**Celcius Income**

**Problem:** The prediction task is to determine whether a person makes over $50K a year*.*

**Solution:**

1. Task is to predict whether a given person is getting a salary more than $50K.
2. We are given the below information to analyse and build machine learning algorithms.

'Age', 'Workclass', 'Fnlwgt', 'Education\_num', 'Marital\_status',   
'Occupation', 'Relationship', 'Race', 'Sex', 'Capital\_gain',   
'Capital\_loss', 'Hours\_per\_week', 'Native\_country', 'Income'

1. We can see that we can drop Education information as there is duplicate column for the same in number representation. Educaiton\_num contains the same information that we have in Education.
2. From workclass variable, we can see that most of the data is from private sector followed by self employment and local government employees
3. Coming to education most of them have an education of 9
4. Most of the data is of white people
5. Most of them have capital gain zero and capital loss as zero
6. Majority of the information is of United States so our algorithm might tilt towards their nation
7. We have 24719 less than 50K while we have only 7841 data points for >50K
8. All the null values have been replaced with ?

**Data Cleaning**

* 1. We have a lot of unknown data. We need to replace ? with null values so that we can identify actual null values
  2. We have lot of null values in Occupation.
  3. We have two options. One is filling up the null values with some know methods like KNN imputer, min or max or mean etc., But we must be careful with that because that might have a lot of impact on our Machine learning algorithm.
  4. Since occupation is a main deciding factor for how much do they earn, we should ideally drop the data with null values in the occupation field.
  5. After dropping the null values from Occupation, we can see that other null values automatically got handled, Null values in other columns are also cleared by dropping. Other null values are present in the same columns as same as the null values in Occupation.

**EDA**

* + 1. Age starts from 16 years and is similar to normal distribution, most of the people are from middle age. between 20 and 50 years, So we have very less people whose age is above 50. Till mid or end thirty’s, we have increasing count of people while for less than 30 and greater than 40 we have decreasing population.
    2. Private secret have more employees than any other employees that might cause our model tilt towards private more, we can even that out by oversampling other sectors. All other classes have similar range of values. So we are good with that values. Only private sector is highly populated.
    3. Final weights is uneven but can be converted to normal distribution. We have most of the data is around 0.2 and after that we can see that there is sharp decline after that. Before that there is sharp increases and sharp decreases in the spikes.
    4. Most of them are educated. Also we can see that most of the education is of value nine and ten. Other than that we have good amount of people completed thirteen followed by fourteen and other values in less densities
    5. Other columns like married/ occupation, etc., are categorical
    6. Hours per week is discrete distribution
    7. Most of them are US citizens which means we are more tilted towards their income.
    8. Here we can see that there is no direct correlation between income and other variables.
    9. We need to transform the labels using label encoder
    10. Once we encode labels, we can use standard scalar to encode the data and get it into single format. It will help us do analysis much better.
    11. We need to check outlier influence and skewness

**MULTIVARIATE** **EDA**

* + - 1. Age – age variable have a positive correlation with our target variable, that is 0.24 percentage.
      2. Work class have a minimal positive correlation on income variable. There will be less effect on our model but will have.
      3. Final weights have very little negative correlation with our target variable.
      4. Education number have good correlation with Income. It have a positive correlation on our target variable which is income. We have 0.33 correlation which is highest of all. So this will have significant effect on Income variable.
      5. Marital status have negative correlation with our target variable. The value is -0.19
      6. Occupation have 0.05 positive correlation with our target variable.
      7. Relationship have negative correlation with a value of 0.25 which is a significant number. This number is one of the highest negative correlations.
      8. Race have positive correlation with target. The value is 0.071\
      9. Sex and capital gain have a good positive correlation that is 0.22 which is good number. With that we have good effect on our target
      10. Capital loss have positive correlation with target variable. 0.15
      11. Hours per week have positive correlation with the income. The more hours they work, the more income they will get. The value is 0.23 which have significant impact on our model.
      12. Native country have positive correlation on our model. The value is 0.017.

**Handling skewness and outliers**

* + 1. After checking variance inflation factor, we can see that there is no much outlier influence in our data
    2. We can see that there is a lot of skewness in Capital gain, capital loss, followed by native country, race work class, final weights
    3. We need to reduce skewnessfor which we can use square root transformation or yeo Johnson transformer

The Johnson Transformation relies on the fact that any function of a normal random variable is also normal. Suppose we have a random variable X that is not normally distributed. We can create a new random variable Y=g(X) where g() is a function that transforms X into a new random variable Y that is normally distributed. If we can find such a function g(), then we can use Y in place of X in any statistical analysis that assumes normality.

After applying johnson transformation, we can see that there is significant reduction in skewness.

**Training and model validation:**

* + - 1. I’ve defined a grid search cv model to automate the training process. I’ve used Random forest, gradient descent, svc, knn and decision tree models to evaluate which models give the best performance of all.
      2. Upon training we got the best accuracy to be 0.8622 when we use Decision tree classifer with a maximum depth of 150 and maximum leaf nodes of 100. We have finalized the model and used it for validation.