Early prediction of outcast people in the Afro-American community using machine learning technique

Animesh Samanta[1], Akash Chowdhury[2], Argha Ghosh[3], Sulekha Das[4], Dr. Avijit Kumar Chaudhuri[5]

[1] UG-Computer Science and Engineering, Techno Engineering College Banipur,
[2] UG-Computer Science and Engineering, Techno Engineering College Banipur,
[3] UG-Computer Science and Engineering, Techno Engineering College Banipur,
[4] Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur,
[5] Assistant Professor, Computer Science and Engineering, Techno Engineering College Banipur,

[1]Corresponding Author Orcid ID: 0000-0003-3685-2646
[2]Corresponding Author Orcid ID: 0000-0002-9156-1553
[3]Corresponding Author Orcid ID: 0000-0002-1469-9882

ABSTRACT

The study of Homeless people is challenging, and describing Black Homelessness is especially difficult because of spotty record keeping[**Error! Reference source not found.**]. The story of homeless Afro-Americans is often absent from the usual textbook study of homeless people. Homelessness is incompatibly experienced among historically marginalized groups in the united states, but for Black Americans, the distinction is especially stark[**Error! Reference source not found.**]. Homeless condition is often occurred to the condition of poverty, less number of house, and less number of beds in the room. In most homes, there are fewer than three bedrooms and the income of each family is less, most of the rooms are already occupied. Some people in the Afro-American community are living their lives as renters what are the rent cost including the utilities like electricity, gas, water, etc is also very high. The communication system of the caucasian population is bad also, there is no phone in the family. The main reason for homelessness is that a huge number of people are immigrating within the last 5 years. The economic condition of Afro-Americans is at the poverty level, with few services, and often very poor educational opportunities. Here the data was analyzed using Multiple Regression Analysis (MRA). The proposed model is tested on the "Communities and Crime Data Set" from the UCI Machine Learning Repository: which is available at https://archive.ics.uci.edu/ml/datasets/communities+and+crime. Through this research, approximately 99.49 percent of the data was predicted correctly.

Keywords— Homeless, Poverty, Afro-American community, Machine Learning, Crime Data-set, Multiple Linear Regression, Feature Selection, Confusion Matrix, Cross-Validation.

1. Introduction

In our present society, there are two types of people who are homeless and another type of people who have homes. Many people are still homeless in this present society and it is a bad sign for society. It may be that their economic condition is very bad. Based on this survey it will be known how many people are still homeless. Using this survey it will be also known how many people occupied less than 3 bedroom houses and the total income of those households from all sources. Some people in the Afro-American community are living their lives as renters what are the rent cost including the utilities like electricity, gas, water. This research work will also provide information about the income of the renters and how many people whose ages are 16 and above are working as an employee in a management or any professional organization. We also will know what is the salary income in present and also in 1989. There are some places where crime increased frequently but there were no police officers still now. So we can know from this paper how many total requests were there for the requirement of police officer per poofficersfice. Also, this paper will clarify the Caucasian race in the afro-American population which means how many people fall under the Caucasian race. And the last thing that will be also known by this paper is the number of immigrants who immigrated within the last 5 years.

Machine learning can simply be defined as using data instead of logic to perform tasks by a machine. We use data to train the machine, as in, tell it what it has to do and then test the trained model on different tasks to see whether the training has been successful or not. When it comes to data mining, the term classification plays an important role as it assigns class values to new instances found during data mining.

Multiple regression is an applied math technique that will be accustomed analyze the link between one dependent variable and several other independent variables. The target of multiple regression analysis is to use the independent variables whose worth square measure is acknowledged to predict the worth of the only dependent value.

Cross-validation (CV) is a popular technique for selecting the algorithm. The main concept behind Cross-validation is to split the whole data set once or several times, for calculating the risk of each algorithm. Part of the data (the training sample) is used for training each algorithm, and the remaining part (the validation sample) is used for estimating the risk of the algorithm. Then, CV selects the algorithm with the smallest estimated risk. Besides the common case, where a general measure of the reliability and accuracy of a system is needed, evaluation often becomes necessary to choose the best one out of different methods and/or parameter sets.

Statistical analysis is widely used in all aspects such as in science, medicine, crime, and English literature, employed in professional sources, and also in social sciences. There are many methods in statistics and one of them is regression. There are six types of linear regression analyses which are simple linear regression, multiple linear regression, logistic regression, ordinal regression, multinomial regression, and discriminate analysis. Multiple linear regression was selected to build a model of early prediction of outcast people in the Afro-American community using machine learning techniques. One method that is categorized in the stepwise-type process is stepwise regression also used in this paper. The main purpose of this paper is to select the suitably controlled variables in the forecast for the Early prediction of outcast people in the Afro-American community using machine learning techniques.

Feature selection is the process of selecting a subset of relevant features (also called predictors or independent variables) from a larger set of features that are available in a dataset. The goal of feature selection is to improve the performance of a machine learning model by reducing the dimensionality of the feature space, which can help to mitigate the risk of overfitting and improve the interpretability of the model. There are several methods for feature selection, including:

Filter methods: These methods use statistical tests to rank the importance of each feature and select the top features based on their ranking. Examples of filter methods include correlation-based feature selection and chi-squared feature selection.

Wrapper methods: These methods use the performance of a machine learning model as a criterion for selecting the best subset of features. Wrapper methods evaluate all possible subsets of features by training and testing the model on each subset. Examples of wrapper methods include recursive feature elimination and forward selection.

Embedded methods: These methods perform feature selection as part of the model training process. Examples of embedded methods include Lasso regression and decision tree-based methods such as Random Forest and Gradient Boosting.
It is important to note that feature selection should be used with caution and in conjunction with other data preprocessing techniques. The choice of feature selection method and the number of selected features should be based on the specific problem and the nature of the data.

2. Literature review

Multiple regression is an applied math technique, the target of multiple regression analysis is to use the independent variables whose worths square measure acknowledged to predict the worth of the only dependent value. The main aim of this project is prediction of outcast people in the Afro-American community using machine learning technique. Mr. M. S. BARTLETT had done on "FURTHER ASPECTS OF THE THEORY OF MULTIPLE REGRESSION".
Intan Martina Md Ghani,, Sabri Ahmad (2010) had done a research to Forecast Fish Landing.
Isık Yilma and ,Oguz Kaynar (2011) had done research prediction of the swell potential of clayey soils using multiple linear regression.
We have gathered some specific ideas about machine learning and multiple linear regression. So we were interested very much to do a project based on it. So we collected some real-life data on the "Communities and Crime Data Set" from the UCI, which is available at https://archive.ics.uci.edu/ml/datasets/communities+and+crime and try to prediction of outcast people in the Afro-American community using machine learning technique.

3. Experimental Methods or Methodology

3.1 Data

In this pape,r data were taken from UCI Machine Learning Repository. Here we work in the following field…

Table.1. Data attribute description

Attribute	Description	Mean	Standard Deviation
MedRentPctHousInc	median gross rent as a percentage of household income	0.449	0.187
PctOccupMgmtProf	percentage of people 16 and over who are employed in management or professional occupations	0.441	0.186
PctHousOccup	percent of housing occupied	0.719	0.193
PolicReqPerOffic	total requests for police per police officer	0.054	0.148
racePctWhite	percentage of the population that is caucasian	0.075	0.243
PctImmigRec5	percentage of _immigrants_ who immigrated withinthe last 5 years	0.360	0.210
MedRent	median gross rent	0.384	0.213
PctHousLess3BR	percent of housing units with less than 3 bedrooms	0.495	0.172
pctWWage	percentage of households with wage or salary income in 1989	0.558	0.182

PctHousIncPovLevel	Household income in the poverty level	0.490	0.169
perCapInc	per capita income	0.350	0.191

3.2 Research Method

Multiple simple regression is the strategy of statistics in regression that's familiar to analyzing the link between one response variable (dependent variable) with 2 or additional controlled variables (independent variables). This methodology was selected for this analysis as a result of there have been quite controlled variables.

During this analysis, the response variable is Homeless people (Y), Household income as renters (X_1), Employees in Professional occupations or management (X_2), Occupied houses(X_3) Requests for police officers (X_4) Population of Caucasian (X5) Immigrated people within last 5 years (X6), rent includes utilities (X7), less than room(X8), Household salary income (X9), Household income in poverty level (X10), Average income of per person (X11), Sworn full-time officers (X12), Owner occupied house lower quartile value (X13), immigrated people within last 8 years (X14), Household income for owners without a mortgage (X15), House occupied people without pa hone (X16), Large family household (X17), Hispanic heritage population (X18), People immigrated recently (X19), People in owner-occupied households (X20) are controlled variables in the primary case.

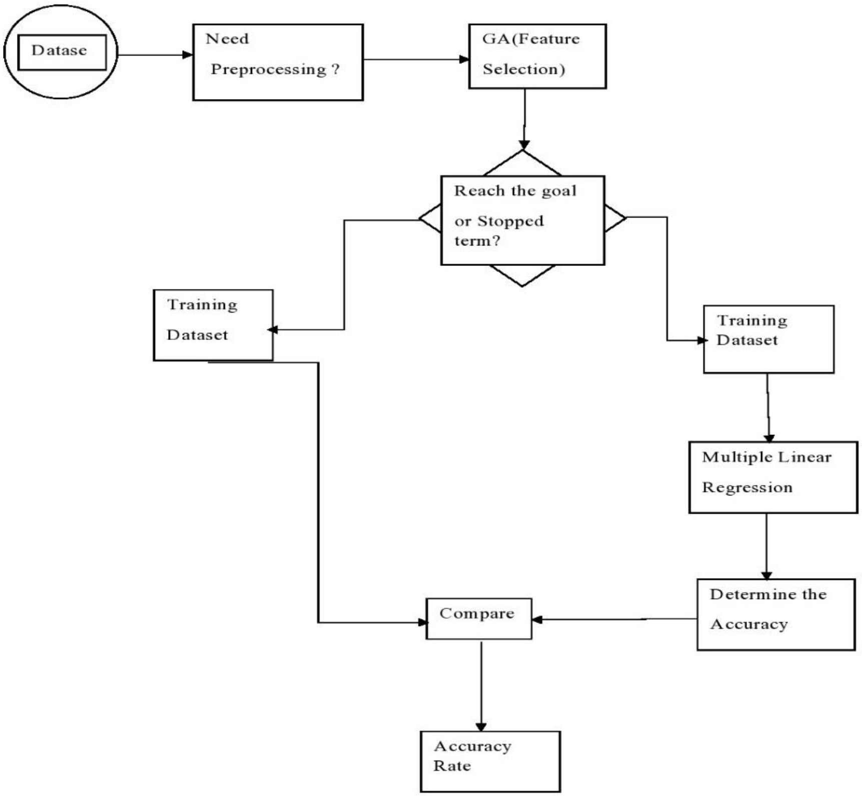


Fig.1. Feature Selection Model

3.3 Feature Selection

In machine learning, feature selection is the method of selecting an important feature from the data set so that the output of the model can be appropriate. The first goal of the feature selection method is to reach at least the same accuracy rate as the whole features provided. The second goal of the feature selection is having improve the accuracy rate. In many times it happens that there are many features in the data set which is not required and those features reduce the performance of the model. Here, not only is the collection of extensive information on features very costly in terms of both time and money but the additional information results in wasted classification and diagnosis time. It is better to reduce the dimension in terms of the number of features to get a better response and to find a better relationship between features and results.

Then we have applied different features to get optimal accuracy. Here we have reduced the independent field to get better accuracy, In between the accuracy was fluctuated frequently. So after optimizing, the following fields have given the optimal accuracy, finally the, controlled variables are: Household income as renters (X1), Employee in Professional occupations or management (X2), Occupied houses(X3), Request for police officer (X4), Population of Caucasian (X5), Immigrated people within last 5 years (X6), Rent includes utilities (X7), House less than 3 bedroom (X8), Household salary income (X9), Household income in poverty level (X10), Average income of per person (X11).

3.4 Accuracy of difference between Actual data and Calculated data

In this research, the hypotheses used:

H_0: $b_1=b_2=b_3=b_4=0$

H_a: At least one of the b_1, b_2, b_3, and b_4 does not equal to 0 which says that

H_0: None of the controlled variables X_1, X_2, X_3, and X_4 is significantly related to Y, H_a: At least one of the controlled variables X_1, X_2, X_3, and X_4 is significantly related to Y The model of multiple regression can be represented as:

$$\mathbf{Y = a + b_1X_1 + b_2X_2 + \ldots\ldots\ldots\ldots\ldots\ldots + b_nX_n}$$

And the a and b formulas are:

$$b_i = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$a = \frac{(\sum Y \times \sum X^2) - (\sum X \times \sum XY)}{(n \times \sum X^2) - (\sum X \times \sum X)}$$

Fig.2. Formula of constant(a) and coefficient(b) of the control variable

Where

y= Dependent variable(Homeless people)

a= Constant variable

b_i=Coefficient of the X_i control variable,

x_1=controlled variable (Household income as renters)

x_2=controlled variable (Employee in Professional occupations or management)

x_3=controlled variable (Occupied houses)

x_4=controlled variable (Request for police officer)

x_5=controlled variable (Population of Caucasian)

x_6=controlled variable (Immigrated people within last 5 years)

x_7=controlled variable (Rent includes utilities)

x_8=controlled variable (House less than 3 bedrooms)

x_9=controlled variable (Household salary income)

x_{10}=controlled variable (Household income in poverty level)

x_{11}=controlled variable (Average income per person)

4 Confusion-Matrix

After finding the accuracy of the difference between actual data and calculated data we did the Confusion Matrix. In this confusion matrix it can be seen that,[2] we find the **TP** – which stands for '**TRUE POSITIVE'** means the accuracy of classified positive data, **TN** – which stands for '**TRUE NEGATIVE'**_means the accuracy of classified negative data, **FP** – which stands for '**FALSE POSITIVE**', means which remark that actual value is negative but predicted data is positive, **FN** – which stands for '**FALSE NEGATIVE'**_means that actual data and the predicted data both are negative and append the TP, TN, FP, FN value in 2*2 matrix(mat1). After that, we find the accuracy, sensitivity, precision, and specificity. This matrix contains all the raw information about the predictions done by a classification model on a given data set.[3]

Accuracy- It's the ratio of the correctly labeled subjects to the whole pool of subjects. Accuracy is intuitional

Precision- Precision is the ratio of the correctly +ve labeled by our program to all +ve labeled.

Sensitivity- Sensitivity means out of the total positive, what percentage are predicted positive.

Specificity- Specificity is calculated as the number of correct negative predictions divided by the total number of negatives.

- **ACCURACY = $(TP+TN/ \sum K)$* 100**
- **PRECISION = (TP/FP+TP)*100**
- **SENSITIVITY = (TP/FN+TP)*10**
- **SPECIFICITY = $(TN/TN+FP)$* 100**

5 Cross-Validation

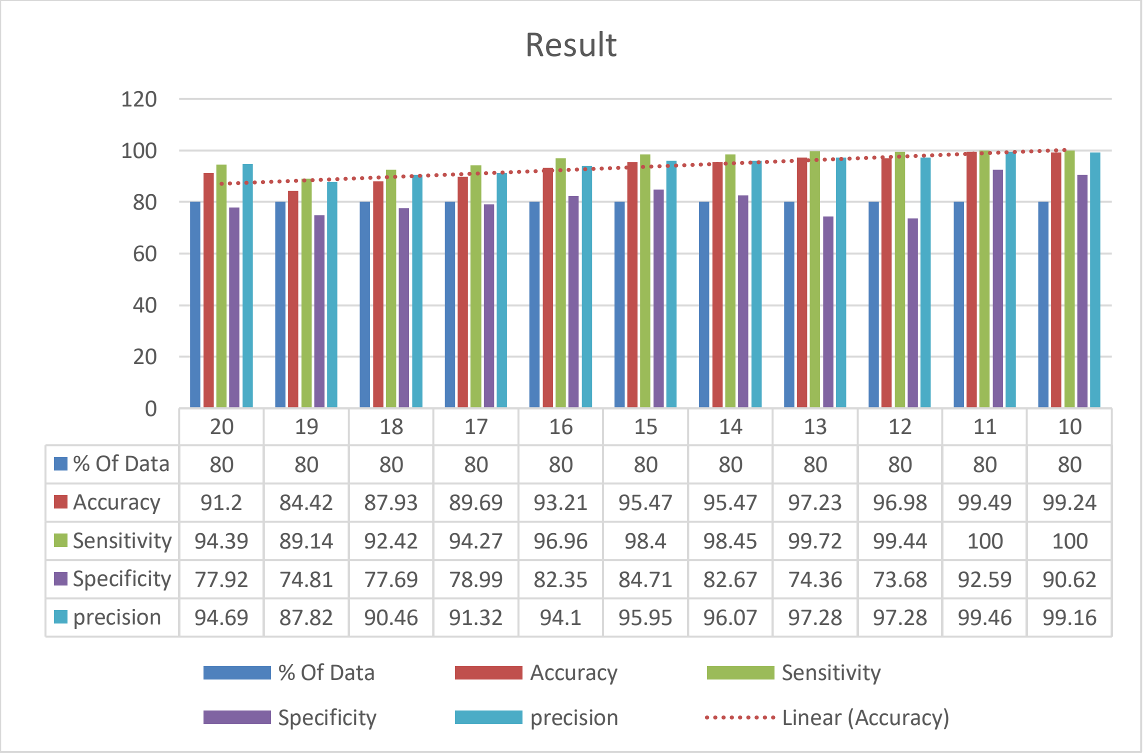
After finding the accuracy of the difference between actual data and calculated data we did cross-validation. In this cross-validation process first, we divide the whole list into 10 sub-list and then we find the accuracy of 10 sub-list elements we also find the Confusion Matrix of each Sub-list and we find the accuracy, sensitivity, precision, and specificity.

6 Results

Table.2. Result of the particular number of the data fields

No Of Data-field	% Of Data	Accuracy	Sensitivity	Specificity	Precision
20	80	91.20	94.39	77.92	94.69
	66	87.00	92.38	70.30	90.61
	50	84.93	90.88	68.2	88.95
19	80	84.42	89.14	74.81	87.82
	66	79.17	84.58	70.61	82.01
	50	75.10	82.21	64.5	77.53
18	80	87.93	92.42	77.69	90.46
	66	83.01	88.29	74.0	85.29
	50	79.01	85.92	68.05	81.02
17	80	89.69	94.27	78.99	91.32
	66	85.37	91.12	75.5	86.47
	50	81.82	89.26	70.33	82.32
16	80	93.21	96.96	82.35	94.1

	66	89.21	95.18	76.92	89.48
	50	87.04	93.91	74.72	86.98
15	80	95.47	98.40	84.71	95.95
	66	92.46	96.65	82.41	92.96
	50	90.36	95.67	79.51	90.52
14	80	95.47	98.45	82.67	96.07
	66	93.20	96.96	83.15	93.91
	50	90.76	96.08	78.83	91.06
13	80	97.23	99.72	74.36	97.28
	66	96.60	98.78	84.76	97.25
	50	95.48	98.13	84.38	96.38
12	80	96.98	99.44	73.68	97.28
	66	96.60	98.77	85.32	97.23
	50	95.38	98.13	84.1	96.21
11	80	99.49	100	92.59	99.46
	66	98.37	99.66	90.43	98.47
	50	97.38	99.02	90.0	97.82
10	80	99.24	100.0	90.62	99.16
	66	97.78	99.29	90.27	98.07
	50	96.88	98.98	89.2	97.12



	20	19	18	17	16	15	14	13	12	11	10
% Of Data	80	80	80	80	80	80	80	80	80	80	80
Accuracy	91.2	84.42	87.93	89.69	93.21	95.47	95.47	97.23	96.98	99.49	99.24
Sensitivity	94.39	89.14	92.42	94.27	96.96	98.4	98.45	99.72	99.44	100	100
Specificity	77.92	74.81	77.69	78.99	82.35	84.71	82.67	74.36	73.68	92.59	90.62
precision	94.69	87.82	90.46	91.32	94.1	95.95	96.07	97.28	97.28	99.46	99.16

Fig. 3. Graph of Accuracy, Sensitivity, Specificity, Precision

CONCLUSIONS

This paper uses multiple regressions (MLR) to predict the outcast people in the Afro-American community. Data has been collected from UCI Machine Learning Repository based on that we made a relationship between the dependent variable and the independent variable after that we perform Confusion Matrix where we compare the actual target values with those predicted by the machine learning model. After checking the Confusion Matrix, we move to the Cross Validation where we find the accuracy of 10 sub-list elements and we also find the Confusion Matrix of each Sub-list. we predict the accuracy as well as sensitivity, precision, and specificity for user choice test data and the 10 sub-list. After that, we have optimized the independent data field to get the optimal accuracy by selected features. This type of project may help in the future to find any kind of prediction from any data field.

Reference

1. Samanta, A., Chowdhury, A., Das, D., Dey, A. K., & Das, M. S. Prediction through machine learning on the dependence of job prospects in the Afro-American community on proficiency in English.

2. McClendon, L., & Meghanathan, N. (2015). Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ)*, *2*(1), 1-12.

3. Ajagbe, S. A., Idowu, I. R., Oladosu, J. B., & Adesina, A. O. (2020). Accuracy of machine learning models for mortality rate prediction in a crime dataset. *International Journal of Information Processing and Communication (IJIPC)*, *10*(1), 150-160.

4. Shermila, A. M., Bellarmine, A. B., & Santiago, N. (2018, May). Crime data analysis and prediction of perpetrator identity using machine learning approach. In *2018 2nd international conference on trends in electronics and informatics (ICOEI)* (pp. 107-114). IEEE.

5. Ghani, I. M. M., & Ahmad, S. (2010). Stepwise multiple regression method to forecast fish landing. *Procedia-Social and Behavioral Sciences*, *8*, 549-554.

6. Yilmaz, I., & Kaynar, O. (2011). Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. *Expert systems with applications*, *38*(5), 5958-5966.

7. Hong, B., Malik, A., Lundquist, J., Bellach, I., & Kontokosta, C. E. (2018). Applications of machine learning methods to predict readmission and length-of-stay for homeless families: The case of win shelters in new york city. *Journal of Technology in Human Services*, *36*(1), 89-104.

8. Chelmis, C., Qi, W., Lee, W., & Duncan, S. (2021). Smart homelessness service provision with machine learning. *Procedia Computer Science*, *185*, 9-18.

9. Visa, S., Ramsay, B., Ralescu, A. L., & Van Der Knaap, E. (2011). Confusion matrix-based feature selection. *Maics*, *710*(1), 120-127.

10. Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, *81*(3), 429-450.

11. Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, *81*(3), 429-450.

12. Caelen, O. (2017). A Bayesian interpretation of the confusion matrix. *Annals of Mathematics and Artificial Intelligence*, *81*(3), 429-450.

13. Zhang, D., Shen, D., & Alzheimer's Disease Neuroimaging Initiative. (2012). Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *NeuroImage*, *59*(2), 895-907.

14. Nusrath Unnisa A; Manjula Yerva; Kurian M Z. "Review on Intrusion Detection System (IDS) for Network Security using Machine Learning Algorithms". *International Research Journal on Advanced Science Hub*, 4, 03, 2022, 67-74. doi: 10.47392/irjash.2022.014

15. Chaudhuri, A. K., Banerjee, D. K., & Das, A. (2021). A Dataset Centric Feature Selection and Stacked Model to Detect Breast Cancer. *International Journal of Intelligent Systems and Applications (IJISA)*, *13*(4), 24-37.

16. Marom, N. D., Rokach, L., & Shmilovici, A. (2010, November). Using the confusion matrix for improving ensemble classifiers. In *2010 IEEE 26-th Convention of Electrical and Electronics Engineers in Israel* (pp. 000555-000559). IEEE.

17. Mahalakshmi G.; Shimaali Riyasudeen; Sairam R; Hari Sanjeevi R; Raghupathy B.. "A Survey: Effective Machine Learning Based Classification Algorithm for Medical Dataset". *International Research Journal on Advanced Science Hub*, 3, Special Issue 9S, 2021, 28-33. doi: 10.47392/irjash.2021.245

18. Toomula Srilatha; Jyothi Sree C.. "Survey on Plant Diseases Prediction using Machine learning for better Crop Yield". *International Research Journal on Advanced Science Hub*, 3, Special Issue 6S, 2021, 1-5. doi: 10.47392/irjash.2021.156