

Data Analytics in Cyber Security

(CT115-3-M) (Version E)

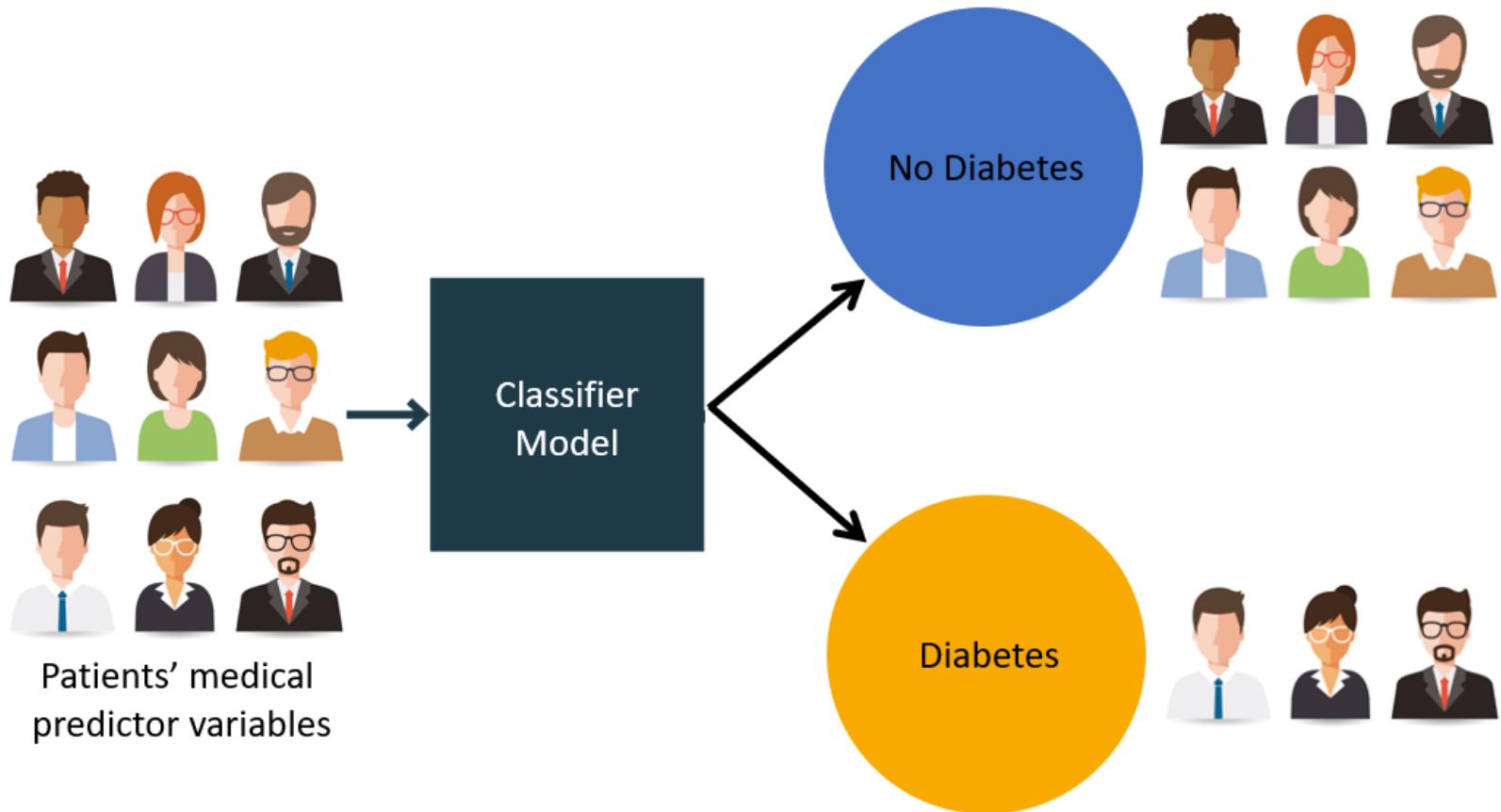
Machine Learning Pipeline

TOPIC LEARNING OUTCOMES

At the end of this topic, you should be able to:

1. Understand various Machine Learning terminologies
2. Understand the Machine Learning pipeline.

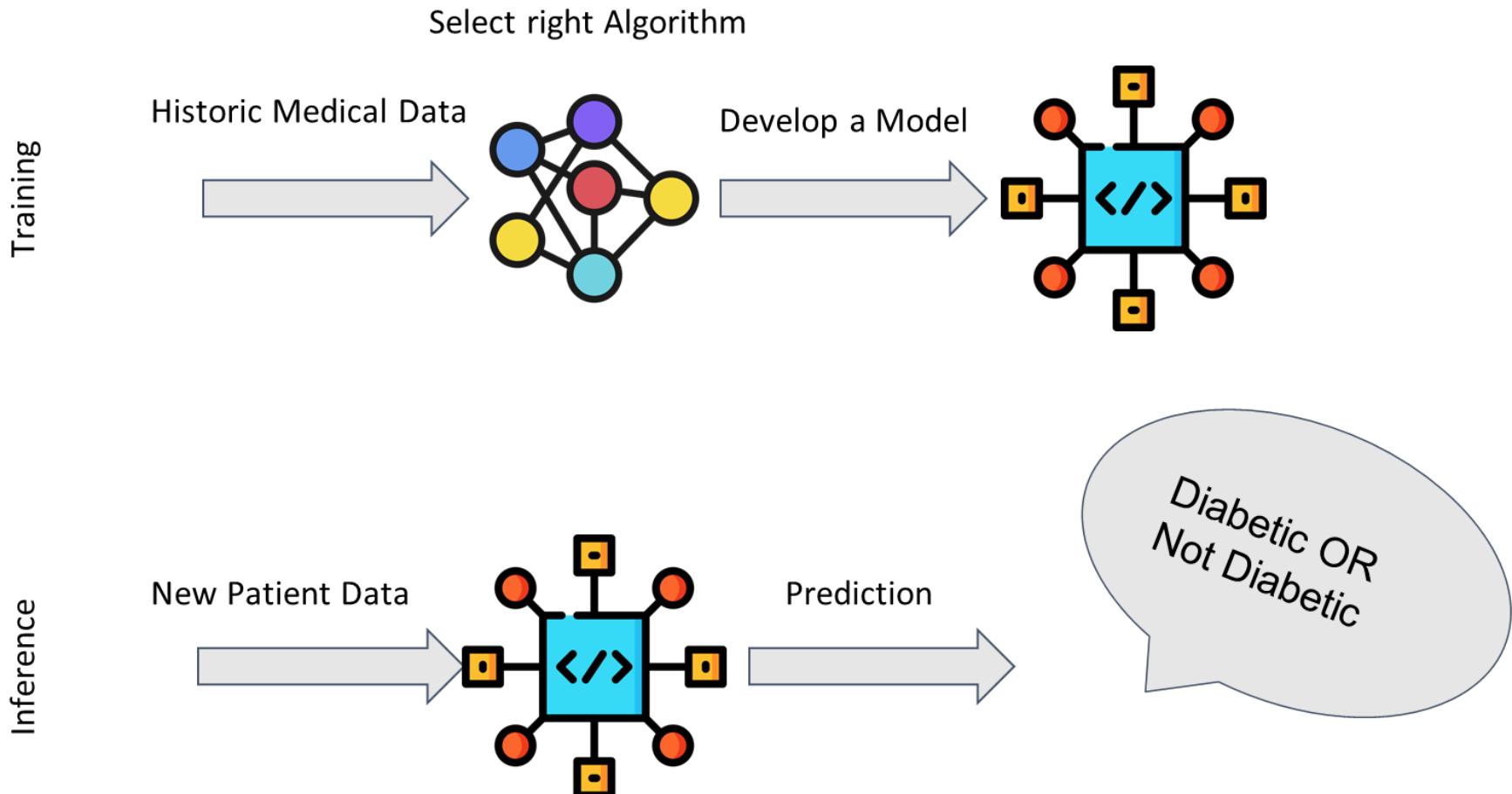
Develop A Diabetes Prediction System (DPS)



Sample Medical Data

Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1
11	143	94	33	146	36.6	0.254	51	1
10	125	70	26	115	31.1	0.205	41	1
7	147	76	0	0	39.4	0.257	43	1
1	97	66	15	140	23.2	0.487	22	0

Workflow For Developing DPS



Contents & Structure

- ML Terminology
- ML Pipeline



ML Terminology

- **Dataset**: A sample of real-world observations, organised into a table - rows and columns
- **Feature**: a column in the dataset table. Also called a predictor, variable, input, attribute, covariate
- **Training example**: A row in the table representing a set of feature values. Also called an observation, record, training instance
- **Target**: What we want to predict. Also called ground truth, (class) label, desired or expected output, response variable, dependent variable

ML Terminology

- **Model**: A function in the form $y = f(x)$ that we believe (or hope) is similar to the true function that represents the relationship between the target (y) and the observations in the dataset (x). Also called the **target function** or **objective function**
- **Prediction** or **Output**: Outcome from applying the model to the dataset - used to distinguish from targets, which are desired or expected outputs
- **Classifier**: An implementation that combines a *Learning algorithm*, a *Loss function*, and an *Optimiser* to create a model and generate predictions

ML Terminology

- **Learning algorithm:** A set of (mathematical) instructions that create a model using the training dataset. Most classifiers are named for the learning algorithm they implement to find or approximate the target function
- **Loss function:** Measures how far the predicted output for a single training example is from its true value. Also called the error function.
- **Optimiser:** An algorithm used to minimize the average or summed output of the Loss function for the entire dataset (the **Cost function**)

Dataset

- Hyper-Parameters
- Loss Function
- Optimiser
- Behavior Tuning

Input Row

Classifier
(algorithms)

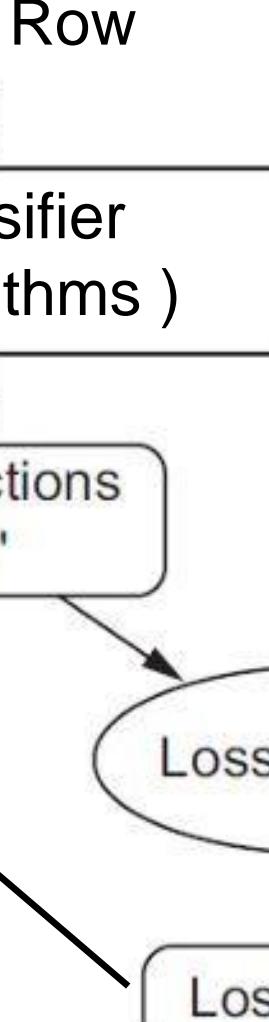
Predictions
 Y'

Essentially, the loss function and optimizer work together to fit the algorithm to the data in the best way possible.

True targets
 Y

Loss function

Loss score



ML Terminology

- **Model Parameters**: The parameters that the learning algorithm “learns” from the training data - for example, the slope (weight coefficients) and y-axis intercept of a linear regression line
- **Fitting a Model**: The process of learning the Model Parameters
- **Hyperparameters**: The tuning parameters of a classifier that affect its behavior, such as error tolerance, number of iterations, or options between variants of how the algorithm behaves. The programmer will specify defaults, and the user can pass new values with the function call.

ML Terminology

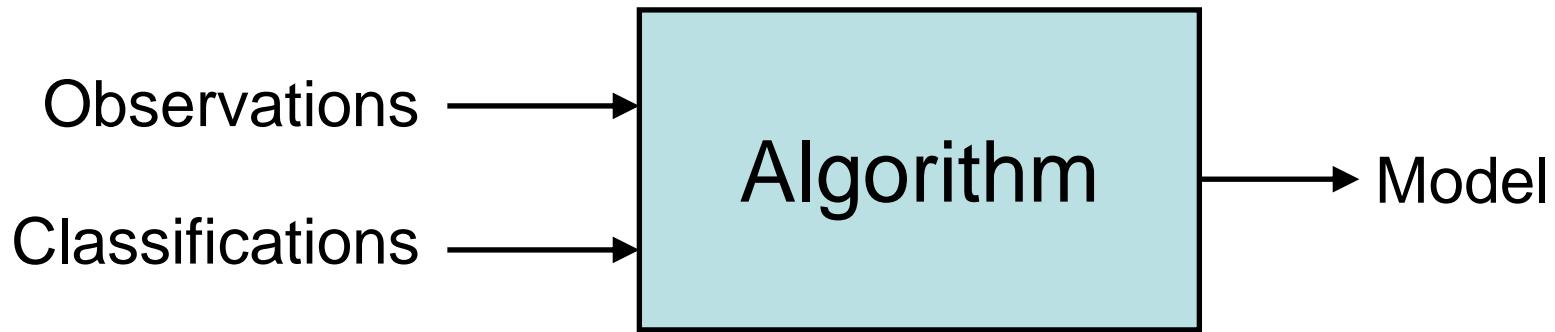
- **Model Evaluation:** Model performance is usually evaluated by counting the fraction of correctly classified instances out of all instances that the model attempted to classify.
 - For example, if we have a test dataset of 10,000 instances and a model classified 7,000 instances correctly, then we say that the model has a 70% accuracy on that dataset.

Note that optimization and evaluation measures are usually not the same in practice.

ML Terminology

- Training/fitting a model typically involves an algorithm that optimizes an objective function (e.g., maximizing the log-likelihood or minimizing mean squared error)
- In practice, we are often interested in the accuracy of a model on new, unseen data that has the same distribution as the training data (the **generalization performance**)
- There are a number of different algorithms for model parameter optimization, as well as a number of different techniques for estimating the generalization performance of a model

Supervised Learning



Machine learning (ML) pipeline consists of several steps to train a model.

ML pipeline is iterative: steps are repeated to continuously improve the accuracy of the model

Pipeline of Machine Learning Task

- Machine learning (ML) pipelines consist of several steps to train a model.
- ML pipelines are iterative as every step is repeated to continuously improve the accuracy of the model and achieve a successful algorithm.
- A typical machine learning pipeline:
 - Data collection
 - Data cleaning
 - Feature extraction (labelling and dimensionality reduction)
 - Model validation
 - Visualisation

Machine Learning Pipeline

What do I have
What do I need
What can I get

- Problem Definition
- Data collection

“Objective” in “objective function”

- Feature extraction
 - Flattening, Labeling
- Data preparation
 - Normalisation by data type
 - Dimensionality reduction

Exploratory Data Analysis (EDA)

Selection of metrics

- Algorithm Selection

– Train and Test

May have special data preparation requirements

- Performance Evaluation
 - Visualisation
 - Parameter tuning
- Model Validation

Repeat Train and Test

Machine Learning Pipeline

- Problem Definition
- Data collection
- Feature extraction
 - Flattening, Labeling
- Data preparation
 - Normalisation by data type
 - Dimensionality reduction
- Algorithm Selection
 - Train and Test
- Performance Evaluation
 - Visualisation
 - Parameter tuning
- Model Validation

Problem Definition

- **Goal:** Decide what information is needed
- **Methods:** Typically used are committee meetings, brainstorming, and analysis of business objectives
- **Outcome:** A project specification for the Data Scientists.

Machine Learning Pipeline

- Problem Definition
- Data collection
- Feature extraction
 - Flattening, Labeling
- Data preparation
 - Normalisation by data type
 - Dimensionality reduction
- Algorithm Selection
 - Train and Test
- Performance Evaluation
 - Visualisation
 - Parameter tuning
- Model Validation

Data Collection

- **Goal:** Representative sample of the population
- **Methods:** Extract existing data from databases, gather new data
- **Outcome:** A large dataset

Machine Learning Pipeline

- Problem Definition
- Data collection
- Feature extraction
 - Flattening, Labeling
- Data preparation
 - Normalisation by data type
 - Dimensionality reduction
- Algorithm Selection
 - Train and Test
- Performance Evaluation
 - Visualisation
 - Parameter tuning
- Model Validation

Feature Extraction

- **Goal:** Specifying features useful for prediction
- **Methods:** Check distributions, correlation/covariance checks, drop unique identifiers
- **Outcome:** A “flat” (single table) dataset with selected features.

Machine Learning Pipeline

- Problem Definition
- Data collection
- Feature extraction
 - Flattening, Labeling
- Data preparation
 - Normalisation by data type
 - Dimensionality reduction
- Algorithm Selection
 - Train and Test
- Performance Evaluation
 - Visualisation
 - Parameter tuning
- Model Validation

Data Preparation

- **Goal:** Clean up (normalize, regularize) the data values
- **Methods:** Scaling, dummy variables, interpolating missing values
- **Outcome:** Normalized data in numeric form appropriate for modeling

Machine Learning Pipeline

- Problem Definition
- Data collection
- Feature extraction
 - Flattening, Labeling
- Data preparation
 - Normalisation by data type
 - Dimensionality reduction
- Algorithm Selection
 - Train and Test
- Performance Evaluation
 - Visualisation
 - Parameter tuning
- Model Validation

Algorithm Selection

- **Goal:** No a-priori best algorithm; must experiment
- **Methods:** Draw on problem statement, domain knowledge, knowledge of ML and statistical methods, and library help files to choose candidate algorithms
- **Outcome:** A set of algorithms that should be appropriate for the analytical task

Machine Learning Pipeline

- Problem Definition
- Data collection
- Feature extraction
 - Flattening, Labeling
- Data preparation
 - Normalisation by data type
 - Dimensionality reduction
- Algorithm Selection
 - Train and Test
- Performance Evaluation
 - Visualisation
 - Parameter tuning
- Model Validation

Performance Evaluation

- **Goal:** Test candidate algorithms and evaluate predictive accuracy
- **Methods:** Fit each algorithm on the same set of historic training data, analyse confusion matrix and various other metrics
- **Outcome:** Comparison of results from selected algorithms

Machine Learning Pipeline

- Problem Definition
- Data collection
- Feature extraction
 - Flattening, Labeling
- Data preparation
 - Normalisation by data type
 - Dimensionality reduction
- Algorithm Selection
 - Train and Test
- Performance Evaluation
 - Visualisation
 - Parameter tuning
- Model Validation

Model Validation

- **Goal:** Evaluate robustness of the predictive model, check for over-fitting, refine hyperparameters
- **Methods:** Grid search CV, k-fold CV
- **Outcome:** Optimization of the objective function, good generalisation performance

Machine Learning Pipeline

- Problem Definition
- Data collection
- Feature extraction
 - Flattening, Labeling
- Data preparation
 - Normalisation by data type
 - Dimensionality reduction
- Algorithm Selection
 - Train and Test
- Performance Evaluation
 - Visualisation
 - Parameter tuning
- Model Validation

Iteration: Algorithm Selection

- Certain algorithms may be selected because they return a measure of feature importance which can be fed back to the feature extraction phase

Machine Learning Pipeline

- Problem Definition
- Data collection
- Feature extraction
 - Flattening, Labeling
- Data preparation
 - Normalisation by data type
 - Dimensionality reduction
- Algorithm Selection
 - Train and Test
- Performance Evaluation
 - Visualisation
 - Parameter tuning
- Model Validation

Iteration: Model Validation

- The model validation phase should lead to a new round of performance evaluation, and the set of candidate algorithms will gradually be reduced until one algorithm (or one ensemble of algorithms) is selected.

Further Reading

- **What is a Machine Learning Pipeline?**

<https://youtu.be/HWWxtVL-D9k>

- **The 7 Steps of Machine Learning**

<https://youtu.be/nKW8Ndu7Mjw>

Review Questions

1. What are the different terminology in Machine Learning?
2. What are the key steps (pipeline) in Machine Learning?

Summary / Recap of Main Points

1. Various terminologies in Machine Learning.
2. Various steps in the Machine Learning pipeline.