**Workshop**

**1) Explain each features**

|  |  |  |
| --- | --- | --- |
| Features | Data Type | Description |
| age | Numeric | Describes the age of individuals. |
| workclass | Nominal | Private, Self-emp-not-inc, Self-emp-inc. |
| fnlwgt | Numeric | The weights on the Current Population Survey (CPS) files are controlled to independent estimates of the civilian noninstitutional population of the US. |
| education | Ordinal | The education of individuals. |
| education-num | Numeric | Number of years spent in education. |
| marital-status | Nominal | Relationship in their family. |
| occupation | Nominal | Occupation of individuals. |
| relationship | Nominal | Relationship in their family. |
| race | Nominal | Race of individuals. |
| sex | Nominal | Gender of individuals. |
| Capital-gain | Numeric | The increase in the value of an asset or investment. |
| Capital-loss | Numeric | The reduction in the value of the capital assets or investment. |
| hours-per-week | Numeric | Work hour spent in a week. |
| native-country | Nominal | Place of birth. |

**2) What kind of DS task, and what are we going to classify and why?**

Classification, We are going to classify people in 2 classes which have income more than $50K/year and have income less than $50K/year because we need to know what features that people who have income more than/ less than $50K/year have. To make the prediction model.

**3) Which feature(s) are not important based on your domain knowledge?**

* Fnlwgt
* Marital-status
* Relationship
* Race
* Native-country

**4) How many miss values, which attribute, where are located?**

|  |  |
| --- | --- |
| Attribute | Miss values |
| Workclass | 1836 |
| Occupation | 1843 |
| Nativecountry | 583 |

**5) How many extreme values/outliers?**

|  |  |  |
| --- | --- | --- |
| Attribute | Extreme value | Outlier |
| fnlwgt | 16 | 992 |
| Capital-gain | 2712 | 2712 |
| Capital-loss | 1519 | 1519 |
| hours-per-week | 816 | 9008 |
| age | 0 | 143 |