Miscarriage Prediction

(Spontaneous Abortion Prediction)

Name:	Manju Yadav
Subject:	Machine Learning
Teachers Name:	Prof. Maria Achary

Acknowledgement

I would like to express my deep gratitude to Professor Maria Achary, my research supervisor, for her patient guidance, enthusiastic encouragement and useful critiques of this research work.

Finally, I wish to thank my parents for their support and encouragement throughout my study.

I am thankful for their aspiring guidance, invaluably constructive criticism and friendly advice during the project work. I am sincerely grateful to them for sharing their truthful and illuminating views on a number of issues related to the project.

Index

Introduction

Around 70% of miscarriages reported in Maharashtra every year are from rural areas, which has 55% of state's population (as per 2011 census), revealed data available with public health department.

Updated: Jun 05, 2018 12:03 IST Sadaguru Pandit Hindustan Times

Gynecologist Nikhil Datar said miscarriage, being a multifactorial issue, causes cannot merely be related to availability of medical services and infrastructure. "In 95% cases, which take place in first trimester of pregnancy, chances of its survival are low. Hence, one has to look at the data with perspective of total number of pregnancies. There is a possibility that because number of pregnancies in rural parts are high, they report more cases of miscarriages"

What do we understand about Spontaneous Abortion?

Well it means the death of fetus, it increases the risk of of having an abortion, fetal death or early delivery is defined as follows: Death of the fetus or passage of products of conception (fetus and placenta) before 20 weeks gestation.

Threatened abortion is vaginal bleeding without cervical dilation occurring during this time frame and indicating that spontaneous abortion may occur in a woman with a confirmed viable intrauterine pregnancy.

Diagnosis is by clinical criteria and ultrasonography. Treatment is usually expectant observation for threatened abortion and, if spontaneous abortion has occurred or appears unavoidable, observation or uterine evacuation.

Fetal death and early delivery are classified as follows:

- Abortion: Death of the fetus or passage of products of conception (fetus and placenta) before 20 weeks gestation
- Fetal demise (stillbirth): Fetal death after 20 weeks
- Preterm delivery: Passage of a live fetus between 20 weeks and 36 weeks/6 days

Abortions may be classified as follows:

- Early or late
- Spontaneous or induced for therapeutic or elective reasons
- Threatened or inevitable
- Incomplete or complete
- Recurrent (also called recurrent pregnancy loss)
- Missed
- Septic

About 20 to 30% of women with confirmed pregnancies bleed during the first 20 weeks of pregnancy; half of these women spontaneously abort. Thus, incidence of spontaneous abortion is up to about 20% in confirmed pregnancies. Incidence in all pregnancies is probably higher because some very early abortions are mistaken for a late menstrual period.

Symptoms and signs:

- Crampy pelvic pain , bleeding & eventually expulsion of tissue.

- Late spontaneous abortion begins with a gush of fluid when membranes rupture
- Hemorrhage is rarely massive
- A dilated cervix indicates that abortion is inevitable.

If products of conception remain in the uterus after spontaneous abortion, vaginal bleeding may occur, sometimes after a delay of hours to days. Infection may also develop, causing fever, pain, and sometimes sepsis (called septic abortion).

A woman has a higher risk of miscarriage if she:

- Is over age 35
- Has certain diseases, such as diabetes or thyroid problems
- Has had three or more miscarriages

Cervical Insufficiency

A miscarriage sometimes happens because there is a weakness of the cervix, called a cervical insufficiency, which means the cervix cannot hold the pregnancy. A miscarriage from a cervical insufficiency usually occurs in the second trimester.

There are usually few symptoms before a miscarriage caused by cervical insufficiency. A woman may feel sudden pressure, her "water" may break, and tissue from the fetus and placenta may be expelled without much pain. An insufficient cervix can usually be treated with a "circling" stitch in the cervix in the next pregnancy, usually around 12 weeks. The stitch holds the cervix closed until it is pulled out around the time of delivery. The stitch may also be placed even if there has not been a previous miscarriage if cervical insufficiency is discovered early enough, before a miscarriage does occur

Threatened miscarriage

When your body is showing signs that you might miscarry, that is called a ‘threatened miscarriage’. You may have a little vaginal bleeding or lower abdominal pain. It can last days or weeks and the cervix is still closed.

- The pain and bleeding may go away and you can continue to have a healthy pregnancy and baby. Or things may get worse and you go on to have a miscarriage.
- There is rarely anything a doctor, midwife or you can do to protect the pregnancy. In the past bed rest was recommended, but there is no scientific proof that this helps at this stage.

Inevitable miscarriage

Inevitable miscarriages can come after a threatened miscarriage or without warning. There is usually a lot more vaginal bleeding and strong lower stomach cramps. During the miscarriage your cervix opens and the developing fetus will come away in the bleeding.

Complete miscarriage

A complete miscarriage has taken place when all the pregnancy tissue has left your uterus. Vaginal bleeding may continue for several days. Cramping pain much like labour or strong period pain is common – this is the uterus contracting to empty.

Incomplete miscarriage

Sometimes, some pregnancy tissue will remain in the uterus. Vaginal bleeding and lower abdominal cramping may continue as the uterus continues trying to empty itself. This is known as an ‘incomplete miscarriage’.

Missed miscarriage

Sometimes, the baby has died but stayed in the uterus. This is known as a ‘missed miscarriage’.

If you have a missed miscarriage, you may have a brownish discharge. Some of the symptoms of pregnancy, such as nausea and tiredness, may have faded. You might have noticed nothing unusual. You may be shocked to have a scan and find the baby has died.

Recurrent miscarriage

A small number of women have repeated miscarriages. If this is your third or more miscarriage in a row, it’s best to discuss this with your doctor who may be able to investigate the causes, and refer you to a specialist.

Types of pregnancy loss

Other types of pregnancies that result in a miscarriage are outlined below.

Ectopic pregnancy

An ectopic pregnancy occurs when the embryo implants outside the uterus, usually in one of the fallopian tubes. A fetus does not usually survive an ectopic pregnancy.

Molar pregnancy

A molar pregnancy is a type of pregnancy that fails to develop properly from conception. It can be either complete or partial and usually needs to be surgically removed.

Blighted ovum

With a blighted ovum the sac develops but there is no baby inside. It is also known as an ‘anembryonic pregnancy’.

This condition is usually discovered during a scan. In most cases, an embryo was conceived but did not develop and was reabsorbed into the uterus at a very early stage.

Misdiagnosis and etc

There is also misdiagnosis of spontaneous miscarriage where its evitable that there was no err and there would be a lot of factors affecting to it, as there is sometimes also infrastructure or even misdiagnosis on the doctors part, there’s internal and external factors relating to it where a diet should be followed and some things should be avoided at all cost and these consumption factors can be like tobacco, drugs, alcohol etc, even stress is a factor which leads to the situation

3. Review of the project

The main goal of this project is to understand the factors affecting the miscarriage and how it would lead to loosing of the fetus as there is only diagnosis based where the doctor declares of spontaneous miscarriage.

Here the data collected is from rurals of india and where its data collected from many states and the data is used asper.

This dataset contains data on Annual Health Survey :

Woman Schedule.

Woman Schedule comprised two sections. Section-I (this dataset) contains information relating to the outcome of pregnancy(s) (live birth/still birth/abortion); birth history; type of medical attention at delivery; details of maternal health care(ante-natal/natal/post-natal); immunization of children; breast feeding practices including supplements; occurrence of child diseases (Pneumonia, Diarrhoea and fever); registration of births, etc. use, sources and practices of family planning methods; details relating to future use of contraceptives and unmet need;awareness about RTI/STI, HIV/AIDS, administration of HAF/ORT/ORS during diarrhoea and danger signs of ARI/Pneumonia.

It also includes more information relating the Ever Married Women (EMW) like conception details, usage of NPT kit, registration of pregnancy, health problems and subsequent treatments during ante-natal/natal/post-natal period, cost incurred by the woman during delivery etc.

Algorithms study

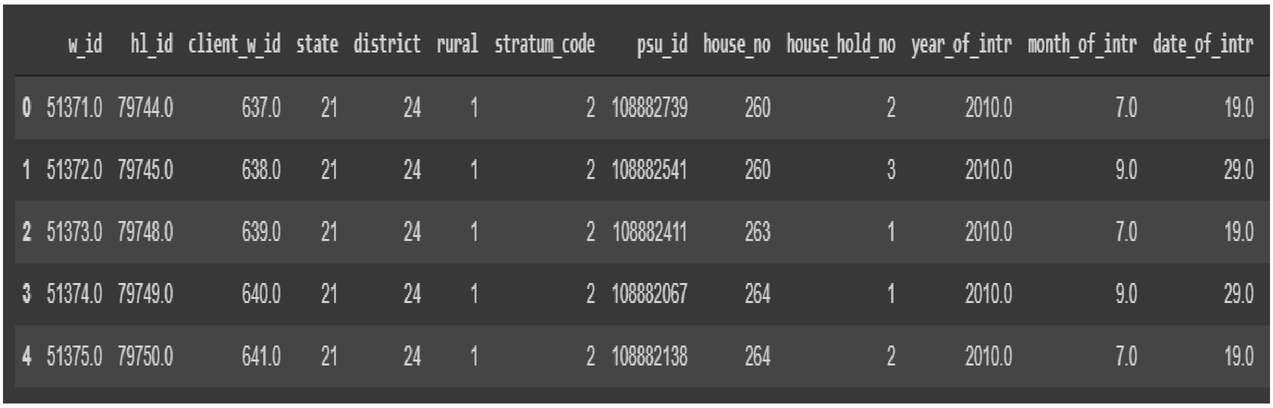
Here used machine learning algorithms such as Logistic Regression and Decision tree algorithm and Random forest applied / implemented the analysis in python to predict the outcome of pregnancy on the factors based on illness, consumption of alcohol and etc where it affects the pregnancy at some level.

This is a classification problem with a variety of input feature parameters.

Description of datasets

Have used here data from Kaggle where its named as Prediction of outcome of pregnancy where data is of 9 states where the rows are 15258 and 201 attributes where further on moving the data was converted into Boolean and nominal variables and data further sorted according to the prediction of the survival of pregnancy or miscarriage.

The file contain the following fields:



	w_id	hl_id	client_w_id	state	district	rural	stratum_code	psu_id	house_no	house_hold_no	year_of_intr	month_of_intr	date_of_intr
0	51371.0	79744.0	637.0	21	24	1	2	108882739	260	2	2010.0	7.0	19.0
1	51372.0	79745.0	638.0	21	24	1	2	108882541	260	3	2010.0	9.0	29.0
2	51373.0	79748.0	639.0	21	24	1	2	108882411	263	1	2010.0	7.0	19.0
3	51374.0	79749.0	640.0	21	24	1	2	108882067	264	1	2010.0	9.0	29.0
4	51375.0	79750.0	641.0	21	24	1	2	108882138	264	2	2010.0	7.0	19.0

Parameters considered

1. age:	age of woman
2. surviving_total:	survival of the fetus
3. mother_age_when_baby_was_born:	mothers age during birth of child
4. outcome_pregnancy:	nominal survival variables
5. is_currently_pregnant:	woman is pregnant or not
6. aware_of_the_danger_signs:	aware of danger signs
7. occupation_status:	occupation of the mother
8. disability_status:	if mother is disable or not
9. injury_treatment_type:	treatment taken for injury
10. illness_type:	mother has an illness
11. diagnosed_for:	illness diagnosis
12. regular_treatment:	took treatment for illness
13. chew:	consumption of tabacco
14. smoke:	smoking
15. alcohol:	consumption of alcohol

Which algorithms are used & why?

1. Logistic Regression

LR is one of the most popular methods used to classify binary data. LR is based on the assumption that the value of dependent variable is predicted by using independent variables. In the model, Y is the dependent variable we are trying to predict by observing X which is the input or set of the independent variables (xi, …, xk). The value of Y that corresponds to the fetus has either survived (Y=1) or not survived (Y=0) and is summarized by (X=x). From this definition, the conditional probability follows a logistic distribution given by P (Y = 1|X =xi). This function called as regression function we need to predict Y.

Were using this algorithm as its implementation follows dataset in Boolean where it was easy to achieve here already and the dataset easily could be applied

2. Decision Tree

Decision trees with their fairly simple structure to create are one of the most used classifiers. A decision tree is a tree structured model with decision nodes and prediction nodes. Decision nodes are used to branch and prediction nodes specify class labels. C4.5 is a kind of decision tree algorithm builds a decision tree from training data by using the information gain. When building decision trees C4.5 uses divide and conquer approach.

Here it was easy to apply as the dataset wasn't required to be defined here and further classification could be applied and the data applied is easy to manipulate.

3. Random Forest

RF is a classification algorithm developed by Bierman and Cutler that uses an ensemble of tree predictors. It is one of the most accurate learning algorithms and for many datasets; it achieves a highly accurate classifier. In RF, each tree is constructed by bootstrapping the training data and for each split randomly selected subset of features are used. Splitting is made based on purity measure. This classification method estimates missing data and large proportion of the data are missing it still maintains accuracy.

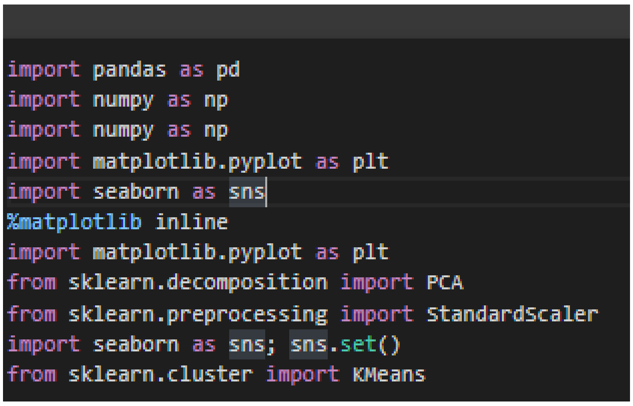
Here we used this because the data defined here could be easily used here no new definition is required of the dataset or grouping and etc no need to apply.

Comparative study of algorithm and accuracy

Here the implementation of the data is done shown how its applied and which festures are highlighted here for understanding the data and dataset altogether.

Implementation

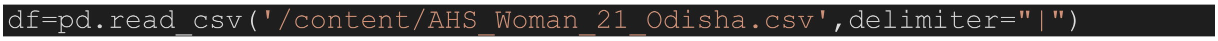
1. Importing libraries
2. NumPy: works with arrays
3. pandas: works with CSV files and data frames
4. matplotlib: Creates charts using pyplot, define parameters using rcParams and color them with cm.rainbow
5. seaborn: For a high-level interface for drawing attractive and informative statistical graphics.



```
import pandas as pd
import numpy as np
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
import seaborn as sns; sns.set()
from sklearn.cluster import KMeans
```

Importing Dataset

After downloading the dataset from Kaggle, saved the Data file into the working directory with the name ‘AHS_woman_odisa’ and imported it using pandas read_csv() function and stored it into an object (df1).



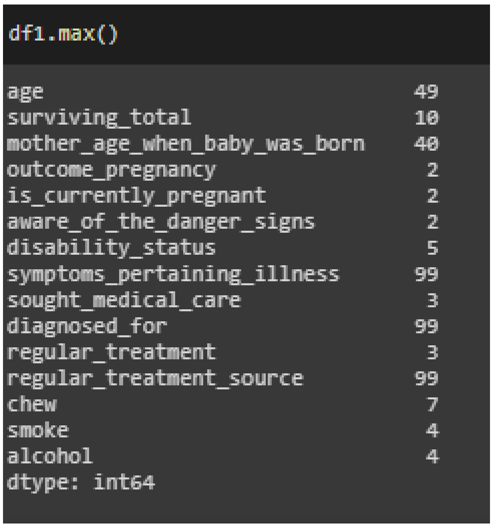
```
df=pd.read_csv('/content/AHS_Woman_21_Odisha.csv',delimiter="|")
```

(15258, 201) Shape of the dataset

Where after deleting columns and having only which is important hence then editing in excel and uploading the file as df1

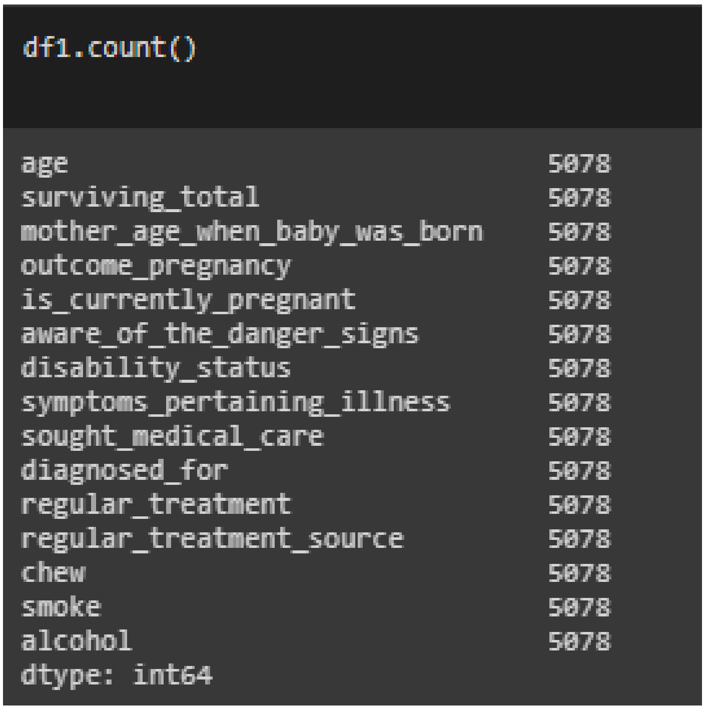
Data Exploration

Here, checked the max, sample and description of dataset.



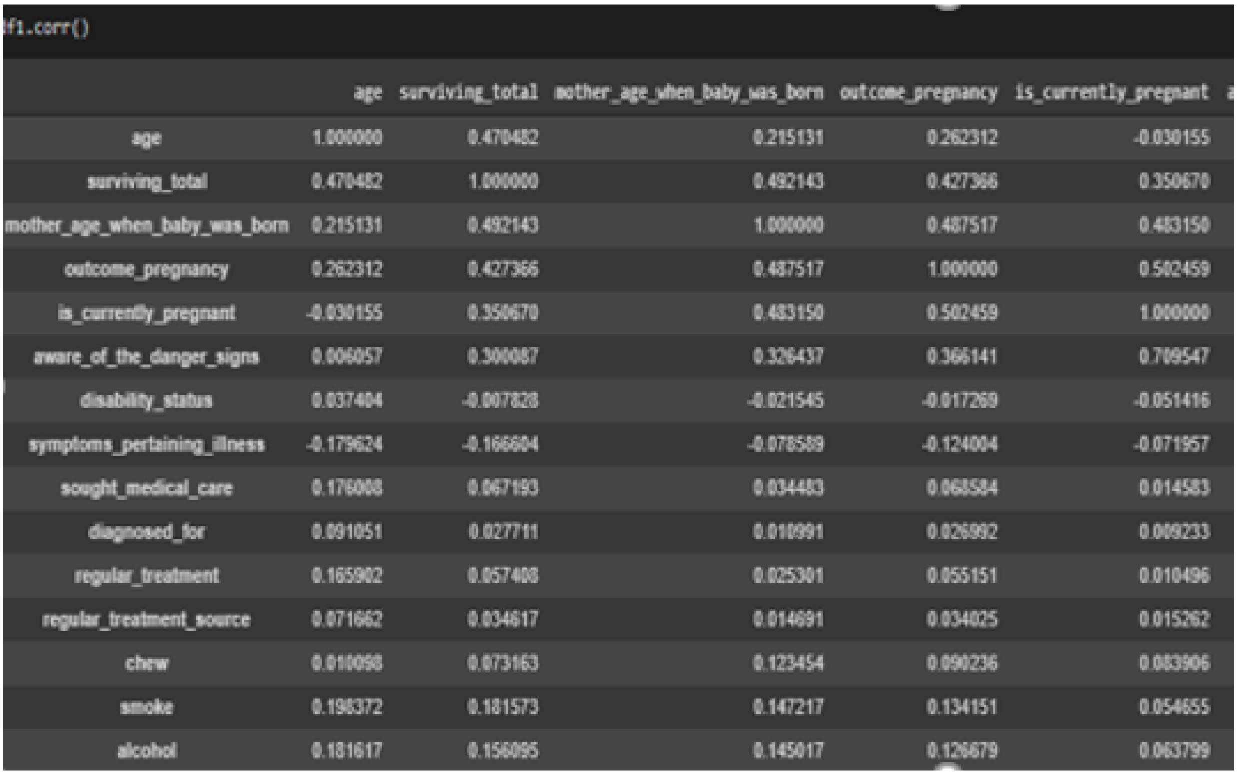
```
df1.max()

age                              49
surviving_total                  10
mother_age_when_baby_was_born    40
outcome_pregnancy                 2
is_currently_pregnant             2
aware_of_the_danger_signs         2
disability_status                 5
symptoms_pertaining_illness      99
sought_medical_care               3
diagnosed_for                    99
regular_treatment                 3
regular_treatment_source         99
chew                              7
smoke                             4
alcohol                           4
dtype: int64
```

Here it’s the count of the data present in the dataset



```
df1.count()

age                              5078
surviving_total                  5078
mother_age_when_baby_was_born    5078
outcome_pregnancy                5078
is_currently_pregnant            5078
aware_of_the_danger_signs        5078
disability_status                5078
symptoms_pertaining_illness      5078
sought_medical_care              5078
diagnosed_for                    5078
regular_treatment                5078
regular_treatment_source         5078
chew                             5078
smoke                            5078
alcohol                          5078
dtype: int64
```

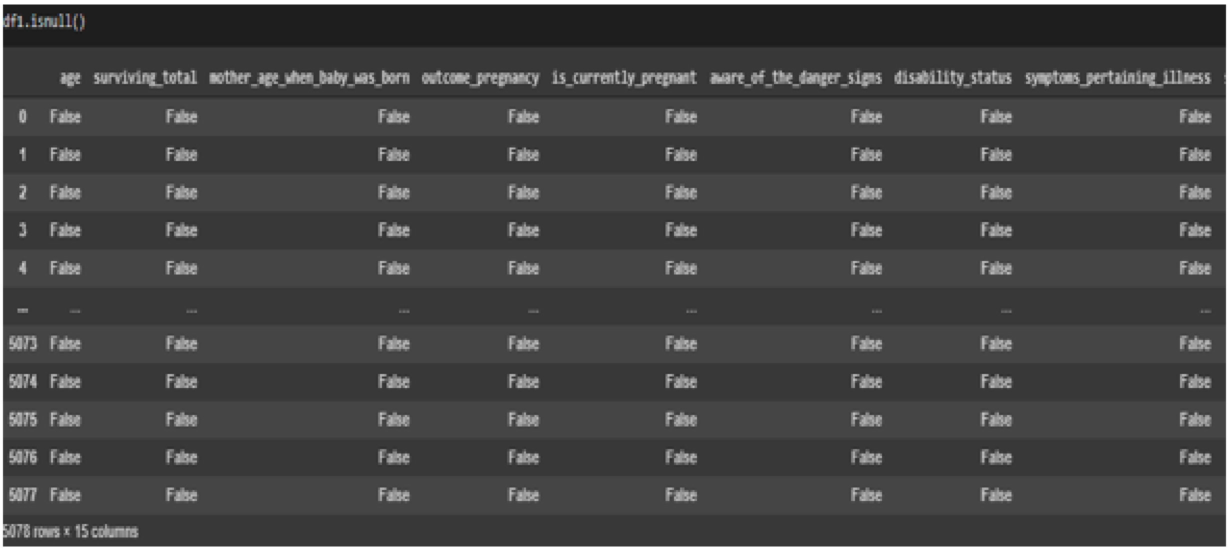
Here have checked the correlation between the variables/attributes



```
df1.corr()
```

	age	surviving_total	mother_age_when_baby_was_born	outcome_pregnancy	is_currently_pregnant
age	1.000000	0.470482	0.215131	0.262312	-0.030155
surviving_total	0.470482	1.000000	0.492143	0.427366	0.350670
mother_age_when_baby_was_born	0.215131	0.492143	1.000000	0.487517	0.483150
outcome_pregnancy	0.262312	0.427366	0.487517	1.000000	0.502459
is_currently_pregnant	-0.030155	0.350670	0.483150	0.502459	1.000000
aware_of_the_danger_signs	0.006057	0.300087	0.326437	0.366141	0.709547
disability_status	0.037404	-0.007828	-0.021545	-0.017269	-0.051416
symptoms_pertaining_illness	-0.179624	-0.166604	-0.078589	-0.124004	-0.071957
sought_medical_care	0.176008	0.067193	0.034483	0.068584	0.014583
diagnosed_for	0.091051	0.027711	0.010991	0.026992	0.009233
regular_treatment	0.165902	0.057408	0.025301	0.055151	0.010496
regular_treatment_source	0.071662	0.034617	0.014691	0.034025	0.015262
chew	0.010098	0.073163	0.123454	0.090236	0.083906
smoke	0.198372	0.181573	0.147217	0.134151	0.054655
alcohol	0.181617	0.156095	0.145017	0.126679	0.063799

Checked values if null/not

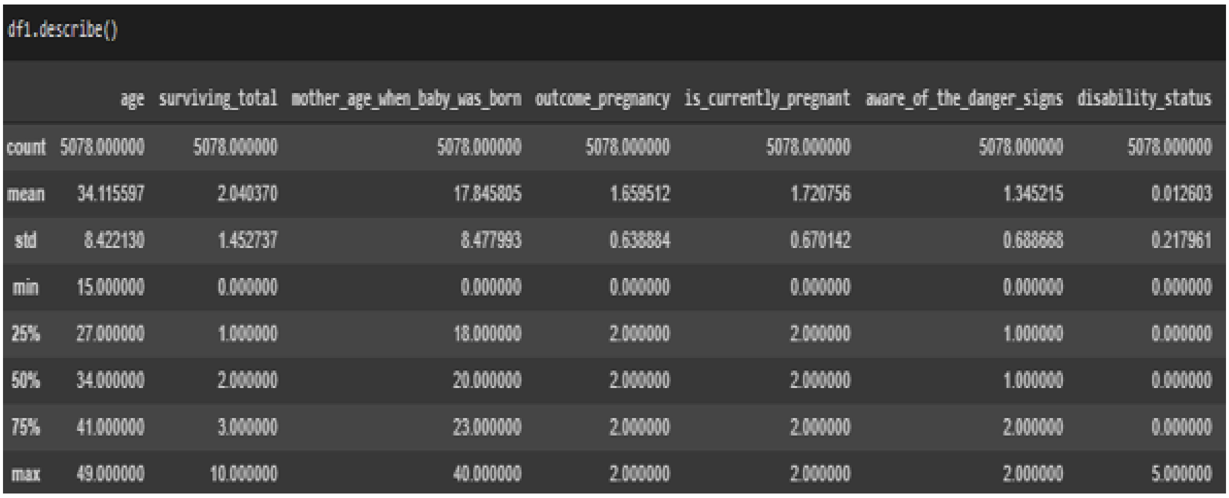


```
df1.isnull()
```

	age	surviving_total	mother_age_when_baby_was_born	outcome_pregnancy	is_currently_pregnant	aware_of_the_danger_signs	disability_status	symptoms_pertaining_illness
0	False	False	False	False	False	False	False	False
1	False	False	False	False	False	False	False	False
2	False	False	False	False	False	False	False	False
3	False	False	False	False	False	False	False	False
4	False	False	False	False	False	False	False	False
...	...	...	...	...	...	...	...	...
5073	False	False	False	False	False	False	False	False
5074	False	False	False	False	False	False	False	False
5075	False	False	False	False	False	False	False	False
5076	False	False	False	False	False	False	False	False
5077	False	False	False	False	False	False	False	False

5078 rows × 15 columns

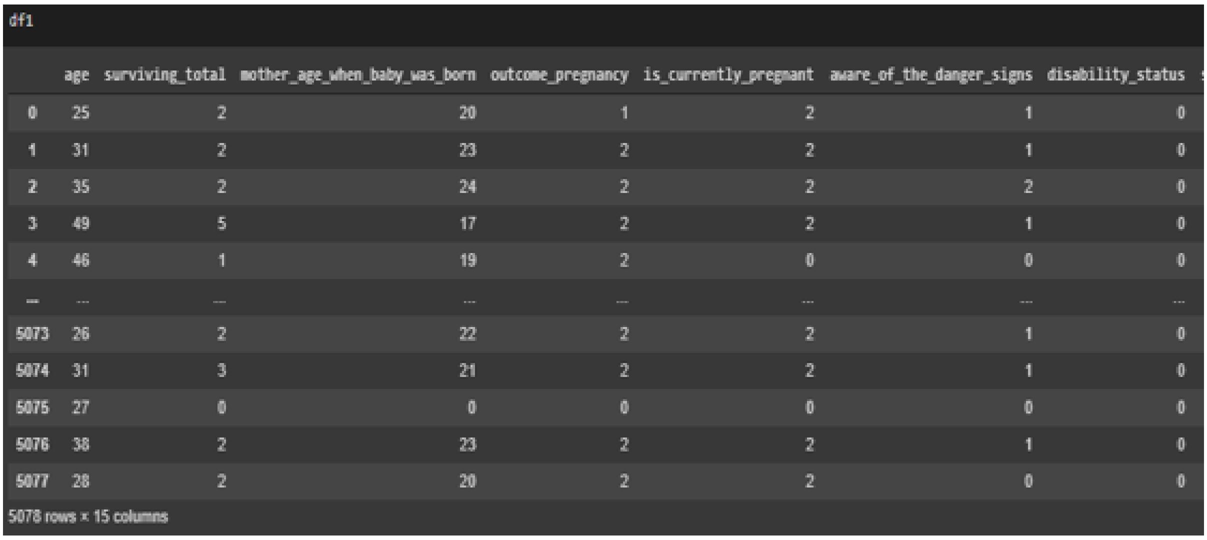
Here we see description of the dataset df1



```
df1.describe()
```

	age	surviving_total	mother_age_when_baby_was_born	outcome_pregnancy	is_currently_pregnant	aware_of_the_danger_signs	disability_status
count	5078.000000	5078.000000	5078.000000	5078.000000	5078.000000	5078.000000	5078.000000
mean	34.115597	2.040370	17.845805	1.659512	1.720756	1.345215	0.012603
std	8.422130	1.452737	8.477993	0.638884	0.670142	0.688668	0.217961
min	15.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	27.000000	1.000000	18.000000	2.000000	2.000000	1.000000	0.000000
50%	34.000000	2.000000	20.000000	2.000000	2.000000	1.000000	0.000000
75%	41.000000	3.000000	23.000000	2.000000	2.000000	2.000000	0.000000
max	49.000000	10.000000	40.000000	2.000000	2.000000	2.000000	5.000000

Further sorting the dataset and deleting /dropping the unwanted attributes / preprocessing the data led to the data having 20 attributes, and also removing the Nan values and dropping and converting the values using advanced excel tricks to easily survey the extracted dataset and understand what nots of it.



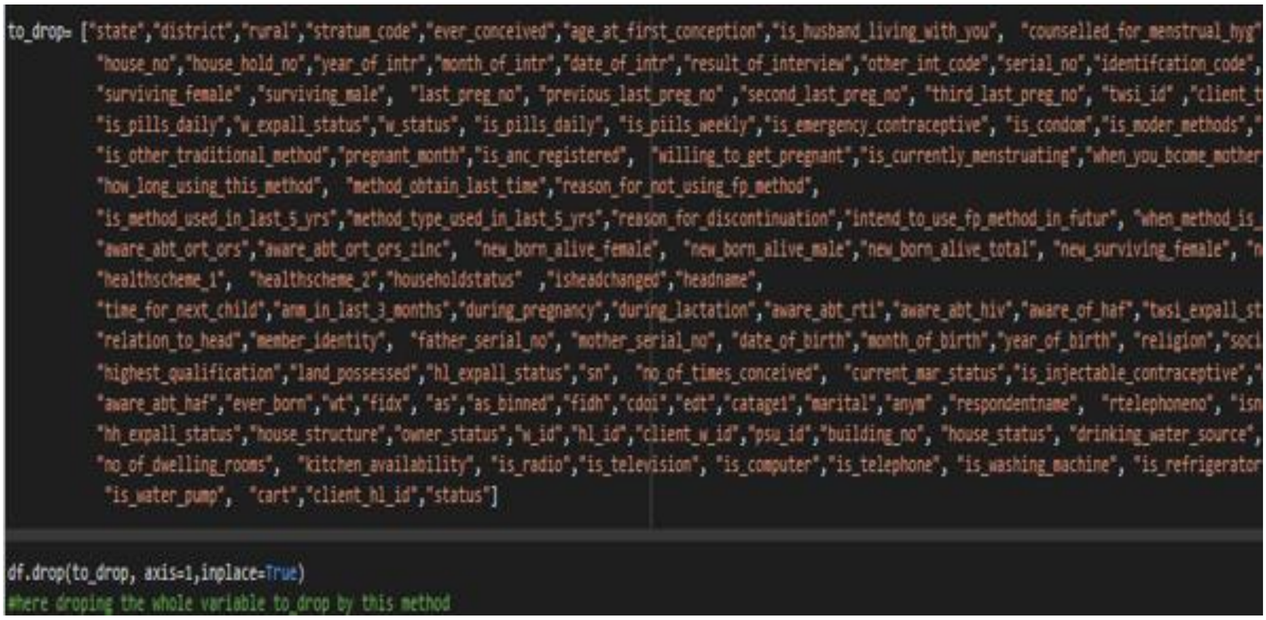
```
df1
```

	age	surviving_total	mother_age_when_baby_was_born	outcome_pregnancy	is_currently_pregnant	aware_of_the_danger_signs	disability_status
0	25	2	20	1	2	1	0
1	31	2	23	2	2	1	0
2	35	2	24	2	2	2	0
3	49	5	17	2	2	1	0
4	46	1	19	2	0	0	0
...	...	...	...	...	...	...	...
5073	26	2	22	2	2	1	0
5074	31	3	21	2	2	1	0
5075	27	0	0	0	0	0	0
5076	38	2	23	2	2	1	0
5077	28	2	20	2	2	0	0

5078 rows × 15 columns

I. Preprocessing Data

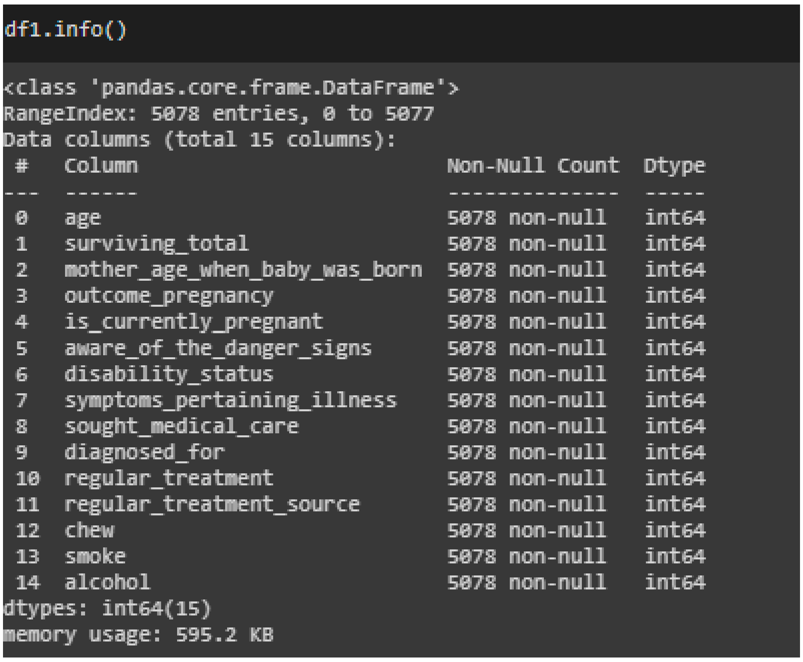
Here dropping all the columns which do not relate to the goal of our project



```
to_drop= ["state","district","rural","stratum_code","ever_conceived","age_at_first_conception","is_husband_living_with_you",  "counselled_for_menstrual_hyg"
        "house_no","house_hold_no","year_of_intr","month_of_intr","date_of_intr","result_of_interview","other_int_code","serial_no","identifcation_code",
        "surviving_female" ,"surviving_male",  "last_preg_no", "previous_last_preg_no" ,"second_last_preg_no", "third_last_preg_no", "twsi_id" ,"client_t
        "is_pills_daily","w_expall_status","w_status", "is_pills_daily", "is_piils_weekly","is_emergency_contraceptive", "is_condom","is_moder_methods","
        "is_other_traditional_method","pregnant_month","is_anc_registered",  "willing_to_get_pregnant","is_currently_menstruating","when_you_bcome_mother
        "how_long_using_this_method",  "method_obtain_last_time","reason_for_not_using_fp_method",
        "is_method_used_in_last_5_yrs","method_type_used_in_last_5_yrs","reason_for_discontinuation","intend_to_use_fp_method_in_futur", "when_method_is_
        "aware_abt_ort_ors","aware_abt_ort_ors_zinc",  "new_born_alive_female",  "new_born_alive_male","new_born_alive_total", "new_surviving_female", "n
        "healthscheme_1",  "healthscheme_2","householdstatus"  ,"isheadchanged","headname",
        "time_for_next_child","anm_in_last_3_months","during_pregnancy","during_lactation","aware_abt_rti","aware_abt_hiv","aware_of_haf","twsi_expall_st
        "relation_to_head","member_identity",  "father_serial_no", "mother_serial_no", "date_of_birth","month_of_birth","year_of_birth", "religion","soci
        "highest_qualification","land_possessed","hl_expall_status","sn",  "no_of_times_conceived",  "current_mar_status","is_injectable_contraceptive","
        "aware_abt_haf","ever_born","wt","fidx", "as","as_binned","fidh","cdoi","edt","catage1","marital","anym" ,"respondentname",  "rtelephoneno", "isn
        "hh_expall_status","house_structure","owner_status","w_id","hl_id","client_w_id","psu_id","building_no", "house_status", "drinking_water_source",
        "no_of_dwelling_rooms",  "kitchen_availability", "is_radio","is_television", "is_computer","is_telephone", "is_washing_machine", "is_refrigerator
         "is_water_pump",  "cart","client_hl_id","status"]

df.drop(to_drop, axis=1,inplace=True)
#here droping the whole variable to_drop by this method
```

There were missing data which were excluded as there were less data present

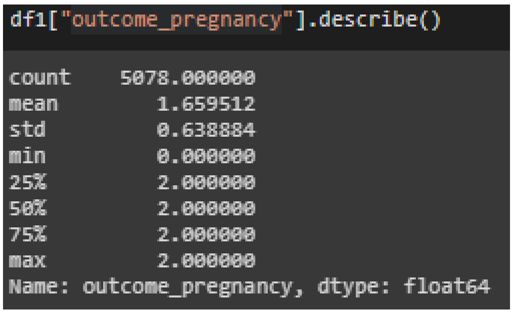
Here we see the values are all int64



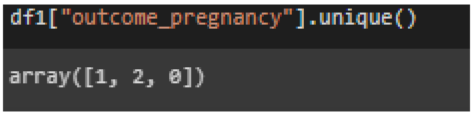
```
df1.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5078 entries, 0 to 5077
Data columns (total 15 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   age                           5078 non-null   int64
 1   surviving_total               5078 non-null   int64
 2   mother_age_when_baby_was_born 5078 non-null   int64
 3   outcome_pregnancy             5078 non-null   int64
 4   is_currently_pregnant         5078 non-null   int64
 5   aware_of_the_danger_signs     5078 non-null   int64
 6   disability_status             5078 non-null   int64
 7   symptoms_pertaining_illness   5078 non-null   int64
 8   sought_medical_care           5078 non-null   int64
 9   diagnosed_for                 5078 non-null   int64
 10  regular_treatment             5078 non-null   int64
 11  regular_treatment_source      5078 non-null   int64
 12  chew                          5078 non-null   int64
 13  smoke                         5078 non-null   int64
 14  alcohol                       5078 non-null   int64
dtypes: int64(15)
memory usage: 595.2 KB
```

II. Analysis of ‘Target’ Variable:

Here it is the description of outcome of pregnancy to identify the right algorithm to apply



```
df1["outcome_pregnancy"].describe()

count    5078.000000
mean        1.659512
std         0.638884
min         0.000000
25%         2.000000
50%         2.000000
75%         2.000000
max         2.000000
Name: outcome_pregnancy, dtype: float64
```

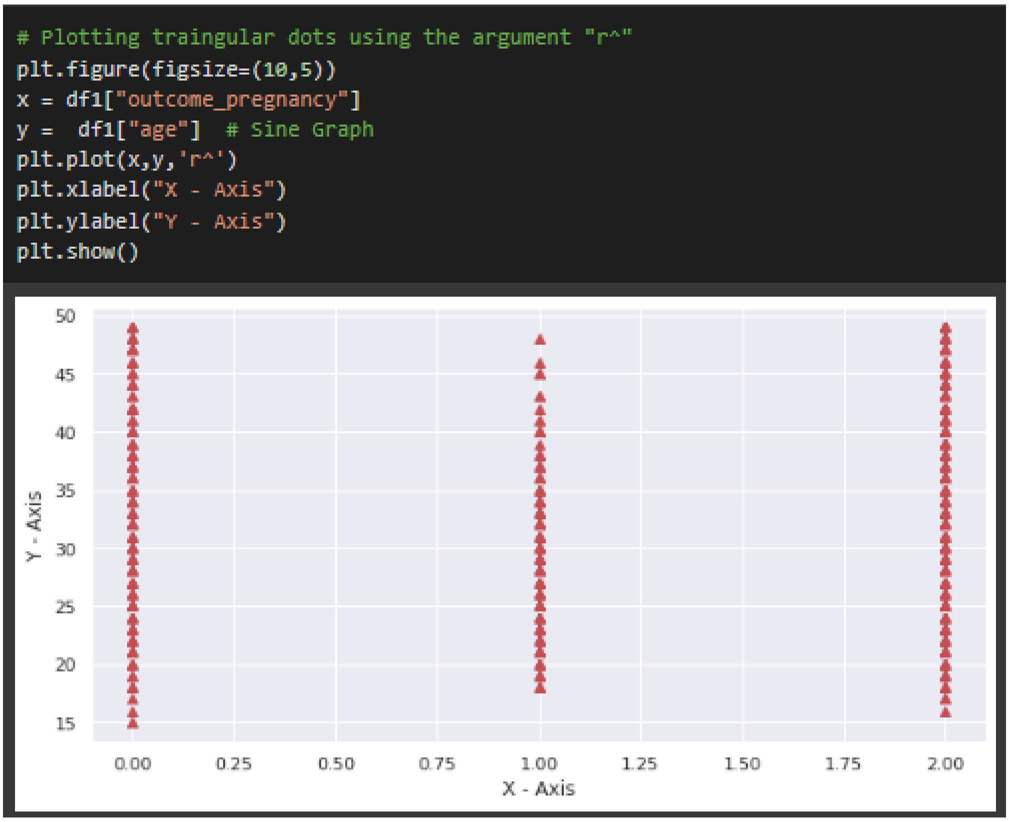


```
df1["outcome_pregnancy"].unique()

array([1, 2, 0])
```

We can understand from here that the values are nominal and 0,1,2 here means of the pregnancy is miscarriage which is 0 and 1 and 2 kids simultaneously

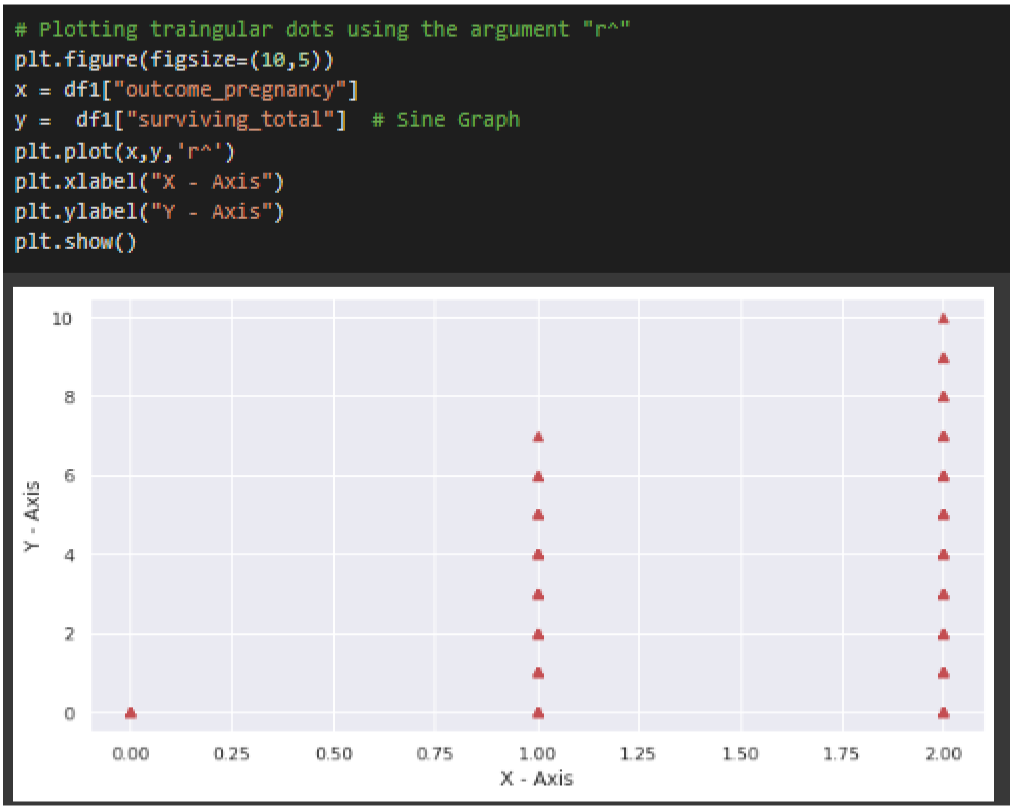
III. Exploratory Data Analysis

Here checked the age vs outcome of pregnancy

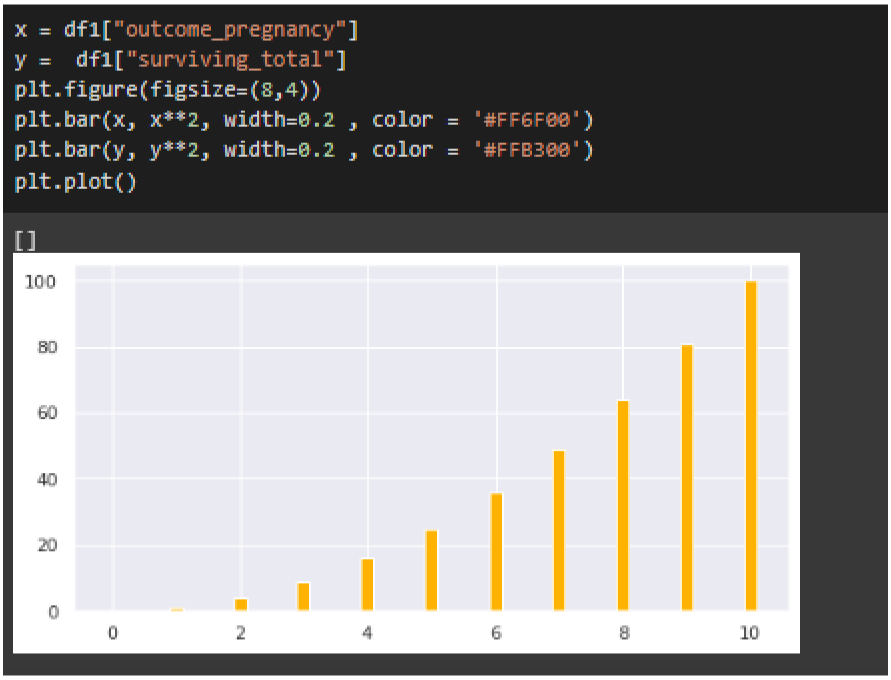


```
# Plotting traingular dots using the argument "r^"
plt.figure(figsize=(10,5))
x = df1["outcome_pregnancy"]
y =  df1["age"]  # Sine Graph
plt.plot(x,y,'r^')
plt.xlabel("X - Axis")
plt.ylabel("Y - Axis")
plt.show()
```

Survival of the fetus is very low it shows from the graph as surviving total is more at 1 and 2 then none or near to nill at 0



```
# Plotting traingular dots using the argument "r^"
plt.figure(figsize=(10,5))
x = df1["outcome_pregnancy"]
y =  df1["surviving_total"]  # Sine Graph
plt.plot(x,y,'r^')
plt.xlabel("X - Axis")
plt.ylabel("Y - Axis")
plt.show()
```

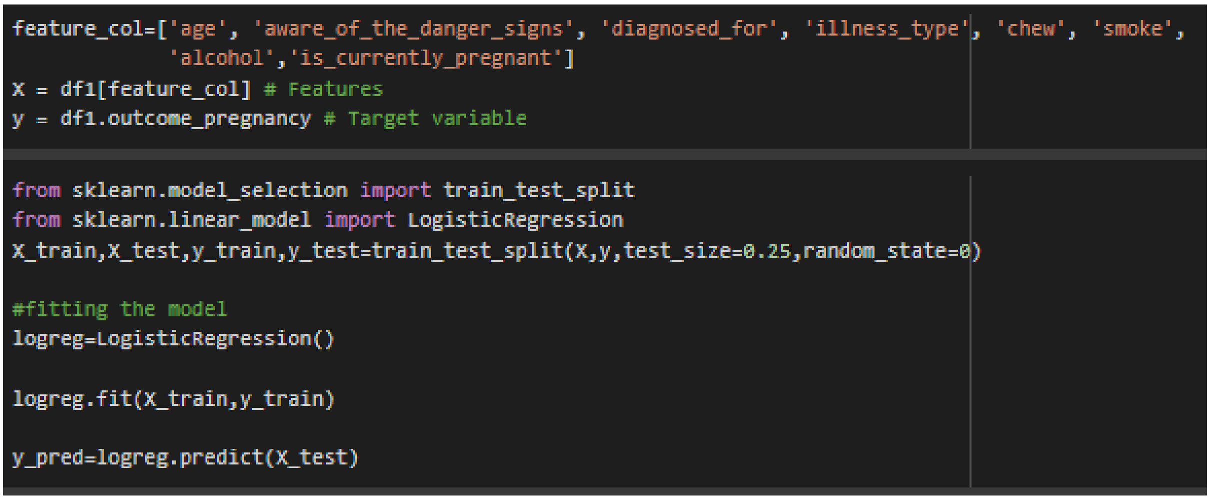


```
x = df1["outcome_pregnancy"]
y =  df1["surviving_total"]
plt.figure(figsize=(8,4))
plt.bar(x, x**2, width=0.2 , color = '#FF6F00')
plt.bar(y, y**2, width=0.2 , color = '#FFB300')
plt.plot()
```

```
[]
```

IV. Model Fitting

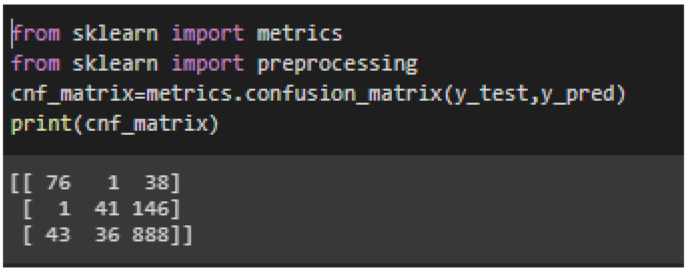
Here the data is classification type hence we use the value such as logistic regression



```
feature_col=['age', 'aware_of_the_danger_signs', 'diagnosed_for', 'illness_type', 'chew', 'smoke',
             'alcohol','is_currently_pregnant']
X = df1[feature_col] # Features
y = df1.outcome_pregnancy # Target variable
```

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.25,random_state=0)

#fitting the model
logreg=LogisticRegression()

logreg.fit(X_train,y_train)

y_pred=logreg.predict(X_test)
```

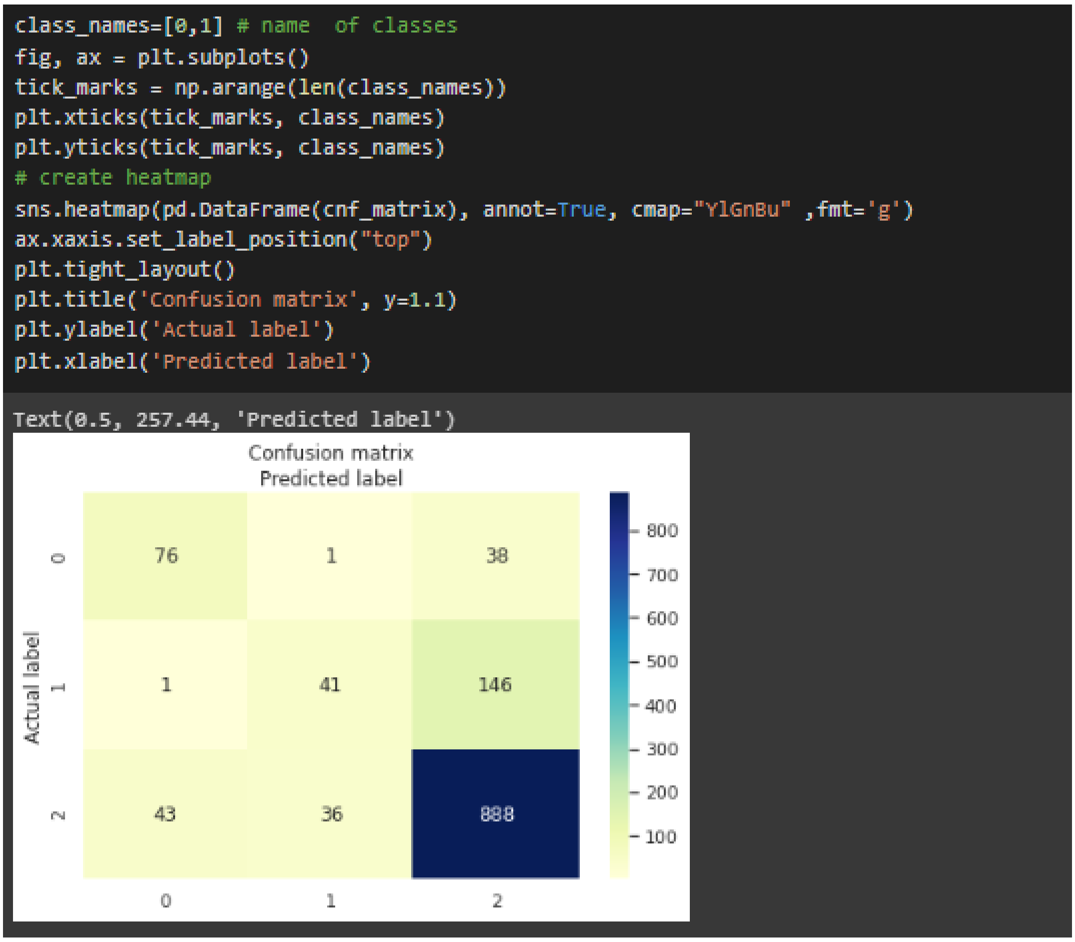
Here we see that the data in the diagnol hence we can say that the prediction is going in the right path as the data is grater in the diagnol



```
from sklearn import metrics
from sklearn import preprocessing
cnf_matrix=metrics.confusion_matrix(y_test,y_pred)
print(cnf_matrix)
```

```
[[ 76   1  38]
 [  1  41 146]
 [ 43  36 888]]
```

Here we plotted the lines and then we can see easily that outcome value of pregnancy is due to three nominal categories such as 0,1,2 where it indicates survival of the fetus and not



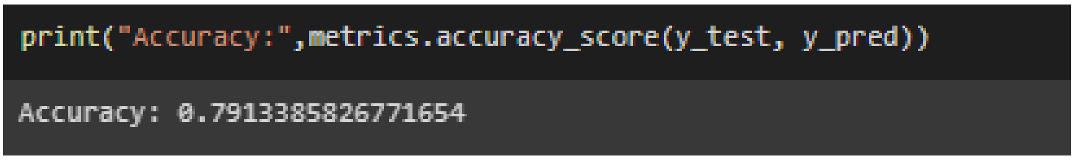
```
class_names=[0,1] # name  of classes
fig, ax = plt.subplots()
tick_marks = np.arange(len(class_names))
plt.xticks(tick_marks, class_names)
plt.yticks(tick_marks, class_names)
# create heatmap
sns.heatmap(pd.DataFrame(cnf_matrix), annot=True, cmap="YlGnBu" ,fmt='g')
ax.xaxis.set_label_position("top")
plt.tight_layout()
plt.title('Confusion matrix', y=1.1)
plt.ylabel('Actual label')
plt.xlabel('Predicted label')
```

```
Text(0.5, 257.44, 'Predicted label')
```

Confusion matrix

	0	1	2
0	76	1	38
1	1	41	146
2	43	36	888

Here we can see the Accuracy where it predicts 79%

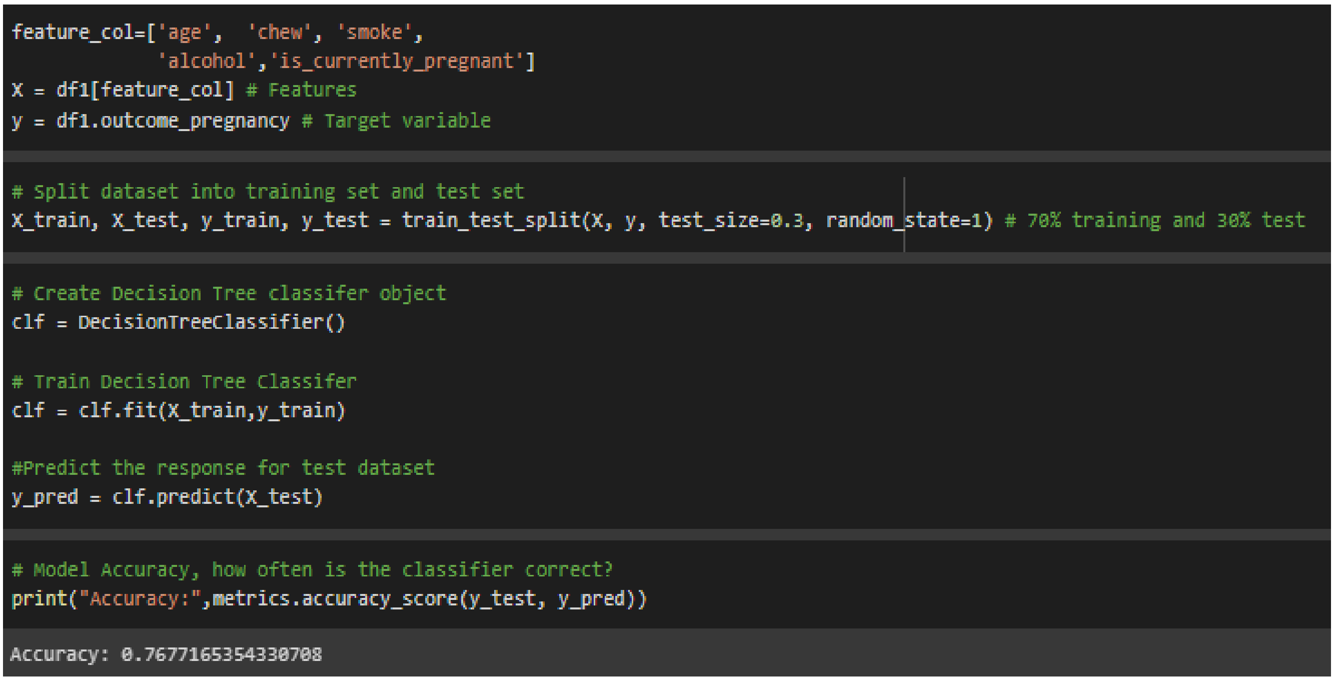


```
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7913385826771654
```

Decision tree classifier algorithm

Here tried the decision tree algorithm as we predicted where we have stated an object as feature_col where the factors taken as age,chew, smoke, alcohol, is_currently_pregnant have been taken to see how the algorithm behaves and our output label given is also outcome_pregnancy



```
feature_col=['age',  'chew', 'smoke',
             'alcohol','is_currently_pregnant']
X = df1[feature_col] # Features
y = df1.outcome_pregnancy # Target variable
```

```
# Split dataset into training set and test set
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1) # 70% training and 30% test
```

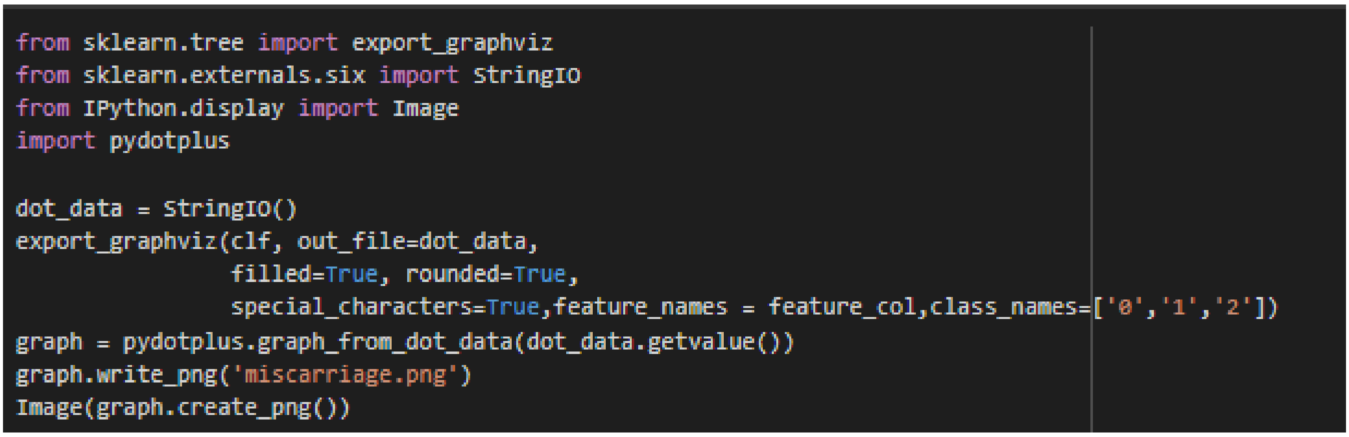
```
# Create Decision Tree classifer object
clf = DecisionTreeClassifier()

# Train Decision Tree Classifer
clf = clf.fit(X_train,y_train)

#Predict the response for test dataset
y_pred = clf.predict(X_test)
```

```
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

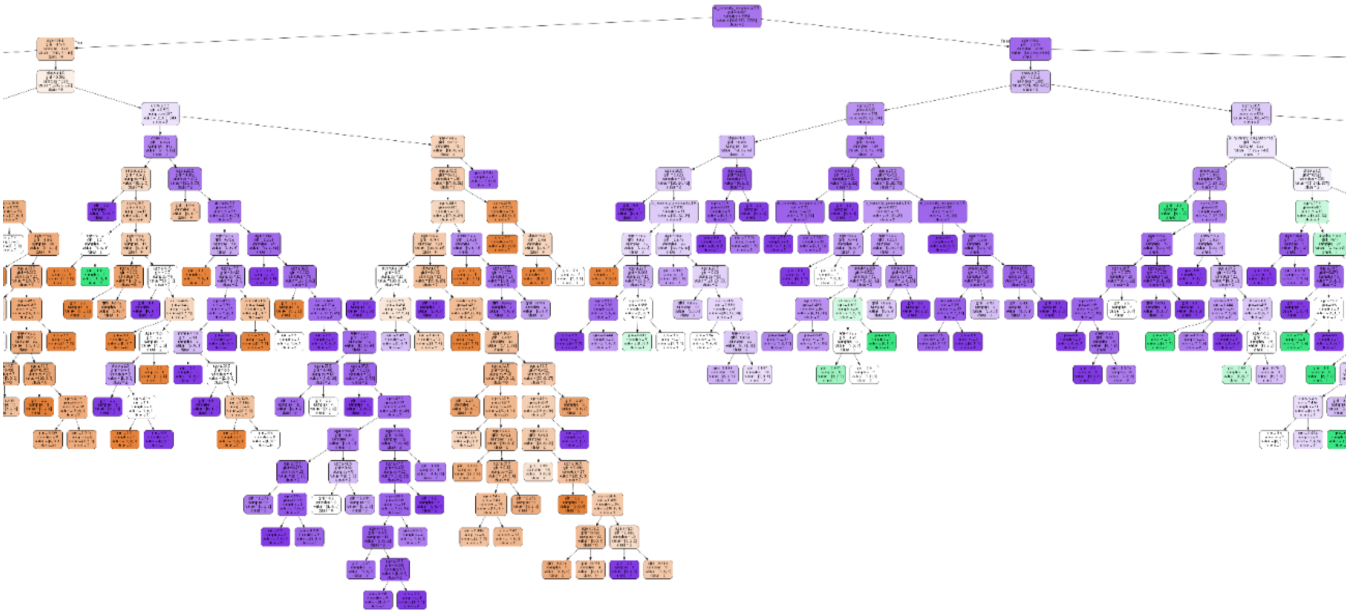
```
Accuracy: 0.7677165354330708
```

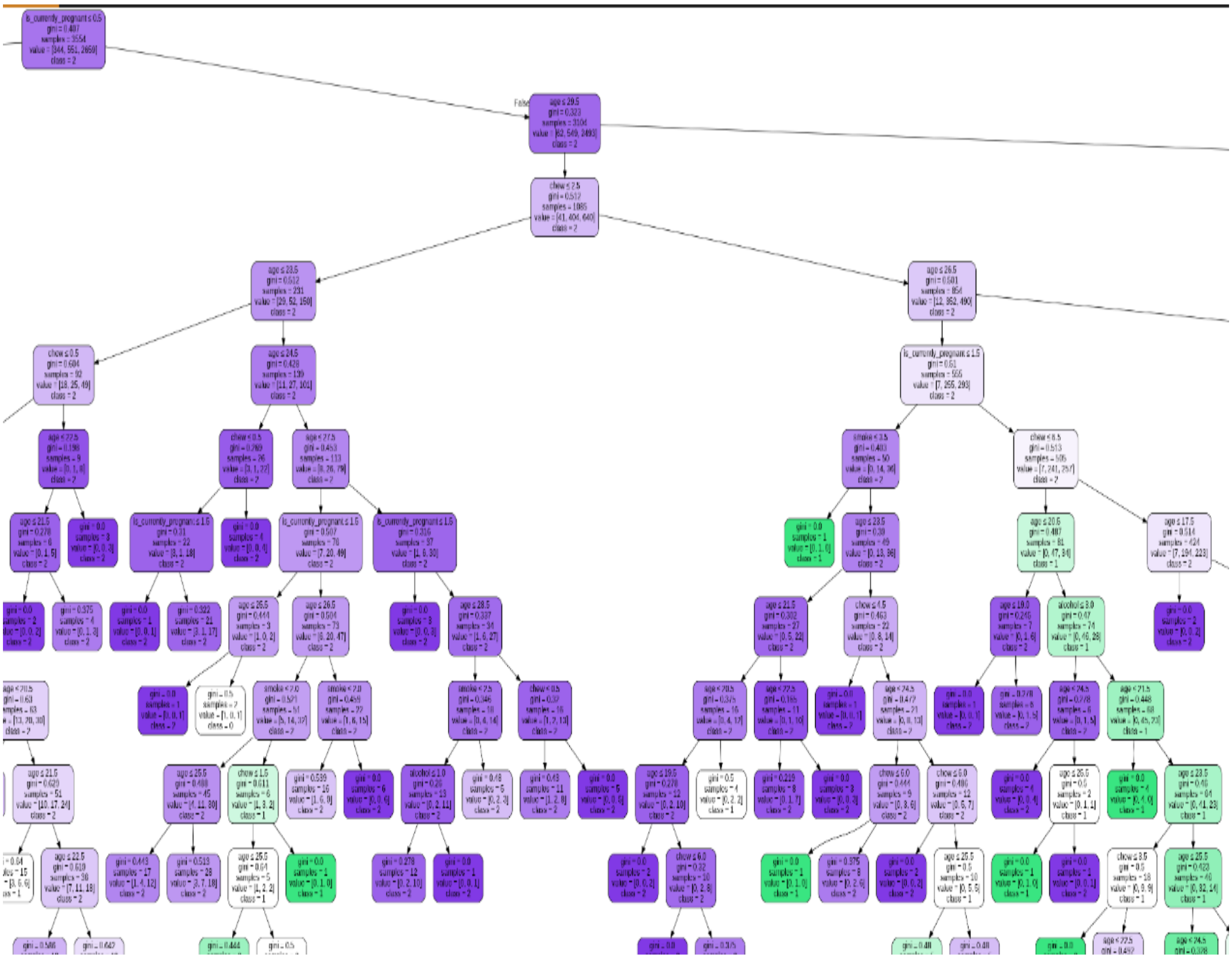
Here we can see the Accuracy which after the factors passing through the algo is 76% which is not bad after all we can always tune it more precisely due to data as the data also not overfit and underfit

Now below we can see that the code is for forming the decision tree and as our data is large upto 5000 values hence the decision tree has turned out to be huge hence it will be given in the file for reference, here in the decision tree the gini index is also present.



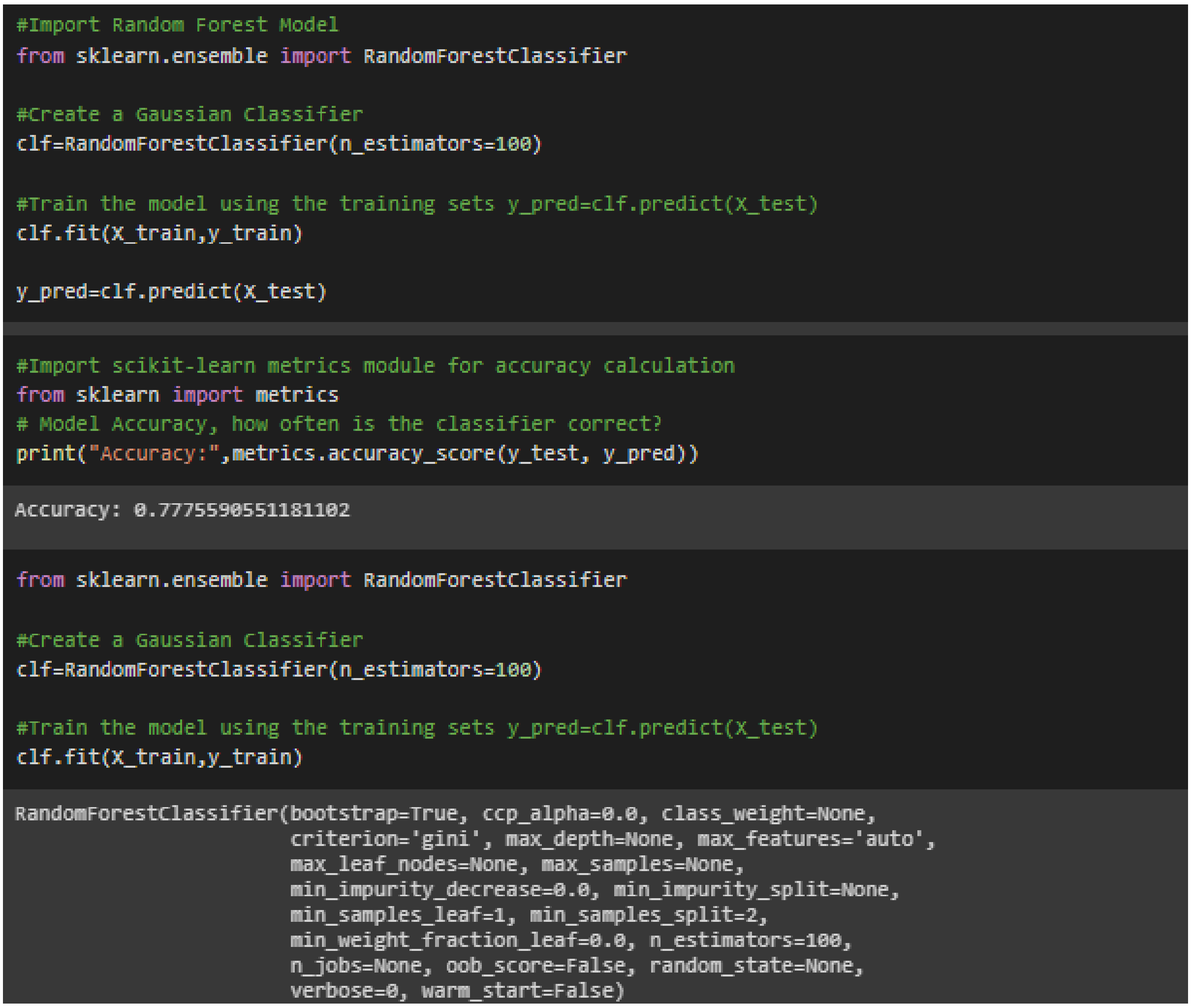
```
from sklearn.tree import export_graphviz
from sklearn.externals.six import StringIO
from IPython.display import Image
import pydotplus

dot_data = StringIO()
export_graphviz(clf, out_file=dot_data,
                filled=True, rounded=True,
                special_characters=True,feature_names = feature_col,class_names=['0','1','2'])
graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
graph.write_png('miscarriage.png')
Image(graph.create_png())
```





is_currently_pregnant ≤ 0.5
gini = 0.407
samples = 3554
value = [344, 551, 2659]
class = 2
False
age ≤ 29.5
gini = 0.323
samples = 3104
value = [62, 549, 2493]
class = 2
chew ≤ 2.5
gini = 0.512
samples = 1085
value = [41, 404, 640]
class = 2
age ≤ 23.5
gini = 0.512
samples = 231
value = [29, 52, 150]
class = 2
age ≤ 26.5
gini = 0.501
samples = 854
value = [12, 352, 490]
class = 2
chew ≤ 0.5
gini = 0.604
samples = 92
value = [18, 25, 49]
class = 2
age ≤ 24.5
gini = 0.428
samples = 139
value = [11, 27, 101]
class = 2
is_currently_pregnant ≤ 1.5
gini = 0.51
samples = 555
value = [7, 255, 293]
class = 2
age ≤ 22.5
gini = 0.198
samples = 9
value = [0, 1, 8]
class = 2
chew ≤ 0.5
gini = 0.269
samples = 26
value = [3, 1, 22]
class = 2
age ≤ 27.5
gini = 0.453
samples = 113
value = [8, 26, 79]
class = 2
smoke ≤ 3.5
gini = 0.403
samples = 50
value = [0, 14, 36]
class = 2
chew ≤ 8.5
gini = 0.513
samples = 505
value = [7, 241, 257]
class = 2

Random Forest



```
#Import Random Forest Model
from sklearn.ensemble import RandomForestClassifier

#Create a Gaussian Classifier
clf=RandomForestClassifier(n_estimators=100)

#Train the model using the training sets y_pred=clf.predict(X_test)
clf.fit(X_train,y_train)

y_pred=clf.predict(X_test)
```

```
#Import scikit-learn metrics module for accuracy calculation
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print("Accuracy:",metrics.accuracy_score(y_test, y_pred))
```

```
Accuracy: 0.7775590551181102
```

```
from sklearn.ensemble import RandomForestClassifier

#Create a Gaussian Classifier
clf=RandomForestClassifier(n_estimators=100)

#Train the model using the training sets y_pred=clf.predict(X_test)
clf.fit(X_train,y_train)
```

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                       criterion='gini', max_depth=None, max_features='auto',
                       max_leaf_nodes=None, max_samples=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=100,
                       n_jobs=None, oob_score=False, random_state=None,
                       verbose=0, warm_start=False)
```

Here we understand now again experimenting as how the model will fit hence we take this as the input is more easily to apply and the input used is same and not much confused to understand after applying random classifier and check the accuracy is 77% and we use the gaussian classifier.

#Accuracy

Logistic Regression	79%
Decision tree algorithm	76%
Random Forest	77%

Here we can see the frequency of the three algorithms altogether and can be seen that logistic regression has the highest frequency

Future work

After identifying the work where prediction can be done due to inputs from various pregnancy datasets and specifically spontaneous abortion related here the work could be done for making a web crawler and downloading the data and just like cancer detection from images, pregnancy sonogram could be read and the misdiagnosis done could be prevented and could be beneficial for the mases as the human intervention would be limited and easily a fault could be corrected

Conclusion

Now after implementing the algorithm we understand that the highest accuracy attained was by logistic regression and that our data is viable for miscarriage prediction where the woman being pregnant and consumption of alcohol, chewing of tobacco affects pregnancy of the woman

Rural women present they are not immuned to the effects of miscarriage as the statistics goes 25% of women are prone to have miscarriage and also that after 1st miscarriage the woman is able to have a normal healthy baby the second time and also there are reason for not having a baby are due to various factors and could be checked by the factors and the algorithm supports our theory that the factors do affect the pregnancy and results to miscarriage where a woman having factors such as stress and unhealthy habits hence could lead to spontaneous abortion

Bibliography

1. https://www.datacamp.com/community/tutorials/random-forests-classifier-python
2. https://towardsdatascience.com/improving-random-forest-in-python-part-1-893916666cd
3. https://data-flair.training/blogs/machine-learning-classification-algorithms/
4. https://www.kaggle.com/rajanand/ahs-woman-1
5. https://www.datacamp.com/community/tutorials/machine-learning-python
6. https://www.hindustantimes.com/mumbai-news/70-miscarriages-in-maharashtra-are-from-rural-areas-says-health-dept/story-kNEw2rgRV6ASUnFz2yFtGJ.html
7. https://www.uptodate.com/contents/spontaneous-abortion-management
8. https://data.gov.in/resources/percentage-pregnancy-resulted-live-birth-still-birth-induced-abortion-and-spontaneous
9. https://www.medicalnewstoday.com/articles/322634#pregnancy-loss-rates-by-week
10. https://www.mayoclinic.org/diseases-conditions/pregnancy-loss-miscarriage/symptoms-causes/syc-20354298
11. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5992995/
12. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5741961/
13. https://www.nhp.gov.in/disease/gynaecology-and-obstetrics/early-pregnancy-loss
14. https://indianexpress.com/article/cities/mumbai/in-maharashtra-60495-miscarriage-cases-reported-from-april-17-to-march-18-5204256/