# 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: Nukala Ayyappa Bharthwaj**

**Batch Id: DSWDMCOS 21012022**

**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.

**Sol:**

**Business Objective:** to predict the customers will buy the properties or not with respect to their salaries and other factors by using Support vector machine.

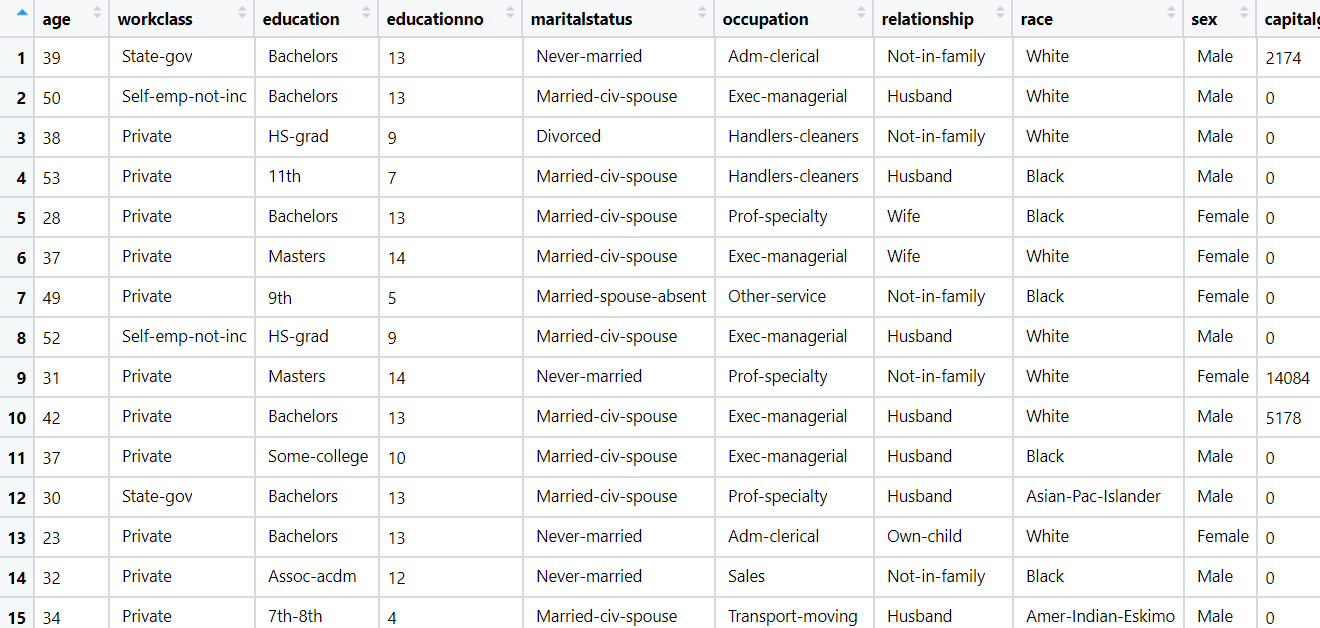
**Constraints:** Lack of analysis of the previous data related to the customers.

**Data Types:** The given data and its types are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data Type | Relevance |
| Age | Age of the customer | Ratio | Relevant |
| Workclass | Work Class of the customer | Nominal | Relevant |
| education | Education type of the customer | Nominal | Relevant |
| educationno | Rank given to the education of the customer | Ordinal | Relevant |
| maritalstatus | Whether the customer is married or not | Nominal | Relevant |
| occupation | Occupation of the customer | Nominal | Relevant |
| relationship | Relationship with the customer | Nominal | Relevant |
| Race | Racism of the customer | Nominal | Relevant |
| sex | Sex of the customer | Nominal | Relevant |
| capitalgain | Capital amount gained by the customer | Ratio | Relevant |
| capitalloss | Capital loss by the customer | Ratio | Relevant |
| hoursperweek | Number hours worked per week | Ratio | Relevant |
| native | Native country of the customer | Nominal | Relevant |
| Salary | Salary of the customer whether >50K or <=50K | Nominal | Relevant |

**Data Pre-Processing:** All the variables of the given data is used for doing the analysis. Some of the variables in the given data is in categorical format so the same is converted into numeric data in order to do the analysis.

**Support Vector Machine:** After cleaning the data the same is used for using the analysis by taking the output variable as salary. SVM model is generated for the given data and the accuracy of the linear kernel model is 84.62% and the accuracy RBF kernel model is 85.43%.



**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.

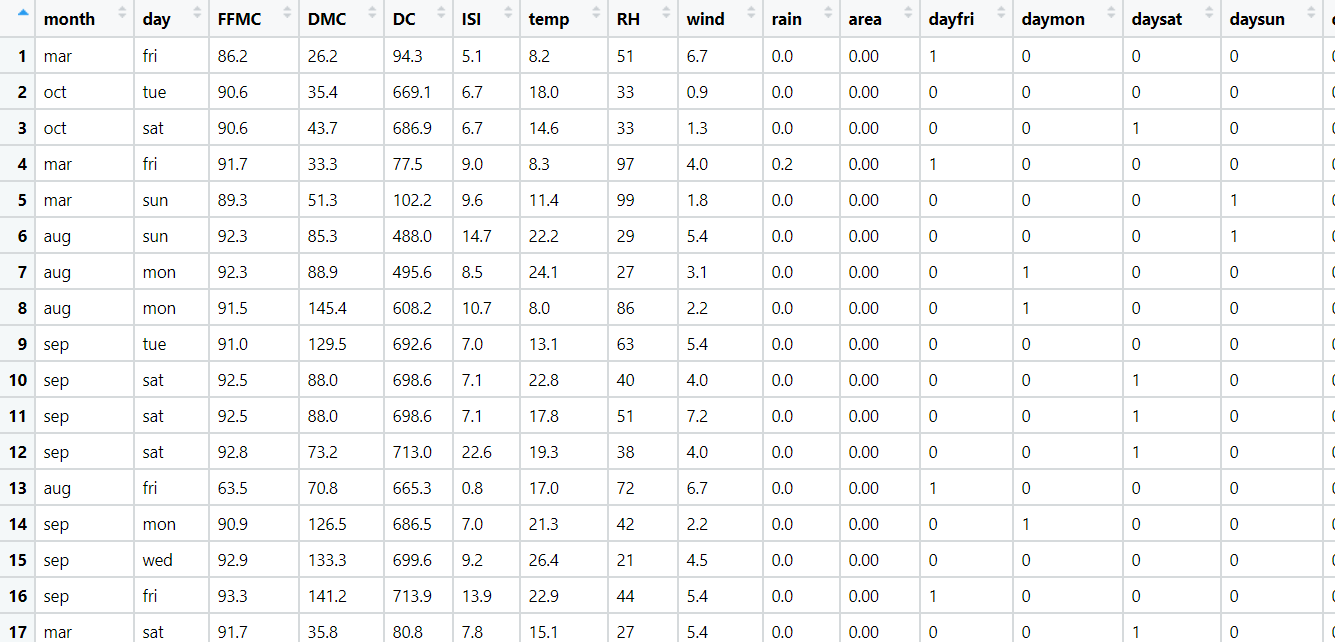
Note: - Size\_ category is the output variable.

**Business Objective:** To predict the forest area with other factors using support vector machine model.

**Constraints:** Lack of analysis of the previous forest fire data.

**Data Types:** Given data and its types are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data Type | Relevance |
| Month | Month when fire occurred | Nominal | Relevant |
| Day | Day when fire occured | Nominal | Relevant |
| FFMC | Fine Fuel Moisture Code for moisture content in fire data | Ratio | Relevant |
| DMC | Duff Moisture Code for moisture content in fire data | Ratio | Relevant |
| DC | third moisture index for moisture content in fire data | Ratio | Relevant |
| ISI | Initial Spread Index of the fire data | Ratio | Relevant |
| Temp | Temperature of the area | Ratio | Relevant |
| RH | Relative humidity of the fire area | Ratio | Relevant |
| Wind | Wind speed of the fire area | Ratio | Relevant |
| Rain | Rain forecast of the fire area | Ratio | Relevant |
| Area | Area of the fired region | Ratio | Relevant |
| Day type | All the days are converted into 1 or 0 based on type of day | Nominal | Relevant |
| Month type | All the months are converted into 1 or 0 based on type of months | Nominal | Relevant |
| size\_category | Size of the area whether small or bir | Nominal | Relevant |

**Data Pre-Processing:** All the given data is used for doing the analysis. Some of the data are in categorical format so the same is converted into numeric data to do the analysis on the data.

**Support Vector Machine:** After cleaning the data the same is used for using the analysis by taking the output variable as size category. SVM model is generated for the given data and the accuracy of the linear kernel model is 96% and the accuracy RBF kernel model is 82%.