Classification of Stars and Galaxies

Saarthak Gupta - June 14, 2019

SPACE HAS BILLIONS OF STARS AND GALAXIES. THIS ALGORITHM USES MACHINE LEARNING TO PREDICT IF AN OBJECT IS A STAR OR GALAXY.

THE DATASET

- The dataset used for this project is called Sloan Digital Sky Survey DR14.
- It was downloaded from here.
- The dataset contains 10,000 total items. However, for the purpose of this program, the data pertaining to Quasars was removed. This left us with 4998 Galaxies and 4152 stars.
- The dataset has 14 features for each object. (objid, specobjid , rerun and classes were dropped as they do not help us in predicting the type of the object.)
- The dataset has 5 features, which represent different wavelengths of light. (u, g, r, i and z):

Ultraviolet				
U	365 nm	66 nm	u, u', u*	"U" stands for ultraviolet.
Visible				
B	445 nm	94 nm	b	"B" stands for blue.
V	551 nm	88 nm	v, v'	"V" stands for visual.
G[3]	464 nm	128 nm	g'	"G" stands for green.
R	658 nm	138 nm	r, r', R', R_c, R_e, R_j	"R" stands for red.
Near-Infrared				
I	806 nm	149 nm	i, i', I_c, I_e, I_j	"I" stands for infrared.
Z	900 nm[4]		z, z'	
Y	1020 nm	120 nm	y	
J	1220 nm	213 nm	J', J_s	

.

- **Redshift** is the displacement of the spectrum of an astronomical object toward longer (red) wavelengths.

- In **astronomy**, **declination** (abbreviated **dec**; symbol δ) is one of the two angles that locate a point on the celestial sphere in the equatorial coordinate system, the other being hour angle.
- **Right ascension** (ra) is the angular distance of a particular point measured eastward along the celestial equator from the Sun at the March equinox to the point above the earth in question.
- **Run, rerun, camcol and field** are features which describe a field within an image taken by the SDSS. A field is basically a part of the entire image corresponding to 2048 by 1489 pixels. A field can be identified by: - run number, which identifies the specific scan, - the camera column, or "camcol," a number from 1 to 6, identifying the scanline within the run, and - the field number. The field number typically starts at 11, and can be as large as 800 for particularly long runs. - An additional number, rerun, specifies how the image was processed.
- Each spectroscopic exposure employs a large, thin, circular metal **plate** that positions optical fibers via holes drilled at the locations of the images in the telescope focal plane.
- **Modified Julian Date (mjd)**, used to indicate the date that a given piece of SDSS data (image or spectrum) was taken.
- The SDSS spectrograph uses optical fibers to direct the light at the focal plane from individual objects to the slithead. Each object is assigned a corresponding **fiberID**.

MACHINE LEARNING ALGORITHMS USED

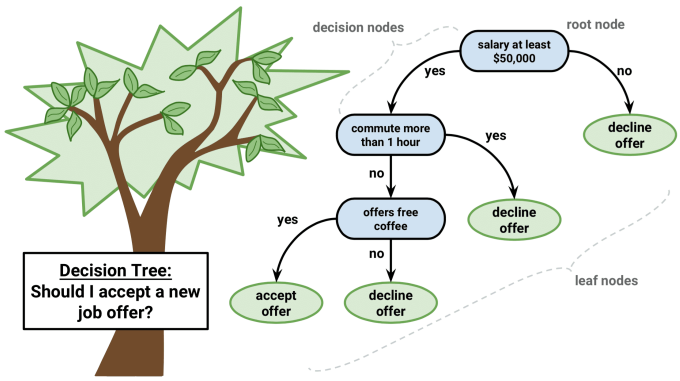
What is Machine Learning?

Machine learning is the scientific study of algorithms and statistical models that computer systems use in order to perform a specific task effectively without using explicit instructions, relying on patterns and inference instead.

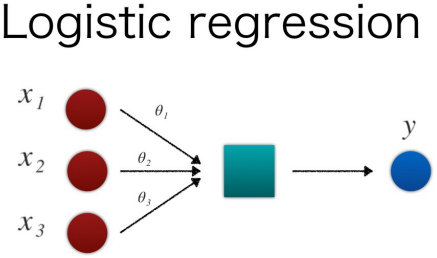
For this project, 6 Machine Learning Algorithms were explored.

ALGORITHMS:

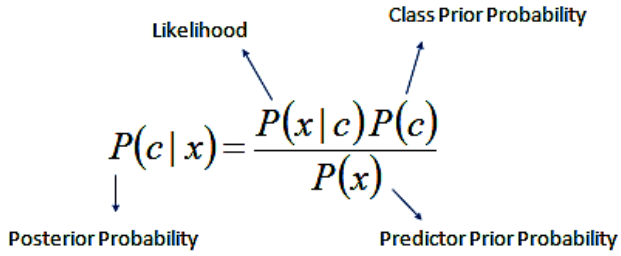
1. **Decision Tree Classifier:** In computer science, Decision tree learning uses a decision tree to go from observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used in statistics, data mining and machine learning.



2. **Logistic Regression:** Logistic regression is a classification algorithm used to assign observations to a discrete set of classes.

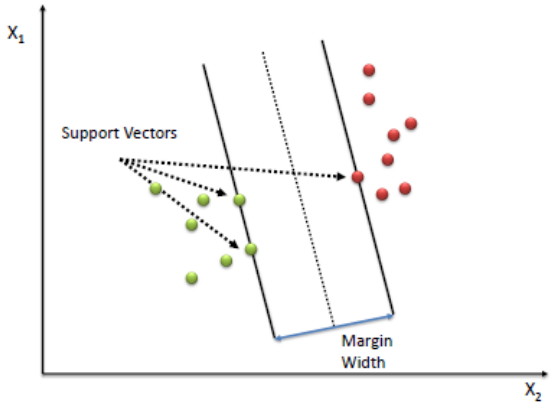


3. **Naive Bayes:** In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

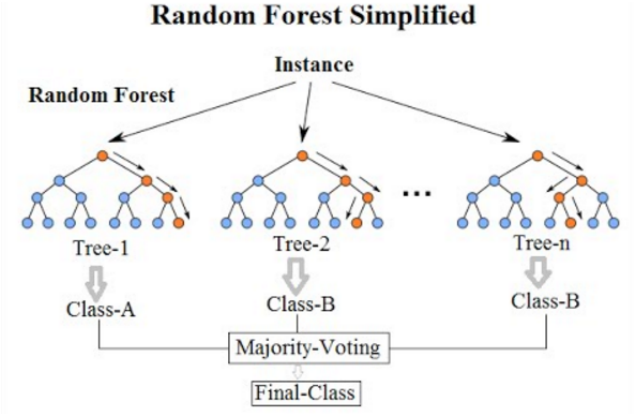


$$P(c \mid \mathrm{X}) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

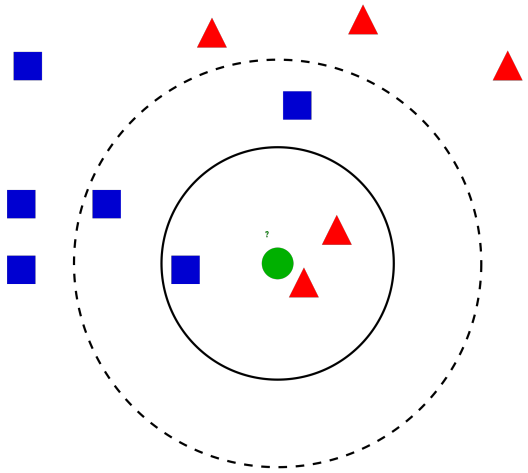
4. **Support Vector Machines:** The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space(N—the number of features) that distinctly classifies the data points.



5. **Random Forest:** Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees.

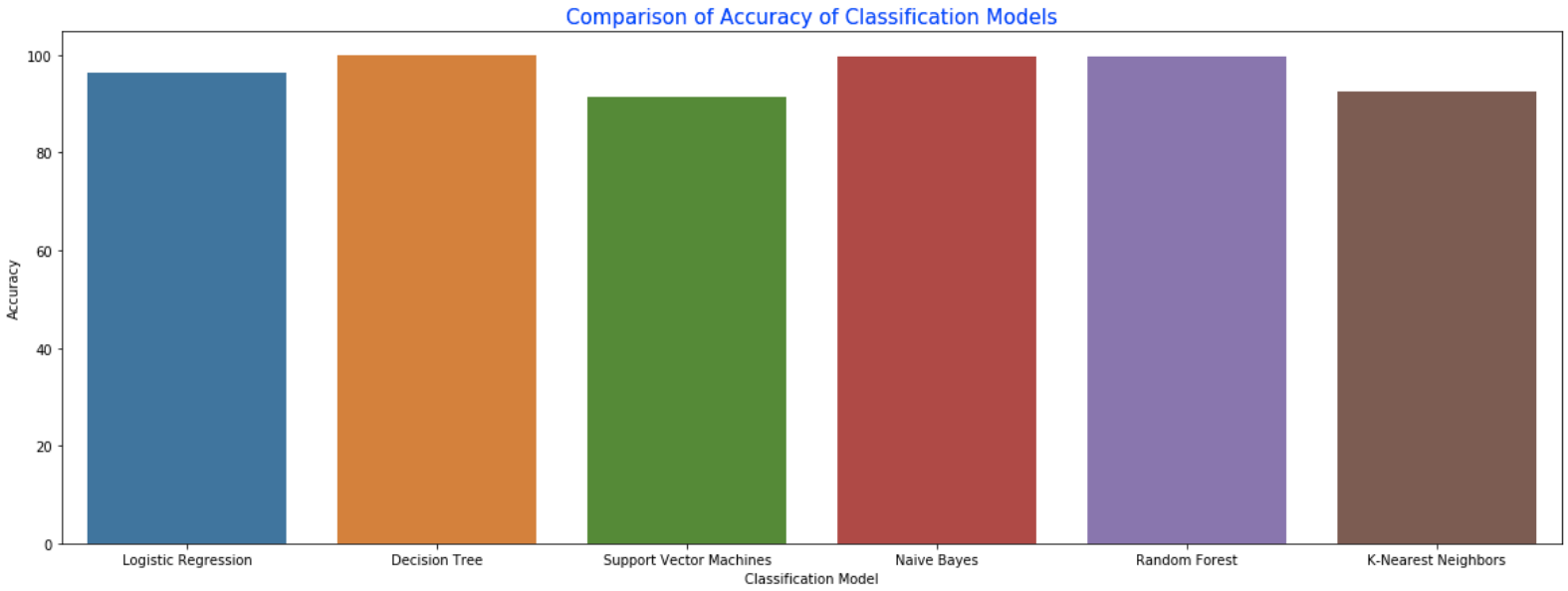


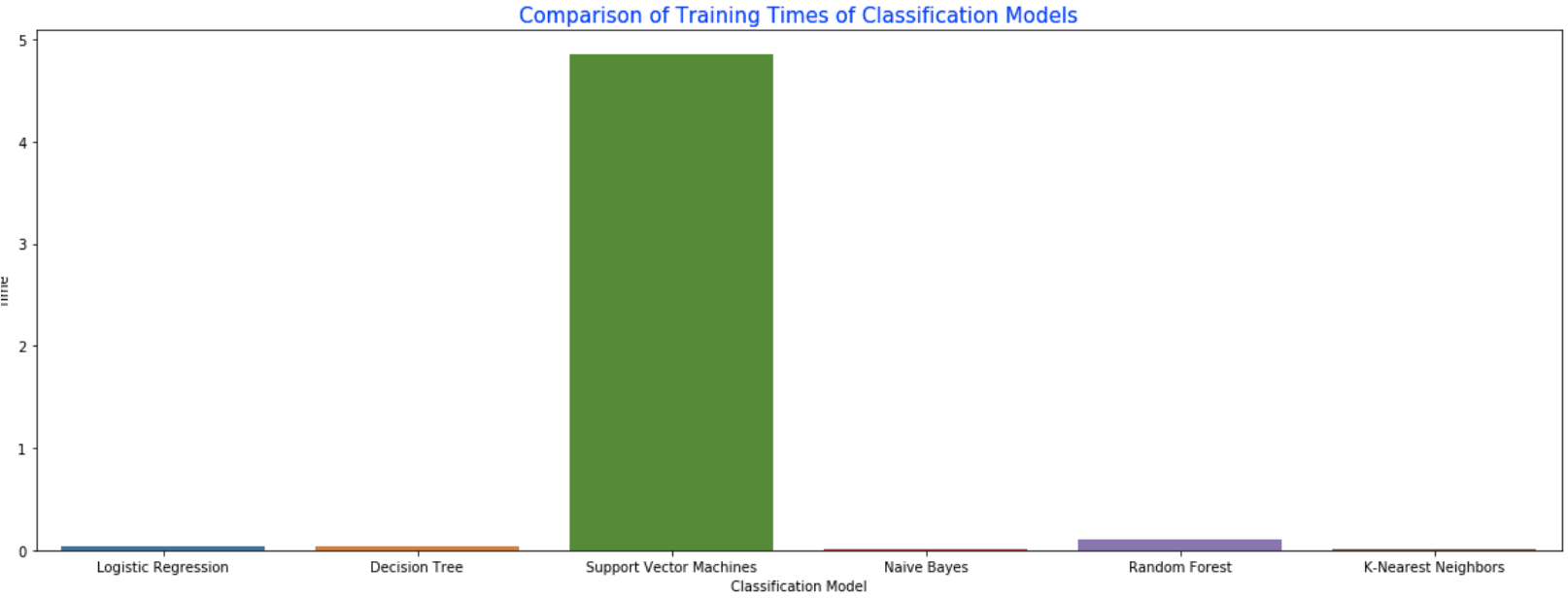
6. **K-Nearest Neighbors:** In pattern recognition, the *k*-nearest neighbors algorithm (*k*-NN) is a non-parametric method used for classification and regression.[1] In both cases, the input consists of the *k* closest training examples in the feature space.



COMPARISON OF DIFFERENT ALGORITHMS

Algorithm	Accuracy	Training Time
Logistic Regression	96.39%	0.049 s
Decision Tree	99.89%	0.042 s
Support Vector Machines	91.42%	5.289 s
Naive Bayes	99.56%	0.006 s
Random Forest	99.61%	0.104 s
K - Nearest Neighbors	92.45%	0.007 s



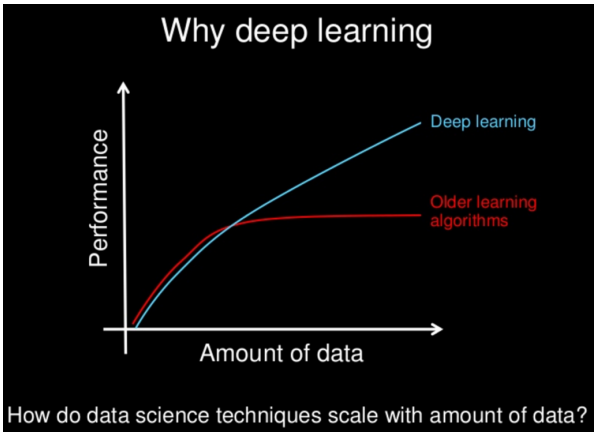


From this data, it was concluded that the **Decision Tree Algorithm** is best suited for the classification of this particular data set.

Deep Learning and Neural Networks

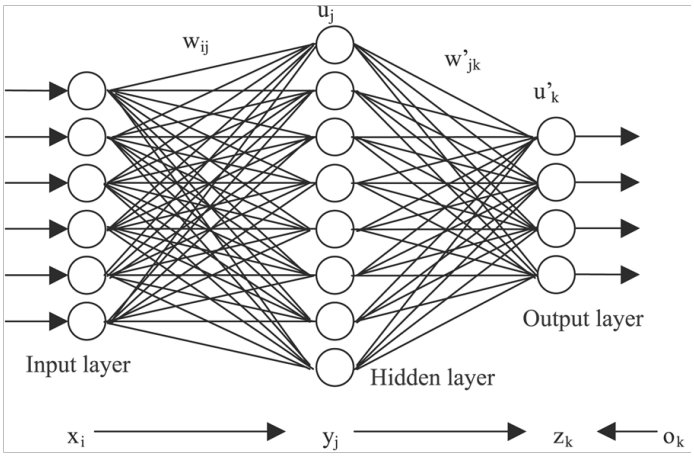
What is Deep Learning?

Deep learning (also known as **deep structured learning** or **hierarchical learning**) is part of a broader family of machine learning methods based on artificial neural networks. Learning can be supervised, semi-supervised or unsupervised. This project use the Keras deep learning module, which is built on top of Google's TensorFlow.



What are Neural Networks?

Artificial neural networks or **connectionist systems** are computing systems that are inspired by, but not necessarily identical to the biological neural networks that constitute animal brains. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules.



This is the Neural Network used for this project:

```
#Neural Network
model = keras.Sequential([
        keras.layers.Dense(14 , activation=tf.nn.relu),
        keras.layers.Dense(128 , activation=tf.nn.relu),
        keras.layers.Dense(1 , activation=tf.nn.sigmoid)
        ])

model.compile(optimizer = 'adam',
              loss = 'binary_crossentropy',
              metrics = ['accuracy'])

start = time.time()
model.fit(x_train.values , y_train , epochs=10 , batch_size=20)
end=time.time()

scores = model.evaluate(x_test , y_test)
```

Training And Testing The Network:

```
Epoch 1/10
7320/7320 [==============================] - 3s 380us/sample - loss: 0.3968 - acc: 0.8422
Epoch 2/10
7320/7320 [==============================] - 2s 207us/sample - loss: 0.1915 - acc: 0.9321
Epoch 3/10
7320/7320 [==============================] - 1s 204us/sample - loss: 0.1124 - acc: 0.9671
Epoch 4/10
7320/7320 [==============================] - 1s 205us/sample - loss: 0.0794 - acc: 0.9786
Epoch 5/10
7320/7320 [==============================] - 2s 224us/sample - loss: 0.0632 - acc: 0.9847
Epoch 6/10
7320/7320 [==============================] - 2s 239us/sample - loss: 0.0536 - acc: 0.9872
Epoch 7/10
7320/7320 [==============================] - 1s 204us/sample - loss: 0.0440 - acc: 0.9899
Epoch 8/10
7320/7320 [==============================] - 2s 224us/sample - loss: 0.0395 - acc: 0.9908
Epoch 9/10
7320/7320 [==============================] - 1s 203us/sample - loss: 0.0374 - acc: 0.9914
Epoch 10/10
7320/7320 [==============================] - 1s 202us/sample - loss: 0.0341 - acc: 0.9923
1830/1830 [==============================] - 1s 597us/sample - loss: 0.0371 - acc: 0.9880

Training Time: 26.435540199279785
```

Accuracy: 98.80% [10 epochs, 3 dense layers (units = 14-128-1)]

Conclusion

After training and evaluation, the system asks the user to enter values for the various astronomical features. Using this, it predicts whether the supplied data corresponds to a Star or Galaxy.

```
ra= 183.531

dec= 0.089693

u= 19.4741

g= 17.0424

r= 15.947

i= 15.5034

z= 15.2253

run= 752

camcol= 4

field= 267

redshift= -0.00000896

plate= 3306

mjd= 54922

fiberid= 491
        ra       dec        u        g  ...     redshift plate    mjd fiberid
0  183.531  0.089693  19.4741  17.0424  ...  -0.00000896  3306  54922     491

[1 rows x 14 columns]
Decision Tree Prediction:
[1]
GALAXY
Prediction Probability:
[[100.   0.]]
```

The Algorithm successfully predicts that the supplied data is from a Galaxy.