# **What are regression and classification ML techniques? Explain in about 500 words how these techniques are helpful in different sectors?**

**There are two popular categories of machine learning techniques: regression and classification. .**

A supervised learning method called regression is employed to forecast a continuous result. Finding the link between the variable we are attempting to predict and one or more independent variables is required (the variables we are using to make the prediction). Finding a best fit line that minimises the difference between the expected and actual values may be done with linear or nonlinear regression models (Tigga and Garg, 2020).

**The following are a few incidents of regression issues:**

* Estimating a house's cost based on its dimensions, location, and other characteristic
* Estimating a company's stock price based on its fiscal performance and market circumstances
* Calculating the quantity of rainfall that will fall in a specific place depending on the weather

An example of supervised learning is classification, which is used to forecast a categorical result. It consists of classifying a data point into one of a limited number of groups or classes. A classification model could be used, for instance, to identify the kind of animal in an image or to determine whether or not an email is spam.

**Problems with categorization include the following examples:**

* Determining a customer's likelihood to churn quit to use a product or service based on historical behaviour and other characteristics
* Determining a patient's likelihood of having a certain ailment according on their health records and test findings
* Estimating if a sentence is positive, negative or neutral

**Techniques like regression and classification may be applied in a range of fields including marketing, finance, and healthcare.**

Regression models can be used in finance to forecast stock prices or evaluate credit risk. Classification models may be used in the healthcare industry to forecast a patient's risk of contracting a certain disease or to forecast which course of therapy will be most successful for that patient. Regression models in marketing may be used to estimate a customer's lifetime value or to determine how much to charge for a product, whereas classification models may be used to estimate customer churn or to determine the categories of goods  buyers is most likely to buy.

Overall, data-based decision-making and outcome prediction certainly benefit from the use of regression and classification algorithms. They may assist companies and organisations in streamlining their operations, enhancing decision-making and gaining a competitive edge in their marketplaces.

# **Explain overfitting and underfitting? How to combat Overfitting and Underfitting problems?**

**Whenever the model is trained on either a dataset in machine learning, overfitting and underfitting were frequent issues.**

**Overfitting**

Overfitting is when a model performs badly on fresh, untrained data because it was trained well on training data. The model cannot generalise these patterns to fresh data because it has learnt the noise and unpredictable variations inside the dataset too well.

**Underfitting**

When a model is unable to identify the fundamental trend of the data, underfitting occurs, and as a result, the model performs badly on both the training data and new, unforeseen data. This occurs because your model is inadequately complicated to account again for diversity of the data (Li et al., 2020).

There are numerous strategies for overcoming both over- and underfitting:

* In cross-validation, the training data is divided into many folds, the model is trained on part of a folds, and the model is then validated on the other folds. This makes it easier to determine how well the model generalises.
* Regularization: This includes introducing a penalty into the objective function of the model to prevent it from picking up too many intricate patterns from the data. In doing so, overfitting is fought.
* Expanding the training dataset will aid in the model's capacity to learn additional reliable patterns, which will aid in preventing overfitting.
* Bringing down the model's complexity: A smaller model is less likely to overfit, but it could also be less correct.
* Ensembling: This process includes developing several models, integrating their outputs, and then making a final forecast. This can aid in lowering the model's variance, which can aid in preventing overfitting.
* Early halting: This entails keeping an eye on the model's performance on a validation data during training and preventing the approach when the performance begins to deteriorate. This may aid in avoiding overfitting.

# **Can you explain the difference between validation set and a test set?**

**The validation set or a test set is 2 different sets of the dataset that are utilized for various reasons in ML and data science.**

In machine learning and data science, a validation set is a set of data used to evaluate a model during the training phase. The model is trained on a training set, and then evaluated on the validation set to tune the hyperparameters (such as the learning rate or regularization strength) and ensure that the model is not overfitting to the training data.

A test set, on the other hand, is a set of data used to evaluate the final performance of a trained model. The test set is usually held out from the training process, and is used to evaluate the model's performance on unseen data. This is important because we want to know how well the model will perform on new, previously unseen data in the real world.

It is important to use both a validation set and a test set when developing and evaluating machine learning models. The validation set is used to tune the model and ensure that it is not overfitting, while the test set is used to evaluate the final performance of the model on unseen data. Using both a validation set and a test set helps ensure that the model is able to generalize to new data and will perform well in the real world.

# **Explain SVM algorithm with an example?**

Support Vector Machines (SVMs) are a type of supervised learning algorithm that can be used for classification or regression. SVMs are based on the idea of finding the hyperplane in a high-dimensional space that maximally separates different classes.

To understand how SVMs work, let's consider a simple example of binary classification (classifying data points into two classes). Imagine we have a dataset with two classes, "positive" and "negative," that are linearly separable.We can plot the data points on a two-dimensional graph, with the x-axis representing one feature and the y-axis representing another feature.

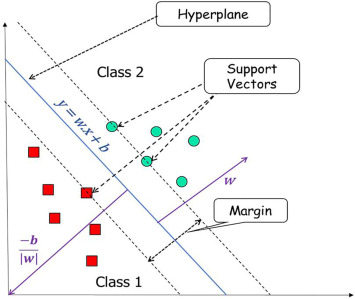


Figure SVM Plane

Source:- Sciencedirect.com

In this example, we can see that the two classes can be separated by a straight line. This line is called the decision boundary, and it is used to classify new data points as either positive or negative.

However, in many real-world datasets, the classes are not always linearly separable. In these cases, we can use a technique called the kernel trick to transform the data into a higher-dimensional space where it is linearly separable.

The goal of the SVM algorithm is to find the maximum margin hyperplane that maximally separates the two classes. This is done by optimizing an objective function that tries to maximize the margin between the two classes while simultaneously minimizing the number of misclassified data points.

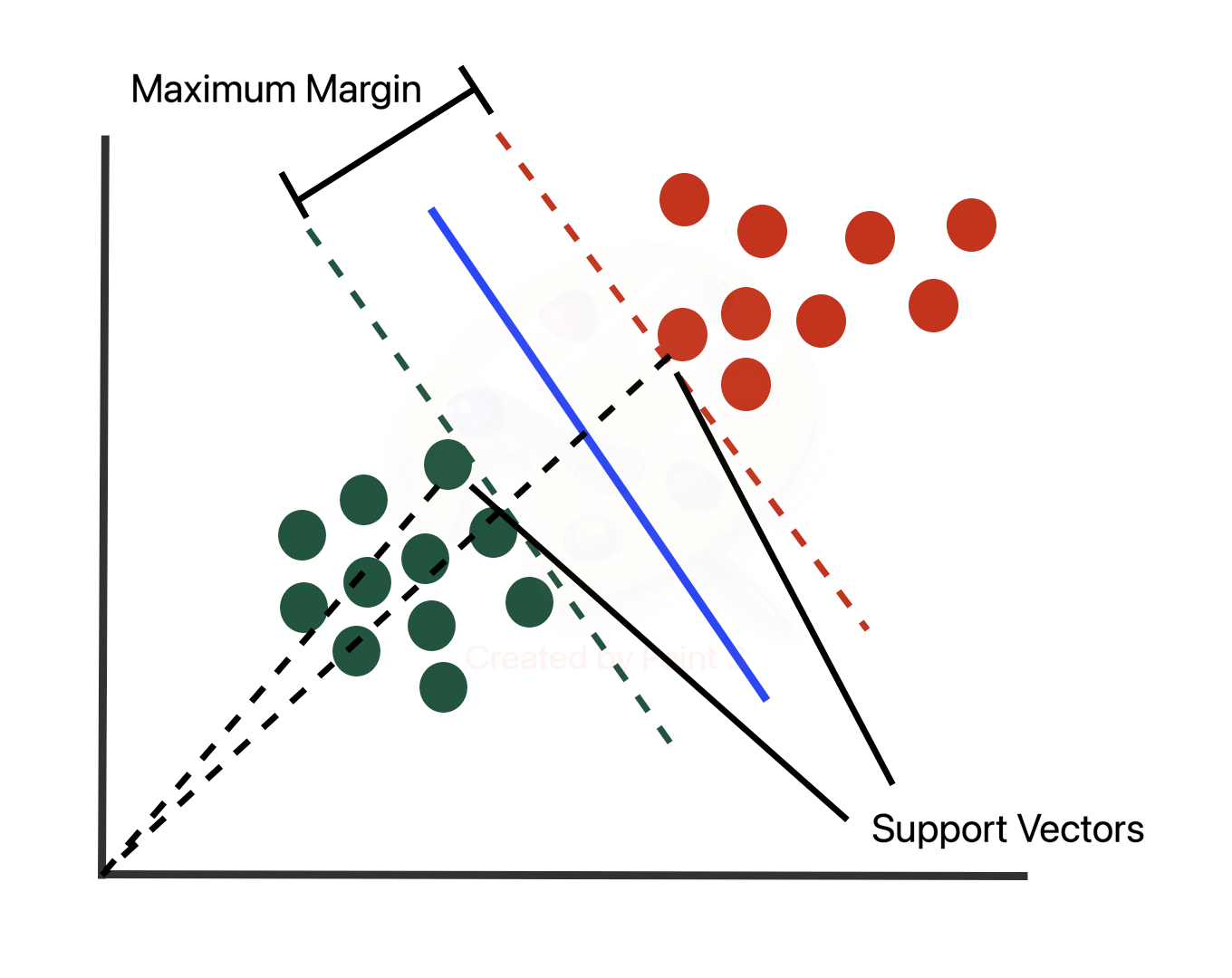


Figure SVM Example with margin and seperate line

Source:- Towards Data science

In the above example, the solid line represents the maximum margin hyperplane, and the dotted lines represent the margins. The points that lie on the margins are called support vectors, and they play a crucial role in defining the maximum margin hyperplane.

Once the SVM model has been trained, we can use it to classify new data points by projecting them onto the hyperplane and determining which side of the hyperplane they fall on. Data points that fall on one side of the hyperplane are classified as one class, while data points that fall on the other side are classified as the other class.

SVMs have several advantages, including their ability to handle high-dimensional data and their ability to perform well even when the number of features is much greater than the number of samples. They are also robust to noise and can handle outliers well. However, they can be sensitive to the choice of hyperparameters and may require careful tuning to achieve good performance.

In summary, Support Vector Machines are a powerful and widely-used machine learning algorithm that can be used for classification or regression. They work by finding the maximum margin hyperplane that separates different classes in a high-dimensional space, using a technique called the kernel trick to handle non-linearly separable data. SVMs are known for their ability to handle high-dimensional data and their robustness to noise and outliers, but they can be sensitive to the choice of hyperparameters and may require careful tuning to achieve good performance.

# **Explain the Decision Tree algorithm and how it works?**

Decision trees are a type of supervised learning algorithm that can be used for classification or regression. They work by creating a tree-like model of decisions and splits based on certain features.

To understand how decision trees work, let's consider a simple example of binary classification (classifying data points into two classes). Imagine we have a dataset with two classes, "positive" and "negative," and a number of features that we can use to make predictions.

The decision tree algorithm begins at the root node and splits the data into two or more subsets based on a certain feature. For example, the algorithm might split the data based on the feature "hair length," with data points having "short" hair going to the left branch and data points having "long" hair going to the right branch.

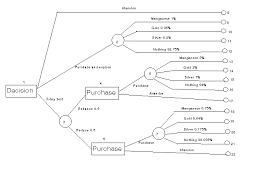


Figure Decesion Tree Examplr

Source:- People brunet.ac.uk

The algorithm then continues to split the data at each node based on the most important feature, until it reaches a leaf node. A leaf node is a node that does not split the data any further, and the prediction is made at the leaf node based on the majority class of the data points that reach that node.

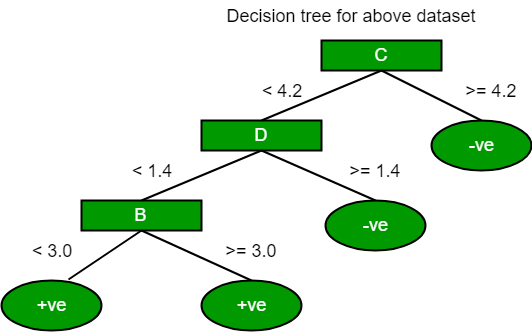


Figure Decesion Tree

Source:- Algorythem Class.com

In this example, the decision tree has made a prediction of the "positive" class at the leaf node because the majority of the data points at that node belong to the "positive" class.

One of the main advantages of decision trees is their interpretability. It is easy to understand how the model is making predictions because the splits and decisions are clearly visible in the tree structure. Decision trees are also easy to implement and require little data preparation.

However, decision trees can be prone to overfitting, especially if they are allowed to grow too deep. To prevent overfitting, we can use techniques such as pruning or setting a maximum depth for the tree.

In summary, decision trees are a simple and interpretable machine learning algorithm that can be used for classification or regression. They work by creating a tree-like model of decisions and splits based on certain features, with the goal of creating the most pure leaf nodes (i.e., leaf nodes with the highest percentage of a single class). Decision trees are easy to understand and implement, but can be prone to overfitting if they are allowed to grow too deep.

# **. What is forward and backward propagation and explain its working?**

**Forward Propagation**

Forward propagation and backward propagation are two key concepts in neural network training. They refer to the process of forward and backward passing of input data and error gradients through a neural network during the training process.

Forward propagation refers to the process of passing input data through a neural network to compute the output predictions. This is done by applying the weights and biases at each layer to the input data, and then applying an activation function to the resulting outputs. The output of each layer becomes the input to the next layer, and this process continues until the final output prediction is produced.

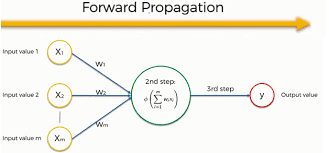


Figure 5 Forward Propagation

Source:- RPub.com

In the above diagram, we can see the process of forward propagation through a simple neural network with three layers (an input layer, a hidden layer, and an output layer). The input data is passed through the network, and the weights and biases at each layer are applied to produce the output predictions.

**Backward propagation**

Backward propagation, on the other hand, refers to the process of computing the error gradients at each layer in the neural network in order to update the weights and biases. This is done using the gradient descent optimization algorithm, which minimizes the loss function (the difference between the predicted output and the true output) by adjusting the weights and biases in the opposite direction of the error gradients.

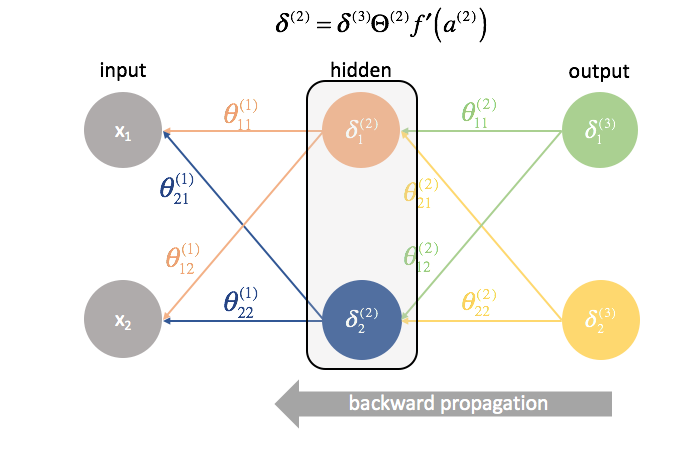


Figure Backward Propagation

Source:- RPub.com

In the above diagram, we can see the process of backward propagation through the same neural network as in the forward propagation example. The error gradients are computed at each layer and used to update the weights and biases in the opposite direction.

The process of forward and backward propagation is repeated for multiple epochs (passes through the entire dataset) during the training process, with the goal of minimizing the loss function and improving the accuracy of the model.

In summary, forward propagation and backward propagation are two key concepts in neural network training. Forward propagation refers to the process of passing input data through a neural network to compute the output predictions, while backward propagation refers to the process of computing the error gradients at each layer in order to update the weights and biases using the gradient descent optimization algorithm. Both processes are essential for training a neural network and improving its accuracy.

# **References**

Hirasawa, K., Ohbayashi, M., Koga, M. and Harada, M., 2018, June. Forward propagation universal learning network. In *Proceedings of international conference on neural networks (ICNN'96)* (Vol. 1, pp. 353-358). IEEE.

Charbuty, B. and Abdulazeez, A., 2021. Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends*, *2*(01), pp.20-28.

Abedini, M., Ghasemian, B., Shirzadi, A. and Bui, D.T., 2019. A comparative study of support vector machine and logistic model tree classifiers for shallow landslide susceptibility modeling. *Environmental Earth Sciences*, *78*(18), pp.1-15.

Mishra, P., Varadharajan, V., Tupakula, U. and Pilli, E.S., 2018. A detailed investigation and analysis of using machine learning techniques for intrusion detection. *IEEE communications surveys & tutorials*, *21*(1), pp.686-728.