Roll no = 2018201083
Name = Ankush Nagpal

##
###################################Question 1-1#######################
##

-Used n-ary tree with entropy as impurity value.
-For each feature calculated entropy value, the one with max info gain is selected
-A feature once selected cannot occur on a path from root to leaf more than once.
-Used categorical features namely = 'Work_accident','promotion_last_5years', 'sales', 'salary'

******************RESULTS************************
count= 1720
total= 2248
accuracy= 0.7651245551601423
tp fp fn= [0, 0, 528]
recall= 0.0
precision= 0.0
f1 score= 0

###
################################   END   ######################
###

###
################################Question 1-2#####################
###
-Used binary tree for numerical features.For each unique value of each feature, calculated entropy and info gain value
the one with max info gain is the best feature and best value to split
-Each feature can occur multiple times on a path until all the unique values of feature are not exhausted.

******************RESULTS************************
count= 2200
total= 2248
accuracy= 0.9786476868327402
tp fp fn= [509, 21, 27]
recall= 0.9496268656716418
precision= 0.960377358490566
f1 score= 0.954971857410882
sklearn precision= 0.960377358490566

###
################################   END   ######################
###

###
################################Question 1-3#####################
###
Question)Contrast the effectiveness of Misclassification rate, Gini, Entropy as impurity measures in terms of precision, recall and accuracy

(i)Entropy

```
******************RESULTS************************
count=  2200
total=  2248
accuracy=  0.9786476868327402
tp fp fn=  [509, 21, 27]
recall=  0.9496268656716418
precision=  0.960377358490566
f1 score=  0.954971857410882
sklearn precision=  0.960377358490566
```

(ii)Gini

```
******************RESULTS************************
count=  2187
total=  2248
accuracy=  0.972864768683274
tp fp fn=  [509, 34, 27]
recall=  0.9496268656716418
precision=  0.9373848987108656
f1 score=  0.9434661723818349
sklearn precision=  0.9373848987108656
```

(iii)Missclassification rate

```
******************RESULTS************************
count=  2175
total=  2248
accuracy=  0.9675266903914591
tp fp fn=  [516, 53, 20]
recall=  0.9626865671641791
precision=  0.9068541300527241
f1 score=  0.9339366515837105
sklearn precision=  0.9068541300527241
```

```
#################################################################
#################################    END    ###################
#################################################################
```

```
#################################################################
###################################Question 1-4##################
#################################################################
```

After calculating impurity of all the features , the most important features came out to be
satisfaction_level and number_project

So plotting graph between these two features with different colors for label ==1 and label==0

```
#################################################################
#################################    END    ###################
#################################################################
```

```
#################################################################
###################################Question 1-5##################
#################################################################
```

For each depth value is predicted by max(positive_samples,negative_samples)
error value is calculated for each level

##
################################   END   ###################
##

##
################################Question 1-6###################
##
For handling missing values , maintained number of positive and negative samples
at each node at every level.
suppose value = [a1,b1,c1,d1] are values and d1 is missing
We will give the output as max(positive_samples, negative_samples) at c1 level.

##
################################   END   ###################
##