Machine learning approaches classify clinical malaria outcomes based on haematological parameters

#Number of columns

fatemeh.columns

```
Index(['SampleID', 'consent_given', 'location', 'Enrollment_Year', 'bednet',
       'fever_symptom', 'temperature', 'Suspected_Organism',
       'Suspected_infection', 'RDT', 'Blood_culture', 'Urine_culture',
       'Taq_man_PCR', 'parasite_density', 'Microscopy', 'Laboratory_Results',
       'Clinical_Diagnosis', 'wbc_count', 'rbc_count', 'hb_level',
       'hematocrit', 'mean_cell_volume', 'mean_corp_hb', 'mean_cell_hb_conc',
       'platelet_count', 'platelet_distr_width', 'mean_platelet_vl',
       'neutrophils_percent', 'lymphocytes_percent', 'mixed_cells_percent',
       'neutrophils_count', 'lymphocytes_count', 'mixed_cells_count',
       'RBC_dist_width_Percent'],
      dtype='object')
```

#shows first 5 rows of data set

fatemeh.head()

	SampleID	consent_given	location	Enrollment_Year	bednet	fever_symptom	temperature	Suspected_Organism	Suspected_infection	RDT	...	platelet_count
0	CCS20043	yes	Navrongo	2004	NaN	Yes	38.0	Not Known / Missing entry	NaN	Positive	...	156.0
1	CCS20102	yes	Navrongo	2004	NaN	Yes	38.2	Not Known / Missing entry	NaN	Positive	...	55.0
2	CCS20106	yes	Navrongo	2004	NaN	Yes	37.7	Not Known / Missing entry	NaN	Positive	...	20.0
3	CCS20147	yes	Navrongo	2004	NaN	Yes	37.7	Not Known / Missing entry	NaN	Positive	...	132.0
4	CCS20170	yes	Navrongo	2004	NaN	Yes	37.1	Not Known / Missing entry	NaN	Positive	...	85.0

5 rows × 34 columns

platelet_distr_width	mean_platelet_vl	neutrophils_percent	lymphocytes_percent	mixed_cells_percent	neutrophils_count	lymphocytes_count	mixed_cells_count
8.2	6.8	61.8	31.7	6.5	3.6	1.8	0.3
16.5	7.6	68.5	23.6	7.9	5.4	1.8	0.6
2.3	5.9	32.8	53.3	13.9	2.8	4.3	1.1
17.2	6.2	82.6	11.5	5.9	13.2	1.8	0.9
16.1	6.8	83.7	11.3	5.0	3.8	0.5	0.2

#shows last 5 rows of data set

fatemeh.tail()

	SampleID	consent_given	location	Enrollment_Year	bednet	fever_symptom	temperature	Suspected_Organism	Suspected_infection	RDT	...	platelet_count
2202	KC366	yes	Kintampo	2017	yes	No	37.1	Bacteria/Protozoa	Malaria/LRTI	Positive	...	277.0
2203	KC368	yes	Kintampo	2017	no	No	36.7	Bacteria/Protozoa	Helminthiasis	Negative	...	340.0
2204	KC369	yes	Kintampo	2017	yes	No	36.4	Bacteria	Dermatitis	Negative	...	300.0
2205	KC370	yes	Kintampo	2017	yes	No	37.4	Not Known / Missing entry	URTI	Negative	...	136.0
2206	KC375	yes	Kintampo	2017	yes	No	36.4	Protozoan	Instetinal flagellates	Negative	...	272.0

5 rows × 34 columns

platelet_distr_width	mean_platelet_vl	neutrophils_percent	lymphocytes_percent	mixed_cells_percent	neutrophils_count	lymphocytes_count	mixed_cells_count
12.3	7.1	71.3	22.6	6.1	8.9	2.7	0.7
15.2	7.2	73.6	21.0	5.4	6.0	1.7	0.4
14.3	6.5	43.6	49.4	7.0	4.3	4.6	0.6
13.3	7.1	35.3	58.1	6.6	4.4	7.1	0.8
12.9	7.3	59.1	35.4	5.5	2.9	1.6	0.2

#getting to know about rows and columns print("Diabetes data set dimensions :{} ".format(diabetes.shape))

fatemeh data set dimensions :(2207, 34)

*data set contain 2207 rows and 34 columns

#knowledge of data type helps for computation diabetes.dtypes

SampleID	object
consent_given	object
location	object
Enrollment_Year	int64
bednet	object
fever_symptom	object
temperature	float64
Suspected_Organism	object
Suspected_infection	object
RDT	object
Blood_culture	object
Urine_culture	object
Taq_man_PCR	object
parasite_density	float64
Microscopy	object
Laboratory_Results	object
Clinical_Diagnosis	object
wbc_count	float64
rbc_count	float64
hb_level	float64
hematocrit	float64
mean_cell_volume	float64
mean_corp_hb	float64
mean_cell_hb_conc	float64
platelet_count	float64
platelet_distr_width	float64
mean_platelet_vl	float64

```
neutrophils_percent       float64
lymphocytes_percent       float64
mixed_cells_percent       float64
neutrophils_count         float64
lymphocytes_count         float64
mixed_cells_count         float64
RBC_dist_width_Percent    float64
dtype: object
```

#This method prints information about a Data Frame including the index dtype and columns, #non-null values and memory usage.

fatemeh.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2207 entries, 0 to 2206
Data columns (total 34 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   SampleID                2207 non-null   object
 1   consent_given           2207 non-null   object
 2   location                2207 non-null   object
 3   Enrollment_Year         2207 non-null   int64
 4   bednet                  1676 non-null   object
 5   fever_symptom           2200 non-null   object
 6   temperature             2197 non-null   float64
 7   Suspected_Organism      2207 non-null   object
 8   Suspected_infection     1569 non-null   object
 9   RDT                     2065 non-null   object
 10  Blood_culture           122 non-null    object
 11  Urine_culture           112 non-null    object
 12  Taq_man_PCR             176 non-null    object
 13  parasite_density        2173 non-null   float64
 14  Microscopy              2170 non-null   object
 15  Laboratory_Results      2207 non-null   object
 16  Clinical_Diagnosis      2207 non-null   object
 17  wbc_count               2207 non-null   float64
 18  rbc_count               2207 non-null   float64
 19  hb_level                2207 non-null   float64
 20  hematocrit              2207 non-null   float64
 21  mean_cell_volume        2207 non-null   float64
```

```
 22  mean_corp_hb            2204 non-null   float64
 23  mean_cell_hb_conc       2205 non-null   float64
 24  platelet_count          2198 non-null   float64
 25  platelet_distr_width    2175 non-null   float64
 26  mean_platelet_vl        2190 non-null   float64
 27  neutrophils_percent     2207 non-null   float64
 28  lymphocytes_percent     2207 non-null   float64
 29  mixed_cells_percent     2207 non-null   float64
 30  neutrophils_count       2195 non-null   float64
 31  lymphocytes_count       2196 non-null   float64
 32  mixed_cells_count       2196 non-null   float64
 33  RBC_dist_width_Percent  2198 non-null   float64
dtypes: float64(19), int64(1), object(14)
memory usage: 586.4+ KB
```

#NaN

fatemeh.isna().sum()

```
SampleID                  0
consent_given             0
location                  0
Enrollment_Year           0
bednet                  531
fever_symptom             7
temperature              10
Suspected_Organism        0
Suspected_infection     638
RDT                     142
Blood_culture          2085
Urine_culture          2095
Taq_man_PCR            2031
parasite_density         34
Microscopy               37
Laboratory_Results        0
Clinical_Diagnosis        0
wbc_count                 0
rbc_count                 0
hb_level                  0
hematocrit                0
mean_cell_volume          0
mean_corp_hb              3
mean_cell_hb_conc         2
platelet_count            9
```

```
platelet_distr_width        32
mean_platelet_vl            17
neutrophils_percent          0
lymphocytes_percent          0
mixed_cells_percent          0
neutrophils_count           12
lymphocytes_count           11
mixed_cells_count           11
RBC_dist_width_Percent       9
dtype: int64
```

 # count :- the number of NoN-empty rows in a feature.

 # mean :- mean value of that feature.

 # std :- Standard Deviation Value of that feature.

 # min :- minimum value of that feature.

 # max :- maximum value of that feature.

 # 25%, 50%, and 75% are the percentile/quartile of each features.

fatemeh.describe()

	Enrollment_Year	temperature	parasite_density	wbc_count	rbc_count	hb_level	hematocrit	mean_cell_volume	mean_corp_hb	mean_cell_hb_conc
count	2207.000000	2197.000000	2.173000e+03	2207.000000	2207.000000	2207.000000	2207.000000	2207.000000	2204.000000	2205.000000
mean	2013.123244	37.869822	6.175196e+04	10.734209	3.890689	9.360222	29.101541	74.635850	24.102704	32.304259
std	5.701969	1.252016	3.258399e+05	5.924517	1.139474	2.680846	8.912130	8.239094	3.227082	2.893977
min	2002.000000	34.200000	0.000000e+00	0.500000	0.500000	1.400000	4.300000	7.800000	2.100000	15.700000
25%	2012.000000	36.800000	0.000000e+00	6.850000	3.300000	7.800000	23.700000	69.800000	22.100000	30.600000
50%	2017.000000	38.000000	4.800000e+02	9.300000	4.150000	10.100000	31.600000	75.000000	24.100000	32.100000
75%	2017.000000	38.900000	3.688000e+04	12.900000	4.640000	11.300000	35.400000	80.000000	26.200000	33.500000
max	2019.000000	41.100000	1.011400e+07	53.900000	6.670000	18.700000	52.700000	121.000000	38.800000	46.600000

platelet_count	platelet_distr_width	mean_platelet_vl	neutrophils_percent	lymphocytes_percent	mixed_cells_percent	neutrophils_count	lymphocytes_count
2198.000000	2175.000000	2190.000000	2207.000000	2207.000000	2207.000000	2195.000000	2196.000000
213.672611	14.124184	8.026119	58.486951	33.119574	8.393521	6.435157	3.450660
129.661849	3.092620	1.196604	16.561085	14.938599	3.466975	4.244165	2.560436
3.000000	0.000000	3.300000	9.300000	3.800000	0.300000	0.100000	0.300000
104.000000	12.900000	7.200000	45.900000	20.800000	5.800000	3.600000	1.700000
199.500000	14.900000	7.900000	59.200000	31.800000	8.100000	5.400000	2.800000
299.000000	15.600000	8.800000	72.100000	44.200000	10.600000	8.000000	4.400000
1087.000000	23.900000	18.600000	93.300000	81.500000	27.300000	42.000000	28.100000

mixed_cells_count	RBC_dist_width_Percent
2196.000000	2198.000000
0.856179	16.381797
0.637455	2.610800
0.000000	10.600000
0.400000	14.500000
0.700000	15.800000
1.100000	17.700000
5.600000	29.000000

fatemeh.shape

(2207, 34)

#Check for missing values

```
print((fatemeh[['SampleID', 'consent_given', 'location', 'Enrollment_Year', 'bednet',
       'fever_symptom', 'temperature', 'Suspected_Organism',
       'Suspected_infection', 'RDT', 'Blood_culture', 'Urine_culture',
       'Taq_man_PCR', 'parasite_density', 'Microscopy', 'Laboratory_Results',
       'Clinical_Diagnosis', 'wbc_count', 'rbc_count', 'hb_level',
       'hematocrit', 'mean_cell_volume', 'mean_corp_hb', 'mean_cell_hb_conc',
       'platelet_count', 'platelet_distr_width', 'mean_platelet_vl',
       'neutrophils_percent', 'lymphocytes_percent', 'mixed_cells_percent',
       'neutrophils_count', 'lymphocytes_count', 'mixed_cells_count',
       'RBC_dist_width_Percent']] == 0).sum())
```

```
SampleID                  0
consent_given             0
location                  0
Enrollment_Year           0
bednet                    0
fever_symptom             0
temperature               0
Suspected_Organism        0
Suspected_infection       0
RDT                       0
Blood_culture             0
Urine_culture             0
Taq_man_PCR               0
parasite_density       1021
Microscopy                0
Laboratory_Results        0
Clinical_Diagnosis        0
wbc_count                 0
rbc_count                 0
hb_level                  0
hematocrit                0
mean_cell_volume          0
mean_corp_hb              0
mean_cell_hb_conc         0
platelet_count            0
platelet_distr_width      4
mean_platelet_vl          0
```

```
neutrophils_percent          0
lymphocytes_percent          0
mixed_cells_percent          0
neutrophils_count            0
lymphocytes_count            0
mixed_cells_count            2
RBC_dist_width_Percent       0
dtype: int64
```

#Dealing with missing values

#Drop rows having NaN

data_nan=diabetes.copy()

print("Size before dropping NaN rows",data_nan.shape,"\n")

nan_dropped = data_nan.dropna()

print(nan_dropped.isnull().sum())

print("\nSize after dropping NaN rows",nan_dropped.shape)

```
Size before dropping NaN rows (2207, 34)

SampleID                  0
consent_given             0
location                  0
Enrollment_Year           0
bednet                    0
fever_symptom             0
temperature               0
Suspected_Organism        0
Suspected_infection       0
RDT                       0
Blood_culture             0
Urine_culture             0
Taq_man_PCR               0
parasite_density          0
Microscopy                0
Laboratory_Results        0
Clinical_Diagnosis        0
wbc_count                 0
rbc_count                 0
hb_level                  0
hematocrit                0
mean_cell_volume          0
mean_corp_hb              0
mean_cell_hb_conc         0
platelet_count            0
```

```
platelet_distr_width      0
mean_platelet_vl          0
neutrophils_percent       0
lymphocytes_percent       0
mixed_cells_percent       0
neutrophils_count         0
lymphocytes_count         0
mixed_cells_count         0
RBC_dist_width_Percent    0
dtype: int64

Size after dropping NaN rows (4, 34)
```

print(fatemeh.isnull().sum())