**ASSIGMENT 1: ML & DS**

1. Machine learning addresses the question of how to build computers that improve automatically through experience (Brookings, 2021; ScienceDirect, 2023). It is one of today’s most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science (Udacity, 2015). Recent progress in machine learning has been driven both by the development of new learning algorithms and theory and by the ongoing explosion in the availability of online data and low-cost computation (All-About-Industries, n.d.). The adoption of data-intensive machine-learning methods can be found throughout science, technology and commerce (Citizenside, n.d.).
2. **Supervised Learning**

**Definition:**

* Supervised learning occurs when a model is trained on **labeled data**, meaning each input has a known correct output.
* Think of it as the computer having a “teacher” that shows the right answer for every example.

**Key Tasks:**

1. **Regression** → Predicting continuous values (numbers).
   * *Example:* Estimating the price of a car based on mileage, age, and brand.
2. **Classification** → Predicting discrete categories (classes).
   * *Example:* Detecting whether a message is “Spam” or “Not Spam”.

**Use Cases:**

* Financial forecasting, medical diagnosis, email filtering, and sales prediction.

**Unsupervised Learning**

**Definition:**

* Unsupervised learning occurs when the data is **unlabeled**, so the algorithm must **identify patterns and structure** by itself.
* There is **no teacher** guiding the output.

**Key Tasks:**

1. **Clustering** → Grouping similar items together.
   * *Example:* Segmenting customers in e-commerce by shopping behavior.
2. **Dimensionality Reduction** → Reducing the number of variables while keeping important information.
   * *Example:* Compressing image data for faster processing or visualizing complex datasets.

**Use Cases:**

* Market segmentation, anomaly detection, data compression, and visualization of complex datasets.

| **Aspect** | **Supervised Learning** | **Unsupervised Learning** |
| --- | --- | --- |
| **Data** | Labeled (with correct answers) | Unlabeled (no correct answers) |
| **Goal** | Predict outcomes / categories | Discover patterns or structures |
| **Teacher** | Yes | No |
| **Examples** | House price prediction, Spam detection | Customer segmentation, PCA |
| **Output** | Known (numbers or categories) | Unknown patterns or groups |

1. **Overfitting .**

**Overfitting** happens when a machine learning model learns **too much from the training data**, including the **noise or random details**.

This means the model **memorizes the data** instead of learning the **general patterns**.

**Result:** Performs very well on training data but **poorly on new, unseen data**.

**Simple Example:**

A student memorizes all the answers to a past test instead of understanding the concepts. When a new test comes, the student struggles because the questions are different.

**Causes of Overfitting**

1. **Too complex model**
   * Example: Very deep neural networks or highly detailed decision trees.
2. **Too little training data**
   * Small datasets → model learns random noise instead of general patterns.
3. **Noisy data**
   * Errors, outliers, or irrelevant features can be treated as important by the model.
4. **Too many features**
   * More input variables than necessary can confuse the mode

**How to Prevent Overfitting**

1. **Use more training data**
   * Larger datasets help the model learn general patterns.
2. **Simplify the model**
   * Reduce features or parameters. Use simpler algorithms if possible.
3. **Regularization**
   * Penalize overly large coefficients in models (e.g., **L1 / L2 regularization**).
4. **Cross-validation**
   * Split data into training and validation sets to check performance on unseen data.
5. **Pruning / Early stopping**
   * Stop training when performance on validation data starts to drop.
6. **Dropout (for neural networks)**
   * Randomly ignore some neurons during training to prevent memorization.

**Summary Table**

| **Aspect** | **Overfitting** |
| --- | --- |
| **What happens?** | Model memorizes training data → fails to generalize to new data. |
| **Causes** | Complex model, small dataset, noisy data, too many features. |
| **Prevention** | More data, simpler model, regularization, cross-validation, pruning/dropout. |

1. **What are Training Data and Test Data**

** Training Data:**

* This is the dataset given to the model to learn patterns and the relationship between inputs (features) and outputs (labels).
* The model uses this data to adjust its parameters during training.

** Test Data:**

* This is data the model has never seen before during training.
* It is used to evaluate the model’s performance on new, unseen data.
* This shows the generalization ability of the model.

**How Training and Test Data are Split**

* Datasets are usually split into:
  + **Training set:** 70% – 80% of the data.
  + **Test set:** 20% – 30% of the data.
* Sometimes, a **validation set** is also used:
  + About 10% – 20% of the data, used for **tuning hyperparameters** or **early stopping**.

**Methods of Splitting:**

1. **Random split:**
   * Data is randomly divided into training and test sets.
2. **Stratified split:**
   * Used when class distribution is imbalanced.
   * Ensures the same proportion of each class in training and test sets.
3. **Cross-validation:**
   * The dataset is divided into k-folds (e.g., 5 or 10).
   * The model is trained on k-1 folds and validated on the remaining fold, repeated k times.
   * Provides an average performance metric.

**Why Splitting is Necessary**

1. **To prevent overfitting:**
   * If the model sees all the data during training, it may **memorize the dataset** rather than learn general patterns.
2. **To measure generalization:**
   * Test data shows how the model performs on **unseen data**, which is critical for real-world usage.
3. **To tune hyperparameters:**
   * The validation set is used to choose the **best model configuration** without touching the test set.
4. **To compare models:**
   * Using the same test set allows a fair comparison of different models’ performance.

| **Aspect** | **Training Data** | **Test Data** |
| --- | --- | --- |
| **Purpose** | Learn patterns & train model | Evaluate performance on unseen data |
| **Usage** | Adjust model parameters | Not used during training |
| **Size** | 70–80% of dataset | 20–30% of dataset |
| **Function** | Fit the model to data | Test generalization ability |

1. **Case Study: Machine Learning in Healthcare**

**Overview**

This study explores the integration of machine learning tools in two critical areas of healthcare:

1. **Sepsis Diagnosis**: Early identification of sepsis to prevent complications.
2. **Suicide Prediction**: Assessing patient data to predict and prevent suicide attempts.

**Key Findings**

* **Sepsis Diagnosis**: ML models demonstrated the ability to analyze patient data in real-time, identifying early signs of sepsis more swiftly than traditional methods. This early detection is crucial for timely intervention and improved patient outcomes.
* **Suicide Prediction**: By analyzing electronic health records, social determinants, and behavioral health data, ML models identified patterns that could predict suicide risk. This predictive capability allows healthcare providers to intervene proactively.

**Challenges Identified**

* **Human Factors**: There is a lack of comprehensive studies on how healthcare providers interact with ML tools, which is essential for effective implementation.
* **Integration into Clinical Practice**: Incorporating ML tools into existing healthcare workflows presents challenges, including training staff and ensuring the tools complement clinical judgment.

**Conclusion**

The study underscores the potential of ML in enhancing healthcare delivery, particularly in critical areas like sepsis diagnosis and suicide prevention. However, it also highlights the need for further research into human factors and integration strategies to maximize the effectiveness of ML tools in clinical settings