Lab 1

1. Load the training and testing dataset provided with this lab and join the Kaggle competition. (1 point) https://www.kaggle.com/t/cdd98e14cf1c471ea95e47da59afdd06

   Data Description: The dataset contains information about people's jobs such as their demographic information etc. and the target variable is to predict whether the person is looking for a job change or not, the dataset is imbalanced. There are 14,368 rows in the training set and 4790 rows in the testing set. The columns in the dataset are as follows:

   - index: Unique ID for candidate
   - city: City code
   - city_development_index: Development index of the city (scaled)
   - gender: Gender of candidate
   - relevent_experience: Relevant experience of candidate
   - enrolled_university: Type of University course enrolled if any
   - education_level: Education level of candidate
   - major_discipline : Education major discipline of candidate
   - experience: Candidate total experience in years
   - company_size: No of employees in current employer's company
   - company_type : Type of current employer
   - last*new*job: Difference in years between previous job and current job
   - training_hours: training hours completed
   - target: 0 – Not looking for job change, 1 – Looking for a job change

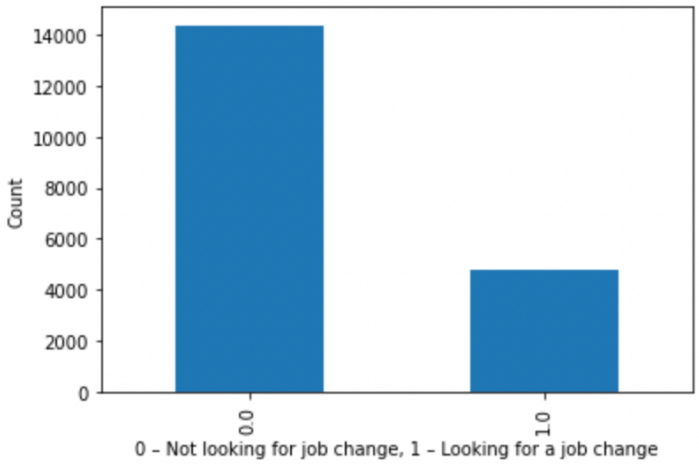
   Sample code for plotting distribution of target in the entire dataset:

```
In [5]: import pandas as pd
        import matplotlib.pyplot as plt

In [6]: data = pd.read_csv("aug_train.csv")

In [ ]: data

In [7]: data.target.value_counts().plot.bar()
        plt.xlabel("0 – Not looking for job change, 1 – Looking for a job change")
        plt.ylabel("Count")
```



2. Explain in your own words what is NOIR classification of data. (2 points)

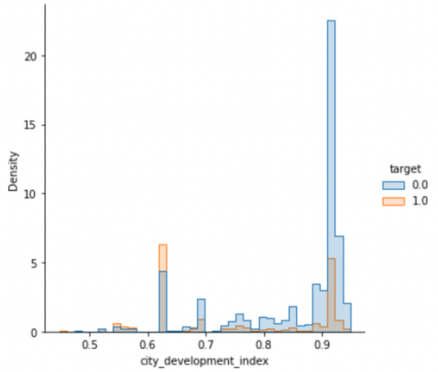
3. Classify the given dataset features into NOIR categories. (2 points)

4. Summarize the dataset: (10 points)
    a. Number of columns and rows present (1 point)
    b. Min, max, avg, std dev etc. stats for continuous features (hint: use pandas describe function) (1 point)
    c. Number of unique values for categorical features (1 point)
    d. Number of nulls and Nans in each column (1 point)
    e. Visualize distribution of each feature using graphs (6 points)

5. Visualize the relationship of each feature with target variable (hint: create density plots for continuous features and cross tables for categorical features) (6 points)

    Sample plot with code:

```
In [6]: sns.displot(data, x = "city_development_index", hue = "target", stat = "density", element = "step")
Out[6]: <seaborn.axisgrid.FacetGrid at 0x7fa1eaf27460>
```



6. Handle missing values: Use any three methods to handle missing values. (6 points)

   Sample code and output for checking missing values in each column

```
In [9]: data.isnull().sum()

Out[9]: enrollee_id                0
        city                       0
        city_development_index     0
        gender                  4508
        relevent_experience        0
        enrolled_university      386
        education_level          460
        major_discipline        2813
        experience                65
        company_size            5938
        company_type            6140
        last_new_job             423
        training_hours             0
        target                     0
        dtype: int64
```

7. Create new features using combinations / transformations of existing features (Optional) (0 points)

8. Scale the features for models that require scaling and perform required pre-processing (such as one hot encoding etc.) (5 points)

9. Explain the following terms in detail (9 points)
   a. Pearson's correlation (3 points)
   b. T – test (3 points)
   c. Chi squared test (3 points)

10. Perform Pearson correlation between continuous features and plot the heatmap of the correlation matrix. (3 points)

11. Perform t test on continuous features after dividing them using target variable. (3 points)

12. Perform Chi squared test among categorical variables and with the target variable (3 points)

13. Compare the features selected using visualization from question 5 and using the statistical tests from the previous three questions. (5 points)

14. Explain the following terms: (6 points)
    a. Forward Selection (2 points)
    b. Backward Elimination (2 points)
    c. Recursive feature elimination (2 points)

15. Perform the above methods of feature reduction if you want. (Optional)

16. Train at least 5 different classification models on the final data (Perform hyper parameter tuning using a validation set for each of the model). (10 points)

17. Write any one classification algorithm without using scikit-learn and train on the final data. (10 points)

18. Submit the test predictions using your model from question 14 to the Kaggle competition. (**Note**: Actual ranking would be based on a private leader board data (40% of the test data), the live ranking is given only using the public leader board data (60% of the test data)) (9 points)

    a. 95 – 100 percentiles: 9 points
    b. 90 – 94.99 percentile: 8.5 points
    c. 85 – 89.99 percentile: 8 points
    d. 80 – 84.99 percentile: 7.5 points
    e. 75 – 79.99 percentile: 7 points
    f. 70 – 74.99 percentile: 6.5 points
    g. 65 – 69.99 percentile: 6 points
    h. 60 – 64.99 percentile: 5.5 points
    i. 55 – 59.99 percentile: 5 points
    j. 50 – 54.99 percentile: 4.5 points
    k. 45 – 49.99 percentile: 4 points
    l. 40 – 44.99 percentile: 3.5 points
    m. 35 – 39.99 percentile: 3 points
    n. 30 – 34.99 percentile: 2.5 points
    o. 25 – 29.99 percentiles: 2 points
    p. 20 – 24.99 percentile: 1.5 points
    q. 15 – 19.99 percentile: 1 point
    r. 10 – 14.99 percentile: 0.5 points
    s. 0 – 9.99 percentile: 0 points

19. Create a detailed report of the lab. (10 points)

20. Give your comments on how you found the lab tasks. (0 points)