www.multisoftvirtualacademy.com
info@multisoftvirtualacademy.com
+91-8130666206 / 209

multisoft

# DATA ANALYTICS & MACHINE LEARNING

# About Me

- Priyanka Talla
- Working in Analytics from past 4 years.

  - 2 years in SAS Development
- Background in BI, Developer and Analyst.

# What we will learn

1. Need of Data Science
2. What is Data Science
3. Use case of Data Science
4. Business Intelligence vs. Data Science
5. Tools used in Data Science
6. Life cycle of Data Science

# NEED OF DATASCIENCE

Problem

Data Flow

Unstructured Data

Data Storage

Lack of Predictive Analytics

Lack of Scientific insights

What we can do with Data Science

- Decision Making
- Prediction
- Pattern Discovery

Then

- Structured Data
- Data Warehouse
- Traditional BI
- Predetermined Report Only

NOW

- unstructured & structured data
- Hadoop
- Data Science Algorithms
- Scientific Discovery

# You can use Data Science to

- Recommend the right product to the right customer to enhance business.
- Predict the characteristics of high LTV customers and helps in customer segmentation.
- Build intelligence and ability in machines.
- Predict fraudulent transactions beforehand.
- Perform sentiment analysis to predict the outcome of elections.

# WHAT IS DATA SCIENCE

- Data science is a blend of various tools, algorithms, and machine learning principles with the goal to discover hidden patterns from the raw data.
- Data science is primarily used to make decisions and predictions.

# Lets understand data science with some examples

Sports Analytics & Data Science

-Winning the game with methods and models

- ➢ Basketball teams are using data for tracking team strategies and outcome of matches.
- ➢ Below parameters will be used for model building
- - Average pass time of ball.
- - Number of successful passes.
- - Speed and accuracy of successful baskets.
- - Area of court the player on average is shadowing.
- ➢ Models build on the basis of data science algorithms help in pattern discovery of player game.

# ECOMMERCE

- Amazon has huge amount of consumer purchasing data.
- The data consists of consumer demographics (age,gender,location), purchasing history, past browsing history.
- Based on this data, Amazon segments its customers, draws a pattern and recommends the right product to the right customer at the right time.

# GOOGLE CAR

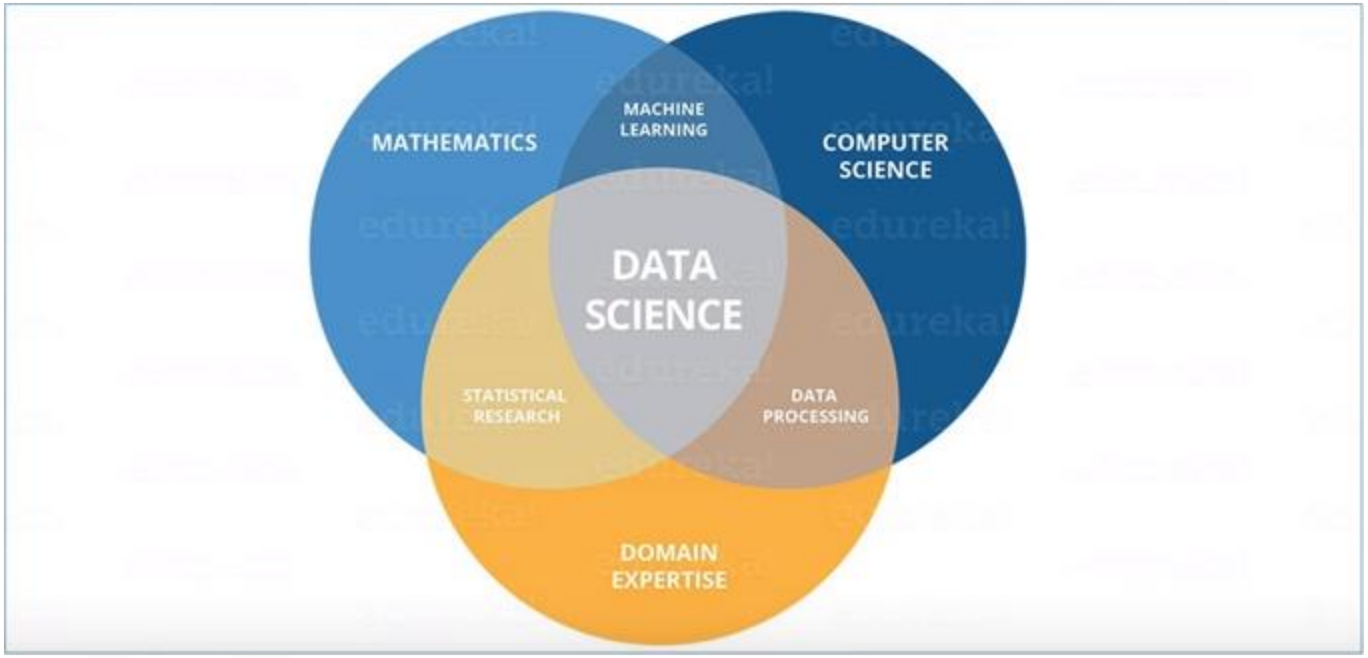
- Google self driving car is a smart, driverless car.
- It collects data from environment through sensors.
- Takes decisions like when to speed up, when to speed down, when to overtake and when to turn.

# USE CASES OF DATA SCIENCE

1. Travel
   - Dynamic pricing
   - Predicting flight delay
2. Marketing
   - cross selling
   - predicting lifetime value of customer
3. Healthcare
   - Disease prediction
   - Medication effectiveness
4. Social media
   - Sentiment Analysis
   - Digital Marketing
5. Sales
   - Discount offering
   - Demand forecasting
6. Automation
   - Self diving cars
   - pilotless aircrafts, drones

7.Credit & Insurance
   - Claims prediction
   - Fraud & risk detection

# SKILLS OF DATASCIENTIST



# ROLE OF A DATA SCIENTIST

The Data Scientist will be responsible for designing and creating processes and layouts for complex, large-scale data sets used for modeling, data mining and research purposed.

RESPONSIBILITIES

- Selecting features, building and optimizing classifiers using machine learning techniques.
- Data mining using state-of-the-art methods.
- Extending company's data with third party sources of information when needed.
- Processing, Cleansing and verifying the integrity of data for analysis.
- Building predictive models using Machine Learning algorithms.

# BI Vs. Data Science

| Characteristics | Business Intelligence | Data Science |
|---|---|---|
| Perspective | Looking Backward | Looking Forward |
| Data Sources | Structured (Usually SQL, often Data Warehouse) | Both Structured and Unstructured (logs, cloud data, SQL, No SQL, text) |
| Approach | Statistics and Visualization | Statistics, Machine Learning, Graph Analysis, Neuro- linguistic Programming (NLP) |
| Focus | Past and Present | Present and Future |
| Tools | Pentaho, Microsoft BI, QlikView, R | RapidMiner, BigML, Weka, R |

# Tools Used In Data Science

Commonly used tools by Data Scientists

1. Data Analysis
   - R
   - Spark
   - Python
   - SAS
2. Data Warehousing
   - Hadoop
   - SQL
   - Hive
3. Data Visualization
   - R
   - Tableau
   - Raw
4. Machine Learning
   - Spark
   - Mahout
   - Azure ML Studio

PROBLEM :

What if we could predict the occurrence of diabetes and take appropriate measures beforehand to prevent it?

SOLUTION :

Definitely! Let me take you through the steps to predict the vulnerable patients.

# LIFECYCLE OF DATA SCIENCE

1. Discovery
2. Data Preparation
3. Model Planning
4. Model Building
5. Operationalize
6. Communicate Results

# 1. DISCOVERY

- Discovery involves acquiring data from all the identified internal and external sources that can help answer the business question.
- This data could be

  - logs from webservers

  - social medial data

  - census datasets

  - data streamed from online sources via APIs

# Problem

Doctor gets this data from the medical history of the patient.

Attributes:

Npreg – number of times pregnant

Glucose – Plasma glucose concentration

Bp – blood pressure

Skin – Triceps skinfold thickness

Bmi – Body mass index

Ped – Diabetes pedigree function

Age – Age

Income – income

# 2. DATA PREPARATION

- The data can have a lot of inconsistencies like missing values, blank columns, abrupt values and incorrect data format which need to be cleaned.
- It is required to explore, preprocess and condition data prior to modelling
- This will help you to spot the outliers and establish a relationship between the variables.

# 3. MODEL PLANNING

- Here, we determine the methods and techniques to draw the relationships between variables.
- Apply Exploratory Data Analytics (EDA) using various statistical formulas and visualization tools.

COMMON TOOLS FOR MODEL PLANNING

- SQL (Analysis Services)
- R
- SAS

Use of visualization techniques like histograms, line graphs, box plots to get a fair idea of the distriution of data.

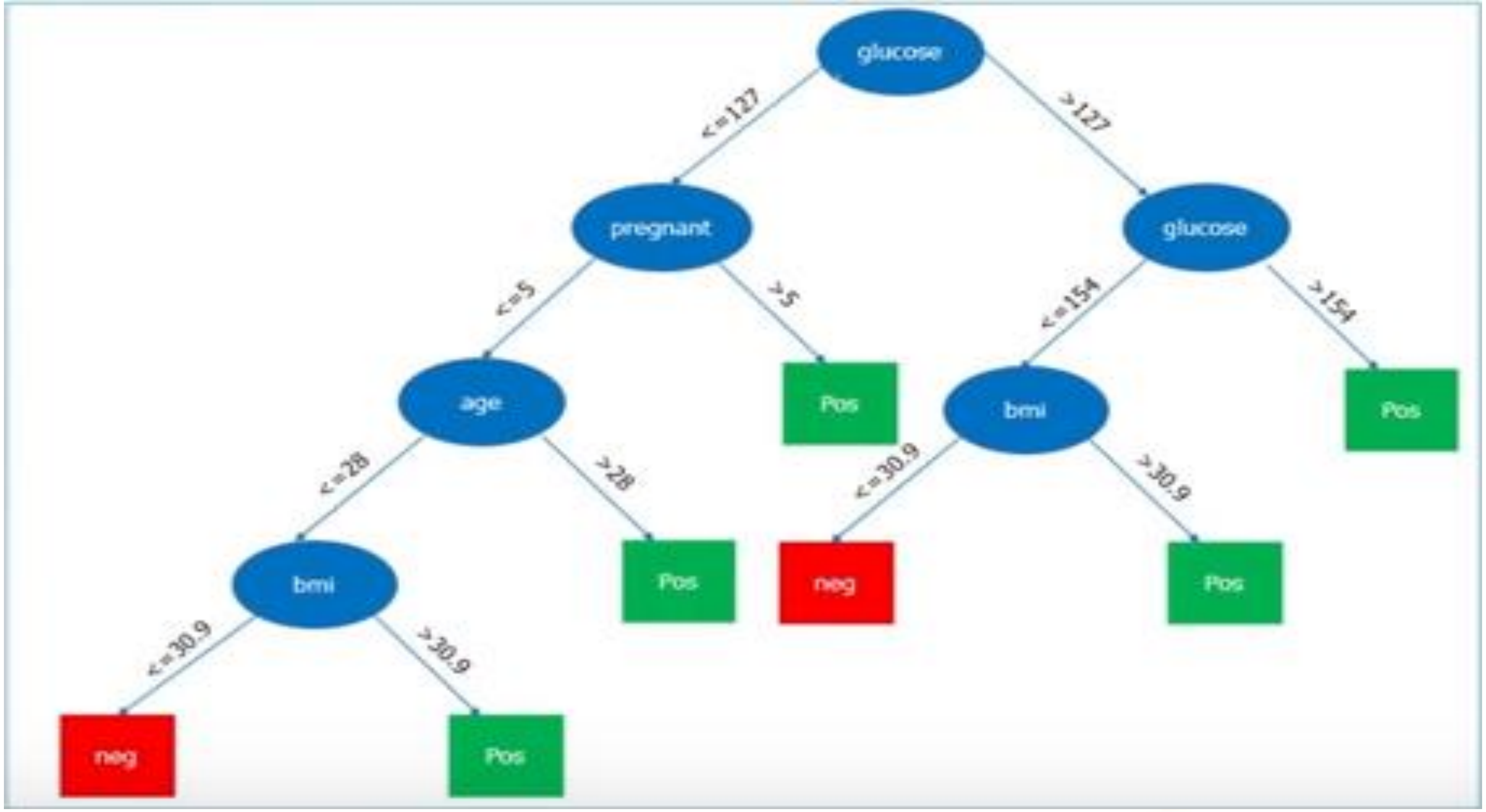
# 4. MODEL BUILDING

- Develop datasets for training and testing purposes.
- Consider whether existing tools will suffice for running the models.
- Analyze various learning techniques like classification, association and clustering to build the model.

COMMON TOOLS FOR ODEL BUILDING

- SAS Enterprise Miner
- Weka
- SPCS Modeler
- R
- Python
- Statistica

Below is a decision tree based on different attributes.



# 5. OPERATIONALIZE

- Deliver final reports, briefings, code and technical documents.
- Implement pilot project in a real-time production environment.
- Look for performance constraints if any.

# 6. COMMUNICATE RESULTS

- Identify all the key findings and communicate to the stakeholders.
- Explaining the model and result to medical authorities.
- Determine if the results of the project are a success or a failure based on the criteria developed.

# FINAL RESULT

- Diabetes Positive set :
    - glucose>154
    - glucose>127&<=154+bmi>30.9
    - glucose<=127+pregnant>5
    - glucose<=127+pregnant<=5+age>28
    - glucose<=127+pregnant<=5+age<=28+bmi>30.9
- Diabetes Negative set :
    - glucose>154
    - glucose>127&<=154+bmi<=30.9
    - glucose<=127+pregnant<=5+age<=28+bmi<=30.9
- We can use this decision tree result to know whether the patient is vulnerable to diabetes or not.

# Machine Learning Algorithms

1. What is an algorithm?
2. What is Machine Learning?
3. How is a problem solved using Machine Learning?
4. Types of Machine Learning
5. Machine Learning Algorithms

# WHAT IS AN ALGORITHM

- To tell a computer what it has to do, you need a program.
- A program is nothing but logic in some language's syntax
- Logic

  - This logic is what an algorithm is

A process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer

# WHAT IS MACHINE LEARNING?

Machine learning is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programed. Machine learning focuses on the development of computer programs that can change when exposed to new data.

# MACHINE LEARNING TYPES

CATEGORIES OF ALGORITHMS :

Types of Learning

1. Supervised Learning
2. Reinforcement Learning
3. Unsupervised Learning

# Supervised Learning

Supervised learning is a type of machine learning algorithm that uses a known dataset (called the training dataset) to make predictions. The training dataset includes input data and response values. From it, the supervised learning algorithm seeks to build a model that can make predictions of the response values for a new dataset.

# UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.

# REINFORCEMENT LEARNING

Reinforcement learning is an area of machine learning inspired by behaviorist psychology, concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

# HOW A PROBLEM IS SOLVED USING MACHINE LEARNING

1. Is this A or B?

    - Classification Algorithm

2. Is this weird?

    - Anomaly Detection Algorithm

3. How much or how many?

    – Regression Algorithm

4. How is this organized?

    – Clustering Algorithm

5. What should I do next?

    – Reinforcement Learning

# MACHINE LEARNING ALGORITHMS

1. CLASSIFICATION ALGORITHM

Classification algorithms are used to classify a record.

It is used for questions which can have only a limited number of answers.

For example:

Is it cold?

- Yes or no

Will you go to work today?

- yes, no or maybe

When you have only two choices, its called 2 class classification, if you have more then 2 choices it call Multi Class Classification.

# ANOMALY DETECTION ALGORITHMS

- It analyzes a certain pattern and alerts you whenever there is change in the pattern.

Example :

In real life, your credit card company uses these anomaly detection algorithms, and flag any transaction, which is not usual as per your transaction history

REGRESSION ALOGORITHMS

- Regression Algorithms are used to calculate numeric values

Example :

- What will the temperature be tomorrow?
- How much discount can you give on a particular item?

# CLUSTERING ALGORITHMS

- It helps you understand the structure of a dataset.
- These algorithms separates the data into groups or clusters, to ease out the interpretation of the data.
- By understanding how data is organized, you can better predict the behavior of a particular event.

# REINFORCEMENT ALGORITHM

- These algorithms were designed as to how brains of humans or rats respond to punishments and rewards, they learn from outcomes, and decide on next action.
- They are good for systems which have to make lot of small decisions without human guidance.

Example :

1. A system which plays chess.
2. A temperature control system, when it has to decide whether temperature should be increased or decreased.

multisoft

# Thank you

www.multisoftvirtualacademy.com

info@multisoftvirtualacademy.com

+91-8130666206 / 209

CCNA® and other certifications are registered trademarks of Cisco® Systems, Inc