EcoMinds

Co-funded by the European Union

Introduction to Machine Learning for Environmental Data

Data All Around Us

In this module, you will learn what data is and why it matters for understanding the environment. You don't need any technical background. The goal is to help you recognize data in your daily life and see how they form the foundation for digital technologies and environmental decisions.

EcoMinds

Co-funded by the European Union

What Are Data?

- Data are collected observations or measurements
- They describe what is happening in the real world
- Data are created through observation and measurement
- Data are the starting point of digital technologies

Data are essential recorded facts about the world around you. They are generated when something is observed or measured, such as temperature, air quality, or energy consumption. Data help us describe reality as it is and form the basis of digital systems and artificial intelligence.

EcoMinds

Co-funded by the European Union

Data as Raw Values

- Data are single values or records
- On their own, data do not explain situations
- Context and analysis give data meaning
- Multiple data points are usually needed

At this stage, data are simply values. A single number rarely tells the whole story. For example, the value “18°C” is data, but it doesn't indicate whether the day is warm or cold. When we gather many data points and analyze them together, data become meaningful.

EcoMinds

Co-funded by the European Union

Why Do We Need Data?

- To describe real-world situations
- To compare values over time
- To detect changes and patterns
- To support understanding and decisions

We need data to understand what is truly happening around us. Without data, decisions would be based on assumptions. By gathering data over time, we can identify changes and patterns, which is especially vital in environmental monitoring and protection.

EcoMinds

Co-funded by the European Union

Forms of Data

- Numbers (e.g. temperature values)
- Text (e.g. weather reports)
- Images (e.g. satellite photos)
- Sensor data (e.g. air quality sensors)
- Data can appear in different formats

Not all data are numbers. Environmental data come in many forms, including text, images, and continuous sensor readings. Each form of data captures a different aspect of the environment and helps us describe reality more accurately.

EcoMinds

Co-funded by the European Union

Examples of Data Forms

- A weather forecast written as text
- A satellite image of a forest area
- Temperature values recorded every hour
- Air quality measurements from sensors

These examples demonstrate how environmental data are collected and stored in various ways. Text describes conditions, images show spatial information, and numerical and sensor data capture precise measurements over time.

EcoMinds

Co-funded by the European Union

Why Different Data Forms Matter

- Each data form provides different information
- Combining data gives a clearer picture
- Real systems use multiple data types

Different types of data complement each other. For example, sensor measurements provide precise values, while images reveal where changes are occurring. In environmental monitoring, combining different data types results in better understanding and more accurate conclusions.

EcoMinds

Co-funded by the European Union

Environmental Data in Everyday Life

- Data are collected continuously
- Many data are publicly available
- You interact with environmental data every day
- Data support awareness and decision-making

You already interact with environmental data, even if you don't always see it as data. Many environmental measurements are collected automatically and shared with the public. This data helps people understand current environmental conditions and make informed decisions.

EcoMinds

Co-funded by the European Union

Common Examples of Environmental Data

- Temperature and weather data
- Air quality measurements
- Energy consumption data
- Water quality indicators
- Noise and pollution levels

Weather forecasts, air quality apps, and energy use information all rely on continuous data collection and analysis. These data are often gathered by sensors, monitoring stations, and public institutions, and they are updated regularly.

EcoMinds

Co-funded by the European Union

Who Uses Environmental Data?

- Companies and industries
- Individuals (daily decisions)
- Cities and local communities
- Governments and public institutions
- Scientists and environmental experts

Environmental data are utilized by various groups. Individuals depend on them to plan daily activities, while cities and governments use them to develop policies and address environmental issues. Scientists examine environmental data to understand long-term changes and evaluate potential risks.

Environmental data are also widely used by companies and industries. For example, energy companies analyze consumption and production data, transport companies use weather data to plan routes, and manufacturing companies monitor emissions to comply with environmental regulations. As a result, environmental data are becoming increasingly important for sustainable and responsible business decisions.

EcoMinds

Co-funded by the European Union

Data vs Information

- Data are raw values
- Information is data with meaning
- Data describe what is measured
- Information helps us understand

Data and information are not the same. Data are raw measurements collected from the environment. Information is created when we analyze data and understand what they represent. Without analysis, data remain just values without context.

EcoMinds

Co-funded by the European Union

From Data to Information

- Single data points give limited insight
- Multiple data points reveal patterns
- Analysis adds context and meaning

- Example:
- Temperature readings -> daily average temperature

A single temperature reading reveals little on its own. When multiple measurements are analyzed together, patterns begin to emerge. For instance, calculating a daily average temperature helps you determine if a day was generally warm or cold.

EcoMinds

Co-funded by the European Union

Why Information Matters

- Supports understanding of the environment
- Enables comparison over time
- Helps decision-making and prediction

Information enables us to interpret environmental data. It helps us compare conditions over time, identify changes, and inform decisions. Later in this seminar, you will see how machine learning uses information from data to make predictions.

EcoMinds

Co-funded by the European Union

Why Do Environmental Data Matter?

- Environmental data describe changes over time
- They help us understand what is happening in nature
- Data provide evidence, not assumptions

Environmental data are crucial because they help us track how the environment changes over time. Instead of relying on opinions or guesses, data provide evidence about what is actually happening in nature, like changes in temperature, air quality, or water conditions.

EcoMinds

Co-funded by the European Union

Detecting Changes and Trends

- Monitoring environmental conditions
- Identifying long-term trends
- Recognising unusual events

By continuously collecting environmental data, we can identify trends and changes. For example, we can notice rising average temperatures, increasing pollution levels, or sudden unusual events. Detecting these patterns is crucial for understanding environmental risks and challenges.

EcoMinds

Co-funded by the European Union

Supporting Decisions and Predictions

- Data support informed decision-making
- Decisions can be based on evidence
- Data are the basis for prediction methods

Environmental data inform decisions made by individuals, companies, and governments. When decisions rely on data, they tend to be more dependable. These data also provide the basis for prediction methods, including machine learning, which you will examine later in this seminar.

EcoMinds

Co-funded by the European Union

Your Turn: Think About Environmental Data

- Think about the following questions:
  - Which environmental data do you encounter in your daily life?
  - Where do these data come from?

Take a few minutes to think about these questions. Consider applications, websites, public displays, or information you encounter in the media. Focus on finding real examples from your daily life instead of searching for “correct” answers.

EcoMinds

Co-funded by the European Union

Your Turn: Apply What You Learned

- Try to identify:
  - At least two types of environmental data
  - At least two different data sources
  - One example of how these data are used

Write a brief response or prepare to discuss your ideas with others. For each example, consider who collects the data and how they use it. This task helps you actively connect the content of this module with real-world situations and gets you ready for the upcoming modules of the seminar.

EcoMinds

Co-funded by the European Union

Introduction to Machine Learning for Environmental Data

Basic Statistics for Understanding Data



In this module, you'll learn how basic statistical measures help us understand data. You will apply statistics to simple environmental examples and see how numbers describe patterns, variability, and trends.

EcoMinds

Co-funded by the European Union

Why Do We Need Statistics?

- Environmental data sets can be large
- Data are often collected continuously
- Raw data are difficult to read and interpret
- We need a way to simplify data

Environmental data are typically collected over long periods and include many values. Viewing raw data, like hundreds of temperature readings, is impractical. Statistics help us simplify large data sets so we can understand them more easily.

EcoMinds

Co-funded by the European Union

What Do Statistics Help Us Do?

- Summarise many values into a few numbers
- Compare different situations or time periods
- Detect similarities and differences
- Support understanding and decisions

Statistics convert large amounts of data into a few meaningful values. This helps us compare different days, locations, or conditions. Using statistics, we can better understand environmental changes and make informed decisions.

EcoMinds

Co-funded by the European Union

From (Environmental) Data to Statistics

- Environmental data often include many values
- Large data sets are hard to interpret directly
- Statistics help reduce complexity
- A few numbers can describe many values

Environmental data are often collected continuously and can include many measurements. Analyzing each value individually is impractical. Statistics help us simplify this complexity by summarizing numerous data points into a few meaningful values.

EcoMinds

Co-funded by the European Union

What Do Statistics Describe?

- Typical or central values
- Differences between data sets
- Patterns and changes over time
- General behaviour of data

Statistics help characterize what is typical in a data set and how values vary from each other. They also enable us to compare different periods or locations and detect patterns or changes in environmental conditions.

EcoMinds

Co-funded by the European Union

Average (Mean)

- Represents a typical value
- Uses all measured values
- One number summarises many values
- Very common in environmental data

The average, also called the mean, is one of the most used statistical measures. It combines all measured values into a single number that represents the entire data set. This simplifies understanding large data sets.

The average helps us understand what is typical in a data set. For example, comparing average temperatures allows us to see whether one day, month, or year was warmer or cooler than another. This is why averages are frequently used in weather reports and environmental summaries

EcoMinds

Co-funded by the European Union

Example: Calculating the Average

- The arithmetic average of all data points.
- It's like finding the "balance point" of your data.

- Daily temperatures: 2 4 4 6 8 °C

- Average temperature: (2 + 4 + 4 + 6 + 8) / 5 = 4.8 °C

This example shows how multiple temperature measurements are combined into one representative value. The average allows us to compare different days or periods easily.

EcoMinds

Co-funded by the European Union

Minimum and Maximum

- Minimum: lowest recorded value
- Maximum: highest recorded value
- Show extreme conditions

- Example:
- Daily temperatures: 16, 18, 21, 27 °C
  - Minimum = 16 °C
  - Maximum = 27 °C

Minimum and maximum values show the extreme environmental conditions in a data set. These extremes are significant because they can indicate risks, like heat waves or cold spells. However, extreme values alone do not describe typical conditions and should be considered alongside other statistics, such as the average or median.

EcoMinds

Co-funded by the European Union

Range

- Difference between maximum and minimum
- Shows how much values vary
- Simple measure of data spread

- Example:
- Daily temperatures: 16, 18, 21, 27 °C
  - Range = 27 − 16 = 11 °C

The range shows how spread out the data values are. A small range indicates stable conditions, while a large range suggests significant variation. The range is easy to calculate and provides a quick way to see how much environmental conditions fluctuate within a given period.

EcoMinds

Co-funded by the European Union

Mode

- The most frequent value in a data set
- Shows which value occurs most often
- Easy to identify in small data sets

- Example:
- Data: 2, 4, 4, 6, 8
  - Mode = 4

The mode is the value that appears most often in a data set. You can think of it as the “most popular” value. In the example shown, the number 4 appears more frequently than the others, so it is the mode.
In environmental data, the mode is sometimes less useful than the average or median, especially when values are continuous, but it can still offer quick insights in simple cases.

EcoMinds

Co-funded by the European Union

Median

- Middle value of an ordered data set
- Half of the values are lower, half are higher
- Less affected by extreme values
- Odd Number of Points:
  - Example: 1, 3, 5
  - Median = 3 (Middle value)
- Even Number of Points:
  - Example: 1, 3, 5, 7
  - Median = (3+5)/2 = 4 (Average of two middle values)

The median is the middle value of a data set when the values are arranged in order. Half of the values are below it, and half are above. The median is especially helpful when extreme values might skew the average because it emphasizes the center of the data instead of all the values.

EcoMinds

Co-funded by the European Union

Comparing Mean and Median

- Mean uses all values
- Median focuses on the middle
- Differences can indicate extreme values

- Example:
- Data: 10, 11, 12, 13, 50
- Mean > Median

Comparing the mean and median helps us understand how data are spread out. When there are extreme values, they influence the mean more than the median. In the example shown, one very large value pulls the mean up, while the median stays closer to the typical values. This difference indicates that the data are not evenly distributed.

EcoMinds

Co-funded by the European Union

Standard Deviation (Concept)

- Measures variability in data
- Describes the spread around the mean
- Small value -> similar data
- Large value -> high variation

- Example:
- Data A: 19, 20, 21 -> small standard deviation
- Data B: 10, 20, 30 -> large standard deviation

Standard deviation indicates how much data points differ from the mean. When most values are near the average, the standard deviation is small. When values are spread over a wide range, the standard deviation is large. This metric helps understand the stability or variability of environmental conditions.

EcoMinds

Co-funded by the European Union

Standard Deviation (Calculation)

- Calculate the mean
- Subtract the mean from each value
- Square the differences
- Calculate the average of squared differences (**variance**)
- Take the square root

- **Example** (Temperature Data):
- Values: 18, 20, 22, 19, 21 °C
- **Mean** = 20 °C

EcoMinds

Co-funded by the European Union

Standard Deviation (Calculation)

Value	Difference (x – mean)	Squared difference
18	−2	4
20	0	0
22	2	4
19	−1	1
21	1	1

Average of squared differences (variance): 4+0+4+1+15 / 5 = 2 °C

Standard deviation: $\sqrt{2}$ ≈1.41°C

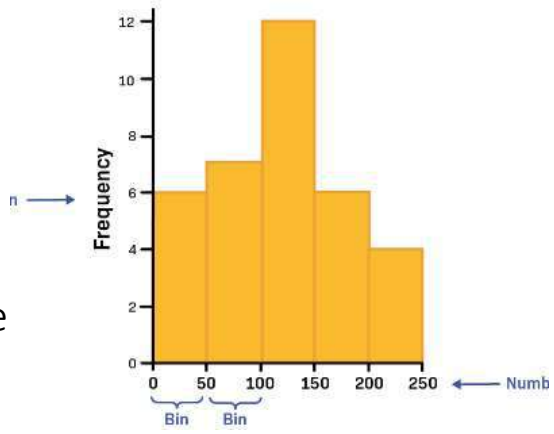
This slide demonstrates the step-by-step calculation of standard deviation. Each step of the algorithm is connected to the numerical values from the example. First, we find how far each value is from the mean. Then we square the differences to prevent cancellation. The average of the squared differences represents the variance, and taking the square root brings the result back to the original units. The final value indicates how much temperatures typically fluctuate around the average.

EcoMinds

Co-funded by the European Union

Histograms: The Data's Photo Album!

- **Bins** (Intervals): Designated spots for numbers to stand.
- **Height**: Shows how popular each spot is.
- **Tall Spot**: Crowded! Lots of numbers here.
- **Short Spot**: Fewer numbers, more space.



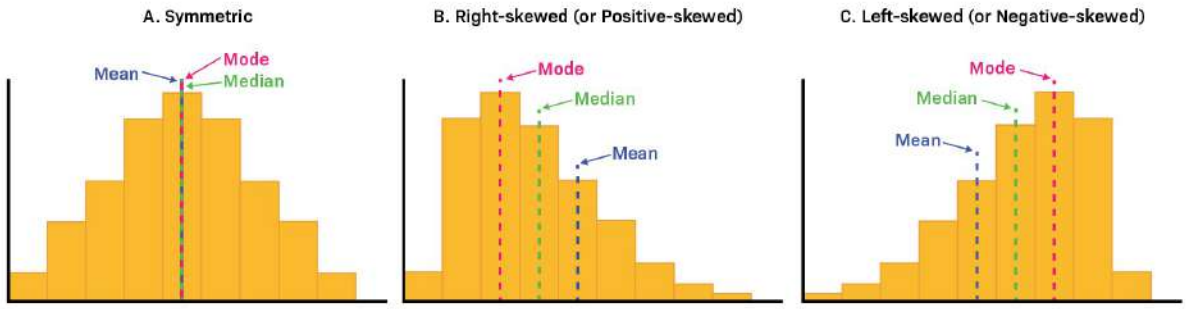
A histogram visualizes how data values are spread out. Instead of listing individual numbers, values are grouped into ranges, and the height of each bar indicates how frequently values appear. Histograms help us understand whether data is clustered around a typical value or spread across a wide range.

EcoMinds

Co-funded by the European Union

Histograms: The Data's Photo Album!

- **Symmetry and Skewness**: A symmetric histogram is balanced, while a skewed histogram leans to one side



This slide illustrates how histograms can reveal data distribution shapes. A histogram is symmetric when values are evenly spread around the center. In such cases, the mean, median, and mode are roughly at the same point, indicating balanced data.

In a right-skewed (positively skewed) histogram, most values are concentrated on the left side, while a few large values extend to the right. These extreme values pull the mean to the right, making the mean larger than the median and the mode.

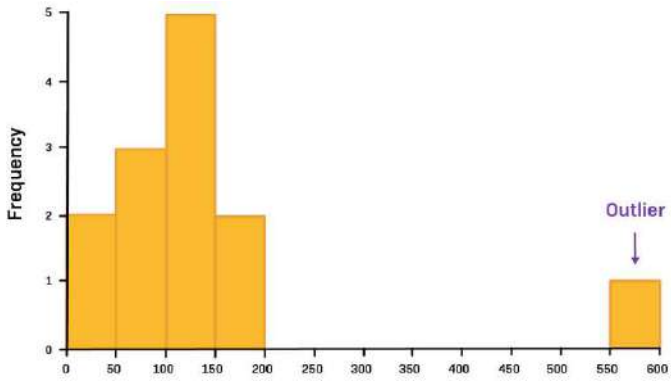
In a left-skewed (negatively skewed) histogram, most values are focused on the right side, with a few small values stretching to the left. In this case, the mean is pulled to the left and is less than the median and the mode.

EcoMinds

Co-funded by the European Union

Histograms: The Data's Photo Album!

- **Outliers**: Data points far from the rest can stand out as isolated bars.



This slide demonstrates how histograms can identify outliers. Outliers are data points that are far from most values in a data set. In a histogram, outliers often show up as isolated bars separated from the main group of data.

Outliers can happen for different reasons. They might reflect rare but genuine events, like extreme weather conditions. Sometimes, they are caused by measurement mistakes or wrong data recording.

It is important to recognize outliers because they can significantly impact statistical measures. Extreme values can shift the mean away from typical values and raise the standard deviation, while the median is generally less affected.

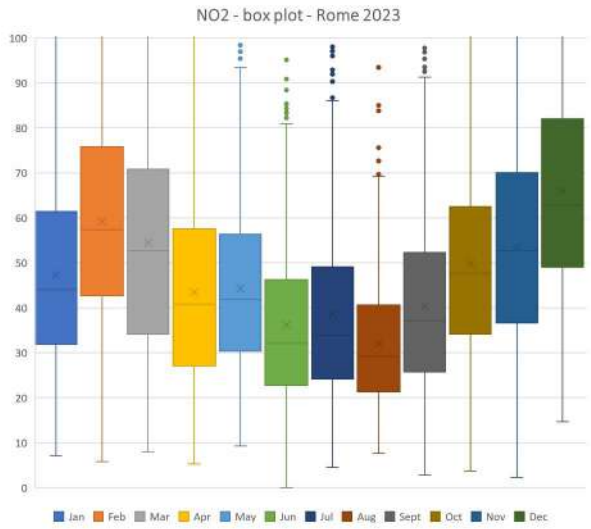
When analyzing environmental data, outliers should not be automatically removed. Instead, we should first try to understand why they happen and what they indicate.

EcoMinds

Co-funded by the European Union

Box Plot

- Summarises data using key statistics
- Shows median and data spread
- Highlights variability and extreme values
- What to observe:
  - Position of the median
  - Length of the box and whiskers



This slide introduces the box plot as a concise way to summarize data using key statistics. A box plot displays the median of the data and how values are distributed around it, allowing us to quickly evaluate variability.

The position of the median within the box shows where the central value is located. If the median is nearer to the top or bottom of the box, it indicates that the data are not evenly spread around the center.

The length of the box shows the spread of the middle part of the data. A longer box indicates more variability, while a shorter box suggests more stable values. The whiskers extend to show how far values reach beyond the central range, and individual points may represent extreme values.

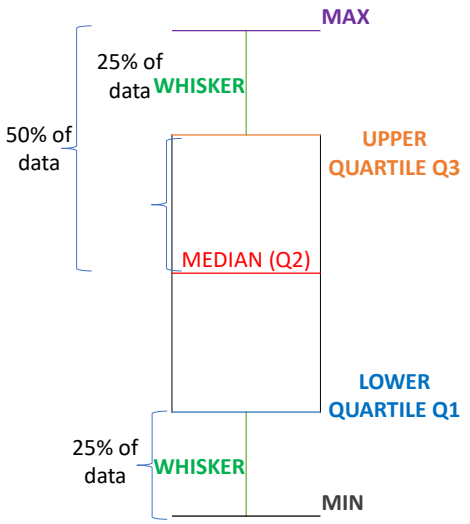
Box plots are particularly helpful for comparing multiple data sets side by side. In this example, different months can be compared to observe changes in median values and variability over time.

EcoMinds

Co-funded by the European Union

Box Plot

- From Q3 to the maximum value (Fourth Quartile): Contains the highest 25% of the data, completing 100% of the data set.
- From the Median to Q3 (Third Quartile): Contains the next 25% of the data. By the end of Q3, you've covered 75% of the data.
- From Q1 to the Median (Second Quartile): Contains the next 25% of the data. By the median's end, you've covered 50% of the data.
- From the minimum value to Q1 (First Quartile): Contains the lowest 25% of the data.



This slide explains how a box plot separates data into four equal parts called quartiles. Each quartile accounts for 25% of the data, arranged from the smallest to the largest values.

The lower quartile, Q1, indicates the point below which 25% of the data falls. This marks the lower portion of the data set. From Q1 to the median, another 25% of the data is included, meaning that by the time we reach the median, we have covered 50% of all values.

From the median to the upper quartile, Q3, we cover the next 25% of the data. By the time we reach Q3, 75% of the data is included. Finally, from Q3 to the maximum value, the last 25% of the data is represented.

The box contains the middle 50% of the data, between Q1 and Q3. The whiskers extend to show the range of the remaining values towards the minimum and maximum. This structure helps us quickly understand the data distribution and where most values are concentrated.

EcoMinds

Co-funded by the European Union

Understanding Environmental Trends

- Comparing averages over time
- Monitoring changes in variability
- Identifying unusual situations

By analyzing statistical data over time, we can start to recognize environmental trends. For instance, comparing average temperatures across days, months, or years helps us see long-term changes instead of focusing on isolated events.

Monitoring changes in variability is equally important. Even if the average stays the same, increased variability can signal more unstable or extreme environmental conditions. That's why measures like range and standard deviation are helpful when analyzing trends.

Statistics also help us identify unusual situations, such as sudden spikes or drops in values. These cases may signal rare events, environmental hazards, or potential measurement errors. Using basic statistics to understand trends is often the first step before applying more advanced analytical or machine learning methods.

EcoMinds

Co-funded by the European Union

Why Statistics Matter for Machine Learning

- Data must be understood before modelling
- Variability affects model performance
- Statistics guide data preparation

Machine learning models do not work directly with raw data. Before building a model, it is important to understand the data using basic statistics. Measures like the average, median, range, and standard deviation help us see what is typical, what is unusual, and how much values vary.

Variability is crucial for model performance. Data sets with high variability or extreme values can influence how effectively a model learns patterns. By analyzing variability, we can determine if data needs cleaning, scaling, or further investigation.

Statistics also guide data preparation. Understanding data distributions helps us select appropriate methods and avoid misleading results. For this reason, statistics form a foundation for any machine learning application.

EcoMinds

Co-funded by the European Union

Your Turn: Apply Statistics

- Choose an environmental data set
- Identify useful statistics
- Explain what they tell you

In this activity, you will apply what you've learned about basic statistics. First, select a simple environmental data set, such as daily temperatures, air quality measurements, or energy consumption values. You can think of a real example or use data you've seen in previous slides.

Next, decide which statistical measures are useful for understanding these data. For example, you might choose the average to describe typical conditions, the range or standard deviation to describe variability, or the median if extreme values are present. Finally, describe what these statistics reveal about the data. Emphasize interpretation rather than calculation. The purpose of this activity is to demonstrate how statistics convert raw data into meaningful information that can support further analysis and machine learning.

EcoMinds

Co-funded by the European Union

Introduction to Machine Learning for Environmental Data

What Is Artificial Intelligence?

This module will explore what artificial intelligence truly is, in a straightforward and practical manner.

We'll examine how AI differs from traditional computer programs, where computers operate based on strict, pre-written rules.

You will also see everyday examples of AI that you already use, sometimes without even realizing it.

Finally, we'll connect AI with environmental protection and sustainability, and see how AI can help us better understand environmental data and make smarter decisions for the planet.

EcoMinds

Co-funded by the European Union

Programs, Algorithms, and AI

- Programs follow explicit instructions
- Algorithms define step-by-step procedures
- AI systems learn patterns from data

To understand artificial intelligence, we first need to explain how computers have typically operated.

A program consists of a set of instructions created by humans. The computer executes these instructions precisely as they are provided. An algorithm is a systematic method of outlining instructions. It establishes clear, step-by-step processes to resolve a problem.

Artificial intelligence brings a fundamental change. Instead of setting all rules beforehand, AI systems learn patterns from data. This means that behavior is not entirely programmed by the developer. The system modifies its actions based on examples and experience. This single difference explains why AI is valuable in complex, dynamic environments, such as human behavior or environmental systems.

EcoMinds

Co-funded by the European Union

Why Rules Are Not Enough

- Real-world systems are complex
- Conditions change over time
- Not all situations can be predefined

Traditional programs work effectively when problems are straightforward and predictable. For instance, calculating a tax rate or sorting a list of names can be managed with clear, defined rules.

However, many real-world systems are complex. They involve numerous variables, interactions, and constant changes.

Human behavior is one such system. Environmental systems are another. In these cases, it is impossible to write rules for every possible situation.

Conditions evolve, new patterns emerge, and unforeseen events happen.

EcoMinds

Co-funded by the European Union

Learning from Data

- Systems learn from examples
- Past data influence future behaviour
- Learning improves with experience

In complex systems, instead of trying to define every possible rule, we adopt a different approach - we learn from data.

An AI system is presented with many examples from the past. These examples show how the system behaved or what occurred in different situations. From these examples, the system recognizes patterns. When new data appear, it uses what it has learned to respond, predict, or decide.

Importantly, learning does not stop after the initial use. As more data becomes available, the system's behavior can improve over time. This capacity to learn from experience enables AI systems to function effectively in complex and evolving environments.

EcoMinds

Co-funded by the European Union

Everyday Examples of AI

- Recommendation systems (Netflix, TikTok …)
- Voice assistants
- Image and speech recognition
- Navigation and route planning

Now that we understand learning from data, let's explore examples you already see in daily life. Recommendation systems, like Netflix or TikTok, learn from your viewing habits and adjust what they show you. Voice assistants learn from extensive speech datasets to identify words and provide suitable responses.

Image and speech recognition systems learn patterns from examples, enabling them to recognize faces, objects, or spoken language. Navigation apps analyze historical and real-time data to forecast traffic and recommend efficient routes. In all these cases, the system gets better as it processes more data. This is the learning process we talked about earlier, now clearly seen in action.

EcoMinds

Co-funded by the European Union

AI and Data Quality

- AI relies on historical data
- Data can be incomplete or biased
- Poor data lead to poor outcomes

AI systems learn from data, but not all data are reliable. The data used to train AI systems depict past situations and behaviors. If these data are incomplete, outdated, or biased, the system will learn an inaccurate picture of reality.

This can result in inaccurate predictions or biased decisions. Therefore, data quality is essential. Before using data for AI, we need to understand where the data originates, what they represent, and what might be missing. This directly relates to the previous module, where you learned about basic statistics and data analysis as tools for understanding data before applying them in AI systems.

EcoMinds

Co-funded by the European Union

AI in Environmental Protection

- Monitoring air and water quality
- Predicting environmental changes
- Detecting pollution and anomalies
- Supporting sustainable decisions

Artificial intelligence is especially helpful in environmental protection because environmental systems produce large amounts of data. Sensors, satellites, and monitoring stations constantly gather data about air quality, water quality, weather conditions, and ecosystems. AI systems assist in analyzing these datasets to detect patterns and trends that would be very difficult for humans to identify manually. AI can also be used to forecast environmental changes, such as pollution levels or weather-related risks.

Another key application is detecting anomalies, such as sudden pollution spikes that could indicate an incident. By transforming data into insights, AI enables authorities, organizations, and communities to make better and more sustainable decisions.

EcoMinds

Co-funded by the European Union

Example: Environmental Monitoring System

- Sensors collect environmental data
- AI analyses patterns and trends
- Alerts or predictions are generated

Let's look at a simple example that ties everything together. In an environmental monitoring system, sensors are installed in the environment. These sensors gather data like temperature, air pollution levels, or water quality indicators. Over time, they accumulate large amounts of data.

An AI system analyzes this data to identify patterns and trends. For example, it can learn what normal pollution levels look like and recognize when something unusual happens. Based on this analysis, the system can produce alerts or forecasts, such as predicting higher pollution levels during specific weather conditions. This enables authorities or communities to respond more quickly and effectively, rather than reacting only after a problem becomes visible.

EcoMinds

Co-funded by the European Union

AI Is Not Magic

- AI does not understand like humans
- AI depends on data and models
- Wrong data lead to wrong results

Artificial intelligence can be very powerful, but it's crucial to understand its limitations. AI systems do not think, reason, or understand the world like humans do. They identify patterns and generate outputs based on data and models.

This shows that AI lacks common sense and understanding of context outside the data it was trained on. If the data is incomplete, biased, or wrong, the system will give unreliable results. For environmental uses, this is especially critical because AI-based decisions can impact people, ecosystems, and long-term sustainability.

Therefore, AI should be viewed as a support tool, not as a substitute for human judgment and responsibility.

EcoMinds

Co-funded by the European Union

Your Turn: Identify AI Around You

- Identify one AI system you use
- What data does it use?
- What decision or prediction does it make?

To conclude this module, spend a moment reflecting on what you've learned.

Think about one AI system you use daily. It could be related to entertainment, navigation, communication, or work.

Try to answer three questions:

- What kind of data does this system use?
- What does the system predict or decide?
- How might the quality of data affect its behaviour?

EcoMinds

Co-funded by the European Union

Introduction to Machine Learning for Environmental Data

How Machines Learn from Data?

In the previous module, we talked about what artificial intelligence is and why learning from data is important. Now we take one step further and ask: How does this learning actually happen?

We will look at:

- how machines learn from examples,
- why data quality and errors are part of learning,
- and how simple prediction scenarios work.

This module does not include mathematics or programming. The goal is to understand the idea of learning from data, using simple and familiar examples, including environmental ones.

EcoMinds

Co-funded by the European Union

Learning from Examples

- Machines learn from past data
- Examples include inputs and correct outcomes
- Learning means finding patterns

Machines do not learn like humans do. They do not understand explanations or instructions in words. Instead, machines learn from examples.

Each example contains two parts:

- input data (what the system sees)
- a correct outcome (what the correct answer is)

For example:

- a photo of a tree + label “tree”
- temperature and pollution values + label “high pollution”
- past traffic conditions + actual travel time

By analysing many such examples, the system looks for patterns that connect inputs with outcomes. Once these patterns are learned, the system can make predictions for new data it has never seen before.

EcoMinds

Co-funded by the European Union

Learning from Examples

- Weather forecast
  - Past weather data -> observed temperature tomorrow
- Email spam filter
  - Email text -> spam / not spam
- Air quality prediction
  - Sensor readings -> pollution level category
- Traffic prediction
  - Time + location + traffic history -> travel time

These examples show how learning functions across various fields. Although the applications appear different, the basic principle is the same: historical input data are matched with known outcomes.

The model analyzes these relationships and learns how particular patterns cause specific outcomes. Once trained, the system no longer requires the correct answer - it can estimate it on its own for new situations.

In environmental settings, this predictive ability is especially useful, as it enables early detection, planning, and risk mitigation.

EcoMinds

Co-funded by the European Union

Training Data and Learning

- Data used for learning
- Must represent real situations
- More data ≠ better data

Training data are the examples a machine uses to learn. These data describe real situations from the past. If the data do not represent reality well, the system will learn a distorted picture of the world. For example: If we train a system to predict air quality, but we only use data from one season, the system will not perform well during the rest of the year. Therefore, more data does not automatically mean better learning. A small but representative dataset is often more useful than a large but biased or incomplete one.

Training data should:

- cover different situations
- reflect real conditions
- include relevant variables

EcoMinds

Co-funded by the European Union

Training Data and Learning

- Weather data
  - Only summer data -> poor winter predictions
- Air pollution sensors
  - Sensors only in city centre -> poor predictions for suburbs
- Traffic data
  - Data collected only on working days -> poor weekend predictions
- School example
  - Learning only from easy exam questions -> poor performance on harder ones

These scenarios highlight the significance of representativeness. A model can only perform well in situations it has previously experienced. If specific scenarios are absent during training, its predictions become unreliable in those areas.

This is known as distribution mismatch—when training conditions differ from real-world conditions. In environmental modeling, this problem can greatly lower reliability, especially when systems operate differently across seasons, locations, or time periods.

A careful data collection strategy is therefore just as important as the algorithm itself.

EcoMinds

Co-funded by the European Union

From Data to Prediction

- Data contain patterns
- Learning extracts patterns
- Patterns are used for prediction

Data by themselves are just numbers, values, or measurements. Learning happens when a system looks at many examples and extracts patterns from them. These patterns describe relationships in the data.

For example:

- higher temperature often means higher electricity consumption
- certain weather conditions are linked to higher pollution levels

Once these patterns are learned, the system can use them to make predictions for new data. Importantly, the system does not memorise individual examples. Instead, it learns general relationships that can be applied to situations it has not seen before.

EcoMinds

Co-funded by the European Union

From Data to Prediction

- Weather prediction
  - Past temperatures -> predicted temperature tomorrow
- Air quality
  - Sensor data + weather -> pollution level category
- Energy consumption
  - Time + temperature -> expected electricity demand
- School analogy
  - Many solved exercises -> ability to solve a new one

These examples demonstrate how learned relationships are used in new situations. The system does not memorize past cases. It determines how variables are related.

When new data appears, the model uses those connections to estimate the most probable outcome. This ability to generalize forms the basis of predictive modeling.

EcoMinds

Co-funded by the European Union

Data Quality Matters

- Incomplete data
- Incorrect measurements
- Biased data sets

Machine learning systems depend heavily on data quality. If the data are incomplete, serious situations may be missing from the learning process. If the data contains incorrect measurements, the system will learn wrong relationships. Biased data are especially problematic.

Bias means that some situations are overrepresented while others are underrepresented or missing entirely. In such cases, the system may perform well in some situations but fail badly in others. This is why understanding and checking data quality is a crucial step before any learning takes place.

EcoMinds

Co-funded by the European Union

Data Quality Matters

- Environmental sensors
  - Broken sensor -> wrong pollution predictions
- Weather data
  - Missing extreme events -> poor risk prediction
- Traffic data
  - Data collected only during daytime -> poor night predictions
- School analogy
  - Learning only from correct answers -> no understanding of mistakes

These examples demonstrate how data limitations directly impact model reliability. If vital information is missing or incorrect, the system might arrive at misleading conclusions.

Machine learning does not recognize that data may be incomplete. It assumes that what it receives accurately reflects reality. Therefore, monitoring data quality is as crucial as creating the prediction model.

EcoMinds

Co-funded by the European Union

Errors Are Part of Learning

- Predictions are not always correct
- Errors help improve models
- Learning is an iterative process

Machine learning systems do not produce perfect results immediately. At the beginning, predictions are often inaccurate. This is normal and expected.

What makes learning possible is feedback. When the system makes a prediction, we can compare it with the correct outcome and measure the error. This error is then used to adjust the model so that future predictions become better. This process is repeated many times. Learning is therefore iterative: predict → measure error → improve → predict again.

Over time, performance improves, but it is never perfect.

EcoMinds

Co-funded by the European Union

Errors Are Part of Learning

- Weather forecast
  - First prediction is wrong -> model adjusted using new data
- Navigation apps
  - Travel time estimate improves as more trips are completed
- Air quality prediction
  - Model learns after comparing predicted vs. measured pollution
- Human analogy
  - Learning to ride a bike -> falling helps improve balance

These examples show how feedback promotes progress. When predictions differ from actual results, the system calculates the error and updates its internal parameters.

Over time, repeated comparison of expected and actual results decreases inaccuracies. This iterative refinement process enables machine learning systems to improve their accuracy over time.

EcoMinds

Co-funded by the European Union

Simple Learning Scenario

- **Input data**: temperature, humidity, pollution sensors
- **Learning**: system detects relationships and patterns
- **Output**: predicted value or risk category

Let's apply the learning scenario to a concrete environmental example. First, we have input data. These come from environmental sensors measuring temperature, humidity, and pollution levels over time.

Second, the system goes through a learning phase. It analyses past data and learns relationships between variables. For example, it may learn that certain weather conditions often lead to higher pollution levels.

Finally, the system produces an output. This output can be:

- a predicted pollution value, or
- a risk category such as low, medium, or high.

This allows environmental authorities or communities to react earlier and plan appropriate measures.

EcoMinds

Co-funded by the European Union

Limits of Learning from Data

- Models depend on past data
- Unexpected situations are difficult
- Learning does not guarantee correctness

Machine learning systems learn from past data. This means their knowledge is limited to situations that are represented in those data.

When something new or unusual happens, the system may struggle to respond correctly. For example, extreme weather events that rarely occurred in the past are difficult to predict accurately.

Learning from data also does not guarantee that predictions are always correct. AI systems provide estimates and probabilities, not certainty. Therefore, results must always be interpreted carefully, especially in environmental and sustainability contexts.

EcoMinds

Co-funded by the European Union

Limits of Learning from Data

- Extreme weather
  - Few historical examples -> weak predictions
- Sudden pollution incident
  - No similar past data -> late or inaccurate detection
- New sensor type
  - Model trained on old sensors -> unreliable results

These examples show the boundaries of data-driven learning. Machine learning depends on past observations to identify patterns. When events are rare, unexpected, or significantly different from training data, the accuracy of predictions drops.

Models are not meant to predict entirely new conditions. They work best when future data is similar to past experiences.

Understanding these boundaries is crucial for the responsible use of predictive systems.

EcoMinds

Co-funded by the European Union

Your Turn: Think Like a Model

- Choose input data
- Define what should be predicted
- Explain how learning would occur

To conclude this module, let's change perspective. Instead of thinking like a user of AI, try to think like a learning system. Imagine an environmental situation you are familiar with.

For example:

- predicting air quality,
- estimating temperature,
- identifying pollution risk.

First, decide what data would be used as input. These could be sensor measurements, weather data, or historical records. Second, decide what the system should predict or decide. This could be a value, a category, or a warning. Finally, explain how learning would occur. What kind of examples would the system need from the past? Focus on the idea of learning from data, not on technical details or algorithms.

EcoMinds

Co-funded by the European Union

Introduction to Machine Learning for Environmental Data

Machine Learning Tasks

In this module, we concentrate on machine learning tasks. A task explains what kind of question a machine learning system aims to answer.

Rather than questioning how the system functions internally, we focus on the desired output. We will concentrate on two basic types of tasks:

- classification, where the output is a category,
- regression, where the output is a number.

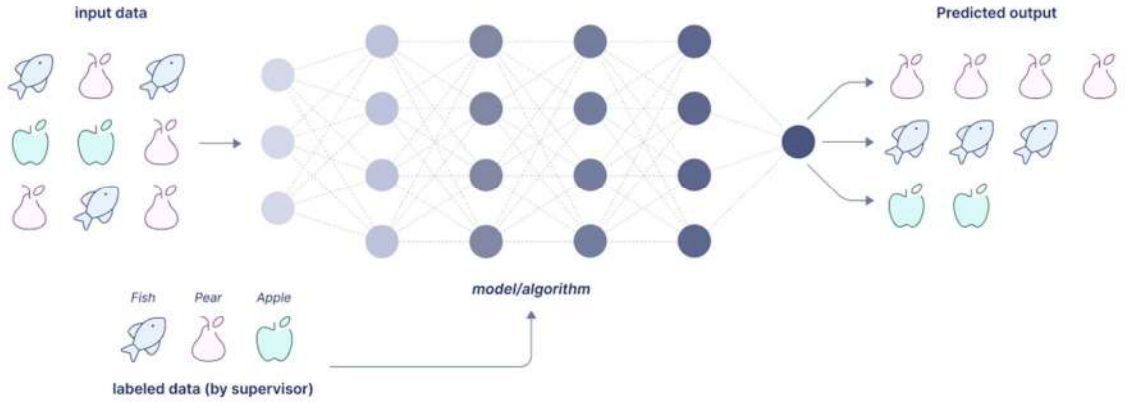
Throughout the module, we will use environmental examples to clarify these ideas and make them easier to understand. Importantly, in this module, we still do not focus on algorithms or calculations. Instead, we focus on selecting the appropriate task for a given problem.

EcoMinds

Co-funded by the European Union

Types of Machine Learning

- **Supervised learning** - learning from labeled data (***e.g. regression, classification***)



In supervised learning, the system learns from data with known correct answers. Each input is paired with a label that guides the learning process.

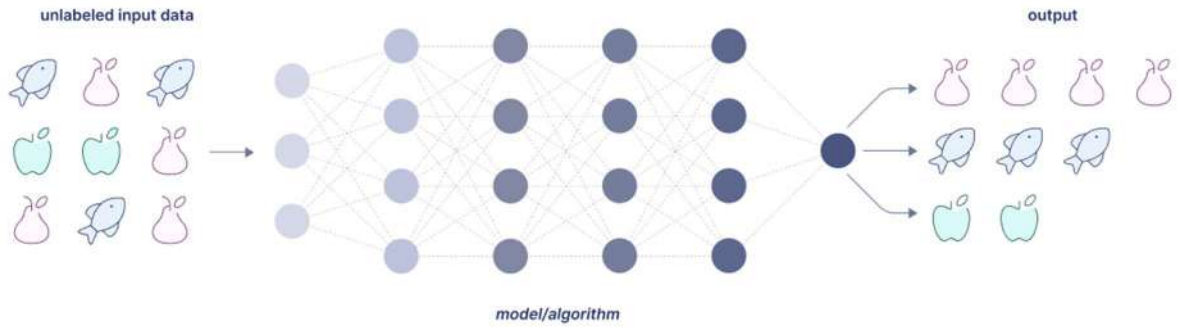
This is exactly the case for the two tasks we focus on in this module: regression and classification. In both cases, the system learns by comparing its predictions with known correct outcomes and improving over time.

EcoMinds

Co-funded by the European Union

Types of Machine Learning

- Unsupervised learning – finding patterns in data (*e.g. clustering*)



In unsupervised learning, the system processes data without predefined answers or labels. Instead of predicting a known result, the goal is to identify patterns or structure in the data.

The diagram illustrates this idea: the input data are unlabeled, and the system automatically groups similar items together. A common example of unsupervised learning is clustering, where data are organized into groups based on similarity.

EcoMinds

Co-funded by the European Union

What Is a Machine Learning Task?

- Defines what the model should predict
- Determines the type of output
- Comes before choosing a model

Before discussing algorithms or models, we need to clearly define the machine learning task. A task explains what the system is expected to predict or decide. This step is one of the most crucial in machine learning. If the task is unclear or poorly defined, even highly advanced algorithms will not produce useful results.

Defining the task also determines the type of output:

- are we predicting a category, or
- are we predicting a number?

Only after the task is clearly defined does it make sense to choose a specific model or algorithm.

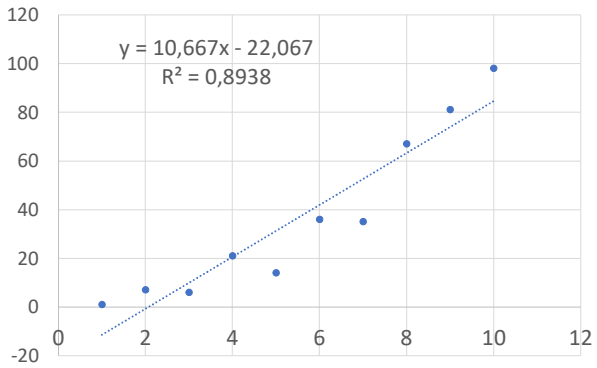
EcoMinds

Co-funded by the European Union

(Simple) Linear Regression

- Predicts a number
- Learns a trend from data
- Answers: *„how much?”*

$$Y = b0 + b1*X$$



Let’s examine a very basic example of regression, known as simple linear regression. The word simple is important because it indicates we are analyzing the relationship between one input and one output.

For example:

- Input (X): temperature
- Output (Y): energy consumption

We gather numerous examples from the past. Each example includes a temperature value and the corresponding energy usage. Simple linear regression aims to find a straight line that best describes how the output varies as the input changes. This line shows a trend in the data. After the trend is learned, we can use it to predict a numerical value for new input data. For example, if we know today’s temperature, the model can estimate how much energy will be needed today.

EcoMinds

Co-funded by the European Union

(Simple) Linear Regression - Calculation

Suppose we have the following data:

- X = [1, 2, 3, 4, 5]
- Y = [2, 4, 5, 4, 5]

Step 1: Calculate the mean of X and Y

- mean(X) = (1 + 2 + 3 + 4 + 5) / 5 = 3
- mean(Y) = (2 + 4 + 5 + 4 + 5) / 5 = 4

Step 2: Calculate the difference of each value of X and Y from their respective means

- differences of X = [1-3, 2-3, 3-3, 4-3, 5-3] = [-2, -1, 0, 1, 2]
- differences of Y = [2-4, 4-4, 5-4, 4-4, 5-4] = [-2, 0, 1, 0, 1]

Step 3: Calculate the sum of products of the differences of X and Y

Sum of products of differences =
(-2 * -2) + (-1 * 0) + (0 * 1) + (1 * 0) + (2 * 1) = 6

Let's quickly examine a small numerical example to understand what occurs behind the scenes. We begin with two basic data sets: one input variable X and one output variable Y.

First, we find the average, or mean, of X and Y. This provides a reference point for the data. Next, we examine how each value varies from the average. These variations indicate whether a value is above or below the mean.

Next, we multiply the differences between X and Y and then add them.

EcoMinds

Co-funded by the European Union

(Simple) Linear Regression - Calculation

Step 4: Calculate the sum of squares of differences of X

sum of squares of difference of X =

$=(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2 = 10$

Step 5: Calculate the slope of the regression line (b1)

$$b1 = \frac{\text{sum of products of differences of x}}{\text{sum of squares of differences of X}}$$

b1 = 6 / 10 = 0.6

Step 6: Calculate the intercept of the regression line (b0)

b0 = mean(Y) - b1 * mean(X) = 4 - 0.6 * 3 = 2.2

Step 7: Write the equation of the regression line

Y = b0 + b1 * X = 2.2 + 0.6 * X

In the next steps, we use the values we've already calculated to describe the regression line. First, we examine how spread out the values of X are. This is done by squaring the differences of X and summing them up.

Next, we determine the slope of the regression line. The slope indicates how much Y changes when X increases by one unit. In this example, the slope is 0.6, which shows that when X goes up, Y tends to go up as well.

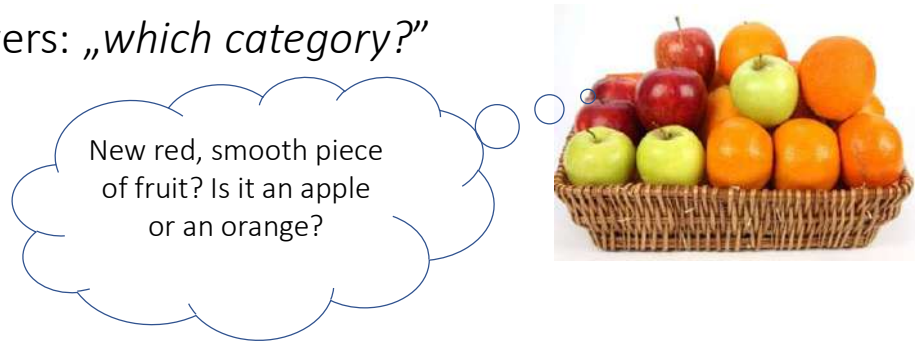
Next, we calculate the intercept. The intercept indicates where the regression line crosses the Y-axis. Then, we write the equation of the line. This equation represents the trend learned from the data and can be used for simple predictions.

EcoMinds

Co-funded by the European Union

Classification - Predicting Categories

- Assigns data to categories
- Output is a label or class
- Answers: *„which category?"*



In many situations, however, we don't need an exact value. Instead, we want to determine which category something falls into. This type of task is called classification. In classification, the output is not a number but a label or a class.

These classes are typically predefined, such as low, medium, or high, or good and bad. Classification is often used when the goal is decision-making rather than precise measurement.

The key question in classification is simple: Which category does this new data point belong to?

EcoMinds

Co-funded by the European Union

Classification - Predicting Categories

- Air quality -> good / moderate / poor
- Flood risk -> low / medium / high
- Land cover -> forest / water / urban
- Warning systems -> safe / warning / danger

These examples demonstrate real-world environmental classification tasks. In each case, the goal is to assign an observation to a specific risk or condition level.

Unlike regression, where we predict exact values, classification reduces complex measurements into clear decision categories.

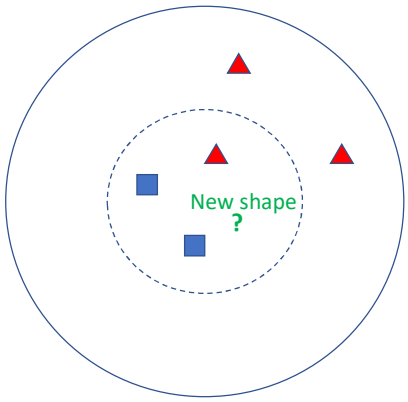
This is especially helpful in environmental management, where rapid interpretation and clear communication of risk levels are crucial.

EcoMinds

Co-funded by the European Union

Classification: k-Nearest Neighbours (KNN)

- Choose the number of nearest neighbor's „k”
- Calculate the distance between the new data point and all the data points in the training dataset.
- Select the "k" data points that are closest to the new data point.
- The new data point is assigned the class that is most common among the "k" nearest neighbors.



Imagine you're in a room filled with triangles and squares, and you see a new shape. You want to decide whether this new shape is more like a triangle or a square. This is the basic idea behind k-nearest neighbors, or KNN. First, we select a number called k, which indicates how many nearby examples we will consider. In this illustration, the dashed circle shows k = 3, and the larger solid circle indicates k = 5. Next, we identify which existing shapes are closest to the new shape. If we use k = 3, the three nearest neighbors include two blue squares and one red triangle. Since the majority are blue squares, the new shape is classified as a square. If we increase the value to k = 5, the five nearest neighbors now include three red triangles and two blue squares. In this case, the new shape is classified as a triangle.

This example shows two important ideas: classification is based on similarity, and the choice of k can influence the result.

Again, we do not focus on how distance is calculated. What matters is the intuition: new data points are classified based on nearby examples.

EcoMinds

Co-funded by the European Union

Clustering

- The goal is to group objects that are more alike with each other than with objects in other clusters.
- Think of it like organizing a messy room by creating neat piles of similar things.

The goal of clustering is to group similar data points together so that items within the same group are more like each other than those in different groups.

A simple way to understand clustering is like organizing a messy room. Without labels, you instinctively group similar items together: clothes with clothes, books with books.

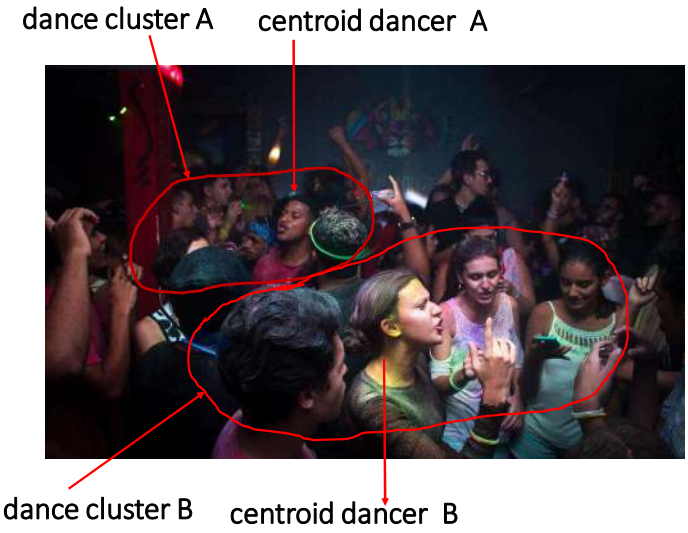
Clustering functions similarly. It helps us identify structure and patterns in data without labels.

EcoMinds

Co-funded by the European Union

Clustering: K-means

- K-means groups similar data points together.
- It assigns each point to the nearest cluster centroid.
- The algorithm continues until the clusters are stable.



One of the most common clustering algorithms is K-means. K-means groups data points based on similarity and distance. First, we choose a number K, which indicates how many groups we want.

Each group has a center called a centroid. Every data point is assigned to the closest centroid. The centroids are then updated to reflect the center of their groups. This process repeats until the groups stop changing.

Think of it like a dance floor. People naturally gather around the dancer closest to them. As people move, the center of each group shifts. Eventually, the groups stabilize. Again, the goal here is intuition, not calculation or implementation.

EcoMinds

Co-funded by the European Union

Same Data, Different Tasks

- Same input data
- Different prediction goals
- Different task types

- Example:
  - Sensor data -> pollution value (regression)
  - Sensor data -> pollution category (classification)

The same data can be applied to various machine learning tasks. Imagine we gather air quality sensor data:

- temperature
- humidity
- pollution measurements

If our goal is to determine the exact pollution level, we use a regression task. However, if our goal is to aid decision-making, we might opt for a classification task, such as low, medium, or high pollution.

The data remain unchanged. What varies is what we expect from the system.

EcoMinds

Co-funded by the European Union

Your Turn: Choosing the Right ML Task

- Choose an environmental problem
- Decide: regression or classification?
- Explain your choice

To conclude this module, consider an environmental issue you know well, such as air quality monitoring, temperature prediction, flood risk assessment, or energy use.

The first step is to determine what kind of output is required. If the goal is to predict an exact numerical value, the task is regression. If the goal is to assign a category or risk level, the task is classification.

Try to explain your choice in one or two sentences, focusing on the type of output rather than the algorithm or technical details.

EcoMinds

Co-funded by the European Union

Introduction to Machine Learning for Environmental Data

Environmental Data in Practice

In this module, we take a step closer to real-world practice. We focus on environmental data: where they come from, how they are collected, and the challenges they present.

You will also see how actual environmental data can be used to create a straightforward machine learning scenario, similar to the ones we discussed earlier.

The goal of this module is not to focus on specific tools or software, but to understand how real data relates to machine learning concepts in practice.

EcoMinds

Co-funded by the European Union

Where Do Environmental Data Come From?

- Sensors and monitoring stations
- Satellites and remote sensing
- Weather stations
- Field measurements

Environmental data are gathered from various sources, depending on what we aim to observe and the scale involved. Sensors and monitoring stations continuously record variables such as temperature, air pollution, water quality, or noise levels, often generating significant amounts of data over time. Satellites and remote sensing systems observe the Earth from above, enabling us to monitor large areas such as forests, oceans, and atmospheric conditions—something impossible to do with only ground-based measurements. Weather stations use multiple sensors to track atmospheric conditions and are among the most common and long-established sources of environmental data. In addition to automated systems, environmental data are also collected through field measurements, where experts manually gather samples or observations at specific sites. Together, these various sources generate diverse data sets that serve as the foundation for environmental analysis and machine learning applications.

EcoMinds

Co-funded by the European Union

Sensors and Monitoring Systems

- Continuous data collection
- High data volume
- Real-time or near real-time data

Many environmental data sets are produced by sensor and monitoring systems that run continuously for long periods. Unlike occasional measurements, sensors collect data at regular intervals, sometimes every few seconds or minutes, which quickly results in very large amounts of data.

This continuous flow of information enables monitoring changes over time, identifying trends, and detecting sudden events like pollution spikes or extreme weather. Often, sensor data is transmitted in real time or nearly real time, allowing for prompt analysis and quick responses during critical situations.

Because of this, environmental monitoring systems are not only focused on data collection but also on making sure that data can be processed, stored, and interpreted efficiently to support monitoring and decision-making in practice.

EcoMinds

Co-funded by the European Union

Open Data and Public Datasets

- Publicly available data
- Provided by governments and institutions
- Used for research and education

Besides data collected by individual projects or organizations, a large amount of environmental data is available as open or public data. These datasets are published by governments, research institutions, international organizations, and environmental agencies to support transparency, scientific research, and education.

Open environmental data enable students, researchers, and practitioners to work with real-world information without needing to collect data themselves. This allows for analyzing long-term trends, comparing different regions, and testing ideas using authentic data sources.

For educational purposes, open data are especially valuable because they expose learners to the complexity and diversity of real environmental information rather than simplified or artificial examples.

EcoMinds

Co-funded by the European Union

Open Data: Opportunities and Challenges

- Easy access to real data
- Large and diverse data sets
- Missing values and errors
- Data may require cleaning

Open environmental data present many opportunities, especially for learning and experimentation, because they provide easy access to large and diverse data sets from real monitoring systems. These data enable us to explore realistic scenarios, analyze long-term trends, and develop meaningful machine learning tasks without the need to collect data ourselves.

At the same time, working with open data presents challenges that should not be overlooked. Data sets can have missing values, measurement errors, or inconsistent formats, and they are often collected under different conditions or following different standards.

Because of this, open data usually require careful inspection and cleaning before they can be reliably analyzed or used for machine learning. Understanding both the opportunities and the limitations of open data is an important step toward responsibly using environmental data in practice.

EcoMinds

Co-funded by the European Union

Environmental Data Are Not Perfect

- Measurement errors
- Missing or noisy values
- Differences between locations

In real-world environmental monitoring, data are rarely perfect or clean. Measurements can be impacted by sensor calibration problems, environmental factors, or technical issues, which can introduce errors or noise into the data. Sometimes, values may be missing because sensors temporarily stop functioning or data transmission fails.

Environmental data often varies significantly between locations, even when measuring the same variable, because local conditions such as geography, weather, or human activity influence the results.

Recognizing that these imperfections are a natural part of real data is crucial because machine learning models and statistical methods can only be as reliable as the data they are based on. This is why understanding data limitations is just as important as understanding learning algorithms.

EcoMinds

Co-funded by the European Union

From Data to a Learning Scenario

- Choose input data
- Define the prediction goal
- Decide the task type

Before applying machine learning to environmental data, it is important to clearly define the learning scenario. This begins with choosing which data will be used as input, such as sensor measurements, satellite observations, or weather variables.

The next step is to choose what the system should predict, such as a future value, a risk level, or a category. Based on this choice, we determine whether the task is regression or classification, as discussed in the previous module.

These choices are made before selecting any algorithm and significantly influence how the data will be interpreted and used. Setting the learning scenario this way helps ensure that machine learning is applied intentionally and that the results are relevant in a real environmental context.

EcoMinds

Co-funded by the European Union

Example Scenario – Linear Regression Model

- Context and Data Source
  - Average annual air temperatures
  - City: Zadar, Croatia
  - Period: 1991–2023
  - Data source: Historical weather data retrieved from: Open-Meteo Historical Weather API

YEAR	AVG_YEAR_TEMP	YEAR	AVG_YEAR_TEMP
1991	13,6	2007	15,1
1992	14,6	2008	15,2
1993	14,1	2009	14,9
1994	15,1	2010	14,2
1995	13,8	2011	15,2
1996	13,6	2012	15,0
1997	14,2	2013	14,9
1998	14,3	2014	15,4
1999	14,6	2015	15,3
2000	15,4	2016	15,1
2001	14,7	2017	15,0
2002	14,7	2018	15,6
2003	15,2	2019	15,5
2004	14,5	2020	15,4
2005	13,7	2021	15,1
2006	14,6	2022	15,7
		2023	15,8

The analysis is based on average annual air temperatures recorded for Zadar, Croatia, covering the period from 1991 to 2023. These values represent yearly averages rather than daily or monthly data, making them suitable for identifying long-term climate trends.

All temperature data were sourced from historical weather records provided by the Open-Meteo Historical Weather API. This service offers open-access meteorological data and is commonly used for educational, research, and applied data analysis purposes.

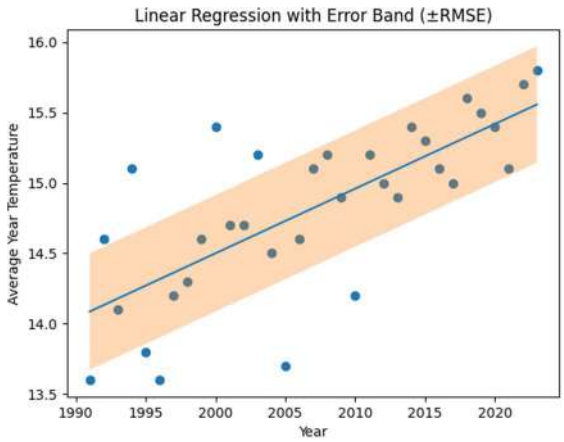
On the right side of the slide, we see the complete dataset used in the analysis, listing each year along with its corresponding average annual temperature. This dataset serves as the input for the regression model shown in the following slides.

EcoMinds

Co-funded by the European Union

Model and Prediction

- Estimated warming trend:
  - +0.046 °C per year
- Predicted average temperature for 2024:
  - 15.6 °C
- Model uncertainty (RMSE):
  - ±0.4 °C
- Measured average temperature in 2024:
  - 15.9 °C
- Prediction error:
  - +0.3 °C (within expected error range)



Average temperature =0.046·YEAR−77.50

The estimated warming trend from the regression analysis is about 0.046 degrees Celsius per year, showing a consistent long-term rise in the average annual air temperature for the city of Zadar.

Based on this trend, the model predicts an average annual temperature of 15.6 degrees Celsius for 2024. However, as with any statistical model, there is inherent uncertainty. This uncertainty is measured using the root mean squared error, which in this case is about plus or minus 0.4 degrees Celsius.

The average temperature measured for 2024 was 15.9 degrees Celsius. The difference between the predicted and observed values is 0.3 degrees Celsius, which is well within the model's expected error range.

The chart on the right shows the regression line along with an error band that indicates the typical prediction uncertainty. The scatter of points around the line highlights the natural year-to-year variability, which a simple linear model cannot fully capture.

EcoMinds

Co-funded by the European Union

Example Scenario-Temperature-Based Clustering

- Same dataset as used for the linear regression model
- Clustering method: k-means (k = 3)
- Clustering based only on temperature values
- Time (year) not used as an input variable
- Goal
  - Identify distinct temperature regimes
  - Explore changes in temperature patterns without assuming a linear trend

The analysis uses the same dataset as the linear regression example, which consists of average annual air temperatures measured in Zadar, Croatia, from 1991 to 2023.

Although the underlying data remains unchanged, a different analytical approach is used. Instead of modeling temperature as a function of time, k-means clustering groups years solely based on their average temperature values.

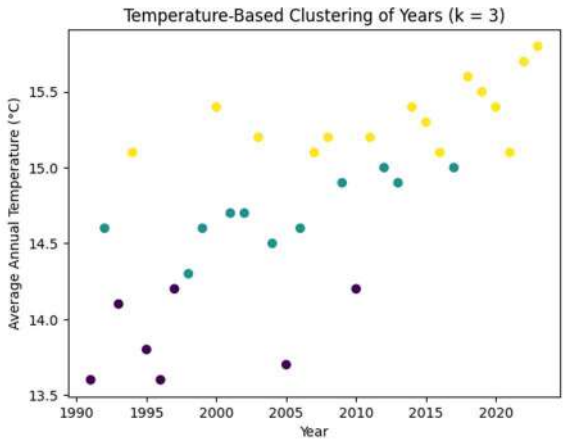
By omitting time as an explicit input variable, this method lets temperature regimes arise naturally from the data. The goal is to determine whether years with similar thermal patterns form meaningful groups and if these groups display a temporal pattern when viewed over time.

EcoMinds

Co-funded by the European Union

Results - Temperature Clusters (k = 3)

- Cool years
  - Mean temperature ≈ 13.9 °C
- Moderate years
  - Mean temperature ≈ 14.7 °C
- Warm years
  - Mean temperature ≈ 15.3 °C
- Key Observation:
  - Earlier years are predominantly classified as cool
  - Recent years are increasingly classified as warm
  - Indicates a shift between temperature regimes over time



The temperature-based clustering results reveal three separate groups of years (cool, moderate, and warm temperature regimes) characterized solely by average annual temperatures.

Although time isn't directly used as an input in the clustering process, the resulting clusters closely align with the dataset's temporal structure. Earlier years are mainly classified as cooler, while recent years are more often associated with the warm cluster.

This pattern offers an intuitive and visually compelling confirmation of long-term warming, independent of linear regression assumptions. At the same time, it shows how unsupervised learning techniques can be effectively used on climate data to explore underlying structures and detect regime shifts.

EcoMinds

Co-funded by the European Union

Your Turn: Explore Environmental Data

- Use the provided aggregated data
- Or retrieve data from Open-Meteo
- Design a simple analysis scenario

To conclude this module, you're encouraged to explore environmental data independently. You can begin with the aggregated temperature data provided here and consider how to analyze it using either regression or clustering.

Alternatively, you can gather historical weather data for a city of your choice using the Open-Meteo API and build your own simple dataset.

The goal is not to create a perfect model, but to practice defining an analysis scenario, selecting input data, and deciding whether a regression or clustering approach makes sense. Concentrate on understanding the data and the purpose of the analysis rather than on technical implementation details.

EcoMinds

Co-funded by the European Union

Introduction to Machine Learning for Environmental Data

Responsible Use of AI

So far, you have seen how environmental data are collected, analyzed, and used in simple AI models. However, even the most accurate models can cause problems if they are used without understanding their limitations or ethical implications.

This module encourages you to think critically about:

- how environmental data are used,
- how AI results should be interpreted,
- and why human responsibility remains essential when AI supports environmental decisions.

This is not a technical module. Instead, it helps you develop awareness and judgment when using AI for environmental protection and sustainability.

EcoMinds

Co-funded by the European Union

Why Responsibility Matters?

- AI influences real-world decisions
- Mistakes can affect people and systems
- Environmental decisions have long-term impact

AI systems are increasingly used to assist decision-making in various fields like healthcare, transportation, finance, and public services. These decisions can impact people's safety, resources, and quality of life.

When AI systems make errors or are used without proper understanding, the outcomes can include incorrect decisions, loss of trust, or inefficient resource use. That's why responsibility is crucial in any AI application.

In environmental settings, this responsibility becomes even more crucial. Environmental decisions often impact entire ecosystems, large populations, and future generations. For example, AI-based predictions of pollution, flooding, or climate trends can influence environmental policies or emergency responses.

Because environmental systems are complex and slow to recover, mistakes can have lasting effects. Responsible use of AI involves combining AI outputs with human judgment, environmental understanding, and ethical considerations.

EcoMinds

Co-funded by the European Union

Data Privacy and Protection

- Some data may be sensitive
- Data must be collected responsibly
- Privacy must be respected

At first glance, data used in AI systems might seem harmless, especially environmental data like temperature, air quality, or rainfall. However, not all data are neutral or anonymous. Generally, data privacy refers to the responsible collection, storage, and use of information in accordance with ethical principles and legal regulations. People and communities should not be harmed or exposed to data misuse.

In environmental applications, privacy concerns may still emerge. For example:

- collecting detailed location data when only regional trends are needed,
- storing long-term historical data without a clear purpose,
- combining environmental data with other data sources in ways that were not originally intended.

Responsible data use involves understanding who or what might be impacted by data collection and analysis. Respecting privacy helps foster trust in environmental monitoring systems and AI-driven decision support.

EcoMinds

Co-funded by the European Union

Data Reliability and Trust

- AI depends on data quality
- Incorrect data lead to incorrect results
- Reliability must be evaluated

AI systems depend completely on the data they receive. If the data are wrong, incomplete, or outdated, the AI's results will also be unreliable. This rule applies to all AI systems, no matter how advanced the model is.

A common misconception is that more data automatically means better results. In reality, unreliable or poorly collected data can decrease the quality of predictions, even if large volumes of data are available.

In environmental applications, the reliability of data is particularly crucial. Environmental data typically originate from sensors, monitoring stations, or satellites, which may: produce measurement errors, fail or stop working temporarily and be unevenly distributed across regions.

For example, an AI system predicting air pollution may perform well in areas with many sensors but poorly in regions with limited monitoring. Evaluating data reliability helps users understand how much confidence they should place in AI-supported environmental predictions.

EcoMinds

Co-funded by the European Union

Bias in AI Systems

- Bias comes from data
- Uneven data lead to uneven results
- Bias can affect decisions

Bias in AI systems generally stems from the data used for training rather than the technology itself. When certain situations, locations, or conditions are underrepresented in the data, the AI model may perform less effectively in those areas. In general, bias means an AI system may perform better in some scenarios than others, even if unintended. This can result in uneven or unfair outcomes.

In environmental applications, bias may occur when: sensors are placed mainly in urban areas, data are collected only during certain seasons and extreme or rare environmental events are missing from the data.

For example, an AI model trained mainly on data from large cities might not accurately predict air quality in rural or coastal areas. Recognizing bias helps users interpret AI results more carefully and avoid overconfidence in predictions.

EcoMinds

Co-funded by the European Union

Limitations of AI

- AI learns from the past
- Unexpected situations are difficult
- AI does not understand context

AI systems identify patterns from past data. This means their predictions rely on what has already occurred, not on a genuine understanding of the world. Generally, AI works best when future situations resemble past ones. However, unexpected or rare events are hard to predict, especially if they are not well represented in the data.

In environmental contexts, this limitation is especially significant. Environmental systems are complex and continually evolving. Examples include:

- extreme weather events,
- sudden pollution incidents,
- long-term climate changes that go beyond historical patterns.

AI systems do not grasp causes, effects, or context like humans do. They cannot explain why something happens or determine if a result makes sense. Therefore, AI predictions should always be interpreted with caution and validated by human expertise and environment knowledge.

EcoMinds

Co-funded by the European Union

AI as a Decision-Support Tool

- AI supports, not replaces, humans
- Humans remain responsible
- Expert judgement is essential

AI systems should be viewed as tools that assist human decision-making, not as systems that replace human responsibility. AI can analyze large amounts of data, identify patterns, and offer predictions, but it cannot make value-based or ethical decisions.

In general, humans remain responsible for interpreting AI results, verifying their accuracy, and deciding how to respond. This is true across all application areas. In environmental contexts, expert judgment is especially crucial. Environmental decisions often require:

- knowledge of local conditions,
- understanding of ecological relationships,
- consideration of social and economic impacts.

For example, if an AI system predicts high pollution levels, experts must decide how to respond, when to issue warnings, and which measures are appropriate. AI provides information, but humans decide and take responsibility.

EcoMinds

Co-funded by the European Union

Example: Environmental Decision Support

- AI predicts pollution levels
- Experts interpret results
- Decisions consider multiple factors

Imagine an AI system used to forecast air pollution levels in a city. The system analyzes data from air quality sensors, weather conditions, and historical pollution patterns to estimate pollution levels for the following day.

The AI can forecast when pollution levels are likely to be high. However, this prediction by itself is not a decision. Environmental experts need to interpret the results and consider other factors, such as:

- current weather changes,
- local traffic patterns,
- industrial activity,
- vulnerable groups (children, elderly).

For example, if AI predicts high pollution, experts may decide whether to issue a public warning, restrict traffic, or take no action if uncertainty is high. This illustrates how AI assists decision-making, while humans remain responsible for the final decisions.

EcoMinds

Co-funded by the European Union

Your Turn: Think Critically

- Identify a possible risk of AI use
- Explain how it could be reduced
- Discuss the role of humans

Now it's your turn to think critically about responsible AI use. Pick a simple example of AI used in an environmental setting, like air quality monitoring, flood prediction, or energy usage forecasting.

First, identify a potential risk, for example:

- unreliable or incomplete data,
- bias due to missing locations or seasons,
- overreliance on AI predictions.

Next, consider ways to reduce this risk. This could involve enhancing data quality, merging multiple data sources, or consulting experts for interpretation.

Finally, reflect on humans' role. Who should interpret the results? Who should make the final decision? This activity helps you understand why responsibility and human oversight are crucial in AI-supported environmental systems.