**Machine Learning**

**CIA -2 A**

**Submitted By:**

Angel Mary Abraham (2048027)

Namitha Antony (2048044)

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| 1. Introduction | 01 | |
| 1.1 Domain | 01 |  | |
| 1.2 Problem Statement  1.3 Need to solve the problem  1.4 Scope of the proposed work  1.5 Benefits from the proposed work | 01  02  02 |  | |
| 2. About the data | 03 |  | |
| 2.1 Data Description | 03 |  | |
| 2.2 Feature Description | 03 |  | |
| 3. Implementation Details | 04 |  | |
| 3.1 Exploratory data analysis | 04 |  | |
| 3.2 Data Pre-Processing | 14 |  | |
| 3.3 Model Building | 15 |  | |
| 3.4 Model Evaluation | 19 |  | |
| 3.5 Visualisation and Analysis | 20 |  | |
| 4. Results and Discussion | 21 |  | |
| 4.1 Results from Pie-Charts and Count Plots | 21 |  | |
| 4.2 Results of Multiplot wrt Target variable  4.3 Results obtained from different models | 22 |  | |
| 5. Limitations | 23 |  | |
| 6. Future Works | 24 |  | |
| Acknowledgement  References | 25 |  | |

**A Digital Survey on the Acceptance, Affordability, and Willingness to take COVID 19 Vaccine**

**------------------------------------------------------------------------------------------------**

**1. INTRODUCTION**

**1.1 Domain - Healthcare**

The healthcare sector consists of businesses that provide medical services, manufacture medical equipment or drugs, provide medical insurance, or otherwise facilitate the provision of healthcare to patients. The healthcare industry is one of the world's largest and fastest-growing [industries](https://en.wikipedia.org/wiki/Industry_(economics)). Consuming over 10 percent of the [gross domestic product](https://en.wikipedia.org/wiki/Gross_domestic_product) (GDP) of most [developed nations](https://en.wikipedia.org/wiki/Developed_nations), health care can form an enormous part of a country's economy. Growing incidence of lifestyle diseases, rising demand for affordable healthcare delivery systems due to the increasing healthcare costs, technological advancements, the emergence of telemedicine, rapid health insurance penetration, and government initiatives like e-health together with tax benefits and incentives are driving the healthcare market in India.

**1.2 Problem Statement**

The rising number of Covid Positive cases has been one of the major concerns in the country. Reducing and Controlling the number of Covid cases in India has become the need of the hour. The Indian Medical Research Institutes have successfully designed Covaxin and Covishied vaccines. But it has been observed that a lot of vaccines manufactured are wasted because the vaccines are not easily available at adorable prices to the public. Also, many rumors and misinformation regarding the side effects of Vaccines have been spread among the public. Hence, this survey is conducted to estimate the number of people who are willing to take vaccines at a certain locality, so that the number of vaccines for the locality can be manufactured accordingly and the wastage of vaccines can be reduced.

**1.3 Need to solve the problem**

It is expected that the COVID-19 pandemic will begin to inflict immense morbidity and mortality pressures thus seriously affecting populations and economies worldwide. Governments must be prepared to ensure access to and delivery of COVID-19 vaccines on a wide scale, in an equal manner, if and when safe and reliable vaccines become available. This would include ample capacity for the health sector, as well as strategies to increase trust in and understanding of the vaccine and those that administer it.

Vaccine wastage is an expected component of any large vaccination drive, and a vaccine is procured from the maker with an estimated wastage. For each vaccine type, the wastage has to be within recommended limits. In general, high vaccine wastage inflates vaccine demand and increases unnecessary vaccine procurement and supply chain costs. Vaccine wastage is directly linked to vaccine usage, which is the proportion of vaccines administered against vaccines issued to a vaccination site. Hence, it the need of the hour to estimate an approximate number of people, who are willing to take the Covid vaccine as per their availability, affordability, and willingness. This can help in the reduction of Vaccine wastage and help in determining the count of the public that trust the Indian Vaccines.

**1.4 Scope of the proposed work**

Effective vaccine utilization is an integral component of vaccine security and vaccine wastage is one of the key factors to be considered with regards to vaccine forecasting and need estimation. The objectives of the vaccine wastage assessment were to provide an estimation of vaccine wastage rate, type, and place of occurrence and recommend measures to reduce wastage at various levels. A better sense of vaccine utilization and wastage rates can lead to better planning and management of vaccine stocks. This assessment will give some information about the current wastage level, which could then lead to appropriate guidance and training to reduce vaccine wastage.

**1.5 Benefits from the proposed work**

If this study is proved to be successful, then it can be of great benefit to the vaccine manufacturers and health care industry of India. Currently, our country is undergoing a major crisis of Coronavirus pandemic. Vaccinating the majority of the Indian public has become the need of the hour. However, while conducting the vaccination drives in the country it is important to make sure the proper distribution and availability of the vaccine amongst the public. The results obtained from this survey of ‘Acceptance, Affordability and Willingness to take COVID 19 Vaccine’ can help in recognizing the preferences of people while taking the vaccine. It can help to identify the majority of the public sectors who are willing to take the vaccine. Hence, this survey can help in estimating the number of people who are willing to take vaccines at a certain locality, so that the number of vaccines for the locality can be manufactured accordingly and the wastage of vaccines can be reduced.

**2. ABOUT THE DATA**

The data was collected from different age groups of people across the states of India. It was a digital survey conducted via google forms. The questionnaire in the google forms consisted of personal information and collecting the details stating people’s attitude regarding acceptance. affordability and willingness to take covid vaccine.

**2.1 Dataset Description**

Table 1: Dataset Description

|  |  |
| --- | --- |
| **Name** | **Willingness to take COVID 19 Vaccine** |
| Type | Multivariate |
| No of Rows | 540 |
| No of Columns | 20 |
| Missing Values | No |
| Target Type | Binary |
| Application Technique | Classification |

**2.2 Feature Description**

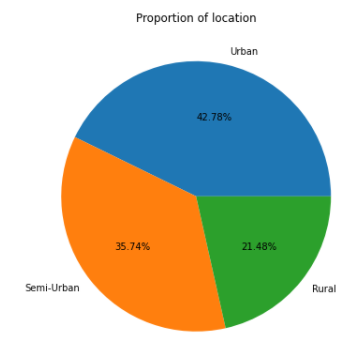
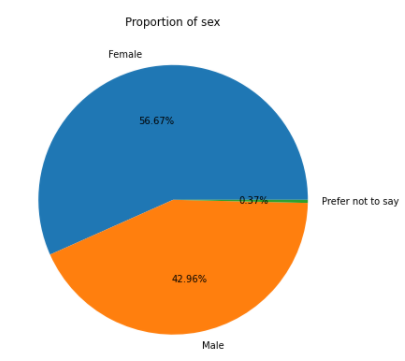
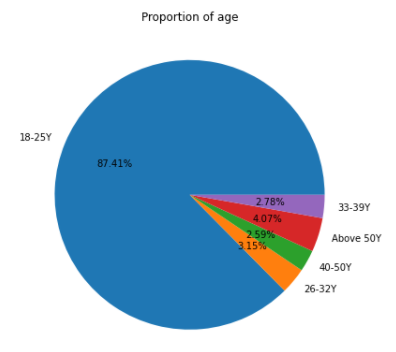
Table 2: Feature Description

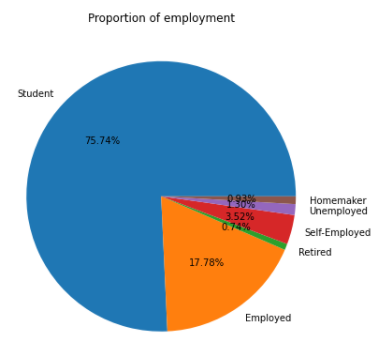
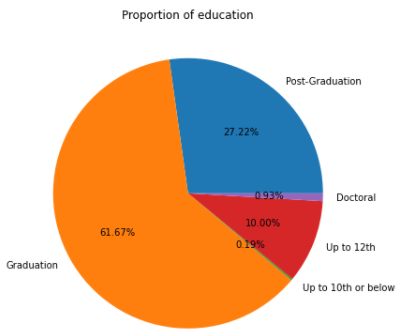
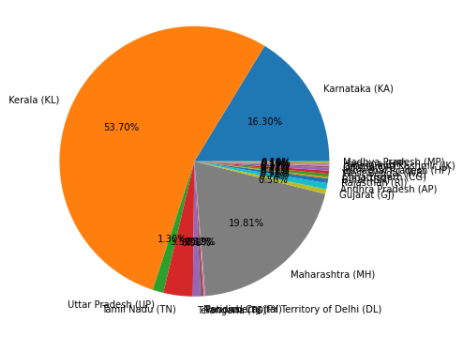
|  |  |
| --- | --- |
| **Feature Name** | **Description** |
| Timestamp | It indicates the time at which they have filled the form |
| Name | It indicates the name of the participant |
| Age | It indicates the age of the participant |
| State/UT | It indicates the state where they live |
| Where do you live? | It indicates the location where they live (Rural, Semi-Urban, or Urban) |
| Educational qualification? | It indicates the educational Qualification of the participant |
| What is your Employment? | It indicates the employment Status of the participant |
| Marital Status? | It indicates whether the participant is married or not |
| Do health centers near you currently have the ability to test patients for COVID-19? | It indicates whether the health centers near them have the provision to check for COVID-19 |
| Which Indian vaccine is availed by your state or UT? | It indicates which vaccine is available in their locality |
| On average for each week, how quickly are health centers near you, able to obtain COVID-19 test results for SARS-CoV-2 virus detection (PCR, antigen)? | It indicates the time the health centers take to give the test results |
| How many people from your locality have tested positive for COVID-19 in the last week? | It indicates the number of patients reporting in their locality in a week |
| Which of the following define your concerns about getting the vaccine? | It indicates the concerns about the providing vaccine |
| What challenges does your locality face in deploying the COVID-19 vaccine? | It indicates the challenges of their health center in giving the vaccine to everyone |
| Which of the following would be helpful if you did not decide yet about taking the vaccine? | It indicates when they will be ready to take the vaccine |
| Your affordable cost of the vaccine? | It indicates the price they are ready to spend for the vaccine |
| Are you planning to get the COVID-19 vaccine? | It indicates whether they are willing to take the vaccine |
| If Yes, then You want a vaccine? | It indicates which vaccine they are preferring to take |
| Please provide any additional information, comments, or challenges you are experiencing due to the supply or intake of the COVID-19 vaccine. | Additional information about the pandemic, vaccine, etc. |

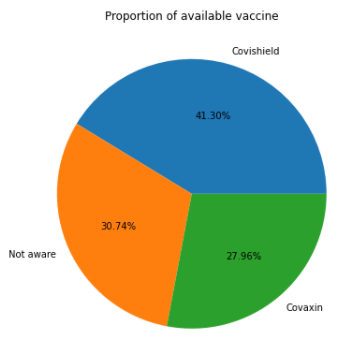
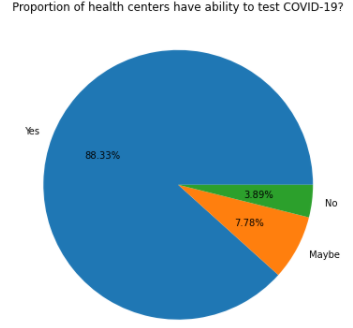
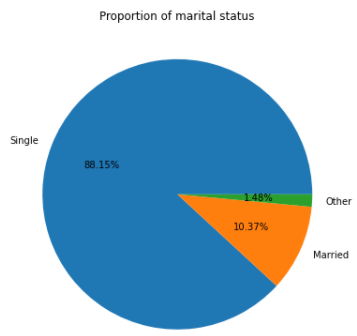
**3. IMPLEMENTATION DETAILS**

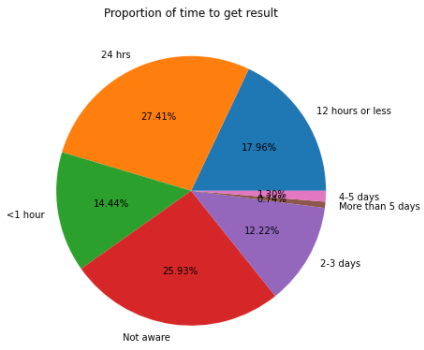
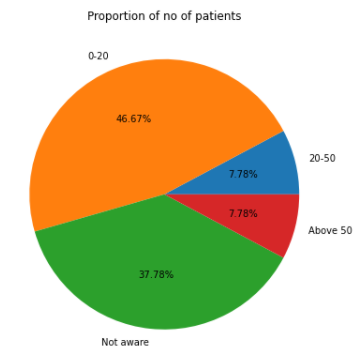
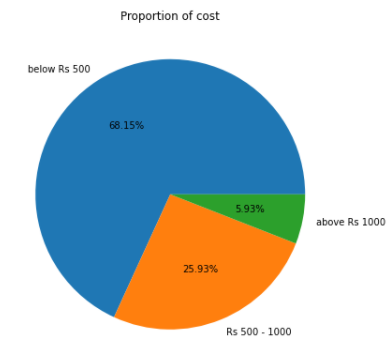
**3.1 Exploratory Data Analysis**

Visualization using pie-charts and count plots has been implemented for several variables such as age, sex, state, location, education, employment, marital status, health center ability to test COVID-19, available vaccine, time to get test results, the number of patients tested positive in the locality, cost of the vaccine, the willingness of people to take the vaccine and which is the more preferred vaccine.







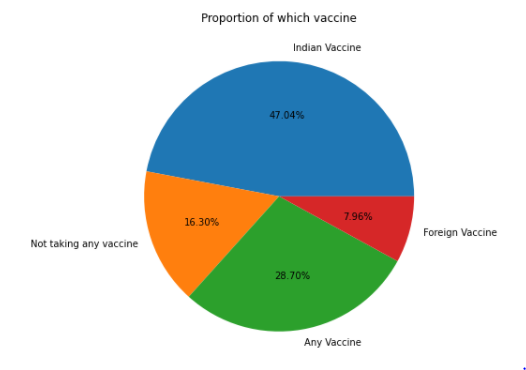
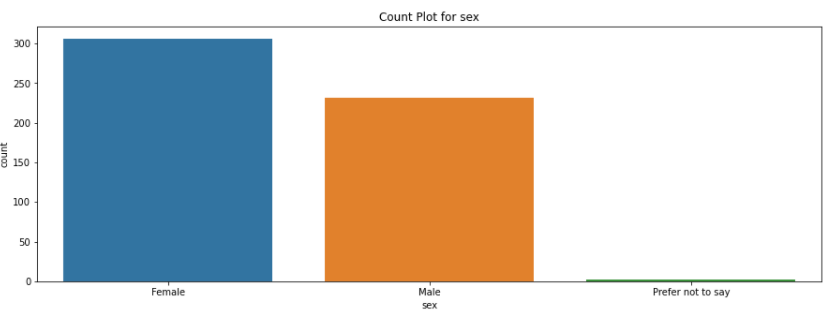
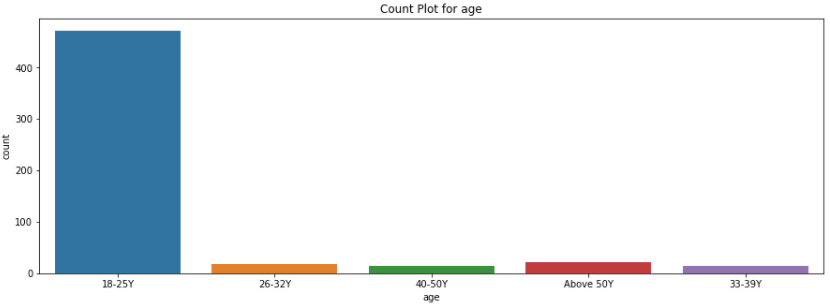
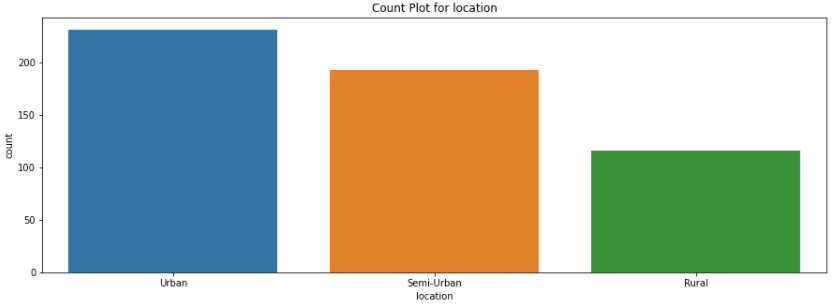
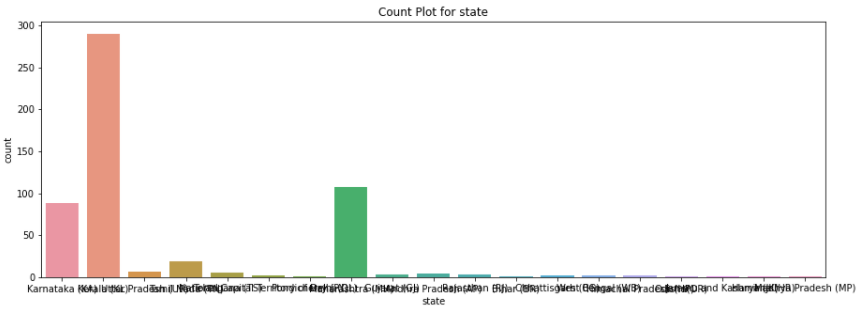
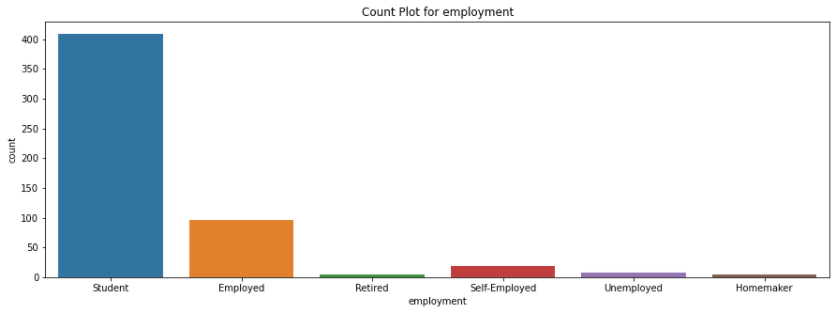
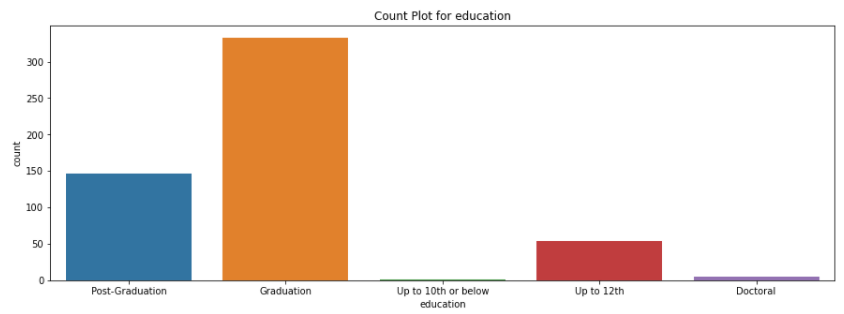
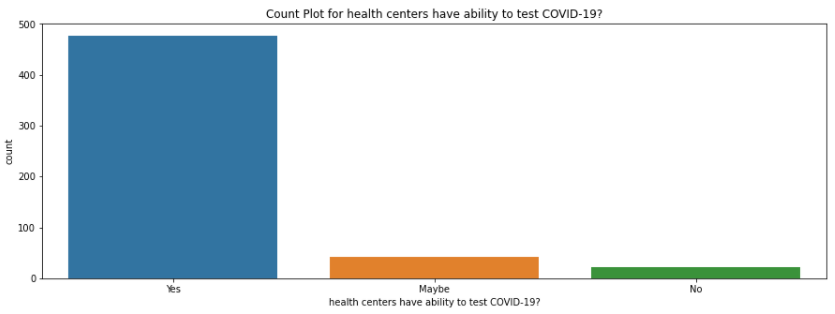
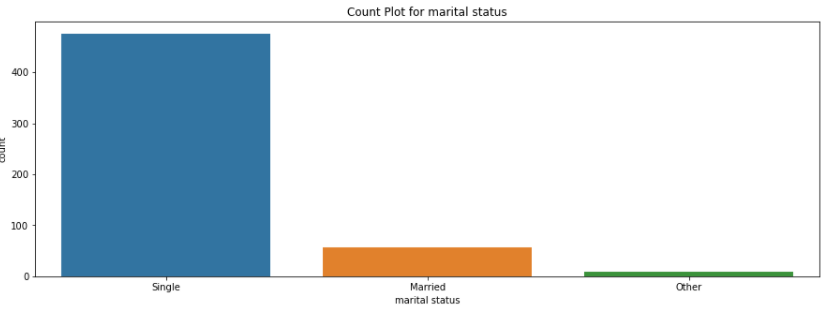


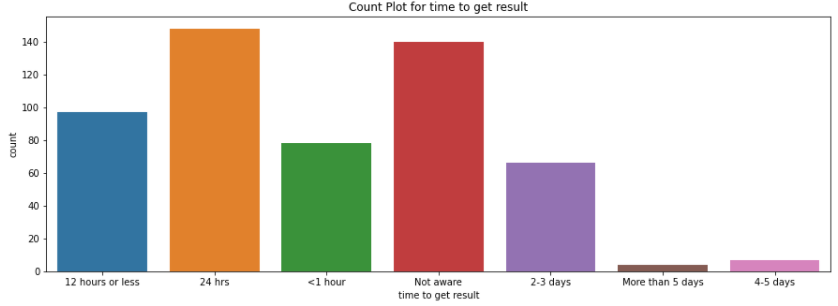
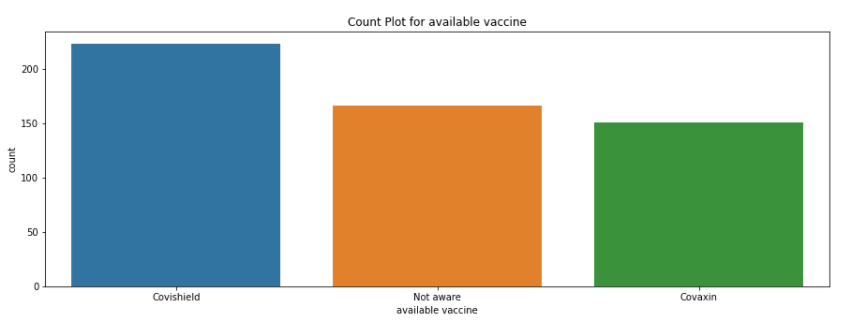
Fig 1: Pie Chart of all Variables

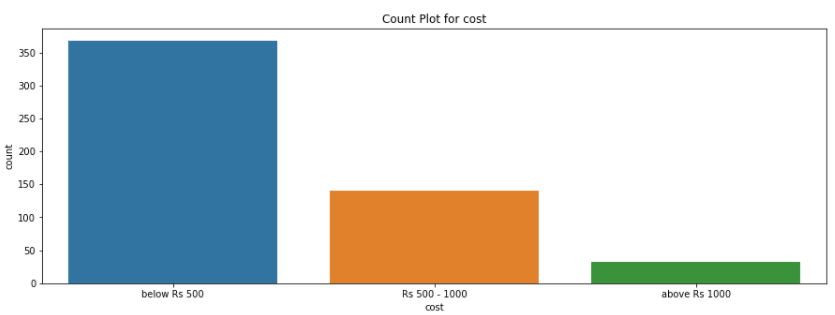
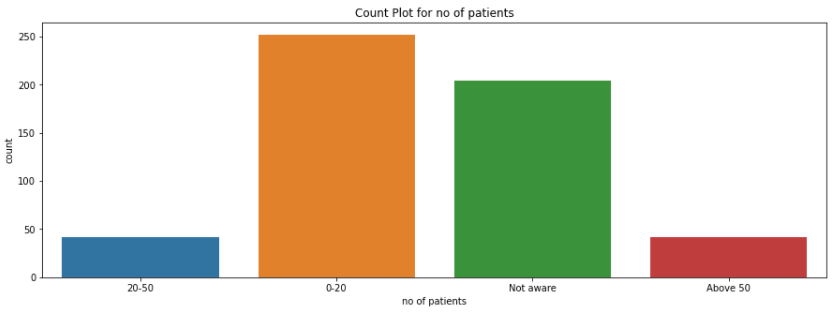












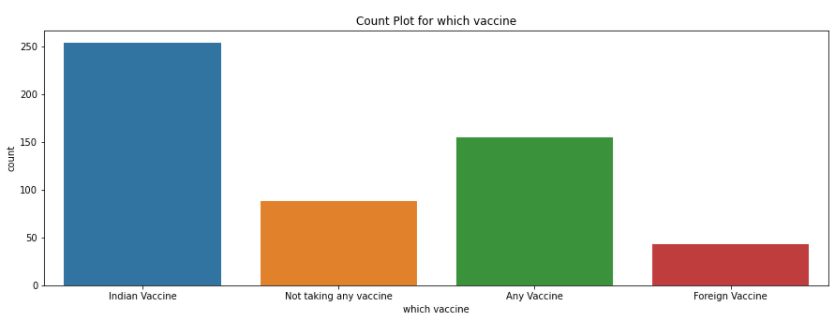
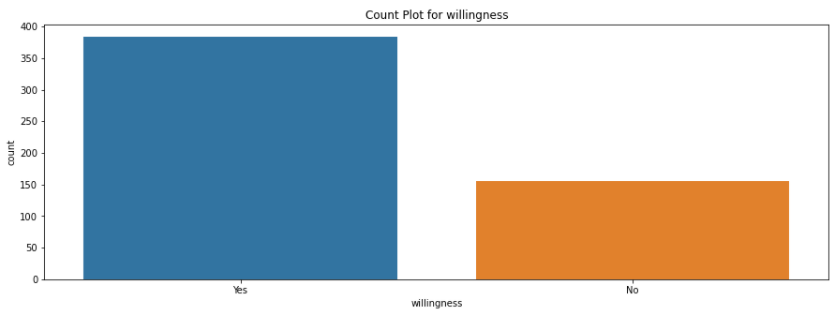
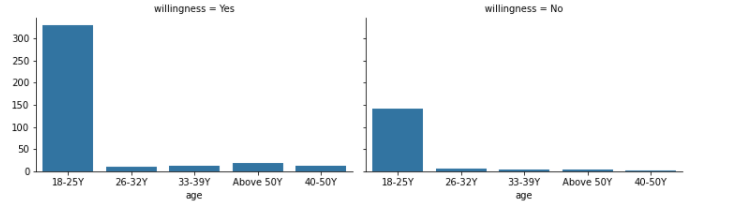
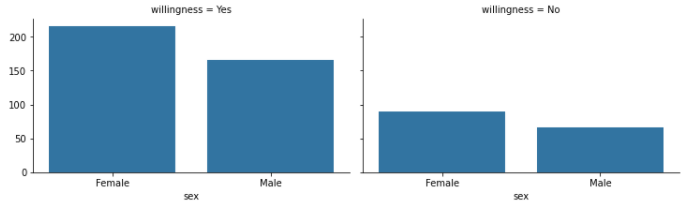
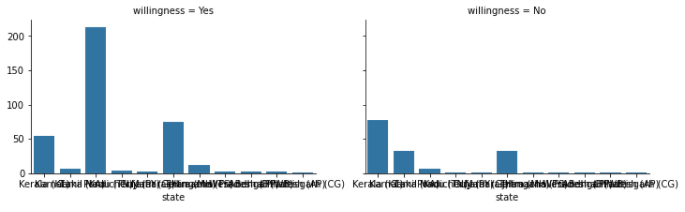


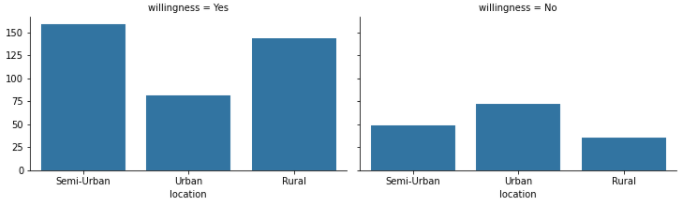
Fig 2: Count Chart of all variables

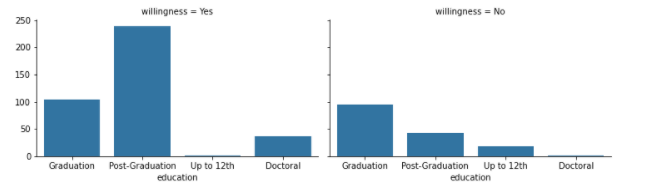
The multiplot for all factors for the target variable that is ‘Willingness to take the vaccine’ has been obtained. It depicts the willingness of people to take the vaccine, considering factors such as age, sex, state, location, education, employment, marital status, health centre ability to test COVID-19, available vaccine, time to get test results, number of patients tested positive in the locality, cost of the vaccine, the willingness of people to take the vaccine and which is a more preferred vaccine.

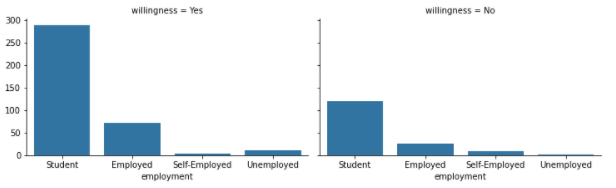


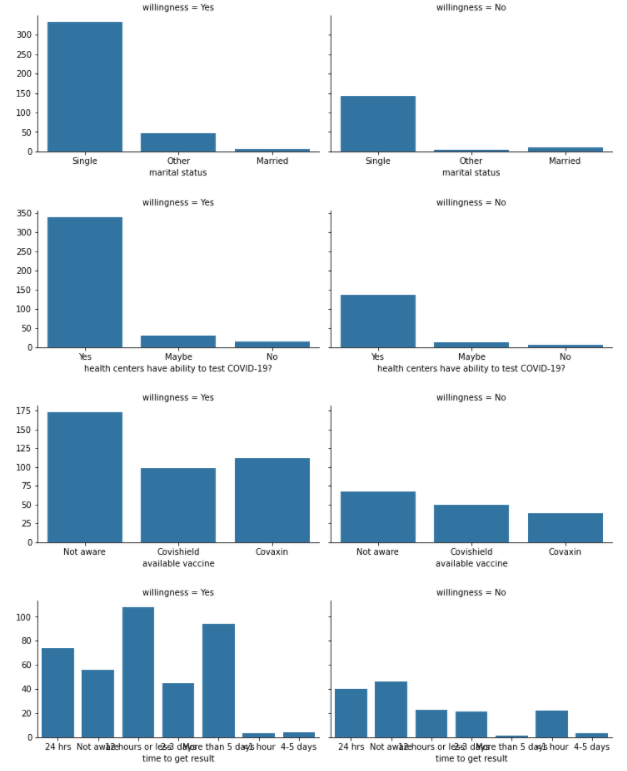












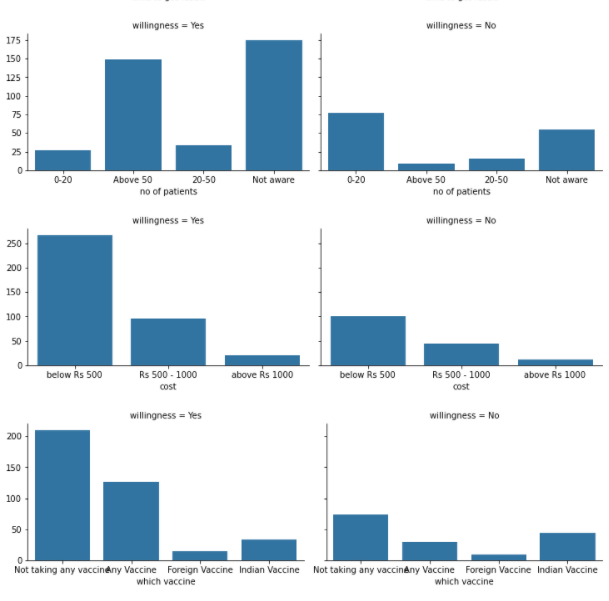
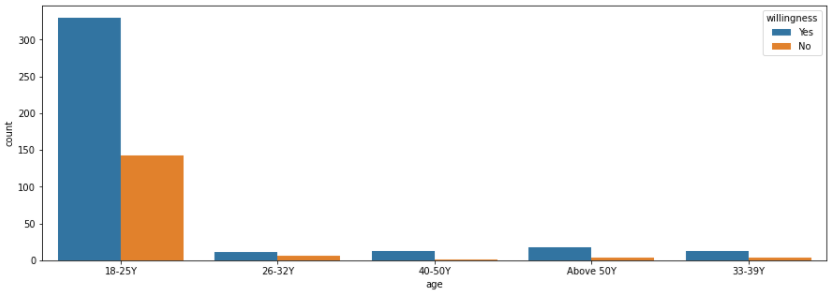
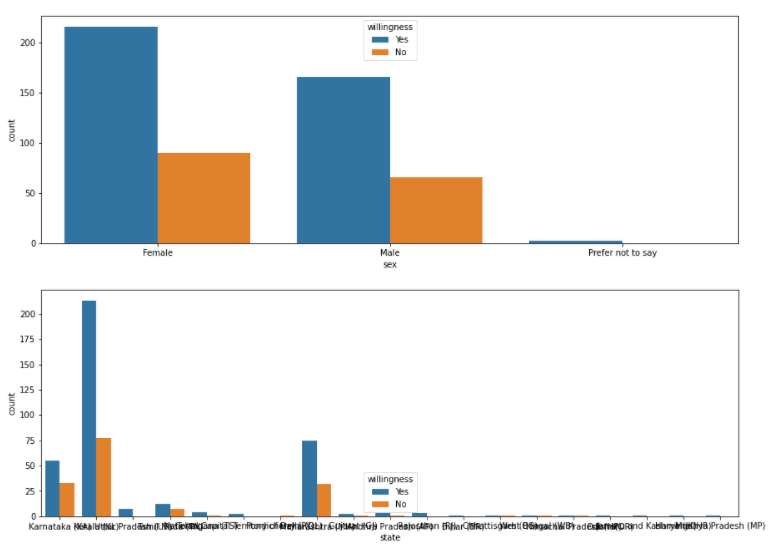
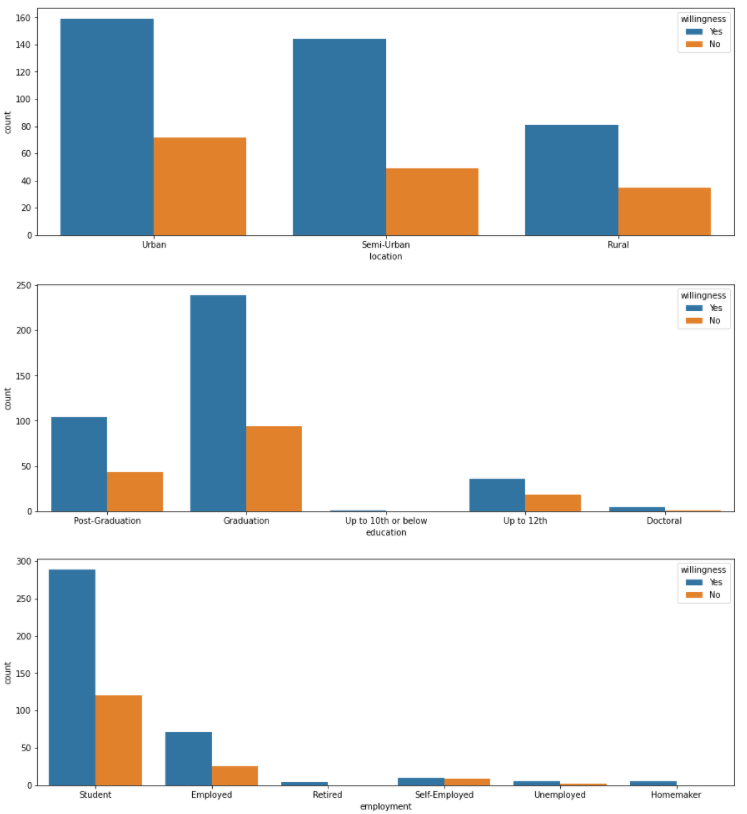


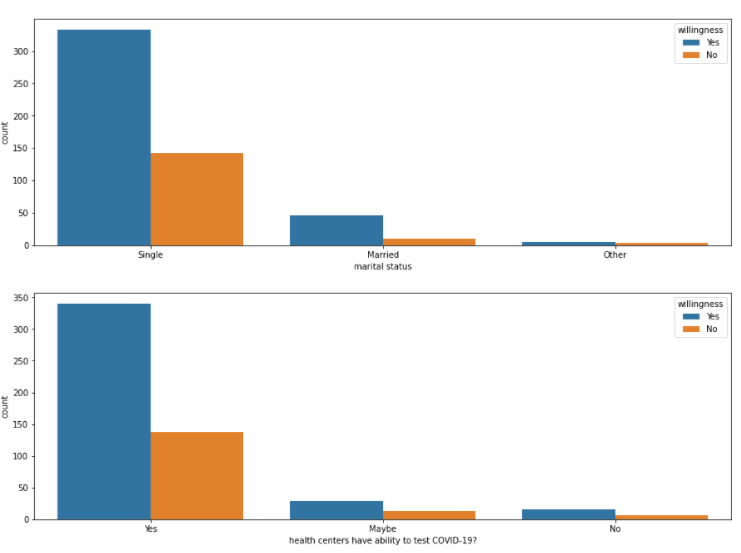
Fig 3: Multi-plot of all features wrt target variable

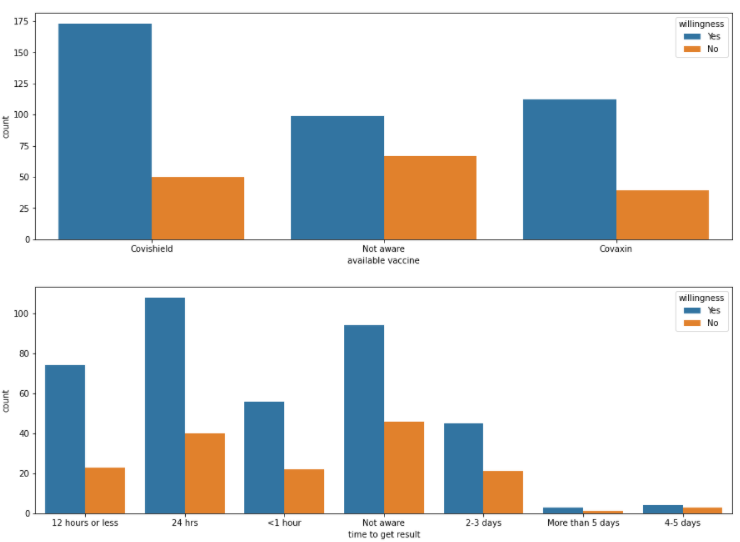
The count plot for all factors concerning the target variable that is ‘Willingness to take the vaccine’ has been obtained with the categorization as ‘Yes’ or ‘No’. These graphs depict the willingness of people to take the vaccine, concerning their age, sex, state, location, education, employment, marital status, health centre ability to test COVID-19, available vaccine, time to get test results, the number of patients tested positive in the locality, cost of the vaccine, the willingness of people to take the vaccine and which is the more preferred vaccine.

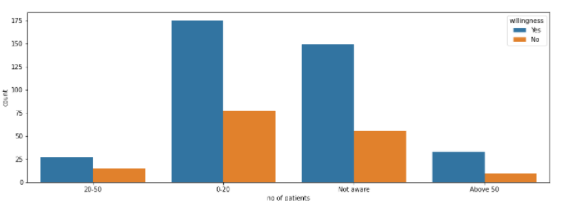












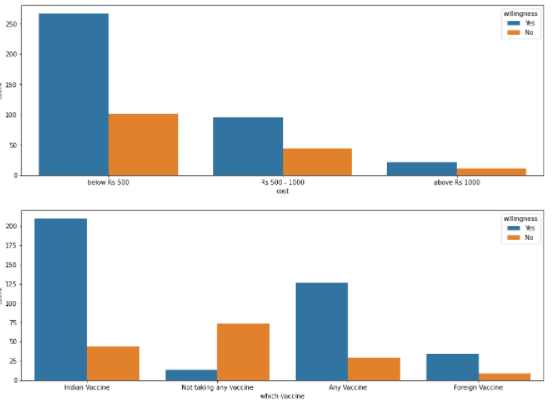


Fig 4: Count plot of all features with the target variable

**Dependency check of factors on the target variable**

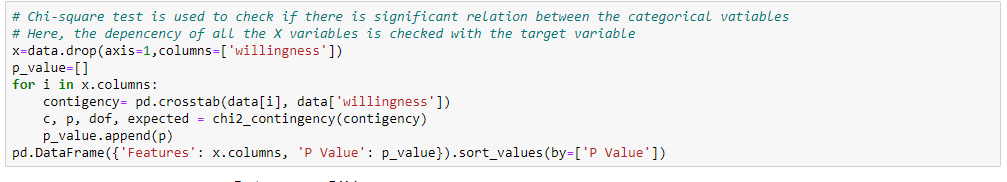
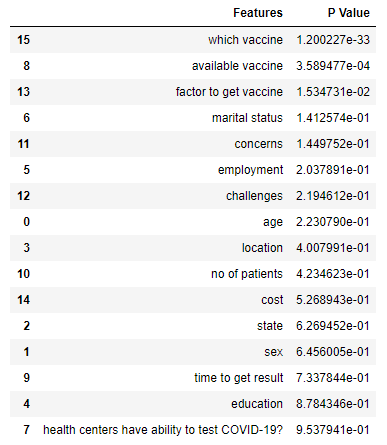


Table 3: P-value of each variable in Chi square test

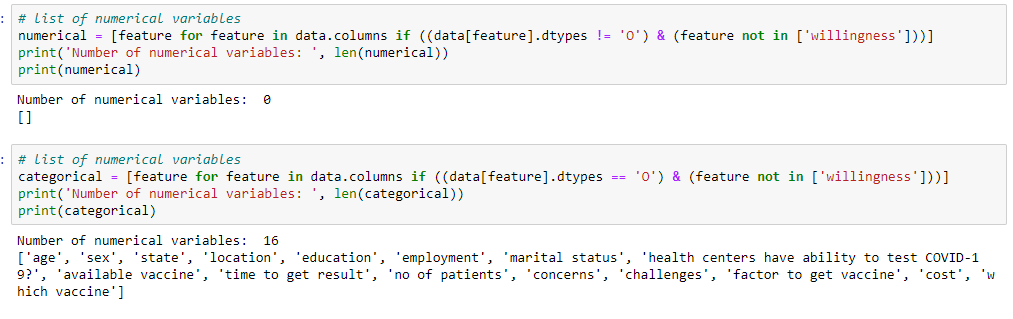


The variables with a P value less than 0.05 have a significant relationship with the target variable.

**3.2 Data Pre-processing**

**3.2.1 Finding number of numerical and categorical variables**

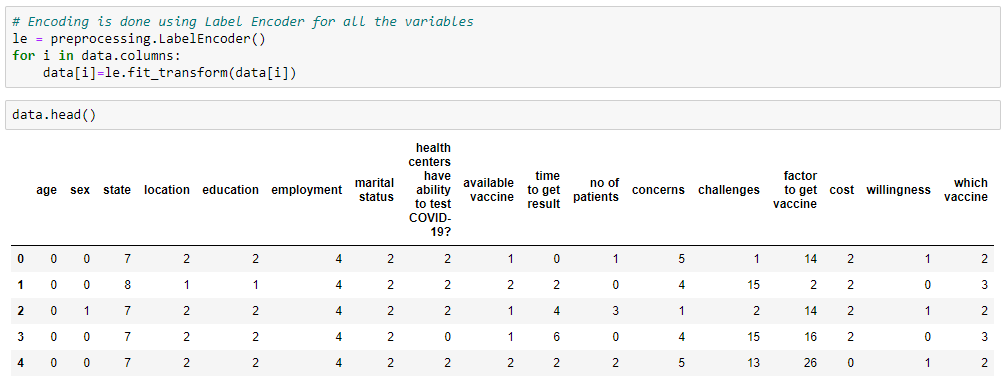
Numerical variables can be discrete or continuous. Discrete variables are those where the pool of possible values is finite and are generally whole numbers, such as 1, 2, and 3. Categorical variables are values that are selected from a group of categories, also called labels. Examples of categorical variables include gender, which takes values of male and female, or country of birth, which takes values of Argentina, Germany, and so on.



There are no numerical columns. All the columns in the dataset are categorical.

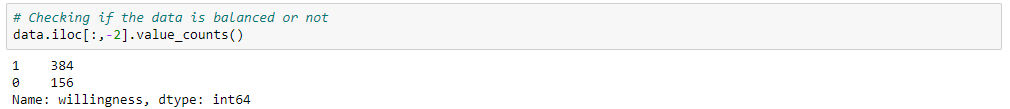
**3.2.2 Encoding the categorical variables**

Label encoding means that if the data contains categorical data, we must encode it to numbers before fitting and evaluating a model. This categorical data encoding method transforms the categorical variable into a set of binary variables (also known as dummy variables).

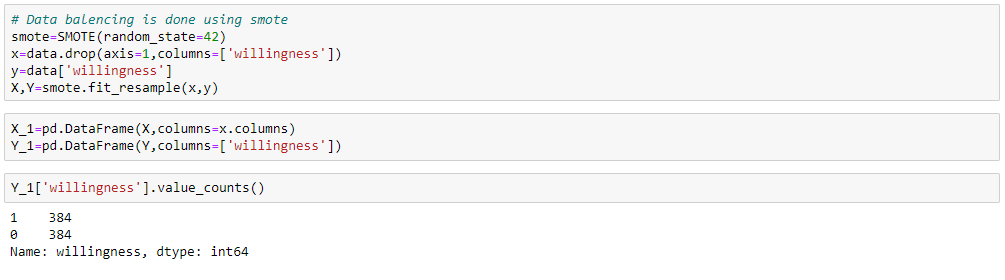


**3.2.3 Data Balancing**

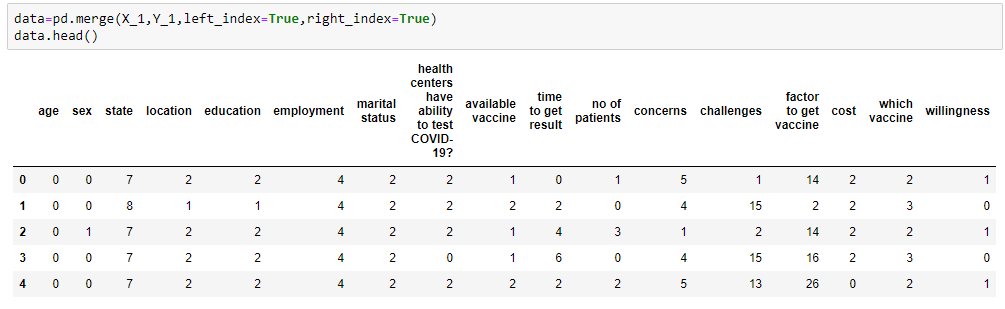
A balanced data set is a set that contains all elements observed in all time frames.



There is a big difference in the frequency of each class. Therefore, data balancing is needed



 Data is balanced using [data augmentation](https://machinelearningmastery.com/how-to-configure-image-data-augmentation-when-training-deep-learning-neural-networks/) for the minority class and is referred to as the Synthetic Minority Oversampling Technique, or SMOTE.



**3.3 Model Building**

**3.3.1 Logistic regression**

Logistic regression is used to obtain an odds ratio in the presence of more than one explanatory variable. The procedure is quite similar to multiple linear regression, with the exception that the response variable is binomial. The result is the impact of each variable on the odds ratio of the observed event of interest.

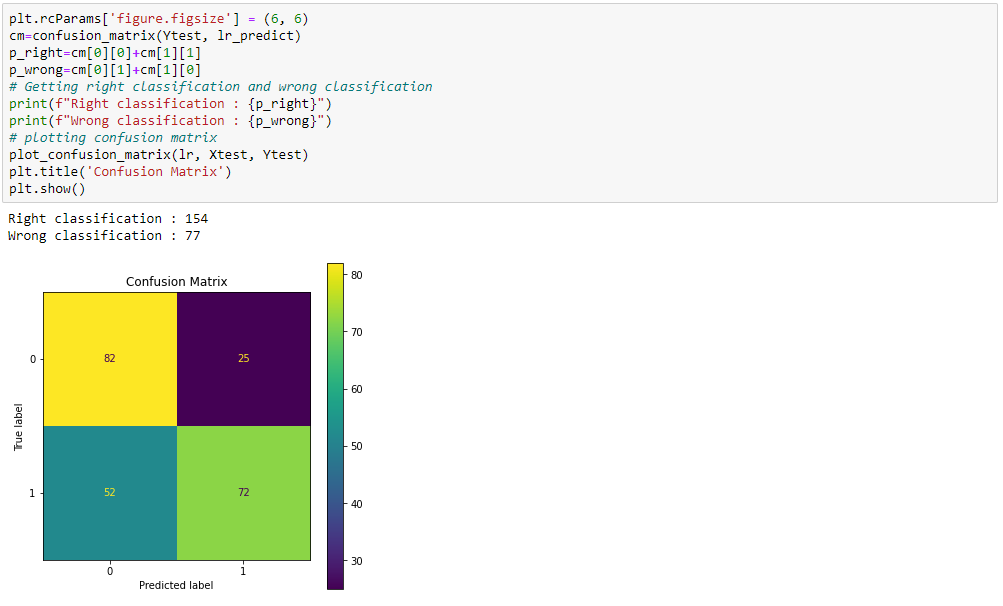


Fig 5: Confusion matrix of Logistic Regression

The logistic regression model provides a 67% accuracy rate.

**3.3.2 Naive Bayes classifiers**

Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified is independent of each other.

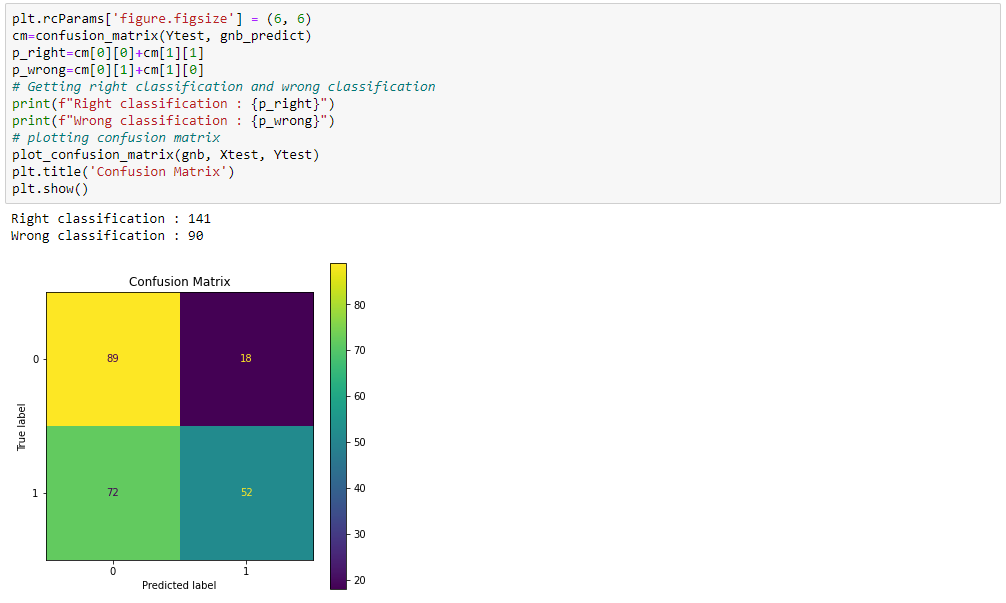


Fig 6: Confusion matrix of Gaussian Naïve Bayes Classifier

Naive Bayesian Classifier Model yields 61% accuracy

**3.3.3 Support vector machine**

A support vector machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labelled training data for each category, they're able to categorize new text. So you're working on a text classification problem

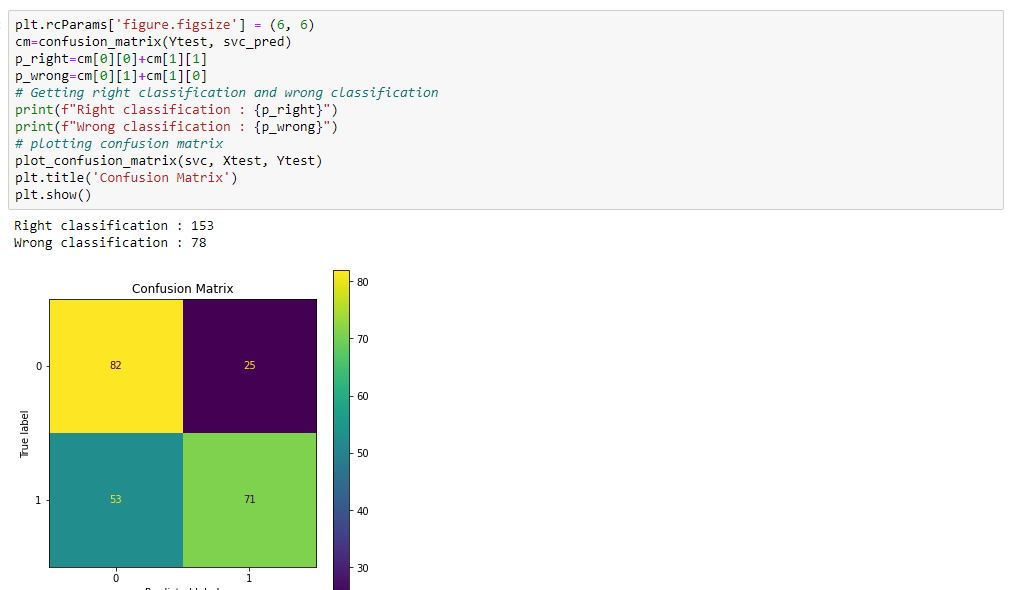


Fig 7: Confusion matrix of SVM Classifier

Support vector classifier model yields 66% accuracy.

**3.3.4 K nearest neighbours**

K nearest neighbours is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure. KNN has been used in statistical estimation and pattern recognition

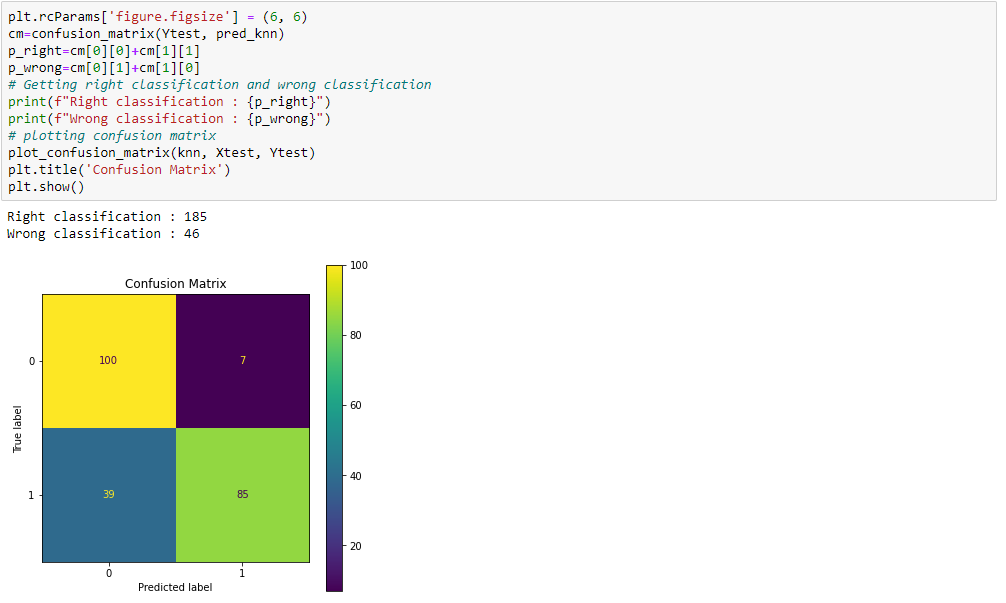


Fig 8: Confusion matrix of KNN Classifier

K-Nearest Neighbour classifier model yields 80% accuracy

**3.3.5 Random forests**

Random forests or random decision forests are an ensemble learning method for classification, regression, and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.

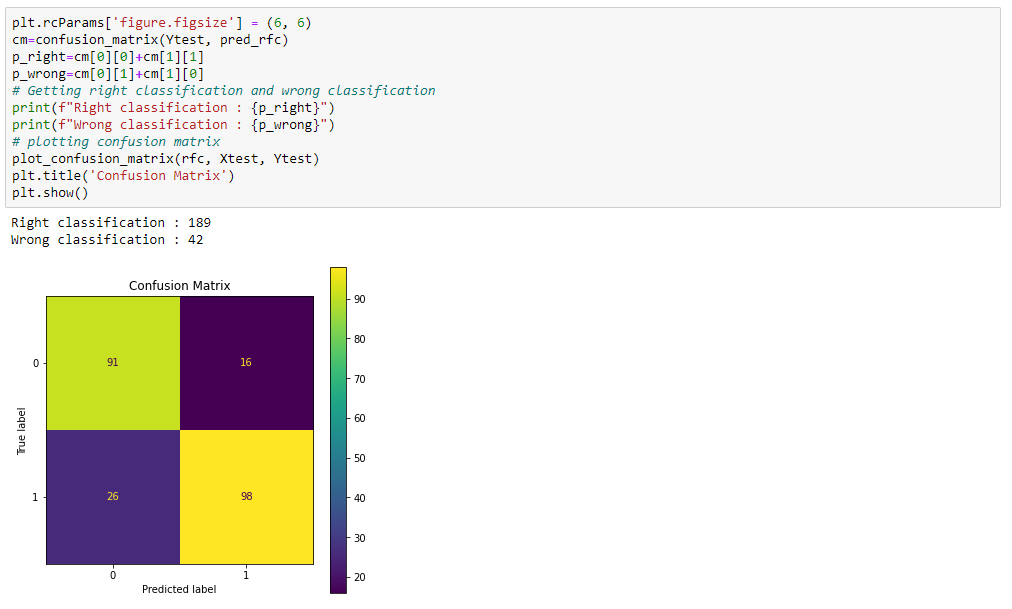


Fig 9: Confusion Matrix of Random Forest Classifier

Random Forest classifier model yields 82% accuracy

**3.4 Model Evaluation**

The best model has been selected based on its accuracy rate as well as the ROC curve.

**3.4.1 Accuracy**

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition: Accuracy = Number of correct predictions Total number of predictions.

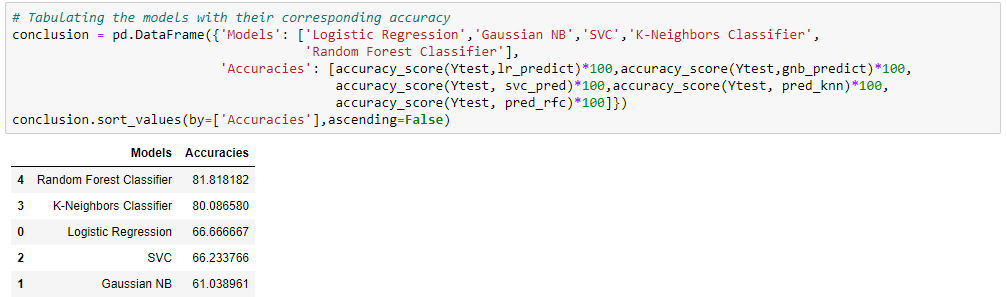
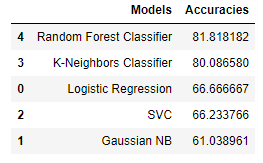


Table 4: Accuracy of each model



The Random Forest Classifier model have the maximum accuracy among all the fitted models

**3.4.2 ROC Curve**

A ROC curve shows the relationship between clinical sensitivity and specificity for every possible cut-off. The ROC curve is a graph with: The x-axis showing 1 – specificity (= false-positive fraction = FP/(FP+TN)) The y-axis showing sensitivity (= true positive fraction = TP/(TP+FN)). ROC curve of all the fitted models have plotted.

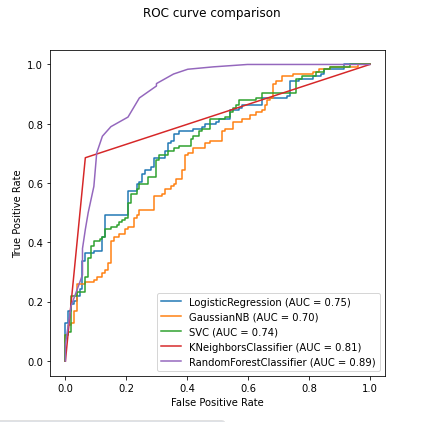


Fig 10: ROC curve of all the models

The ROC curve of random forest classifier is closer to top-left among all the models. Hence the random forest classifier is considered as the best model with respect to ROC curve.

**3.5 Visualisation and Analysis**

**3.5.1 Feature importance**

Feature importance in random forest describes which features are relevant. It can help with a better understanding of the solved problem and sometimes lead to model improvements by employing the feature selection

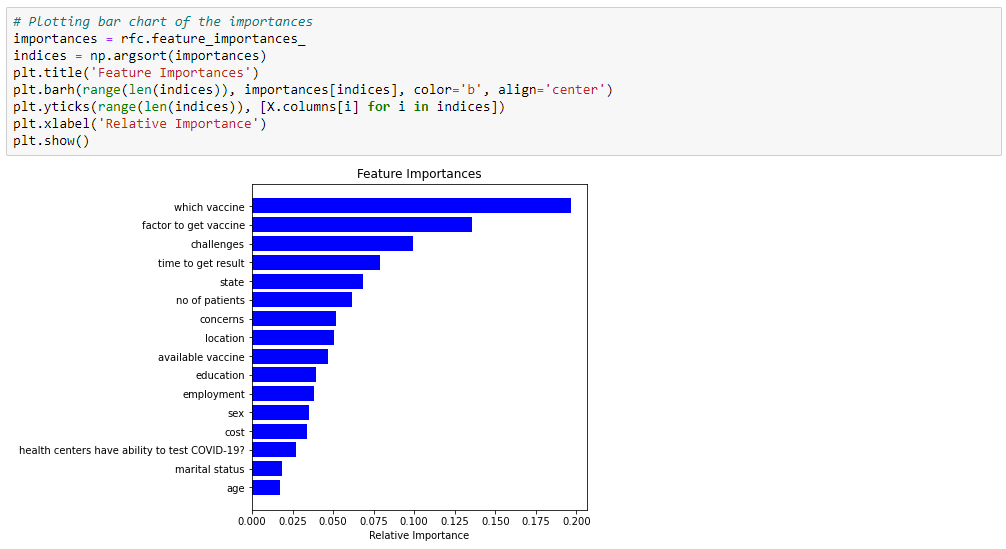
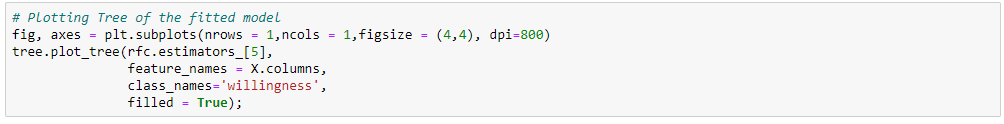


Fig 11: Feature Importance bar graph

The variable 'which vaccine' is of the highest importance and 'age' is of lowest importance in the fitted random forest classifier model.

**3.5.2 Plotting tree**

Plotting trees in the random forest gives a hunch of basically how a model predicts the value of a target variable by learning simple decision rules inferred from the data features. Every decision at a node is made by classification using a single feature



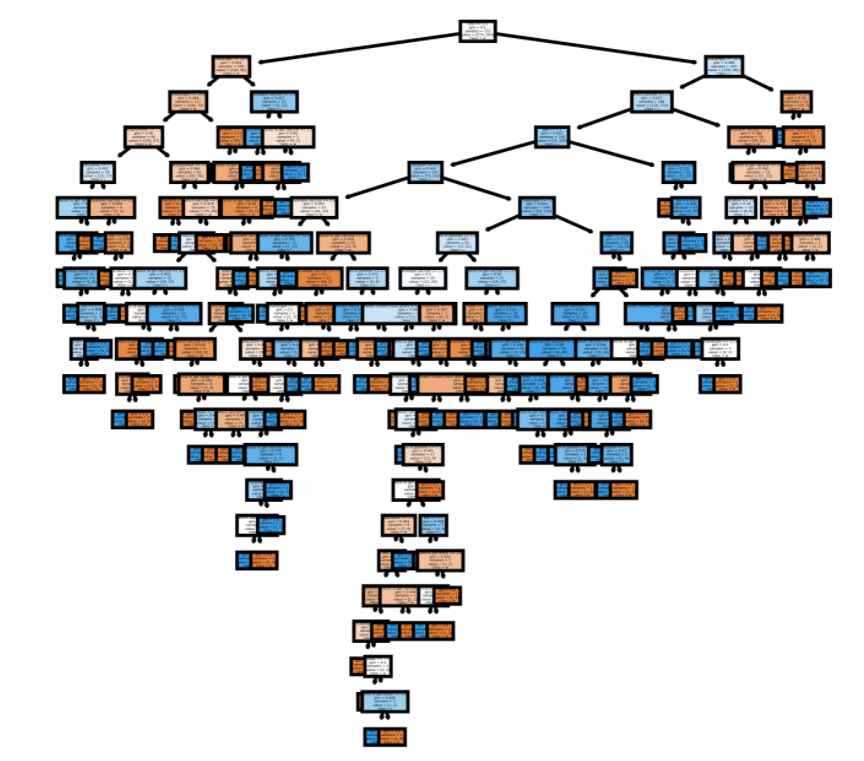


Fig 12: Decision Tree of the fitted model

**4. RESULTS & DISCUSSION**

**4.1 Results from Pie-Charts and Count Plots**

The visualized count plots and pie charts depict the approximate data about the nature of participants in our survey, their attitude towards the vaccination drive, and also their acceptance, affordability, and willingness to take COVID 19 vaccine. The major insights from this visualization are as follows:

1. The major of the people who participated in our survey are between the age of 18-25 years.
2. About 56.67% of the survey participants included females and 42.96% of them included males.
3. Majority of the people who participated in our survey hail from the states of Kerala, Maharashtra, and Karnataka.
4. Most of the survey participants are from the Urban or Sem-Urban sectors of India.
5. The majority of the people who participated in the survey have an education qualification till graduation level.
6. The majority of the survey participants are students and are single individuals.
7. About 88.33% of the survey participants have accepted that the health centers near them have the facility of testing Covid-19.
8. The vaccines availed by the states of the survey participants are Covisheld-41.30%, Covaxin 27.96%. Whereas, 30.74% of the participants are unaware of it.
9. Most of the participants claim that the health centers near them take a minimum of 24 hours to give the results of Covid tests. At the same time, about 23.96% of the people are unaware of the test time.
10. About 46.67% of people have claimed that the locality near them has reported at least 0-20 patients in the previous week of the survey.
11. About 68.15% of people in the survey prefer the cost of the vaccine as below Rs 500.
12. The majority of the people who have decided to take the vaccine prefer the Indian vaccine over others.

**4.2 Results of Multi-plot wrt Target variable**

The multiple count plots considering the willingness of people to take the vaccine have the following interpretations and results:

1. The visualization depicts that most of the participants between the age of 18-25 years are willing to take the vaccine. The overall interpretation of all the age groups also shows that most people are willing to take the vaccine.
2. The sex group categorization shows that the majority of the females, as well as males, are willing to take the vaccine. However, the survey participants consist of the majority as females.
3. The state-wise grouping of the survey participants shows that majority of the people from Kerala, Karnataka, and Maharashtra are willing to take the vaccine when available.
4. Most of the people from Urban, Sem-Urban, and Rural localities have decided to take the vaccine. However, the majority of the participants consisted of Urban and Semi-Urban areas.
5. The majority of the people who are graduated, single, and are students are willing to take the vaccine.
6. The majority of the people with covid testing health centers near them are willing to take the vaccine.
7. Most of the people whose states have availed of the Covisheild vaccine are willing to take the vaccine.
8. The majority of the people whose nearest health centers provide the covid test results within 24 hours are willing to take the vaccine.
9. The majority of the people whose locality has accounted for at least 20 covid positive patients in the past week of the survey have claimed that they are willing to take the vaccine.
10. Most of the people who have chosen the preferred vaccine price below Rs. 500 have decided to take the vaccine.
11. Most of the people who prefer Indian vaccines have decided to take the vaccine.

**4.3 Results obtained from different models**

Model accuracies have been obtained as follows: Random Forest Classifier 81.8181823, K-Neighbours Classifier 80.0865800, Logistic Regression 66.6666672, SVC 66.2337661, and Gaussian NB61.038961. Both accuracy rates and ROC curve depict the Random Forest as the best fit model suitable for this survey data. Hence, we can conclude that the Random Forest method with an accuracy of 81.8181823% is the best fit model for this survey of Acceptance, Affordability, and Willingness to take COVID 19 Vaccine.

**5. LIMITATIONS**

1. The majority of the participants who filled the survey forms belong to the age group of 18-25 years. Hence, the data obtained from other age groups were limited.
2. Majority of the people who participated in the survey hail from the states of Maharashtra, Kerala, and Karnataka. Hence, the data obtained from other Indian states were limited.
3. Most of the people who filled the survey details have claimed the details asked in the form as unaware. Hence, the majority of the public has only limited information regarding the vaccination drive.
4. The majority of the people who filled the survey forms belong to the urban or semi-urban sectors. Hence, the data obtained from the rural sectors was limited.
5. The majority of the people who filled the survey forms were students and had educational qualifications till graduation level. Hence, the data obtained from other professions were limited.

**6. CONCLUSION**

Vaccine wastage is an expected component of any large vaccination drive, and a vaccine is procured from the maker with an estimated wastage. Vaccine wastage is directly linked to vaccine usage, which is the proportion of vaccines administered against vaccines issued to a vaccination site. The documentation of vaccine wastage is poor at all levels. The reporting of vaccine wastage is still observed in many states in India. However, an assessment of the wastage at the supply chain should be conducted for a larger period (at least one year). The reasons for vaccine wastage vary from person to person. The main reasons including the misinformation spread about the vaccine among the public. However, from the survey, it can be observed that most people can afford to take the vaccine if its market availability is at the price below Rs. 500. Hence, reducing the vaccine cost can encourage more people to use the available vaccine and reduce the vaccine cost. Many people who participated in the survey also seemed to be unaware of the details of the vaccination drive. Hence, spreading more valid information about the vaccine through campaigns and surveys can clear the rumors and encourage more people to take the vaccine.

**7. FUTURE WORKS**

1. We are planning to expand this project by collecting more data from various age groups focusing more on the older population as well.
2. The people who participated in this survey are mostly from the states of Maharashtra, Kerala, and Karnataka. We are planning to collect more data from other states of India for expanding our analysis.
3. We are planning to develop a Web Application to identify the ‘Acceptance, Affordability and Willingness to take COVID 19 Vaccine’, through which people can fill in their preferences regarding the vaccine. Hence, the vaccine can be manufactured for each locality accordingly to the preference of the public and the wastage of vaccine can be reduced.

**ACKNOWLEDGEMENT**

We would like to express gratitude to our teacher Mrs. Ummesalma M for giving us an opportunity to do this project on Machine Learning and guiding us until the completion of our project. We would like to extend my gratitude to the Coordinator, Head of the department, Dean of Sciences and the Management at Christ University for such a useful opportunity.

**REFERENCES**

[1] J. Ali, R. Khan, N. Ahmad, and I. Maqsood, “Random Forests and Decision Trees,” no. December 2013, 2012.

[2] G. Pipis, “How to Run the Chi-Square Test in Python.” [Online]. Available: https://medium.com/swlh/how-to-run-chi-square-test-in-python-4e9f5d10249d.

[3] J. Brownlee, “SMOTE for Imbalanced Classification with Python.” [Online]. Available: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/.

[4] Scikit-Learn, “sklearn.linear\_model.LogisticRegression.” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.linear\_model.LogisticRegression.html.

[5] Scikit-Learn, “sklearn.naive\_bayes.GaussianNB.” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.naive\_bayes.GaussianNB.html.

[6] Scikit-Learn, “sklearn.svm.SVC.” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html.

[7] J. Korstanje, “The k-Nearest Neighbors (kNN) Algorithm in Python.” [Online]. Available: https://realpython.com/knn-python/.

[8] Scikit-Learn, “sklearn.ensemble.RandomForestClassifier.” [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[9] J. Brownlee, “How to Use ROC Curves and Precision-Recall Curves for Classification in Python.” .

[10] W. Koehrsen, “How to Visualize a Decision Tree from a Random Forest in Python using Scikit-Learn.” [Online]. Available: https://towardsdatascience.com/how-to-visualize-a-decision-tree-from-a-random-forest-in-python-using-scikit-learn-38ad2d75f21c.

[11] J. Brownlee, “How to Calculate Feature Importance With Python.” [Online]. Available: https://machinelearningmastery.com/calculate-feature-importance-with-python/.

[12] W. Koehrsen, “An Implementation and Explanation of the Random Forest in Python.” [Online]. Available: https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76.