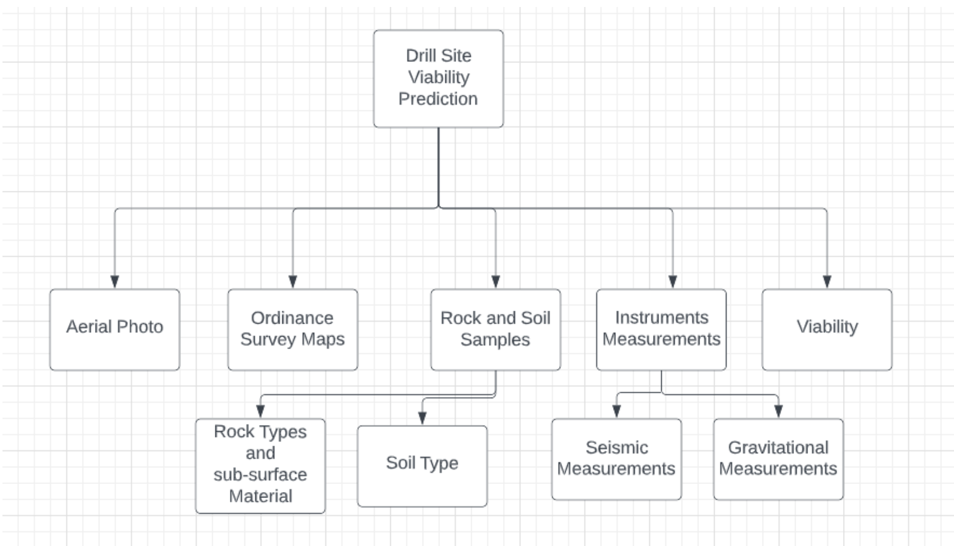
Chapter 2

7. Select one of the predictive analytics models that you proposed in your answer to the previous question about the oil exploration company for exploration of the design of its analytics base table.

a. **What is the prediction subject for the model that will be trained using this ABT?**
   Ans: For the oil exploration prediction model, the prediction subject for the model is drilling site. We are accessing the likelihood that an exploratory drill performed at a drilling site will be usable or not.

b. **Describe the domain concepts for this ABT.**
   Ans: The key domain concepts for this ABT are:
   - Rock and Soil Samples: The aim of this project is to take samples that represent sub surface conditions for entire site. By doing analysis of soil and rock samples, we can find out a large number of soil and rock features
   - Measuring Instruments : Measurements obtained by using specialised instruments (such as Gravitational and Seismic)should be included in the ABT.
   - Viability : It is crucial not to forget the target variable. The can be obtained by accessing the likelihood that a potential drilling site will be viable or not.
   - Aerial Photo: Aerial photo is one of the most important method as this involves taking hundreds of from above most probably using drone or helicopter . These photographs help us to locate potential drilling sites.
   - Ordnance survey Map: These types of survey maps contain detailed information regarding physical locations such as mountains , rivers etc. which are presented using symbols. They help us to visualise potential sites along with near by infrastructure .

c. **Draw a domain concept diagram for the ABT.**
   Ans:



d. Are there likely to be any legal issues associated with the domain concepts you have included?

Ans: There will be no legal complications for the domain concepts described above.

Chapter 3

5. The table below shows the scores achieved by a group of students on an exam.

ID	1	2	3	4	5	6	7	8	9	10
SCORE	42	47	59	27	84	49	72	43	73	59
ID	11	12	13	14	15	16	17	18	19	20
SCORE	58	82	50	79	89	75	70	59	67	35

Using this data, perform the following tasks on the SCORE feature:

a. **A range normalization that generates data in the range (0, 1)**

Ans: Formula for range normalisation (0,1) is :

$A_{i=}$ $(A_i - \min(A)/ \max(A)- \min(A)\ ) * (high - low) + low$

ID	Normalised Score	ID	Normalised Score
1	0.24	11	0.5
2	0.32	12	0.89
3	0.52	13	0.37
4	0	14	0.84
5	0.92	15	1
6	0.35	16	0.77
7	0.73	17	0.69
8	0.26	18	0.52
9	0.74	19	0.65
10	0.52	20	0.13

```python
import numpy as np
from sklearn.preprocessing import MinMaxScaler

# create an array
values = np.array([42,47,59,27,84,49,72,43,73,59,58,82,50,79,89,75,70,59,67,35]).reshape(-1,1)

range_scaler = MinMaxScaler()
range_normalized_values = range_scaler.fit_transform(values)


print(np.round(range_normalized_values,2))
```

```
[[0.24]
 [0.32]
 [0.52]
 [0.  ]
 [0.92]
 [0.35]
 [0.73]
 [0.26]
 [0.74]
 [0.52]
 [0.5 ]
 [0.89]
 [0.37]
 [0.84]
 [1.  ]
 [0.77]
 [0.69]
 [0.52]
 [0.65]
 [0.13]]
```

b. A range normalization that generates data in the range (−1, 1)

Ans : Formula for range normalisation (-1,1) is :

$$A_{i=} (A_i – min(A)/ max(A)- min(A)) * (high – low) + low$$

Where high =1 and low = -1

ID	Normalised Score	ID	Normalised Score
1	−0.52	11	0.0
2	-0.35	12	0.77
3	0.03	13	-0.26
4	-1.0	14	0.68
5	0.84	15	1.0
6	-0.29	16	0.55
7	0.45	17	0.39
8	-0.48	18	0.03
9	0.48	19	0.29
10	0.03	20	-0.74

```python
values = [42,47,59,27,84,49,72,43,73,59,58,82,50,79,89,75,70,59,67,35]

mx= max(values)
mn = min(values)

high= 1
low = -1

normalised_val= []

for i in values:
    sol= (i-mn)/(mx-mn)
    sol= sol*(high-low)
    sol= sol +low
    sol= round(sol,2)
    normalised_val.append(sol)

print(normalised_val)
```

```
[-0.52, -0.35, 0.03, -1.0, 0.84, -0.29, 0.45, -0.48, 0.48, 0.03, 0.0, 0.77, -0.26, 0.68, 1.0, 0.55, 0.39, 0.03, 0.29,
-0.74]
```

c. A standardization of the data

Ans: Formula for standardisation is :

$$A_{i} = (A_i - A\_bar)/std(A)$$

ID	Scaled Score	ID	Scaled Score
1	-1.098428	11	-0.170995
2	-0.808605	12	1.220154
3	-0.113031	13	-0.634712
4	-1.967896	14	1.046260
5	1.336083	15	1.625905
6	-0.692676	16	0.814402
7	0.640508	17	0.524579
8	-1.040464	18	-0.113031
9	0.698473	19	0.350685
10	-0.113031	20	-1.504180

c. A standardization of the data

```
import numpy as np

# create an array
values = np.array([42,47,59,27,84,49,72,43,73,59,58,82,50,79,89,75,70,59,67,35]).reshape(-1,1)

scaled_df= pd.DataFrame(values, columns=["val"])

mean = scaled_df['val'].mean()
stdev = scaled_df['val'].std()
scaled_df=scaled_df.assign(norm_score= lambda x: (x['val'] - mean)/stdev)
norm_df=scaled_df[['norm_score']].copy(deep=True)
print(norm_df)
```

```
    norm_score
0    -1.098428
1    -0.808605
2    -0.113031
3    -1.967896
4     1.336083
5    -0.692676
6     0.640508
7    -1.040464
8     0.698473
9    -0.113031
10   -0.170995
11    1.220154
12   -0.634712
13    1.046260
14    1.625905
15    0.814402
16    0.524579
17   -0.113031
18    0.350685
19   -1.504180
```

6. The following table shows the IQs for a group of people who applied to take part in a television general knowledge quiz.

ID	1	2	3	4	5	6	7	8	9	10
IQ	92	107	83	101	107	92	99	119	93	106

ID	11	12	13	14	15	16	17	18	19	20
IQ	105	88	106	90	97	118	120	72	100	104

Using this dataset, generate the following binned versions of the IQ feature:

a. **An equal-width binning using 5 bins.**

Ans : To perform an equal width binning we calculate the bin size i.e. range/ no of bins.

Range = max - min

Max= 120

Min = 72

No. of bins = 5

So, bin size= (120-72) / 5 = 9.6

ID	Bins	ID	Bins
1	Bin 3	11	Bin 4
2	Bin 4	12	Bin 2
3	Bin 2	13	Bin 4
4	Bin 4	14	Bin 2
5	Bin 4	15	Bin 3
6	Bin 3	16	Bin 5
7	Bin 3	17	Bin 5
8	Bin 5	18	Bin 1
9	Bin 3	19	Bin 3
10	Bin 4	20	Bin 4

```
values= [92,107,83,101,107,92,99,119,93,106,105,88,106,90,97,118,120,72,100,104]
```

```
mx= max(values)
mn= min(values)
bin_number=5
bin_size= (mx-mn) / bin_number


```

```
import pandas as pd

bins= {"bin1":[72.0,81.6], "bin2": [81.6,91.2], "bin3":[91.2,100.8], "bin4":[100.8,110.4], "bin5":[110.4,120.0]}
```

```
df_bins= pd.DataFrame(bins)
df_bins= df_bins.T
df_bins
```

	0	1
bin1	72.0	81.6
bin2	81.6	91.2
bin3	91.2	100.8
bin4	100.8	110.4
bin5	110.4	120.0

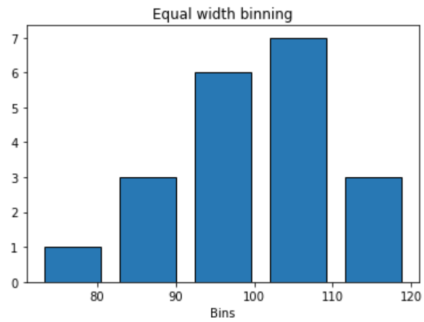
```
df= pd.DataFrame(values, columns= ["val"])
def f(x):
    if x>= 72.0 and x< 81.6:
        return "bin1"
    elif x>= 81.6 and x< 91.2:
        return "bin2"
    elif x>= 91.2 and x< 100.8:
        return "bin3"
    elif x>= 100.8 and x< 110.4:
        return "bin4"
    else:
        return "bin5"
df["bins"]= df.val.apply(f)
```

```
df['bin_cut'] = pd.cut(df.val, bins=5,labels=["bin1","bin2","bin3","bin4","bin5"])
df
```

	val	bins	bin_cut
0	92	bin3	bin3
1	107	bin4	bin4
2	83	bin2	bin2
3	101	bin4	bin4
4	107	bin4	bin4
5	92	bin3	bin3
6	99	bin3	bin3
7	119	bin5	bin5
8	93	bin3	bin3
9	106	bin4	bin4
10	105	bin4	bin4
11	88	bin2	bin2
12	106	bin4	bin4
13	90	bin2	bin2
14	97	bin3	bin3
15	118	bin5	bin5
16	120	bin5	bin5
17	72	bin1	bin1
18	100	bin3	bin3
19	104	bin4	bin4

```
seabrn = sns.distplot(values,bins=5, kde=False, hist_kws={"rwidth":0.75,'edgecolor':'black', 'alpha':1.0})
seabrn.set(xlabel = "Bins" , title ='Equal width binning')
```

```
[Text(0.5, 0, 'Bins'), Text(0.5, 1.0, 'Equal width binning')]
```



b. An equal-frequency binning using 5 bins

Ans : bins = 5

Number of data points in each bin = 20/ 5 = 4

Then we will sort the data and assign bins to each values i.e. bins1 to first four values, then bins2 to next four values and so on.

The sorted values are :

72,83,88,90,92,92,93,97,99,100,101,104,105,106,106,107,107,118,119,120.

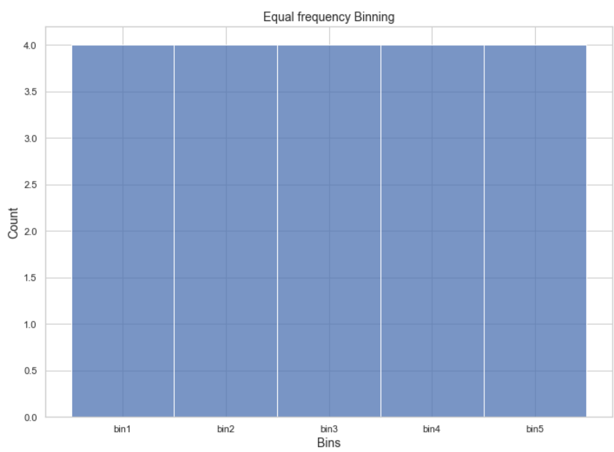
Values	Bins	Values	Bins
72	Bin 1	101	Bin 3
83	Bin 1	104	Bin 3

88	Bin 1	105	Bin 4
90	Bin 1	106	Bin 4
92	Bin 2	106	Bin 4
92	Bin 2	107	Bin 4
93	Bin 2	107	Bin 5
97	Bin 2	118	Bin 5
99	Bin 3	119	Bin 5
100	Bin 3	120	Bin 5

```
df_freq= pd.DataFrame(values, columns= ["val"])
df_freq["qcut"]= pd.qcut(df.val, q=5,labels=['bin1','bin2','bin3','bin4','bin5'])
df_freq= df_freq.sort_values("val")
df_freq
```

0	72	Bin1
1	83	Bin1
2	88	Bin1
3	90	Bin1
4	92	Bin2
5	92	Bin2
6	93	Bin2
7	97	Bin2
8	99	Bin3
9	100	Bin3
10	101	Bin3
11	104	Bin3
12	105	Bin4
13	106	Bin4
14	106	Bin4
15	107	Bin4
16	107	Bin5
17	118	Bin5
18	119	Bin5
19	120	Bin5

```
Text(0, 0.5, 'Count')
```



7. Comment on the distributions of the features shown in each of the following histograms.

a. **The height of employees in a truck driving company**.
   Ans:
   The distribution of height of the employees in a truck driving company follows a normal distribution. Mean is around 175 . As per the empirical rule in normal distribution, around 68% of the values are in range (mean + standard deviation, mean - standard deviation) and as per the graph 68% of the values are in the range (150,200) so
   standard deviation is close to 25.
   This type of distribution is well suited for analytics as there are a few outliers in the dataset.

b. **The number of prior criminal convictions held by people given prison sentences In a city district over the course of a full year.**
   Ans: As per the graph, the distribution of convictions held by people over the course of full year follows an exponential distribution. The central tendency is around 0 and the values are exponentially decreasing. This type of data has many outliers values which decreases the machine learning model efficiency such as in the graph one of the outlier values is 40 convictions which is very odd.

c. **The LDL cholesterol values for a large group of patients, including smokers and non-smokers.**
   Ans: This type of distribution is multimodal (it is distribution in which there is more than one mode). There are two distinct groups ie smoker and non-smoker. The first one has mean around 100 and the other one has mean around 160. The first group is larger than the other group. From the graph we can infer that first group might be smokers with high cholesterol values whereas the smaller group might be non-smokers.

d. **The employee ID numbers of the academic staff at a university.**
   Ans: This employeeID in the academic staff is uniformly distributed. The probability is constant since each ID has equal chances of being the outcome.

e. **The salaries of motor insurance policy holders.**
   Ans: This salaries of motor insurance policy holders are right skewed. In this type of distribution the mean is greater than median and mode. From the graph, we can infer that mean will be close to 34000 but there will be many outlier values too. It is crucial to deal with these outlier before implementing ML models. One ideal way is that we can define the upper and lower threshold values using IQR and then can remove the outliers.

9. **Tachycardia is a condition that causes the heart to beat faster than normal at rest. The occurrence of tachycardia can have serious implications including increased risk of stroke or sudden cardiac arrest. An analytics consultant has been hired by a major hospital to build a predictive model that predicts the likelihood that a patient at a heart disease clinic will suffer from tachycardia in the month following a visit to the clinic.**

The hospital will use this model to make predictions for each patient when they visit the clinic and offer increased monitoring for those deemed to be at risk. The analytics consultant has generated an ABT to be used to train this model.[17] The descriptive features in this dataset are defined as follows:

AGE: The patient's age

GENDER: The patient's gender (male or female)

WEIGHT: The patient's weight

HEIGHT: The patient's height

BMI: The patient's body mass index (BMI) which is calculated as where weight is measured in kilograms and height in meters.

SYS. B.P.: The patient's systolic blood pressure

DIA. B.P.: The patient's diastolic blood pressure

HEART RATE: The patient's heart rate

H.R. DIFF.: The difference between the patient's heart rate at this visit and at their last visit to the clinic

PREV. TACHY.: Has the patient suffered from tachycardia before?

TACHYCARDIA: Is the patient at high risk of suffering from tachycardia in the next month?

The following table contains an extract from this ABT—the full ABT contains 2,440 instances.

ID	AGE	GENDER	WEIGHT	HEIGHT	BMI	SYS. B.P.	DIA. B.P.	HEART RATE	H.R. DIFF.	PREV. TACHY.	TACHYCARDIA
1	6	male	78	165	28.65	161	97	143			true
2	5	m	117	171	40.01	216	143	162	17	true	true
		⋮			⋮			⋮			
143	5	male	108	1.88	305,568.13	139	99	84	21	false	true
144	4	male	107	183	31.95	1,144	90	94	-8	false	true
		⋮			⋮			⋮			
1,158	6	female	92	1.71	314,626.72	111	75	75	-5		false
1,159	3	female	151	1.59	596,495.39	124	91	115	23	true	true
		⋮			⋮			⋮			
1,702	3	male	86	193	23.09	138	81	83		false	false
1,703	6	f	73	166	26.49	134	86	84	-4		false
		⋮			⋮			⋮			

Feature	Count	% Miss.	Card.	Mode	Mode Freq.	Mode %	2^{nd} Mode	2^{nd} Mode Freq.	2^{nd} Mode %
GENDER	2,440	0.00	4	male	1,591.00	65.20	female	647.00	26.52
PREV. TACHY.	2,440	44.02	3	false	714.00	52.27	true	652.00	47.73
TACHYCARDIA	2,440	2.01	3	false	1,205.00	50.40	true	1,186.00	49.60

Feature	Count	% Miss.	Card.	Min.	1^{st} Qrt.	Mean	Median	3^{rd} Qrt.	Max.	Std. Dev.
AGE	2,440	0.00	7	1.00	3.00	3.88	4.00	5.00	7.00	1.22
WEIGHT	2,440	0.00	174	0.00	81.00	95.70	95.00	107.00	187.20	20.89
HEIGHT	2,440	0.00	109	1.47	162.00	162.21	171.50	179.00	204.00	41.06
BMI	2,440	0.00	1,385	0.00	27.64	18,523.40	32.02	38.57	596,495.39	77,068.75
SYS .B.P.	2,440	0.00	149	62.00	115.00	127.84	124.00	135.00	1,144.00	29.11
DIA. B.P.	2,440	0.00	109	46.00	77.00	86.34	84.00	92.00	173.60	14.25
HEART RATE	2,440	0.00	119	57.00	91.75	103.28	100.00	110.00	190.40	18.21
H.R. DIFF.	2,440	13.03	78	-50.00	-4.00	3.00	1.00	8.00	47.00	12.38

Discuss this data quality report in terms of the following:
a. Missing values
Ans: From the ABT above, we can conclude that missing values are present in Prev. Tachy.(44.02%) , Tachycardia(2.01%), H.R.Diff (13.03%).
For H.R.Diff which contains 13.03% missing values, should be replace by mean. We can implement SimpleImputer module from Sklearn to replace missing values with mean. In this way, the data quality will improved which shall increase ML model efficiency.

For Tachycardia which contains a few missing values around(2%) and it is a target feature. The missing data points should be removed from the dataset as we cannot implement imputation methods on target values

For Prev. Tachy. Which has 44.02% missing values, this feature should be removed from analysis because if we implement mean imputation to replace missing values we might alter the overall dataset.

b. Irregular cardinality

Ans: Most of the features are numeric and have regular cardinalities such as dia. B.P. However, irregular cardinality is observed in age and gender. Age has a cardinality 7 which is less for a numeric measure. Also, as per the plot we can see that there are 7 distinct values in age and it is normally distributed. Hence, Age feature should be stored as categorical variable with seven ordinal values.
Also, the Gender feature has cardinality 4 which is odd. From the bar plot we can see the gender contains “male”, “female”, “m”, “f” and this is data quality issue. The values should be either stored as (m and f) or (male and female) to maintain consistency in the data.

c. Outliers

Ans: As per the ABT and the graphs, we can see that outliers are present in height, BMI and sys. Blood pressure.

With reference to feature Height, we can see that there is a huge difference between min value and the first quartile and between median and first quartile. Also, in the bar plot we can that the graph is skewed and the outliers are in the range 0-10.From the table above, we can see that in row 143, 1158 and 1159 have values of very low magnitude ie. 1.88,1.71,1.59 respectively. This is rare because as per the ABT table majority of the values are of higher magnitude. So this might be a glitch in data entry. If we can find the correct values then these values should be corrected otherwise they can be multiplied by 100 to come in the desired range.

With reference to BMI, we can see in the ABT that there are very large values present which is odd because the BMI values should range between 15 and 60. Also, there is a huge difference between max and third quartile and between third quartile and median. The max value in the BMI is 596,495.39 on row 1159 and this is totally unreasonable. Also, this value occurs on the same row(1159) where outlier is present for height feature. Also, it can be observed from the table that the large BMI values occur from low values for Height because BMI is calculated from weight and height. So, we can infer that the incorrect values for Height are causing outlier values for BMI as well. These values can be corrected by correcting the height values and the recalculating the BMI.

With reference to Sys. Blood pressure feature, the maximum value is 1144 which is not a correct value as per the blood pressure range. Also, in the table we can see that this value is present in row 144. We can replace this value with mean imputation

d. Feature distributions

Ans: The target feature seems to be evenly distributed which is good for a ML model. Also, most of the continuous descriptive features are normally distributed. However, the graph for H.R. Diff seems to be a bimodal. Also, there are a few outliers present in a few features(height, BMI and sys. Blood pressure) which we discussed above and that can be replaced by imputation methods.