School of Computer Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

LAB-1

COURSE ID: ITA6016

FACULTY: Dr. Kiruthika S

SUBMITTED BY:

AVANTIKA SHARMA(21MCA1001)

ADITI VERMA(21MCA1086)

ARYAN MITTAL(21MCA1097)

Abstract:

For the medical cost dataset, a support vector regression model will be constructed. The dataset includes independent and dependent features such as age, sex, BMI (body mass index), children, smokers, and geography. We'll forecast how much each patient's medical bills from insurance will be.

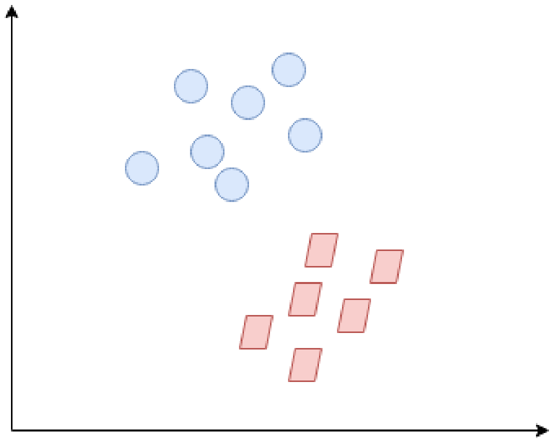
Definition & Working principle

SVR:

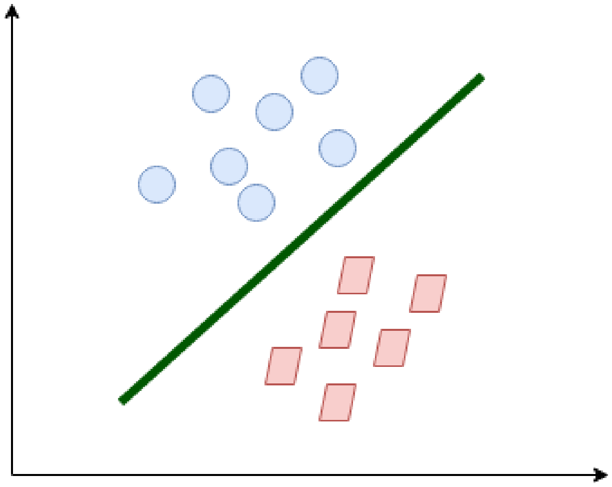
Support Vector Machines (SVMs in short) are machine learning algorithms that are used for classification and regression purposes. SVMs are one of the powerful machine learning algorithms for classification, regression and outlier detection purposes. An SVM classifier builds a model that assigns new data points to one of the given categories. Thus, it can be viewed as a non-probabilistic binary linear classifier.

Basic Explanation:

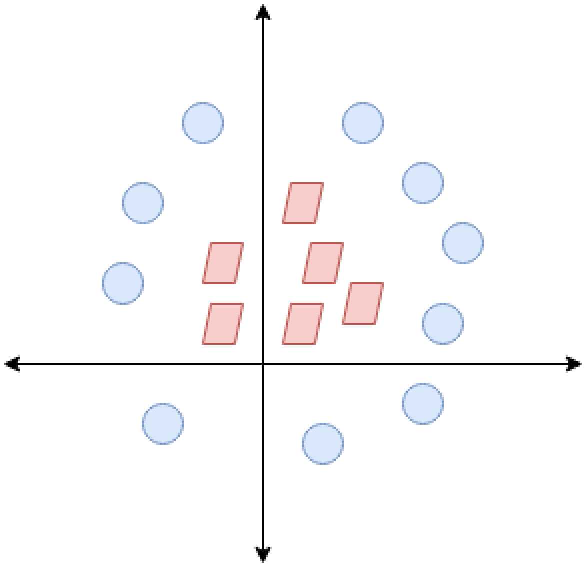
Let’s say we have a plot of two label classes as shown in the figure below:



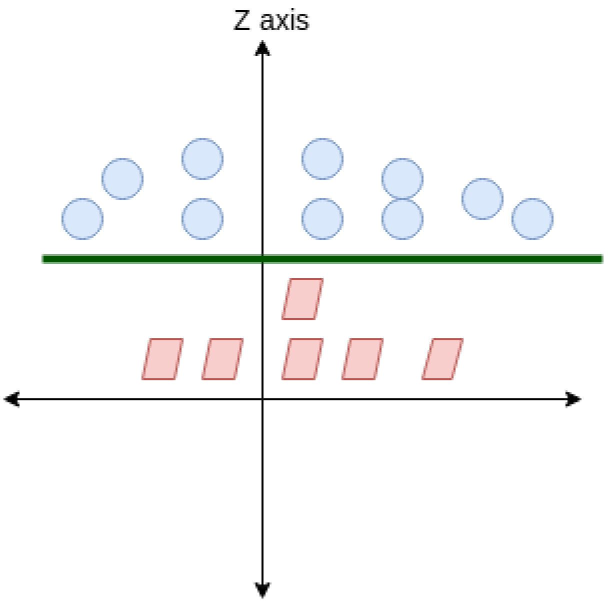
If we try to separate the classes , we might come up with this:



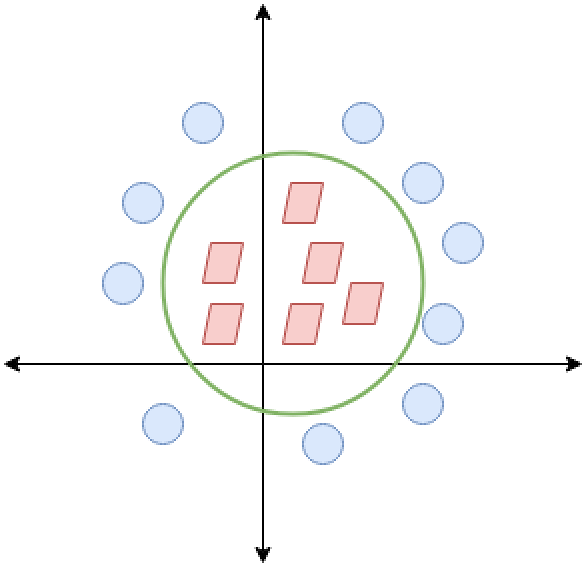
The line fairly separates the classes. This is what SVM essentially does – simple class separation. Now, what is the data was like this:



Here, we don't have a simple line separating these two classes. So we'll extend our dimension and introduce a new dimension along the z-axis. We can now separate these two classes:



When we transform this line back to the original plane, it maps to the circular boundary as I've shown here:

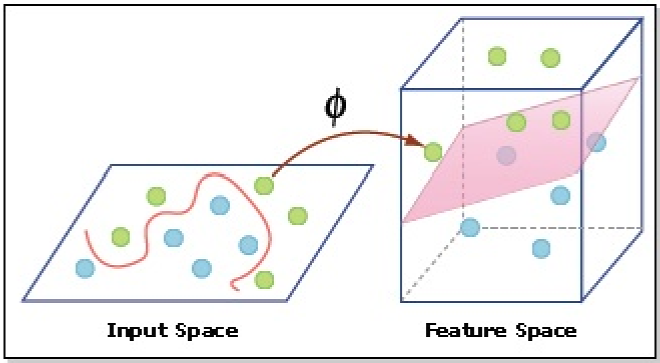


SVM performs just this! In multidimensional space, it looks for a line or hyperplane that divides these two classes. Then, using the classes to forecast, it assigns the new point a classification based on whether it is located on the positive or negative side of the hyperplane.

Hyperparameters of the Support Vector Machine (SVM) Algorithm

There are a few important parameters of SVM that you should be aware of before proceeding further:

- **Kernel:** Without raising the computing cost, a kernel aids in the discovery of a hyperplane in the higher dimensional space. Usually, as the dimension of the data grows, the computing cost grows as well. When we can't go in a certain dimension because there isn't a dividing hyperplane there, we must move in a higher dimension instead.

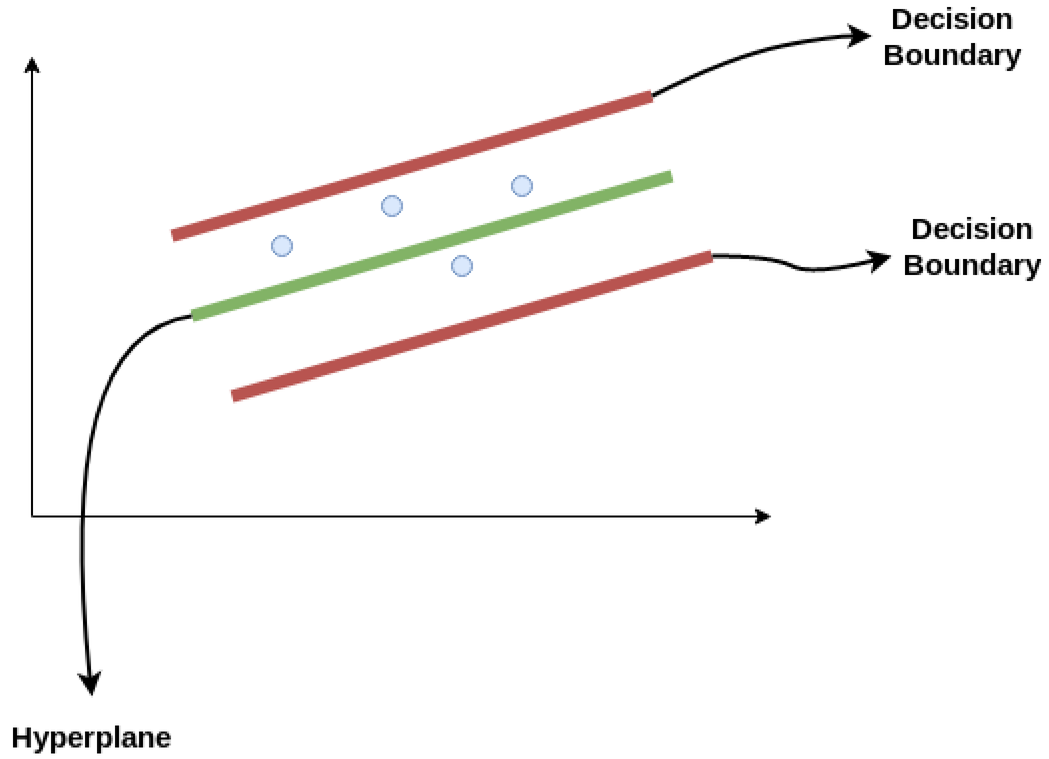


- **Hyperplane:** In SVM, this essentially acts as a boundary between two data classes. However, this line will be utilised in Support Vector Regression to forecast the continuous output.

- **Decision Boundary:** A decision boundary can be conceptualised as a demarcation line (for the sake of simplification) where the positive examples are on one side and the negative examples are on the other. The instances can be categorised as either positive or negative along this exact line. Support Vector Regression will also use this same SVM concept.

Support Vector Regression (SVR) uses the same principle as SVM, but for regression problems.

The problem of regression is to find a function that approximates mapping from an input domain to real numbers on the basis of a training sample. So let’s now dive deep and understand how SVR works actually.



Think of the green line as the hyperplane, and the two red lines as the decision boundaries. When using SVR, our goal is to essentially take into account the points that are inside the decision boundary line. The hyperplane with the most points serves as our best fit line.

The decision boundary (the hazardous red line above!) is the first concept that we will comprehend. Think of these lines as being at any distance from the hyperplane, let's say 'a'. The lines that we draw at '+a' and '-a' distances from the hyperplane are thus as follows. The text basically refers to this 'a' as epsilon.

Hyperplane

A hyperplane is a decision boundary which separates between given set of data points having different class labels. The SVM classifier separates data points using a hyperplane with the maximum amount of margin. This hyperplane is known as the maximum margin hyperplane and the linear classifier it defines is known as the maximum margin classifier.

Support Vectors

Support vectors are the sample data points, which are closest to the hyperplane. These data points will define the separating line or hyperplane better by calculating margins.

Margin

A margin is a separation gap between the two lines on the closest data points. It is calculated as the perpendicular distance from the line to support vectors or closest data points. In SVMs, we try to maximize this separation gap so that we get maximum margin.

DATASET DESCRIPTION:

Columns:

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height,
  objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of children covered by health insurance / Number of dependents
- smoker: Smoking

- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance.

Number of rows and columns in the data set: (1338, 7)

LIBRARIES USED:

NUMPY:

A general-purpose array processing package is called NumPy. It offers a high-performance multidimensional array object as well as utilities for interacting with these arrays. It is the core Python package for scientific computing. It is open-source software. It has a number of characteristics, including these crucial ones.An effective N-dimensional array object sophisticated (broadcasting) operations.Tools for combining C/C++ and Fortran code useful linear algebra, fourier transform, and random number capabilities.

NumPy can be used as a productive multi-dimensional container of generic data in addition to its apparent scientific applications. Numpy's ability to establish any data-types enables NumPy to quickly and easily interact with a wide range of databases.

PANDAS:

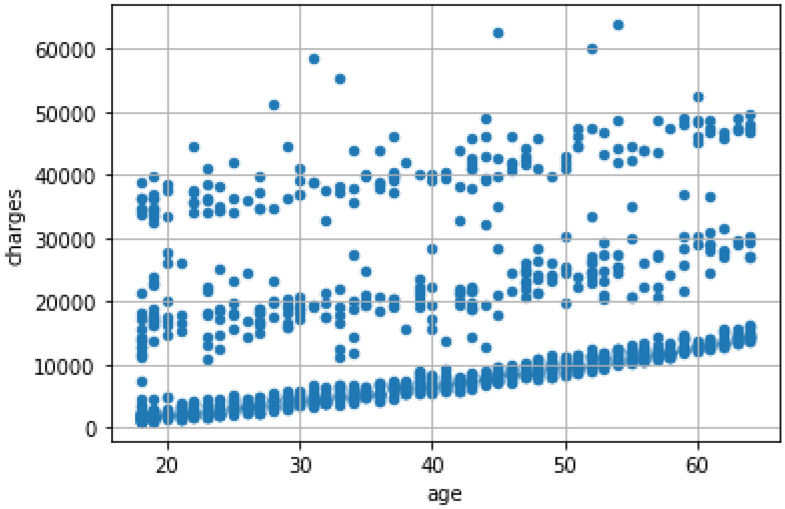
Pandas is an open-source library designed primarily for working quickly and logically with relational or labeled data. It offers a range of data structures and procedures for working with time series and numerical data.

The NumPy library serves as the foundation for this library. Pandas is quick and offers its users exceptional performance & productivity.

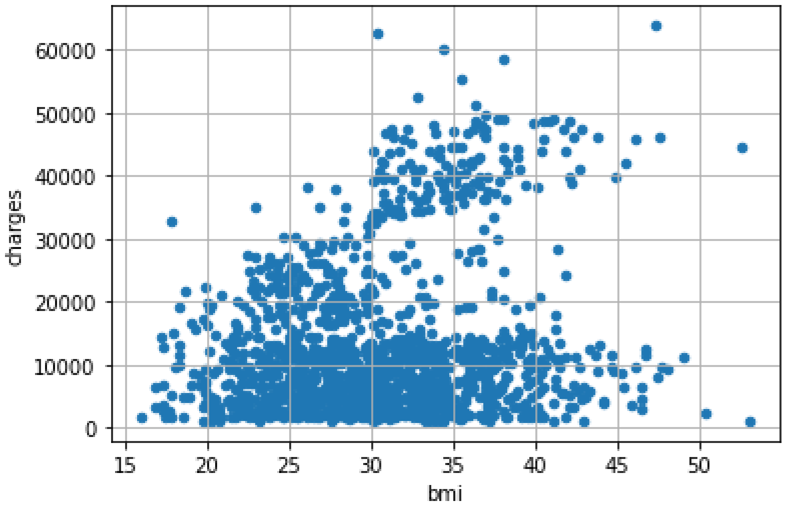
MATPLOTLIB:

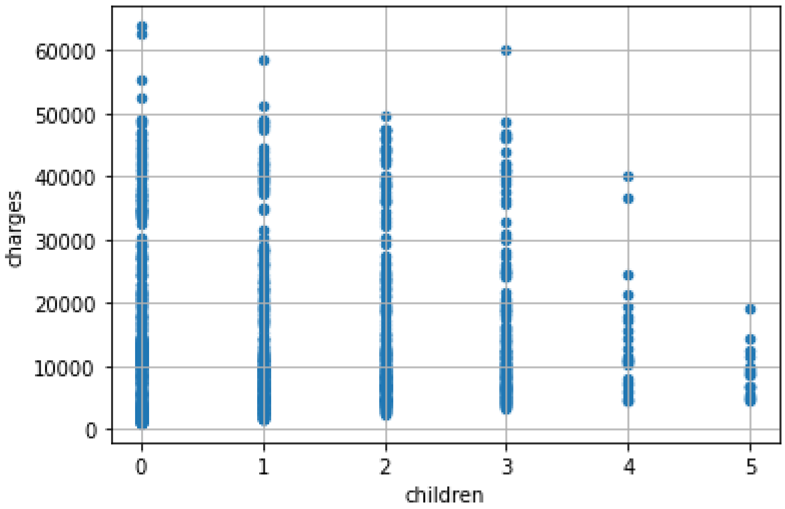
Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

VISUALS OF THE DATASET:



We have plotted the points where the x axis represents **age** and y axis represents **medical charges**.We have done this to see the relation of our target attribute on independent attribute.





Before we train our model , first we have to preprocess the dataset. Our SVR algorithm does not take any string values so we have to modify the sex and smoker attribute.

For sex attribute:

Male -> 0

Female ->1

For smoker attribute:

Non-smokers->0

Smoker-> 1

RESULTS:

We have analyzed our model using the three kernels and get the RMSE(Root mean square error) scores with all the three kernels on the same dataset.

	RBF	Linear	Polynomial
RMSE Score	12859.0127515	13017.656314	12787.468416

Conclusion:

We can see that the RBF kernel is more efficient as compared to other kernels.RBF kernel shows less RMSE score compared to Linear and Polynomial kernels.

APPENDIX:

Link to the dataset:

https://www.kaggle.com/datasets/noordeen/insurance-premium-prediction