# 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: Prajay B. Urkude Batch ID: 16092021**

**Topic: SVM**

**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.



Ans:- Business Objective:

To build a model which predict the income level of the residents.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| age | Age of resident | Quantitative, ratio | Relevant |
| workclass | Class of work | Qualitative, Nominal | Relevant |
| education | Education of resident | Qualitative, ordinal | Relevant |
| educationno | The education no of the resident | Quantitative, ordinal | Relevant |
| maritalstatus | Marital status of the resident | Qualitative, Nominal | Relevant |
| occupation | Occupation of the resident | Qualitative, Nominal | Relevant |
| relationship | Relationship of the residents | Qualitative, Nominal | Relevant |
| race | Race of the residents | Qualitative, Nominal | Irrelevant |
| sex | Sex male/female | Qualitative, Nominal | Relevant |
| capitalgain | Capital income of resident | Quantitative, ratio | Relevant |
| capitalloss | Expenditure of the resident | Quantitative, ratio | Relevant |
| hoursperweek | Hours per week work do by resident | Quantitative, ratio | Relevant |
| native | Native place of the resident | Qualitative, Nominal | Irrelevant |
| Salary | Salary class <=50k, >50k | Quantitative, Nominal | Relevant |

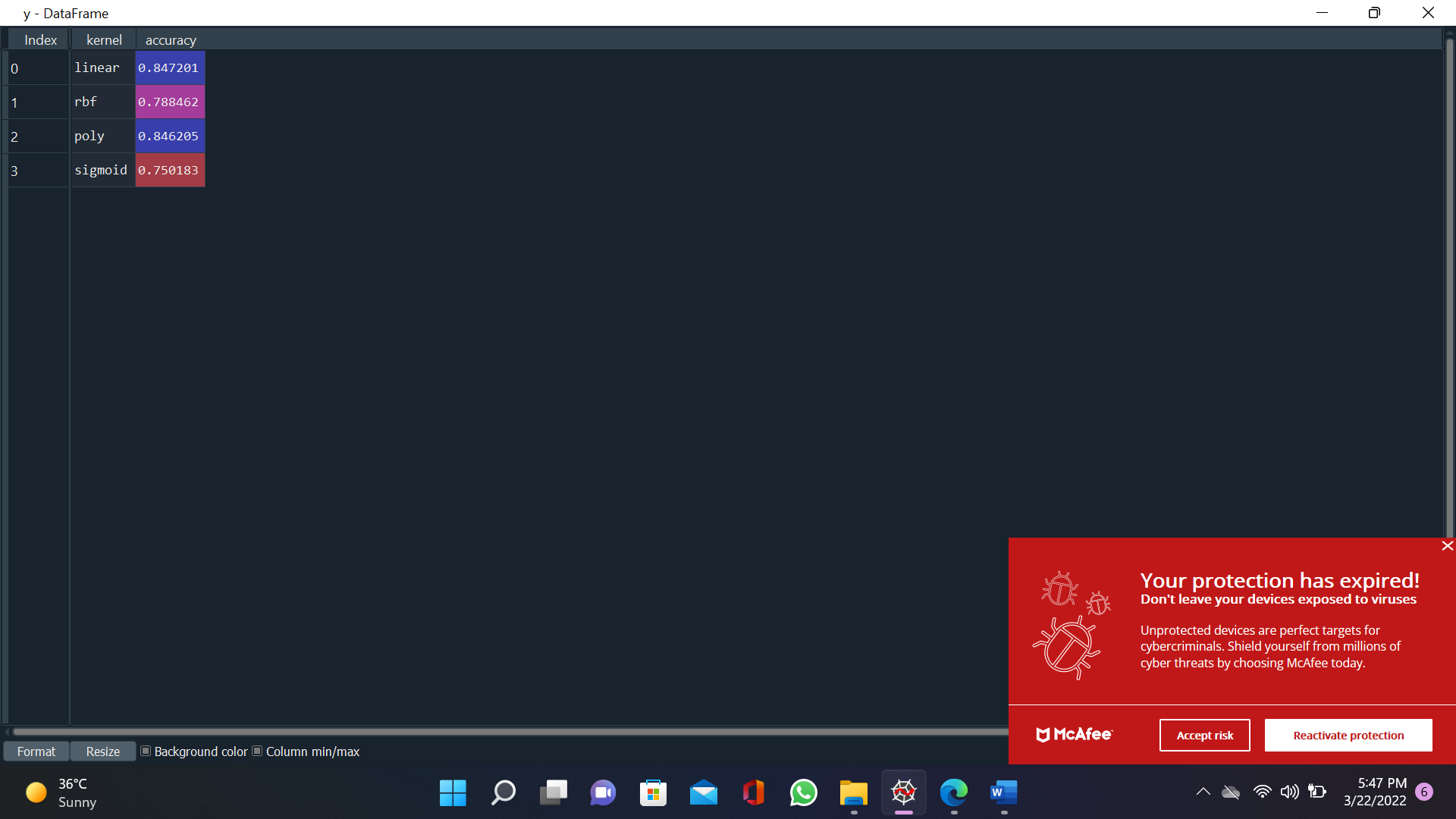
**Steps to build a model by using the support vector machine**

* Import the libraries such as numpy, pandas, matplotlib, sklearn, statsmodel and import the packages.

From sklearn library

**Support vector machines (SVMs)** are a set of supervised learning methods used for classification, regression and outliers’ detection.

* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Splitting the data into train and test data by using the train\_test\_split function.
* Initialize the model by using the SVC function by taking different kernels like linear, rbf, poly, sigmoid, precomputed.
* Fit the model in the train datasets and then evaluate the model on the test datasets and find out the accuracy.
* Evaluate the model on the train datasets and find out the accuracy and compare with the test accuracy.
* From the linear kernel we get the highest accuracy i.e.84.89% which is the highest among all the kernel models.

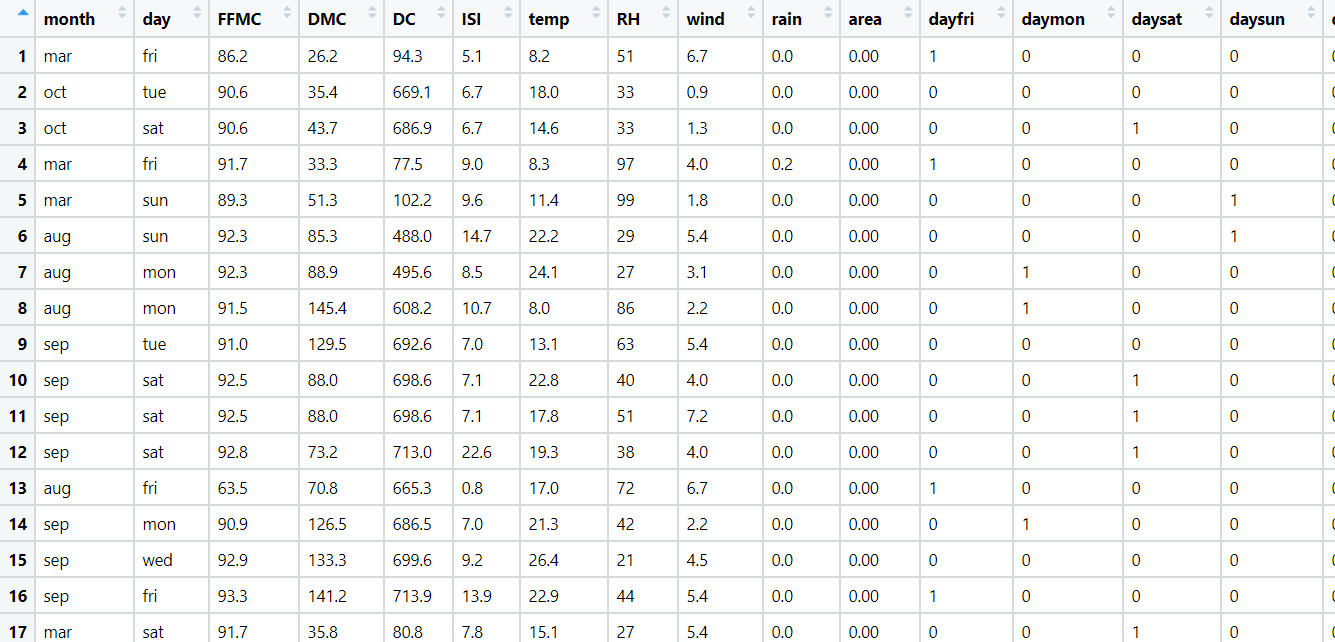


**Problem Statement: -**

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 outp



Ans: Business Objective:

To build a model to predict the size of the burnt area because of the forest burns.

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of features** | **Description** | **Type** | **Relevance** |
| month | Month in which the forest fire happen | Qualitative, Nominal | Relevant |
| day | Day on which the forest fire happen | Qualitative, Nominal | Relevant |
| FFMC | FFMC index from the FWI system: 18.7 to 96.20 | Quantitative, Ratio | Relevant |
| DMC | DMC index from the FWI system: 1.1 to 291.3 | Quantitative, Ratio | Relevant |
| DC | DC index from the FWI system: 7.9 to 860.6 | Quantitative, Ratio | Relevant |
| ISI | ISI index from the FWI system: 0.0 to 56.10 | Quantitative, Ratio | Relevant |
| temp | temperature in Celsius degrees: 2.2 to 33.30 | Quantitative, Ratio | Relevant |
| RH | relative humidity in %: 15.0 to 100 | Quantitative, Ratio | Relevant |
| wind | wind speed in km/h: 0.40 to 9.40 | Quantitative, Ratio | Relevant |
| rain  area | outside rain in mm/m2 : 0.0 to 6.4 | Quantitative, Ratio | Relevant |
| size\_category | the burned area size | Qualitative, Nominal | Relevant |

**Steps to build a model by using the support vector machine**

* Import the libraries such as numpy, pandas, matplotlib, sklearn, statsmodel and import the packages.

From sklearn library

**Support vector machines (SVMs)** are a set of supervised learning methods used for classification, regression and outliers’ detection.

* Doing the univariate analysis and Exploratory data analysis.
* Checking the head i.e., top 5 rows of the datasets
* Checking the columns names of the datasets
* Checking the null values if any available in dataset.
* Checking the duplicate values in the datasets
* Checking the information i.e., datatypes of the datasets
* Exploratory data analysis. mean, median, mode, count, min, max etc.
* Check the distribution of the data.
* Dropping the unwanted column which is not useful for the analysis.
* Converting the nonnumerical data into numerical data by using one hot encoding or Label Encoder or pandas get\_dummies function as per the requirement
* Converting the continuous data into discrete form if necessary.
* Splitting the data into train and test data by using the train\_test\_split function.
* Initialize the model by using the SVC function by taking different kernels like linear, rbf, poly, sigmoid, precomputed.
* Fit the model in the train datasets and then evaluate the model on the test datasets and find out the accuracy.
* Evaluate the model on the train datasets and find out the accuracy and compare with the test accuracy.
* From the poly kernel we get the highest accuracy i.e.86.5% which is the highest among all the kernel models and the accuracy is nearly equal to the train accuracy.

