Praxis Business School
CELEBRATE YOUR WORTH

DNA Sequence Classification

Using Machine Learning

Ctrl + Click here to visit the Notebook for this project

Contents

Introduction to DNA

DNA sequencing is the process of determining the sequence of nucleotides (As, Ts, Cs, and Gs) in a piece of DNA.

3 Billion

The human genome contains about 3 billion base pairs that spell out the instructions for making and maintaining a human being.

In the DNA double helix, the four chemical bases always bond with the same partner to form "base pairs." Adenine (A) always pairs with thymine (T); cytosine (C) always pairs with guanine (G). This pairing is the basis for the mechanism by which DNA molecules are copied when cells divide, and the pairing also underlies the methods by which most DNA sequencing experiments are done.

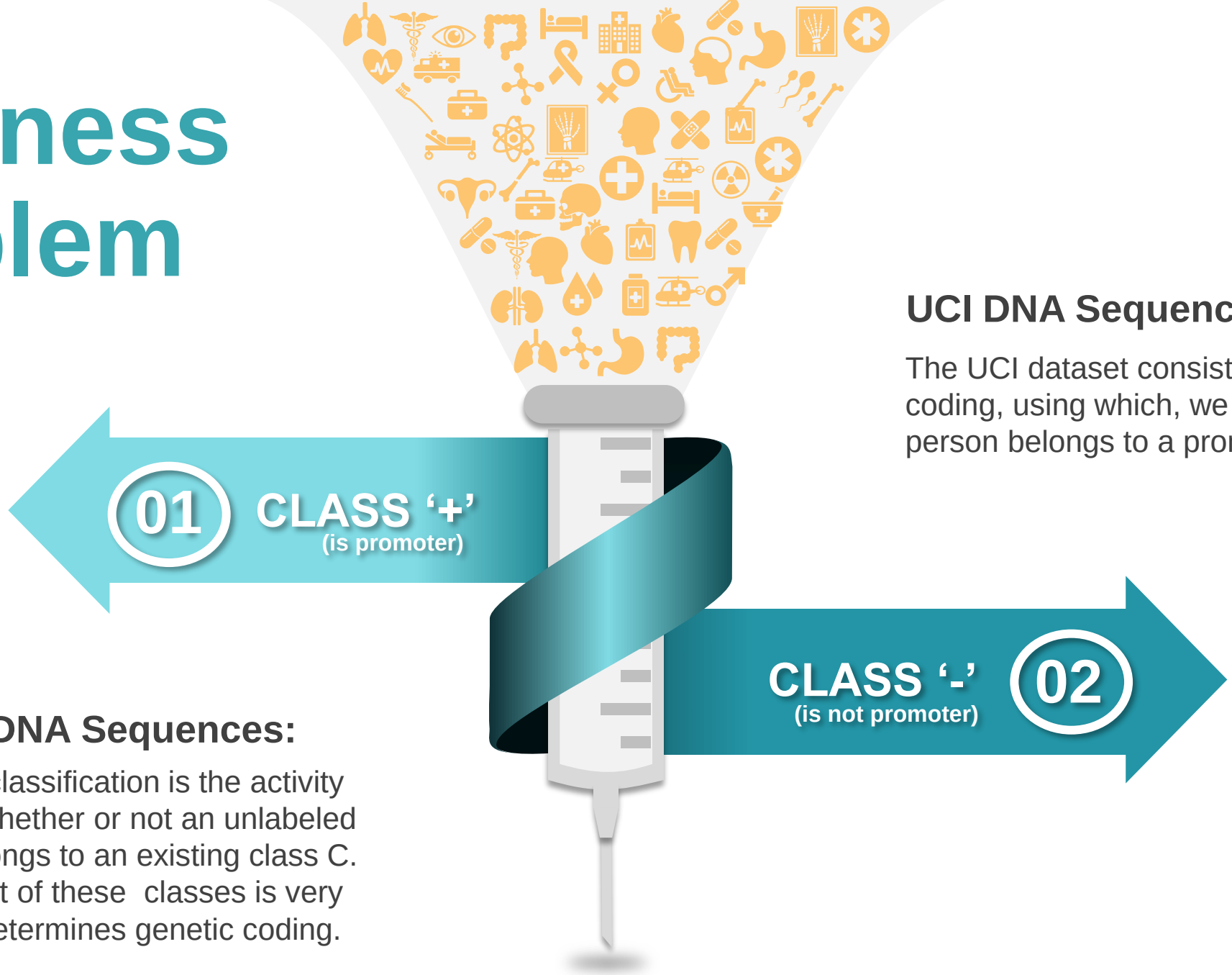
How new is DNA Sequencing

Since the completion of the Human Genome Project (1990-2003), technological improvements and automation have increased speed and lowered costs to the point where individual genes can be sequenced routinely, and some labs can sequence well over 100,000 billion bases per year, and an entire genome can be sequenced for just a few thousand dollars.

What do improvements in DNA sequencing mean for human health?

Researchers now are able to compare large stretches of DNA - 1 million bases or more - from different individuals quickly and cheaply. Such comparisons can yield an enormous amount of information about the role of inheritance in susceptibility to disease and in response to environmental influences. In addition, the ability to sequence the genome more rapidly and cost-effectively creates vast potential for diagnostics and therapies.

Business Problem



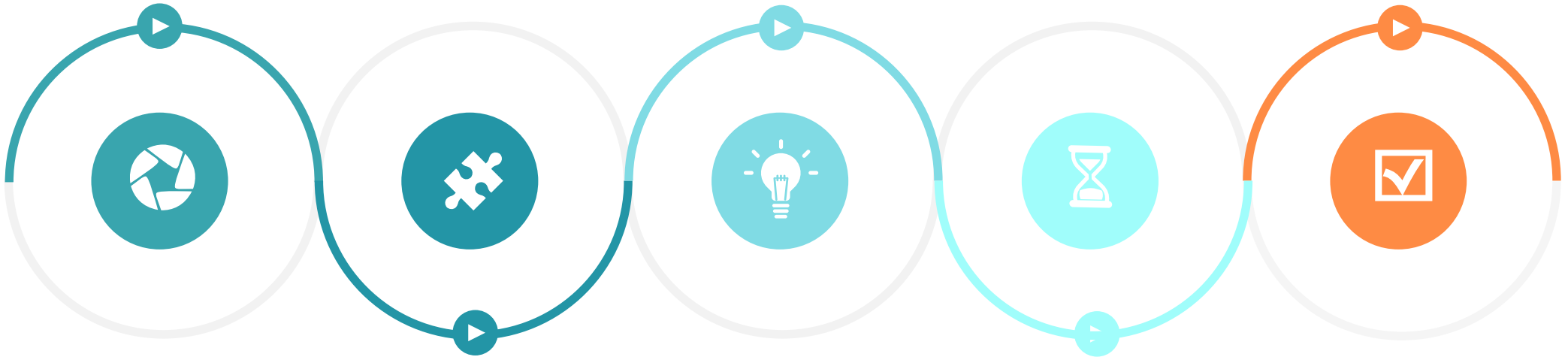
UCI DNA Sequence Data:

The UCI dataset consists of this genetic coding, using which, we determine if a person belongs to a promoter class or not.

Classifying DNA Sequences:

DNA sequence classification is the activity of determining whether or not an unlabeled sequence S belongs to an existing class C. The arrangement of these classes is very important as it determines genetic coding.

Training Methodology



Data-set

The data is a 1990 generated collection of 106 instances across 58 attributes. All the instances are labelled with binary labels i.e. ‘+’ and ‘-’

Train-test split

The data is split into training and test data for validation of the model later.
The training data size is set to 25% of the original data

Model Selection

To select which model performs best for the given dataset, we train a host of different models, like Logistic Regression, SVM, Naïve Bayes, Neural Networks etc.

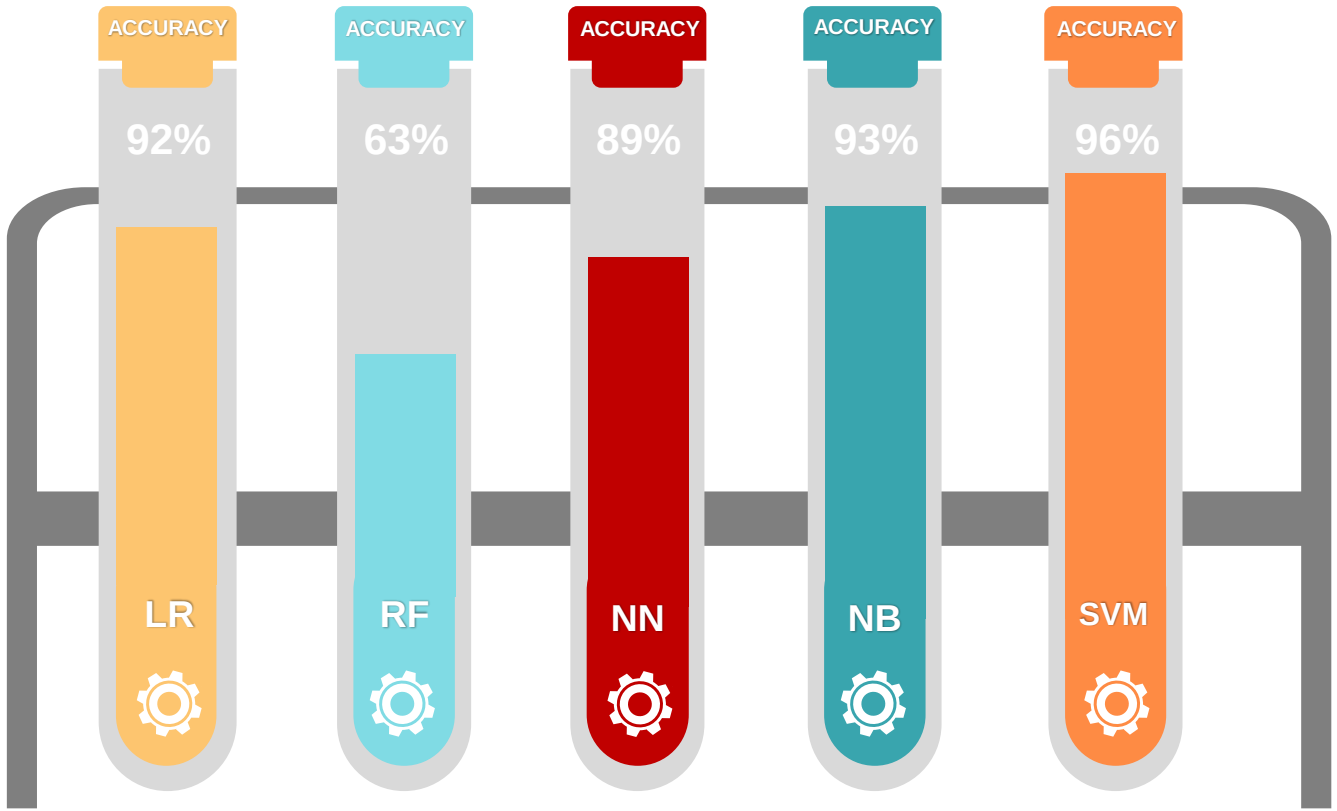
Model Training

The models are trained one-by one with the training data. The model parameters are updated using *gradient descent*.

Model Performance

Model performances are compared using the metric of Accuracy Score, which tells us the proportion of correct classifications made by the model.

Modelling Techniques Deployed



01

Logistic Regression

Logistic regression models the probabilities for classification problems with two possible outcomes. It's an extension of the linear regression model for classification problems.

02

Random Forest

Random forests is an ensemble learning method for classification, that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees.

03

Neural Networks

Artificial neural networks are computing systems inspired by the biological neural networks that "learn" to perform tasks by considering examples, generally without being programmed with task-specific rules.

04

Naïve Bayes

Naive Bayes classifiers are a collection of algorithms based on Bayes' Theorem where every pair of features being classified is independent of each other.

05

Support Vector Machines

Support-Vector Machines are supervised learning models that use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form.

Model Selection

	Training Accuracy	Std. Deviation	Testing Accuracy
K-Nearest Neighbor	82%	0.1139	78%
Gaussian Process	87%	0.0561	89%
Random Forest	61%	0.0682	63%
Neural Networks	88%	0.0968	89%
Naïve Bayes	84%	0.1375	93%
Support Vector Machine	85%	0.1089	96%

Highest Testing Accuracy

The Support Vector Machine provides the highest testing accuracy. This means it correctly classified the highest number of samples in the test data

Lowest Std. Deviation

The Gaussian Process gave the smallest amount of Std. Deviation in it's test results. This means it has the most consistent accuracy score

Highest Training Accuracy

The Neural Networks offer highest amount of testing accuracy. This mean that the neural network might be better suited for a number of new unseen data.

Hybrid Results

Gaussian process model offers good amount of training accuracy and low std. deviation, coupled with relatively good testing accuracy.

Conclusion

Business Problem

We are able to build machine learning models that can classify any new incoming data about a DNA sequence into the Binary Classes with relatively high accuracy.

Technical Problem

We are able to achieve good performance in different metrics using different models. We can do a critical analysis of these models and compare the performances on different test data.

Which Model to select for Deployment?

Generally it is observed that there exist no such model that offers, train once, fit all capability. Different models perform differently for distinct data-sets. It is up to the end user to decide what he/she desires out of the model. In our case:

- If accuracy is of paramount importance, then Support Vector Machines is the choice.
- If consistency of results I required, then Gaussian Process should be chosen.
- If the end user requires a model that is scalable, then Neural Networks fit the role.
- If there is requirement of a model that performs relatively good on all parameters but does not excel in any, then Gaussian process can be used.

Thank You