```
# Importing Libraries
import numpy as np # Linear algebra -> To perform mathematical operations
import pandas as pd # Data pre-processing, CSV file I/O (e.g. pd.read_csv)
import seaborn as sns # Statistical Graphs
import matplotlib.pyplot as plt # Visualizations
import matplotlib.image as mpimg

import warnings # Warnings do not cause a program to terminate, Errors do
warnings.filterwarnings("ignore")


# Importing Dataset
df_train=pd.read_csv("/content/train.csv") # Train Dataset
df_test=pd.read_csv("/content/test.csv") # Test Dataset
df_train
```

	Patient Id	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Patient First Name	Family Name	Father's name	...	Birth defects	White Blood cell count (thousand per microliter)	
0	PID0x6418	2.0	Yes	No	Yes	No	4.760603	Richard	NaN	Larre	...	NaN	9.857562	
1	PID0x25d5	4.0	Yes	Yes	No	No	4.910669	Mike	NaN	Brycen	...	Multiple	5.522560	
2	PID0x4a82	6.0	Yes	No	No	No	4.893297	Kimberly	NaN	Nashon	...	Singular	NaN	
3	PID0x4ac8	12.0	Yes	No	Yes	No	4.705280	Jeffery	Hoelscher	Aayaan	...	Singular	7.919321	inco
4	PID0x1bf7	11.0	Yes	No	NaN	Yes	4.720703	Johanna	Stutzman	Suave	...	Multiple	4.098210	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
22078	PID0x5598	4.0	Yes	Yes	Yes	No	5.258298	Lynn	NaN	Alhassane	...	Multiple	6.584811	inco
22079	PID0x19cb	8.0	No	Yes	No	Yes	4.974220	Matthew	Farley	Dartanion	...	Multiple	7.041556	inco
22080	PID0x3c4f	8.0	Yes	No	Yes	No	5.186470	John	NaN	Cavani	...	Singular	7.715464	
22081	PID0x13a	7.0	Yes	No	Yes	Yes	4.858543	Sharon	NaN	Bomer	...	Multiple	8.437670	a
22082	PID0x9332	11.0	Yes	No	No	No	4.738067	Andrew	Mose	Eban	...	Singular	11.188371	

22083 rows × 45 columns

```
df_train.tail()
```

	Patient Id	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Patient First Name	Family Name	Father's name	...	Birth defects	White Blood cell count (thousand per microliter)	r
22078	PID0x5598	4.0	Yes	Yes	Yes	No	5.258298	Lynn	NaN	Alhassane	...	Multiple	6.584811	inconc
22079	PID0x19cb	8.0	No	Yes	No	Yes	4.974220	Matthew	Farley	Dartanion	...	Multiple	7.041556	inconc
22080	PID0x3c4f	8.0	Yes	No	Yes	No	5.186470	John	NaN	Cavani	...	Singular	7.715464	r
22081	PID0x13a	7.0	Yes	No	Yes	Yes	4.858543	Sharon	NaN	Bomer	...	Multiple	8.437670	abr

```
# Information about the Training Dataset
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22083 entries, 0 to 22082
Data columns (total 45 columns):
 #   Column                                          Non-Null Count  Dtype  
---  ------                                          --------------  -----  
 0   Patient Id                                      22083 non-null  object 
 1   Patient Age                                     20656 non-null  float64
 2   Genes in mother's side                          22083 non-null  object 
 3   Inherited from father                           21777 non-null  object 
 4   Maternal gene                                   19273 non-null  object 
 5   Paternal gene                                   22083 non-null  object 
 6   Blood cell count (mcL)                          22083 non-null  float64
 7   Patient First Name                              22083 non-null  object 
 8   Family Name                                     12392 non-null  object 
 9   Father's name                                   22083 non-null  object 
 10  Mother's age                                    16047 non-null  float64
 11  Father's age                                    16097 non-null  float64
 12  Institute Name                                  16977 non-null  object 
 13  Location of Institute                           22083 non-null  object 
 14  Status                                          22083 non-null  object 
 15  Respiratory Rate (breaths/min)                  19934 non-null  object 
 16  Heart Rate (rates/min                           19970 non-null  object 
 17  Test 1                                          19956 non-null  float64
 18  Test 2                                          19931 non-null  float64
 19  Test 3                                          19936 non-null  float64
 20  Test 4                                          19943 non-null  float64
 21  Test 5                                          19913 non-null  float64
 22  Parental consent                                19958 non-null  object 
 23  Follow-up                                       19917 non-null  object 
 24  Gender                                          19910 non-null  object 
 25  Birth asphyxia                                  19944 non-null  object 
 26  Autopsy shows birth defect (if applicable)      21057 non-null  object 
 27  Place of birth                                  19959 non-null  object 
 28  Folic acid details (peri-conceptional)          19966 non-null  object 
 29  H/O serious maternal illness                    19931 non-null  object 
 30  H/O radiation exposure (x-ray)                  19930 non-null  object 
 31  H/O substance abuse                             19888 non-null  object 
 32  Assisted conception IVF/ART                     19961 non-null  object 
 33  History of anomalies in previous pregnancies    19911 non-null  object 
 34  No. of previous abortion                        19921 non-null  float64
 35  Birth defects                                   19929 non-null  object 
 36  White Blood cell count (thousand per microliter)  19935 non-null  float64
 37  Blood test result                               19938 non-null  object 
 38  Symptom 1                                       19928 non-null  float64
 39  Symptom 2                                       19861 non-null  float64
 40  Symptom 3                                       19982 non-null  float64
 41  Symptom 4                                       19970 non-null  float64
 42  Symptom 5                                       19930 non-null  float64
 43  Genetic Disorder                                19937 non-null  object 
 44  Disorder Subclass                               19915 non-null  object 
dtypes: float64(16), object(29)
memory usage: 7.6+ MB
```

```
# Information about the Testing Dataset
df_test.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9465 entries, 0 to 9464
Data columns (total 43 columns):
 #   Column                                          Non-Null Count  Dtype  
---  ------                                          --------------  -----  
 0   Patient Id                                      9465 non-null   object 
 1   Patient Age                                     9465 non-null   int64  
 2   Genes in mother's side                          9465 non-null   object 
 3   Inherited from father                           8914 non-null   object 
 4   Maternal gene                                   5742 non-null   object 
 5   Paternal gene                                   9465 non-null   object 
 6   Blood cell count (mcL)                          9465 non-null   float64
 7   Patient First Name                              9465 non-null   object 
 8   Family Name                                     148 non-null    object 
 9   Father's name                                   9465 non-null   object 
 10  Mother's age                                    9465 non-null   int64  
 11  Father's age                                    9465 non-null   int64  
 12  Institute Name                                  7429 non-null   object 
 13  Location of Institute                           9465 non-null   object 
 14  Status                                          9465 non-null   object 
 15  Respiratory Rate (breaths/min)                  6579 non-null   object 
 16  Heart Rate (rates/min                           6565 non-null   object 
 17  Test 1                                          9465 non-null   int64  
 18  Test 2                                          9465 non-null   int64  
 19  Test 3                                          9465 non-null   int64  
 20  Test 4                                          9465 non-null   int64  
 21  Test 5                                          9465 non-null   int64  
 22  Parental consent                                9465 non-null   object 
 23  Follow-up                                       9465 non-null   object 
 24  Gender                                          9465 non-null   object 
 25  Birth asphyxia                                  9465 non-null   object 
 26  Autopsy shows birth defect (if applicable)      9465 non-null   object 
 27  Place of birth                                  9465 non-null   object 
 28  Folic acid details (peri-conceptional)          9465 non-null   object 
 29  H/O serious maternal illness                    9465 non-null   object 
 30  H/O radiation exposure (x-ray)                  9465 non-null   object 
 31  H/O substance abuse                             9465 non-null   object 
 32  Assisted conception IVF/ART                     9465 non-null   object 
 33  History of anomalies in previous pregnancies    9465 non-null   object 
 34  No. of previous abortion                        9465 non-null   int64  
 35  Birth defects                                   9465 non-null   object 
 36  White Blood cell count (thousand per microliter)  9465 non-null   float64
 37  Blood test result                               9465 non-null   object 
 38  Symptom 1                                       9465 non-null   bool   
 39  Symptom 2                                       9465 non-null   bool   
 40  Symptom 3                                       9465 non-null   bool   
 41  Symptom 4                                       9465 non-null   bool   
 42  Symptom 5                                       9465 non-null   bool   
dtypes: bool(5), float64(2), int64(9), object(27)
memory usage: 2.8+ MB
```

```
df_train["Genetic Disorder"].unique()
```

```
array(['Mitochondrial genetic inheritance disorders', nan,
       'Multifactorial genetic inheritance disorders',
       'Single-gene inheritance diseases'], dtype=object)
```

```
df_train["Disorder Subclass"].unique()
```

```
array(["Leber's hereditary optic neuropathy", 'Cystic fibrosis',
       'Diabetes', 'Leigh syndrome', 'Cancer', 'Tay-Sachs',
       'Hemochromatosis', 'Mitochondrial myopathy', nan, "Alzheimer's"],
      dtype=object)
```

```
# Total Columns
df_train.columns
```

```
Index(['Patient Id', 'Patient Age', 'Genes in mother's side',
       'Inherited from father', 'Maternal gene', 'Paternal gene',
       'Blood cell count (mcL)', 'Patient First Name', 'Family Name',
       'Father's name', 'Mother's age', 'Father's age', 'Institute Name',
       'Location of Institute', 'Status', 'Respiratory Rate (breaths/min)',
       'Heart Rate (rates/min', 'Test 1', 'Test 2', 'Test 3', 'Test 4',
       'Test 5', 'Parental consent', 'Follow-up', 'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)', 'Place of birth',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
```

```
       'History of anomalies in previous pregnancies',
       'No. of previous abortion', 'Birth defects',
       'White Blood cell count (thousand per microliter)', 'Blood test result',
       'Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5',
       'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

```
# Pre-Processing Starts
# Dropping unwanted columns
df_train.drop("Patient Id",axis=1,inplace=True)
df_train.drop("Family Name",axis=1,inplace=True)
df_train.drop("Patient First Name",axis=1,inplace=True)
df_train.drop("Father's name",axis=1,inplace=True)
df_train.drop("Institute Name",axis=1,inplace=True)
df_train.drop("Location of Institute",axis=1,inplace=True)
df_train.drop("Place of birth",axis=1,inplace=True)
```

```
# Checking total null values
df_train.isna().sum()
```

```
Patient Age                                        1427
Genes in mother's side                                0
Inherited from father                               306
Maternal gene                                      2810
Paternal gene                                         0
Blood cell count (mcL)                                0
Mother's age                                       6036
Father's age                                       5986
Status                                                0
Respiratory Rate (breaths/min)                     2149
Heart Rate (rates/min                              2113
Test 1                                             2127
Test 2                                             2152
Test 3                                             2147
Test 4                                             2140
Test 5                                             2170
Parental consent                                   2125
Follow-up                                          2166
Gender                                             2173
Birth asphyxia                                     2139
Autopsy shows birth defect (if applicable)         1026
Folic acid details (peri-conceptional)             2117
H/O serious maternal illness                       2152
H/O radiation exposure (x-ray)                     2153
H/O substance abuse                                2195
Assisted conception IVF/ART                        2122
History of anomalies in previous pregnancies       2172
No. of previous abortion                           2162
Birth defects                                      2154
White Blood cell count (thousand per microliter)   2148
Blood test result                                  2145
Symptom 1                                          2155
Symptom 2                                          2222
Symptom 3                                          2101
Symptom 4                                          2113
Symptom 5                                          2153
Genetic Disorder                                   2146
Disorder Subclass                                  2168
dtype: int64
```

```
df_train["Patient Age"]
```

```
0         2.0
1         4.0
2         6.0
3        12.0
4        11.0
         ...
22078     4.0
22079     8.0
22080     8.0
22081     7.0
22082    11.0
Name: Patient Age, Length: 22083, dtype: float64
```

```
# Filling Null values with mode
df_train["Patient Age"].fillna(str(df_train["Patient Age"].mode().values[0]),inplace=True)
df_train["Inherited from father"].fillna(str(df_train["Inherited from father"].mode().values[0]),inplace=True)
df_train["Maternal gene"].fillna(str(df_train["Maternal gene"].mode().values[0]),inplace=True)
df_train["Mother's age"].fillna(str(df_train["Mother's age"].mode().values[0]),inplace=True)
```

```
df_train["Father's age"].fillna(str(df_train["Father's age"].mode().values[0]),inplace=True)
df_train["Respiratory Rate (breaths/min)"].fillna(str(df_train["Respiratory Rate (breaths/min)"].mode().values[0]),inplace=True)
df_train["Heart Rate (rates/min"].fillna(str(df_train["Heart Rate (rates/min"].mode().values[0]),inplace=True)
df_train["Test 1"].fillna(str(df_train["Test 1"].mode().values[0]),inplace=True)
df_train["Test 2"].fillna(str(df_train["Test 2"].mode().values[0]),inplace=True)
df_train["Test 3"].fillna(str(df_train["Test 3"].mode().values[0]),inplace=True)
df_train["Test 4"].fillna(str(df_train["Test 4"].mode().values[0]),inplace=True)
df_train["Test 5"].fillna(str(df_train["Test 5"].mode().values[0]),inplace=True)
df_train["Parental consent"].fillna(str(df_train["Parental consent"].mode().values[0]),inplace=True)
df_train["Follow-up"].fillna(str(df_train["Follow-up"].mode().values[0]),inplace=True)
df_train["Gender"].fillna(str(df_train["Gender"].mode().values[0]),inplace=True)
df_train["Birth asphyxia"].fillna(str(df_train["Birth asphyxia"].mode().values[0]),inplace=True)
df_train["Autopsy shows birth defect (if applicable)"].fillna(str(df_train["Autopsy shows birth defect (if applicable)"].mode().values[0]),in
df_train["Folic acid details (peri-conceptional)"].fillna(str(df_train["Folic acid details (peri-conceptional)"].mode().values[0]),inplace=Tr
df_train["H/O serious maternal illness"].fillna(str(df_train["H/O serious maternal illness"].mode().values[0]),inplace=True)
df_train["H/O radiation exposure (x-ray)"].fillna(str(df_train["H/O radiation exposure (x-ray)"].mode().values[0]),inplace=True)
df_train["H/O substance abuse"].fillna(str(df_train["H/O substance abuse"].mode().values[0]),inplace=True)
df_train["Assisted conception IVF/ART"].fillna(str(df_train["Assisted conception IVF/ART"].mode().values[0]),inplace=True)
df_train["History of anomalies in previous pregnancies"].fillna(str(df_train["History of anomalies in previous pregnancies"].mode().values[0]
df_train["No. of previous abortion"].fillna(str(df_train["No. of previous abortion"].mode().values[0]),inplace=True)
df_train["Birth defects"].fillna(str(df_train["Birth defects"].mode().values[0]),inplace=True)
df_train["White Blood cell count (thousand per microliter)"].fillna(str(df_train["White Blood cell count (thousand per microliter)"].mode().v
df_train["Blood test result"].fillna(str(df_train["Blood test result"].mode().values[0]),inplace=True)
df_train["Symptom 1"].fillna(str(df_train["Symptom 1"].mode().values[0]),inplace=True)
df_train["Symptom 2"].fillna(str(df_train["Symptom 2"].mode().values[0]),inplace=True)
df_train["Symptom 3"].fillna(str(df_train["Symptom 3"].mode().values[0]),inplace=True)
df_train["Symptom 4"].fillna(str(df_train["Symptom 4"].mode().values[0]),inplace=True)
df_train["Symptom 5"].fillna(str(df_train["Symptom 5"].mode().values[0]),inplace=True)
df_train["Genetic Disorder"].fillna(str(df_train["Genetic Disorder"].mode().values[0]),inplace=True)
df_train["Disorder Subclass"].fillna(str(df_train["Disorder Subclass"].mode().values[0]),inplace=True)


# Checking if any null value is present
df_train.isna().sum()
```

```
Patient Age                                     0
Genes in mother's side                          0
Inherited from father                           0
Maternal gene                                   0
Paternal gene                                   0
Blood cell count (mcL)                          0
Mother's age                                    0
Father's age                                    0
Status                                          0
Respiratory Rate (breaths/min)                  0
Heart Rate (rates/min                           0
Test 1                                          0
Test 2                                          0
Test 3                                          0
Test 4                                          0
Test 5                                          0
Parental consent                                0
Follow-up                                       0
Gender                                          0
Birth asphyxia                                  0
Autopsy shows birth defect (if applicable)      0
Folic acid details (peri-conceptional)          0
H/O serious maternal illness                    0
H/O radiation exposure (x-ray)                  0
H/O substance abuse                             0
Assisted conception IVF/ART                     0
History of anomalies in previous pregnancies    0
No. of previous abortion                        0
Birth defects                                   0
White Blood cell count (thousand per microliter)    0
Blood test result                               0
Symptom 1                                       0
Symptom 2                                       0
Symptom 3                                       0
Symptom 4                                       0
Symptom 5                                       0
Genetic Disorder                                0
Disorder Subclass                               0
dtype: int64
```

```
df_train.head()
```

	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Mother's age	Father's age	Status	Respiratory Rate (breaths/min)	...	Birth defects	White Blood cell count (thousand per microliter)	
0	2.0	Yes	No	Yes	No	4.760603	23.0	20.0	Alive	Normal (30-60)	...	Singular	9.857562	
1	4.0	Yes	Yes	No	No	4.910669	23.0	23.0	Deceased	Tachypnea	...	Multiple	5.52256	
2	6.0	Yes	No	No	No	4.893297	41.0	22.0	Alive	Normal (30-60)	...	Singular	3.0	
3	12.0	Yes	No	Yes	No	4.705280	21.0	20.0	Deceased	Tachypnea	...	Singular	7.919321	inc
4	11.0	Yes	No	Yes	Yes	4.720703	32.0	20.0	Alive	Tachypnea	...	Multiple	4.09821	

```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22083 entries, 0 to 22082
Data columns (total 38 columns):
 #   Column                                          Non-Null Count  Dtype  
---  ------                                          --------------  -----  
 0   Patient Age                                     22083 non-null  object 
 1   Genes in mother's side                          22083 non-null  object 
 2   Inherited from father                           22083 non-null  object 
 3   Maternal gene                                   22083 non-null  object 
 4   Paternal gene                                   22083 non-null  object 
 5   Blood cell count (mcL)                          22083 non-null  float64
 6   Mother's age                                    22083 non-null  object 
 7   Father's age                                    22083 non-null  object 
 8   Status                                          22083 non-null  object 
 9   Respiratory Rate (breaths/min)                  22083 non-null  object 
 10  Heart Rate (rates/min                           22083 non-null  object 
 11  Test 1                                          22083 non-null  object 
 12  Test 2                                          22083 non-null  object 
 13  Test 3                                          22083 non-null  object 
 14  Test 4                                          22083 non-null  object 
 15  Test 5                                          22083 non-null  object 
 16  Parental consent                                22083 non-null  object 
 17  Follow-up                                       22083 non-null  object 
 18  Gender                                          22083 non-null  object 
 19  Birth asphyxia                                  22083 non-null  object 
 20  Autopsy shows birth defect (if applicable)      22083 non-null  object 
 21  Folic acid details (peri-conceptional)          22083 non-null  object 
 22  H/O serious maternal illness                    22083 non-null  object 
 23  H/O radiation exposure (x-ray)                  22083 non-null  object 
 24  H/O substance abuse                             22083 non-null  object 
 25  Assisted conception IVF/ART                     22083 non-null  object 
 26  History of anomalies in previous pregnancies    22083 non-null  object 
 27  No. of previous abortion                        22083 non-null  object 
 28  Birth defects                                   22083 non-null  object 
 29  White Blood cell count (thousand per microliter)  22083 non-null  object 
 30  Blood test result                               22083 non-null  object 
 31  Symptom 1                                       22083 non-null  object 
 32  Symptom 2                                       22083 non-null  object 
 33  Symptom 3                                       22083 non-null  object 
 34  Symptom 4                                       22083 non-null  object 
 35  Symptom 5                                       22083 non-null  object 
 36  Genetic Disorder                                22083 non-null  object 
 37  Disorder Subclass                               22083 non-null  object 
dtypes: float64(1), object(37)
memory usage: 6.4+ MB
```

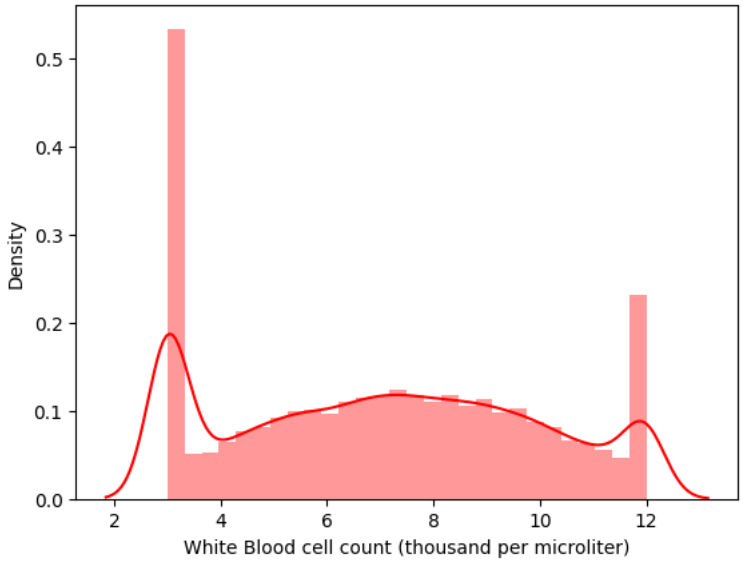
```
# Optional Column name change
# for column in df_train:
#     columnSeriesObj = df_train[column]
#     print('Column Name : ', column)
#     print('Column Contents : ', columnSeriesObj.values)
#     print("----------------------------------------------")
```

```
df_train.columns
```

```
Index(['Patient Age', 'Genes in mother's side', 'Inherited from father',
       'Maternal gene', 'Paternal gene', 'Blood cell count (mcL)',
       'Mother's age', 'Father's age', 'Status',
       'Respiratory Rate (breaths/min)', 'Heart Rate (rates/min', 'Test 1',
       'Test 2', 'Test 3', 'Test 4', 'Test 5', 'Parental consent', 'Follow-up',
       'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies',
       'No. of previous abortion', 'Birth defects',
       'White Blood cell count (thousand per microliter)', 'Blood test result',
       'Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5',
       'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

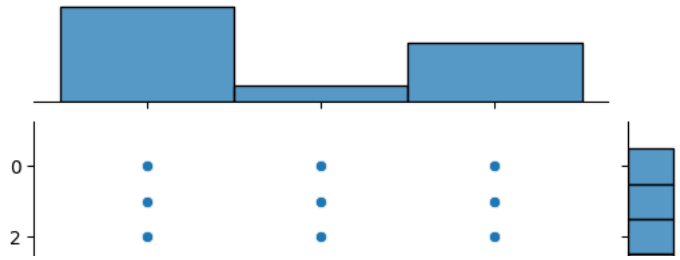
```
# Plotting
sns.distplot(df_train["White Blood cell count (thousand per microliter)"],color = "red")
```

```
<Axes: xlabel='White Blood cell count (thousand per microliter)', ylabel='Density'>
```



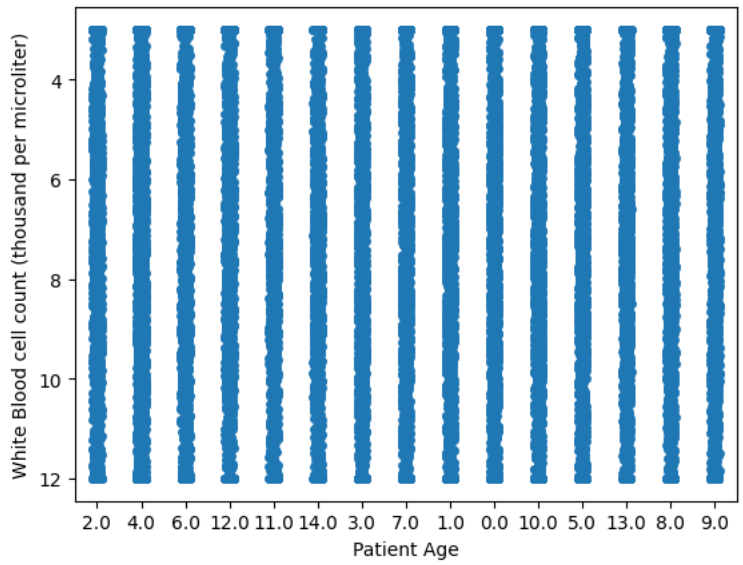

```
sns.jointplot(x="Genetic Disorder",y="Patient Age",data=df_train)
```

```
<seaborn.axisgrid.JointGrid at 0x7b433f735810>
```



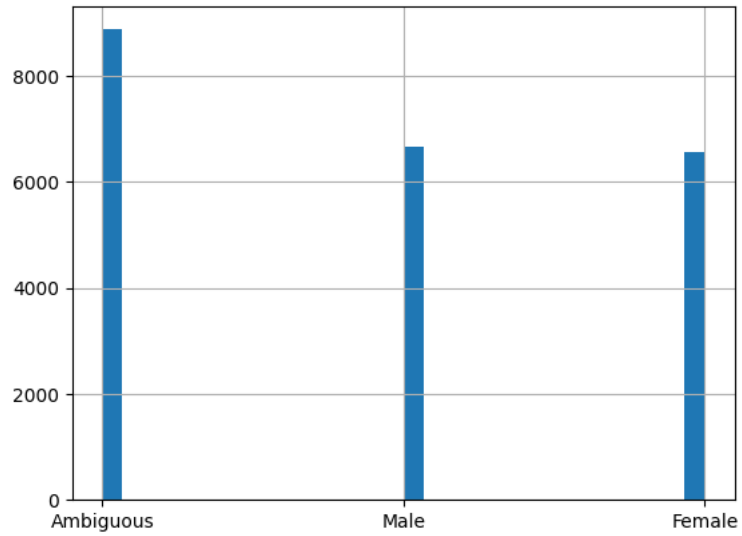

```
sns.stripplot(x="Patient Age",y="White Blood cell count (thousand per microliter)",data=df_train,jitter=True)
```

```
<Axes: xlabel='Patient Age', ylabel='White Blood cell count (thousand per microliter)'>
```




```
df_train["Gender"].hist(bins=30)
```

```
<Axes: >
```




```
df_train['Gender'].value_counts()
```

```
Ambiguous    8868
Male         6666
Female       6549
Name: Gender, dtype: int64
```

```
# Changing from yes or no[Categorical] to numerical(1 or 0)
df_train["Genes in mother's side"]=[1 if i.strip()== "Yes" else 0 for i in df_train["Genes in mother's side"]]
df_train["Inherited from father"]=[1 if i.strip()== "Yes" else 0 for i in df_train["Inherited from father"]]
df_train["Maternal gene"]=[1 if i.strip()== "Yes" else 0 for i in df_train["Maternal gene"]]
df_train["Paternal gene"]=[1 if i.strip()== "Yes" else 0 for i in df_train["Paternal gene"]]
df_train["Parental consent"]=[1 if i.strip()== "Yes" else 0 for i in df_train["Parental consent"]]
df_train["Birth asphyxia"]=[1 if i.strip()== "Yes" else 0 for i in df_train["Birth asphyxia"]]
df_train["Folic acid details (peri-conceptional)"]=[1 if i.strip()== "Yes" else 0 for i in df_train["Folic acid details (peri-conceptional)"
df_train["H/O radiation exposure (x-ray)"]=[1 if i.strip()== "Yes" else 0 for i in df_train["H/O radiation exposure (x-ray)"]]
df_train["H/O substance abuse"]=[1 if i.strip()== "Yes" else 0 for i in df_train["H/O substance abuse"]]
df_train["Assisted conception IVF/ART"]=[1 if i.strip()== "Yes" else 0 for i in df_train["Assisted conception IVF/ART"]]
df_train["History of anomalies in previous pregnancies"]=[1 if i.strip()== "Yes" else 0 for i in df_train["History of anomalies in previous
df_train["H/O serious maternal illness"]=[1 if i.strip()=="Yes" else 0 for i in df_train["H/O serious maternal illness"]]
```

```
df_train.head()
```

	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Mother's age	Father's age	Status	Respiratory Rate (breaths/min)	...	Birth defects	White Blood cell count (thousand per microliter)	
0	2.0	1	0	1	0	4.760603	23.0	20.0	Alive	Normal (30-60)	...	Singular	9.857562	
1	4.0	1	1	0	0	4.910669	23.0	23.0	Deceased	Tachypnea	...	Multiple	5.52256	
2	6.0	1	0	0	0	4.893297	41.0	22.0	Alive	Normal (30-60)	...	Singular	3.0	
3	12.0	1	0	1	0	4.705280	21.0	20.0	Deceased	Tachypnea	...	Singular	7.919321	inc
4	11.0	1	0	1	1	4.720703	32.0	20.0	Alive	Tachypnea	...	Multiple	4.09821	

5 rows × 38 columns

```
# Check if you changed the column name
# for column in df_train:
#     columnSeriesObj = df_train[column]
#     print('Column Name : ', column)
#     print('Column Contents : ', columnSeriesObj.values)
#     print("----------------------------------------------")
```

```
# Checking the unique elements in Categorical Columns
print("Status: ",df_train["Status"].unique())
print("Respiratory Rate (breaths/min): ",df_train["Respiratory Rate (breaths/min)"].unique())
print("Heart Rate (rates/min: ",df_train["Heart Rate (rates/min"].unique())
print("Follow-up: ",df_train["Follow-up"].unique())
print("Gender: ",df_train["Gender"].unique())
print("Autopsy shows birth defect (if applicable): ",df_train["Autopsy shows birth defect (if applicable)"].unique())
print("Birth defects: ",df_train["Birth defects"].unique())
print("Blood test result: ",df_train["Blood test result"].unique())
print("Genetic Disorder: ",df_train["Genetic Disorder"].unique())
print("Disorder Subclass: ",df_train["Disorder Subclass"].unique())
```

```
Status:  ['Alive' 'Deceased']
Respiratory Rate (breaths/min):  ['Normal (30-60)' 'Tachypnea']
Heart Rate (rates/min:  ['Normal' 'Tachycardia']
Follow-up:  ['High' 'Low']
Gender:  ['Ambiguous' 'Male' 'Female']
Autopsy shows birth defect (if applicable):  ['Not applicable' 'None' 'No' 'Yes']
Birth defects:  ['Singular' 'Multiple']
Blood test result:  ['slightly abnormal' 'normal' 'inconclusive' 'abnormal']
Genetic Disorder:  ['Mitochondrial genetic inheritance disorders'
```

```
 'Multifactorial genetic inheritance disorders'
 'Single-gene inheritance diseases']
Disorder Subclass:  ["Leber's hereditary optic neuropathy" 'Cystic fibrosis' 'Diabetes'
 'Leigh syndrome' 'Cancer' 'Tay-Sachs' 'Hemochromatosis'
 'Mitochondrial myopathy' "Alzheimer's"]
```

```
# plots
df_train.head()
```

	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Mother's age	Father's age	Status	Respiratory Rate (breaths/min)	...	Birth defects	White Blood cell count (thousand per microliter)	
0	2.0	1	0	1	0	4.760603	23.0	20.0	Alive	Normal (30-60)	...	Singular	9.857562	
1	4.0	1	1	0	0	4.910669	23.0	23.0	Deceased	Tachypnea	...	Multiple	5.52256	
2	6.0	1	0	0	0	4.893297	41.0	22.0	Alive	Normal (30-60)	...	Singular	3.0	
3	12.0	1	0	1	0	4.705280	21.0	20.0	Deceased	Tachypnea	...	Singular	7.919321	inc
4	11.0	1	0	1	1	4.720703	32.0	20.0	Alive	Tachypnea	...	Multiple	4.09821	

5 rows × 38 columns

```
# Changing Categorical Values to Numerical Values
#Alive':1 'Deceased:0'
df_train["Status"]=[1 if i.strip()== "Alive" else 0 for i in df_train["Status"]]
#Normal (30-60):1' 'Tachypnea:0
df_train["Respiratory Rate (breaths/min)"]=[1 if i.strip()== "Normal (30-60)" else 0 for i in df_train["Respiratory Rate (breaths/min)"]]
#Normal:1' 'Tachycardia:0
df_train["Heart Rate (rates/min"]=[1 if i.strip()== "Normal" else 0 for i in df_train["Heart Rate (rates/min"]]
#High:1, Low:0
df_train["Follow-up"]=[1 if i.strip()== "High" else 0 for i in df_train["Follow-up"]]
#['Singular' 'Multiple']
df_train["Birth defects"]=[1 if i.strip()== "Singular" else 0 for i in df_train["Birth defects"]]
#1: male 0: female 2: ambiguous
df_train["Gender"]=[1 if i.strip()== "Male" else 0 if i.strip() == "Female" else 2 for i in df_train["Gender"]]
#Not applicable:3' 'None:2' 'No:0' 'Yes:1'
df_train["Autopsy shows birth defect (if applicable)"]=[1 if i.strip()== "Yes" else 0 if i.strip() == "No" else 2 if i.strip()=="None" else 3
#'slightly abnormal':1, 'normal':0, 'inconclusive':2 'abnormal:3']
df_train["Blood test result"]=[1 if i.strip()== "slightly abnormal" else 0 if i.strip() == "normal" else 2 if i.strip()=="inconclusive" else
#'Mitochondrial genetic inheritance disorders':1,'Multifactorial genetic inheritance disorders':0'Single-gene inheritance diseases:2'
df_train["Genetic Disorder"]=[1 if i.strip()== "Mitochondrial genetic inheritance disorders" else 0 if i.strip() == "Multifactorial genetic i
#Leber's hereditary optic neuropathy:1
#Cystic fibrosis:0
#Diabetes:2
#Leigh syndrome:3
#Cancer:4
#Tay-Sachs:5
#Hemochromatosis:6
#Mitochondrial myopathy:7
#Alzheimer's:8
df_train["Disorder Subclass"]=[1 if i.strip()== "Leber's hereditary optic neuropathy"
                               else 0 if i.strip() == "Cystic fibrosis"
                              else 2 if i.strip()=="Diabetes"
                              else 3 if i.strip()=="Leigh syndrome"
                              else 4 if i.strip()=="Cancer"
                              else 5 if i.strip()=="Tay-Sachs"
                              else 6 if i.strip()=="Hemochromatosis"
                              else 7 if i.strip()=="Mitochondrial myopathy"
                              else 8 for i in df_train["Disorder Subclass"]]
```

```
df_train["total symptom"]=(df_train["Symptom 1"]+df_train["Symptom 2"]+df_train["Symptom 3"]+df_train["Symptom 4"]+df_train["Symptom 5"]) / 5
df_train.drop(["Symptom 1","Symptom 2","Symptom 3","Symptom 4","Symptom 5"],axis=1,inplace=True)
```

```
---------------------------------------------------------------------------
TypeError                                 Traceback (most recent call last)
/usr/local/lib/python3.10/dist-packages/pandas/core/ops/array_ops.py in _na_arithmetic_op(left, right, op, is_cmp)
    164     try:
--> 165         result = func(left, right)
    166     except TypeError:

 10 frames 

TypeError: unsupported operand type(s) for +: 'float' and 'str'

During handling of the above exception, another exception occurred:

TypeError                                 Traceback (most recent call last)
/usr/local/lib/python3.10/dist-packages/pandas/core/ops/array_ops.py in _masked_arith_op(x, y, op)
    108         # See GH#5284, GH#5035, GH#19448 for historical reference
    109         if mask.any():
--> 110             result[mask] = op(xrav[mask], yrav[mask])
    111
    112     else:

TypeError: unsupported operand type(s) for +: 'float' and 'str'
```

SEARCH STACK OVERFLOW

```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22083 entries, 0 to 22082
Data columns (total 38 columns):
 #   Column                                          Non-Null Count  Dtype  
---  ------                                          --------------  -----  
 0   Patient Age                                     22083 non-null  object 
 1   Genes in mother's side                          22083 non-null  int64  
 2   Inherited from father                           22083 non-null  int64  
 3   Maternal gene                                   22083 non-null  int64  
 4   Paternal gene                                   22083 non-null  int64  
 5   Blood cell count (mcL)                          22083 non-null  float64
 6   Mother's age                                    22083 non-null  object 
 7   Father's age                                    22083 non-null  object 
 8   Status                                          22083 non-null  int64  
 9   Respiratory Rate (breaths/min)                  22083 non-null  int64  
 10  Heart Rate (rates/min                           22083 non-null  int64  
 11  Test 1                                          22083 non-null  object 
 12  Test 2                                          22083 non-null  object 
 13  Test 3                                          22083 non-null  object 
 14  Test 4                                          22083 non-null  object 
 15  Test 5                                          22083 non-null  object 
 16  Parental consent                                22083 non-null  int64  
 17  Follow-up                                       22083 non-null  int64  
 18  Gender                                          22083 non-null  int64  
 19  Birth asphyxia                                  22083 non-null  int64  
 20  Autopsy shows birth defect (if applicable)      22083 non-null  int64  
 21  Folic acid details (peri-conceptional)          22083 non-null  int64  
 22  H/O serious maternal illness                    22083 non-null  int64  
 23  H/O radiation exposure (x-ray)                  22083 non-null  int64  
 24  H/O substance abuse                             22083 non-null  int64  
 25  Assisted conception IVF/ART                     22083 non-null  int64  
 26  History of anomalies in previous pregnancies    22083 non-null  int64  
 27  No. of previous abortion                        22083 non-null  object 
 28  Birth defects                                   22083 non-null  int64  
 29  White Blood cell count (thousand per microliter)  22083 non-null  object 
 30  Blood test result                               22083 non-null  int64  
 31  Symptom 1                                       22083 non-null  object 
 32  Symptom 2                                       22083 non-null  object 
 33  Symptom 3                                       22083 non-null  object 
 34  Symptom 4                                       22083 non-null  object 
 35  Symptom 5                                       22083 non-null  object 
 36  Genetic Disorder                                22083 non-null  int64  
 37  Disorder Subclass                               22083 non-null  int64  
dtypes: float64(1), int64(22), object(15)
memory usage: 6.4+ MB
```

```
# Changing the datatype to float
df_train = df_train.apply(pd.to_numeric,downcast="float")
```

```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 22083 entries, 0 to 22082
Data columns (total 38 columns):
 #   Column                                          Non-Null Count  Dtype
---  ------                                          --------------  -----
 0   Patient Age                                     22083 non-null  float32
 1   Genes in mother's side                          22083 non-null  float32
 2   Inherited from father                           22083 non-null  float32
 3   Maternal gene                                   22083 non-null  float32
 4   Paternal gene                                   22083 non-null  float32
 5   Blood cell count (mcL)                          22083 non-null  float32
 6   Mother's age                                    22083 non-null  float32
 7   Father's age                                    22083 non-null  float32
 8   Status                                          22083 non-null  float32
 9   Respiratory Rate (breaths/min)                  22083 non-null  float32
 10  Heart Rate (rates/min                           22083 non-null  float32
 11  Test 1                                          22083 non-null  float32
 12  Test 2                                          22083 non-null  float32
 13  Test 3                                          22083 non-null  float32
 14  Test 4                                          22083 non-null  float32
 15  Test 5                                          22083 non-null  float32
 16  Parental consent                                22083 non-null  float32
 17  Follow-up                                       22083 non-null  float32
 18  Gender                                          22083 non-null  float32
 19  Birth asphyxia                                  22083 non-null  float32
 20  Autopsy shows birth defect (if applicable)      22083 non-null  float32
 21  Folic acid details (peri-conceptional)          22083 non-null  float32
 22  H/O serious maternal illness                    22083 non-null  float32
 23  H/O radiation exposure (x-ray)                  22083 non-null  float32
 24  H/O substance abuse                             22083 non-null  float32
 25  Assisted conception IVF/ART                     22083 non-null  float32
 26  History of anomalies in previous pregnancies    22083 non-null  float32
 27  No. of previous abortion                        22083 non-null  float32
 28  Birth defects                                   22083 non-null  float32
 29  White Blood cell count (thousand per microliter)  22083 non-null  float32
 30  Blood test result                               22083 non-null  float32
 31  Symptom 1                                       22083 non-null  float32
 32  Symptom 2                                       22083 non-null  float32
 33  Symptom 3                                       22083 non-null  float32
 34  Symptom 4                                       22083 non-null  float32
 35  Symptom 5                                       22083 non-null  float32
 36  Genetic Disorder                                22083 non-null  float32
 37  Disorder Subclass                               22083 non-null  float32
dtypes: float32(38)
memory usage: 3.2 MB
```

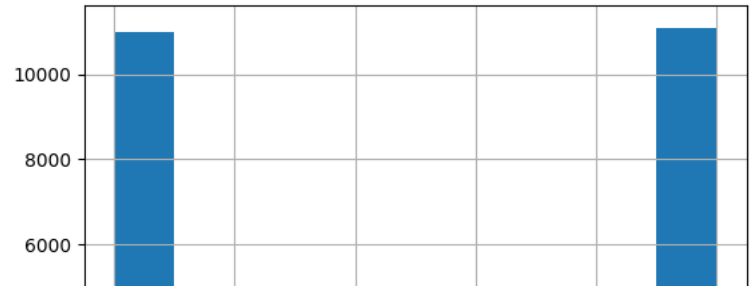
```
df_train.head()
```

	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Mother's age	Father's age	Status	Respiratory Rate (breaths/min)	...	Birth defects	White Blood cell count (thousand per microliter)	Bloc tes resul
0	2.0	1.0	0.0	1.0	0.0	4.760603	23.0	20.0	1.0	1.0	...	1.0	9.857562	1
1	4.0	1.0	1.0	0.0	0.0	4.910669	23.0	23.0	0.0	0.0	...	0.0	5.522560	0
2	6.0	1.0	0.0	0.0	0.0	4.893297	41.0	22.0	1.0	1.0	...	1.0	3.000000	0
3	12.0	1.0	0.0	1.0	0.0	4.705280	21.0	20.0	0.0	0.0	...	1.0	7.919321	2
4	11.0	1.0	0.0	1.0	1.0	4.720703	32.0	20.0	1.0	0.0	...	0.0	4.098210	1

5 rows × 38 columns

```
df_train["Status"].hist()
```

```
<Axes: >
```

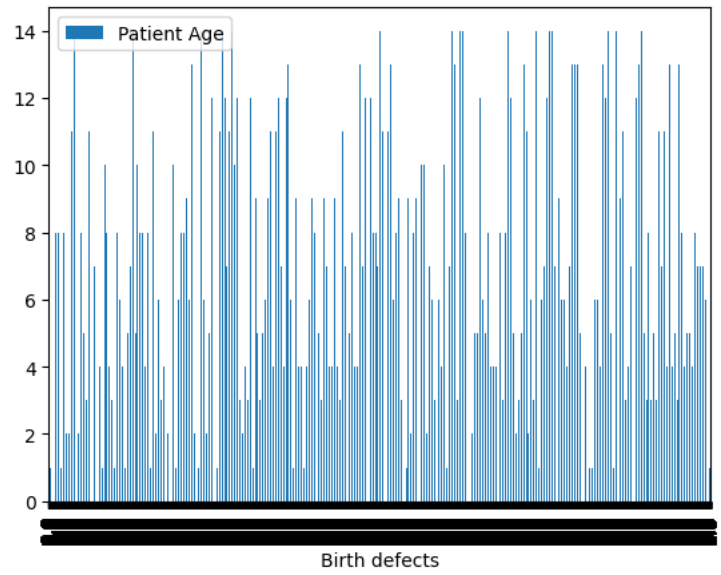



```
df_train.Status.value_counts()
```

```
1.0    11083
0.0    11000
Name: Status, dtype: int64
```

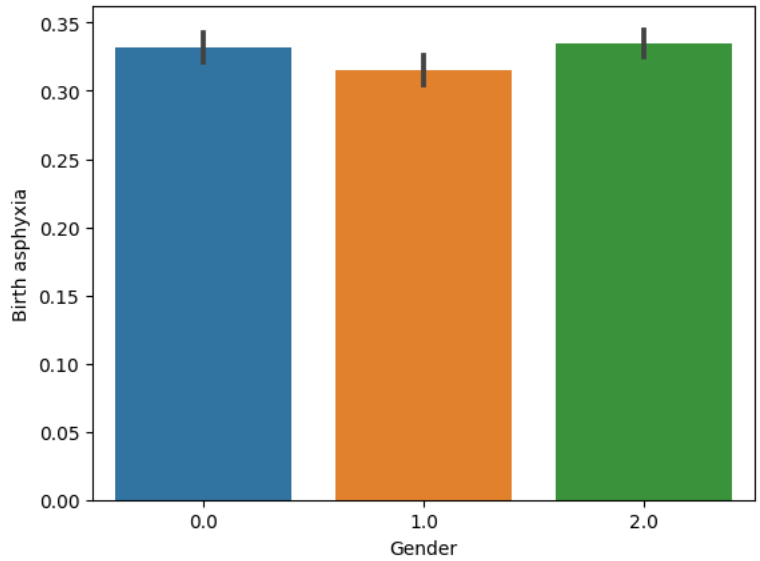
```
df_train.plot.bar(y="Patient Age",x="Birth defects")
```

```
<Axes: xlabel='Birth defects'>
```



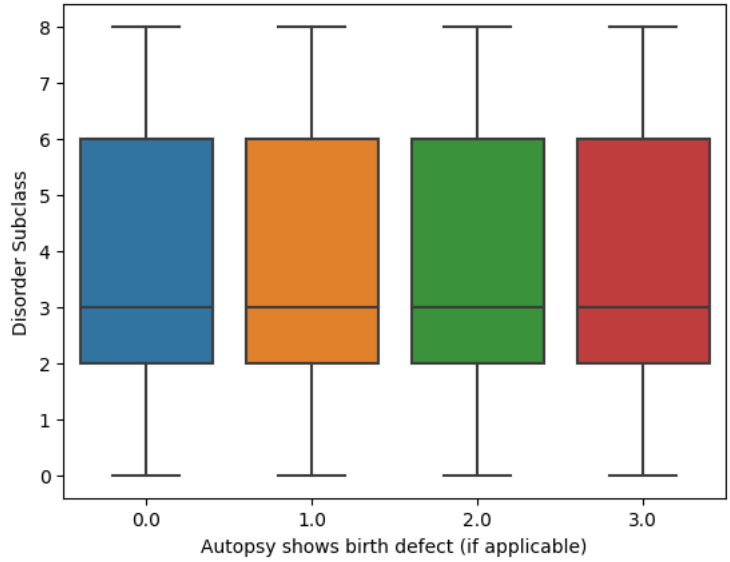

```
sns.barplot(x="Gender",y="Birth asphyxia",data=df_train)
```

```
<Axes: xlabel='Gender', ylabel='Birth asphyxia'>
```



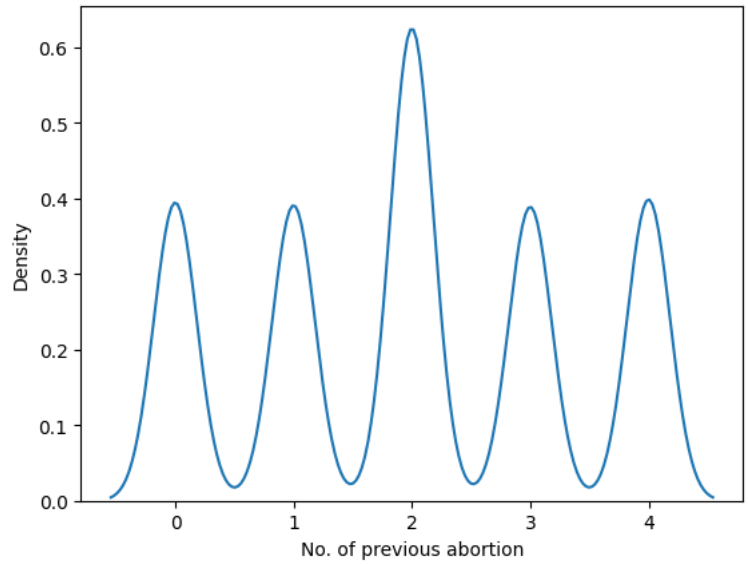

```
sns.boxplot(x="Autopsy shows birth defect (if applicable)",y="Disorder Subclass",data=df_train)
```

```
<Axes: xlabel='Autopsy shows birth defect (if applicable)', ylabel='Disorder Subclass'>
```




```
sns.kdeplot(df_train["No. of previous abortion"],palette="dark")
```

```
<Axes: xlabel='No. of previous abortion', ylabel='Density'>
```

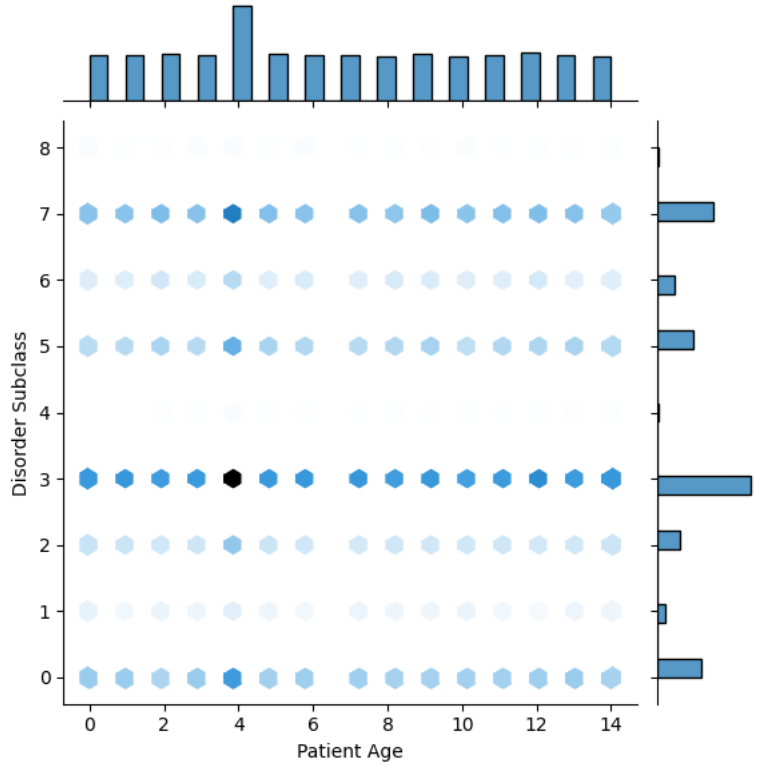



```
# Distplot
sns.distplot(df_train['Disorder Subclass'],color="green",bins=30)
```

```
<Axes: xlabel='Disorder Subclass', ylabel='Density'>
```

```
#JointPlot
plt.figure(figsize=(12,6))
sns.jointplot(x=df_train["Patient Age"],y=df_train['Disorder Subclass'],kind="hex")
```

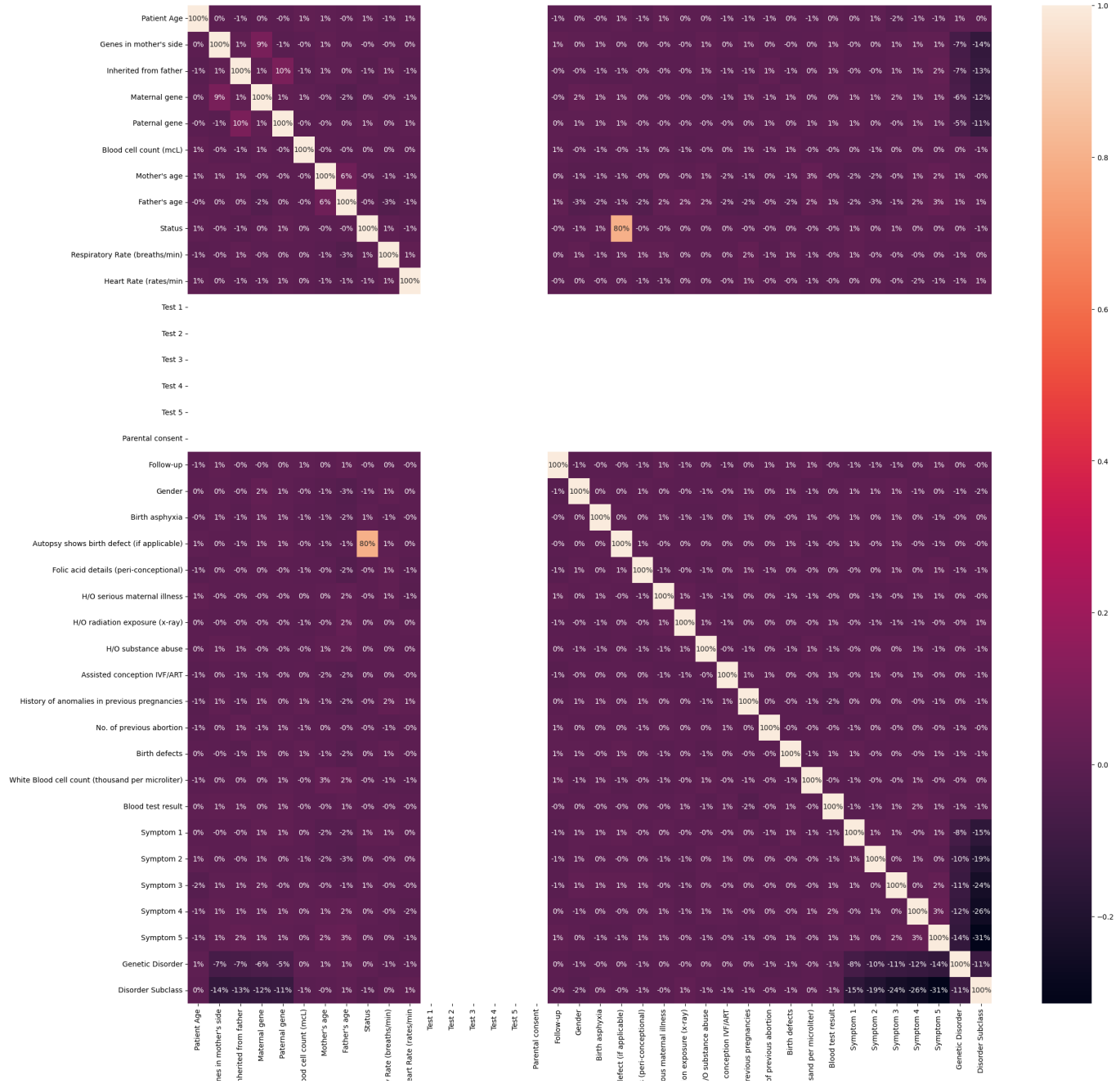
```
<seaborn.axisgrid.JointGrid at 0x7b4335479600>
<Figure size 1200x600 with 0 Axes>
```




```
df_train.corr()
```

	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Mother's age	Father's age	Status	Respiratory Rate (breaths/min)	...	Birth defects
Patient Age	1.000000	0.003452	-0.008275	0.001492	-0.004422	0.010155	0.008203	-0.000949	0.007764	-0.011186	...	0.002748
Genes in mother's side	0.003452	1.000000	0.008960	0.089605	-0.007389	-0.002403	0.010247	0.000520	-0.000221	-0.001917	...	-0.002398
Inherited from father	-0.008275	0.008960	1.000000	0.009222	0.095115	-0.007677	0.006349	0.003769	-0.012293	0.011765	...	-0.012021
Maternal gene	0.001492	0.089605	0.009222	1.000000	0.011829	0.008119	-0.004247	-0.022747	0.003566	-0.003921	...	0.013352
Paternal gene	-0.004422	-0.007389	0.095115	0.011829	1.000000	-0.003494	-0.001070	0.000636	0.013799	0.000572	...	0.002223
Blood cell count (mcL)	0.010155	-0.002403	-0.007677	0.008119	-0.003494	1.000000	-0.001129	-0.003498	0.003149	0.002200	...	0.011251
Mother's age	0.008203	0.010247	0.006349	-0.004247	-0.001070	-0.001129	1.000000	0.059002	-0.001519	-0.008097	...	-0.005871
Father's age	-0.000949	0.000520	0.003769	-0.022747	0.000636	-0.003498	0.059002	1.000000	-0.000087	-0.030954	...	-0.016907
Status	0.007764	-0.000221	-0.012293	0.003566	0.013799	0.003149	-0.001519	-0.000087	1.000000	0.008206	...	0.001038
Respiratory Rate (breaths/min)	-0.011186	-0.001917	0.011765	-0.003921	0.000572	0.002200	-0.008097	-0.030954	0.008206	1.000000	...	0.009044
Heart Rate (rates/min	0.008489	0.001019	-0.010575	-0.005682	0.005119	0.000185	-0.008515	-0.012575	-0.005491	0.013372	...	-0.001436
Test 1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
Test 2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
Test 3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
Test 4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
Test 5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
Parental consent	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
Follow-up	-0.005463	0.009939	-0.003019	-0.003337	0.003752	0.005366	0.002463	0.009632	-0.000193	0.002213	...	0.008091
Gender	0.000485	0.004320	-0.002209	0.020249	0.006780	-0.004604	-0.011823	-0.026024	-0.006103	0.010699	...	0.006707
Birth asphyxia	-0.001758	0.005706	-0.009471	0.010482	0.006934	-0.007196	-0.014797	-0.017346	0.008079	-0.007212	...	-0.000958
Autopsy shows birth defect (if applicable)	0.007851	0.000360	-0.008126	0.009456	0.013202	-0.000833	-0.013375	-0.010641	0.797307	0.005417	...	0.011019
Folic acid details (peri-conceptional)	-0.008526	0.000115	-0.001314	0.003564	-0.000652	-0.005324	-0.002391	-0.023858	-0.001263	0.012461	...	0.000713
H/O serious maternal illness	0.010857	-0.001007	-0.001980	-0.001020	-0.000146	0.002725	0.004212	0.017155	-0.001300	0.011142	...	-0.006746
H/O radiation exposure (x-ray)	0.003394	-0.003072	-0.004070	-0.004825	-0.002030	-0.009873	-0.000561	0.024242	0.000139	0.000782	...	0.003333
H/O substance abuse	0.001144	0.010171	0.005247	-0.002817	-0.001557	-0.003252	0.012517	0.015525	0.001333	0.001821	...	-0.013496
Assisted conception IVF/ART	-0.007153	0.003390	-0.010749	-0.011799	-0.003710	0.002817	-0.016697	-0.019483	0.002584	0.000868	...	0.003539
History of												

```
plt.figure(figsize=(25,25))
sns.heatmap(df_train.iloc[:,0:39].corr(),annot=True,fmt=".0%")
plt.show()
```




```
df_train.columns
```

```
Index(['Patient Age', 'Genes in mother's side', 'Inherited from father',
       'Maternal gene', 'Paternal gene', 'Blood cell count (mcL)',
       'Mother's age', 'Father's age', 'Status',
       'Respiratory Rate (breaths/min)', 'Heart Rate (rates/min', 'Test 1',
       'Test 2', 'Test 3', 'Test 4', 'Test 5', 'Parental consent', 'Follow-up',
       'Gender', 'Birth asphyxia',
       'Autopsy shows birth defect (if applicable)',
       'Folic acid details (peri-conceptional)',
       'H/O serious maternal illness', 'H/O radiation exposure (x-ray)',
       'H/O substance abuse', 'Assisted conception IVF/ART',
       'History of anomalies in previous pregnancies',
       'No. of previous abortion', 'Birth defects',
       'White Blood cell count (thousand per microliter)', 'Blood test result',
       'Symptom 1', 'Symptom 2', 'Symptom 3', 'Symptom 4', 'Symptom 5',
       'Genetic Disorder', 'Disorder Subclass'],
      dtype='object')
```

```
df_test.head()
```

	Patient Id	Patient Age	Genes in mother's side	Inherited from father	Maternal gene	Paternal gene	Blood cell count (mcL)	Patient First Name	Family Name	Father's name	...	History of anomalies in previous pregnancies	No. of previous abortion	Birth defects
0	PID0x4175	6	No	Yes	No	No	4.981655	Charles	NaN	Kore	...	-99	2	Multiple
1	PID0x21f5	10	Yes	No	NaN	Yes	5.118890	Catherine	NaN	Homero	...	Yes	-99	Multiple
2	PID0x49b8	5	No	NaN	No	No	4.876204	James	NaN	Danield	...	No	0	Singular
3	PID0x2d97	13	No	Yes	Yes	No	4.687767	Brian	NaN	Orville	...	Yes	-99	Singular
4	PID0x58da	5	No	NaN	NaN	Yes	5.152362	Gary	NaN	Issiah	...	No	-99	Multiple

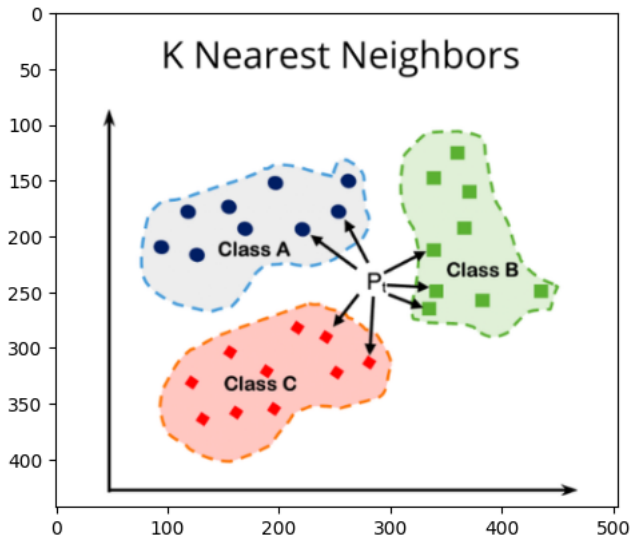
5 rows × 43 columns

```
df_train['Genetic Disorder'].head()
```

```
0    1.0
1    1.0
2    0.0
3    1.0
4    0.0
Name: Genetic Disorder, dtype: float32
```

KNN

```
img = mpimg.imread('knn.png')
imgplot = plt.imshow(img)
plt.show()
```



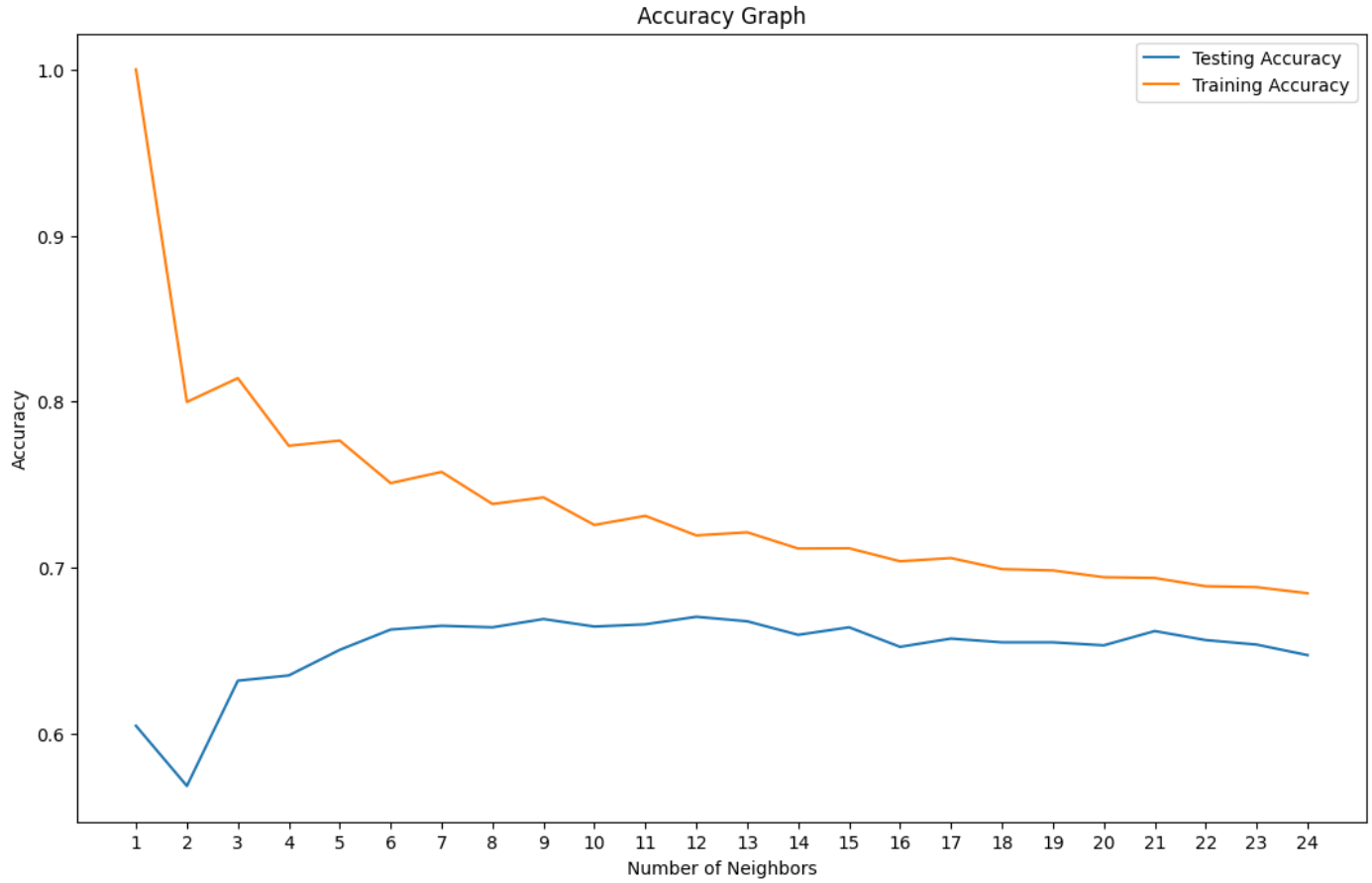

```
# train test split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.model_selection import train_test_split
x,y = df_train.loc[:,df_train.columns != 'Status'], df_train.loc[:,'Status']
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size = 0.2,random_state = 1)
knn = KNeighborsClassifier(n_neighbors = 3)
knn.fit(x_train,y_train) #  Learn and estimate the parameters of the transformation
prediction = knn.predict(x_test)
#print('Prediction: {}'.format(prediction))
print('With KNN (K=3) accuracy is: ',knn.score(x_test,y_test)) # accuracy
```

```
With KNN (K=3) accuracy is:  0.7550373556712701
```

```
# train test split
x,y = df_train.loc[:,df_train.columns != 'Genetic Disorder'], df_train.loc[:,'Genetic Disorder']
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size = 0.1,random_state = 1)
knn = KNeighborsClassifier(n_neighbors = 3)
knn.fit(x_train,y_train)
prediction = knn.predict(x_test)
#print('Prediction: {}'.format(prediction))
print('With KNN (K=3) accuracy is: ',knn.score(x_test,y_test)) # accuracy
```

```
With KNN (K=3) accuracy is:  0.6319601629696695
```

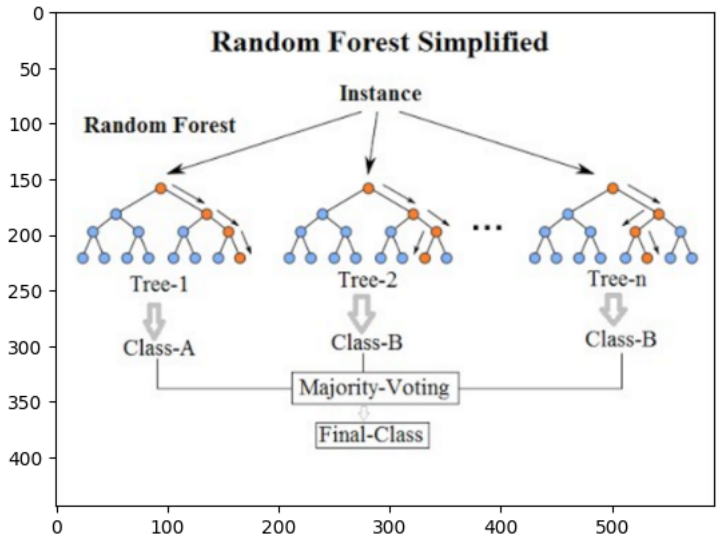
```
neig = np.arange(1, 25)
train_accuracy = []
test_accuracy = []
# Loop over different values of k
for i, k in enumerate(neig):
    # k from 1 to 25(exclude)
    knn = KNeighborsClassifier(n_neighbors=k)
    # Fit with knn
    knn.fit(x_train,y_train)
    #train accuracy
    train_accuracy.append(knn.score(x_train, y_train))
    # test accuracy
    test_accuracy.append(knn.score(x_test, y_test))

# Plot
plt.figure(figsize=[13,8])
plt.plot(neig, test_accuracy, label = 'Testing Accuracy')
plt.plot(neig, train_accuracy, label = 'Training Accuracy')
plt.legend()
plt.title('Accuracy Graph')
plt.xlabel('Number of Neighbors')
plt.ylabel('Accuracy')
plt.xticks(neig)
plt.savefig('graph.png')
plt.show()
print("Best accuracy is {} with K = {}".format(np.max(test_accuracy),1+test_accuracy.index(np.max(test_accuracy))))
```




```
Best accuracy is 0.6704391127206881 with K = 12
```

Random Forest

```
img = mpimg.imread('Random Forest.png')
imgplot = plt.imshow(img)
plt.figure(figsize=(12,8))
plt.show()
```

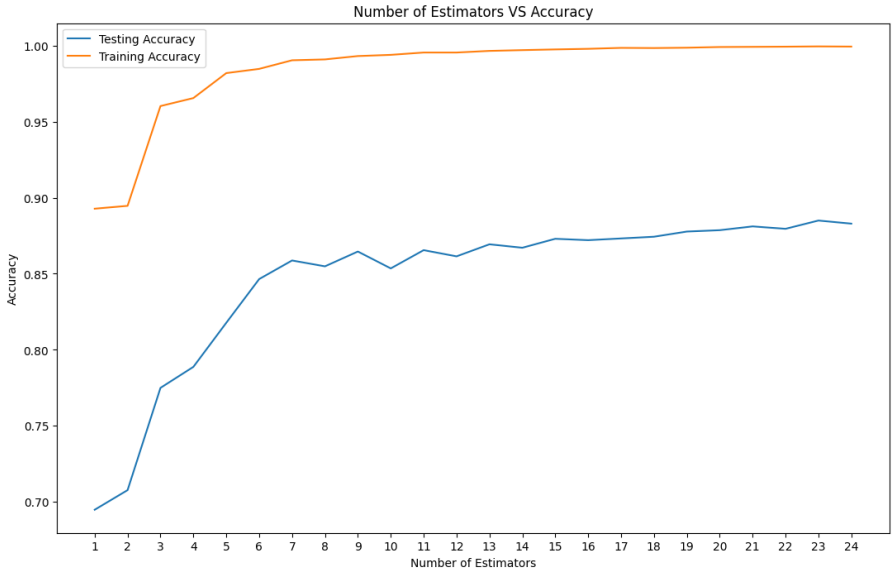



```
<Figure size 1200x800 with 0 Axes>
```

```
x, y = df_train.loc[:, df_train.columns != 'Genetic Disorder'], df_train.loc[:, 'Genetic Disorder']
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.8, random_state=1)

# Create a Random Forest classifier
rf = RandomForestClassifier(n_estimators=120, random_state=1)  # You can adjust n_estimators as needed

rf.fit(x_train, y_train)

prediction = rf.predict(x_test)
accuracy = rf.score(x_test, y_test)
print('Random Forest accuracy for predicting Disorder Subclass is:', accuracy)
```

```
Random Forest accuracy for predicting Disorder Subclass is: 0.8584932359766797
```

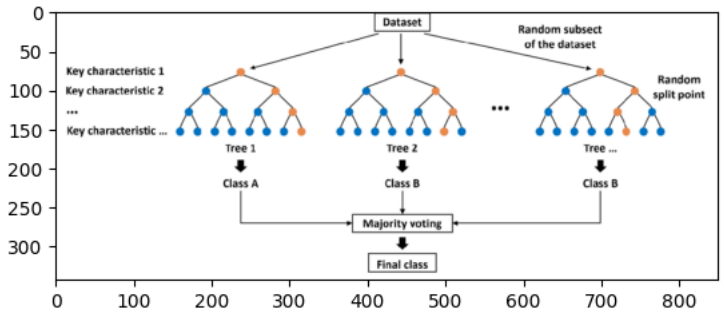
```
from sklearn.ensemble import RandomForestClassifier
import matplotlib.pyplot as plt
import numpy as np

n_estimators = np.arange(1, 25)
train_accuracy = []
test_accuracy = []

for n in n_estimators:
    # Create a Random Forest classifier with 'n' estimators
    rf = RandomForestClassifier(n_estimators=n, random_state=1)
    rf.fit(x_train, y_train)

    # Train accuracy
    train_accuracy.append(rf.score(x_train, y_train))

    # Test accuracy
    test_accuracy.append(rf.score(x_test, y_test))

# Plot
plt.figure(figsize=[13, 8])
plt.plot(n_estimators, test_accuracy, label='Testing Accuracy')
plt.plot(n_estimators, train_accuracy, label='Training Accuracy')
plt.legend()
plt.title('Number of Estimators VS Accuracy')
plt.xlabel('Number of Estimators')
plt.ylabel('Accuracy')
plt.xticks(n_estimators)
plt.savefig('rf_graph.png')
plt.show()

best_accuracy = max(test_accuracy)
best_n_estimators = n_estimators[test_accuracy.index(best_accuracy)]

print("Best accuracy is {} with n_estimators = {}".format(best_accuracy, best_n_estimators))
```



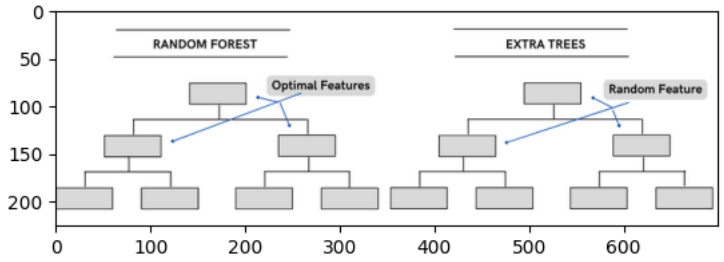

```
Best accuracy is 0.8849898120896537 with n_estimators = 23
```

Extra Tree Classifier

```
img = mpimg.imread('Extra Tree Classifier.png')
plt.imshow(img)
# plt.figure(figsize=(14,12))
plt.show()
```




```
img = mpimg.imread('Extra - Random.png')
plt.imshow(img)
# plt.figure(figsize=(14,12))
plt.show()
```




```
# Import necessary libraries
from sklearn.ensemble import ExtraTreesClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score


#  Split the data into training and testing sets
# X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

x, y = df_train.loc[:, df_train.columns != 'Genetic Disorder'], df_train.loc[:, 'Genetic Disorder']

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=1)


extra_trees_classifier = ExtraTreesClassifier(n_estimators=100, random_state=42)


extra_trees_classifier.fit(x_train,y_train)


y_pred = extra_trees_classifier.predict(x_test)

accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
```

```
Accuracy: 0.8822730359972832
```

```
n_estimators = np.arange(1, 25)
train_accuracy = []
test_accuracy = []

for n in n_estimators:
    # Create a Extra Trees Classifier with 'n' estimators
    rf = ExtraTreesClassifier(n_estimators=n, random_state=1)
    rf.fit(x_train, y_train)

    # Train accuracy
    train_accuracy.append(rf.score(x_train, y_train))

    # Test accuracy
    test_accuracy.append(rf.score(x_test, y_test))

# Plot
plt.figure(figsize=[13, 8])
plt.plot(n_estimators, test_accuracy, label='Testing Accuracy')
plt.plot(n_estimators, train_accuracy, label='Training Accuracy')
plt.legend()
plt.title('Number of Estimators VS Accuracy')
plt.xlabel('Number of Estimators')
plt.ylabel('Accuracy')
plt.xticks(n_estimators)
plt.savefig('rf_graph.png')
plt.show()

best_accuracy = max(test_accuracy)
best_n_estimators = n_estimators[test_accuracy.index(best_accuracy)]

print("Best accuracy is {} with n_estimators = {}".format(best_accuracy, best_n_estimators))
```