the me...

⇒ Feature Engineering:

⇒ Handelling imbalanced datasets

Dataset

Classification (Means output is in the form of categories)

⇓

Binary classification (Means 2 categories)

Let I have dataset of 1000 datapoints

Outputs: Yes, No

900 Yes

100 No

Means

900 : 100 = 9 : 1

So, this is an imbalanced dataset which makes training

model biased. So we have to balance the datapoints making both almost equal.

- Schemes used for balancing dataset:

① Down sampling: Decrease the datapoints of majority

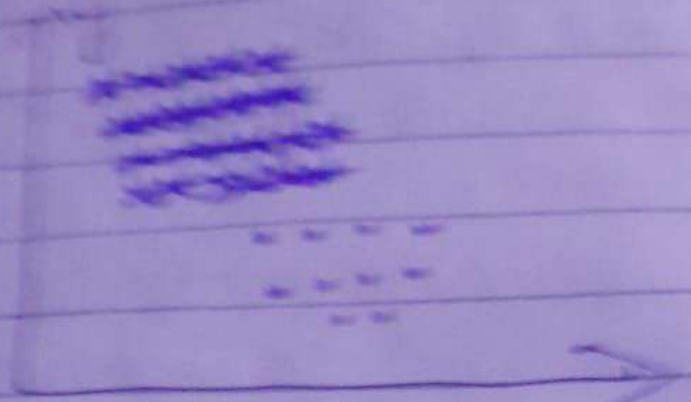
② Up sampling: Increase the datapoints of minority

Further code in repo

⇒ SMOTE (Synthetic Minority Oversampling Technique)

SMOTE is a technique used in Machine Learning to address imbalanced datasets where the minority class has significantly fewer instances than majority class. SMOTE involves generating synthetic instances of minority class by interpolating between existing instances.

Eg



Here → represents majority
and • represents minority

⇒ SMOTE picks the two nearest datapoints of minority and add more datapoints to them then pick next 2 closest where as upsampling just add the datapoints present at random

⇒ SMOTE is better than upsampling

⇒ **Data Encoding:**

When we have categorical feature like Degrees (BS, PHD, Masters) then we understand the meaning of them but ML algorithm or model don't understand it only understands numerical values.

⇒ So, process of converting these categorical information to numerical values is called Data Encoding.

- **Types of Data Encoding:**
  1. Nominal/OHE (One Hot) Encoding
  2. Label and ordinal Encoding
  3. Target Guided ordinal Encoding

1. Nominal/OHE Encoding:

One hot encoding also known as nominal encoding is a technique used to represent categorical data as numerical data which is more suitable for machine learning algorithms. In this technique each category is represented as binary vector where each bit corresponds to a unique category

For example if we have a categorical variable "color" with 3 possible values (red, green, blue) we can represent it using one hot encoding as follows:

Red: [1, 0, 0]
Green: [0, 1, 0]
Blue: [0, 0, 1]

As columns in dataset.

- **Limitations:**

1. Not suitable for data with large number of categorical variables or variable with various categories as there will be lot of features.
2. Sparse matrix (mesh of 0's and 1's) which lead to overfitting.

⇒ Label Encoding and ordinal Encoding:

Label Encoding and ordinal Encoding are two encoding techniques used to encode categorical data as numerical

- **Label Encoding:**
Label encoding involves assigning a unique numerical label to each category in the variable. The labels are usually assigned in alphabetical order or based on frequency of categories. For example, if we have a categorical variable color with 3 possible values ("red", "green", "blue") we can represent it using label encoding as

1- Red: 1
2- Green: 2
3- Blue: 3

- **Limitation of Label Encoder:**
As in label encoding, unique numerical values assigned like 3 to blue, 2 to green the model may consider that blue is greater than green because of higher numerical value. But in this case this problem not occur because we are not assigning ranks

⇒ But in case we have to assign ranks we use

ordinal Encoding:

It is used to encode categorical data that may have an intrinsic order or ranking. In this technique each category is assigned a numerical value based on its position in the order.

For example, if we have a categorical variable "education" with 4 possible values (highschool, college, graduate, post graduate) we can represent it using ordinal encoding.

1. HighSchool: 1
2. College: 2
3. Graduate: 3
4. Post-Graduate: 4

⇒ Target Guided Ordinal Encoding:

It is a technique used to encode categorical values based on their relationship with the target variable. This encoding technique is useful when we have a categorical variable with large number of unique categories, and we want to use this variable as a feature in our Machine Learning model.

In target guided ordinal encoding, we replace each category in categorical variable with a numerical value, based on the mean or median of the target variable for that category. This creates a monotonic relationship b/w categorical variable and target variable, which can improve the predictive power of our model.