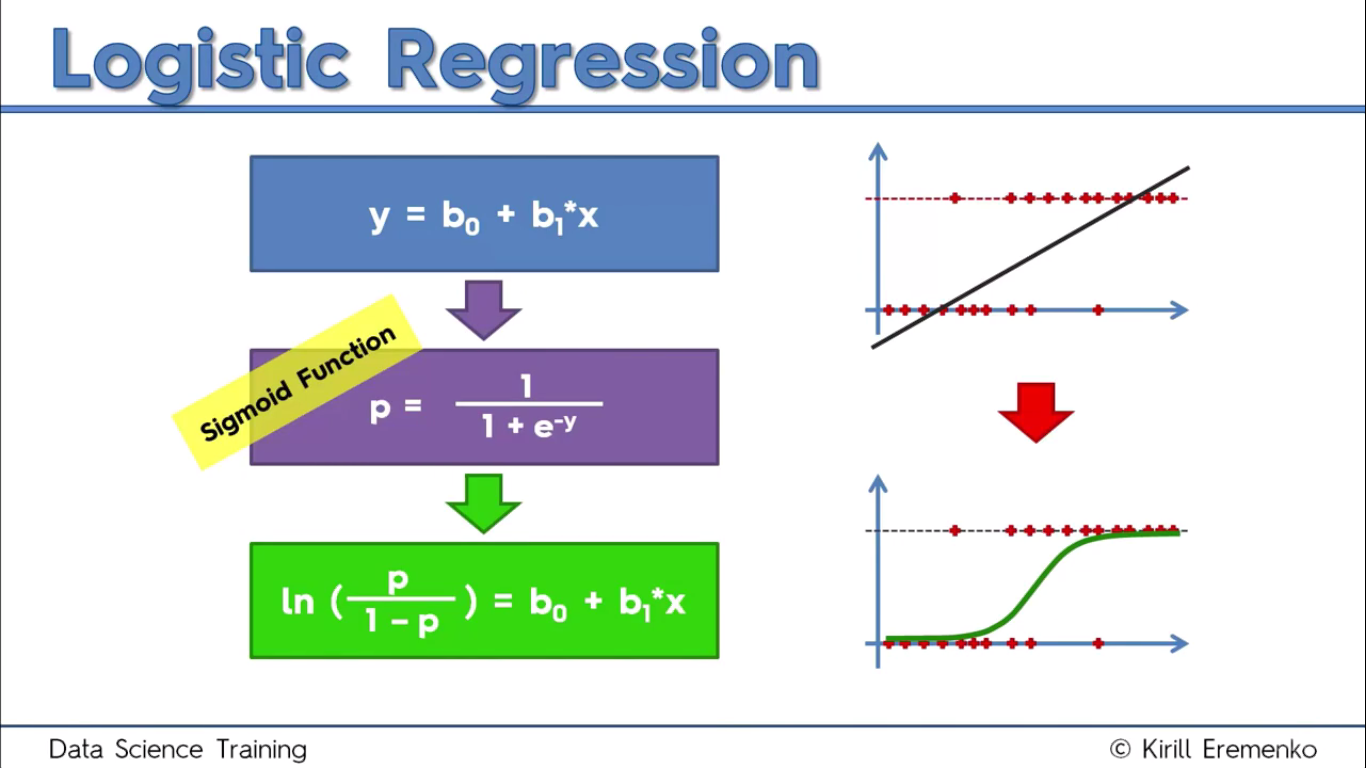
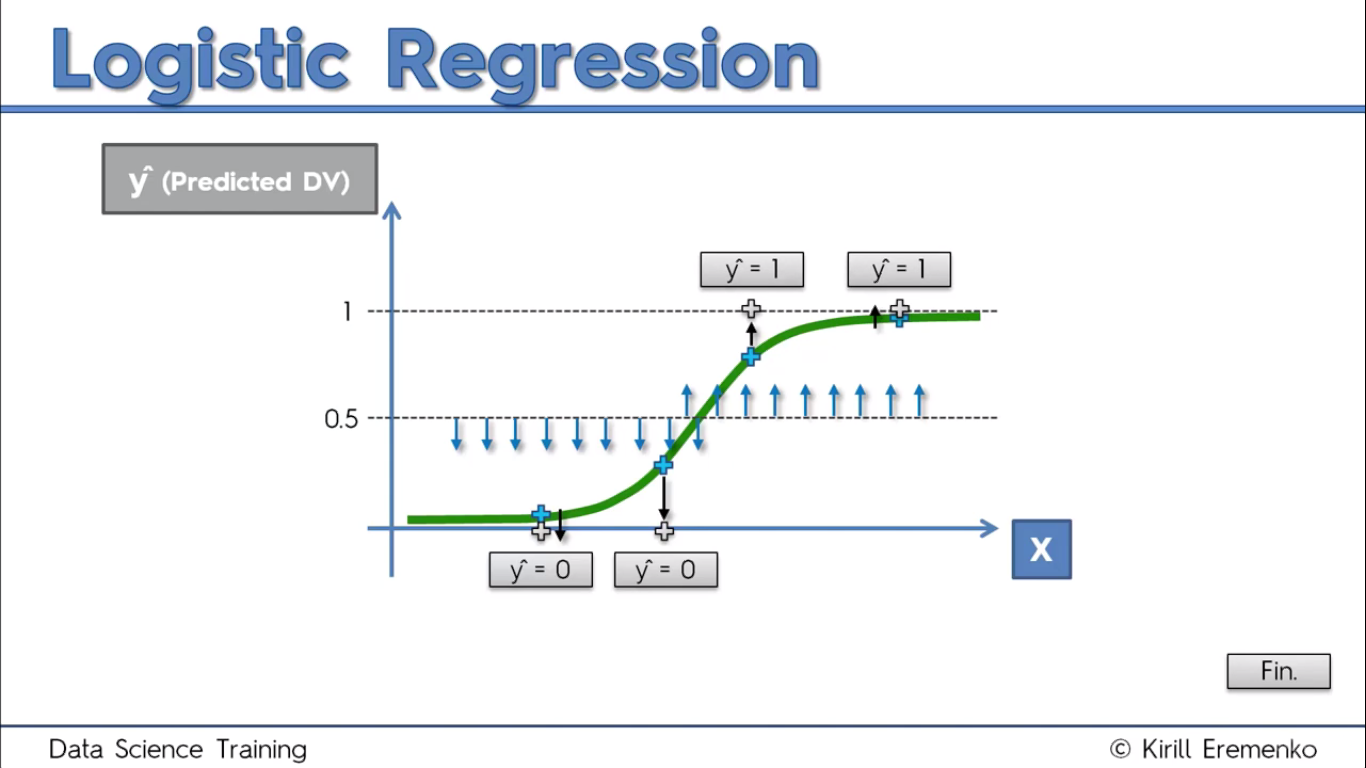
NOTES ON ML Classification

Unlike Regression where we predict a continuous number, we use classification to predict a category. There is a wide variety of classification applications from medicine to marketing. Classification models include linear models like Logistic Regression, SVM, and nonlinear ones like K-NN, Kernel SVM and Random Forests.

Logistic Regression

Logistic Regression is like Linear Regression since it finds the curve of best fit. Consider a company that seeks to predict who will use their product sent via email. For y(Action) its either Y/N :



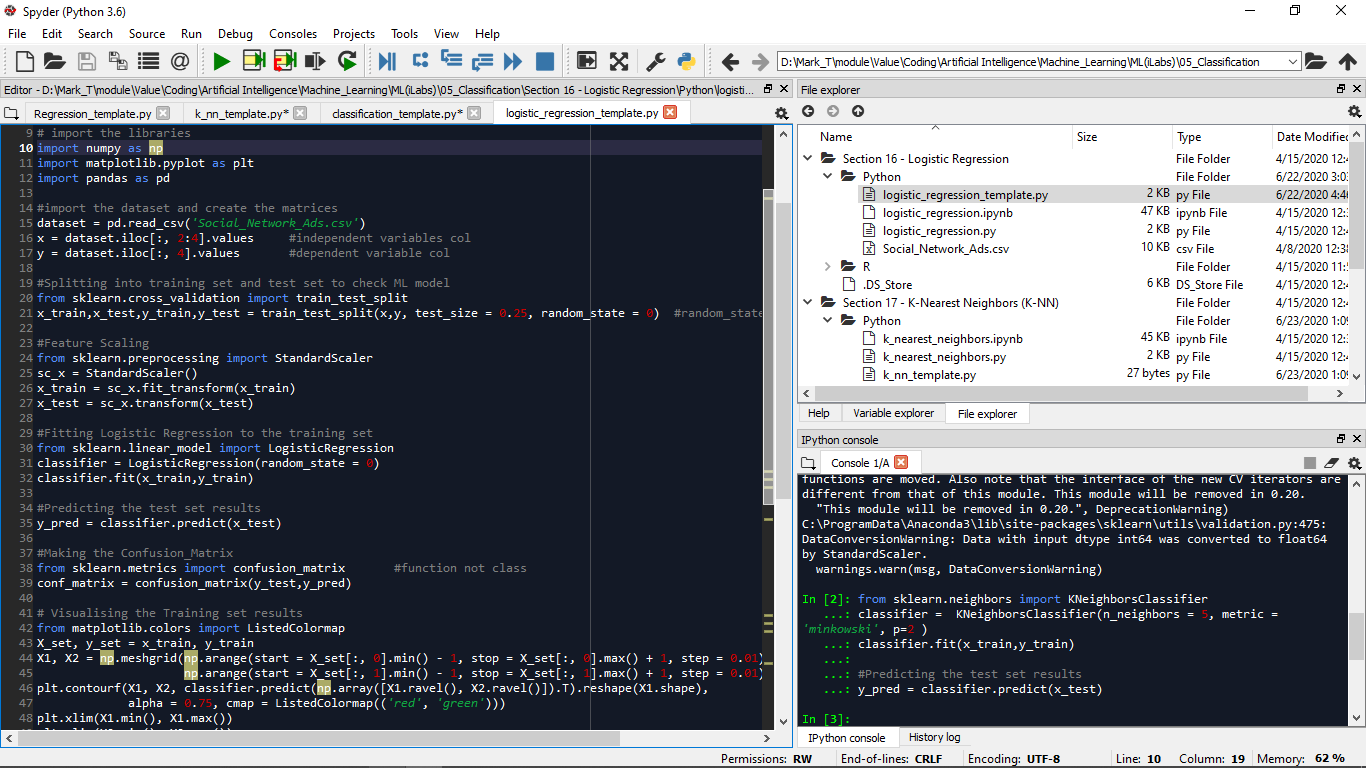
Predicting an exact yes or no can be approached in terms of probability for this problem. Using the sigmoid function puts values of y in a range of 0 – 1. So the y after regression, in terms of p is the formula of the best fit curve. We can acquire the probability of a test value, then classify basing on the range it lies in for different categories.

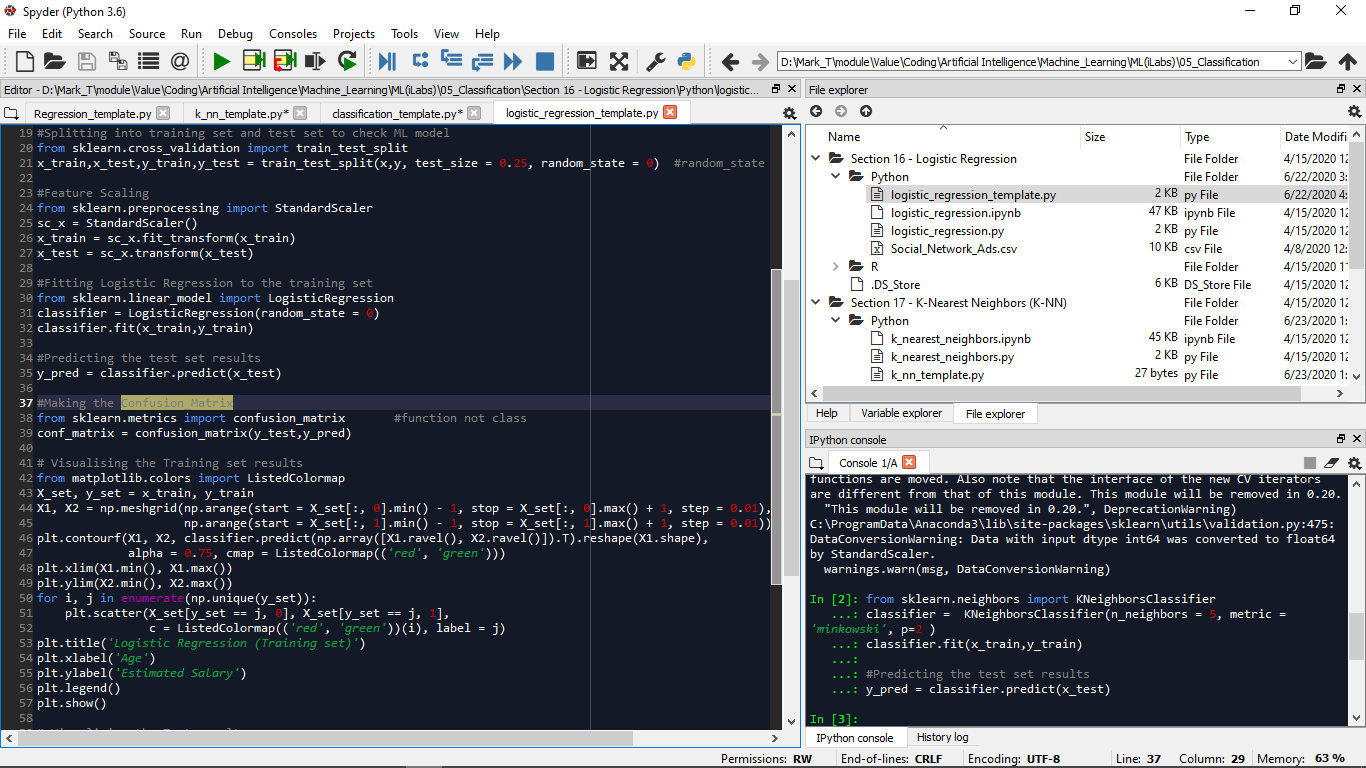
In Python

A company has just put an Ad on their new SUV for a ridiculous price and we seek to predict who among our users will purchase.

Solution:

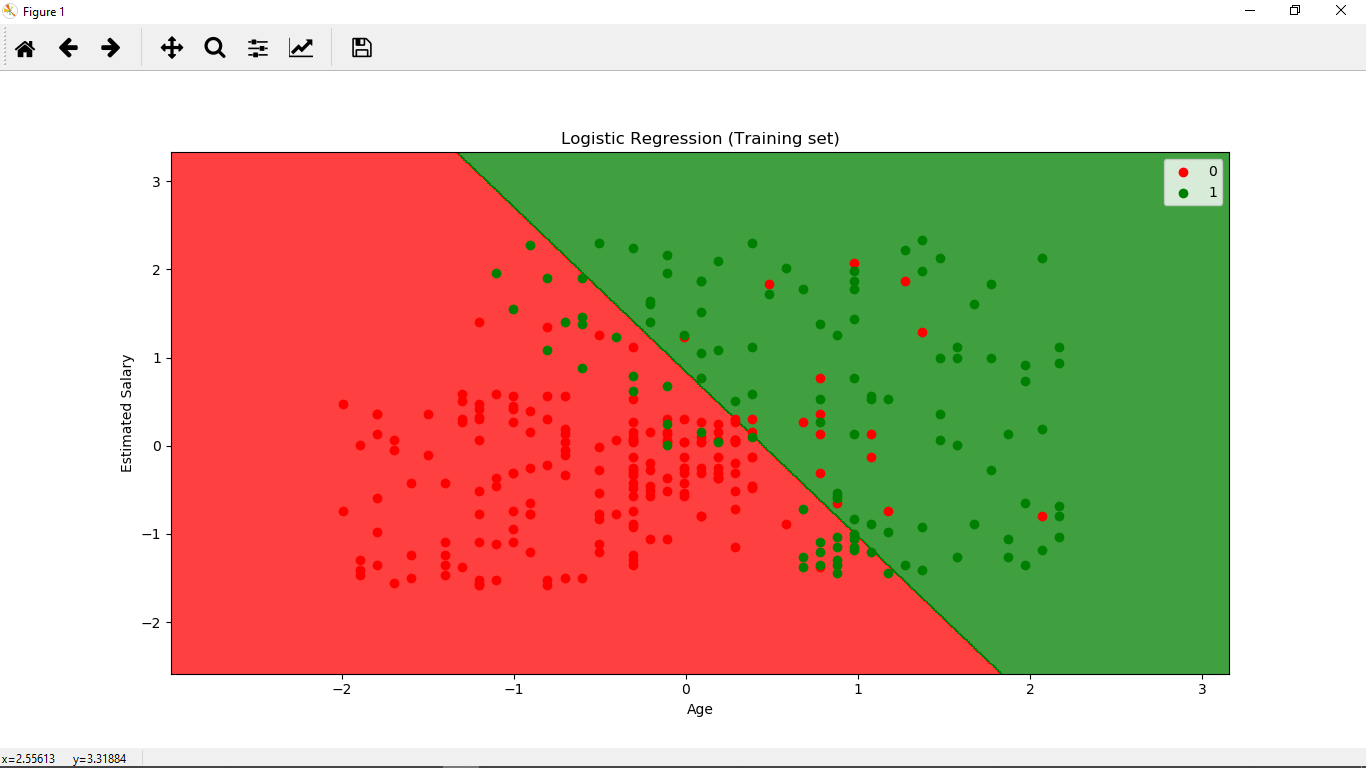
We shall use Age and EstimatedSalary as the IVs and Purchased as the DV.





Because Logistic Regression is a linear classifier, we expect it in sklearn.linear\_model.

For the confusion matrix, we evaluate whether the model understood the correlations between the x\_train and y\_train to see if it can make powerful predictions on future sets. It helps check the model’s success percentage. So we use the sklearn.metrics library. From the code, the leading diagonal are the correct predictions which total to 89/100, and wrong are 11/100. But that’s the beginning; the visual helps you evaluate better.



The straight line is the prediction boundary and is straight coz the model is linear. We see incorrect predictions because our users are not linearly distributed. For the test set, the regions don’t change because that’s the object’s behavior.(Whatif u didn’t need to redo with all old and new data)

**Qns**

**Whats the Sigmoid function**

Using the sigmoid function puts values of y in a range of 0 – 1, with the sum of 1.

**What are current applications of Logistic Regression class**

﻿Approximately 70% of problems in Data Science are classification problems.

Advantages

Logistic regression does work better when you remove attributes that are unrelated to the output variable as well as attributes that are very similar (correlated) to each other. Therefore Feature Engineering plays an important role in regards to the performance of Logistic and also Linear Regression.

Because of its simplicity and the fact that it can be implemented relatively easy and quick, Logistic Regression is also a good baseline that you can use to measure the performance of other more complex Algorithms.

Disadvantages

Also, we can’t solve non-linear problems with logistic regression since it’s decision surface is linear.

Logistic regression will not perform well with independent variables that are not correlated to the target variable and are very similar or correlated to each other.

Applications

In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc. Each object being detected in the image would be assigned a probability between 0 and 1, with a sum of one.

Logistic regression is used in various fields, including machine learning, most medical fields, and social sciences. For example, the Trauma and Injury Severity Score (TRISS), which is widely used to predict mortality in injured patients, was originally developed by Boyd et al. using logistic regression.[4] Many other medical scales used to assess severity of a patient have been developed using logistic regression.[5][6][7][8] Logistic regression may be used to predict the risk of developing a given disease (e.g. diabetes; coronary heart disease), based on observed characteristics of the patient (age, sex, body mass index, results of various blood tests, etc.).[9][10] Another example might be to predict whether a Nepalese voter will vote Nepali Congress or Communist Party of Nepal or Any Other Party, based on age, income, sex, race, state of residence, votes in previous elections, etc.[11] The technique can also be used in engineering, especially for predicting the probability of failure of a given process, system or product.[12][13] It is also used in marketing applications such as prediction of a customer's propensity to purchase a product or halt a subscription, etc.[14] In economics it can be used to predict the likelihood of a person's choosing to be in the labor force, and a business application would be to predict the likelihood of a homeowner defaulting on a mortgage. Conditional random fields, an extension of logistic regression to sequential data, are used in natural language processing.

**I need examples of questions to practice using each model**

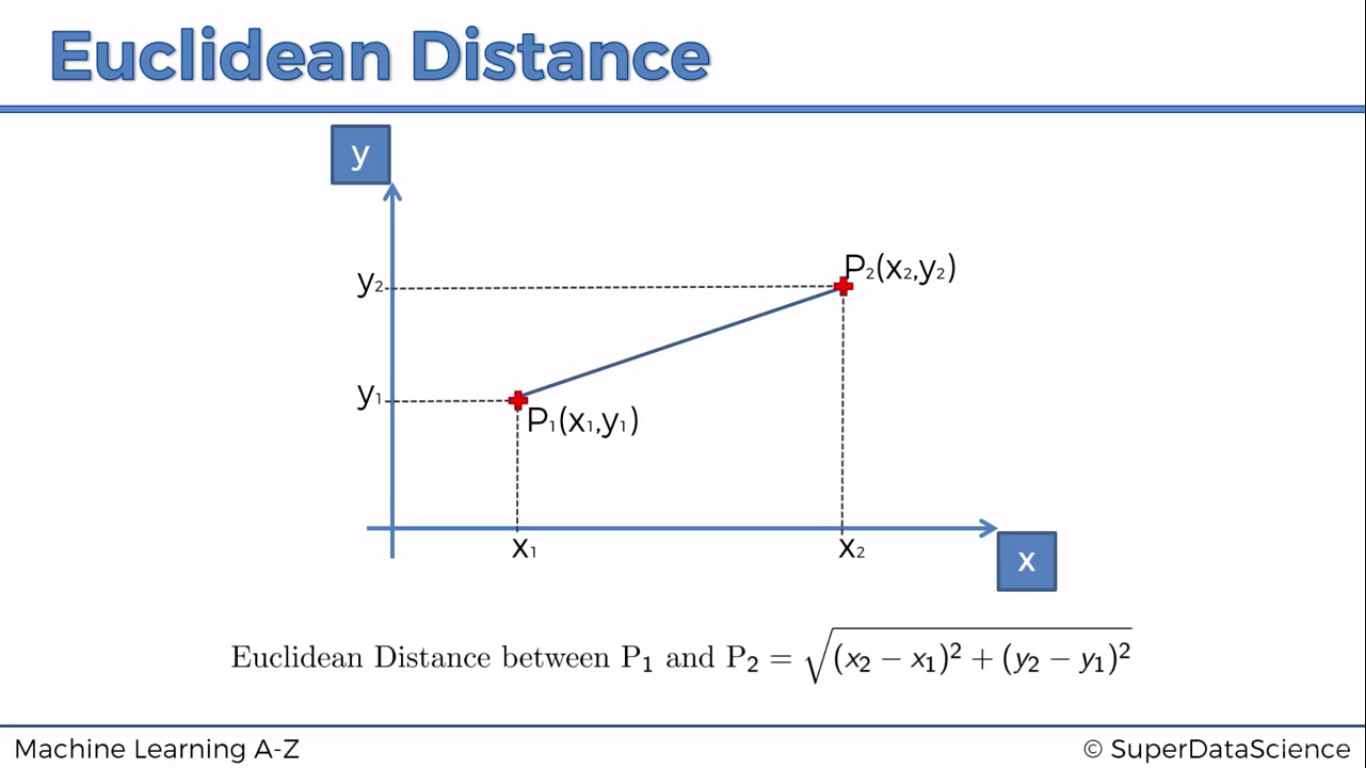
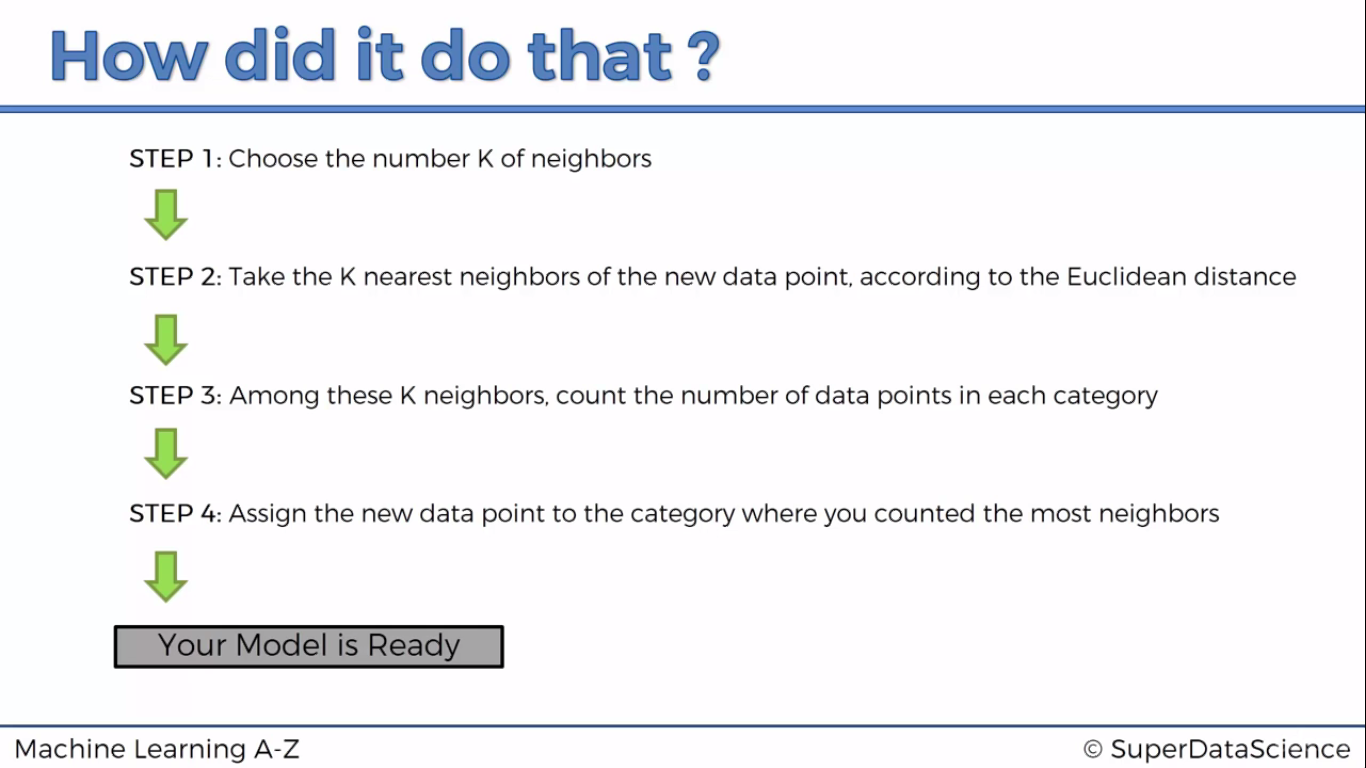
**How does LogisticRegression.fit() work**

logistic regression uses maximum likelihood estimation (MLE) to obtain the model coefficients that relate predictors to the target.

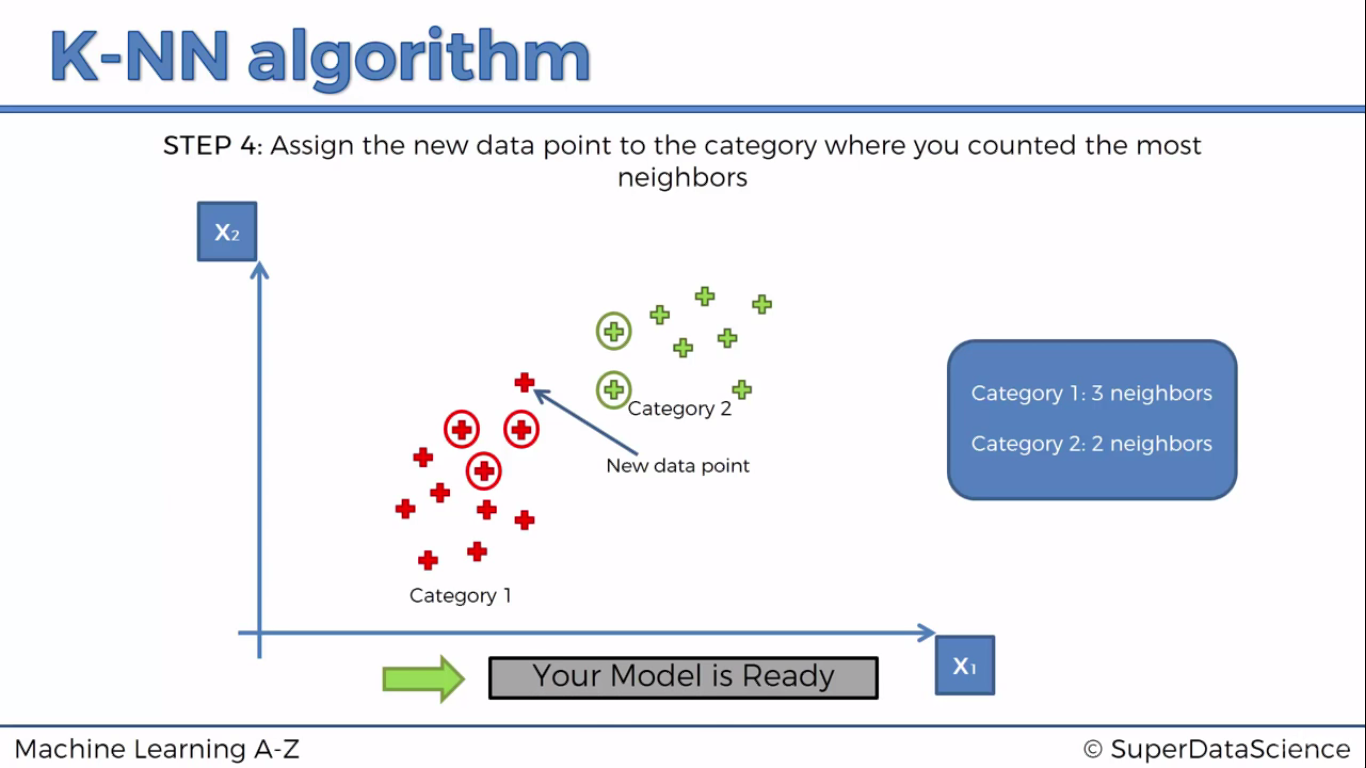
**How to chose whether to use regression or classification**

**K** – Nearest Neighbors Notes

Consider a case where we want to decide the category for a new data point

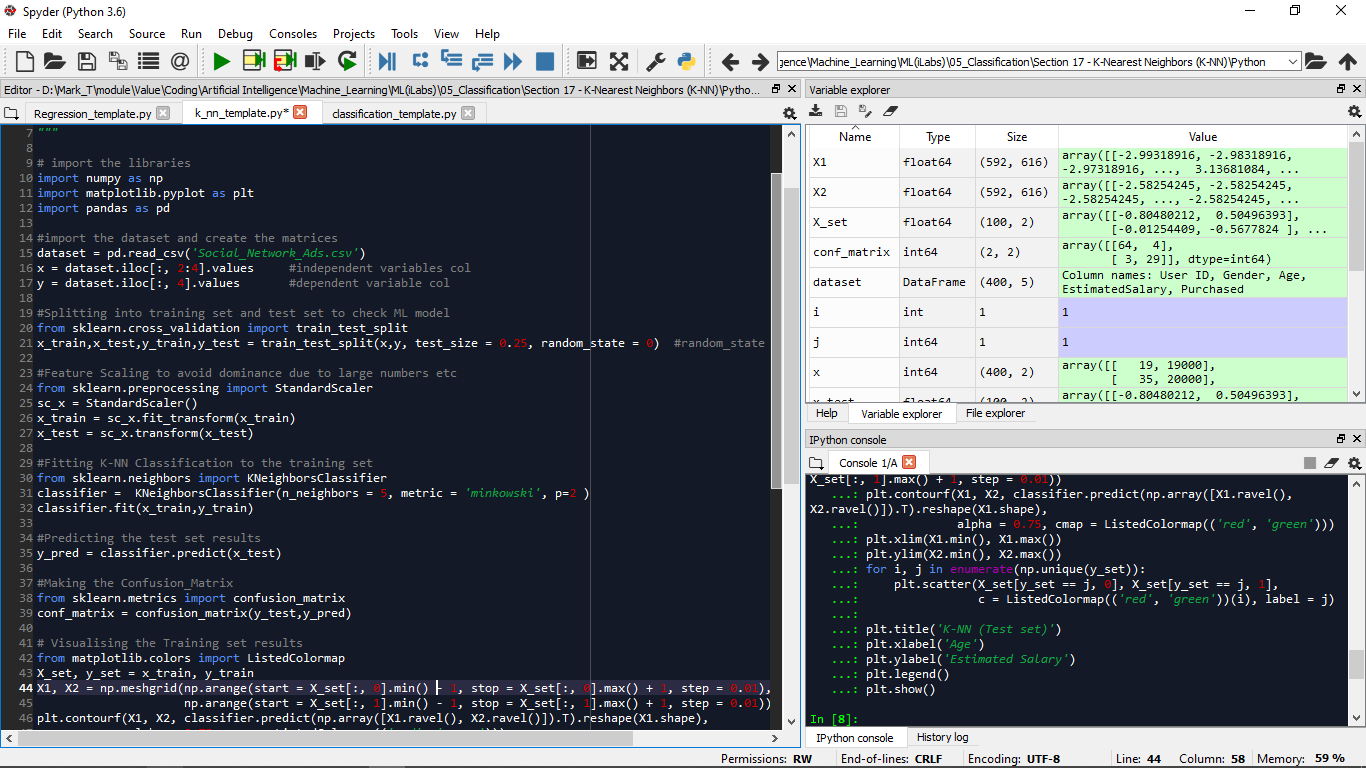


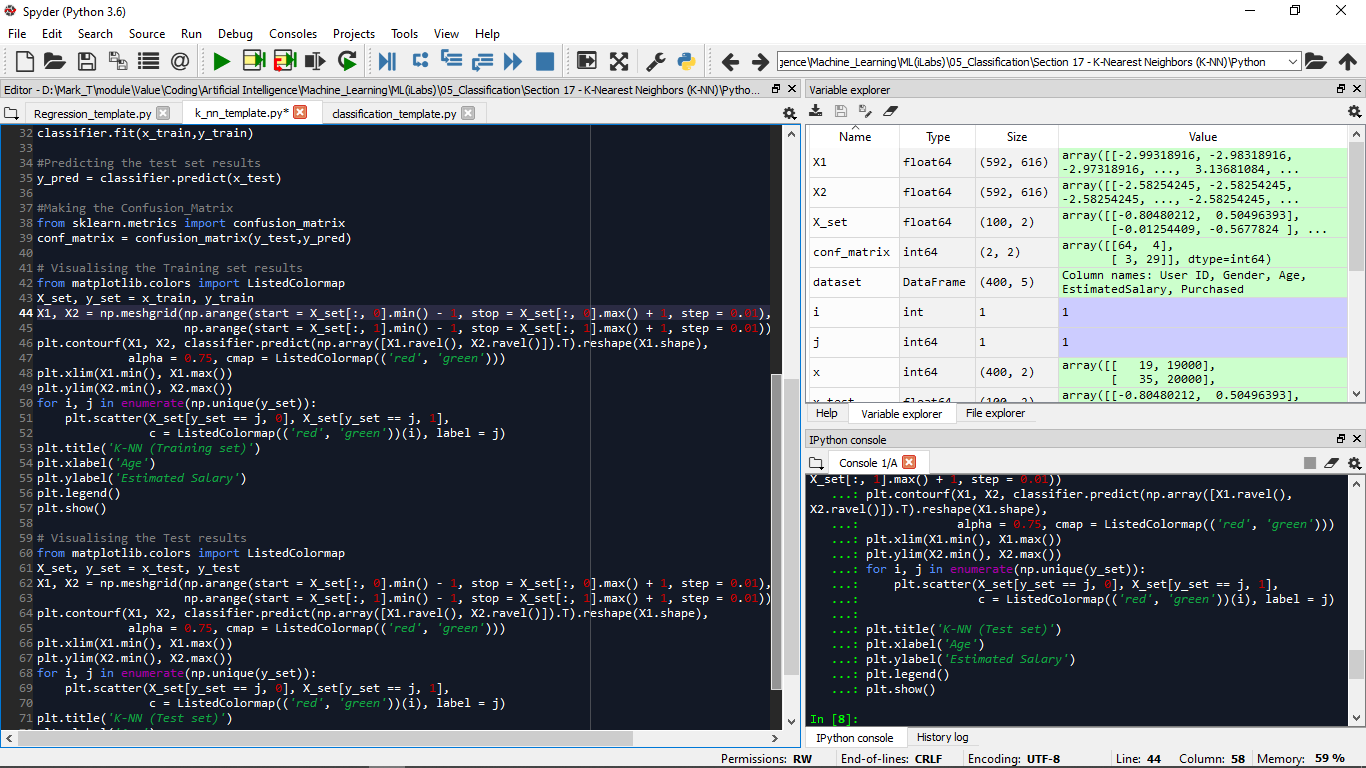
You can use other distances. Euclidean Distance is basically a magnitude of the hypotenuse formula:

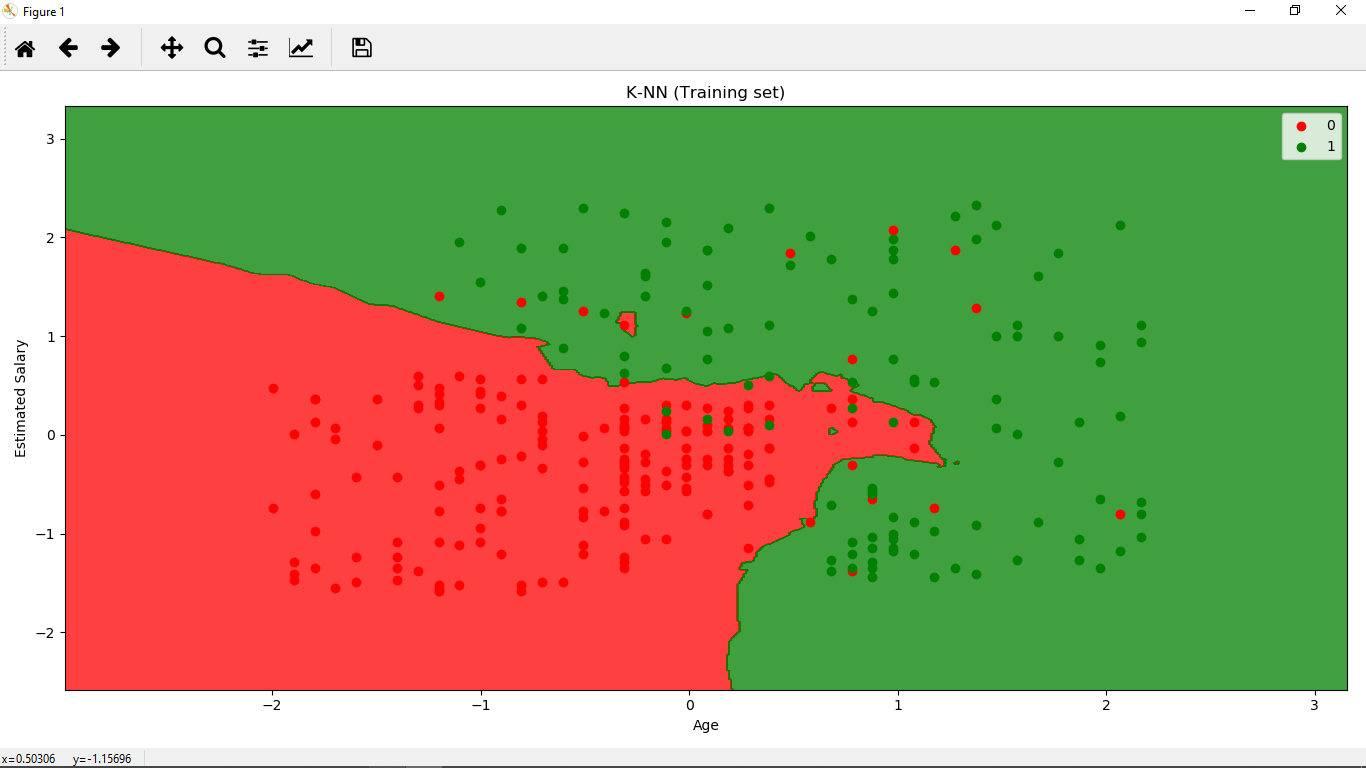


In Python

We use the sklearn.Neighbors.KNeighborsClassifier class.





For the confusion matrix, the wrong prediction percentage is 7% which makes this model seem to be better than Logistic Regression. 

For visualization, if the prediction boundary is straight then it’s a linear classifier; else, its non-linear. On observation, it looks like a country boundary.(Reason?) We have some wrong predictions because the default parameters were selected in such a way that the K-NN algorithm didn’t fit too well.

**Qns**

**What is minkowski distance, Manhattan distance.**

**Applications of K-NN**

1. ﻿﻿﻿﻿﻿﻿﻿﻿﻿Hierarchical clustering algorithms — and nearest neighbor methods, in particular — are used extensively to understand and create value from patterns in retail business data. For example, K-nearest neighbor techniques for pattern recognition are often used for theft prevention in the modern retail business. Today, a modern surveillance system is intelligent enough to analyze and interpret video data on its own, without a need for human assistance.
2. If register data indicates that a lot of customer information is being entered manually rather than through automated scanning and swiping, this could indicate that the employee who’s using that register is in fact stealing customer’s personal information. Or if register data indicates that a particular good is being returned or exchanged multiple times, this could indicate that employees are misusing the return policy or trying to make money from doing fake returns.
3. The modern systems are now able to use k-nearest neighbor for visual pattern recognition to scan and detect hidden packages in the bottom bin of a shopping cart at check-out. If an object is detected that’s an exact match for an object listed in the database, then the price of the spotted product could even automatically be added to the customer’s bill. While this automated billing practice is not used extensively at this time, the technology has been developed and is available for use.
4. Data Scientist Dan got in there, he quickly uncovered a pattern among working middle-aged male adults — they tended to visit the grocery store only during the weekends or at the end of the day on weekdays, and if they came into the store on a Thursday, they almost always bought beer. Well, when Manager Mike was armed with these facts, he quickly used this information to maximize beer sales on Thursday evenings by offering discounts, bundles.

Some of the other applications of KNN in finance are mentioned below:

Forecasting stock market: Predict the price of a stock, on the basis of company performance measures and economic data.

Currency exchange rate

Bank bankruptcies

Understanding and managing financial risk

Trading futures

Credit rating

Loan management

Bank customer profiling

Money laundering analyses

**Advantages and Disadvantages over other classifiers**

Advantages

1. No Training Period: KNN is called Lazy Learner (Instance based learning). It does not learn anything in the training period. It does not derive any discriminative function from the training data. In other words, there is no training period for it. It stores the training dataset and learns from it only at the time of making real time predictions. This makes the KNN algorithm much faster than other algorithms that require training e.g. SVM, Linear Regression etc.

2. Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm.

3. KNN is very easy to implement. There are only two parameters required to implement KNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.)

Disadvantages of KNN

1. Does not work well with large dataset: In large datasets, the cost of calculating the distance between the new point and each existing points is huge which degrades the performance of the algorithm.

2. Does not work well with high dimensions: The KNN algorithm doesn't work well with high dimensional data because with large number of dimensions, it becomes difficult for the algorithm to calculate the distance in each dimension.

3. Need feature scaling: We need to do feature scaling (standardization and normalization) before applying KNN algorithm to any dataset. If we don't do so, KNN may generate wrong predictions.

4. Sensitive to noisy data, missing values and outliers: KNN is sensitive to noise in the dataset. We need to manually impute missing values and remove outliers. A data point of category two may have a weakness in characteristic and exist within datapoints of category 1 hence affecting the classification

**How to make the model more robust or correct**

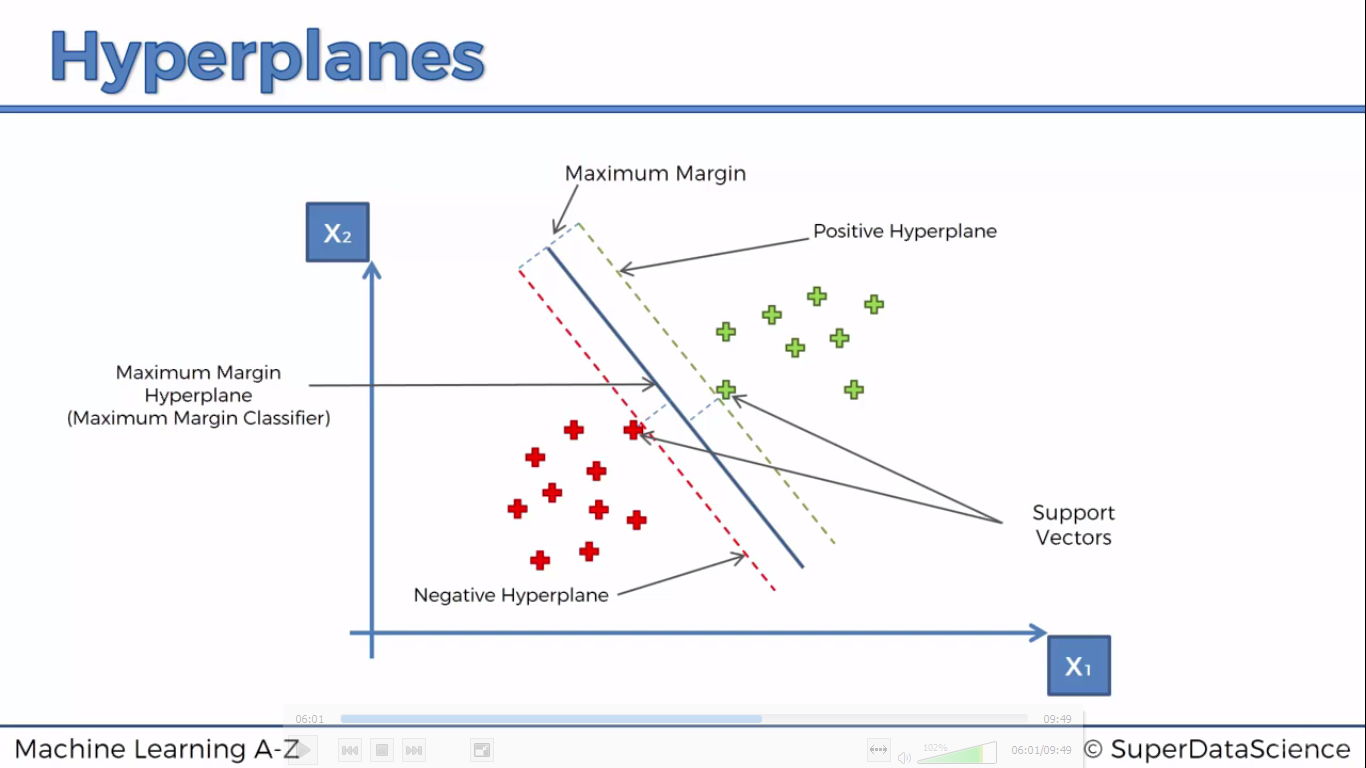
**How does the sklearn.Neighbors.KNeighborClassifier class work?**

**Qns and Answers**

I need examples of questions to practixe using each model

SUPPORT VECTOR MACHINES (SVMs)

The SVM algorithm seeks to place a separator to classify data so it makes use of the maximum margin due to the two support vectors as shown to come up with this separator(Maximum Margin Hyperplane).



The algorithm only needs these two support vectors. In reality if you are to think about this, these support vectors are like an apple which almost looks like an orange and vice versa(closest data to the hyperplane). This means the algorithm makes use of extreme cases(close to the other type) to construct its analysis and this at times can have consequences such as better performance.