.Type ~ df.Installs, family = "binomial", data=down_train)
summary(logitmod)


pred <- predict(logitmod, newdata = testData, type = "response")
```

Fig. 11 -Model on dataset for App Types

5. *Naïve Bayes*
   The naïveBayes() function was employed to specify the predictor and dependent variables before running the model. Please see Fig. 12.

```
set.seed(2)
id <- sample(2, nrow(inst), prob = c(0.7,0.3), replace = T)
inst_train <- inst[id == 1,]
inst_test <- inst[id==2, ]
library(e1071)
library(caret)


instb <- naiveBayes(df.Type ~., data = inst)
instb
```

Fig. 12 – Model on data set for App Types

F. Pattern Evaluation

The patterns obtained by running models on the datasets were evaluated using requisite measures such as RMSE, Confusion matrix and so on. Pattern evaluation is described at length later, in the Evaluation section.

G. Knowledge Presentation

The obtained knowledge is presented in the form of plots and confusion matrix where necessary at is discussed in detail in the Evaluation section.

IV. Evaluation

1. Random Forest
   The output of this model is as shown in Fig. 13.

```
Call:
 randomForest(formula = slides.fatality_count ~ ., data = df)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 432.5919
                    % Var explained: 8.04
```

Fig. 13

As shown in Fig. 13, the the var explained in random forest is 8.04. Thereafter, the importance of the variables is as shown in Fig. 14

```
> importance(land_slides)
                           IncNodePurity
slides.location_accuracy        91693.79
slides.Landslide_size          331658.21
```

Fig. 14 – Variable importance

Thus, it may be concluded that the size of the landslide is an important factor in predicting fatalities due to landslides. The mean squared error and the number of trees are plot to see that the error showed a drastic rise at first and then stayed roughly the same. Please refer to Fig. 15.

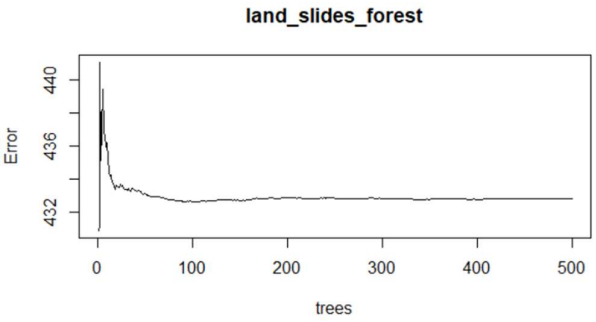

Fig. 15

2. Multiple Regression
   Fig. 16 shows the summary of the model. From the summary it can be seen that both the variables used to predict fatalities due to landslides are indeed statistically significant. However still, the model could not rightly predict the values as from Fig. 17, the blue lines showing the test values and the red ones showing the predicted values have a significant difference.

```
> summary(Model)

Call:
lm(formula = slides.fatality_count ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-26.266 -10.479  -8.536  -3.759 108.308

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              27.67000    1.41330  19.578  < 2e-16 ***
slides.location_accuracy  0.48586    0.08423   5.768 8.24e-09 ***
slides.Landslide_size    -4.77645    0.27625 -17.290  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.28 on 11030 degrees of freedom
Multiple R-squared:  0.03724,    Adjusted R-squared:  0.03706
F-statistic: 213.3 on 2 and 11030 DF,  p-value: < 2.2e-16
```

Fig. 16 – summary

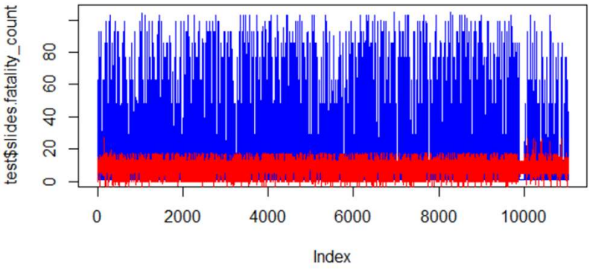

Fig. 17 – Test vs predicted values

The RMSE value was observed to be 1.096007 e-13

```
> rmse
[1] 1.096007e-13
```

Fig. 18

It can thus be concluded that Multiple Regression performed better than Random Forest on this dataset.

3. Simple Linear Regression
   The summary of this model is present in Fig. 19. The summary shows that NO2 concentrations are significant in predicting the PM 2.5 levels in the air. Furthermore, the residual standard error is low as compared to the number of observations considered. Thereafter, the R-squared statistic amounts to 45%, thereby showing a good performance.

```
> summary(Model)

Call:
lm(formula = df.PM2.5 ~ df.NO2, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-285.13  -35.16   -8.40   22.01  767.27

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.272337   0.590363  -5.543    3e-08 ***
df.NO2       1.456561   0.008564 170.077   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.79 on 35062 degrees of freedom
Multiple R-squared:  0.4521,    Adjusted R-squared:  0.452
F-statistic: 2.893e+04 on 1 and 35062 DF,  p-value: < 2.2e-16
```

Fig. 19 – Model summary

Fig. 20 shows the machine's performance on the test set along with the regression line.

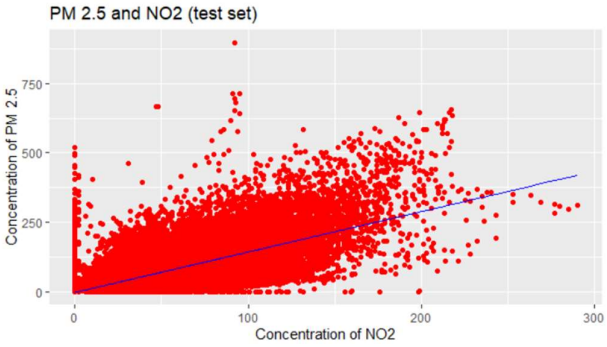

Fig. 20 – Performance on Test set

4. Logistic Regression
   The summary of the model is shown in Fig. 21. The null deviance is observed to be larger than the residual deviance. Thus, it may be said that the model has worked in a good way after including the independent variable that marks the number of installations of an app. The confusion matrix is shown in Fig. 22.

```
> summary(logitmod)

Call:
glm(formula = df.Type ~ df.Installs, family = "binomial", data = down_train)

Deviance Residuals:
    Min      1Q  Median      3Q     Max
-6.4235  -0.9160  -0.4576   1.0087   1.4645

Coefficients:
              Estimate   Std. Error z value             Pr(>|z|)
(Intercept) -0.6535693880  0.0750085807  -8.713 <0.0000000000000002 ***
df.Installs  0.0000021284  0.0000002494   8.536 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1552.6  on 1119  degrees of freedom
Residual deviance: 1204.8  on 1118  degrees of freedom
AIC: 1208.8

Number of Fisher Scoring iterations: 12
```

Fig. 21 – Model summary

```
            Predictedvalue
Actualvalue FALSE TRUE
          0   225   15
          1  1526 1485
```

Fig. 22 – Confusion Matrix

The accuracy of the model thus turns out to be 0.525992.

5. Naïve Bayes
   The confusion matrix and model summary are presented in Fig. 23. The accuracy of Naïve Bayes is seen to be 0.2548 which is less than Logistic Regression and thus, this model has underperformed.

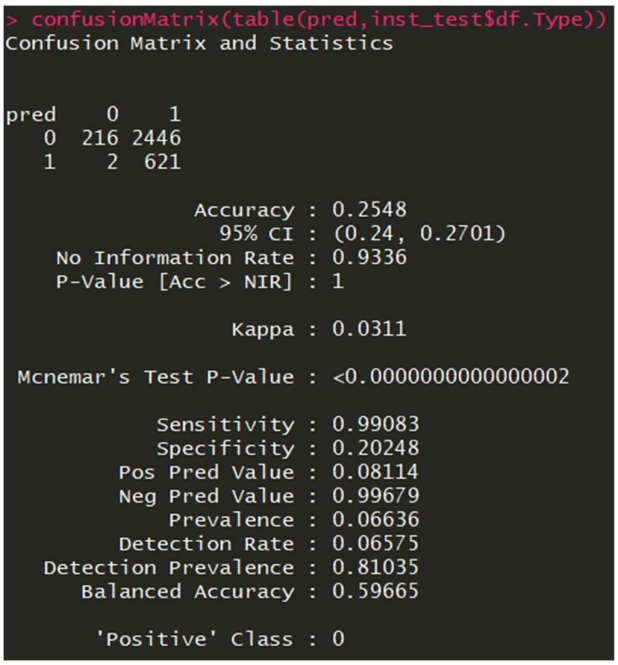

Fig. 23 – Model Summary

V. CONCLUSIONS AND FUTURE WORK

The implementation of five models is therefore successful. Comparing the outputs of all models, it can be concluded that some models perform better on some data. The research objectives were met with implementation of models for comparing their performance. In future, research objectives for some kind of betterment of the society could be implemented.

REFERENCES

[1] T. Bogaard and R. Greco, "Invited perspectives: Hydrological perspectives on precipitation intensity-duration thresholds for landslide initiation: proposing hydro-meteorological thresholds," *Natural Hazards and Earth System Sciences*, vol. 18, no. 1, pp. 31–39, Apr. 2018.

[2] T. Li, Y. Tian, C. Xiao, and W. Zhao, "Slope location-based landslide vulnerability assessment," *2013 21st International Conference on Geoinformatics*, 2013.

[3] S. Miao, Q. Zhu, L. Zhang, C. Liu, B. Zhang, and M. Chen, "A knowledge-guided landslide deformation prediction approach based on SVR," *2017 25th International Conference on Geoinformatics*, 2017.

[4] C. Lian, L. Zhu, Z. Zeng, Y. Su, W. Yao, and H. Tang, "Constructing prediction intervals for landslide displacement using bootstrapping random vector functional link networks selective ensemble with neural networks switched," *Neurocomputing*, vol. 291, pp. 1–10, 2018.

[5] S.-F. Chen and P.-A. Hsiung, "Landslide prediction with model switching," *2017 IEEE Conference on Dependable and Secure Computing*, 2017.

[6] C. D. Fernandez, K. J. A. Mendoza, A. J. S. Tiongson, and M. B. Mendoza, "Development of microcontroller-based landslide early warning system," *2016 IEEE Region 10 Conference (TENCON)*, 2016.

[7] A. Heryana, E. Nugraheni, B. Kusumo, A. F. Rojie, and B. Setiadi, "Applying agile methods in designing an earthquake and landslide early warning system application for Android," *2017 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, 2017.

[8] S. Zhang, B. Guo, A. Dong, J. He, Z. Xu, and S. X. Chen, "Cautionary tales on air-quality improvement in Beijing," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 473, no. 2205, p. 20170457, 2017.

[9] A. Ramírez-Venegas, C. A. Torres-Duque, N. E. Guzmán-Bouilloud, M. González-García, and R. Sansores, "Small Airway Disease in COPD Associated to Biomass Exposure," *Revista de investigaci□n Cl□nica*, vol. 71, no. 1, Apr. 2019.

[10] B. F. Cachon, S. Firmin, A. Verdin, L. Ayi-Fanou, S. Billet, F. Cazier, P. J. Martin, F. Aissi, D. Courcot, A. Sanni, and P. Shirali, "Proinflammatory effects and oxidative stress within human bronchial epithelial cells exposed to atmospheric particulate matter (PM2.5 and PM>2.5) collected from Cotonou, Benin," *Environmental Pollution*, vol. 185, pp. 340–351, 2014.

[11] B. R. Subrahmanyam, A. G. Singh, and D. P. Tiwari, "Air Purification System for Street Level Air Pollution and Roadside Air Pollution," *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*, 2018.

[12] M. Korunoski, B. R. Stojkoska, and K. Trivodaliev, "Internet of Things Solution for Intelligent Air Pollution Prediction and Visualization," *IEEE EUROCON 2019 - 18th International Conference on Smart Technologies*, 2019.

[13] S. Karuchit and P. Sukkasem, "Application of AERMOD Model with Clean Technology Principles for Industrial Air Pollution Reduction," *2018 Third International Conference on Engineering Science and Innovative Technology (ESIT)*, 2018

[14] I. J. M. Ruiz, M. Nagappan, B. Adams, T. Berger, S. Dienst, and A. E. Hassan, "Impact of Ad Libraries on Ratings of Android Mobile Apps," *IEEE Software*, vol. 31, no. 6, pp. 86–92, 2014.

[15] F. Sarro, M. Harman, Y. Jia, and Y. Zhang, "Customer Rating Reactions Can Be Predicted Purely using App Features," *2018 IEEE 26th International Requirements Engineering Conference (RE)*, 2018.

[16] S. A. Licorish, B. T. R. Savarimuthu, and S. Keertipati, "Attributes that Predict which Features to Fix," *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering - EASE17*, 2017.

[17] Y. Liu and Y. Sun, "Can reputation manipulation boost app sales in Android market?," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[18] Y. Xu, M. Lin, H. Lu, G. Cardone, N. Lane, Z. Chen, A. Campbell, and T. Choudhury, "Preference, context and communities," *Proceedings of the 17th annual*

international symposium on International symposium on wearable computers - ISWC 13, 2013.

[19] K. Manoj, T. Sandeep, N. S. Reddy, and P. Alikhan, "Genuine ratings for mobile apps with the support of authenticated users' reviews," *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*, 2018.

[20] M. Ali, M. E. Joorabchi, and A. Mesbah, "Same App, Different App Stores: A Comparative Study," *2017 IEEE/ACM 4th International Conference on Mobile Software Engineering and Systems (MOBILESoft)*, 2017.

[21] Data.nasa.gov,2019.[Online]. Available:https://data.nasa.gov/Earth-Science/Global-Landslide-Catalog-Export/dd9e-wu2v/data.[Accessed:09-Nov- 2019]

[22] Kirschbaum, D. B., Adler, R., Hong, Y.,Hill, S., & Lerner-Lam, A. (2010). A global landslide catalog for hazard applications: method, results, and limitations. Natural Hazards, 52(3),561–575. doi:10.1007/s11069-009-9401-4.

[23] "UCI Machine Learning Repository: Beijing Multi-Site Air-Quality Data Data Set", Archive.ics.uci.edu, 2019. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi-Site+Air-Quality+Data. [Accessed: 09- Nov- 2019]

[24] "Google Play Store Apps", Kaggle.com, 2019. [Online]. Available: https://www.kaggle.com/lava18/google-play-store-apps/. [Accessed: 09- Nov-2019]

[25] J. Han, M. Kamber, and J. Pei, "Advanced Pattern Mining," *Data Mining*, pp. 279–325, 2012.