CHRONIC KIDNEY DISEASE-SURVIVAL ANALYSIS

SUBMITTED BY :-

DEVI ESWAR KUMAR DAMERLA

RUFUS NEMALIKANTI

SAKTHIVEL.V

CONTENT

1.1 Introduction……………………………………………3

1.2 Objectives of research…………………………….5

1.3 Problem Statement…………………………………5

1.4 Industry Profile……………………………………….6

2.Review of Literature…………………………………..6

3.Data Collection………………………………………….7

4.Methodology…………………………………………….9

4.1 Exploratory Data Analysis……………………9

4.1.1 Figures and Tables ……………………9

4.2 Statistical Techniques and Data

Visualization………………………………………………..11

4.3 Data Modelling using Supervised ML techniques…………………………………………………13

5. Findings and Suggestion…………………………16

6. Conclusion………………………………………………16

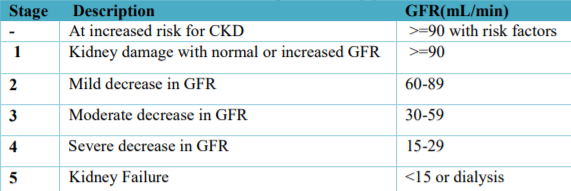
7. References……………………………………………..17

1.1 INTRODUCTION

In today's era everyone is trying to be conscious about health although due to workload and busy schedule one gives attention to the health when it shows any symptoms of some kind. But CKD is a disease which doesn't shows symptoms at all or in some cases it doesn't show any disease specific symptoms it is hard to predict, detect and prevent such a disease and this could be lead to permanently health damage, but machine learning can be hope in this problem it is best in prediction and analysis. By using data of CKD patients with 12 attributes and 400 record we are going to use various machine learning techniques like Decision Tree, SVM, etc. To build a model with maximum accuracy of predicting whether CKD or not and if yes then its Severity.

We all knowthat Kidney is essential organ in human body. Which has main functionalities like excretion and osmoregulation. In simple words we can say that all the toxic and unnecessary material from the body is collected and thrown out by kidney and excretion system. There are approximately 1 million cases of Chronic Kidney Disease (CKD) per year in India. Chronic kidney disease is also called renal failure. It is a dangerous disease of the kidney which produces gradual loss in kidney functionality. CKD is a slow and periodical loss of kidney function over a period of several years. A person will develop permanent kidney failure. If CKD is not detected and cured in early stage then patient can show following Symptoms: Blood Pressure, anaemia, weekbones, poor nutrition health and nerve damage, Decreased immune response because at advanced stages dangerous levels of fluids, electrolytes, and wastes can build up in your blood and body. Hence it is essential to detect CKD at its early stage but it is unpredictable as its Symptoms develop slowly and aren't specific to the disease. Some people have no symptoms at all so machine learning can be helpful in this problem to predict that the patient has CKD or not. Machine learning does it by using old CKD patient data to train predicting model. Glomerular Filtration Rate (GFR) is the best test to measure your level of kidney function and determine your stage of chronic kidney disease. It can be calculated from the results of your blood creatinine, age, race, gender, and other factors. The earlier disease is detected the better chance of showing or stopping its progression.

Based upon GFR the renal damage severity by CKD is categorized into following five stages:



1.2 OBJECTIVES OF REASEARCH

Machine Learning is one such tool which is widely utilized in different domains because it doesn’t require different algorithm for different dataset. In medical science CKD is one of the major challenges; because a lot of parameters and technicality is involved for accurately predicting this disease. Machine learning could be a better choice for achieving high accuracy for predicting not only CKD but also another diseases because this vary tool utilizes feature vector and its various data types under various condition for predicating the CKD, algorithms such as Naive Bayes, Decision Tree, KNN, Neural Network, are used to predicate risk of CKD each algorithm has its specialty such as Naive Bayes used probability for predicting CKD, whereas decision tree is used to provide classified report for the CKD, whereas the Neural Network provides opportunities to minimize the error in prediction of CKD. All these techniques are using old patient record for getting prediction about new patient. This prediction system for CKD helps doctors to predict heart disease in the early stage of disease resulting in saving millions of life.

From this research we learn

* Data Collection
* Data Analysis
* Data Visualisation
* Exploring Machine Learning Models

1.3 PROBLEM STATEMENT

To classify whether the person is suffering from Chronic Kidney Disease using the previous data on CKD.

1.4 INDUSTRY PROFILE

The industry in which this is going to help is medical industry, that too mainly in the human renal system named as Nephrology and human pumping system known as Cardiologyas mentionedearlier if a person is predicted with Chronic Kidney Disease he has a higher chance of getting heart disease so this project will mainly focus on predicting whether the person has CKD and heart related disease

2.REVIEW OF LITERATURE

Chronic kidney disease (CKD) means your kidneys are damaged and can’t filter blood the way they should. The disease is called “chronic” because the damage to your kidneys happens slowly over a long period of time. This damage can cause wastes to build up in your body. CKD can also cause other health problems.

You are at risk for kidney disease if you have

* Diabetes. Diabetes is the leading cause of CKD. High blood glucose, also called blood sugar, from diabetes can damage the blood vessels in your kidneys. Almost 1 in 3 people with diabetes has CKD.
* High blood pressure. High blood pressure is the second leading cause of CKD. Like high blood glucose, high blood pressure also can damage the blood vessels in your kidneys. Almost 1 in 5 adults with high blood pressure has CKD.
* Heart disease. Research shows a link between kidney disease and heart disease . People with heart disease are at higher risk for kidney disease, and people with kidney disease are at higher risk for heart disease. Researchers are working to better understand the relationship between kidney disease and heart disease.
* Family history of kidney failure. If your mother, father, sister, or brother has kidney failure, you are at risk for CKD. Kidney disease tends to run in families. If you have kidney disease, encourage family members to get tested.

Your chances of having kidney disease increase with age. The longer you have had diabetes, high blood pressure, or heart disease, the more likely that you will have kidney disease.

African Americans, Hispanics, and American Indians tend to have a greater risk for CKD. The greater risk is due mostly to higher rates of diabetes and high blood pressure among these groups. Scientists are studying other possible reasons for this increased risk.

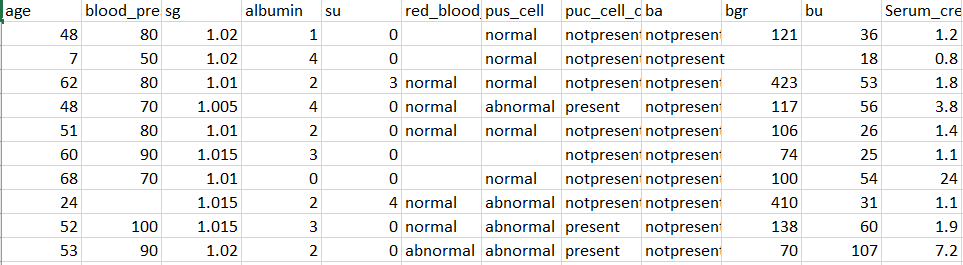
We use this real world cases and factors to determine our features which can be used for the machine learning and deployment of our project.

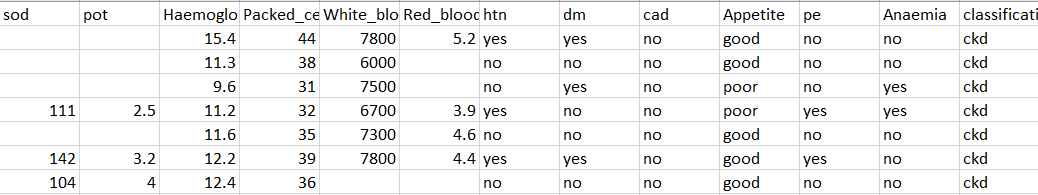
3.DATA COLLECTION

Dataset - Dataset of prediction of chronic kidney disease using machine learning algorithm is downloaded from UCI repository(<https://archive.ics.uci.edu/ml/machine-learning-databases/00336/>) ,Apart from this we can also get it from (<https://www.kaggle.com/mansoordaku/ckdisease>)

In that dataset there are 400 patient records are included. Also they include 25 attributes but we take only 12 attributes for building model. Age, Blood pressure, Albumin, Red blood cells, Pus cell, Puc cells clumps, Sugar Level, SodiumLevel ,Potassium Level, Coronary Artery Disease, Bacteria, Specific Gravity all this 12 attributes are used to build model.

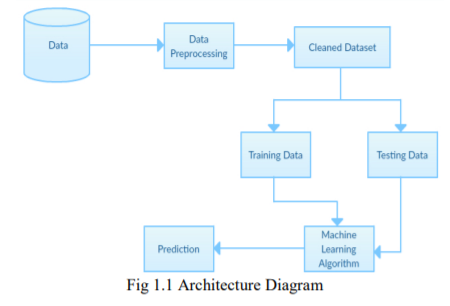
Wecan see the data set below which contains the needed features for predicting the condition of the patient.





The dataset contains 24 features we have selected twelve features from this dataset which are mentioned above. We replaced the Nan values with the nearest value for the numerical values and We dropped the Nan values in the categorical values and encoded them to 0 and 1 and We took the features in accordance with the correlation obtained by the heat map in the figure section

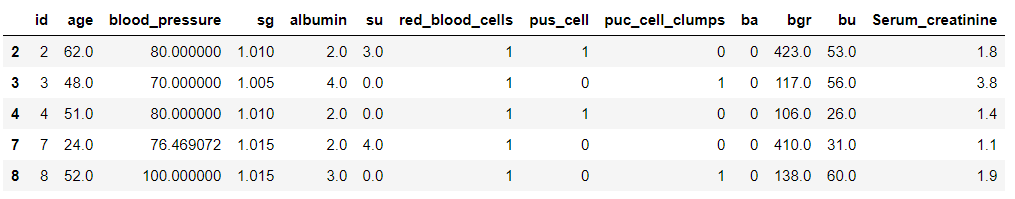
4.METHODOLOGY

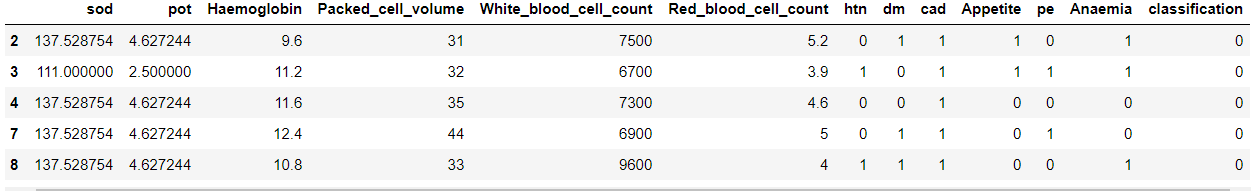


As we have discussed in the above section about the data pre processing and data cleaning are done now are in the part of splitting the data for training and testing ,we took 37.5% of the data for test and the other data of 62.5% was given to train the machine and the machine was trained and we used different machine learning algorithms like Logistic Regression, Decision Tree ,Random Forest, k-Nearest Neighbour, Support Vector Machine, Naïve Bayes and we plotted the of these techniques and we got the ROC-AUC graph as 4.1EXPLORATORY DATA ANALYSIS

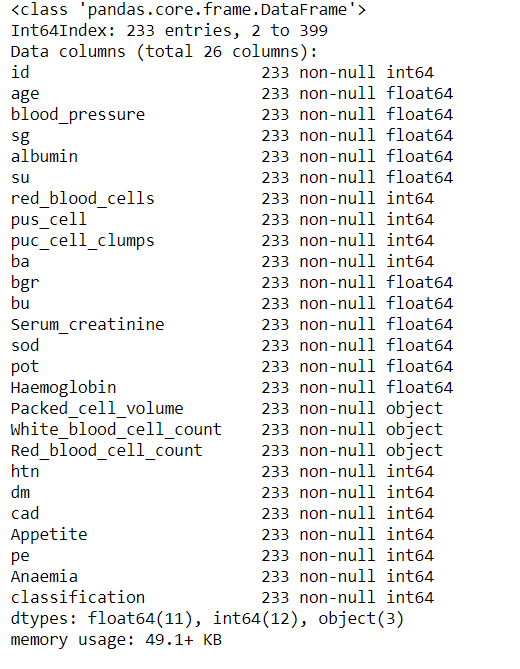
We explore the data using python function like head(),info(),describe().

4.1.1 FIGURES AND TABLES

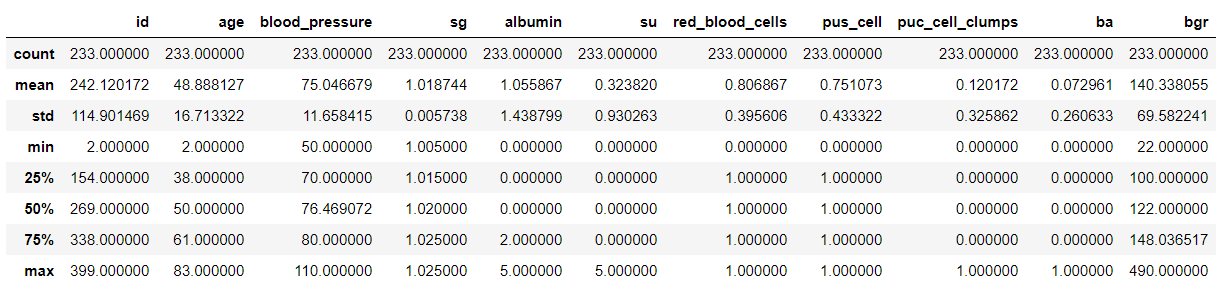


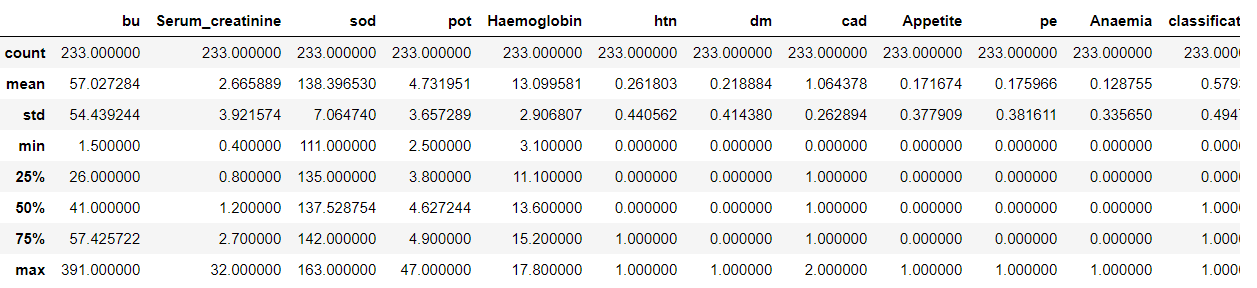


The above image show the head of the data set.



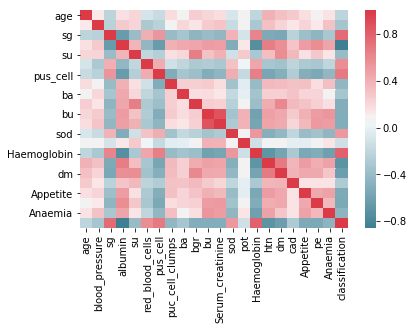
This is the information about the data set





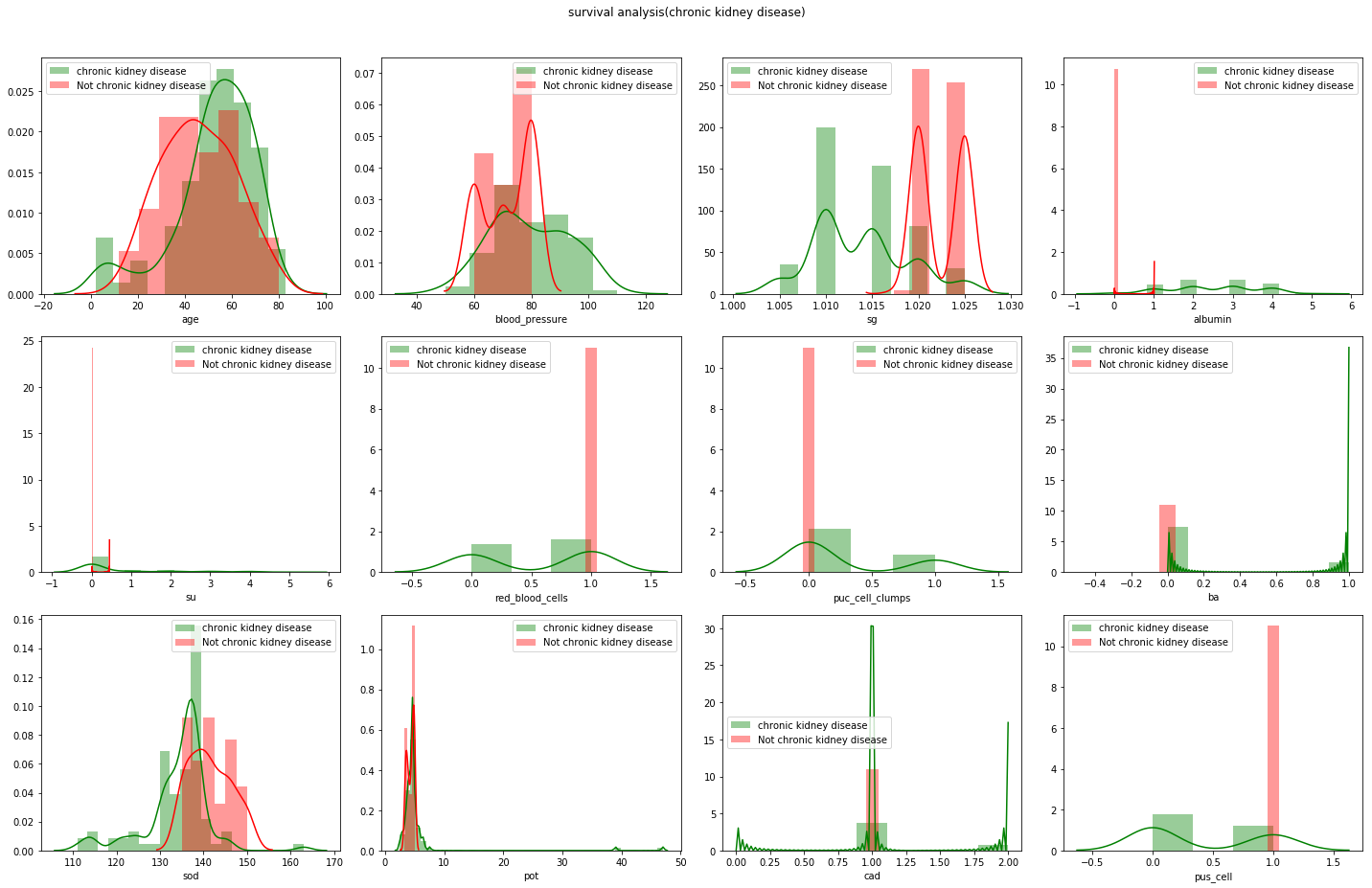
This is the description of the data set

4.2 STATISTICAL TECHNIQUES AND DATA VISUALIZATION



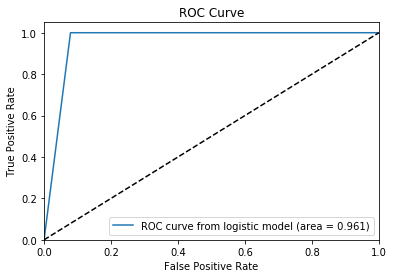
HEAT MAP USED FOR VISUALIZING THE CORRELATION BETWEEN THE FEATURES IN THE DATASET.

From this heat map we are choosing the features with p-value greater than are equal to 0.5, which means the positively correlated features since the features are linearly dependent on one another we can drop any of the features in the selection of features.

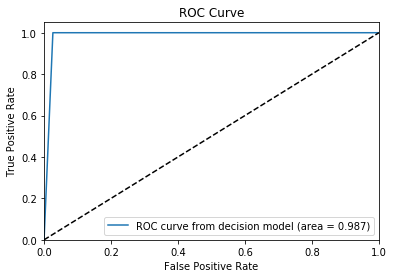


DISTIBUTION BETWEEN THE SELECTED FEATURES AND THE CLASSIFICATION FEATURE BETWEEN THE RANGE OF 0 AND 1

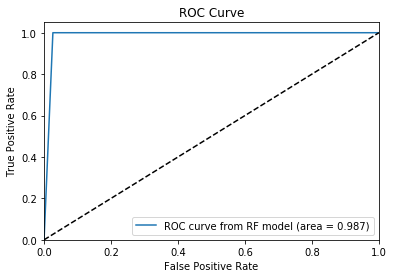
4.3 DATA MODELLING USING SUPERVISED ML TECHNIQUES



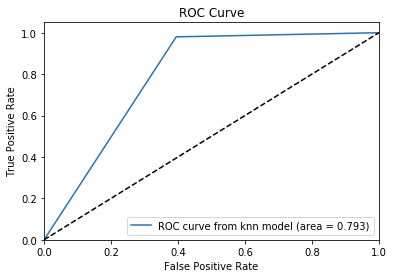
LOGISTIC REGRESSION



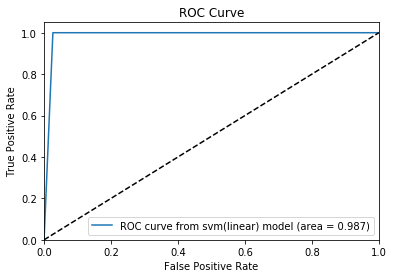
DECISION TREE



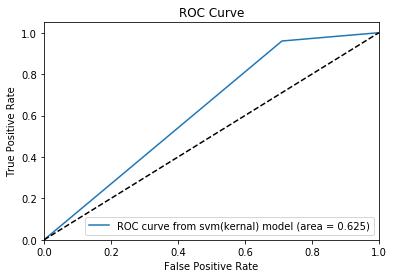
RANDOM FOREST



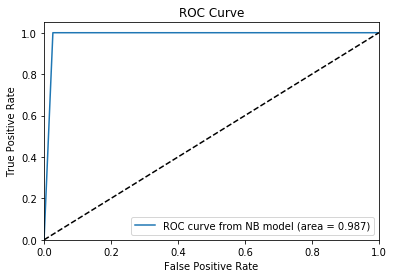
KNN



SVM(LINEAR)

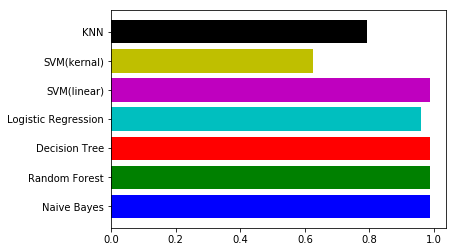


SVM(KERNAL)



NAÏVE BAYES

From the above ROC-AUC graphs we are plotting the bar graph between the areas of different classifiers.



BAR GRAPH OF AREAS OF CLASSIFIERS

From the above graph we can easily select the maximum are among the classifiers which is the area of Naïve Bayes.

5.FINDING AND SUGGESTION

From the above graph we have selected Naïve Bayes algorithm for the prediction ,we found that the CONFUSION MATRIX with True Positive to be 37, False Positive to be 1, False Negative to be 0 and True Negative to be 50 and the PRECISION to be [1.0 0.987] , RECALL to be [0.97363 1.0] ,

F-SCORE [0.9866 0.99009] and SUPPORT [38,50]

We suggest to use SVM(linear) as a optional algorithm as a replacement of Naïve Bayes, since it has similar accuracy score.

6.CONCLUSIONS

We have analysed 12 different attributes related to CKD patients and predicted accuracy for different machine learning algorithms like Logistic Regression, Decision Tree ,Random Forest, k-Nearest Neighbour, Support Vector Machine, Naïve Bayes. From the results analysis, it is observed that the Naïve Byes algorithms gives the accuracy of 98% . The advantage of this system is that, the prediction process is less time consuming. It will help the doctors to start the treatments early for the CKD patients and also it will help to diagnose more patients within a less time period. Limitations of this study are the strength of the data is not higher because of the size of the data set and the missing attribute values. To build a machine learning model targeting chronic kidney disease with overall accuracy of 99.99%, will need millions of records with zero missing values.

REFERENCES

1. <https://www.kaggle.com/mansoordaku/ckdisease>
2. International Journal of Advanced Research in Computer and Communication Engineering(Prediction of Chronic Kidney Disease Using Machine Learning Algorithm).
3. Wikipedia.