PREDICTING AUTISM IN CHILDREN

Project report submitted in partial fulfillment of the requirement
for the degree of Bachelor of Technology
in

Computer Science and Engineering/Information Technology

By

Shubham Kansal (151309)

Under the supervision of
Dr. Geetanjali
Assistant Professor (Senior Grade)

To

Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology Waknaghat, Solan-173234, Himachal Pradesh

Candidate’s Declaration

I hereby declare that the work presented in this report entitled **“ Predicting Autism in Children”** in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering/Information Technology** submitted in the department of Computer Science & Engineering and Information Technology**,** Jaypee University of Information Technology Waknaghat is an authentic record of my own work carried out over a period from August 2018 to May 2019 under the supervision of **Dr. Geetanjali** (Assistant Professor).

The matter embodied in the report has not been submitted for the award of any other degree or diploma.

Shubham Kansal (151309)

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

Dr. Geetanjali
Assistant Professor (Senior Grade)
Department of Computer Science & Engineering and Information Technology
Dated:

Acknowledgement

We express our profound gratitude and indebtedness to **Dr. Geetanjali**, Department of Computer Science & Engineering and Information Technology **Jaypee University of Information Technology** Waknaghat for introducing the present topic and for their inspiring intellectual guidance, constructive criticism and valuable suggestion throughout this journey.

We are also thankful to all other faculty members for their constant motivation and helping us bring in improvements in the project.

Finally we like to thank our family and friends for their constant support. Without their contribution it would have been impossible to complete our work.

Date:

Shubham Kansal

TABLE OF CONTENT

ABSTRACT

Our project aims at designing such a system which can predict the autism of a person given the reports of doctor with symptoms. In this project, three different models such as Logistic Regression, Support Vector Machine and Naive Bayes(through weka) have been investigated for the sole purpose of predicting autism. The data set consists of doctor reports whether the patient is autistic or non autistic . The features include the prediction of the autism given doctor reports and the symptoms from which they are suffering like fever,surgery,age,speech etc.

In this project, we have studied several different ways of forming up input data sequences, as well as different architectures that may lead to effective prediction of autism.

Using the proposed Machine Learning Model, we show that the f1 score and accuracy score of the logistic regression are 0.93 and 0.93.

Our test outcomes have demonstrated that Machine Learning might be utilized for effectively foreseeing the autism in a person. For further expanding the execution of the anticipating calculations, earlier data about every patient would be alluring and the parameters of SVM and logistic regression could likewise be tuned

1. INTRODUCTION

1.1 INTRODUCTION

In this rising above condition of the Internet and its advances, the human personality has its own inclinations for testing and intriguing new thoughts and its execution. Purchaser items have been one of the real territories where we have seen some fascinating development and improvement. Some other vague advancement in the Internet world has additionally been at the position where they have pulled their socks up and sown their seed for achievement regarding transformation with the time, and obviously the potential has been demonstrated with regards to well being concerns.There are incalculable conceivable outcomes of advances done by the web of how it decides the glucose of a patient ,as of late it was discovered that the most precise method for discovering circulatory blood pressure is through various sensors and with utilization of Internet .This is designated "Internet of Medical Things"

One of the greatest transformation in "Internet of Medical Things" has not just put the financial examples of overcoming adversity in the codex of different expert specialists yet additionally advanced the organization on the way to self important triumph. It likewise helped medical industry to make it so natural to accomplish something that set aside individuals a group of effort to manage without it. A portion of its exercises like Statistical computation of all patient in a composed manner was the best age old problem for the general population back in the time before the Internet and when it came, everything appeared to be so easy and great.

With the progression of time, there are some gigantic enhancements as far as Radio Frequency Identification, Machine Learning approach has been acquainted with lower the expense i.e they can speak with frameworks through internet.To outline the previous decade as far as "Internet of Sports", it has risen above exponentially, as far as speed and furthermore for precise expectation in the well being business, which will be our essential worry for project implementation.

The first key goal of any prediction forecasting is exactness and accuracy. Regardless of whether its business, climate, sports or whatever other real zones where the forecast of results is the main undertaking. Furthermore, with the assistance of the Internet which has been utilized as a device in market expectations to gauge future occasions, our key targets are being watched and the rightness is rising above as time passes. A portion of the organizations like Microsoft, Google, IBM and so forth have their very own forecast apparatuses for organization explicit outcomes. This showing up business sectors is predominantly having its development because of their accuracy, in every one of the fields of looking over market development, which is made increasingly proficient with the accessibility of huge informational indexes.

The patient expectation has utilized as a methods for exactness, yet additionally helps for yearly appraisal of patients and the related outcomes. Moreover, unfurling basic leadership of particular specialists and staffs, money related accomplishment by developing income of emergency clinics are additionally discharged.

In spite of the fact that, individuals hold various opinions with respect to internet medical services, however paying little heed to these postulation of shifting sentiments, the fame of this has been augmented to the uttermost dimension.

1.2 PROBLEM STATEMENT

The concern of this report is to predict the autism influence around the globe. To examine the consequences of autism on distinct users.People usually delay their diagnosis as a result therapies will also get postponed. All these problems collectively affect the predictions of the disease and are against the legal standards.

Our alternate motive is to promote the concept of machine learning in predicting the heath concerns on time which is the one of the major problem nowadays i.e. secure transaction of resources without losing our hands on performance and precision by thorough examination of risks, particularly to some of the ill-protected groups and organizations. Data is gathered from “Vaccine adverse event reporting system”(VAERS)

1.3 OBJECTIVE

We will develop an online web application to predict outcomes of autism diagnosed in patients, in addition of assess and implementation of various models of machine learning. We will use a step-by-step approach to analyze the data of past decade. The predicted data of patient with autism and non autism will be observed as new data and will be used for observing results for forthcoming patients. Once the former has been implemented, we will extend our project to showcase some advanced features so that accuracy will be maintained.

The selection highlights will be executed and utilization of models like Logistic Regression, SVM machine, Naive Bayes will be finished by us. Moreover, some propelled devices will be in thought to build the execution and security of the application.

The estimate models usage will be tested and utilized for result forecasts. What's more, the entire structure of the undertaking will be execution and precision driven by utilizing vast measure of informational indexes and stable information outlines.

The remaining of the report will be discussed as follows:

1. Literature Survey ,work analysis and Methodology of the prediction markets.
2. Mathematical and analytical approach used for Systematic Model development .
3. Algorithms implementation.
4. Metrics, Data Sets as a part of Test plan used for the project.
5. Result accuracy analysis and Performance.Calculating accuracy of data prediction along with stable framework environment for the project.

The salient features of all the models and algorithms implementation for future research use will be discussed.

1.4 METHODOLOGY

The base of our project concentrate on the doctor reports and symptoms through which a patient is suffering.First task is to preprocessed the data so that various classifier will be implemented easily and accurately.

Data consist of several CSV files from year 2000-2007 with 30000 instances every year . Various attributes are present in the data.First attribute is patient_id which is unique for every patient . This data set will likewise improve the expectations by radiating increasingly steady and exact outcomes to give greater plausibility of anticipating precisely.

Despite the fact that the high volume informational indexes may cause a few information irregularity, however for the long run, it will be effective. Besides, because of interest of moment engendering of information and data which is satisfied by the critical piece of innovation, there will be a greatness in part of information investigation by utilizing instruments of AI and precise informational collections.

1.5 ORGANIZATION

The whole structure of project has been carved up in sequential approach. There will be detailed overview of the project with the usage of real time data sets featuring the data from (2000-2017) of patients. An Exploratory and spiral analysis of the data will be thoroughly studied to acknowledge the behavior of patients as time passes on.

Besides, we will propose different instruments and prerequisites to execute the wellbeing methodologies related with the datasets. Remedial expectations is our key component and point of view like when to investigate and the amount included will be evaluated by the resultant of our application.

Chapter 4 Algorithm implementation and reports will be implemented.

Chapter 5 Documentation of datasets , analysis, Metrics etc. which will be developed by us.

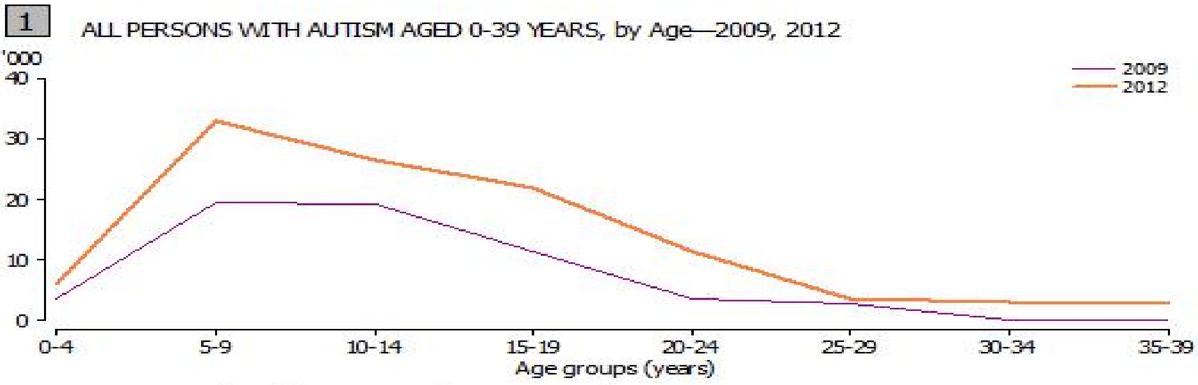
The following parts will chiefly manage result and execution investigation by top to bottom examination of our undertaking project. A powerful scientific and diagnostic calculations will be finished by exuding various sources of info and the examination will be done based on conduct of the outcome.

2. LITERATURE SURVEY

"Because of its growing vogue, the online health applications has turned into an exact and fascinating approach to foresee. With the dynamic scope of health concerns, the online health expectation has been an extension among them and furthermore a business achievement. In any case, with fluctuating assorted variety, the perspective of individuals is additionally getting to appear as something else. Some consider it as exact and an incredible side think it as an exercise in futility, a great way numerical insights contemplate improvement for long run while numerous other think about it as a piece of fake forecast and are against it."

"Not just online health prediction applications are progressively precise, however it likewise gives comfort and results on schedule. The dynamic choice to enter the sum amount you need to be will make the application progressively appropriate to all individuals. Exactness and security will be finished by it and health option will give online health applications a high ground.

Yet, the main thing matters that corrupts the structure of this procedure is legal and administrative issues which change far and wide. This issue put an immediate effect on individuals who may be keen on doing it."



Source: ABS Survey of Disability, Ageing and Carers, 2009 and 2012

Some research paper studies showed that the current method for diagnosing the autism is very tedious.The algorithm used is ADOS-G(gold-standard Autism Diagnostic Observation Schedule-Generic).The drawback of this method is length of complexity which is reduced 72% by machine learning classifier.Due to lengthy procedure all populace will not get this treatment and ultimately disease will not cured properly.Behaviour and language imperfection are the main cause of autism.People are unable to interact with social environment and the mental growth will not be done with age.

"Online health application has just bloomed as an exponentially expanding business sector industry that deals with a multi-billion dollar resources independent of its lawfulness, regarding precision as well as giving the steady economy to the states. It has capacity to make open doors for the American People by giving occupations and will build the all out income for state by billions in the following 7-8 years. However, it won't make the illicit wagering evaporate from sight. Not just this, the complete embodiment of specialist will move from health focused to cash arranged musings and that what health isn't about."

RESEARCH PAPER I:

(http://fadifayez.com/wp-content/uploads/2017/11/Autism-Spectrum-Disorder-Screening-Machine-Learning-Adaptation-and-DSM-5-Fulfillment.pdf)

In this report,there are various tools to diagnose the autism in a person.The current method used DSM-IV tools to detect autism,But there are various debates going on to replace DSM-IV with DSM-V.It was clear from this report that there was no clear way to predict the autism.The method which is on paper is ADOS-G which is known for its complexity and lengthy procedure,so there will be a need for machine learning classifier techniques for prediction.

By using some forecasts based on historical data sets new benchmarks were introduced, DSM-IV as a primary benchmark, with usage of DSM-V ranking of patients as a second benchmark. To compare the accuracy of forecasting to predictions from world rankings, calculations of percentage of correctly predicted autism were done and were stored as hit rate which was nearly 33.33% for random draws.

RESEARCH PAPER II:

(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4463758/)

The primary objective of the study was to assess the accuracy of online health prediction classifier,observation based classifier. The impact on technology was thoroughly assessed and errors of unguarded segments were determined. This whole study was aimed to show the accuracy of health prediction.

Social Responsive Scale(SRS)is calculated for all the patients according to their behaviour and how they respond for the social activity.SRS if >90% high sensitive,if ~8% then very low specificity.SRS is compared with OBC and then accuracy is calculated with the comparison of these two.There are other ways also to calculate accuracy.

The survey was being done on all classes of individuals from teenagers to older and their conduct were completely taken note. The outcomes were put something aside for the evaluations of social economics vulnerabilities with respect to smoking and ensuing issues.

For research techniques, scientist concluded a general intend to get the resultant choices of individuals. Plans included different elucidating contemplates, with longitudinal research structures. Utilizing clear research structure, the pattern was anticipated of games wagering development and its impact on unmistakable gatherings.

Different strategies utilized were quantitative methodology, to evaluate information in numerical amounts. These techniques helped specialists to contemplate the factual quantifiable factors.

Sampling structure classes like Frame, Technique and Size were alluded to decide the example of populace, the determination of test to be watched and the safety buffer resilience in the chose test. This examination utilized different recipes to think of sufficient example measure

Other than those, various data retrival methods were introduced to amplify collections of qualitative as well as quantitative data.

AUTISM TEST APP

(F. (2017). ASDTests. A mobile app for ASD screening. www.asdtests.com [accessed December 20th,2017])

This app was designed for evaluation of all the health autism systems taken in observation. These autism systems were evaluated on the basis of real-life entities like behaviour plans and social strategies and the most optimum solution was taken into account.

The evolution of contemporary health prediction was clearly written in detail. There was exploratory analysis on health market and the study was shown about how gross revenues were made in billions of dollars with shares of health prediction exceeding 43% of all online revenue.

Various behaviour plans were briefed for performance assessment of health prediction systems and one of best simulation tool used for observation “Autism test App” was used to test the performance of all online autism applications by using different combinations of behaviour strategies and plans used.

Results demonstrated gave the proof that the strong choice procedure and staking optimization are just the prime variables for a tuned health expectation application. These elements can possibly create benefits inside the constrained measure of time in perfect conditions.

Be that as it may, for genuine circumstances, other optional variables like probability level and questionable time spans comes set up especially to the particular situation of health applications. By and large, the above technique executed is significant for deciding expanded benefit and lessening hazard components of determined health framework.

RESEARCH PAPER III

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6714480

In this report, various techniques were implemented such as approximate similarity index (sim(APK,PPK)) and usage of different architectures were implemented along with proper study of sequences of data that has to be input.

TF-IDF architectures, which are part of sim(APK,PPK) were thoroughly tuned and tested for effective predictions and observations when some robust test cases were used. Strategies such as ‘many-to-one’ or ‘many-to-many’ were executed where the outcomes accuracy were more than 98% and 88% respectively. The classification accuracy of health outcomes was also increased by feeding previous sequence of data about a patient to the network. Furthermore, some other factors like prior symptoms of patients are weighted and can be given for higher precision in outcomes.

The report mainly aims for simple fundamentals of Machine learning and Neural Networks.The main crux behind this report is to assign weight to different keywords according to their strength in a doctor report or in a sentence,For eg A hotel review system is used to determine whether the hotel is good or bad(ratings).By calculating the similarity index and TF-IDF ratio it will be possible to assign weigths to a specific word in a report.

Mainly GPUs and CPUs were prime factors to achieve mathematical computations and parallel computations (GPUs like AMD and NVIDIA) along with some vast software libraries provided to simply the processes.

3. SYSTEM DEVELOPMENT

Our project mainly focuses on the prediction of autism in patient with symptoms . So for it, we need a structured approach which is more of practical basis than theoretical one. The framework in use is CRISP framework (cross industry process for data mining), a accurate methodology for structural approach to predict the outcomes.

CRISP-DM FRAMWORK:

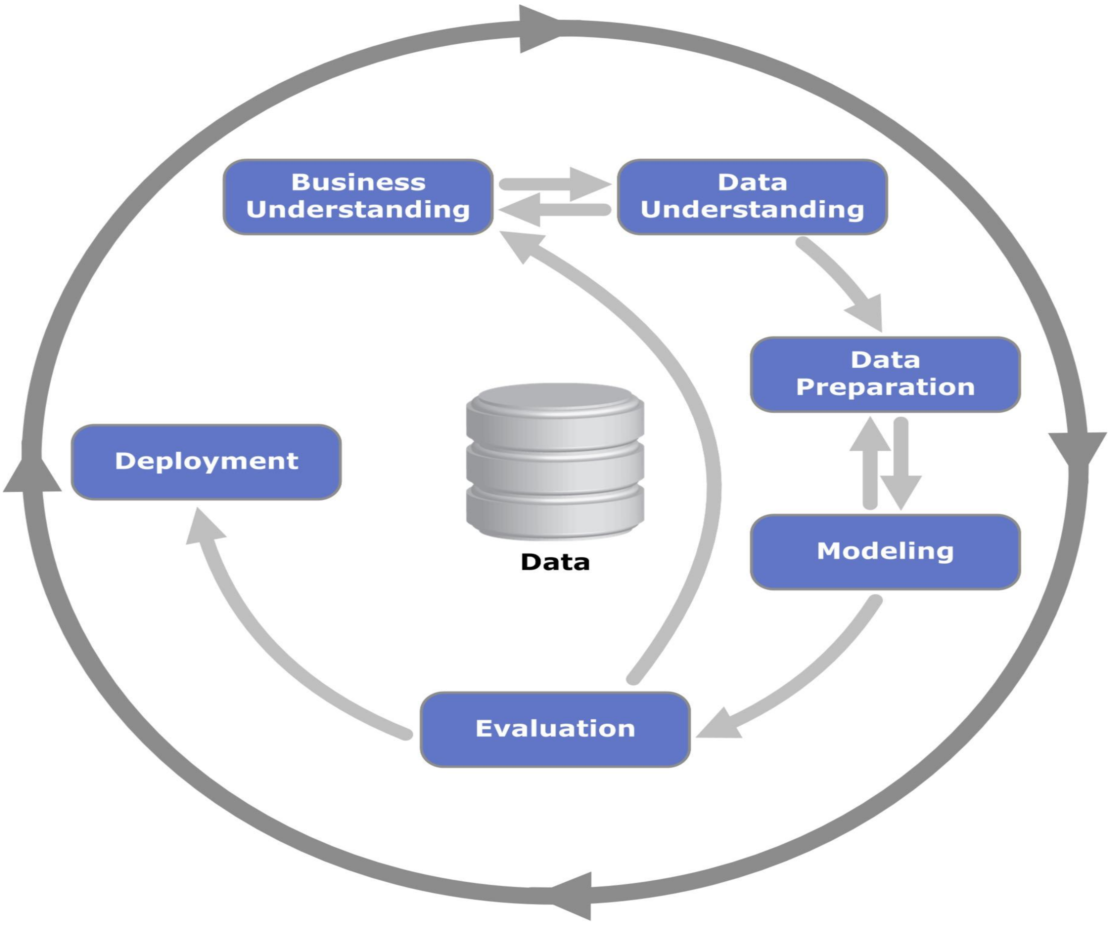


Figure 3.1

CRISP-DM structure has been subdivide into six major stages. These stages need not to be all together and can be utilized flexibly. Arrows, in the outline speaks to some sort of conditions between the stages and not the succession of execution. The external circle shoes the ending perception of datasets for data mining, which is regularly used to study and improve execution just as structure of framework.

In spite of the fact that, these stages are flexible to execute, the means actualized in these stages are requested and are proposed in a particular way. Following are the means of our proposed data mining system.

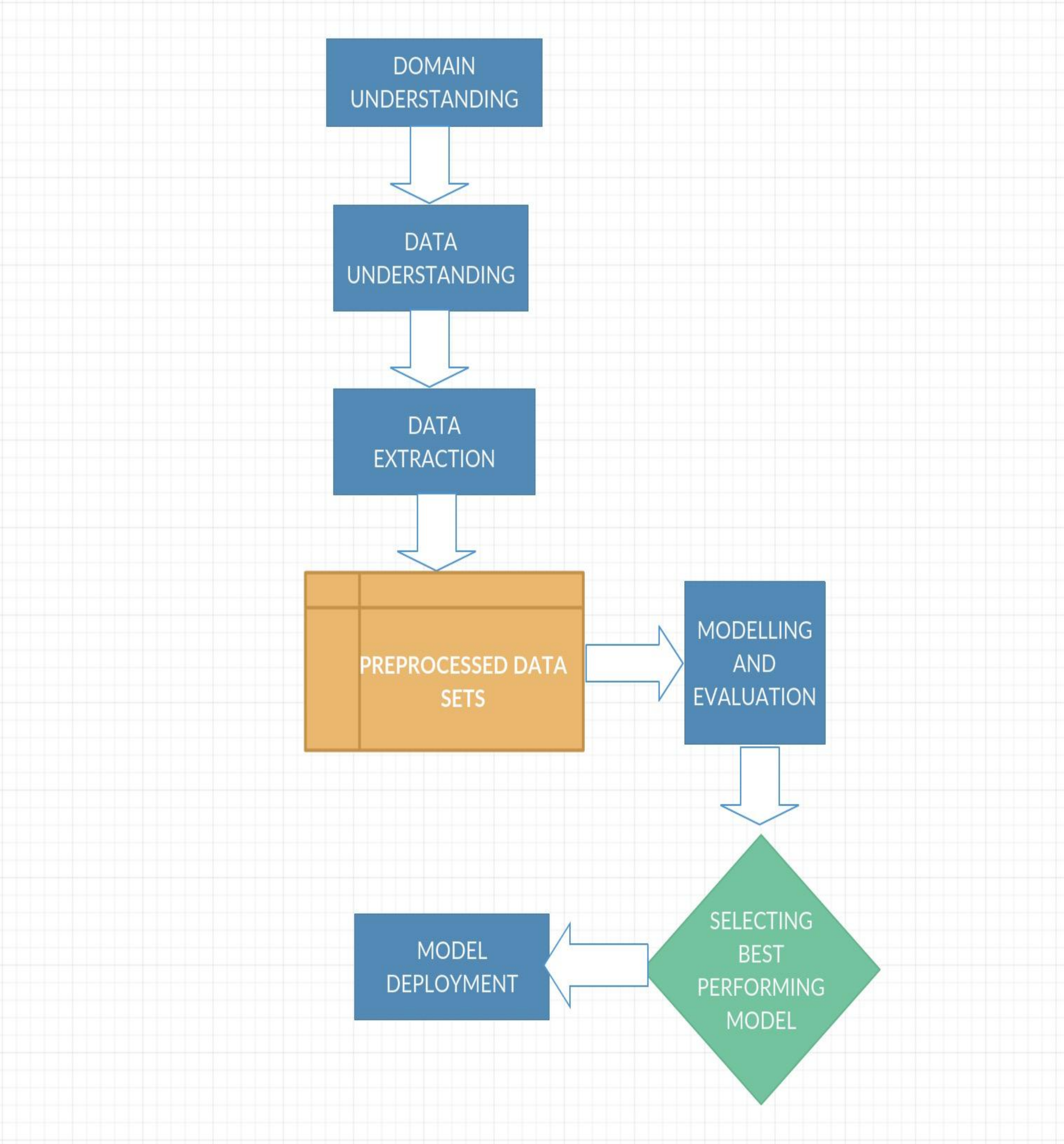


Figure 3.2

3.1 DOMAIN UNDERSTANDING

Getting issues and the target of the issue is the key stage in this part. How an symptom is organized, its effect and what are the variables incorporated into foreseeing the result is resolved. The references by which space grasping works can be additionally broadened by means of individual information of the particular literature or evaluating the writing and research papers.

3.2 DATA UNDERSTANDING

All of the information got can be appreciated by means of accessible assets. These assets may have some earlier information that is mechanized and extricated on the web and loaded into certain databases. This procedure can be additionally moved forward by utilizing end UI in which client can include explicit information and get the anticipated outcome. For every API in machine learning classifier,data is utilized with various spaces.

3.3 DATA EXTRACTION

In this phase, there is creation of featured subsets. These subsets can be patient-related or doctor standings. Some researchers also split the features into subsets such as odds or value provided by expert opinion. But we will mainly focus on patient features and external features which excludes the expert opinion.

Patient-related Features will mainly deal with the arithmetical calculations such as behaviour analysis, symptoms and so on, while external features will determine the analytical calculations like recent behaviour of a patient.

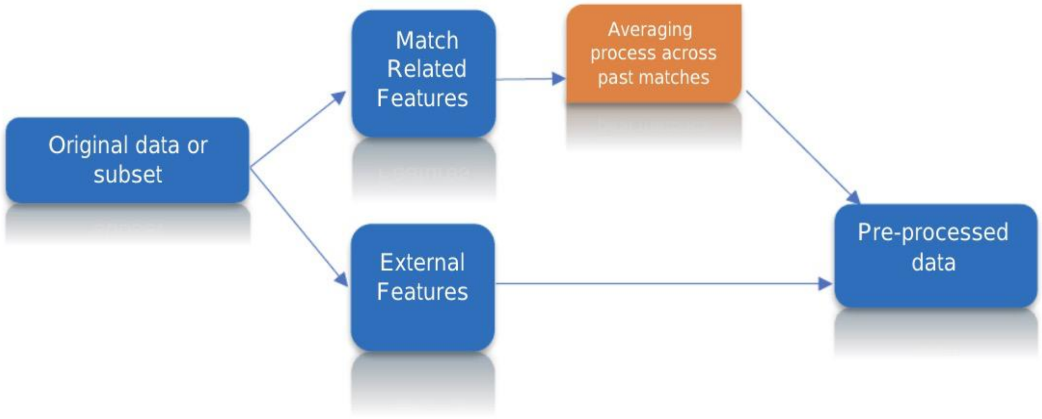


Figure 3.3

3.4 MODELLING AND EVALUATION

By checking on the past literature, diaries, gatherings and connected prescient models that were effective that time, the choice of competitor will be accomplished for the experimentation.

This procedure will additionally recognize the blends of highlights and classifier systems.

For Evaluation purposes, Model execution will be estimated using predefined data collections and a standard grouping framework, this procedure of assessment is best for the information which is adjusted, if the information is exceedingly imbalanced, we will utilize the idea of curve assessment called Receiver Operating Characteristic (ROC).Because the forthcoming patients are anticipated on the premise on past patients, there must be a request protection for the preparation set. For this, we can utilize cross approval procedures which will rearrange the request all things considered. Other than this, we can use WEKA g which is a test suite of AI for example request preservation.Training set includes 66% of information and the rest is for test set.

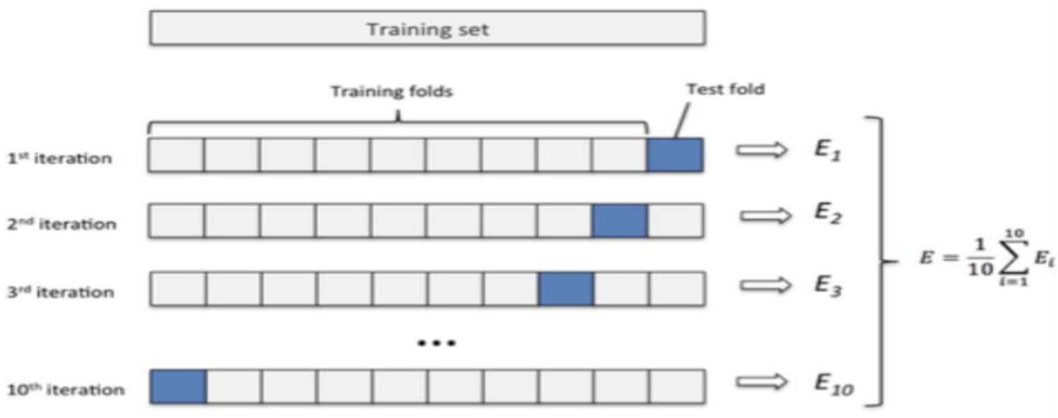


Figure 3.4

3.5 MODEL DEPLOYMENT

Change of training set and test set is done, and with the mechanized procedure the new information is acquired and added to the database either physically by end-client or naturally in coordinated one. With the computations, new preparing sets are made with new forecasts. The outcomes are returned back to the end client. The learning model is additionally refreshed with the preparation set persistently with some time period and must get input information powerfully.

4. ALGORITHMS

When it comes to to anticipating the result of health an expansive number of factors must be mulled over and along these lines it requires a calculation which can relate every one of the factors in a way to such an extent that the result that we get is the ideal one. For the particular issue of foreseeing an outcome given a lot of factors including the past patient measurements one could utilize Artificial Neural Networks as well. Deep learning calculations require an extensive number of preparing precedents, else they won't most likely give great outcomes. In health expectations the quantity of highlights accessible is way a lot bigger than the preparation models that are available. For this issue the data present comprises of past patient measurements of the considerable number of visits containing the one of a kind id which is assign in each visit and other related characteristics.

The main thing required for expectation is the distinguishing proof. The recognizable proof of the learning models to be utilized, the data sources, the required methods for model assessment lastly the particular difficulties that could impede the prediction results.

Two types of Machine Learning algorithms are:

1. **Supervised Learning Algorithms**
2. **Unsupervised Learning Algorithms**

As the name proposes, in supervised learning the entire learning procedure of the model is managed by a specific arrangement of information sources and their comparing yields. While, in unsupervised learning the information is first gathered and the understanding is done dependent on the info information. What's more, for the issue of health forecast administered learning is utilized as it utilizes the aftereffects of past season matches. The past outcomes and comparing insights fill in as the data yield pair for the regulated learning models utilized for the prediction of autism among patients.

Therefore, the three Machine Learning Algorithms that are being used to predict the outcome of a autism among patients are as follows:

1. **Logistic Regression**
2. **Support Vector Machine (SVM)**
3. **Naive Bayes(through Weka)**

4.1 Logistic Regression

Logistic Regression is essentially a statistical model which utilizes a logistic function. This logistic function is in charge of displaying a twofold reliant variable. As far as Regression Analysis, the logistic regression can be thought of as a type of binomial regression, it is utilized for evaluating the parameters of a logistic model. In numerical terms, a binary strategic model comprises of a needy variable which can has values 1 or 0 and a marker variable is utilized to demonstrate these qualities. The logarithm of chances is a direct blend of at least one autonomous factors for the variable esteemed as 1. The free factors can likewise be either continuous or binary themselves where double methods a blend of two classes and coded by a pointer variable that is utilized to show the esteem put away in the needy variable. The logarithm of chances utilizes the unit logit for example logistic unit for the estimation purposes

More or less, the Statistical Regression is utilized to compute the probability of event of an occasion and it does as such by fitting the information to a logistic curve. A few factors known as the indicator factors are utilized by the Logistic Regression model and these factors may either be numerical or categorical. For e.g., the likelihood that an individual shows some kindness assault inside a predefined time frame may be anticipated from learning of individual's sure characteristics like the individual's age, sex, weight file, and so forth. Logistic Regression is broadly utilized in the therapeutic and sociologies just as advertising applications, for example, forecast of a client's penchant to buy an item or stop a membership..

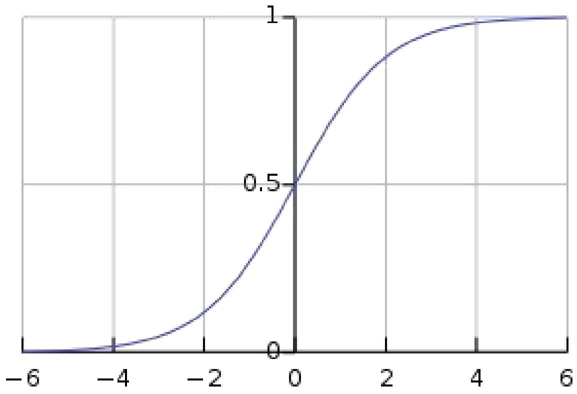
4.1.1 Logistic Model

For understanding the Logistic Regression, initially a logistic model with some given parameters must be viewed as then it is seen that how the coefficients could be assessed from the given data. Presently we should think about a model with two factors, x1 and x2. These two factors can either be indicator capacities for the comparing paired factors or ceaseless factors. The log-chances or the Logarithm of chances (meant by l) can be composed as underneath:

$$l = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Logistic Regression can likewise be clarified by clarifying the standard Statistical Function. The Logistic Function is fundamentally a sigmoid capacity, which takes any genuine number (t) with the end goal that t has a place with the arrangement of Real Numbers and in this way, it gives yield somewhere in the range of zero and one. The logistic function is characterized underneath:

$$\sigma(t) = e^t / (e^t + 1) = 1 / (1 + e^{-t})$$



The graph of statistical function on interval *-6 < t < 6* is shown in the figure 4.1.

Figure 4.1

Here, *x* is an variable therefore, let's assume *t* is a linear function of the variable *x* hence, we can express *t* as follows:

$t = \beta_0 + \beta_1 x$

Therefore, the logistic function now becomes:

$p(x) = 1 / (1 + e^{-(\beta 0 + \beta 1x)})$

The thing to note here is that the variable ***p(x)*** has been interpreted as the occurence of the dependent variable and it tells us about the true and the case instead of telling us about failure.

The inverse of the logistic function can also be defined as

$g(p(x)) = logit\ p(x) = \ln(p(x) / (1 - p(x))) = \beta_0 + \beta_1 x$:

also, after exponenting the both sides we get,

$p(x) / (1 - p(x)) = e^{\beta 0 + \beta 1x}$

4.1.2 Odds

After the exponential function of the linear regression expression has been discovered then the chances of the specific dependent variable can be said to be comparable to the exponential capacity. The two properties, likelihood and the mathematical regression expression can be connected with the assistance of a specific function and logit is the capacity that does this connecting. Changing over logit capacity to chances is very simple and going between negative to positive limitlessness it can likewise give the basis whereupon the Logistic Regression Algorithm can be applied.

Mathematically, odds of a dependent variable can be defined as follows:

$$Odds = e^{\beta 0 + \beta 1x}$$

Also, for an independent variable the ratio is as follows:

$$OR = odds(x + 1) / odds(x) = e^{\beta 0 + \beta 1(x + 1)} / e^{\beta 0 + \beta 1x} = e^{\beta 1}$$

4.2 Support Vector Machine (SVM)

The Support Vector Machines (SVMs or likewise the Support Vector Networks) are supervised learning models. With SVMs are convinced learning calculations and these calculations can examine the information for two purposes, for example Regression analysis and Classification. The SVM model arrange the given training precedents into various classifications and the learning model is in charge of putting new precedents in a single class or the other as indicated by their characteristics. What SVM model does is that it changes over the given examples into focuses in space and plots them on the diagram and arranges them as per their characteristics in a such a way, that there is a reasonable and as wide as conceivable hole between various classifications. The new approaching precedents are then mapped to the specific class without influencing the others.

Alongside the linear classification, the classification of non-linear information can likewise be performed by SVMs by utilizing Kernel Trick, that certainly maps the data sources high-dimensional space.

Essentially, in SVM calculations every data item is plotted as a point in n-dimensional space (where n is the quantity of highlights accessible). Each coordinate speaks to the estimation of the comparing highlight. Henceforth, the grouping is performed by making a hyper-plane that separates between two distinct categories great.

Figure 4.2 speaks to a n-dimensional plane with two unique classes having been separated by a hyper-plane.Support Vectors are the directions of every data point.

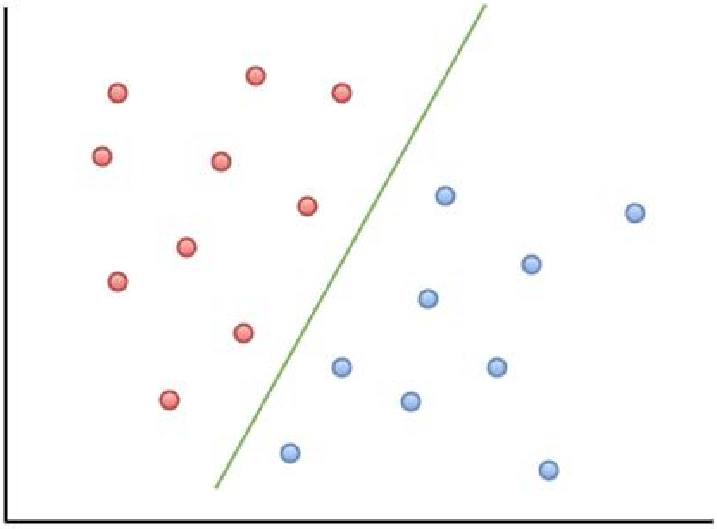


Figure 4.2

4.2.1 Hyper-planes

Different parameters need to be differentiated using a hyper-plane but choosing the right hyper-plane is the difficult task (not so difficult though) and therefore, it has to be chosen wisely.

To understand better:

- **Scenario – 1:** To differentiate the different classes better the hyper-plane ‘B’ will do a better job than any other given hyper-plane.

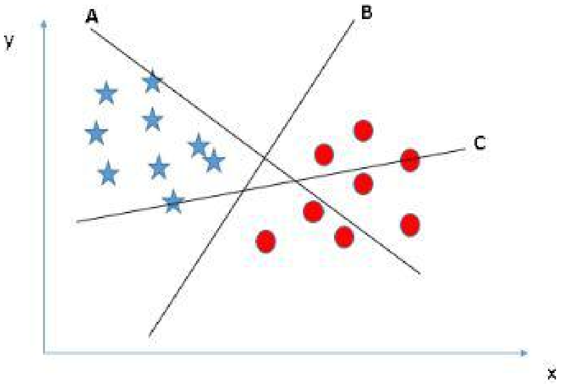


Figure 4.3

- **Scenario – 2:** From the following three hyper-planes 'C' will be chosen as the appropriate one as it separates the two different classes with the most gap with each of the class.

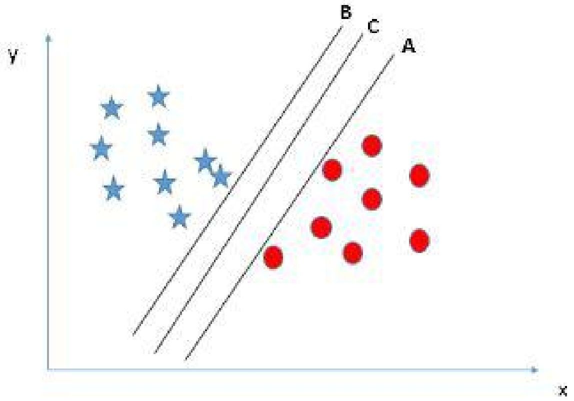


Figure 4.4

- **Scenario – 3:** The following classes can't be segregated with a simple linear hyper-plane therefore; a non-linear hyper-plane has to be chosen.

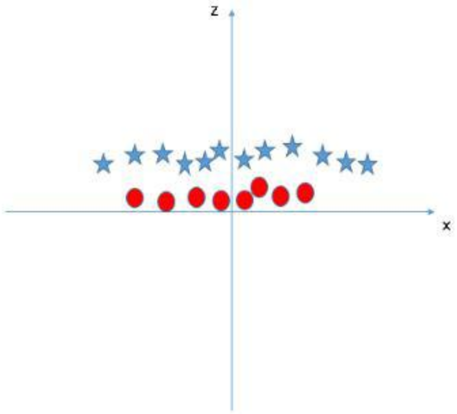
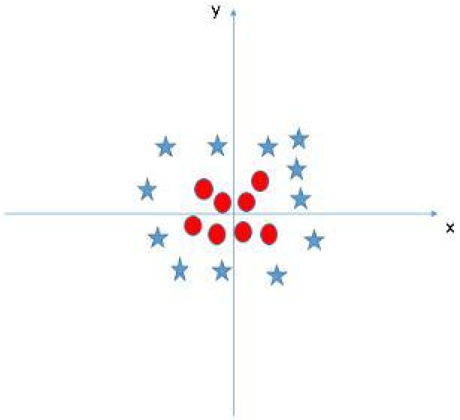


Figure 4.5

SVM has the ability to solve this problem too and it does so by adding a new feature to



the plane i.e., the z-axis where, $z = x^2 + y^2$

Figure 4.6

SVM uses Kernel trick to solve this problem and hyper-plane looks like this :

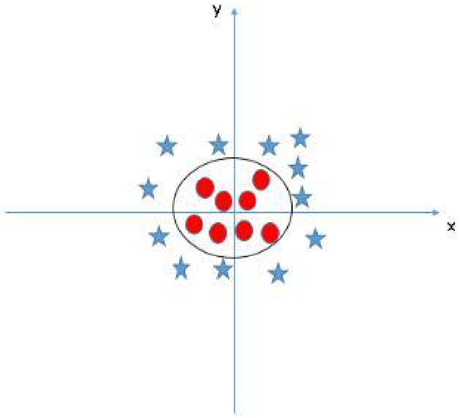


Figure 4.7

4.3 Naive Bayes(through weka)

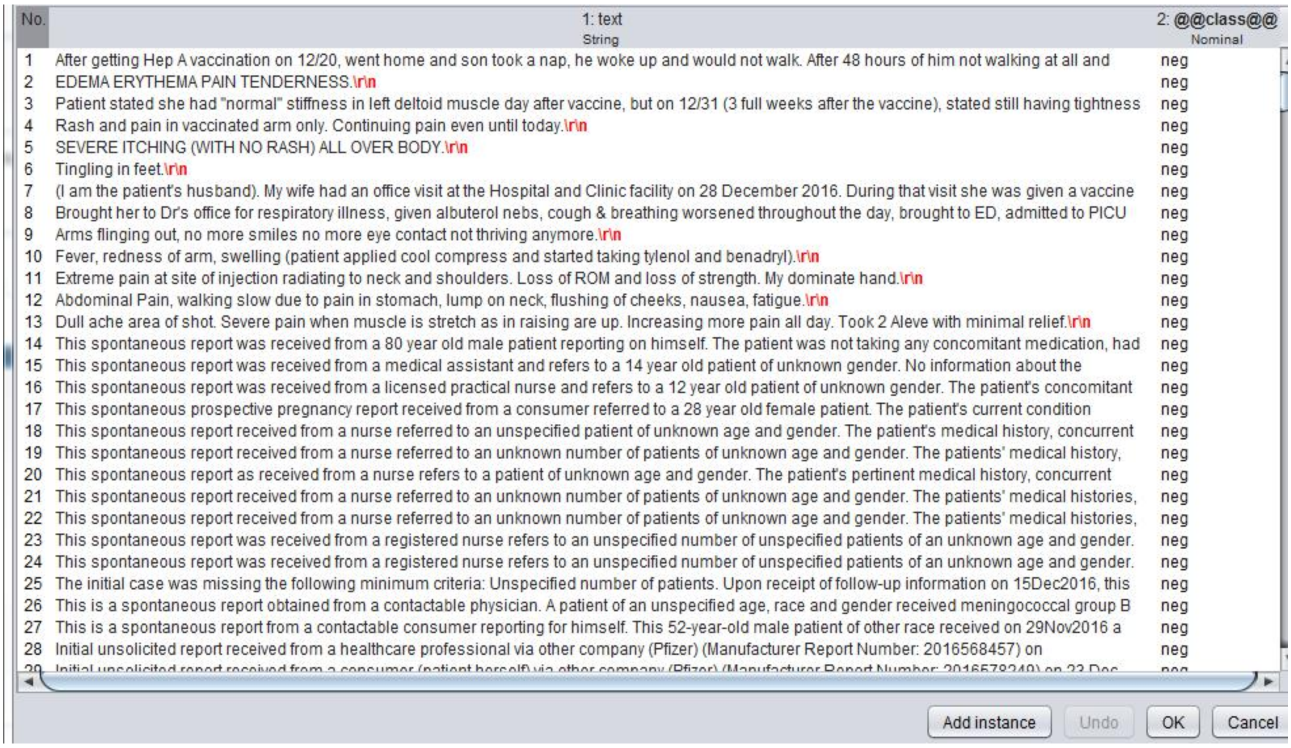
Naive Bayes is a family of classifier algorithm.Classification algorithm based on Bayes theorm.

P(A/B)=(P(B/A)P(A))/P(A/B)

It uses a process in which multiple variants of the training data are created followed by building multiple predictive models which are independent of each other and then combined to finally make predictions on the given set of data.

Figure 4.8

In Weka individual report of every patient is stored in text file and then python script is executed ,for example:Suppose if the Vaers_id of the patient is 225403 then the file is saved either as neg225403 or pos225403.In Weka the classifier is directly executed as it also gave us the option of sampling within the test and training set.



No.	1: text String	2: @@class@@ Nominal
1	After getting Hep A vaccination on 12/20, went home and son took a nap, he woke up and would not walk. After 48 hours of him not walking at all and	neg
2	EDEMA ERYTHEMA PAIN TENDERNESS.\r\n	neg
3	Patient stated she had "normal" stiffness in left deltoid muscle day after vaccine, but on 12/31 (3 full weeks after the vaccine), stated still having tightness	neg
4	Rash and pain in vaccinated arm only. Continuing pain even until today.\r\n	neg
5	SEVERE ITCHING (WITH NO RASH) ALL OVER BODY.\r\n	neg
6	Tingling in feet.\r\n	neg
7	(I am the patient's husband). My wife had an office visit at the Hospital and Clinic facility on 28 December 2016. During that visit she was given a vaccine	neg
8	Brought her to Dr's office for respiratory illness, given albuterol nebs, cough & breathing worsened throughout the day, brought to ED, admitted to PICU	neg
9	Arms flinging out, no more smiles no more eye contact not thriving anymore.\r\n	neg
10	Fever, redness of arm, swelling (patient applied cool compress and started taking tylenol and benadryl).\r\n	neg
11	Extreme pain at site of injection radiating to neck and shoulders. Loss of ROM and loss of strength. My dominate hand.\r\n	neg
12	Abdominal Pain, walking slow due to pain in stomach, lump on neck, flushing of cheeks, nausea, fatigue.\r\n	neg
13	Dull ache area of shot. Severe pain when muscle is stretch as in raising are up. Increasing more pain all day. Took 2 Aleve with minimal relief.\r\n	neg
14	This spontaneous report was received from a 80 year old male patient reporting on himself. The patient was not taking any concomitant medication, had	neg
15	This spontaneous report was received from a medical assistant and refers to a 14 year old patient of unknown gender. No information about the	neg
16	This spontaneous report was received from a licensed practical nurse and refers to a 12 year old patient of unknown gender. The patient's concomitant	neg
17	This spontaneous prospective pregnancy report received from a consumer referred to a 28 year old female patient. The patient's current condition	neg
18	This spontaneous report received from a nurse referred to an unspecified patient of unknown age and gender. The patient's medical history, concurrent	neg
19	This spontaneous report received from a nurse referred to an unknown number of patients of unknown age and gender. The patients' medical history,	neg
20	This spontaneous report as received from a nurse refers to a patient of unknown age and gender. The patient's pertinent medical history, concurrent	neg
21	This spontaneous report received from a nurse referred to an unknown number of patients of unknown age and gender. The patients' medical histories,	neg
22	This spontaneous report received from a nurse referred to an unknown number of patients of unknown age and gender. The patients' medical histories,	neg
23	This spontaneous report was received from a registered nurse refers to an unspecified number of unspecified patients of an unknown age and gender.	neg
24	This spontaneous report was received from a registered nurse refers to an unspecified number of unspecified patients of an unknown age and gender.	neg
25	The initial case was missing the following minimum criteria: Unspecified number of patients. Upon receipt of follow-up information on 15Dec2016, this	neg
26	This is a spontaneous report obtained from a contactable physician. A patient of an unspecified age, race and gender received meningococcal group B	neg
27	This is a spontaneous report from a contactable consumer reporting for himself. This 52-year-old male patient of other race received on 29Nov2016 a	neg
28	Initial unsolicited report received from a healthcare professional via other company (Pfizer) (Manufacturer Report Number: 2016568457) on	neg
29	Initial unsolicited report received from a consumer (patient herself) via other company (Pfizer) (Manufacturer Report Number: 2016578249) on 23 Dec	neg

Figure 4.2: Data after importing into Weka

4.3.1 Feature Extraction in Weka

This is the initial progress towards prediction of autism.The dataset we have contain irrelevant words and paragraphs which is of no use,remove these irrelevant words from the dataset so that dataset is ready for classification as classification is done only when that dataset is suited for the claasifier otherwise it gives false result or error.

Feature extraction in weka is done by Stringtowordvector,wordstokeep,Stopwords, Outputcount,TF-IDF ratio,Tokenizer,MinTermFrequency,AttributeSelectionFilter Wordcount all are contributed towards feature extraction from dataset.

4.3.1.1 StringToWordVector

Various datatype are not supported by machine learning.String is one of them,so string is converted to word vector documents.A table matrix is made in which there are values as input ,output.The table matrix contains text are rows,words as columns.So there is a need to convert string to numeric data which can be done by various ways.

Standardization:Standardization alludes to estimations taken on various scales and re-estimating them on a typical scale.

Stemmer: Stemmer is used to break the words into smaller part i.e remove suffix from the word and break them into stem eg:having is changed to have ,this is done by PorterStemmer class in weka because suffix is of no use in our dataset.

TF-IDF Ratio: Ratio tells us the importance of the words in the dataset i.e find out the total frequency and inverse document frequency which tells the weight of word in the dataset.

StopWords: Remove irrelevant words like is an the which of no use in the dataset.

Tokenizer:Algorithms have diverse methods for parting up the content.They split into tokens.Two types of tokenizer default tokenizer contains @,!,~,&,% and alphatetic tokenizer havind unigrams only,first default tokenizer is used after that we are left with 100% letters which is then taken care of by alphabetic tokenizer.

Wordstokeep:Average number of words we want in our datset.

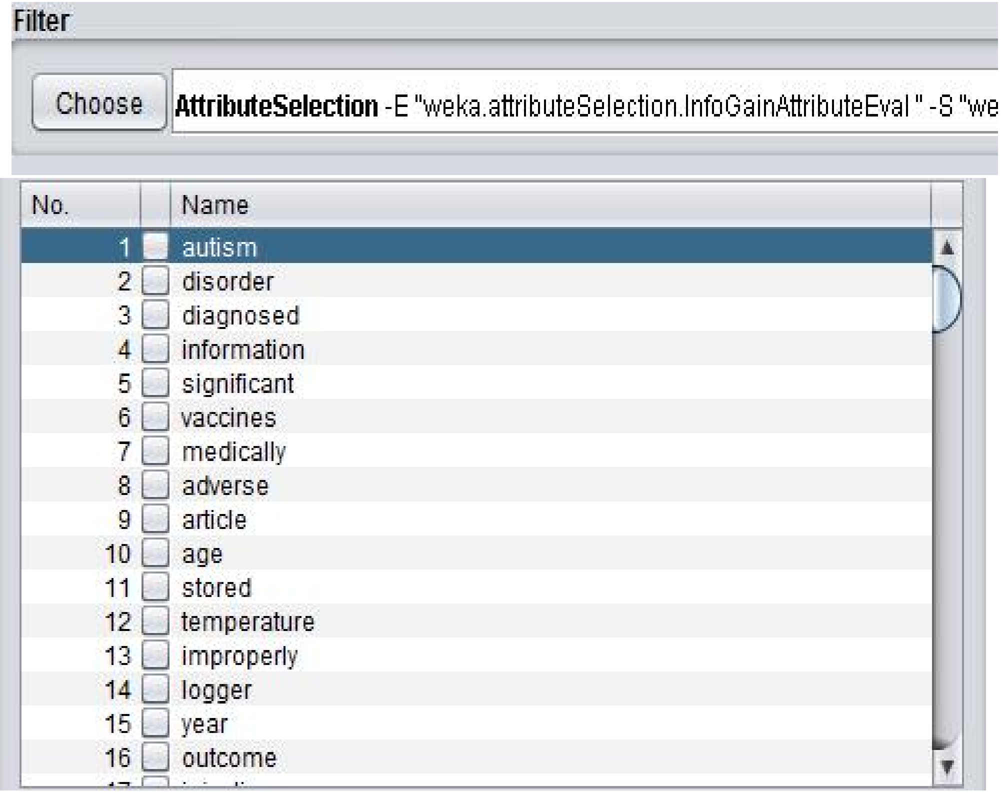
doNotOperateOnPerclassBasis:In the event that this setting is put valid, at that point the quantity of "wordsToKeep" is considered altogether independent of class (Pos/Neg) generally the quantity of "wordsToKeep" is considered on per class premise.

4.3.1.2 Attribute Selection

The AttributeSelectionFilter every now and again compliments the StringToWordVector as splendid data is made. StringToWordVector changed all the symptom_text and their words into file vectors. AttributeSelection is interesting. It doesn't change characters into different numbers. It positions the properties and further improves the data. Under the settings of AttributeSelectionFilter, Evaluator and Search can be picked which are elucidated as seeks after:

Evaluator – InfoGainAttribueEval :Evaluator is the identification for making a decision about the prescient nature of the property.

Inquiry – Ranker :Search counsels the Judge (Evaluator) to settle on an official choice to acknowledge or dismiss the characteristic.



Filter

AttributeSelection -E "weka.attributeSelection.InfoGainAttributeEval " -S "we

No.	Name
1	autism
2	disorder
3	diagnosed
4	information
5	significant
6	vaccines
7	medically
8	adverse
9	article
10	age
11	stored
12	temperature
13	improperly
14	logger
15	year
16	outcome

Figure 4.3: Words after applying AttributeSelectionFilter

From figure 4.3, we can see that ranked # 1-autism, ranked # 2-disorder and so on.

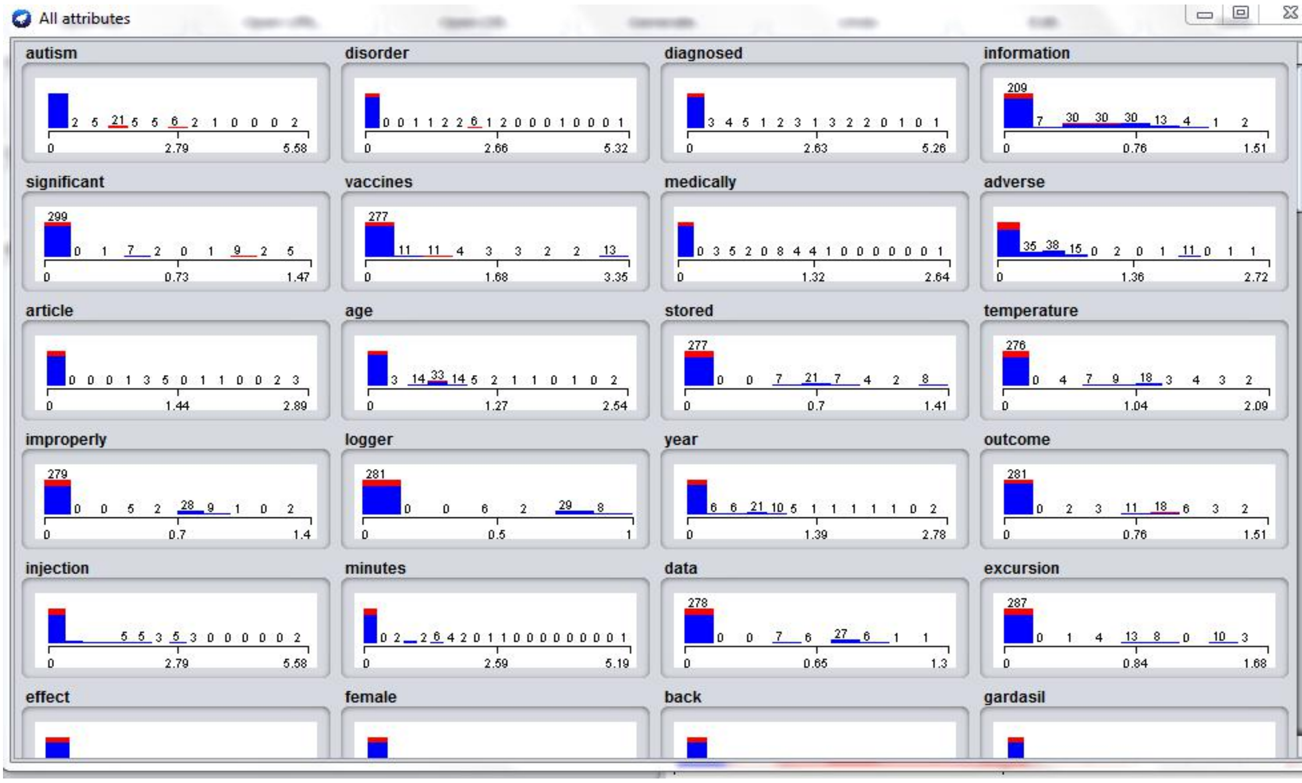
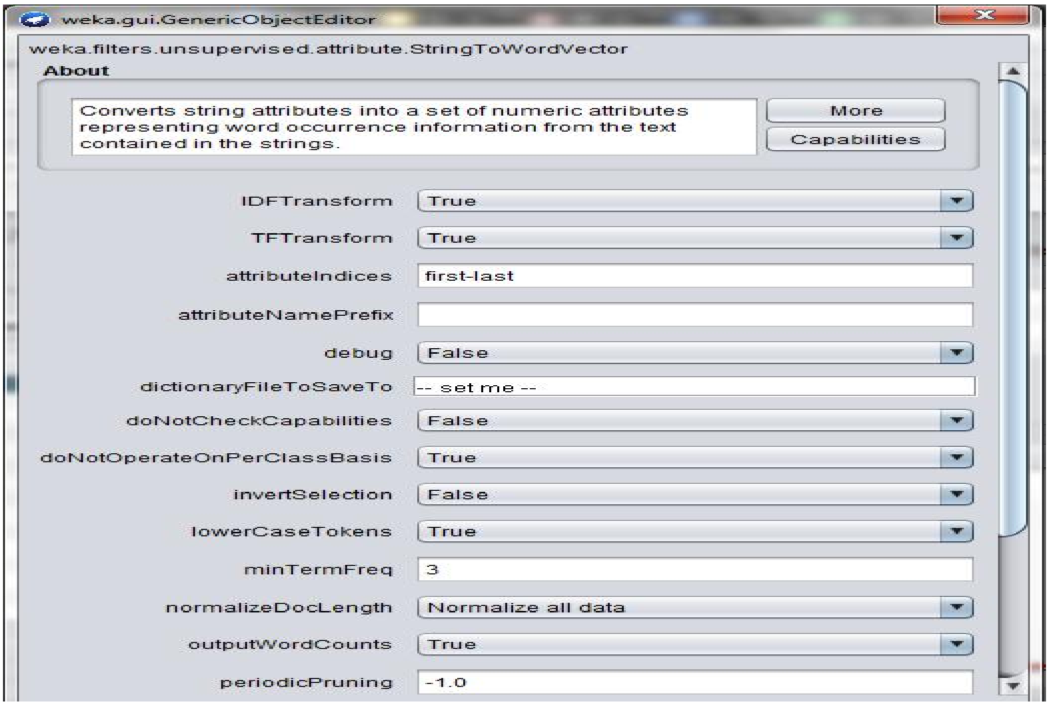


Figure 4.4: Words Distribution



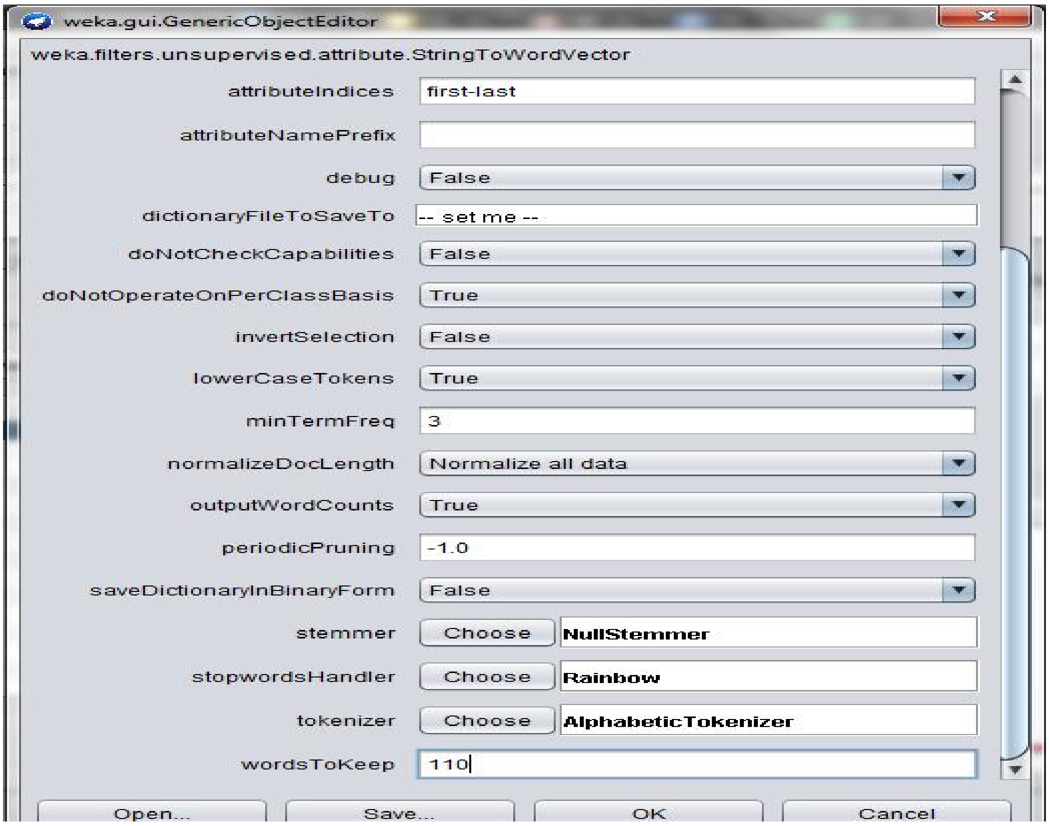
weka.filters.unsupervised.attribute.StringToWordVector

About

Converts string attributes into a set of numeric attributes representing word occurrence information from the text contained in the strings.

IDFTransform	True
TFTransform	True
attributeIndices	first-last
attributeNamePrefix	
debug	False
dictionaryFileToSaveTo	-- set me --
doNotCheckCapabilities	False
doNotOperateOnPerClassBasis	True
invertSelection	False
lowerCaseTokens	True
minTermFreq	3
normalizeDocLength	Normalize all data
outputWordCounts	True
periodicPruning	-1.0

Figure 4.5: (a) :StringToWordVector Settings

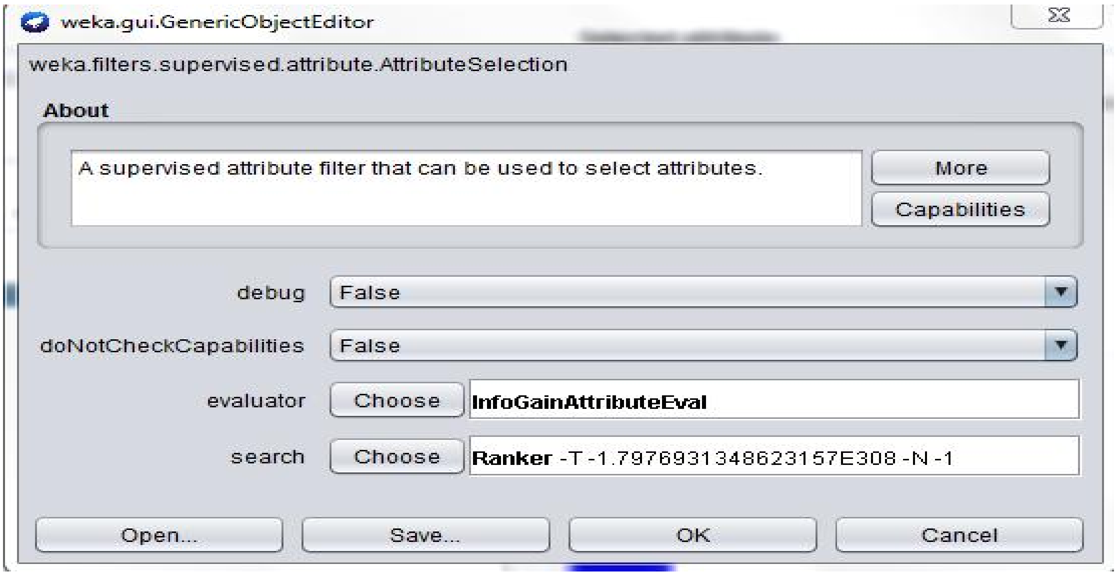


weka.filters.unsupervised.attribute.StringToWordVector

attributeIndices	first-last
attributeNamePrefix	
debug	False
dictionaryFileToSaveTo	-- set me --
doNotCheckCapabilities	False
doNotOperateOnPerClassBasis	True
invertSelection	False
lowerCaseTokens	True
minTermFreq	3
normalizeDocLength	Normalize all data
outputWordCounts	True
periodicPruning	-1.0
saveDictionaryInBinaryForm	False
stemmer	NullStemmer
stopwordsHandler	Rainbow
tokenizer	AlphabeticTokenizer
wordsToKeep	110

Figure 4.5: (b) StringToWordVector Settings

The environment for AttributeSelectionFilter applied to the data in this project are as shown in Figure 4.6.



weka.gui.GenericObjectEditor

weka.filters.supervised.attribute.AttributeSelection

About

A supervised attribute filter that can be used to select attributes.

debug: False

doNotCheckCapabilities: False

evaluator: InfoGainAttributeEval

search: Ranker -T -1.7976931348623157E308 -N -1

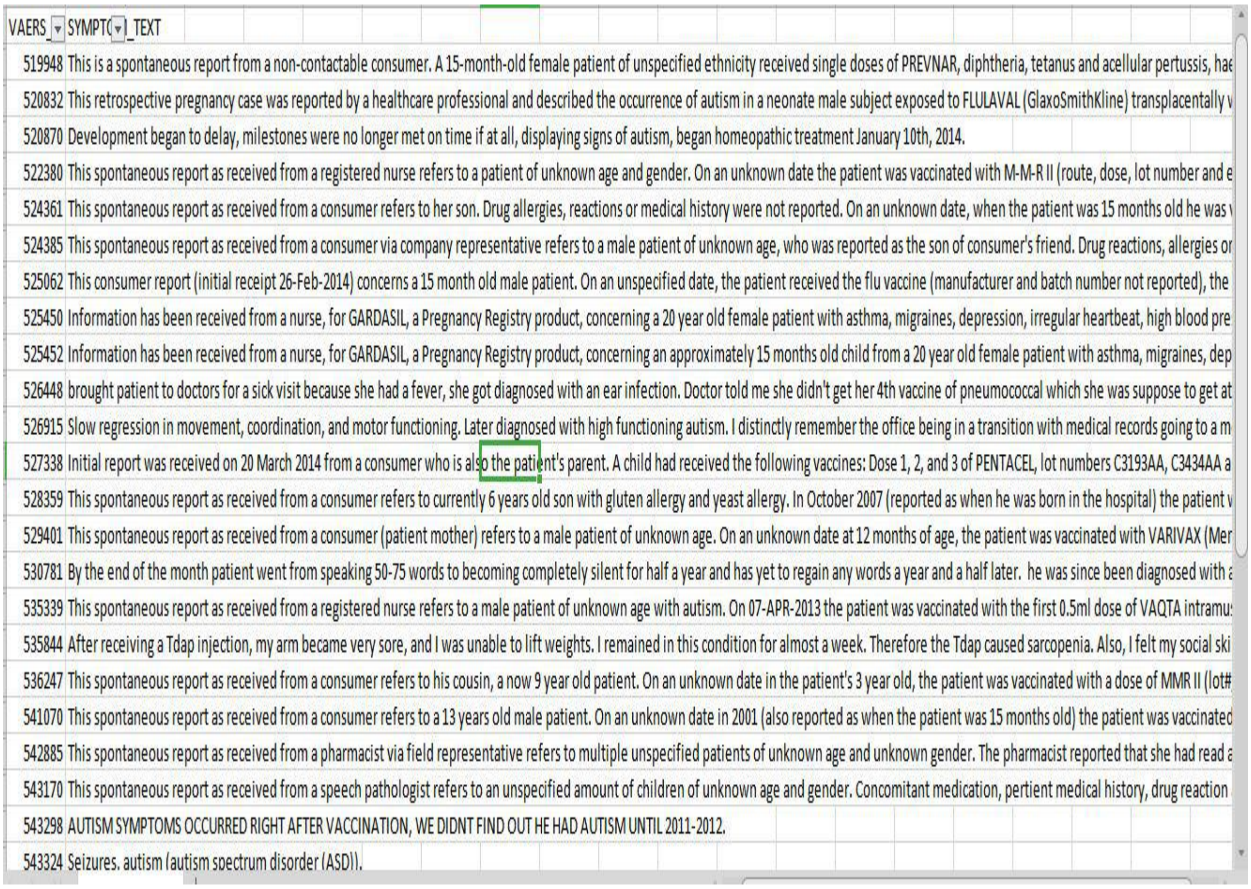
Figure 4.6: AttributeSelectionFilter Settings

5. TEST PLAN

5.1 Data Set

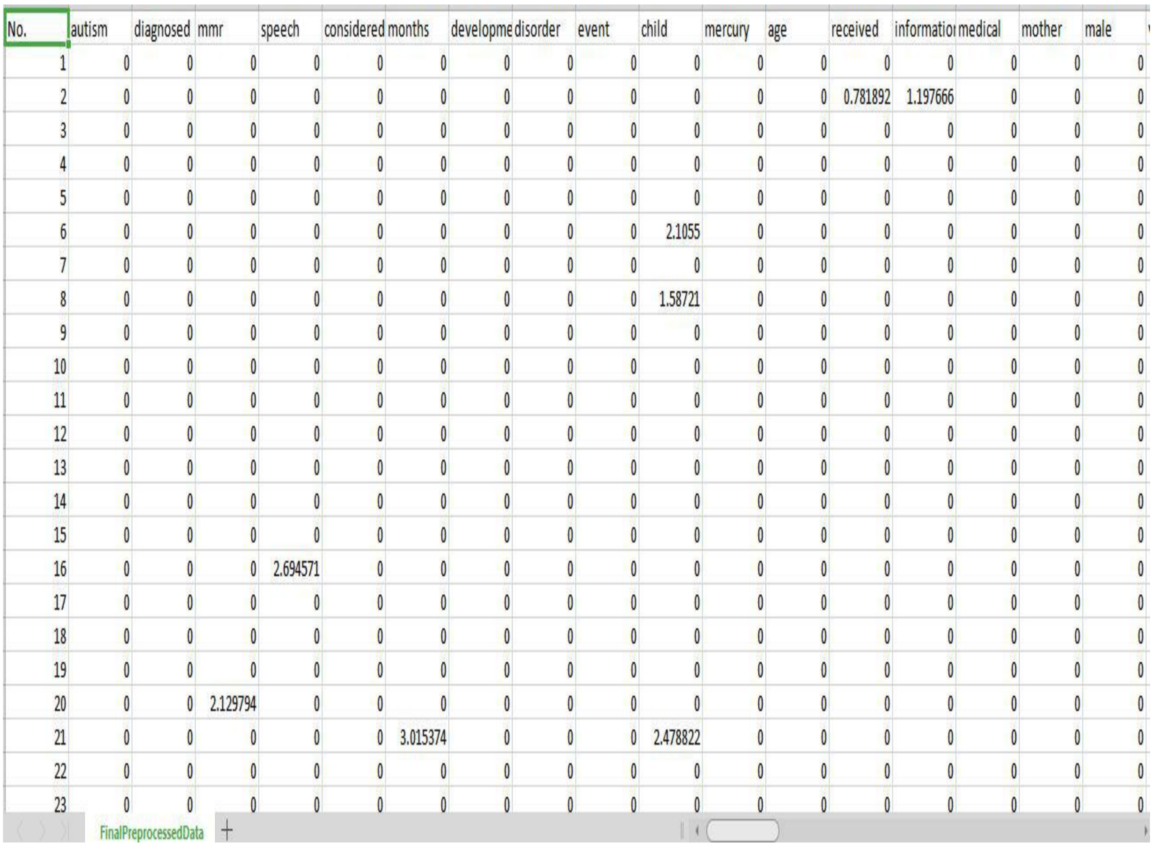
When it comes to health prediction there are a bundle of features that have to be used because the health of patient is so important in its own ways. A prediction of autism can be depended on so many factors such as the symptoms and how to respond to doctor report .Therefore, to predict a autism all of these features have to be used to train the models. And to better train the models huge data set is required hence, for this particular betting app the data of previous years(2000-2017)was collected.

The snippet of the raw data set is shown in the figure 5.1 and the data set after extraction feature from weka is shown in the figure 5.2.



VAERS_	SYMPTO_TEXT
519948	This is a spontaneous report from a non-contactable consumer. A 15-month-old female patient of unspecified ethnicity received single doses of PREVNAR, diphtheria, tetanus and acellular pertussis, hae
520832	This retrospective pregnancy case was reported by a healthcare professional and described the occurrence of autism in a neonate male subject exposed to FLULAVAL (GlaxoSmithKline) transplacentally v
520870	Development began to delay, milestones were no longer met on time if at all, displaying signs of autism, began homeopathic treatment January 10th, 2014.
522380	This spontaneous report as received from a registered nurse refers to a patient of unknown age and gender. On an unknown date the patient was vaccinated with M-M-R II (route, dose, lot number and e
524361	This spontaneous report as received from a consumer refers to her son. Drug allergies, reactions or medical history were not reported. On an unknown date, when the patient was 15 months old he was v
524385	This spontaneous report as received from a consumer via company representative refers to a male patient of unknown age, who was reported as the son of consumer's friend. Drug reactions, allergies or
525062	This consumer report (initial receipt 26-Feb-2014) concerns a 15 month old male patient. On an unspecified date, the patient received the flu vaccine (manufacturer and batch number not reported), the
525450	Information has been received from a nurse, for GARDASIL, a Pregnancy Registry product, concerning a 20 year old female patient with asthma, migraines, depression, irregular heartbeat, high blood pre
525452	Information has been received from a nurse, for GARDASIL, a Pregnancy Registry product, concerning an approximately 15 months old child from a 20 year old female patient with asthma, migraines, dep
526448	brought patient to doctors for a sick visit because she had a fever, she got diagnosed with an ear infection. Doctor told me she didn't get her 4th vaccine of pneumococcal which she was suppose to get at
526915	Slow regression in movement, coordination, and motor functioning. Later diagnosed with high functioning autism. I distinctly remember the office being in a transition with medical records going to a m
527338	Initial report was received on 20 March 2014 from a consumer who is also the patient's parent. A child had received the following vaccines: Dose 1, 2, and 3 of PENTACEL, lot numbers C3193AA, C3434AA a
528359	This spontaneous report as received from a consumer refers to currently 6 years old son with gluten allergy and yeast allergy. In October 2007 (reported as when he was born in the hospital) the patient v
529401	This spontaneous report as received from a consumer (patient mother) refers to a male patient of unknown age. On an unknown date at 12 months of age, the patient was vaccinated with VARIVAX (Mer
530781	By the end of the month patient went from speaking 50-75 words to becoming completely silent for half a year and has yet to regain any words a year and a half later. he was since been diagnosed with a
535339	This spontaneous report as received from a registered nurse refers to a male patient of unknown age with autism. On 07-APR-2013 the patient was vaccinated with the first 0.5ml dose of VAQTA intramus
535844	After receiving a Tdap injection, my arm became very sore, and I was unable to lift weights. I remained in this condition for almost a week. Therefore the Tdap caused sarcopenia. Also, I felt my social ski
536247	This spontaneous report as received from a consumer refers to his cousin, a now 9 year old patient. On an unknown date in the patient's 3 year old, the patient was vaccinated with a dose of MMR II (lot#
541070	This spontaneous report as received from a consumer refers to a 13 years old male patient. On an unknown date in 2001 (also reported as when the patient was 15 months old) the patient was vaccinated
542885	This spontaneous report as received from a pharmacist via field representative refers to multiple unspecified patients of unknown age and unknown gender. The pharmacist reported that she had read a
543170	This spontaneous report as received from a speech pathologist refers to an unspecified amount of children of unknown age and gender. Concomitant medication, pertient medical history, drug reaction
543298	AUTISM SYMPTOMS OCCURRED RIGHT AFTER VACCINATION, WE DIDNT FIND OUT HE HAD AUTISM UNTIL 2011-2012.
543324	Seizures, autism (autism spectrum disorder (ASD)).

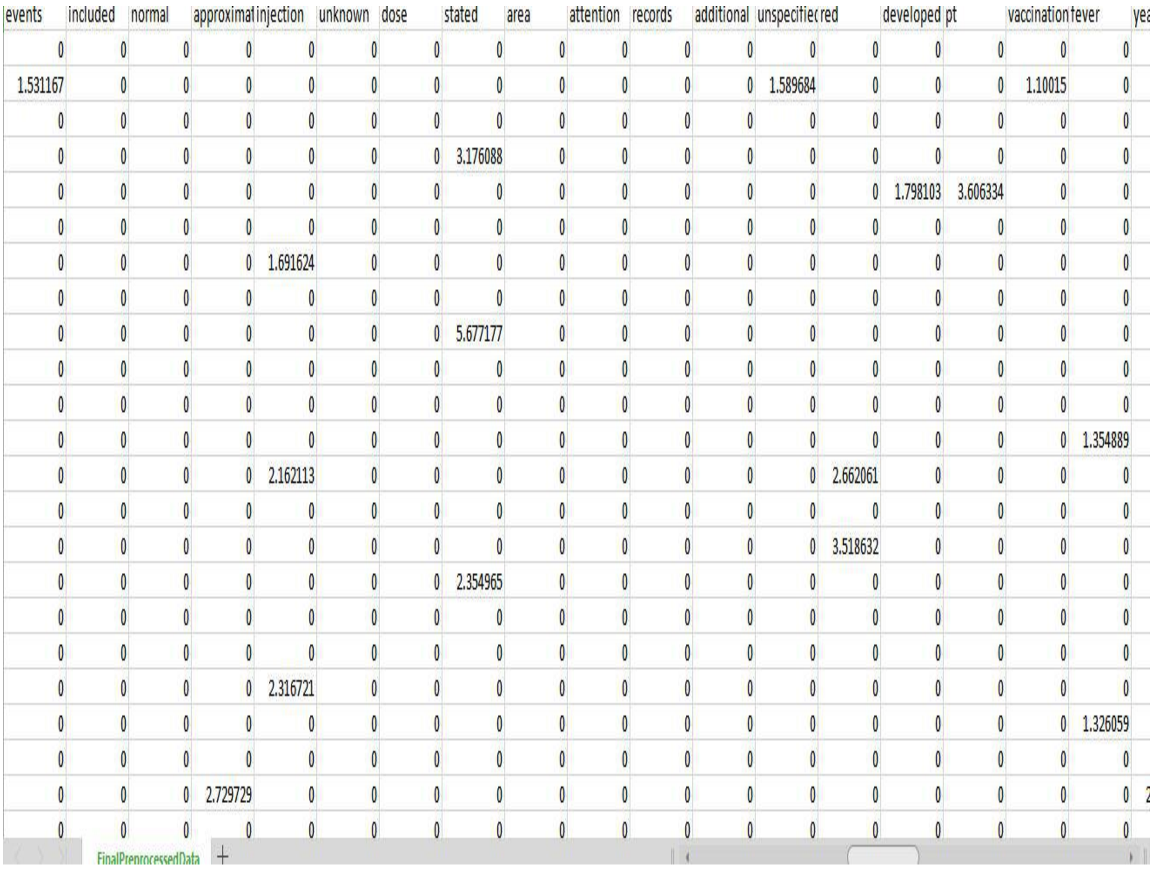
Figure 5.1



No.	autism	diagnosed	mmr	speech	considered	months	developme	disorder	event	child	mercury	age	received	informatio	medical	mother	male
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0.781892	1.197666	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	2.1055	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	1.58721	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	2.694571	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	0	0	2.129794	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	0	0	0	0	0	3.015374	0	0	0	2.478822	0	0	0	0	0	0	0
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 5.2

Another dataset consisting of the symptoms of patients for the seasons from 2000 until 2017 was collected and snippet is shown in the figure 5.3. The zero in the dataset show that the corresponding patient did not undergo with the corresponding symptom.The data which is shown in figure 5.2,5.3 is final preprocessed data which is a csv file which is collected after feature extraction from weka. These matrix represents the weight of a symptom provided the doctor report.

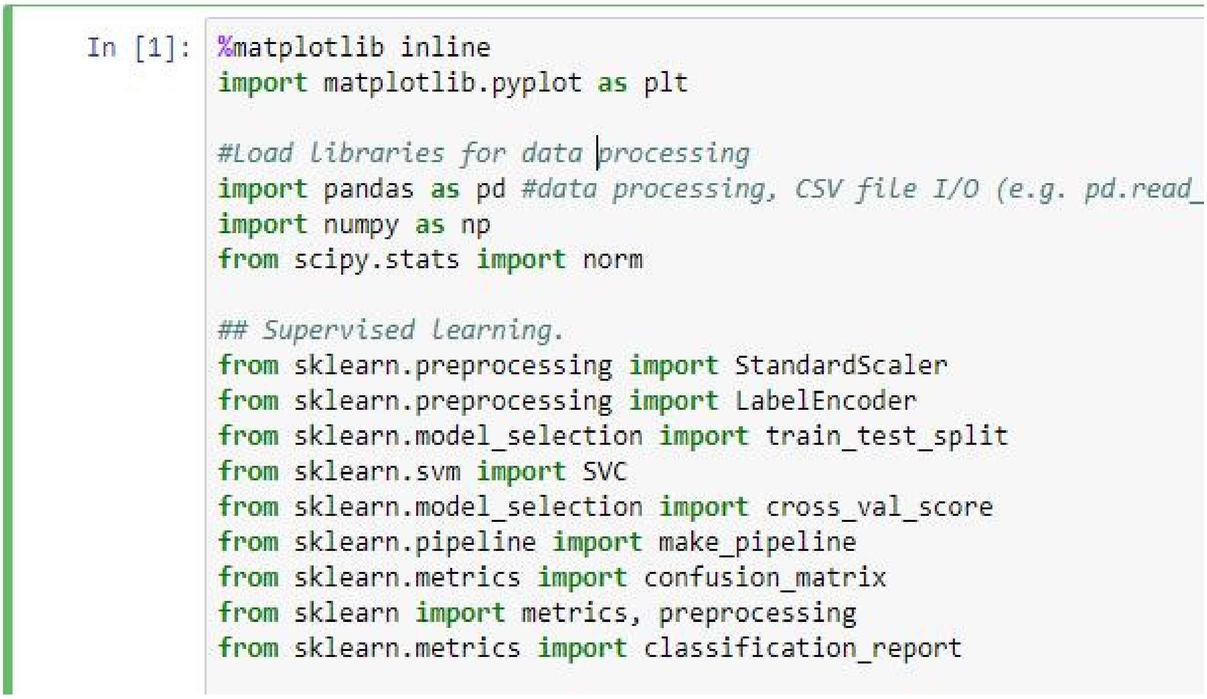


events	included	normal	approximat	injection	unknown	dose	stated	area	attention	records	additional	unspecifiec	red	developed	pt	vaccination	fever	yea
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
1.531167	0	0	0	0	0	0	0	0	0	0	0	1.589684	0	0	0	1.10015	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	3.176088	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.798103	3.606334	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	1.691624	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	5.677177	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.354889	
0	0	0	0	2.162113	0	0	0	0	0	0	0	0	2.662061	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	3.518632	0	0	0	0	
0	0	0	0	0	0	0	2.354965	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	2.316721	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.326059	
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
0	0	0	2.729729	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	

Figure 5.3

5.2 Implementing the Classifier

First we import all the necessary libraries required to implement our machine learning model and also to visualize the results. The import is shown in Figure 5.4.The following code snippet is for the SVM.After that code of both SVM and Logistic Regression will be explained and the performance of both will be compared and the best will be choosen. The performance analysis of naive bayes will also be compared which is done through weka directly.In weka Naive bayes is performed directly no need to write a code for this, Weka run the classifier directly and provided the precision and accuracy which is enough for comparison .

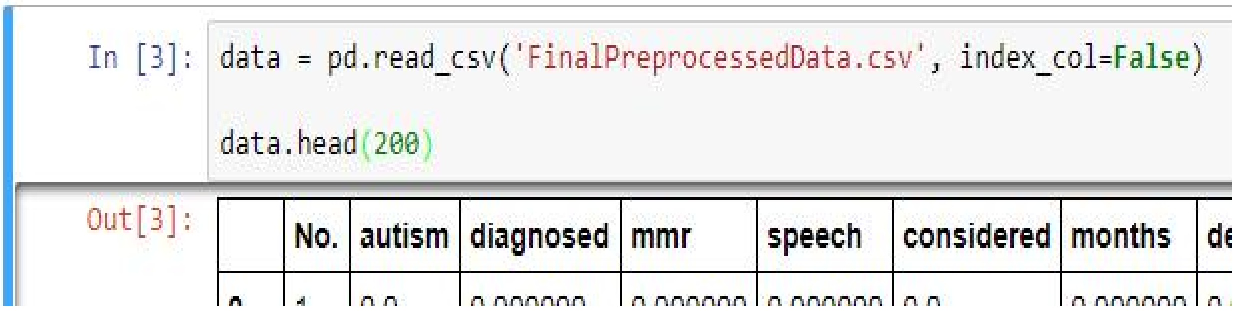


```
In [1]: %matplotlib inline
        import matplotlib.pyplot as plt

        #Load libraries for data processing
        import pandas as pd #data processing, CSV file I/O (e.g. pd.read_
        import numpy as np
        from scipy.stats import norm

        ## Supervised learning.
        from sklearn.preprocessing import StandardScaler
        from sklearn.preprocessing import LabelEncoder
        from sklearn.model_selection import train_test_split
        from sklearn.svm import SVC
        from sklearn.model_selection import cross_val_score
        from sklearn.pipeline import make_pipeline
        from sklearn.metrics import confusion_matrix
        from sklearn import metrics, preprocessing
        from sklearn.metrics import classification_report
```

Figure 5.4. Loading Necessary Libraries

Next, we load the data which we obtained after the pre-processing on weka. We changed the pre-processed data into a CSV (Comma Separated Values) file and then used the read_csv function of pandas to load our data into our ipython notebook. This can be seen in Figure 5.5

.

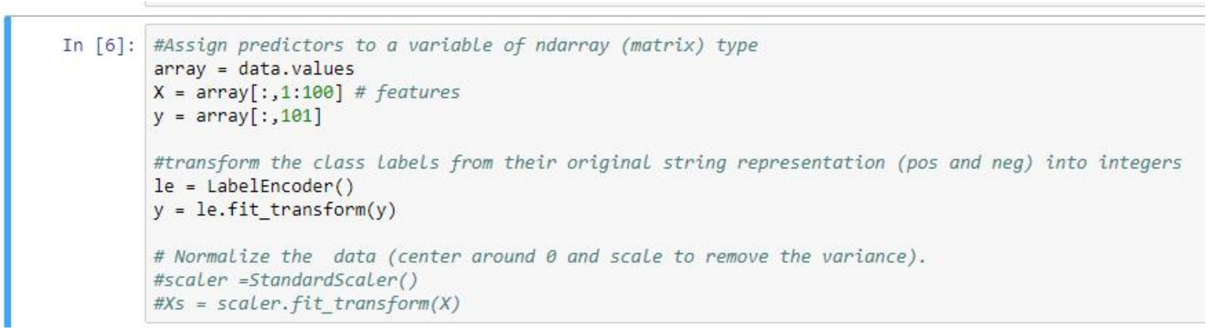


```
In [3]: data = pd.read_csv('FinalPreprocessedData.csv', index_col=False)

        data.head(200)

Out[3]:
```

	No.	autism	diagnosed	mmr	speech	considered	months	de
0	1	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.

Figure 5.5 Loading the pre-processed data

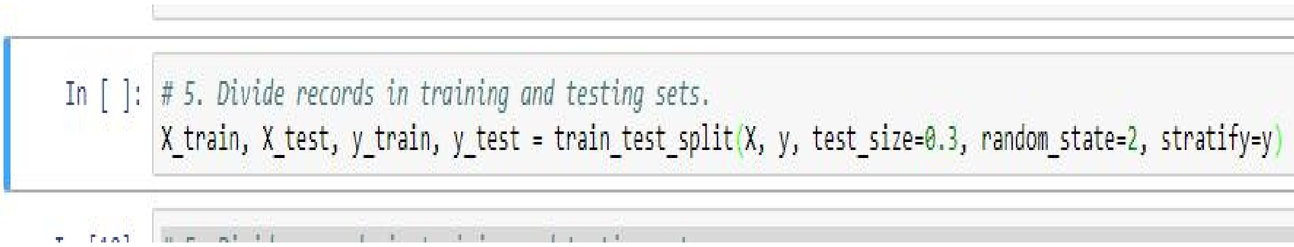
Next, we encode our class labels into integers in order to implement the classifier. The encoding is done using a Label Encoder from sklearn. The implementation of the Label encoder can be seen in Figure 5.6.



```
#Assign predictors to a variable of ndarray (matrix) type
array = data.values
X = array[:,1:100] # features
y = array[:,101]

#transform the class labels from their original string representation (pos and neg) into integers
le = LabelEncoder()
y = le.fit_transform(y)

# Normalize the  data (center around 0 and scale to remove the variance).
#scaler =StandardScaler()
#Xs = scaler.fit_transform(X)
```

Figure 5.6 Code Snippet for Label Encoding

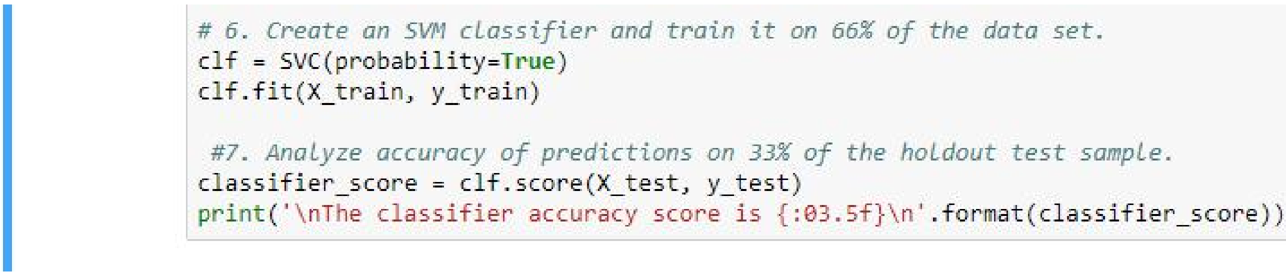
Our next step would be to divide our data into training and test sets. Here, we have used 33% of the information data as the test set and 67% of the information data as the training set. We have used a random variable as our preprocessed data was ordered according to the class labels, i.e., all the tuples with the negative classes were followed by all the tuples with positive classes. The division can be seen in Figure 5.7.



```
# 5. Divide records in training and testing sets.
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=2, stratify=y)
```

Figure 5.7

After this step, we are ready to actually execute the Support Vector Machine on the pre-processed data and get the Accuracy. The efficiency of the algorithm is also determined by factors other than accuracy. These factors include Precision, Recall and the Confusion Matrix which tells us about the false positives and false negatives. The implementation of the classifier and the code for the visualization can be seen in Figure 4.14 and Figure 5.8



```
# 6. Create an SVM classifier and train it on 66% of the data set.
clf = SVC(probability=True)
clf.fit(X_train, y_train)

 #7. Analyze accuracy of predictions on 33% of the holdout test sample.
classifier_score = clf.score(X_test, y_test)
print('\nThe classifier accuracy score is {:03.5f}\n'.format(classifier_score))
```

.

Figure 5.6

5.3 Test Setup and Plan

1.Gathering of the datasets of the season 2000 to 2017 as have already been shown in the figures 5.1, figure 5.2 and figure 5.3.

2.The dataset has to be cleaned by pre-processing all the data and creating a final and single dataset. The process of data scraping and pre-processing is given as above through weka:

3.Finally, after all the scraping and cleaning of the datasets the accumulated data was all appended together into one .csv file and saved as 'final_dataset.csv'. This final dataset is used for the predictions.

4.The final dataset was split into training and testing data having 67% and 33% proportion.

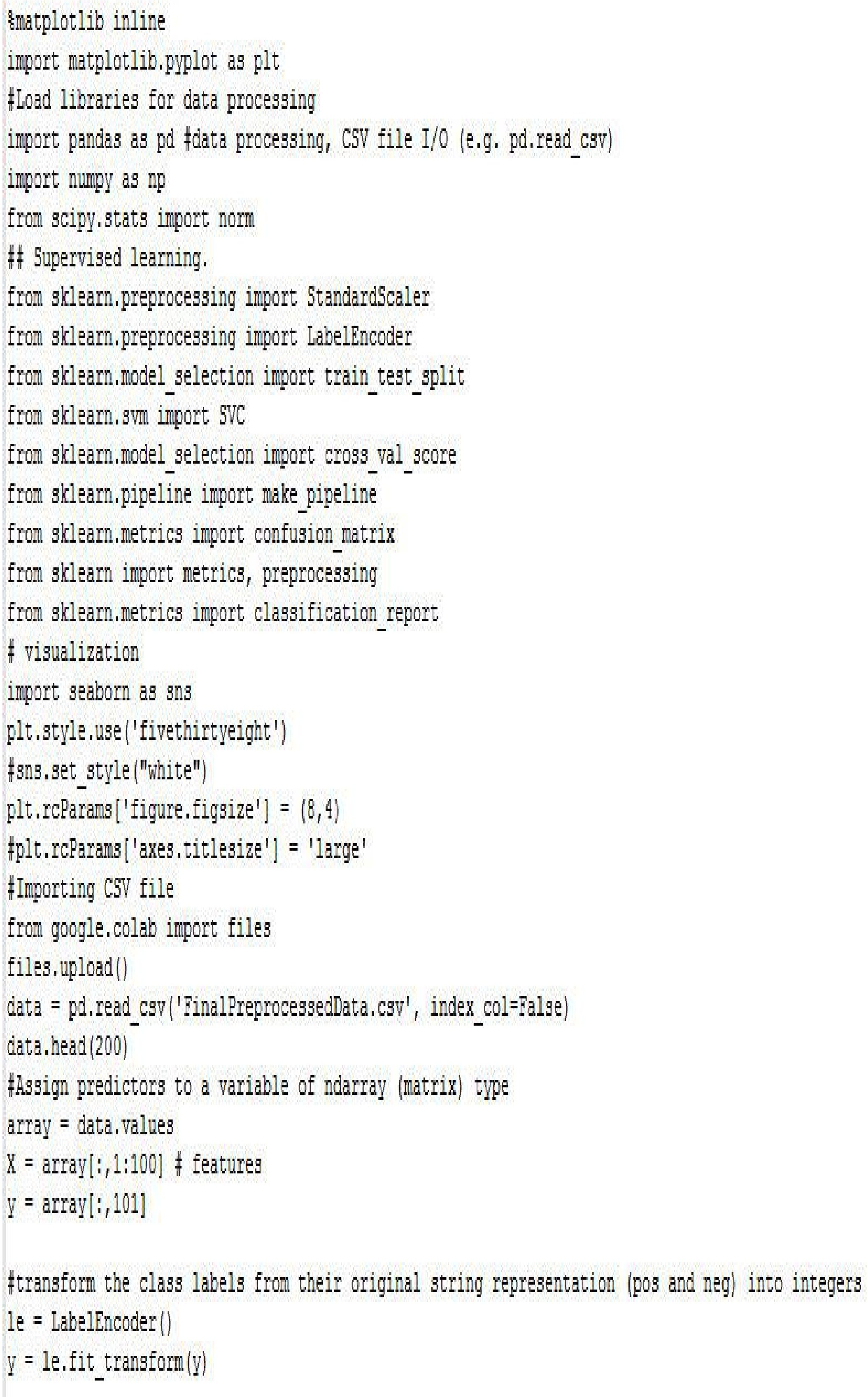
5.Three different classifiers (Logistic Regression, SVM and Naive Bayes) that have already been explained in the Section 4, Algorithms, were trained on the data.

6.The classifier that used up the minimum time to train itself and make predictions was chosen as the best one.

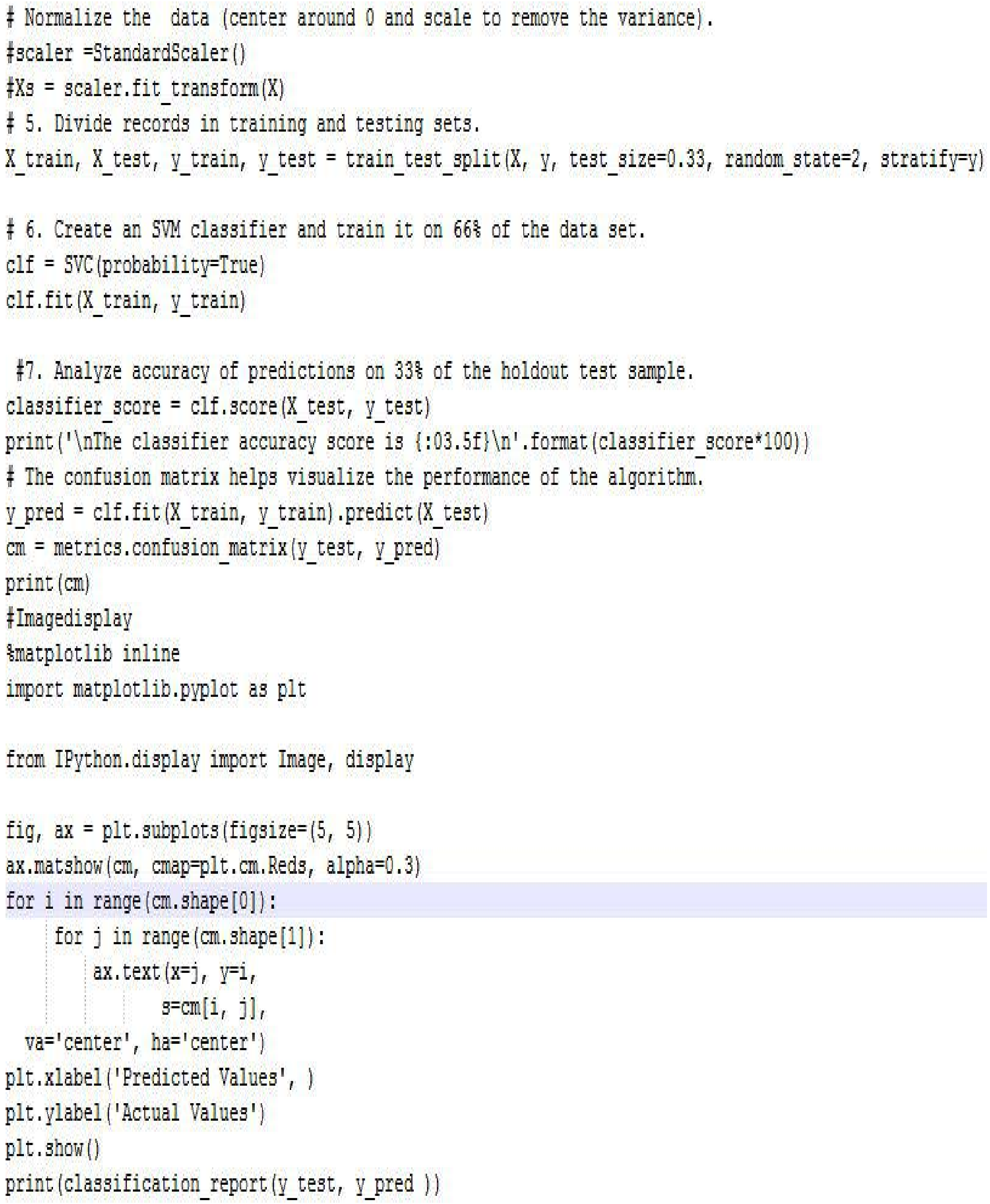
Further optimize the predicting model by choosing its parameters carefully and tuning them. The dataset of the upcoming or the ongoing patient will be downloaded and the predicting model could be used on them to make real time predictions.

Display the odds of a particular patient by using the predictions made so that the prediction will be accurate .

Code for SVM:

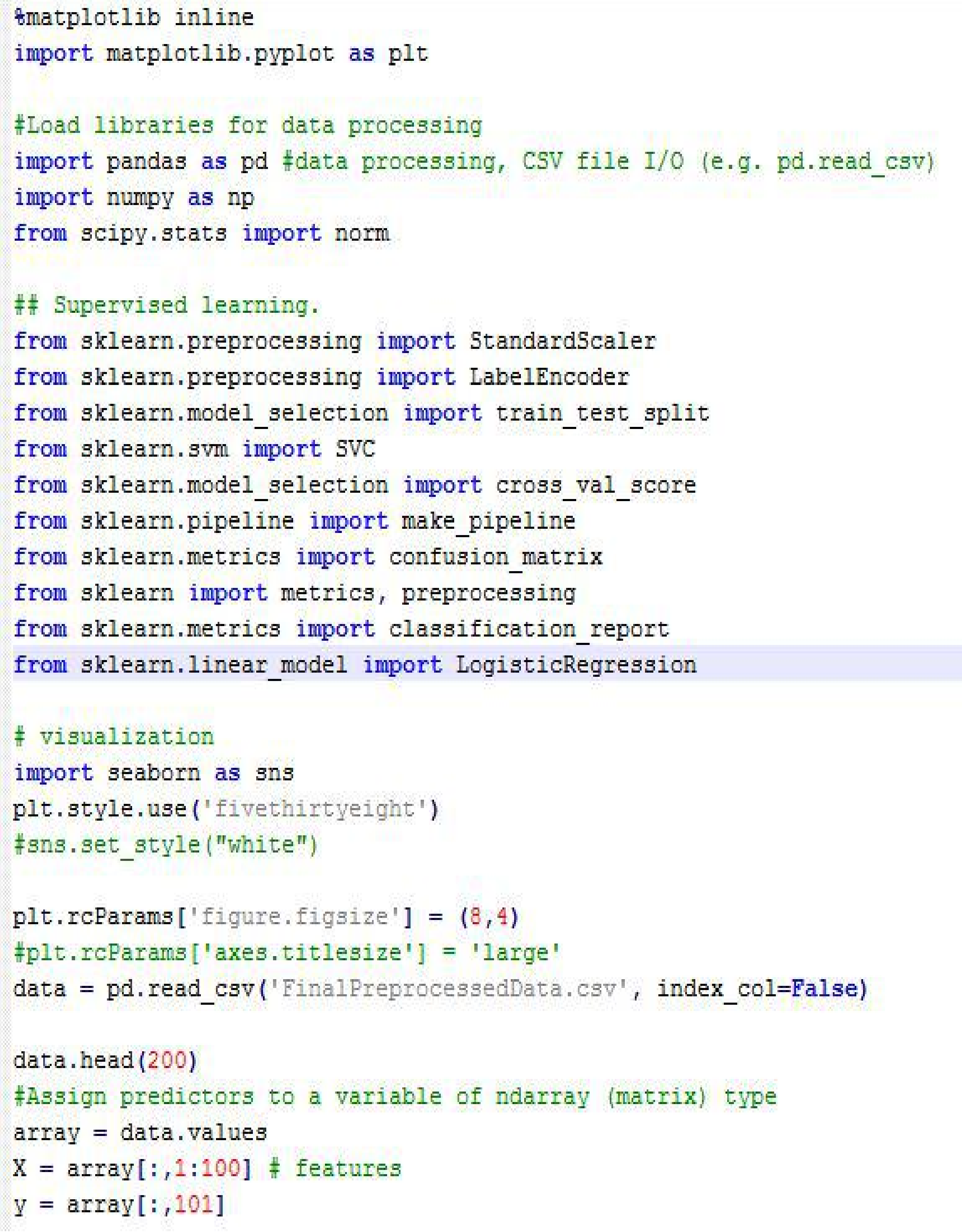


```
%matplotlib inline
import matplotlib.pyplot as plt
#Load libraries for data processing
import pandas as pd #data processing, CSV file I/O (e.g. pd.read_csv)
import numpy as np
from scipy.stats import norm
## Supervised learning.
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.model_selection import cross_val_score
from sklearn.pipeline import make_pipeline
from sklearn.metrics import confusion_matrix
from sklearn import metrics, preprocessing
from sklearn.metrics import classification_report
# visualization
import seaborn as sns
plt.style.use('fivethirtyeight')
#sns.set_style("white")
plt.rcParams['figure.figsize'] = (8,4)
#plt.rcParams['axes.titlesize'] = 'large'
#Importing CSV file
from google.colab import files
files.upload()
data = pd.read_csv('FinalPreprocessedData.csv', index_col=False)
data.head(200)
#Assign predictors to a variable of ndarray (matrix) type
array = data.values
X = array[:,1:100] # features
y = array[:,101]

#transform the class labels from their original string representation (pos and neg) into integers
le = LabelEncoder()
y = le.fit_transform(y)
```

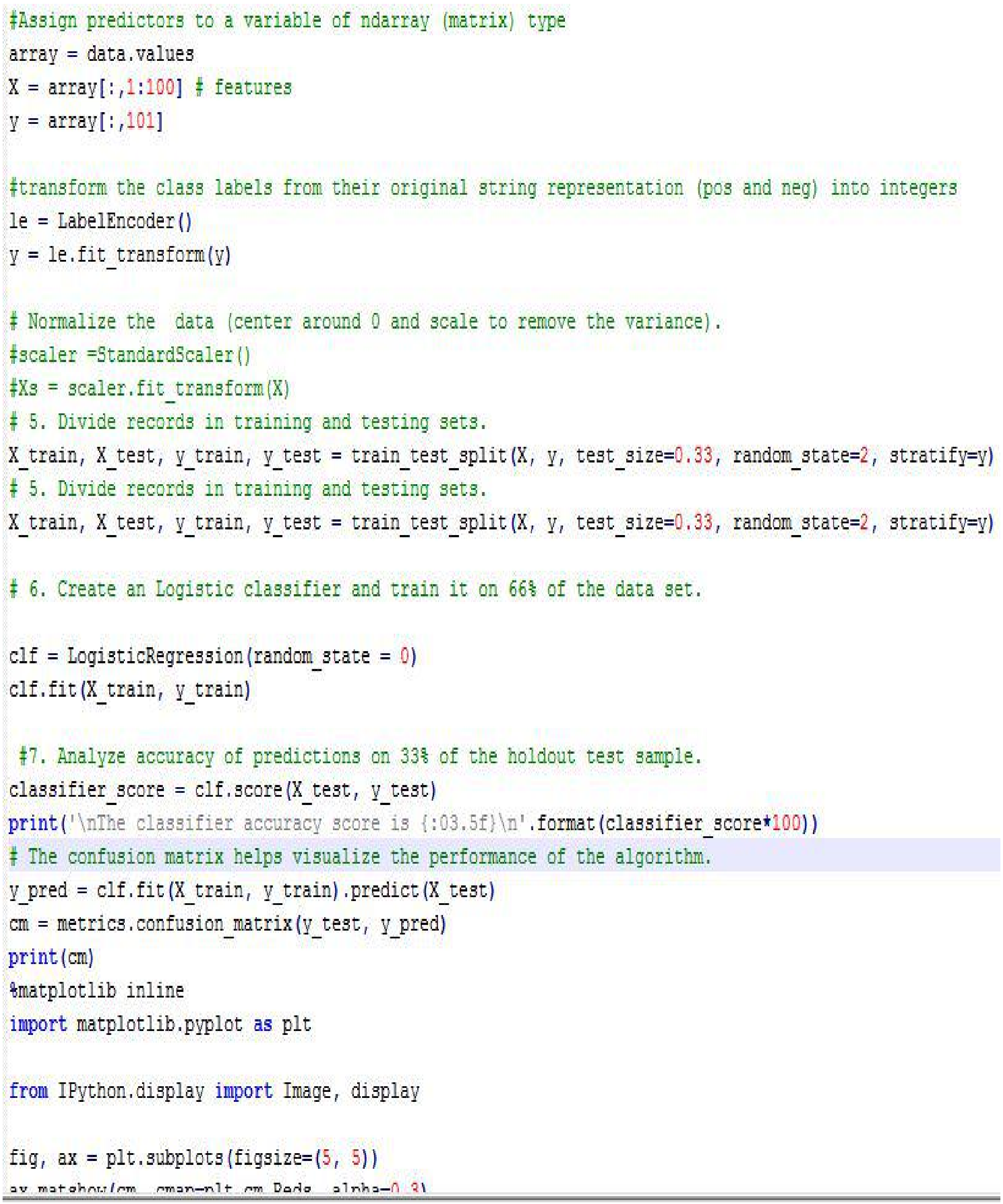


```
# Normalize the  data (center around 0 and scale to remove the variance).
#scaler =StandardScaler()
#Xs = scaler.fit_transform(X)
# 5. Divide records in training and testing sets.
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=2, stratify=y)

# 6. Create an SVM classifier and train it on 66% of the data set.
clf = SVC(probability=True)
clf.fit(X_train, y_train)

 #7. Analyze accuracy of predictions on 33% of the holdout test sample.
classifier_score = clf.score(X_test, y_test)
print('\nThe classifier accuracy score is {:03.5f}\n'.format(classifier_score*100))
# The confusion matrix helps visualize the performance of the algorithm.
y_pred = clf.fit(X_train, y_train).predict(X_test)
cm = metrics.confusion_matrix(y_test, y_pred)
print(cm)
#Imagedisplay
%matplotlib inline
import matplotlib.pyplot as plt

from IPython.display import Image, display

fig, ax = plt.subplots(figsize=(5, 5))
ax.matshow(cm, cmap=plt.cm.Reds, alpha=0.3)
for i in range(cm.shape[0]):
    for j in range(cm.shape[1]):
        ax.text(x=j, y=i,
                s=cm[i, j],
  va='center', ha='center')
plt.xlabel('Predicted Values', )
plt.ylabel('Actual Values')
plt.show()
print(classification_report(y_test, y_pred ))
```

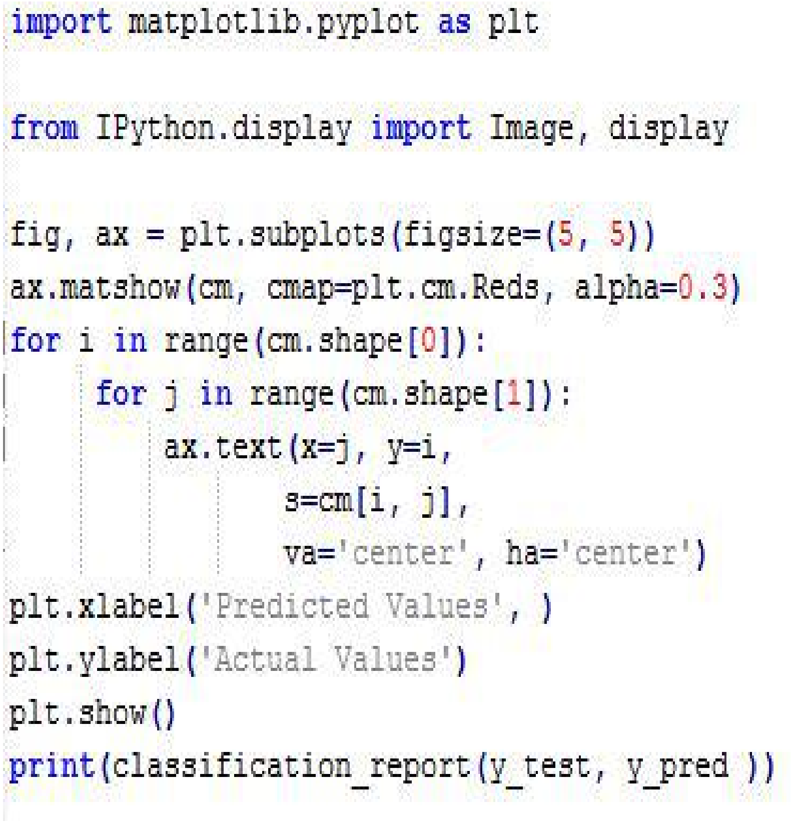
Code for Logistic Regression Classifier



```
%matplotlib inline
import matplotlib.pyplot as plt

#Load libraries for data processing
import pandas as pd #data processing, CSV file I/O (e.g. pd.read_csv)
import numpy as np
from scipy.stats import norm

## Supervised learning.
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.model_selection import cross_val_score
from sklearn.pipeline import make_pipeline
from sklearn.metrics import confusion_matrix
from sklearn import metrics, preprocessing
from sklearn.metrics import classification_report
from sklearn.linear_model import LogisticRegression

# visualization
import seaborn as sns
plt.style.use('fivethirtyeight')
#sns.set_style("white")

plt.rcParams['figure.figsize'] = (8,4)
#plt.rcParams['axes.titlesize'] = 'large'
data = pd.read_csv('FinalPreprocessedData.csv', index_col=False)

data.head(200)
#Assign predictors to a variable of ndarray (matrix) type
array = data.values
X = array[:,1:100] # features
y = array[:,101]
```



```
#Assign predictors to a variable of ndarray (matrix) type
array = data.values
X = array[:,1:100] # features
y = array[:,101]

#transform the class labels from their original string representation (pos and neg) into integers
le = LabelEncoder()
y = le.fit_transform(y)

# Normalize the  data (center around 0 and scale to remove the variance).
#scaler =StandardScaler()
#Xs = scaler.fit_transform(X)
# 5. Divide records in training and testing sets.
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=2, stratify=y)
# 5. Divide records in training and testing sets.
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=2, stratify=y)

# 6. Create an Logistic classifier and train it on 66% of the data set.

clf = LogisticRegression(random_state = 0)
clf.fit(X_train, y_train)

 #7. Analyze accuracy of predictions on 33% of the holdout test sample.
classifier_score = clf.score(X_test, y_test)
print('\nThe classifier accuracy score is {:03.5f}\n'.format(classifier_score*100))
# The confusion matrix helps visualize the performance of the algorithm.
y_pred = clf.fit(X_train, y_train).predict(X_test)
cm = metrics.confusion_matrix(y_test, y_pred)
print(cm)
%matplotlib inline
import matplotlib.pyplot as plt

from IPython.display import Image, display

fig, ax = plt.subplots(figsize=(5, 5))
ax.matshow(cm, cmap=plt.cm.Reds, alpha=0.3)
```



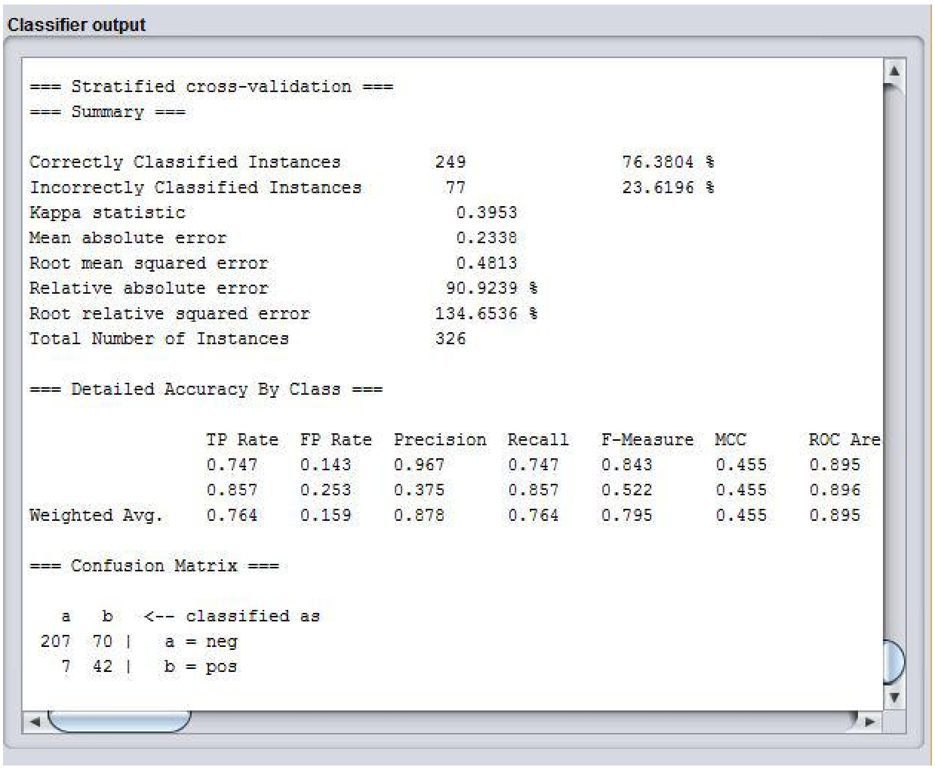
```
import matplotlib.pyplot as plt

from IPython.display import Image, display

fig, ax = plt.subplots(figsize=(5, 5))
ax.matshow(cm, cmap=plt.cm.Reds, alpha=0.3)
for i in range(cm.shape[0]):
    for j in range(cm.shape[1]):
        ax.text(x=j, y=i,
               s=cm[i, j],
               va='center', ha='center')
plt.xlabel('Predicted Values', )
plt.ylabel('Actual Values')
plt.show()
print(classification_report(y_test, y_pred ))
```

6. RESULTS AND PERFORMANCE ANALYSIS

In this section, the accuracy,precision,recall,f-1score are calculated and the comparison is done between all three classifiers .The best one is selected and used for further deployment.

6.1 Naive Bayes Calculation and Results

After performing data preprocessing and feature extraction, we will retrieve the final set of features.



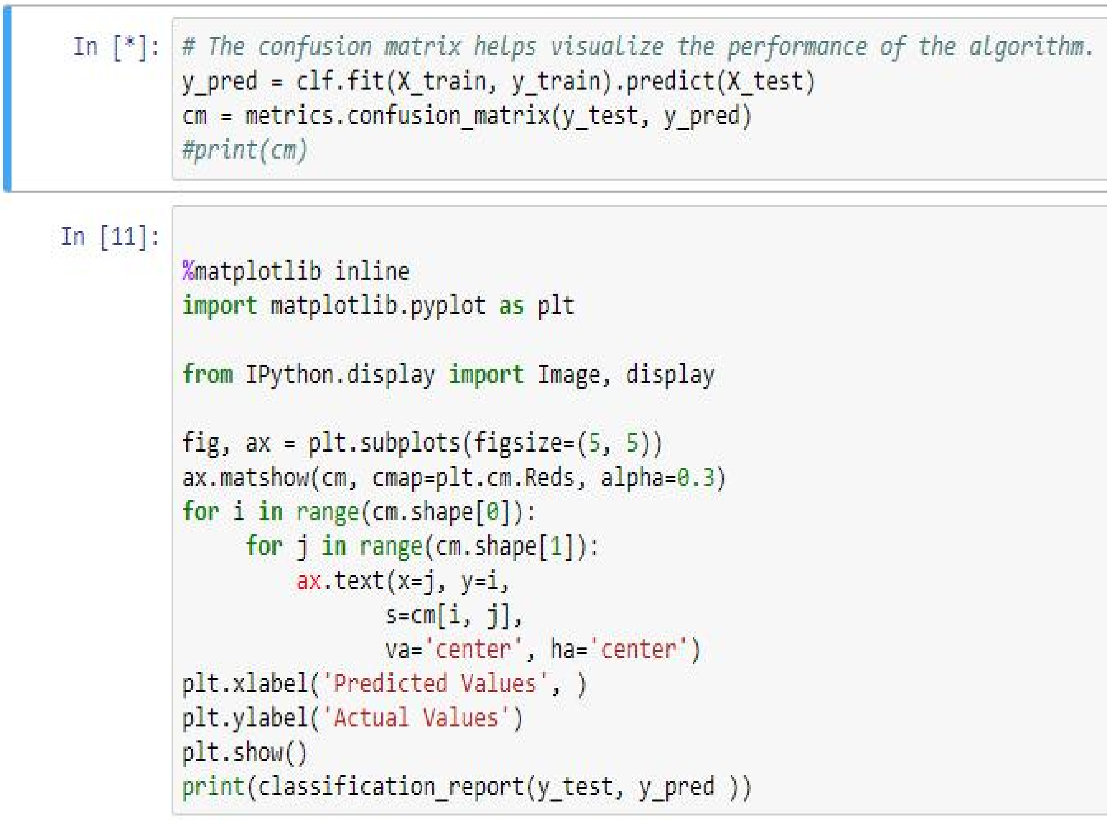
Classifier output

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         249               76.3804 %
Incorrectly Classified Instances        77               23.6196 %
Kappa statistic                          0.3953
Mean absolute error                      0.2338
Root mean squared error                  0.4813
Relative absolute error                 90.9239 %
Root relative squared error            134.6536 %
Total Number of Instances              326

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Are
                 0.747    0.143    0.967      0.747    0.843      0.455    0.895
                 0.857    0.253    0.375      0.857    0.522      0.455    0.896
Weighted Avg.    0.764    0.159    0.878      0.764    0.795      0.455    0.895

=== Confusion Matrix ===

   a   b   <-- classified as
 207  70 |   a = neg
   7  42 |   b = pos
```

Figure 6.1:Naive Bayes Calculation

The accuracy for the naive bayes classifier is 76.33% ,also the FP rate,IP rate,Precision,Recall all are caculated and displayed in figure 6.1.The naive bayes is directly implemented in weka no need for the code,confusion matrix is directly generated in weka.Now caculate the accuracy through SVM and logistic regression.

6.2 SVM Calculations and Results



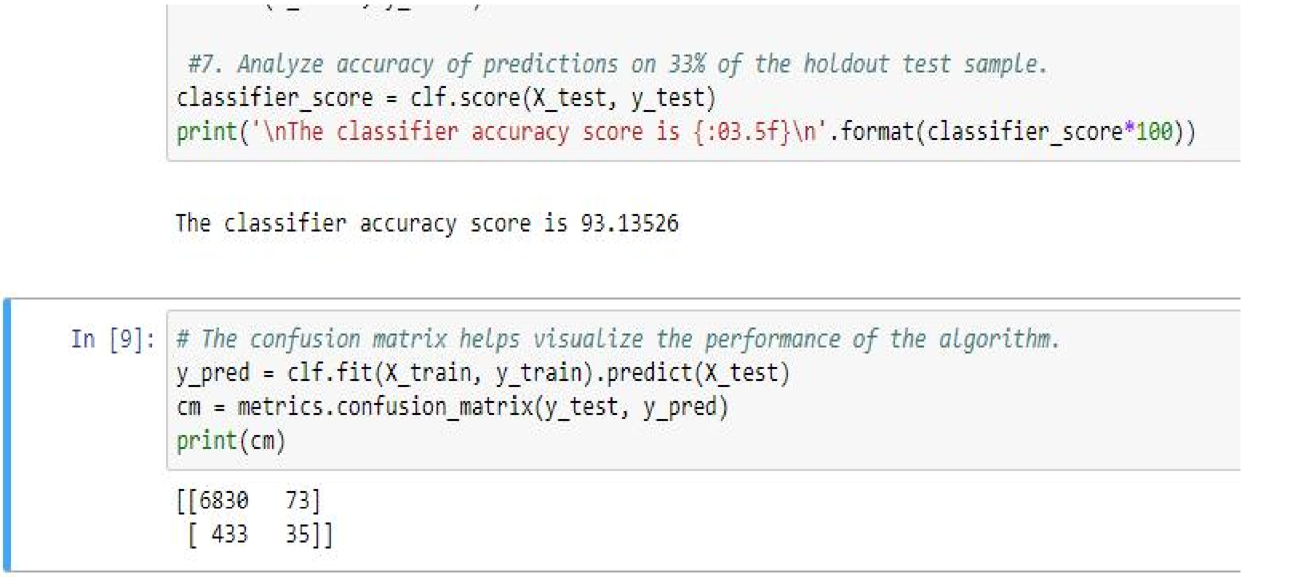
```
In [*]: # The confusion matrix helps visualize the performance of the algorithm.
        y_pred = clf.fit(X_train, y_train).predict(X_test)
        cm = metrics.confusion_matrix(y_test, y_pred)
        #print(cm)
```

```
In [11]:
        %matplotlib inline
        import matplotlib.pyplot as plt

        from IPython.display import Image, display

        fig, ax = plt.subplots(figsize=(5, 5))
        ax.matshow(cm, cmap=plt.cm.Reds, alpha=0.3)
        for i in range(cm.shape[0]):
             for j in range(cm.shape[1]):
                 ax.text(x=j, y=i,
                        s=cm[i, j],
                        va='center', ha='center')
        plt.xlabel('Predicted Values', )
        plt.ylabel('Actual Values')
        plt.show()
        print(classification_report(y_test, y_pred ))
```

Figure 6.2.1:Code Snippet for the visualization of algorithm efficiency

The Result of the classifier model can be visualized in figure 6.2.1 and figure 6.2.2. Figure 6.2.3 shows the Confusion Matrix generated and also the precision and recall after the execution of the Support Vector Machine.

As shown in Figure 6.2.2, the accuracy of the classifier is 93.135%..The precision rate and accuracy of SVM is very good as comparison to Naive Bayes,Because Naive Bayes works on probability and SVM works on hyperplane and good for those dataset which is very sensitive and margin of error is very less.



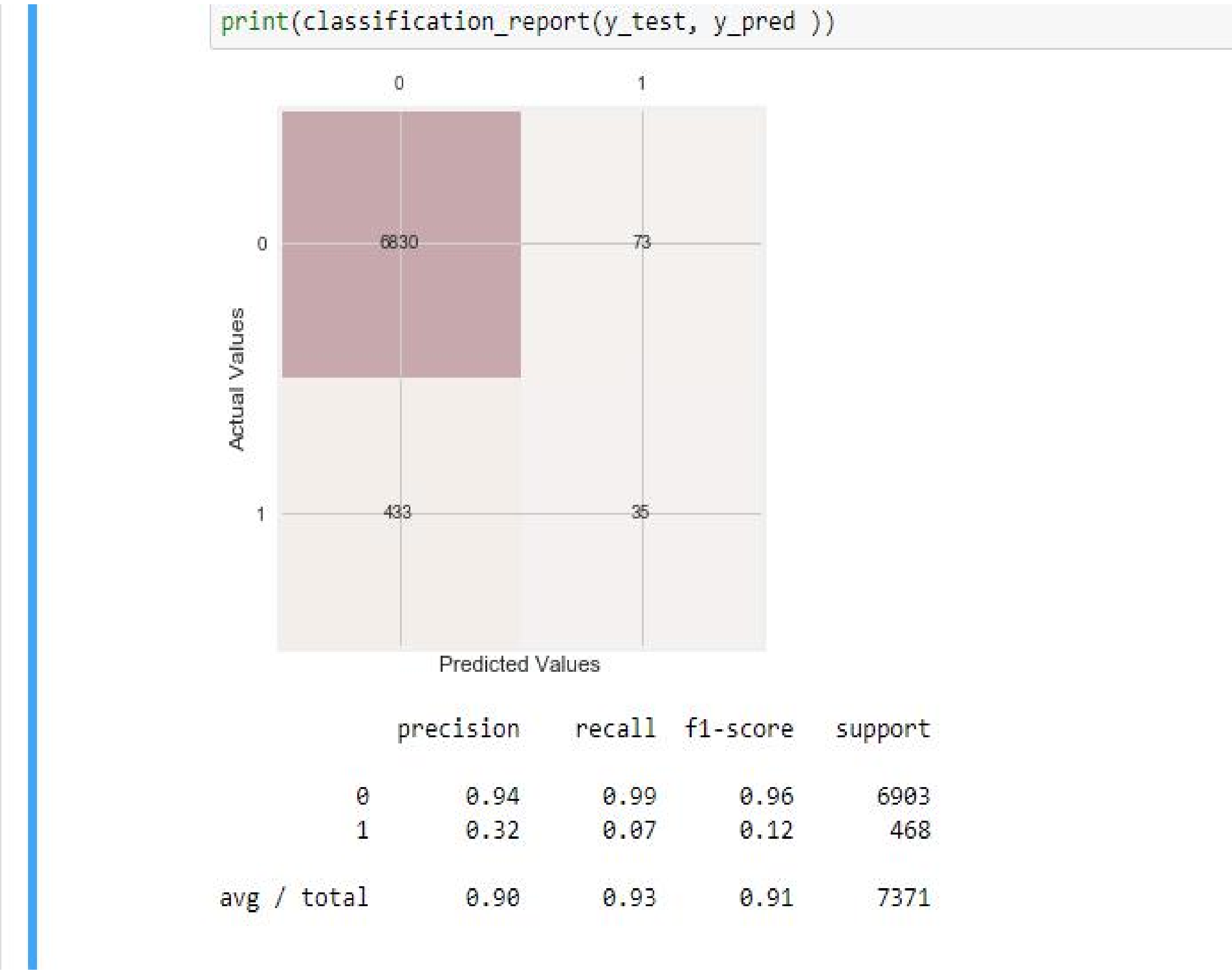
```
#7. Analyze accuracy of predictions on 33% of the holdout test sample.
classifier_score = clf.score(X_test, y_test)
print('\nThe classifier accuracy score is {:03.5f}\n'.format(classifier_score*100))
```

```
The classifier accuracy score is 93.13526
```

```
In [9]: # The confusion matrix helps visualize the performance of the algorithm.
        y_pred = clf.fit(X_train, y_train).predict(X_test)
        cm = metrics.confusion_matrix(y_test, y_pred)
        print(cm)
```

```
[[6830   73]
 [ 433   35]]
```

Figure 6.2.2: Results of the SVM classifier

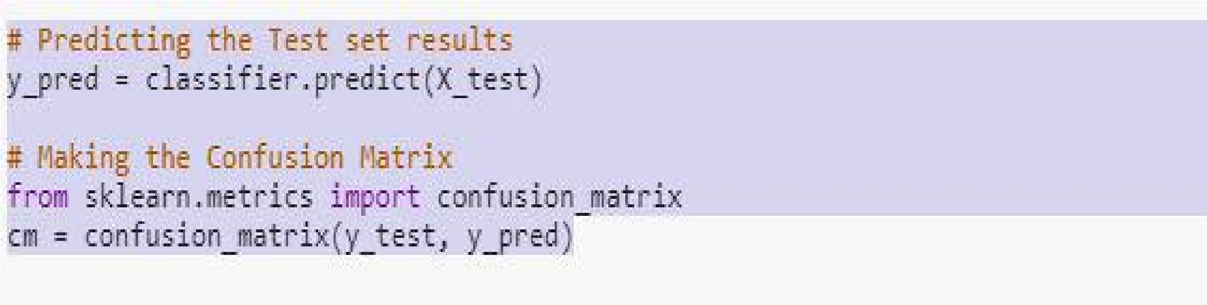


```
print(classification_report(y_test, y_pred ))
```

```
             precision    recall  f1-score   support

          0       0.94      0.99      0.96      6903
          1       0.32      0.07      0.12       468

avg / total       0.90      0.93      0.91      7371
```

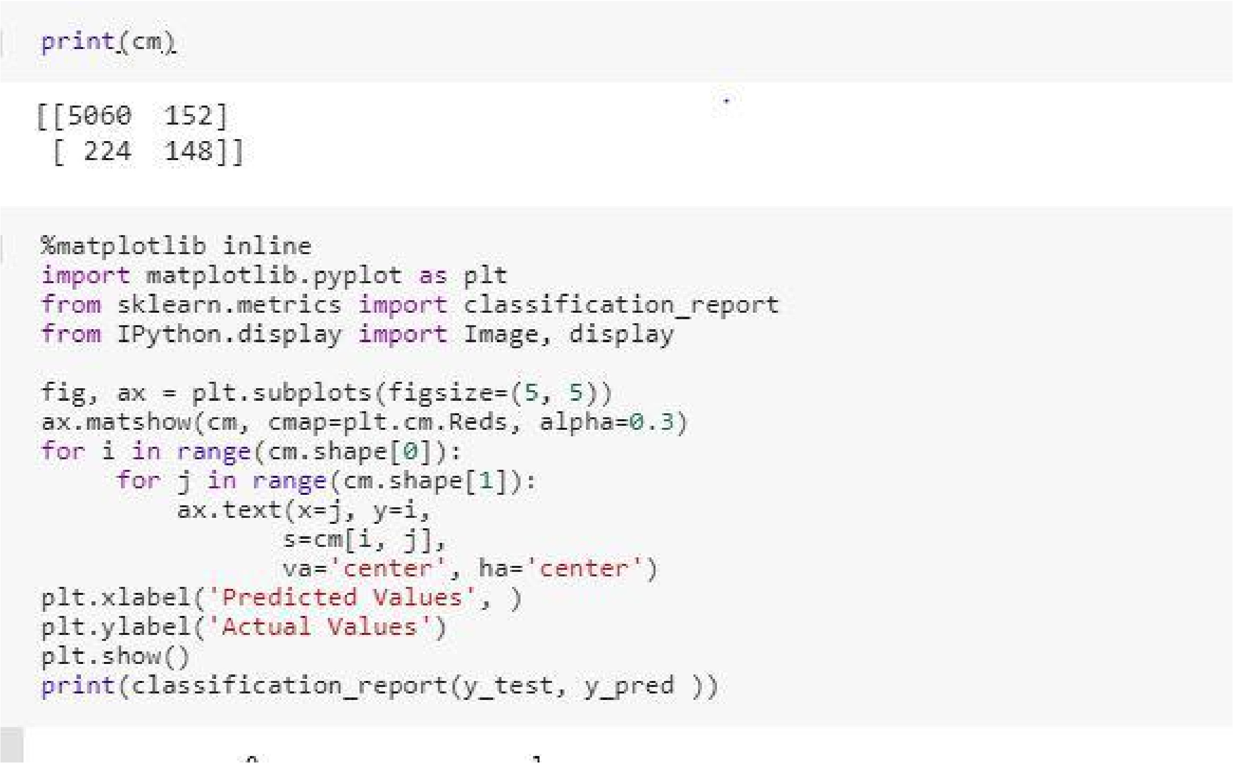
Figure 6.2.3: Visualization of the Confusion Matrix

6.3 Logistic Regression Calculations and Results



```
# Predicting the Test set results
y_pred = classifier.predict(X_test)

# Making the Confusion Matrix
from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, y_pred)
```

Figure 6.3.1:Confusion Matrix for logistic regression



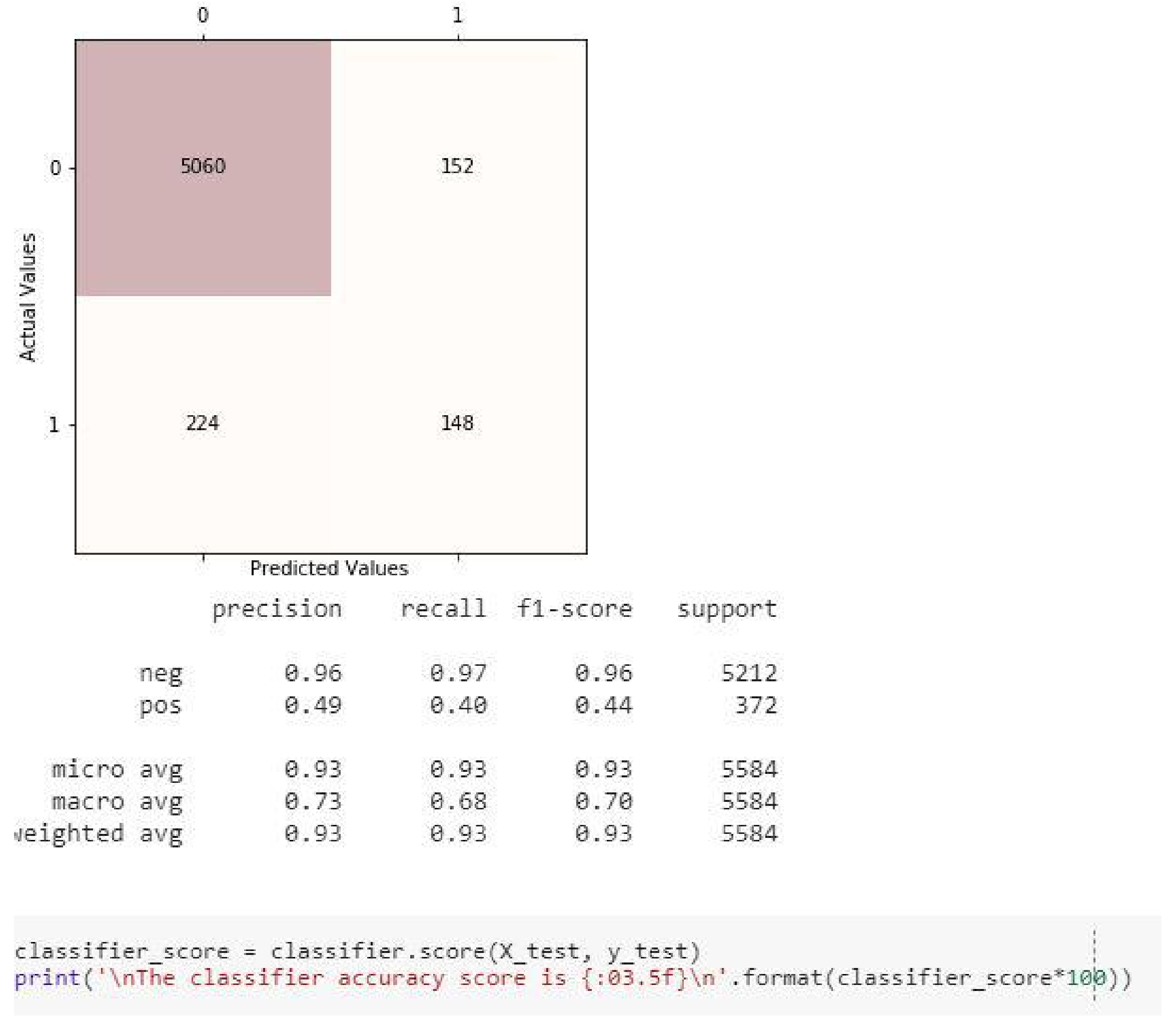
```
print(cm)
```

```
[[5060  152]
 [ 224  148]]
```

```
%matplotlib inline
import matplotlib.pyplot as plt
from sklearn.metrics import classification_report
from IPython.display import Image, display

fig, ax = plt.subplots(figsize=(5, 5))
ax.matshow(cm, cmap=plt.cm.Reds, alpha=0.3)
for i in range(cm.shape[0]):
     for j in range(cm.shape[1]):
         ax.text(x=j, y=i,
                s=cm[i, j],
                va='center', ha='center')
plt.xlabel('Predicted Values', )
plt.ylabel('Actual Values')
plt.show()
print(classification_report(y_test, y_pred ))
```

Figure 6.3.2:Results for Logistic Regression

The following figure 6.3.1,6.3.2,6.3.3 all are depicted about the calculations and result of logistic regression classifier.The precision and f-1 score for the logistic regression is .93and.93 and the accuracy for the logistic regression is 0.932 which is far better than the naive-bayes which has only 0.74 and marginally greater than the SVM.



```
              precision    recall  f1-score   support

         neg       0.96      0.97      0.96      5212
         pos       0.49      0.40      0.44       372

   micro avg       0.93      0.93      0.93      5584
   macro avg       0.73      0.68      0.70      5584
weighted avg       0.93      0.93      0.93      5584
```

```
classifier_score = classifier.score(X_test, y_test)
print('\nThe classifier accuracy score is {:03.5f}\n'.format(classifier_score*100))
```

Figure 6.3.3:Results for Logistic Regression

6.4 Performance Comparison

After the initialization of the models was the step to train the models. A timer was set for each of the model and the time was counted that that each model took for getting trained and making predictions. The models were then tested on the test dataset and the accuracy was noted. All of these details are shown below in the figure 6.4.

Taking in view the performance of each of the models, the fastest and the most accurate model that made the predictions was Logistic Regression. Therefore, Logistic was chosen as the primary model.

Naive Bayes

Accuracy:76.33%

F-measure:0.79

Recall:0.76

Precision:0.878

The accuracy will be calculated through weka and it is 74.33%

SVM

Precision:0.90

Recall:0.93

F-1 Score:0.91

Accuracy:93.1%

Logistic Regression

Precision:0.93

Recall:0.93

F-1 Score:0.93

Accuracy:93.2%

All the calculated parameters are shown above.Precision,accuracy,recall,f-1score all are calculated and the weighted average if all these are taken and comparison is mainly done on the basis of accuracy.Clearly it was seen that logistic regression having highest accuracy of 93.2%. which is good and optimized classifier to predict autism in human given the doctor report which states the symptoms of the patient.As shown in figure 6.4 the exponential rise in logistic regression will help to maintain the accuracy of the classifier.

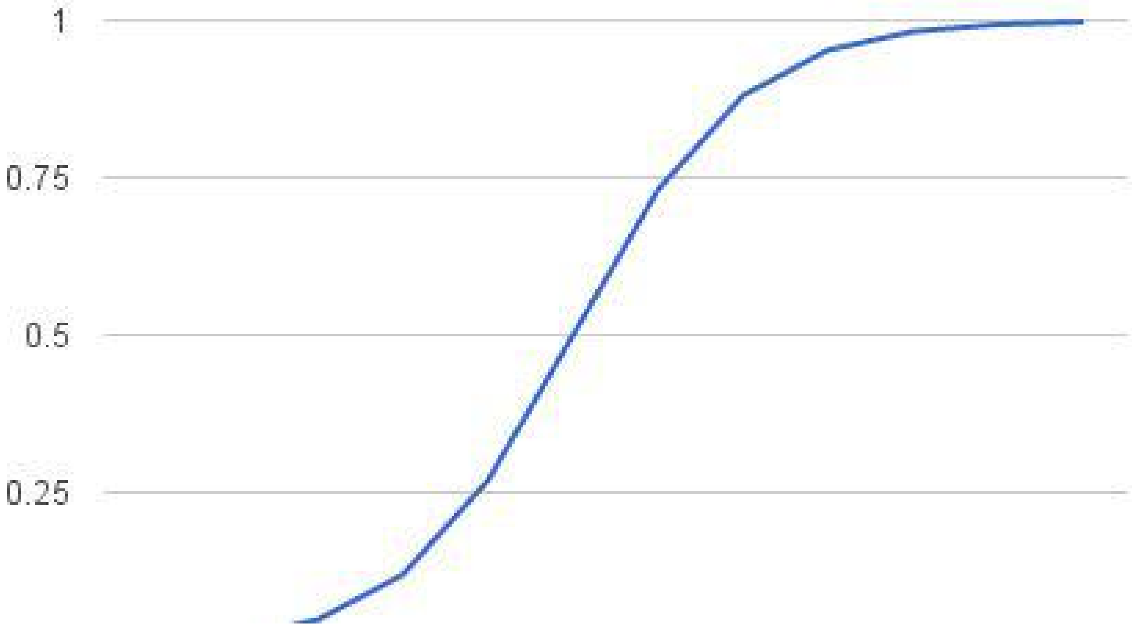


Figure 6.4:Logistic Regression Curve

7. CONCLUSION

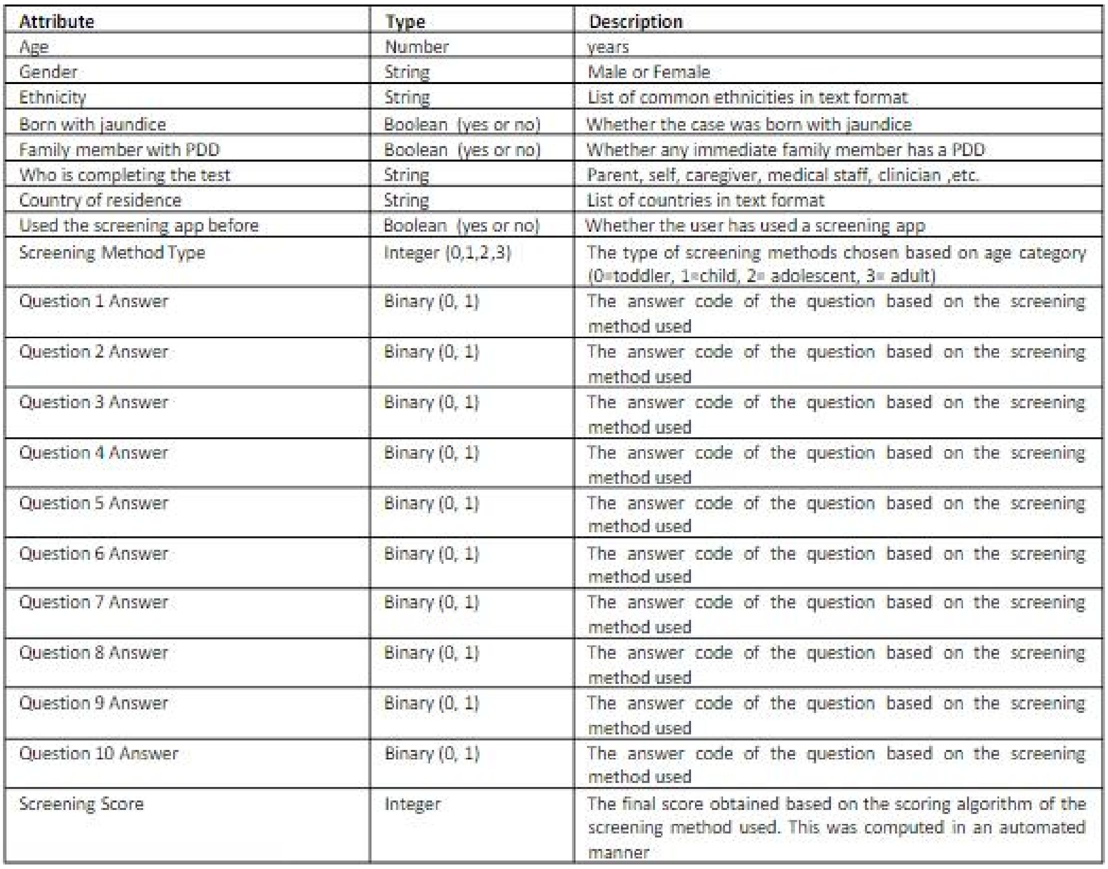
The Logistic algorithm performs well on classification and gave the best performance on making predictions of a patient having autism given a set of features. Therefore, out of all the three algorithms the Logistic algorithm was chosen. Logistic model was trained with a training set of size 18000. It was able to get trained in as less than as 0.4470 seconds. Predictions made by these model for the training set took a total time of 0.0160 seconds and the corresponding f1 score calculated for the model came out to be 0.93 and it got an accuracy of 0.9326.
Future work can be performed on this project; for example, to achieve much more optimization the parameters that are being used in the logistic could be tuned to get better results, along with the prediction of the autism other symptoms will be predicted so that no patient died due to delay in treatment and the results will produce in speedy manner.

Once diagnosed, this project also aims on the proper analysis of the Autistic traits using behavioral tests. A new dataset related to autism screening of children to be utilized for further analysis especially in determining influential autistic traits by calculating a score and improving the classification of ASD cases. In this behavioral test, we record ten behavioral features (ex- how the child responds when you call his/her name) plus ten individuals characteristics (ex- if the child was born with jaundice) that help in detecting the ASD cases from controls in behavior science.

The dataset use for the calculation of the screening source and the identification of the autistic traits is described in Figure 7.1.
Number of Instances in the second dataset (records in your data set): 292
Number of Attributes in the second dataset(fields within each record): 21

The Questions mentioned in the description above would somewhat be like:

1) If the child looks up when called his/her name,
2) How easy it is to get eye contact with the child, etc.



Attribute	Type	Description
Age	Number	years
Gender	String	Male or Female
Ethnicity	String	List of common ethnicities in text format
Born with jaundice	Boolean (yes or no)	Whether the case was born with jaundice
Family member with PDD	Boolean (yes or no)	Whether any immediate family member has a PDD
Who is completing the test	String	Parent, self, caregiver, medical staff, clinician ,etc.
Country of residence	String	List of countries in text format
Used the screening app before	Boolean (yes or no)	Whether the user has used a screening app
Screening Method Type	Integer (0,1,2,3)	The type of screening methods chosen based on age category (0=toddler, 1=child, 2= adolescent, 3= adult)
Question 1 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 2 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 3 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 4 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 5 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 6 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 7 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 8 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 9 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Question 10 Answer	Binary (0, 1)	The answer code of the question based on the screening method used
Screening Score	Integer	The final score obtained based on the scoring algorithm of the screening method used. This was computed in an automated manner

Figure 7.1: Description of the second dataset used

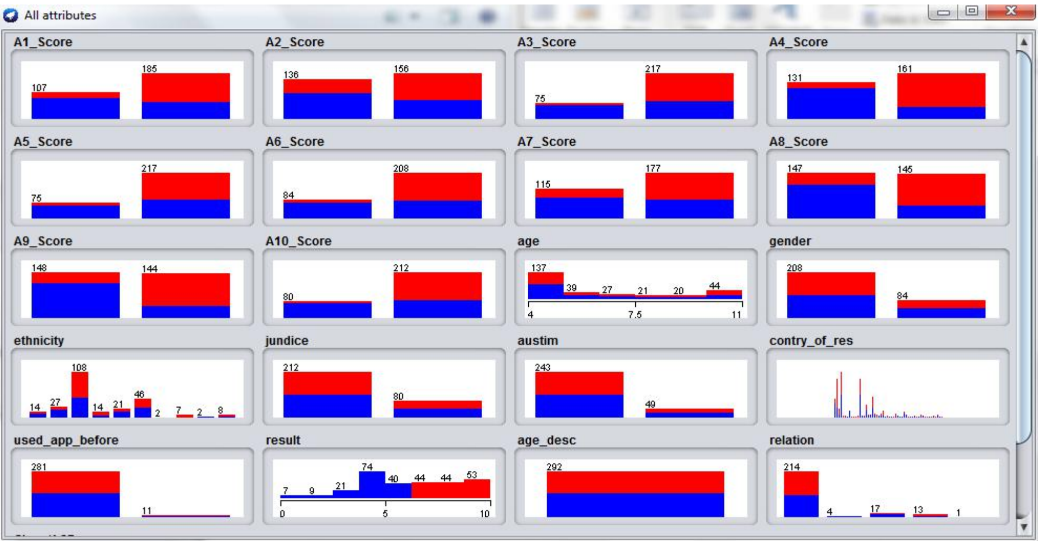


Figure 7.2: (Value, Frequency) Distribution of features in the second dataset

BIBLIOGRAPHY

1.Arya, Arpit, "Predicting Autism over Large-Scale Child Dataset" (2015).Master's Projects.

2..Tabtah, F. “Autism Spectrum Disorder Screening: Machine Learning Adaptation and DSM-5 Fulfillment.” (2017).

3.M Duda, J A Kosmicki, and D P Wall . “Testing the accuracy of an observation-based classifier for rapid detection of autism risk”, Nature Publishing Group .

4.Thabtah, F. (2017). ASDTests. A mobile app for ASD screening. www.asdtests.com [accessed December 20th,2017]

5.Keyword aware system recommendation/ieee.org

Web

1.Vaccine adverse Event Reporting System

2.Autism Spectrum Disorder,Wikipedia

3.Weka text classification,Youtube