# Topic: Support Vector Machine (SVM)

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name:** Nithin Dsouza

**Batch Id:**05012021-10AM

**Topic: Support Vector Machines.**

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.** ****

**Using R and Python codes perform:**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

**3.2 Outlier Imputation**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary**
   2. **Univariate analysis**
   3. **Bivariate analysis**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options)**
   2. **Perform Support Vector Machines.**
   3. **Train and Test the data and compare accuracies by Confusion Matrix and use different Hyper Parameters**
   4. **Briefly explain the model output in the documentation**

****

1. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided**

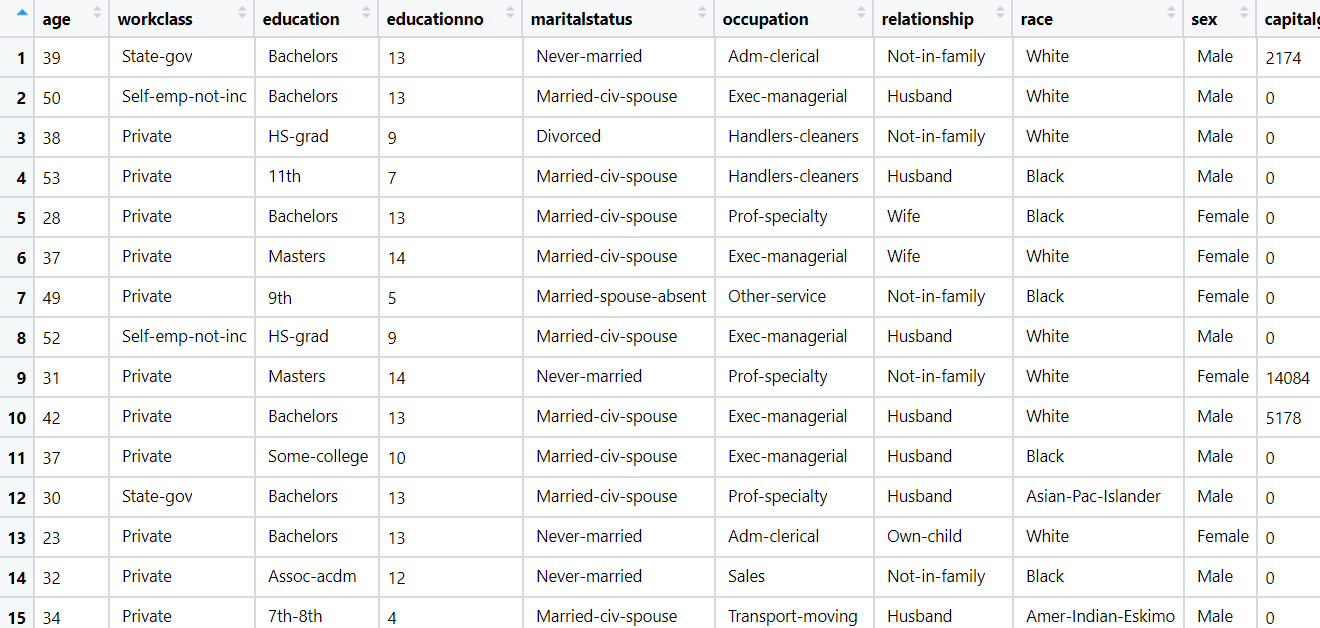
# Note:

**The assignment should be submitted in the following format:**

* **R code**
* **Python code**
* **Code Modularization should be maintained**
* **Documentation of the model building (elaborating on steps mentioned above)**

**Problem Statement: -**

A construction firm wants to develop a suburban locality with new infrastructure but they are faced with a challenge of incurring losses if they cannot sell the properties. To overcome this, they consult an analytics firm and would like to get insights on how densely the area is populated and different level of income group people reside. You as a Data Scientist perform Support Vector Machines Algorithm on the given dataset and bring out informative insights and also comment on if its viable for investment in that area.



**Objective:** Maximize the accuracy in predicting using svm which identify optimal separating hyperplane.

**Constraint:** Minimize error in choosing a good kernel.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of the Feature** | **Type** | **Description** | **Relevance** |
| age | Discrete | Age of a person | Relevant, Provides useful information. |
| workclass | Qualitative, Ordinal | Work class of person | Relevant, Provides useful information. |
| education | Qualitative, Ordinal | Education of a person | Relevant, Provides useful information. |
| educationno | Nominal | Number given to education | Irrelevant, does not Provides useful information. |
| maritalstatus | Qualitative, Ordinal | Marital status of a person | Relevant, Provides useful information. |
| occupation | Qualitative, Ordinal | occupation of a person | Relevant, Provides useful information. |
| relationship | Qualitative, Ordinal | relationship of a person | Relevant, Provides useful information. |
| race | Qualitative, Ordinal | race of a person | Relevant, Provides useful information. |
| sex | Qualitative, Ordinal | Gender of a person | Relevant, Provides useful information. |
| capitalgain | Quantitative, Ratio | Profit of a person | Relevant, Provides useful information. |
| capitalloss | Quantitative, Ratio | Loss of a person | Relevant, Provides useful information. |
| hoursperweek | Quantitative, discrete | Hours per week worked | Relevant, Provides useful information. |
| native | Nominal | Native of a person | Irrelevant, does not Provides useful information. |
| Salary | Quantitative, discrete | Salary of the person | Relevant, Provides useful information. |

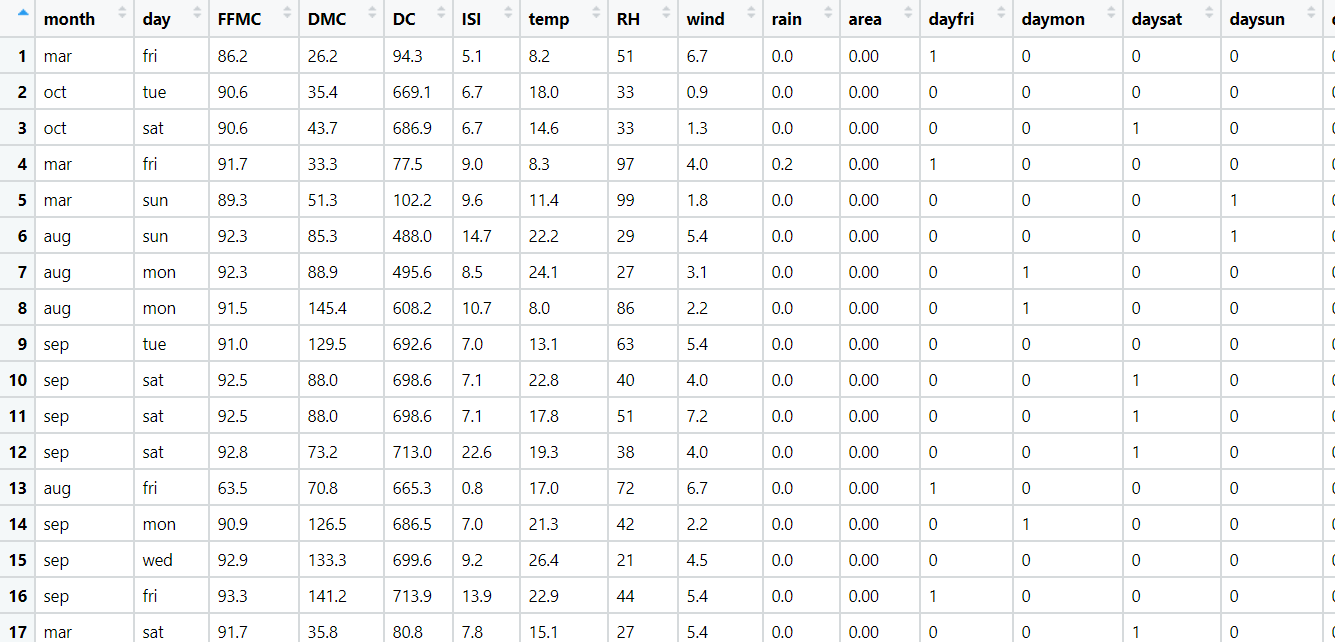
**Problem Statement: -**

In California, annual forest fires can cause huge loss of wild life, human life and property damage can skyrocket in billions. Local officials would like to predict the size burned area in forest fires annually so that they can be better prepared in future calamities.

Build a Support Vector Machines algorithm on the dataset and share your insights on it in the documentation.

Note: - Size\_ category is the output variable.

**Objective:** Maximize the accuracy in predicting using svm which identify optimal separating hyperplane.

**Constraint:** Minimize error in choosing a good kernel.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of the Feature** | **Type** | **Description** | **Relevance** |
| month | Ordinal | Month in which fire occurs | Relevant, Provides useful information. |
| day | Ordinal | Day of the week fire occurs | Relevant, Provides useful information. |
| FFMC | ratio | Fine fuel moisture code | Relevant, Provides useful information. |
| DMC | ratio | Duff moisture code | Relevant, Provides useful information. |
| Dc | ratio | Drought code | Relevant, Provides useful information. |
| ISI | ratio | Initial spread index | Relevant, Provides useful information. |
| Temp | ratio | Temperature of fire | Relevant, Provides useful information. |
| RH | ratio | Relative humidity | Relevant, Provides useful information. |
| Wind | ratio | Wind speed | Relevant, Provides useful information. |
| Rain | ratio | Rain scale | Relevant, Provides useful information. |
| Area | ratio | Area in which fire burns | Relevant, Provides useful information. |