# Topic: Support Vector Machines (SVM)

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: DHEERAJ MISHRA Batch ID:**  DS\_01072021

**Topic: SVM**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline

(OR)

* + Less than 30% of problems in the assignments are submitted before deadline
* **Grade F: (< 50):** No submission (or) malpractice.

**Hints:**

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
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. And if not relevant, provide reasons and a description of the feature.**

**3.Data Pre-processing**

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

**3.2 Outlier Treatment.**

**4. Exploratory Data Analysis (EDA):**

**4.1 Summary.**

**4.2 Univariate analysis.**

**4.3 Bivariate analysis.**

**5. Model Building**

* 1. **Build the model on the scaled data (try multiple options)**
  2. **Use the SVM algorithm.**
  3. **Train and test the model and compare accuracies by building a confusion matrix and use different hyperparameters.**
  4. **Briefly explain the model output in the documentation.**

**6. Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Problem Statement: -**

A construction firm wants to develop a suburban locality with new infrastructure but they might incur losses if they cannot sell the properties. To overcome this, they consult an analytics firm to get insights on how densely the area is populated and the income levels of residents. Use the Support Vector Machines algorithm on the given dataset and draw out insights and also comment on the viability of investing in that area.



1. BUSINESS OBJECTIVE:-

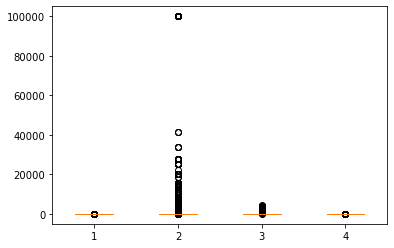
Maximize mapping of salary

1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| age | Age of employers | Discrete | Relevant |
| workclass | Work class of employers | Char | Relevant |
| education | Employers education | Char | Relevant |
| educationno | Employers education code | Discrete | Not relevant |
| maritalstatus | Employers married or not | Char | Relevant |
| occupation | Employers occupation | Char | Relevant |
| relationship | Employers relationship | Char | Relevant |
| race | Employers face | Char | Relevant |
| sex | Employers sex | Char | Relevant |
| capitalgain | Employers capital gain | Discrete | Relevant |
| capitalloss | Employers capital loss | Discrete | Relevant |
| hoursperweek | Working hours per week | Discrete | Relevant |
| native | Employers native place | Char | Relevant |
| Salary | Employers salary class | Discrete , char | Relevant |

1. DATA CLEANSING :-
2. From test and train datasets merged to single datasets
3. Dataset consists of 14 colums and 39239 rows
4. Dropping nominal column education no
5. Obtained duplicates rows and removed
6. All data types are of form int64 and object
7. No null values found in each column
8. From describe function mean , median and standard deviation obtained
9. Outliers detected but retained
10. Log transformation used for normal distribution
11. Dummy column obtained for object columns
12. Scaling is done through normalization techniques
13. EDA:-

Box plot for outliers



1. MODEL BUILDING:-
2. Model builded by SVM
3. For kernel linear
4. Test accuracy = 0.847543160690571
5. Train accuracy = 0.8480488047478532
6. For kernel rbf ( Guassian )
7. Test accuracy = 0.8397078353253652
8. Train accuracy = 0.8480488047478532

Output :-

Model with linear kernel gives better accuracy

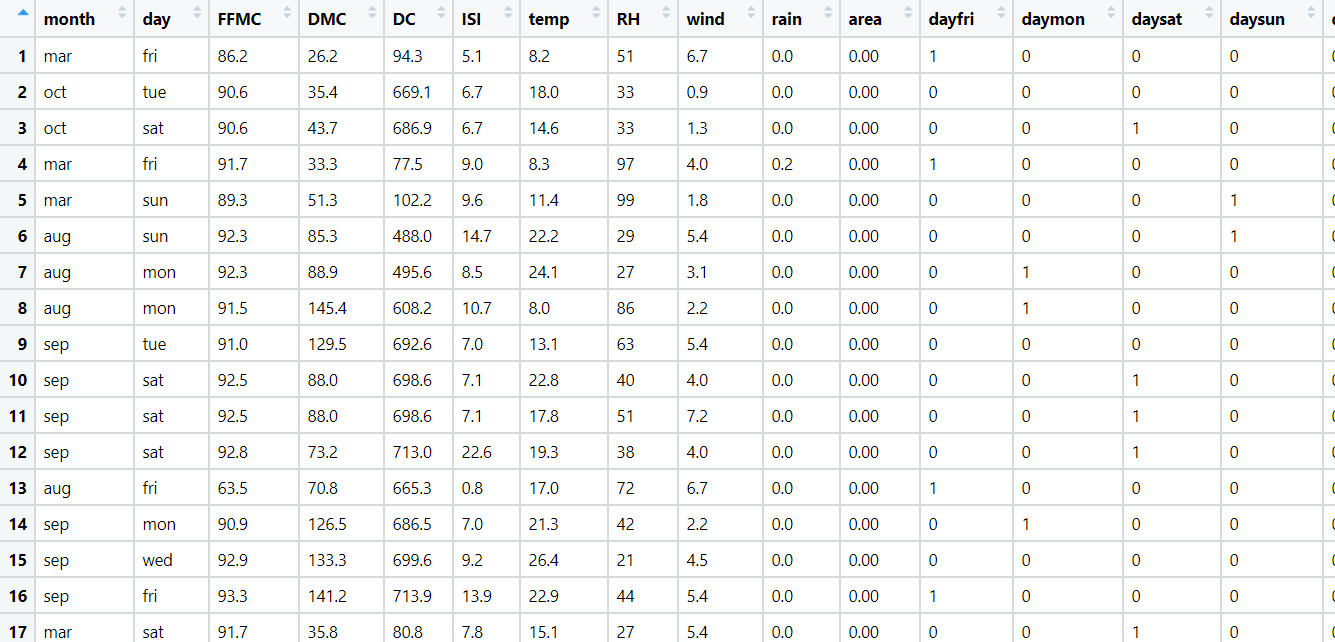
1. BENEFITS :-

From above information we can assign the salary class on basis of new dataset .

**Problem Statement: -**

In California, annual forest fires can cause huge loss of wildlife, human life, and can cost billions of dollars in property damage. Local officials would like to predict the size of the burnt 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.

1. BUSINESS OBJECTIVE:-

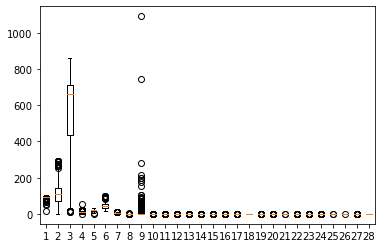
Maximize relationship between dependent variable

1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| month | Month of year | Text | Not Relevant |
| day | Day of week | Text | Not Relevant |
| FFMC | Fine Fuel Moisture Code | Continuous | Relevant |
| DMC | Duff Moisture Code | Continuous | Relevant |
| DC | Drought Code | Continuous | Relevant |
| ISI | Initial Spread Index | Continuous | Relevant |
| temp | temperature noon | Continuous | Relevant |
| RH | Relative Humidity | Continuous | Relevant |
| wind | Wind speed | Continuous | Relevant |
| rain | total day | Continuous | Relevant |
| area | Area burnt | Continuous | Relevant |

1. DATA CLEANSING :-
2. Dataset consists of 30 colums and 517 rows
3. Dropping nominal colums
4. Duplicate row exists and removed
5. All data types are of form float64 and object
6. No null values found in each column
7. From describe function mean , median and standard deviation obtained
8. Outliers present and retained
9. Corelation coefficients calculated
10. Some are positively skewed and some are negatively skewed
11. Scaling is done by normalisation
12. Splitting data to train and test
13. EDA:-

From box plot



1. MODEL BUILDING:-
2. Model builded by SVM
3. For kernel linear
4. Test accuracy = 0.9903846153846154
5. Train accuracy = 1
6. For kernel rbf ( Guassian )
7. Test accuracy = 0.7980769230769231
8. Train accuracy = 1

Output :-

Model with linear kernel gives better accuracy

1. BENEFITS :-

From above information we can assign the burnt area class on basis of new dataset .