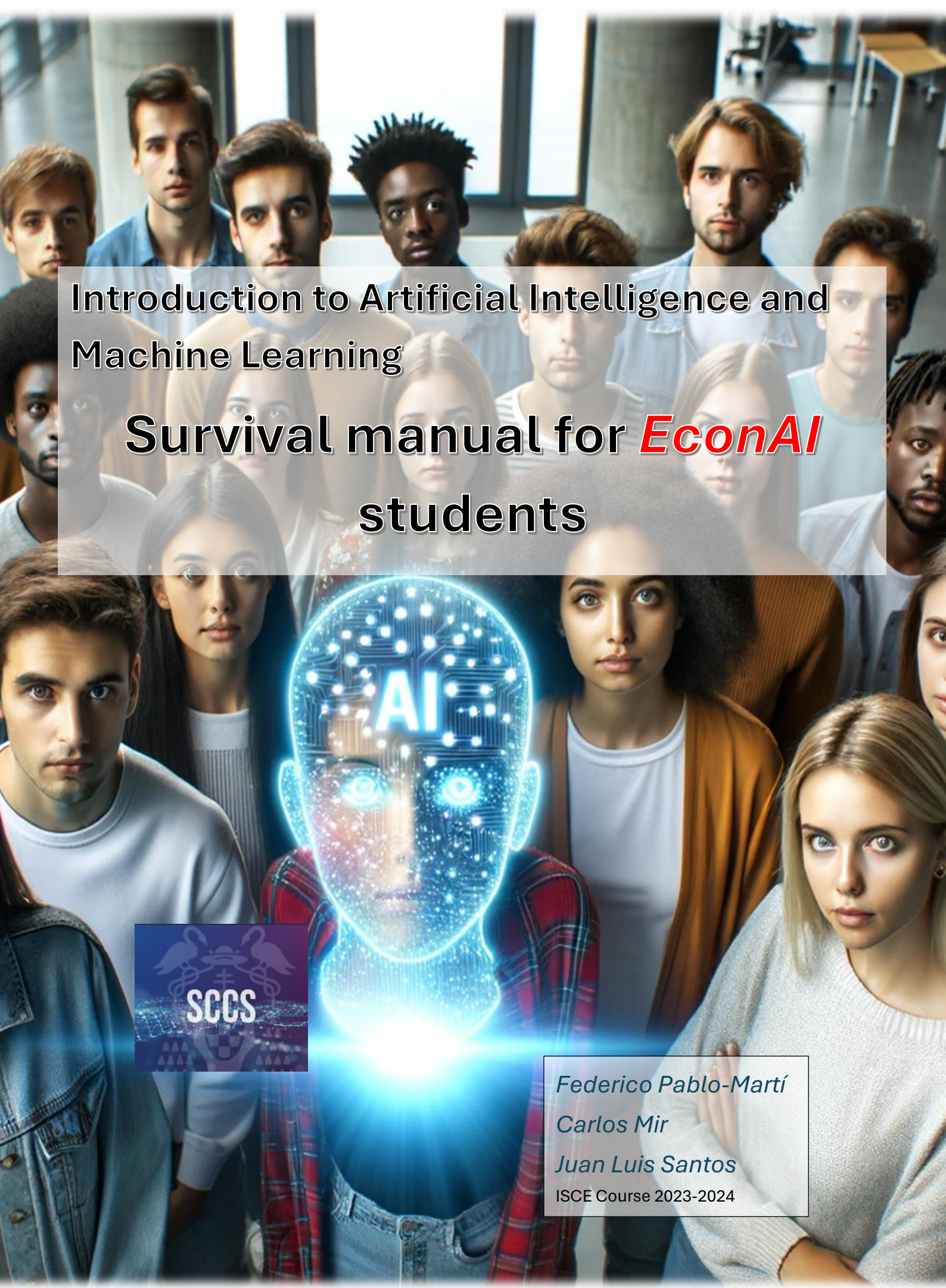


Introduction to Artificial Intelligence and Machine Learning

Survival manual for *EconAI* students

Federico Pablo-Martí

Carlos Mir

Juan Luis Santos

ISCE Course 2023-2024

Index

Foreword

If you are studying economics and feel a bit overwhelmed by all that Artificial Intelligence (AI) and the lesser-known Machine Learning (ML), and they are like an indecipherable labyrinth, this manual is designed for you. Here we will not complicate you with unnecessary technicalities; rather, we will show you how these technologies are not only for computer experts but also incredible tools, which can also enhance your skills as the economist of the future.

Why should you care about this? Well, economics is no longer just charts and theories; it's now full of massive data that needs to be understood and analyzed. And this is where AI and ML come in. Throughout these pages, we are going to explain how these technologies are changing the way we understand economics, from predicting market trends to better understanding human behavior, even within the intricate behavior of irrational exuberance of economic agents, and the Animal Spirits of the market or economic psychology.

In short, this manual is your friendly guide to understanding how AI and ML can be your allies in the world of what we will call the ***EconAI***. After reading it, you will be one step ahead in understanding and using these technologies that are redefining our era.

It's not just about understanding how AI and ML work, but about incorporating these techniques into your professional toolbox. By the end of this journey, you will not only be up to speed on the latest technology trends, but you will also feel empowered to apply this knowledge effectively in your day-to-day economic analysis. Get ready to unlock a new level of economic acumen with the help of AI and ML and discover how these technologies can be your allies in propelling your career into the future!

1. Introduction

On the threshold of a new technological era, we are at a turning point in human history, where artificial intelligence (AI) and *machine learning* (ML) are redefining not only how we interact with technology, but also how we understand and model economic phenomena. For students of Economics, this is an exciting and pivotal time to embark on the study of these emerging disciplines.

Although we will talk about it throughout this text, it is interesting to highlight what artificial intelligence (AI) mean in the context of the development of new technologies. AI refers to the ability of machines to perform tasks that normally require human intelligence. These tasks include learning, reasoning, visual perception, natural language recognition, decision-making, and problem-solving. AI is based on algorithms and mathematical models that enable machines to learn patterns from data, adapt to new situations, and improve their performance over time. Machine learning (ML) is a subfield of AI that focuses on developing algorithms and models that allow computers to learn patterns and perform specific tasks without explicit programming. Instead of following detailed instructions, machine learning systems use data to improve their performance on a particular task as they are exposed to more information.

The history of AI and ML is a fascinating narrative of the human quest to create machines that not only process information, but also learn and make decisions. From the early days of computers in the 1940s, through the invention of the perceptron in the 1950s, to the rise of neural networks and deep learning in the 21st century, we have witnessed a radical transformation in what machines can do (Ponce, 2010). The new projects that the private sector is embarking on are a definite spur to their implementation and development in society and workflows.

In the field of economics, AI and ML offer powerful tools for analyzing large data sets, predicting market trends, and better understanding the complexities of human behavior. Economists have traditionally relied on models based on assumptions and simplifications. However, AI and ML open new avenues to model the economy in more detailed and accurate ways, reflecting the complexity of the real world, being able to analyze data in a massive way that allows obtaining relationships that simplified models could not reach to understand (Olguín Gallardo, 2018).

As we move into the 21st century, economics as a discipline faces unique and complex challenges: from managing global economic crises to understanding the impact of climate change on economic systems. This is where AI and ML come into play, offering advanced methods for analyzing data, predicting economic events, and simulating economic policies. These tools allow economists to explore scenarios, identify patterns, and make informed decisions based on large-scale empirical evidence, entering a terrain that is sure to ring a bell, the so-called Big Data as a starting point.

For students in economics-related degrees, understanding and applying AI and ML concepts is crucial not only to keep up with current trends but also to be prepared for the challenges and opportunities of the future. In this short primer, we will explore the

fundamentals of these disruptive technologies, their application in economics, and how they can radically transform our understanding and approach to economic problems.

The goal is not to learn algorithms and models, but to equip us with a new lens through which to view the economic world: a lens that is more accurate, more analytical, and, most importantly, more prepared for the unpredictable future that awaits us.

2. Machine learning: The island of innovation

In today's technology landscape, two terms are constantly emerging: artificial intelligence (AI) and machine learning (ML). Although they are often used interchangeably, there are fundamental differences between the two concepts that are crucial to understand.

AI can be seen as the vast ocean of possibilities in which ML is just an island. It is a field of computer science dedicated to the creation of systems capable of performing tasks that require intelligence. These tasks can include decision-making, problem-solving, understanding human language, and more. From programs that play chess to virtual assistants like Siri and Alexa, AI encompasses a wide range of applications that mimic or even surpass human capabilities.

Within the vast domain of AI, ML represents a specific methodology. It is the science of getting computers to learn and act like humans, improving their learning over time-based on experience and data. Unlike traditional AI systems, which are programmed specifically to perform tasks, ML systems are trained using large data sets and algorithms that enable them to learn to perform the task (Dimonopoli, 2022).

2.1. Types of learning in machine learning.

In this section, we will explore the different learning methods that are applied in the field of machine learning. Each type of learning offers a different approach to how computational models process and learn from data. These methods are fundamental to understanding how algorithms improve their accuracy and efficiency over time, adapting to various scenarios and requirements. We will start with supervised learning, one of the most common and fundamental approaches in machine learning, and then delve into other forms of learning that complement and broaden the spectrum of possibilities in this field.

2.1.1. Supervised learning

In this type of learning the models are trained using labeled data. This means that each example in the data set is paired with the correct answer. The model learns from this data and makes predictions or judgments about new unseen data. It is like a student learning with the guidance of a teacher who provides examples and correct answers. It allows algorithms to make predictions or classifications. It is a fundamental part of machine learning, which is why it is also sometimes referred to as supervised machine learning.

This approach is based on the idea that, by exposing the model to sufficient and varied examples along with their correct answers, you will learn to identify underlying patterns and relationships that will enable you to make accurate predictions about new data. The quality of the model is determined by testing procedures through cross-validation, confidence

probability, accuracy, or hit rate. To the extent that we can count on more data for practice, better results will be obtained.

In supervised learning, algorithms are basically divided into two groups: classification and regression. The former involves assigning the input data into specific categories. It recognizes certain characteristics and patterns in the data and seeks to find similarities. In the case of the latter, it tries to find relationships between dependent and independent variables.

Principles of supervised learning (Cunningham et al., 2008):

1. **Generalization:** The main objective is for the model to be able to generalize from training data, i.e., to be able to apply what has been learned to previously unseen examples.
2. **Error minimization:** During training, the aim is to minimize the difference between model predictions and actual responses. This process is known as error minimization or loss function minimization.

Supervised learning has a wide range of applications in various fields, including:

1. **Economics and finance:** Stock index forecasting (Patel et al., 2015), credit risk assessment (Borrero-Trigueros & Bedoya-Leiva, 2020), and market trend analysis (Cordero Torres, 2022).
2. **Health:** Diagnosis of diseases from medical images or clinical data.
3. **Marketing:** Customer segmentation and personalization of offers based on buying patterns (Kotler et al., 2022).

Some examples:

1. **Image recognition:** In image recognition, a model is trained with many labeled images (e.g., 'cat' or 'dog') so that it can identify these categories in new images (Oliva Rodriguez, 2018).
2. **Weather prediction:** Using historical weather data, a model can learn to predict future weather conditions.
3. **Fraud detection:** In the banking and auditing sector, models can learn to identify fraudulent transactions by analyzing patterns in past transaction data or detect "creative" accounting criteria that decorate corporate information beyond the regulatory accounting framework (Nisbet et al., 2009).

Challenges and considerations:

- **Data quality:** The quality and representativeness of the training data set are critical. A model can only be as good as the data it is trained on.
- **Over-fitting:** There is a risk that the model fits the training data too closely, losing the ability to generalize to new data. This is known as overfitting.

Supervised learning is a powerful tool that enables models to learn from past examples to make informed and useful predictions about future data. Its application spans a wide range of industries and disciplines, providing valuable insights and improving data-driven

decision-making. For students of Economics, understanding these principles and applications opens a world of possibilities for advanced economic analysis and evidence-based policymaking.

2.1.2 Unsupervised learning

In contrast to the previous system, unsupervised learning ventures into uncharted territory, working with data that is not previously labeled or classified. In this mode, the system is challenged to discover for itself the structure and patterns inherent in the data. Without the right answers or specific examples to guide it, the model must identify correlations, groupings, and features on its own, like an explorer mapping uncharted territory.

Principles of unsupervised learning

1. **Pattern discovery:** The objective is to identify patterns, correlations, and groups in the data.
2. **Self-organization:** Models self-organize based on similarity or dissimilarity of data, thus creating an internal structure.

Applications

Unsupervised learning has fascinating and varied applications, such as:

1. **Customer segmentation in marketing:** grouping customers based on similar characteristics without prior labeling, to better understand preferences and behaviors (Olarte et al., 2018).
2. **Anomaly detection:** Identifying atypical behavior or elements in financial transactions, which is crucial for fraud detection, predictive maintenance, and cyber security (Ameijeiras Sánchez et al., 2021; Álvarez, 2020).
3. **Genomic analysis:** Grouping genes with similar characteristics to better understand biological relationships.

Practical examples:

1. **Grouping or clustering:** For example, in a consumer dataset, a model can identify groups based on purchasing patterns or preferences without having preset labels (Hoz-Dominguez et al., 2019).
2. **Reduction of dimensions:** In the analysis of high-dimensional data, such as genetic data or images, unsupervised learning can help identify the most relevant features (Pérez Verona and Arco García, 2016).
3. **Data generation**: Creation of new data that is statistically like the training data used, allowing the generation of images, text, and other types of content.

Challenges and Considerations:

- **Interpretation of results:** The results of unsupervised learning can be more difficult to interpret, as there are no "correct" answers to validate against.

- **Dependence on data quality:** Data quality and diversity are crucial to avoid erroneous or biased conclusions.

Unsupervised learning opens a world of possibilities for exploring and understanding large datasets without the need for predefined labels. For economists, this branch of ML offers an invaluable tool for uncovering hidden patterns in socioeconomic phenomena, providing insights that might go unnoticed under more traditional approaches. It is an exciting field that promotes a more exploratory and discovering form of analysis, crucial in the era of big data.

2.1.3. Reinforcement learning

This type is, in some ways, like supervised learning. The model learns to make decisions through trials, where it receives rewards or penalties for actions taken. It is like learning to play a game: you improve through practice and adjusting your actions based on wins and losses (Montenegro Meza et al., 2023).

Reinforcement learning is a dynamic facet of *machine learning* that mimics the way humans and other animals learn from the consequences of their actions. In this approach, the model, often referred to as an "agent," learns to make decisions by executing actions within a defined environment. Through a process of trial and error, the agent receives rewards or penalties based on the actions it performs. This method is analogous to learning to play a game, where one improves through practice and adjusting strategies based on the results obtained.

Principles of Reinforcement Learning:

1. **Reward-based feedback:** Learning is guided by a reward system, where beneficial actions result in positive outcomes and detrimental actions result in negative consequences.
2. **Exploration vs. exploitation:** The agent must balance between exploring new strategies and exploiting the ones he already knows work well.

Applications

Reinforcement learning is especially useful in situations that require a sequence of decisions, such as:

1. **Games and simulations:** From classic board games such as Go and chess to complex video games (Aguado Sarrió, 2015).
2. **Robotics:** teaching robots to perform tasks through practice and adaptation to changing environments (Quintía Vidal, 2013).
3. **Process optimization:** In economics, it can be used to model and improve decisions in business processes, such as inventory management or supply chain optimization.

Practical Examples:

1. **Google DeepMind's AlphaGo:** Perhaps the most famous example is AlphaGo, which taught itself to play and beat human champions in the game of Go, a game known for its complexity and strategic depth.

2. **Automated trading:** Systems that learn to perform financial operations based on maximizing rewards (profits) and minimizing risks (Giraldo Escobar, 2021).

Challenges and Considerations:

- **Reward structure design:** How rewards and penalties are designed can significantly influence how the agent learns, so it should be carefully considered.
- **Risk of overfitting:** There is a risk that the agent will overadjust to a specific environment and not perform well in slightly different situations.

Reinforcement learning represents an exciting frontier in the field of *machine learning*, offering a unique approach to the decision-making process. For economists, it provides a powerful tool for simulating and improving decisions in complex and dynamic environments, offering valuable insights for formulating more effective economic strategies and policies. This approach underscores the importance of experience and adaptability in learning, two essential qualities in today's ever-changing economy.

Understanding these definitions and distinctions is essential for any economics student wishing to delve into the field of AI and ML. They provide the tools necessary to navigate this exciting and sometimes overwhelming world of technology, allowing for a deeper understanding of how these innovations can be applied in economic analysis and decision-making.

Machine learning techniques

This table shows the main *machine learning* methodologies grouped according to their type (supervised, unsupervised, and reinforcement), together with examples of their possible uses in the field of economics:

Type of Apprenticeship	Methodologies	Possible Uses in Economics
Supervised	Linear and Logistic Regression	Market Trend Forecasting
	Support Vector Machines (SVM)	Classification of companies according to their credit risk
	Neural Networks	Product demand forecast
	Random Forest (Random Forest)	Financial asset price forecasting
	Decision trees	Demand modeling
Unsupervised	Clustering (K-means)	Market Segmentation, Investment Portfolio Analysis, Real Estate Pricing Modeling

Type of Apprenticeship	Methodologies	Possible Uses in Economics
	Principal Component Analysis (PCA)	Dimensionality reduction in economic data
	Generative Adversarial Networks (GAN)	Synthetic data generation for trend analysis
	Hierarchical Clustering Analysis	
	Self-Organizing Maps	
For reinforcement	Q-Learning, Deep Policy Gradients, Actor-Critic Based Models	Economic Policy Simulations,
	Genetic Algorithms	Design of automated trading strategies
	Deep Learning by Reinforcement (DRL)	Dynamic inventory management in the supply chain
	Policy Gradients	Supply and demand control in economic systems

We will now proceed to briefly explore the main machine learning models. These models play a crucial role in how machines learn, interpret, and process data. From simple linear regression-based models to deep and complex neural networks, each type has its unique characteristics and specific applications. This review will provide an overview of the most common models and their uses, thus providing a basic understanding of the various approaches and techniques that form the basis of machine learning today.

1. Linear regression

Linear regression is one of the simplest and most widely used statistical techniques in machine learning. It is used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data.

How it works: In its simplest form (simple linear regression), it models the relationship between two variables by fitting a straight line to the data points. The general equation is $y=ax+b$, where y is the dependent variable, x is the independent variable, and a and b are the model parameters that are fitted during training.

Applications: Economic forecasting, sales trends, and price impact analysis, among others.

2. Logistic regression

Despite its name, logistic regression is used for binary classification rather than regression. It is used to estimate the probability that an instance belongs to a particular category.

How it works: Logistic regression models the probability that a binary dependent variable belongs to a category (e.g., 0 or 1, true or false). It uses the logistic function to model this probability.

Applications: Medical diagnosis, machinery failure prediction, classification of emails into spam or non-spam, categorical classification of companies and households (Ojeda et al., 2005).

3. Decision trees

Decision trees are predictive models that represent a set of decisions and their possible consequences (Bouza and Santiago, 2012).

How it works: A decision tree consists of nodes, branches, and leaves. Each node represents a characteristic or attribute, each branch represents a decision or rule, and each leaf of the tree represents a result (a classification or prediction).

Decision Process:

1. Attribute Selection: When building a tree, the model selects the attribute that effectively divides the data set into smaller subsets. This selection is based on measures such as Information Gain, Gini Index, or Variance Reduction.
2. Tree Construction: The process starts at the root of the tree and splits according to the best attribute. This process is repeated recursively at each subdivision.
3. Tree Pruning: To avoid overfitting, tree pruning is performed, which removes branches using criteria that do not provide additional information.

Practical Example: Imagine a decision tree used to decide whether an email is spam or not. The root node could be "Does the mail contain the word 'free'?" If the answer is yes, one branch follows, if not, another. This process continues until it reaches a leaf that classifies the email as spam or non-spam.

Applications: evaluate the credit risk of individuals or persons. Analyze different financial and credit factors to predict the probability of default on loan payments. It can also be applied, among many other options, in market segmentation in specific categories: identifying groups of consumers with similar characteristics of special interest in the creation of marketing strategies and personalization of services.

4. Random forests

Random forests are an ensemble method that uses multiple decision trees to improve the robustness and accuracy of predictions.

Operation:

Construction of multiple trees:

1. Bootstrap aggregating (Bagging): Random forests create multiple decision trees using bagging. Each tree is constructed from a random sample of the training data set (with replacement), known as a bootstrap sample.
2. Random feature selection: For each split in each tree, a subset of features is randomly selected. This ensures that the trees are different and reduces the correlation between them.

Prediction Process:

- Voting for classification: Each tree in the forest makes a prediction. The class that gets the most votes from all the trees in the forest is the final prediction.
- Averaging for regression: In regression tasks, the final forecast is the average of the predictions of all the trees.

Advantages of a single tree:

- Lower overfitting: By using multiple trees and averaging their predictions, random forests reduce the risk of overfitting which is common in individual decision trees.
- Robustness: By building each tree from different samples and features, random forests are less sensitive to anomalies and data variability.

Practical example: In a home price prediction model, several decision trees could analyze different aspects (location, size, age of the property, etc.), and the random forest would combine these perspectives to make a more accurate and stable price prediction.

In summary, while decision trees are valuable for their simplicity and ease of interpretation, random forests considerably improve accuracy and robustness, making them suitable for a wide range of classification and regression applications **(**Camacho et al., 2021).

Applications: There are a multitude of them such as credit risk assessment by considering a variety of important financial, historical, and demographic factors, price prediction of stocks or other financial assets, or customer segmentation.

5. Support Vector Machines (SVM)

SVMs are a set of supervised learning methods used for classification and regression. They stand out for their ability to handle high-dimensional data and their effectiveness in high-dimensional spaces (Sánchez Anzola, 2015).

Operation: In the context of classification, the operation of SVMs is both elegant and effective. It focuses on identifying the optimal hyperplane that divides the classes within the feature space. This hyperplane is not chosen randomly or based on a simplistic criterion but is selected in a way that maximizes the margin between classes. This margin maximization is critical as it provides some "breathing space" between the categories, thus allowing for a clearer and more defined classification. Imagine you are trying to separate apples from oranges; SVMs not only draw a dividing line but look for the line that best distinguishes between an apple and an orange, leaving a wide enough space to avoid ambiguity.

As for regression, SVMs take a similar approach, but instead of looking for separation between categories, they focus on finding a function that fits as closely as possible to the distribution of the data, minimizing the error. This is achieved by using the same principle of margin maximization but applied to minimize the discrepancy between the observed data and the prediction function. This approach offers a powerful means of predicting continuous values, which is invaluable in fields such as economics, where price trends can be predicted, or in meteorology for forecasting weather patterns.

Applications: The applications of SVMs are as varied as they are fascinating. In bioinformatics, for example, they are used for pattern recognition in genetic sequences, which can help in the identification of genetic markers for diseases. In the field of natural language processing, SVMs are fundamental for text classification and sentiment analysis, making it possible to discern, for example, whether an opinion expressed on social networks is positive or negative. Furthermore, in the field of computer vision, SVMs contribute significantly to face recognition, distinguishing and classifying facial features in a variety of applications, from security to entertainment.

However, the choice of using SVMs, like any other methodology in machine learning, is inherently dependent on the specific problem at hand, the nature of the data available, and the goals of the analysis. While SVMs are excellent at handling high-dimensional data and complex feature spaces, they may not be the best choice in situations where model interpretation and transparency are a priority, since SVM models can be difficult to interpret. Furthermore, in extremely large datasets, their performance may suffer, requiring a careful balance between accuracy and computational efficiency.

In summary, support vector machines remain a mainstay in the field of supervised learning, offering robust and effective solutions for both classification and regression, successfully addressing the challenges presented by high-dimensional and complex data.

6. Clustering (K-means)

The concept of *clustering* refers to the process of dividing a set of objects into groups so that objects in the same group (or *cluster*) are more similar (in some sense) to each other than to those in other groups. It is a fundamental technique in data analysis and data mining, where one seeks to group objects based on their characteristics without having predefined labels.

K-means is probably the most popular clustering method, both for its efficiency and its ease of use and interpretation. It is used to divide a data set into K distinct clusters, where 'K' is a user-specified number. This method is particularly useful in situations where you have an idea of the number of natural clusters in the data set.

Operation of K-means:

1. **Initialization**: K points are chosen at random from the data set as the initial centers of the clusters. These points are known as centroids.
2. **Cluster assignment**: Each point in the data set is assigned to the cluster whose centroid is closest. This is typically done using Euclidean distance as the proximity metric.

3. **Updating centroids**: After assigning all the points to the clusters, the centroids of these clusters are recalculated. The new centroid of each cluster is the average (or centroid) of all points assigned to that cluster.
4. **Iteration**: The assignment and update steps are repeated iteratively. In each iteration, the centroids are adjusted based on the points assigned to their respective clusters, and the points are reassigned to the clusters with the closest new centroids.
5. **Convergence**: This process is repeated until the cluster centroids stop changing significantly, indicating that the clusters have stabilized, and convergence has been reached.

K-means applications

- **Market Segmentation**: Identify groups of customers with similar characteristics for more targeted marketing campaigns.
- **Biomedical Data Analysis**: Classify cell types or diagnoses based on patterns in the data.
- **Organization of Large Document Databases**: Group similar documents together to improve the search and organization of information.

Importance and Limitations

Despite its usefulness, especially in large data sets, K-means has some limitations that should be kept in mind, such as the need to specify the number of clusters in advance and the sensitivity to the initial starting points of the centroids. In addition, it does not always work well with clusters of non-spherical shapes or very different cluster sizes.

7. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a statistical dimensionality reduction technique that is widely used in the field of data analysis and machine learning. Its main objective is to simplify the complexity of multidimensional data spaces while keeping as much information as possible.

Operation of the PCA

Component extraction:

1. **Identification of principal components**: The first step in PCA is to identify the principal components in the data set. These components are directions in the data space where there is the most variability. In more technical terms, they are the axes that maximize the variance of the data projected onto them.
2. **Calculation of eigenvalues and eigenvectors**: Mathematically, this is done by calculating the eigenvalues and eigenvectors from the covariance matrix of the data or, in some cases, from the correlation matrix or even from matrix decomposition techniques such as SVD (Singular Value Decomposition).

Data transformation:

1. **Projection into new components**: Once the principal components have been identified, the original data are transformed (or projected) into these new components. This translates into changing to a new coordinate system where the axes are now the principal components.
2. **Reduction of dimensions**: By projecting the data onto the principal components, the dimension of the original data is effectively reduced. This is because principal components capture most of the information (variance) present in the data set.

Component selection:

1. **Choice based on variance**: The first principal components capture the most variance. The selection of how many principal components to keep is based on the percentage of total variance you want to retain in the transformed data.
2. **Discard less significant components**: Components that capture less variance (and therefore less information) can be discarded. This simplifies the data set without significantly losing important features.

PCA applications

- **Multidimensional data visualization**: PCA is used to reduce high-dimensional data to 2 or 3 dimensions for visualization.
- **Data compression**: In engineering and data science, PCA is used to reduce the size of data while keeping as much information as possible.
- **Feature extraction in machine learning**: PCA is useful for extracting relevant features that can improve the performance of machine learning models.

Importance and limitations

PCA is powerful for its simplicity and effectiveness in highlighting the fundamental characteristics of the data. However, it has limitations, such as the assumption of linearity and the possible loss of important information in discarded components. In addition, PCA is not always suitable for data with complex nonlinear structures.

In summary, PCA is an essential tool in data analysis, useful for simplifying complex data sets and extracting valuable information. Its correct application can reveal underlying patterns and structures that would otherwise be difficult to identify.

8. Hierarchical clustering analysis

Hierarchical clustering is a method of cluster analysis that seeks to organize a data set into a hierarchical structure, either by merging smaller clusters into larger clusters (agglomerative approach) or by dividing a large cluster into smaller clusters (divisive approach). This method is especially useful when seeking to understand the relationship between elements in a dataset, beyond simple clustering.

How hierarchical clustering works

Agglomerative approach:

1. **Startup**: You start by treating each data point as an individual cluster, so if there are N data points, there are N clusters at startup.
2. **Cluster combination**: In each step, the two clusters that are closest to each other (according to some measure of similarity or distance) are searched for and combined to form a new cluster. This process reduces the total number of clusters by one at each step.
3. **Repetition**: The process is repeated, at each iteration combining the two closest clusters, until all data points have been grouped into a single cluster.

Divisive approach:

1. **Start**: You start with a single cluster containing all the data points.
2. **Cluster splitting**: At each step, the 'larger' or more heterogeneous cluster is split into smaller clusters. This can be done based on some measure of internal cluster dissimilarity.
3. **Repetition**: The process is repeated, splitting at each iteration the most appropriate cluster, until each data point is in its own individual cluster.

Dendrogram:

- **Visual representation**: The result of this process is often represented as a tree called a dendrogram. This tree shows the sequence of mergers (in the agglomerative case) or splits (in the divisive case) and the distance or similarity at which they took place.
- **Interpretation**: The dendrogram allows one to visualize not only how the data points are grouped into clusters, but also how these clusters relate to each other at different levels of the hierarchy.

Applications of hierarchical clustering

- **Genetic analysis**: Used to understand the evolutionary relationships between different species or genetic variants.
- **Document classification**: Group similar documents together to organize large collections of data.
- **Image segmentation**: In image processing, to identify and separate different regions or objects in an image.

Importance and Limitations

Hierarchical clustering is valuable for its ability to provide a detailed, multilevel view of data structure. However, it is computationally more intensive than methods such as K-means, especially for large data sets. In addition, once a decision is made to combine or split clusters at a certain step, that decision cannot be undone in subsequent steps, which can affect the flexibility of the method.

In summary, hierarchical clustering is a powerful tool for data analysis, offering a unique perspective on the structure and relationships between data. Its application can reveal

complex and detailed insights, especially in areas where hierarchical relationships are of particular interest.

9. Self-Organizing Maps (SOM)

Self-Organizing Maps (SOM) is an unsupervised learning technique that uses neural networks to provide a visual, low-dimensional representation of high-dimensional data while maintaining their original topological relationships. This technique was developed by Teuvo Kohonen and is therefore sometimes referred to as "Kohonen maps".

SOM operation

Grid neural network

1. **Network structure**: SOM uses a network of artificial neurons, which are typically organized in a two-dimensional grid. Each neuron in this grid is associated with a vector of weights of the same dimension as the input data.
2. **Initialization**: The initial weights of neurons can be random or based on some specific criterion.

Competitive Learning

1. **Neuron specialization**: During the training process, each neuron "specializes" in a specific pattern of the input data. This is achieved through a process of competition between neurons.
2. **Winning neuron selection**: For each input data, the neuron whose weights are most like the data (the winning neuron) is identified, generally using the Euclidean distance.
3. **Weight update**: The winning neuron adjusts its weights to get closer to the input data. In addition, neighboring neurons in the grid also adjust their weights, but to a lesser extent.

Topological organization

1. **Neighbor matching**: Nearby neurons in the grid tend to adjust to represent similar data. This ensures that the resulting map reflects the distribution and topological relationships of the original data in the high-dimensional space.
2. **Map formation**: At the end of the training process, the grid of neurons forms a "map" that represents the different patterns and relationships present in the data set.

SOM applications

- **Visualization of complex data**: SOM helps visualize and understand high-dimensional data sets in a more accessible format.
- **Pattern recognition**: This can be used to recognize complex patterns in data, which is useful in various applications such as fraud identification.

- **Multidimensional data analysis**: SOM is used to explore and analyze the structure inherent in multidimensional data, such as in market research or bioinformatics.

Importance and limitations

SOM is valuable for its ability to provide an intuitive representation of complex data and for maintaining the topological relationships of the data, which is crucial for understanding the inherent structure of the data. However, the quality of the generated map depends on several factors such as the choice of learning rate, the number of training iterations, and the grid structure, which can be challenging in terms of parameter tuning. In addition, interpreting the results of a SOM can be subjective and requires experience.

In summary, SOM is a powerful tool for the analysis and visualization of high-dimensional data, providing valuable information in a visually accessible and understandable format.

10. Q-learning

Q-learning is a fundamental technique in the field of reinforcement learning (Clifton and Laber, 2020), a branch of artificial intelligence that focuses on how agents must make decisions to maximize some notion of "reward" or "gain" over time. In Q-learning, the goal is to learn an optimal policy: a strategy that tells the agent what action to take under any given circumstance to maximize its total reward.

Q-learning operation

Table Q

1. **State-action representation**: The agent maintains a table, known as a Q-table, which is essentially a state-action representation that estimates the value of taking a specific action in a specific state.
2. **Initialization**: The Q table can be initialized with arbitrary values since the agent will learn the optimal values during the training process.

Updating table Q

1. **Agent experience**: After the agent takes an action and observes the reward and the new state, it updates the Q table.
2. **Update formula**: The Q table is updated using the update formula, which is a form of the Bellman equation. This formula incorporates the reward received and the estimate of the maximum value in the new state, thus adjusting the Q estimate for the action taken.
3. **Learning over time**: Over time and many iterations, the Q-table converges to the optimal values that represent the best action to take in each state.

Action policy

1. **Decision making**: The agent uses the Q-table to decide which action to take. A common policy is ε-greedy, where most of the time, the agent chooses the action with the highest Q value (exploitation), but with a small probability ε, chooses a random action (exploration).

2. **Exploration vs. exploitation**: This strategy helps balance exploration (trying new actions to discover their values) and exploitation (using acquired knowledge to maximize reward).

Q-learning applications

- **Simple games**: Like chess or video games, where the agent learns the best strategy to win.
- **Navigation problems**: Where the agent must learn the optimal path to reach a destination.
- **Environments with manageable state and action space**: Q-learning is particularly effective in environments where state and action spaces are small enough to be represented in a table.

Importance and Limitations

Q-learning is important in the field of reinforcement learning because of its simplicity and effectiveness in environments with a discrete and manageable action-state space. However, in environments with very large or continuous state-action spaces, the Q-table becomes impractical, leading to the need for more advanced techniques such as Deep Q-learning, which uses neural networks to approximate the Q-table.

In summary, Q-learning is a powerful technique for teaching agents how to make optimal decisions in sequential decision environments, providing a solid foundation for many advanced developments in the field of reinforcement learning.

11. Deep Reinforcement Learning (DRL)

Deep Reinforcement Learning is a burgeoning area of research that combines reinforcement learning techniques with the computational power of deep neural networks. This combination has led to significant advances in the ability of reinforcement learning agents to operate in highly complex environments with high-dimensional state and action spaces that were inaccessible to traditional reinforcement learning methods.

DRL operation

Use of deep neural networks

1. **Function approximation**: In DRL, deep neural networks are used to approximate the value function (which indicates how much value it is to be in a particular state) or the policy (which dictates the action to take in each state).
2. **Complexity handling**: Neural networks are particularly useful for handling complex or continuous state/action spaces, where traditional techniques such as Q-tables are not feasible due to their size or continuous nature.

DRL algorithms

1. **Deep Q-Networks (DQN)**: DQN is one of the earliest and best-known DRL architectures. It extends the Q-learning method, using a deep neural network to

approximate the Q-table. This approach is effective in environments with a high number of states, such as those found in video games.

2. **A3C and Variants**: Actor-Critic Asynchronous Advantage (A3C) is another popular algorithm that uses two neural networks: one for policy (actor) and one for value (critical). It offers improvements in efficiency and stability over previous approaches.

DRL applications

- **Complex games**: DRL has been used to achieve outstanding performances in complex games such as Go and chess, where strategies and decisions are extremely intricate, and the state space is huge.
- **Robotics simulations**: In robotics, DRL is used to teach robots to perform complex tasks that require precise and adaptive control.
- **Autonomous vehicle control**: DRL algorithms are critical in the development of autonomous vehicle systems, where real-time decisions must be made in dynamic and often unpredictable environments.

Importance and challenges of DRL

DRL represents a significant step forward in the field of artificial intelligence, as it allows agents to learn and adapt in environments that are much closer to the complexity of the real world. However, it also presents challenges, such as the need for large amounts of data for training, the difficulty in interpreting policies learned by neural networks, and the challenge of ensuring stability and convergence during training.

In summary, deep reinforcement learning is an innovative fusion of deep learning and reinforcement learning techniques, opening new paths for the creation of intelligent systems capable of learning and acting in complex, high-dimensional environments.

12. Policy gradients

Policy gradient methods in reinforcement learning represent an advanced and distinctive approach, differing fundamentally from value-based methods. Their main attraction lies in their ability to directly learn the optimal policy, an essential aspect in contexts where actions and decisions must be continuous and precise, as in robotics.

Direct policy learning: In contrast to value-based methods, which require first estimating the value function and then deriving policy from it, policy gradients focus directly on policy. Here, 'policy' refers to the strategy that the agent employs to decide his actions in each state of the environment. This strategy is modeled as a probability distribution, indicating the probability of selecting each possible action in each state. This approach eliminates the need for a Q-table or a value function, which is particularly advantageous in environments with many states or actions, where estimation of the value function may be computationally demanding or even infeasible.

Use of policy gradients: To fine-tune the policy, Policy Gradient methods apply gradient-based optimization techniques. This involves iteratively adjusting the policy parameters

(which could be, for example, the weights of a neural network) in the direction that maximizes the expected reward. This maximization process is performed by calculating the gradient of the expected reward concerning the policy parameters and adjusting those parameters in the direction that increases the reward. This optimization method is analogous to climbing a hill in search of the highest point, where each step is taken in the direction in which the slope is steepest.

Applications in continuous action spaces: Policy gradient methods are particularly suitable for problems with continuous action spaces. In such environments, actions are not discrete (as in a board game) but can take any value within a range. For example, in robotics, a robotic arm might need to move at a precise angle, requiring continuous control rather than discrete movements. In addition, these methods are useful in tasks that require a delicate balance between actions, where small variations in actions can have large effects on the outcome.

In **summary**, policy gradient methods offer an elegant and effective solution for learning optimal policies in complex, continuously acting environments. Their ability to work directly with policy and use gradient-based optimization techniques makes them particularly useful in advanced applications such as robotics, where accuracy and the ability to adapt to a wide range of situations are crucial.

13. Actor-critic based models

Actor-critic models in reinforcement learning represent a sophisticated synergy between two fundamental approaches: value-based and policy gradient methods. This combination results in a dual-component system: an "actor", who makes decisions, and a "critic", who evaluates these decisions, each playing a crucial role in the agent's learning and decision making.

The actor: learning the optimal policy. The "actor" component in an actor-critical model is responsible for learning the optimal policy. In this context, policy refers to the strategy that the agent follows to choose actions in various states of the environment. The actor, usually modeled as a neural network, generates a probability distribution over the possible actions, indicating which action is most likely to be the best in each state. This process resembles an agent making decisions based on its current experience and knowledge.

The critic: evaluating the actions. The "critic", on the other hand, has the task of evaluating the quality of the actions proposed by the actor. For this, it uses a value function, like value-based reinforcement learning methods. This component evaluates how good an action is in terms of the expected long-term reward, providing a way to measure the success of the decisions made by the actor.

Learning and updating process. At the heart of the actor-critic model is the iterative process of learning and adjustment. The actor is updated using policy gradients, which are calculated based on feedback provided by the critic. This feedback is essential, as it informs the actor about the quality of their actions, allowing it to adjust its policy to improve its performance. In parallel, the critic updates himself using value-based reinforcement learning methods, continuously refining his evaluation of the actions.

Applications in complex tasks Actor-critical models are particularly suitable for complex control and decision tasks. In strategy games, for example, they can effectively handle a wide range of situations and make informed decisions. In autonomous driving simulations, they can accurately balance various objectives, such as safety and efficiency. In addition, they are useful in problems where a careful balance between exploration (trying new actions) and exploitation (using acquired knowledge) is required.

In **summary**, Actor-Critic models represent an advanced methodology in reinforcement learning by combining the directivity of policy gradients with the reflective evaluation of value-based methods. Their dual structure allows for a more nuanced and efficient approach to learning and decision-making in complex environments, making them ideal for applications requiring a high level of adaptability and accuracy.

3. Machine learning and Econometrics

As access to large amounts of data has become more common, machine learning has gained popularity as a tool for analyzing patterns and making predictions. In the context of economics, machine learning can be used to improve the predictive ability of econometric models, especially in situations where the relationships between variables are complex and nonlinear.

Some ML methods, such as neural networks and deep learning algorithms, can be applied to economic problems to identify nonlinear patterns in the data and improve forecasting ability. In addition, ML can be useful in variable selection, managing massive data, and improving the efficiency of parameter estimation.

Econometrics and machine learning can complement each other, as the former provides a sound theoretical and economic basis, while the latter offers advanced tools for data analysis and prediction in complex environments and large data sets.

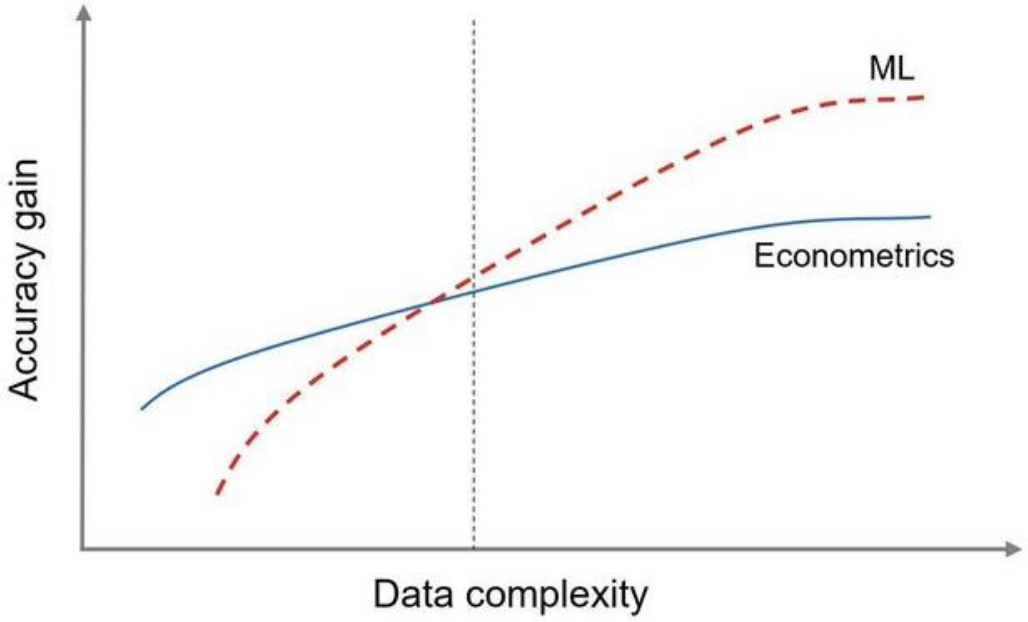
This comparative table defines a broad typology of problems in economics and suggests both econometric and ML tools that might be useful for each case. It also briefly discusses the advantages and disadvantages of each approach:

Problems in Economics	**Econometric Tools**	***Machine Learning* Tools**	**Advantages/Disadvantages**
Economic Trend Forecasting	ARIMA Models, Linear Regression	Neural Networks, Decision Trees	Econometrics: interpretable, ML: higher precision in complex patterns
Market Segmentation	Traditional Cluster Analysis	Clustering Algorithms (K-means, DBSCAN)	Econometrics: based on assumptions, ML: detects nonlinear patterns
Credit Risk Analysis	Logit/Probit Models, Discriminant Analysis	Random Forests, SVM	Econometrics: explicit modeling, ML: best on large datasets

Problems in Economics	Econometric Tools	*Machine Learning* Tools	Advantages/Disadvantages
Price Optimization	Hedonic Pricing Models	Neural Networks, Optimization Algorithms	Econometrics: theory-based, ML: adapts prices in real time
Financial Fraud Detection	Time Series Analysis, Regression	Neural Networks, Reinforcement Learning	Econometrics: good for trends, ML: learns from hidden patterns

This table provides a clear and comparative view of how econometric (Ceballos Mina & Duque García, 2022) and machine learning tools can be applied to different economic problems, highlighting both the strengths and limitations of each approach in the economic context.

Relative advantages of machine learning vs. traditional econometrics in terms of data complexity



Source: Desai (2023) Machine learning for economics research: when what and how

3.1 Causality

Determining causality is a fundamental aspect of economics, crucial for understanding how and why certain economic phenomena occur and for informing the formulation of effective economic and regulatory policies.

Causality implies that a change in one variable (cause) produces a change in another (effect). In economics, establishing causal relationships is essential for understanding

market dynamics, the impact of economic policies, and other key factors that influence the economy.

The main challenge in determining causality in economics is the presence of confounding factors, simultaneity, and endogeneity of variables. Often, the available data are observational rather than experimental, which makes it more difficult to establish definitive causal relationships.

Econometrics uses statistical models that attempt to isolate and quantify causal relationships between variables. Techniques such as instrumental variables regression models, difference-in-differences, and fixed effects models are commonly used to address endogeneity problems and establish causality.

Advantages include its strong emphasis on interpretation and economic theory; it offers proven methods for identifying causal relationships. Drawbacks include its limited ability to deal with data that do not meet certain statistical assumptions; models may be too simplified to capture the complexity of the real world (Quiguiri Daquilema, 2023).

Machine learning:

- Focus on prediction and pattern recognition, traditionally less focused on causality.
- Emerging techniques in ML, such as causal learning and treatment effects models, are beginning to address causality.
- Advantages: capable of handling large volumes of data and complexity; good for pattern detection and prediction.
- Disadvantages: interpretation of models can be difficult; identification of causal relationships is still a developing area.

Comparison chart: Econometrics vs. machine learning in causality

Appearance	Econometrics	Machine learning	Comments
Main Focus	Identification of causal relationships	Prediction and recognition pattern	ML focuses more on prediction and econometrics on causality.
Common Techniques	Regression models, Instrumental variables, Difference-in-differences, Differences in differences	Causal learning, Treatment effects models	Econometrics uses proven techniques; ML explores new methods.
Data Management	Requires specific assumptions	Handles large data sets and complexity	ML is more flexible with data, but econometrics provides greater theoretical clarity.

Appearance	Econometrics	Machine learning	Comments
Economic Policy Applications	Policy impact analysis, Ex ante evaluations, Ex ante evaluations	Data-driven predictions, Simulation models	Econometrics for policy testing, ML to explore future scenarios.
Advantages	Theoretically rigorous, Interpretative	Flexible, Handles complexity and volume of data	Econometrics for understanding causes; ML for patterns and trends.
Inconveniences	Can be limited in unstructured data	Causality still under development	Data-constrained econometrics; ML needs further development in causality.

This table highlights how both econometrics and machine learning offer valuable tools for addressing the question of causality in economics, each with its unique strengths and limitations. While econometrics offers a more theoretical and causality-focused approach, machine learning brings flexibility and the ability to handle large data sets and complexity.

3.2 Differences between econometrics and machine learning

Econometrics and machine learning are disciplines that share the goal of modeling and understanding patterns in data but differ in their fundamental approaches, methods, and objectives. Econometrics, rooted in economics and statistics, focuses on estimating causal relationships between economic variables using statistical models and usually relies on economic theories to specify models and hypothesis testing to validate their results. Machine learning, on the other hand, focuses on developing algorithms that can learn complex patterns directly from data without explicit model specification. The machine learning approach is more predictive and is used in a variety of fields, from speech recognition to online recommendations. While econometrics seeks to understand underlying causal relationships, machine learning focuses on predictive ability and generalization from unobserved data. In short, although they share some methods and techniques, econometrics focuses on causality and theoretical validation, while machine learning emphasizes prediction and the ability to adapt to complex patterns in the data.

Thus, although the tools and techniques used in econometrics and ML may differ, the most significant differences between the two disciplines are in their objectives, methods of analysis, and how they interpret and use the results obtained from the data.

Purpose and application: The main difference between econometrics and ML lies in their objectives and applications. Econometrics seeks to understand relationships and test theories, while ML focuses on making effective predictions from data.

Interpretation vs. accuracy: Econometrics emphasizes the interpretation of model coefficients and statistical significance, while ML prioritizes the accuracy and generalizability of models.

Data handling and models: Econometrics traditionally handles structured data and relies on models that require specific assumptions about the data. In contrast, ML can handle a wider variety of data types, including unstructured data, and often uses models that can learn characteristics directly from the data without predefined assumptions.

	Approach	Optimization method
Econometrics	Econometrics focuses on causal inference and understanding the relationships between variables. It is concerned with estimating causal effectors and testing economic theories.	It uses statistical and mathematical methods to estimate economic relationships. These methods include, but are not limited to, linear regression, time series models, and other models that seek to establish relationships and test hypotheses.
Machine learning	ML focuses on prediction and pattern recognition from data. It is more oriented to maximizing the accuracy of predictions and generally pays less attention to interpreting the underlying relationships between variables.	Although it often uses neural networks (especially in deep learning) and occasionally genetic algorithms, these are not its only methods. ML also includes a variety of techniques such as support vector machines, decision trees, and ensemble methods, among others.

4. Other Artificial Intelligence techniques

Although ML has become synonymous with AI in popular culture, other AI techniques are not based on learning from data. These techniques include symbolic logic, automated planning, non-ML-based natural language processing (NLP), and expert systems, among others. We will explore these techniques in more detail below, culminating in a discussion of Large-Scale Language Models (LLM) and their future developments.

4.1. Main methodologies

From an Artificial Intelligence (AI) perspective, we address the following methodologies:

Symbolic logic

Symbolic logic (formal logic or propositional logic) in AI involves the use of symbols and formal structures to represent knowledge in a precise and structured way (Munarriz, 1994), reflecting propositions. This includes the use of propositions, logical operators, and quantifiers to formulate and manipulate knowledge representations.

Symbolic logic allows machines to make inferences, that is, to derive logical conclusions from a set of premises. This is fundamental in AI for information processing and analysis, allowing systems to make decisions based on logical rules and available data.

Symbolic logic is used in systems that perform automatic reasoning, such as problem-solving and expert systems. These systems can analyze information, apply logical rules, and arrive at new conclusions autonomously.

Systems based on symbolic logic can deduce new and valid conclusions from existing knowledge. This is essential in tasks such as automatic planning, problem diagnosis, and intelligent decision-making.

Symbolic logic serves as a foundation for more advanced fields of AI, such as natural language processing and machine learning, providing a framework for knowledge representation and manipulation.

Symbolic logic has been commonly used in game theory, fuzzy logic, or formal models, in the latter case helping to represent strategies and outcomes with formal symbols.

Symbolic logic applied in AI has important implications in the field of economics. It allows for modeling and analyzing complex economic systems by formally representing economic knowledge and deducing new conclusions (Boden, 2017). This is particularly useful in market simulation, economic policy analysis, and predicting economic trends, where logical and structured reasoning facilitates the understanding and analysis of complex economic interactions.

Automated planning

From an Artificial Intelligence (AI) perspective, automated planning can be explained as follows:

Automated planning in IA refers to the process by which computer systems automatically generate sequences of actions or steps to achieve a given objective. This process involves the identification of objectives, the evaluation of current conditions, and the formulation of a series of actions that, when executed in a specific order, will achieve the proposed objective.

The key to automated planning is the system's ability to define and pursue specific objectives, adapting its actions to environmental conditions and changes that may arise during the execution of the plan.

In the field of robotics, automated planning is crucial for programming robots to perform specific tasks, such as manipulating objects, navigating in unfamiliar environments, or performing operations on assembly lines. Robots use automated planning to evaluate their environment, decide what actions to take, and adapt their movements to the goals and constraints of the environment.

In video game development, automated planning is used to create intelligent behaviors in non-player characters (NPCs). This includes planning movements, tactical decisions, and game strategies that respond dynamically to the player's actions and the game environment.

Automated planning has significant applications in economics, especially in process optimization and strategic decision-making. In economics, this approach can be used to plan and execute investment strategies, optimize supply chains, and in economic policy formulation. By automating the planning of sequences of actions, economists and analysts can create more efficient and effective models to achieve economic objectives, considering a wide range of possible variables and scenarios.

Traditional Natural Language Processing (NLP)

Traditional PLN refers to the set of techniques and methods used for machine processing and interpretation of human language without relying primarily on learning from large amounts of textual data. These techniques are typically based on grammatical rules, syntactic and semantic analysis, and other structured methods for understanding and manipulating language.

Unlike machine learning (ML)-based approaches, traditional PLN focuses more on understanding and applying the linguistic rules and structures inherent in language, which enables machines to process and respond to text inputs in a logical and structured manner.

Before the ML era, machine translation was based on traditional PLN, using dictionaries and grammar rules to convert text from one language to another. Although less fluent and natural than ML-based methods, this approach allowed for basic and functional translations.

Sentiment analysis using traditional PLN involved the identification and classification of opinions and emotions in text based on a predefined set of linguistic rules and markers.

In information retrieval, traditional PLN was used to interpret text queries and extract relevant information from databases or document collections, using techniques such as parsing and pattern matching.

In the economic domain, traditional PLN can be useful for analyzing economic and financial documents, such as reports, articles, and news, to extract relevant information. It can plausibly be used to extract key information from financial statements, such as revenues, expenses, assets, or liabilities with entity and relationship extraction algorithms to automatically identify these elements and structure their information for analysis. It can also be used to identify trends and patterns in the narrative of financial reports as a prelude to identifying potential financial problems, changes in management, or modifications in strategic approaches.

Although ML-based methods have gained predominance nowadays, traditional PLN is still relevant to understanding the evolution of this technology and its applications in contexts where massive data-based models are not applicable or available. In economics, this could include the analysis of historical texts or documents with specific linguistic structures that require a more regimented approach.

Expert systems

Expert systems are computer systems designed to emulate the reasoning and decision-making of a human expert in a specific field. These systems are based on a set of rules and specialized knowledge that reflect the experience and accumulated knowledge of experts in a particular area (Badaró et al., 2013).

A distinctive aspect of expert systems is their focus on specific areas of knowledge. They are programmed to handle complex problems and situations within a particular domain, using rules and heuristics that a human expert would employ in similar situations.

In the medical field, expert systems are used to assist in medical diagnoses by analyzing symptoms, medical records, and other relevant information to suggest possible diagnoses, like the way an experienced physician would.

In finance, these systems can provide investment advice, risk analysis, and financial management recommendations based on a broad base of economic and financial knowledge.

Expert systems are also applied in a variety of other areas where specific expertise is valuable, such as engineering, meteorology, and urban planning, among others.

In economics, expert systems can be extremely valuable for complex analysis and data-driven decision-making. For example, they can be used to assess market conditions, predict economic trends, or advise on economic policies. By emulating expert decision-making in economics, these systems can contribute significantly to the formulation of economic and financial strategies, providing detailed analysis and recommendations based on in-depth knowledge of the field.

Neural networks and Deep learning

Neural networks are a class of machine learning models inspired by biological neural networks. They are characterized by their ability to learn complex patterns from large amounts of data.

Deep learning, a subcategory of neural networks, involves the use of multi-layered neural networks (known as deep neural networks) that enable more sophisticated and abstract learning and modeling of data. This approach has been instrumental in the evolution toward large-scale language models (LLMs) because of its ability to process and learn from vast amounts of information efficiently and effectively.

In the field of voice recognition, neural networks, and deep learning have enabled significant advances, improving accuracy and naturalness in the interpretation of spoken language, which has been crucial for the development of virtual assistants and voice control systems.

These technologies have revolutionized computer vision, enabling machines to interpret and analyze images and videos with an unprecedented level of detail and accuracy. This has had applications in areas such as facial recognition, automatic inspection in industry, and autonomous vehicle systems.

Recently, deep learning has transformed PLN, enabling the development of sophisticated language models that can understand, generate, and translate human language with a degree of complexity and naturalness never seen before. This has been key in the development of technologies such as machine translation, text generation, and advanced chatbots.

In the economic domain, neural networks and deep learning have significant potential for large-scale data analysis, market modeling, and prediction of economic trends. These technologies make it possible to analyze large economic and financial data sets, identify patterns and trends, and make accurate forecasts. In addition, they can be useful in automating tasks such as risk analysis, investment portfolio optimization, and in modeling consumer and market behavior (Romero Martínez et al., 2021). These applications represent an emerging frontier in economics, where data analysis and decision-making benefit greatly from advances in AI and machine learning.

Large-scale language models (LLM)

Large Language Models (LLM) such as GPT (Generative Pre-trained Transformer) are important milestones in the evolution of artificial intelligence. Their development, operation, and applications are detailed here:

LLMs have their roots in the progress of neural networks, especially deep neural networks. These networks, with their ability to process and learn from large volumes of data, lay the foundation for advanced language models.

Deep learning has enabled the creation of more complex and capable models that can understand and generate language with a high degree of sophistication. The depth and breadth of these networks are crucial for large-scale language processing.

A crucial innovation in its development was the discovery of the transformer in 2017. Its ability to handle data streams efficiently has been fundamental to the development of LLM.

Transformers are a model architecture in the field of natural language processing (NLP) and have played a crucial role in the development of Large-Scale Language Models (LLM).

Introduced in the paper "*Attention Is All You Need*" by Vaswani et al. in 2017, Transformers represent a paradigm shift in PLN. Unlike previous models based on recurrent neural networks (RNN) and convolutional neural networks (CNN), Transformers are primarily based on attentional mechanisms.

Attention on transformers allows the model to focus on different parts of the text input to better understand the context and relationships between words. This results in a deeper and more nuanced understanding of the language.

Unlike RNNs, which process data streams sequentially, Transformers allow parallelization of data processing. This means that they can handle larger and more complex data streams much more efficiently.

Their design makes them particularly suitable for training with huge data sets, an essential feature for LLM development.

Transformers are the foundation on which LLMs such as GPT (Generative Pre-trained Transformer) are built. These models use the transformer architecture to learn complex language patterns and generate coherent and contextual text.

Transformers allow pre-training on large text corpora, followed by fine-tuning for specific tasks, which is fundamental to the versatility and effectiveness of LLMs.

Thanks to their ability to handle complex contexts and long text dependencies, Transformers have significantly improved text comprehension and generation in LLMs.

Transformer-based models can generate text that is remarkably natural and fluid, coming closer to the level of human speech.

Although initially designed for PLN tasks, Transformers have found applications in other AI fields, such as computer vision and audio analysis.

In summary, Transformers have been instrumental in the advancement of LLMs, providing an efficient and scalable architecture that has enabled major advances in machine understanding and generation of natural language. They have ushered in a new era in PLN and continue to be a centerpiece in the development of advanced AI technologies.

Functioning of LLMs:

a. **Training in large amounts of text**:

LLMs, during their training phase with large amounts of data, learn linguistic patterns, grammatical structures, and semantic relationships from text data.

Through this training, LLMs develop the ability to generalize and adapt to a variety of topics and linguistic styles.

Upon receiving an input request, or prompt in the usual English terminology, LLMs generate responses that are coherent and relevant to the context provided, applying the knowledge acquired during training. In the scope of these models, a prompt is therefore the instruction or input provided to the model to generate a specific response or output. It is essentially how the user or the system interacts with the model to obtain the desired information or task. The importance of the prompt lies in its ability to direct and condition the output of the model. A well-formulated prompt can influence the generation of the model, causing it to produce more relevant and consistent responses. For example, if you want the model to translate a sentence into French, the prompt could be the sentence in the source language.

In the case of large language models such as GPT-3, the input can be a specific instruction or request that allows the LLM to know which type of task it is expected to perform. The precise formulation of the prompt is crucial for obtaining desired results and avoiding ambiguities. It is therefore a fundamental part of the interaction with large language models, and its importance lies in its ability to influence the output of the model, allowing users to obtain answers or perform specific tasks more effectively.

These models can handle complex requests, demonstrating a degree of creativity and adaptability in their responses.

Applications

LLMs can generate articles, stories, poems, and other types of written content, mimicking various styles and formats.

They can customize the content according to the specifications and style desired by the user.

They are used in the development of chatbots and advanced virtual assistants, providing natural and contextual responses.

They offer advanced machine translation capabilities, handling linguistic and contextual nuances.

Large Scale Language Models (LLM) have transcended beyond their traditional use in natural language processing to make inroads in emerging fields such as education and research. Their integration into these fields is opening new frontiers and transforming the way we interact with information and knowledge.

In the field of **education and learning**, LLMs are revolutionizing the concept of personalized tutoring and learning support (Márquez Benavides et al, 2023). Imagine a scenario in which a student, struggling with a complex mathematical concept or scientific theory, turns to an application powered by an LLM. This app not only provides detailed explanations tailored to the student's level of understanding but also answers specific questions, offering personalized examples and exercises. LLMs can analyze student responses, identify areas of confusion, and adapt their teaching to address these gaps in knowledge. In this way, students receive a tailored learning experience, like having a personal tutor who understands their individual needs and learning styles.

In the field of **research and data analysis**, LLMs are playing a crucial role in providing new insights and in-depth analysis of large data sets. For example, in a research project on climate trends, an LLM can quickly analyze vast amounts of historical and current climate data, identify significant patterns and trends, and suggest possible causes or correlations. This level of analysis, which would traditionally require months of work by a team of researchers, can be performed by an LLM in a much shorter time and with astonishing accuracy. In addition, LLMs can help researchers write reports, summarize findings, and even generate hypotheses relevant to future research. This ability to analyze and synthesize information at unprecedented scale and speed is transforming research in every field, from science and medicine to the humanities and social sciences.

In short, LLMs are paving the way for an era of personalization and efficiency in education and learning, as well as providing powerful tools for data analysis and research. Their ability to process, understand, and generate language is opening new possibilities and redefining the boundaries of what is possible in these fields.

LLMs, therefore, represent a significant advance in the ability of machines to understand and generate human language naturally and efficiently, opening a wide range of applications in various fields.

Future developments

Future developments in AI, particularly in the LLM domain, focus on improving context understanding, accuracy, and the ability to generate more coherent and contextually relevant responses. In addition, greater integration of real-world understanding and abstract reasoning ability is sought. Explanatory AI, which can reason and explain its decisions in a way that is understandable to humans, is another growing field of interest.

AI techniques that do not fall under the ML umbrella, such as symbolic logic and expert systems, have played a crucial role in the development of the field. However, the evolution towards more advanced systems such as LLM has been largely driven by advances in

neural networks and deep learning, pointing to an exciting and promising future for AI in a variety of applications.

5. EconAI (Economics 2.0)

5.1 Impact of AI on the economy and society

The impact of artificial intelligence (AI) on the economy and society is profound and multifaceted, affecting everything from labor markets to political and social decision-making. Each of these areas is discussed in detail below:

1. Economic effects of AI:

In the coming years, the continued development of artificial intelligence promises to have a transformative impact on the global economy. This emerging technology is not only redefining operations and strategies in diverse industries but is also driving a new era of innovation and efficiency. As AI becomes more deeply integrated into the manufacturing, services, and finance sectors, we expect to see significant changes in productivity, labor market dynamics, and industry structures. However, along with these opportunities come substantial challenges, including the need for labor adaptation and the management of economic inequalities. How companies, governments, and societies handle the integration of AI will largely determine the economic landscape of the coming decades. As appreciated in the slogan AI won't take your job, but a worker who knows how to use it will be more than enough to value its use as part of the daily workflow.

a. **Transformation of labor markets and industries**:

- **Automation and skills change**: AI is redefining labor markets by automating routine and predictable tasks, which in turn is changing the demand for skills in the workforce. As some jobs disappear, others emerge, especially those requiring analytical, creative, or interpersonal skills that AI cannot easily replicate (Corvalán, 2019).
- **Reconfiguring industrial sectors**: AI is transforming entire industries, from manufacturing with intelligent robots to financial services with advanced data analytics algorithms. This transformation not only improves the efficiency and quality of products and services but also drives innovation and the creation of new business models.

b. **Impact on productivity and economic growth**:

- **Increased productivity**: AI can significantly increase productivity by optimizing processes, reducing errors, and speeding up decision-making. This productivity improvement is a key driver of economic growth. Productivity gains in certain areas of the labor market will be an important step in redefining jobs.
- **Uneven economic development**: While AI can be a catalyst for economic growth, there is also a risk that it will widen economic inequalities, both between and within countries, due to variability in adoption and adaptation to these technologies.

2. Social implications of AI:

Parallel to its economic impact, the advance of Artificial Intelligence is shaping up to have profound consequences on the social fabric in the years to come. Beyond automation and efficiency, AI raises fundamental questions about ethics, privacy, and the nature of human labor (Cortina Orts, 2019). As this technology becomes a more integral part of our daily lives, from personalized recommender systems to virtual assistants and health applications, the need to address its influence on privacy, inherent biases in algorithms, and the proper governance of these systems emerges. How society adapts to these changes and resolves these ethical and moral dilemmas will be determinant in ensuring a future in which AI aligns with human values and well-being.

a. **Ethical and social challenges: privacy, bias, and AI governance**:

- **Privacy and data use issues**: AI, especially that which relies on the analysis of large amounts of personal data, poses serious challenges in terms of data privacy and security.
- **Biases in AI**: There is growing concern about the biases embedded in AI systems, which can perpetuate and amplify existing inequalities in society.
- **Need for governance and regulation**: These challenges require a governance and regulatory approach that balances innovation and ethical use of AI, involving multiple actors, including governments, businesses, and civil society.

b. **AI and its influence on political and social decision-making**:

- **AI in policy and public administration**: AI has the potential to improve decision-making in policy and public administration, from service optimization to data-driven policy formulation.
- **Risks and challenges**: However, the use of AI in the political sphere also presents risks, such as the possibility of manipulation or misuse of the technology for undemocratic purposes.

On the global stage, the European Union and Spain face a considerable challenge in the race for artificial intelligence (AI), especially in comparison to powers such as the United States and China. This race is not only a measure of technological progress but also a struggle for economic and political influence in an increasingly digitized world.

3. Europe and Spain: one step behind in the AI race

Europe, with Spain as an integral part, has adopted a more cautious approach to AI development, prioritizing regulation and ethics. While this is commendable from a human rights and privacy protection standpoint, it has led to a certain sluggishness in technological adoption and innovation. Meanwhile, the United States and China have advanced rapidly, driven by a combination of strong private-sector investment, a culture of rapid innovation, and, in the case of China, considerable government support.

This difference in approach has led to a gap that appears to be growing. In the United States, companies like Google, Amazon, and Microsoft are at the forefront of AI research and

development, while China is pushing the use of AI on a massive scale, integrating it into everything from urban surveillance to economic planning.

Specifically for Spain, but also to some extent for Europe in general, one of the most significant challenges is the impact of AI on the labor market. With unemployment already high, AI-driven automation threatens to displace even more jobs, especially in sectors with repetitive and predictable tasks. In addition, there is a notable skills gap, where the current workforce is not sufficiently prepared for the jobs of the future demanded by AI.

However, this self-critical view should not lead to pessimism, but to constructive action. Europe and Spain have significant strengths, such as high-quality education systems and a strong tradition of human rights and democracy. These strengths can be the basis for a unique model of AI development, one that balances technological innovation with social responsibility.

To close the gap, Europe and Spain need to increase investments in AI R&D, foster collaborations between the public and private sectors, and adapt education systems to better prepare the next generation of workers. In addition, they can leverage their focus on ethics and regulation as a positive hallmark, attracting those seeking more balanced and sustainable AI development.

In short, while Europe and Spain may currently be lagging in the AI race compared to the United States and China, there is a path forward that can capitalize on their unique strengths and core values (The Economist, 2017). The challenge will be to balance the urgency to innovate with a commitment to uphold ethical and social principles.

In summary, while AI offers significant opportunities for economic advancement and improved quality of life, it also presents complex and multifaceted challenges that require careful and balanced consideration. How society and governments address these challenges will largely determine the long-term impact of AI on our economic and social lives

5.2. **EconAI:** When and how?

The integration of artificial intelligence (AI) and machine learning (ML) in economics is marking a milestone in the way research and analysis in this field are approached. This transformation extends beyond simply increasing the amount of data available, encompassing unprecedented data diversity and complexity. ML, with its ability to process and analyze large volumes of data, is opening new avenues for understanding complex economic dynamics, from monetary policy to forecasting market trends.

One of the most significant contributions of ML in economics is its ability to process unconventional data such as text, images, audio, and video. These types of data, historically difficult to incorporate into conventional economic models, can now be efficiently analyzed thanks to advanced ML techniques. For example, text analysis using natural language processing can extract key information from financial documents or central bank communications, while satellite image analysis can be used to estimate regional or global economic indicators.

In addition, ML is making it easier to capture strong nonlinearities in economic data. Economic relationships, often influenced by multiple interconnected factors interacting in complex ways, are rarely linear or simple. This is where ML shines, offering the ability to capture and model these complexities, which not only improves the accuracy of predictions and analyses but also opens the door to groundbreaking discoveries in economics.

The ability of ML to process large traditional data sets is also improving the accuracy of economic forecasts. With the increasing availability of large-scale data, ML models can identify and use underlying relationships to make more accurate predictions about economic trends, inflation, and GDP growth, among others. This improves the ability of economists to understand the economy and provides more powerful tools for policy formulation and strategic decision-making.

In short, the integration of ML into economics is opening a new landscape of possibilities. It allows economists and analysts to tap into a wider range of data, capture previously inaccessible complexities, and make more accurate predictions. However, this integration also brings challenges, such as the need to handle large volumes of data and understand often complex models. Despite these challenges, ML is proving to be an invaluable tool in the modern economy, redefining how we understand and approach economic issues in an increasingly digitized world.

Preferred ML models in economic applications

Within the vast universe of economic applications, certain machine learning models are emerging as particularly powerful and effective. Among these, deep learning, and natural language processing (NLP) stand out for their ability to transform the way we analyze economic and financial text.

Deep learning models, when applied to PLN, have become indispensable tools for analyzing economic documents. Whether it is market reports, central bank communications, or economic news, these models have the unique ability to process and analyze large amounts of text, extracting valuable insights that were previously inaccessible. What makes PLN even more impressive is its ability to go beyond merely capturing the literal information contained in texts. These models can understand context, discern intent, and unravel underlying economic implications, providing a deeper and more nuanced understanding of economic issues.

On the other hand, reinforcement learning, and unsupervised learning are breaking new ground in economics. Reinforcement learning has proven to be particularly useful in the simulation of economic environments and strategic decision-making. This application of ML can be used to optimize trading strategies or to model the behavior of agents in markets, thus offering new perspectives and strategies in the financial world.

Unsupervised learning, on the other hand, specializes in identifying hidden patterns and correlations in large economic data sets. This type of ML can detect market trends or segment customers, all without the need for prior labeling of the data. Its ability to uncover non-obvious connections and patterns in data makes it invaluable for unraveling the complex web of factors that influence the economy.

Together, these ML models are redefining the field of economics, providing advanced tools for analyzing and understanding an increasingly complex and data-driven economic world. With their help, economists and analysts are equipped to discover new insights and apply them to economic and policy decisions.

Challenges and limitations of the use of ML in economics

The road to integrating machine learning into the economy is paved with both significant opportunities and challenges. As we explore this terrain, we encounter several limitations that require careful attention.

First, ML models, particularly those based on deep learning, have a voracious appetite for data. They require large volumes of data for training, which can be a considerable challenge. This need becomes an obstacle, particularly in areas where data is scarce or difficult to access. Economics, with its diversity of contexts and variables, often finds itself in this situation, raising questions about how and where to obtain sufficient and relevant data.

Furthermore, training these ML models is not only a matter of data but also of computational capacity. These models require many computational resources, ranging from powerful processors to advanced storage and memory capacity. This need can be a significant obstacle, especially for researchers or institutions operating on limited budgets. The barrier of cost and accessibility to appropriate technology can therefore restrict the use and exploration of ML in economics.

Another set of challenges revolves around problems of overfitting, interpretation, and bias. Overfitting occurs when an ML model fits too closely to the data it has been trained on, thus losing its ability to generalize to new situations or data sets. This is a subtle trap, as it can make a model look exceptionally good on paper but fail in practical applications.

Interpretation and transparency are also critical concerns. Many advanced ML models, such as deep neural networks, operate as "black boxes", meaning that their internal processes and the logic behind their decisions are not easily understood. This lack of transparency poses significant challenges, especially when it comes to making economic decisions based on these predictions.

Finally, the bias inherent in ML models is a major concern. These models can perpetuate and even amplify pre-existing biases in the training data. In social and economic applications, this can lead to erroneous or unfair conclusions, perpetuating inequalities, or biases.

In summary, although ML models offer powerful and promising tools for economic analysis and decision-making, their effective implementation requires a careful approach that is aware of their limitations and challenges. Understanding and mitigating these problems is essential to ensure the use of ML in economics that is not only effective but also responsible and ethical.

5.3 ML and AI Research Methodologies

Generating a complete big data analysis scheme, from data processing to the result in prediction or classification, involves following a series of detailed steps. Here is the step-by-step plan to perform this process:

Data collection

Data collection is the first and crucial step in any big data project, as it lays the foundation on which all analyses and predictions will be built. Let's start by exploring this process in detail.

It must be a meticulous process that requires careful planning and the use of advanced tools to ensure that the data collected is complete, accurate, and ready for the next stages of processing and analysis (Mamaqi et al, 2018).

The difference between "data sources" and "data ingestion" in a big data project is fundamental, as both represent distinct and critical stages in the data management workflow.

Data sources

Data sources refer to the various origins or places from which data are obtained. They are the starting point in the data flow.

1. **Source identification**: The process begins by identifying a variety of data sources relevant to the problem or research question. These sources can be internal to the organization, such as customer databases and transaction records, or external, such as demographic or social media data.
2. **Diversity of sources**: It is critical to consider a wide range of sources to get a holistic view of the problem. This includes relational and non-relational databases, files in various formats (such as CSV, JSON, XML), APIs that provide access to real-time data (e.g., weather data, social media feeds), and Internet of Things (IoT) devices, which provide real-time sensor data.
3. **Selection and evaluation**: The selection of appropriate data sources is crucial and is based on the relevance, quality, and reliability of the data they provide. This involves an evaluation process to ensure that the data are relevant to the desired analysis.
4. **Data access**: Once the sources have been identified, mechanisms are established to access this data. This may involve the configuration of database connections, authentication in APIs, or the configuration of networks to collect data from IoT devices.

Data ingestion

Data ingestion refers to the process of collecting and transporting data from its sources to a system where it can be stored, processed, and analyzed.

1. **Collection automation**: Data ingestion is not a one-time task, but a continuous process. Using tools such as Apache NiFi or Kafka, data collection can be

automated. These tools are capable of handling large volumes of data, ensuring that the data flow is constant and reliable.

- **Apache NiFi**: It is a platform that allows the automation of data flow between systems. Its visual drag-and-drop design facilitates the configuration of data flows, including the extraction, transformation, and loading (ETL) of data to and from various sources. NiFi is particularly useful for orchestrating complex data flows and ensuring that data is collected efficiently and error-free.
- **Kafka**: On the other hand, Apache Kafka is a distributed messaging system that excels in the management of real-time data flows. Kafka acts as a kind of backbone that enables data transmission between producers who generate data and consumers who use it. It is ideal for scenarios where data must be processed quickly and in large volumes, such as in financial transaction analysis or IoT sensor monitoring.
- **Cloud services**: In addition, cloud services such as AWS Kinesis, Google Pub/Sub, or Azure Event Hubs can be employed, which offer similar capabilities to Kafka but with the advantage of cloud infrastructure, reducing the need for user maintenance and scalability.

2. **Preliminary processing**: During ingestion, data may undergo preliminary processing such as filtering, basic transformation, and cleaning to prepare them for subsequent stages.
3. **Reliability and scalability**: Regardless of the tool chosen, it is crucial that the data ingest system be robust, reliable, and scalable. It must be able to handle unexpected spikes in data volume and recover from potential failures without losing critical information.

2. Data storage

Data warehousing in big data projects is a crucial stage where it is defined how and where the collected data will be stored for further analysis. This phase involves making strategic decisions about the infrastructure and organization of the data, including the selection of an appropriate infrastructure (Data Lake or Data Warehouse) and the decision on how the data will be structured and organized (read or write schema).

These decisions are critical to ensure that the data are stored efficiently, and securely, and are accessible for further analysis and processing.

Data Lake and Data Warehouse

1. **Concept and Purpose**:
    - A **Data Lake** is a centralized repository that allows storing large volumes of data in its native, unstructured form. It is like a large container that supports all types of data, from raw to processed and is especially useful for storing massive data from various sources in an unmodified format (LaPlante, 2016).

- A **Data Warehouse**, on the other hand, is a system designed to store already processed and structured data, optimized for analysis and reporting. Unlike Data Lakes, Data Warehouses require data to be transformed and loaded into a consistent and structured format.

2. **Technologies and Platforms**:
    - **Hadoop HDFS**: The Hadoop Distributed File System (HDFS) is a common choice for Data Lakes due to its ability to store huge amounts of data and its compatibility with Big Data processing tools such as Apache Spark.
    - **Cloud Services**: Platforms such as Amazon S3 (Simple Storage Service) offer scalable and flexible cloud storage solutions, ideal for Data Lakes. Google BigQuery and Amazon Redshift, for example, are popular choices for Data Warehouses in the cloud, providing large-scale data storage and analytics services.
3. **Design considerations**:
    - The design of a Data Lake or Data Warehouse must consider factors such as scalability, data security, ease of access, and integration with data processing and analysis tools.

Storage Scheme

1. **Scheme in reading vs. scheme in writing**:
    - **Schema in writing**: In this approach, data is transformed and structured before it is stored. It is typical in traditional Data Warehouses, where data is cleansed, transformed into a consistent format, and then loaded into storage. This facilitates subsequent reading and analysis operations, as the data is already in an optimized format.
    - **Schema-on-read**: In contrast, schema-on-read involves storing the data in its raw form and applying the structure or schema at read time. This approach is common in Data Lakes, where flexibility and the ability to store large volumes of data in diverse formats is a priority. Transformation and structuring of the data are performed as needed for specific analyses.
2. **Advantages and disadvantages**:
    - The choice between schema-on-read and schema-on-write depends on several factors such as the nature of the data, the processing and analysis requirements, and the need for flexibility versus optimized query performance.
3. **Implementation considerations**:
    - When deciding on the storage scheme, it is crucial to consider not only current but also future needs, including scalability, query performance, ease of maintenance, and integration with other tools and systems.

3. Data processing and cleaning

In any Big Data project, data processing and data cleansing are essential steps that precede analysis and gaining new insights. This phase is crucial because the data, as captured or collected, is often far from perfect or ready for analysis.

Using advanced tools to process, cleanse, and transform data ensures that data is ready for analysis and modeling, which in turn leads to more accurate and valuable insights.

Processing Tools:

1. **Apache Spark**: This tool is a popular choice for processing large volumes of data due to its speed and efficiency. Spark allows complex data analysis and processing operations to be performed on clusters of computers, distributing the work efficiently. It is especially useful for tasks that require multiple processing steps and data transformations, as it minimizes the need to repeatedly read and write to disk.
2. **Hadoop MapReduce**: This is another essential framework for large-scale data processing. MapReduce divides processing tasks into two phases: 'Map', which processes and transforms input data, and 'Reduce', which performs summarization and aggregation operations. Although it is not as fast as Spark for certain tasks, it is extremely scalable and efficient for processing huge data sets.

Data Cleansing:

1. **Error detection and correction**: Raw data may contain errors, outliers, or inconsistent information. Tools such as Python pandas can be used to inspect the data, identify anomalies or errors, and correct them. This correction may include removing erroneous records, correcting values, or imputing missing values.
2. **Handling missing values**: In many data sets, it is common to find missing values. Depending on the context, these can be handled in different ways, such as imputing a mean or median value or using more sophisticated model-based methods.
3. **Standardization of formats**: Data from different sources may have inconsistent formats. For example, dates and numbers may be in different formats. Standardization of these formats is essential for consistent and accurate analysis.

Data transformation:

1. **Normalization**: This process involves scaling the numerical data to fall within a specific range, such as 0 to 1, which is crucial for some ML algorithms that are sensitive to the scale of the data.
2. **Data type conversion**: Sometimes, it is necessary to convert data types, such as changing a categorical variable (e.g., "red", "blue", "green") into a numerical form that can be used in analysis and modeling.
3. **Creating derived features**: Often, the original features in the data are not sufficient for effective analysis. Here, new derived features can be created from existing data,

such as creating interaction variables in statistical models or calculating new metrics from existing variables.

4. Exploratory Data Analysis (EDA)

Exploratory data analysis is like a detective diving into the world of data to uncover clues, patterns, and anomalies. This phase is vital in any data analysis project, as it allows us to gain an intuitive and deep understanding of what the data is telling us before moving on to more complex stages such as statistical modeling or machine learning (Capa Benítez et al, 2017).

By combining powerful visualizations with solid descriptive statistics, we can gain a deep and comprehensive understanding of our data. This understanding is crucial for any subsequent analysis, ensuring that decisions based on this data are informed and reliable.

Data visualization:

1. **The importance of visualization**: In EDA, a picture is worth a thousand words. Data visualization is the tool that allows us to turn rows and columns of numbers into graphs and images that reveal the stories hidden in the data.
2. **Visualization tools**:
    - **Tableau and PowerBI**: These are powerful and user-friendly tools for business users and analysts. They allow you to create interactive dashboards and complex visualizations without intensive programming. With these tools, you can drag and drop elements to explore different aspects of your data, from general trends to minute details.
    - **Matplotlib in Python**: For those more technically inclined, Matplotlib offers incredible flexibility for customizing visualizations. Although it requires more coding skills, it allows you to create a wide range of plots, from histograms and scatter plots to more advanced visualizations.

Descriptive statistics:

1. **Unraveling the data with statistics**: While visualizations give us a picture, descriptive statistics provide us with the exact numbers and measures that define that picture. These statistics are the first step in quantifying the trends and patterns we observe visually.
2. **Trend and pattern analysis**: We use measures such as mean, median, mode, range, and standard deviation to understand the distribution and centrality of our data. For example, the mean tells us the average value, while the standard deviation shows us how much the data varies around that mean.
3. **Identification of outliers**: Outliers are data that deviate significantly from the rest. They may indicate measurement errors, incorrect inputs, or simply extreme natural variations. Identifying them is crucial, as they can have a significant impact on subsequent analyses.

5. Data preparation for models

Data preparation for models is a painstaking but crucial step in the data modeling process. It involves carefully selecting the most relevant features and strategically partitioning the data to train, refine, and evaluate your model. This phase lays the foundation for building robust and reliable models that can make accurate and useful predictions.

This stage is crucial in any data analysis project and focuses on two main aspects: feature selection and data partitioning.

Feature selection:

1. **Identifying key features**: Imagine that you have a treasure trove of information in front of you, but not all of it is gold. Some features (or variables) in your data are valuable, while others may be redundant or irrelevant. Feature selection involves identifying and keeping only those features that are truly important to your model.
2. **Selection techniques**: There are several techniques for this task, such as correlation analysis, variable significance testing, and automatic methods such as forward or backward selection. The goal is to reduce the size of your data to improve the efficiency of the model and often its performance.
3. **Avoiding overfitting**: A crucial part of feature selection is avoiding overfitting, where the model fits the training data too well and loses the ability to generalize to new data. By choosing only relevant features, you can help prevent this problem.

Data Division:

1. **Creating specific data sets**: Once you have your features selected, the next step is to divide your data into three sets: training, validation, and testing. Each of these sets has a specific purpose in building and evaluating your model.
2. **Training set**: This is the data set on which you will train your model. It is like the training ground where your model learns to identify patterns and make predictions.
3. **Validation set**: This set is used to tune the model parameters and to perform cross-validation. It is your tool for tuning the model, making sure it is learning correctly and not simply memorizing training data.
4. **Test Set**: Finally, the test set is like a final exam for your model. It is used to evaluate how the model performs on data you have never seen before. This set is crucial to get an unbiased estimate of the model's actual performance in the real world.

6. Modeling and machine learning algorithms

Modeling and algorithm selection in machine learning is like choosing and training an athlete for a specific competition. Each algorithm has its strengths and is best suited to certain types of problems and data. This stage of the ML process is where critical decisions are made about which methods to use and how to train them to get the best results.

This stage is essential to ensure that your model can make accurate and valuable predictions, which is the heart of any ML project.

Model Selection:

1. **The art of choosing algorithms**: Choosing the right ML algorithm depends largely on the type of problem you are trying to solve (Mirjalili and Raschka, 2020). For example, if you are trying to predict a numerical value, such as the price of a house, you might choose a regression model. If your goal is to classify items into categories, such as determining whether an email is spam or not, then a classification algorithm would be more appropriate.
2. **Types of algorithms**:
    - **Regression**: Used to predict continuous values. Examples include linear regression and logistic regression.
    - **Classification**: Used to predict discrete categories. Algorithms such as decision trees, support vector machines (SVM), and neural networks are popular in this area.
    - **Clustering**: Used to group similar data without predefined labels. Techniques such as K-means or hierarchical clustering fall into this category.
    - **Neural networks**: They are especially powerful for complex tasks such as image recognition and natural language processing.

Model Training:

1. **Learning from the data**: Once you have selected the right model, the next step is to train it with your training data set. During this process, the model learns to recognize patterns and relationships in the data, adjusting its internal parameters to make the best possible predictions or classifications.
2. **Iterations and improvements**: The training of a model is iterative. It may need to go through the data set several times, adjusting with each iteration to improve its accuracy.

Parameter Setting:

1. **The search for the perfect fit**: Parameter tuning is like tuning a musical instrument. It is about finding the perfect combination of parameters that allows the model to perform to the best of its ability.
2. **Adjustment techniques**:
    - **Grid search**: This technique involves testing several parameter combinations and selecting the one that gives the best performance.
    - **Cross-validation**: It is a method to evaluate the generalization of the model. It involves dividing the data set into several parts, training the model on some of them, and validating its performance on the remaining parts. This helps to ensure that the model not only performs well on the training data but also on new data.

7. Model evaluation

Model evaluation in machine learning is like conducting a thorough review of an athlete's performance after a competition. It is a crucial step that determines how well the model you have trained and tuned can cope with the task for which it was designed. This phase focuses on two key aspects: performance metrics and cross-validation.

Performance metrics and cross-validation techniques ensure that the model is not only accurate but also robust and reliable in different situations. This phase is critical to ensure that ML models are practical and useful in real-world situations, beyond controlled training and testing environments.

Performance metrics:

1. **The role of metrics**: Performance metrics are the indicators that tell us how well a model is performing. They are like scores in a sports competition, providing a quantitative assessment of the model's performance.
2. **Types of metrics**:
    - **Accuracy**: This metric is useful when the consequences of false positives are important. For example, in a model that identifies fraudulent transactions, the accuracy will tell us what percentage of transactions identified as fraudulent are fraudulent.
    - **Sensitivity (Recall)**: This is crucial when it is important to capture all positive cases. For example, in a disease detection model, you will want to capture as many actual cases of the disease as possible.
    - **AUC-ROC**: This is a metric that combines the true positive rate (recall) and false positive rate to give a complete picture of model performance. It is especially useful for binary classifiers and provides a good indication of how the model performs overall, regardless of the classification threshold.

Cross-validation:

1. **Ensuring robustness**: Cross-validation is a technique for evaluating how well a model generalizes to an independent data set. It is like subjecting the athlete to different tracks and conditions to make sure their performance is consistent and not just a fluke in a particular scenario.
2. **How it works**:
    - In cross-validation, you divide your data set into several parts (or "folds"). Then, you train your model on some of these parts and test it on the remaining parts. You repeat this process several times, changing the part of the data set you use for testing each time.
    - This technique allows you to evaluate the model's ability to fit different data and minimizes the risk that your results are biased by the way the data set was originally divided.

3. **Benefits of cross-validation**: In addition to providing a measure of model effectiveness, cross-validation can also help identify problems such as overfitting, where the model fits the training data too well and fails to generalize to new data.

8. Deployment of models

Model deployment in machine learning is the process of putting ML models into action in real environments, either on in-house servers or in the cloud and providing accessible means, such as APIs, to interact with these models. It resembles the final act in the production of a theatrical play, where all the previous work is put into action in front of a real audience. This step turns theoretical models and tests into practical and accessible solutions. In this scenario, the model ceases to be a project under development and becomes an operational tool in a real environment. There are two main components in this phase: the production environments and the APIs to access the models.

Production environments

After a model has been trained, tuned, and evaluated, the next step is to bring it into an environment where it can be used effectively. This involves deploying the model on a server or in the cloud, where it can process real data and provide useful results.

Deployment options:

Servers: Deploying on in-house servers means having complete control over the hardware and software environment. This is useful in cases where data security is a primary concern or when specific configurations are required.

Cloud: Cloud services such as AWS, Google Cloud Platform, or Microsoft Azure offer flexibility, scalability, and reduced maintenance. You can choose from a variety of computing and storage options depending on the needs of your model and pay only for what you use.

The choice of production environment must consider factors such as the amount of data to be processed, performance requirements, security, and ease of maintenance.

APIs for model access

Once the model is in production, you need a way to interact with it. This is where APIs (Application Programming Interfaces) come into play. They are like the interpreters in a concert, facilitating communication between the model and the users or applications that want to use it.

Developing an API for your model involves defining how external users can send data to the model and receive predictions or analysis. This is usually done through HTTP requests, where users send data and receive responses in formats such as JSON.

Benefits of APIs:

- **Accessibility**: APIs allow different applications, from enterprise software to mobile applications, to access the model easily and securely.

- **Flexibility**: With an API, you can update or modify the model without changing the way users interact with it, allowing constant evolution and improvement of the model.

9. Visualization of results

In the data analytics journey, visualization of results is the stage where the stories hidden in the numbers are transformed into understandable visual narratives. It is the moment when the abstract results of Machine Learning models are materialized into forms that are easy to interpret and appealing to the viewer. This phase focuses on two key elements: the use of advanced visualization tools and the creation of interactive dashboards.

Through them, we can ensure that the results of our ML models are accessible, understandable, and useful for decision-making. This phase not only helps to interpret the results but also plays a crucial role in the communication and effective use of the information obtained through data analysis.

Visualization tools:

1. **Bringing data to life**: Advanced visualization tools are the brushes and colors we use to paint the picture of our results. They allow us to transform the complex results of ML models, often in the form of cryptic numbers and tables, into clear and understandable graphs.
2. **Tool selection**:
    - For statistical and scientific visualizations, tools such as Matplotlib and Seaborn in Python are excellent choices. They offer a wide variety of graphs and customizations to effectively represent complex results.
    - For more business and decision-oriented visualizations, tools such as Tableau or PowerBI are ideal. They allow you to create attractive and interactive visualizations that can be easily understood and interpreted by non-technical users.
3. **Effective communication**: The key to good visualization is your ability to communicate results effectively. This means choosing the right type of chart for your data, using colors and labels intelligently, and presenting information in a way that is both informative and engaging.

Interactive dashboards:

1. **Data command centers**: Interactive dashboards are like the dashboard of an airplane, providing all important information in one easy-to-access place. They are dynamic platforms where the results of ML models can be explored and analyzed in real-time.
2. **Dashboard construction**:
    - When building a dashboard, it is important to focus on usability and relevance. This means presenting key performance indicators (KPIs) clearly,

allowing interactions such as filtering and drilling down into data, and making sure the dashboard is intuitive and easy to navigate.

- Tools such as Plotly's Dash or Shiny in R allow you to create customized dashboards that can display data in real-time, update it automatically, and allow the user to interact with the data in a variety of ways.

3. **Benefits of dashboards**:
    - Dashboards are especially valuable in business and decision-making environments, where stakeholders need a fast and reliable way to access the latest results and analysis.
    - They provide an efficient way to continuously monitor model performance, identify trends, and make quick adjustments as needed.

10. Monitoring and maintenance

Once a machine learning model is up and running, the work doesn't stop there. Imagine that the model is a garden that you have planted and now needs constant care and attention to thrive. Monitoring and maintenance of the models are essential to ensure that they remain accurate and relevant over time. These processes include continuous monitoring of model performance and periodic updating of models as needed.

Like a garden, models require care, attention, and periodic adjustments to ensure that they continue to flourish in the changing landscape of data and business needs. These practices not only ensure the model's continued accuracy and effectiveness but also safeguard against potential problems that could arise due to changes in data patterns.

Continuous monitoring

1. **Constant vigilance**: Continuous monitoring of the model's performance is like having a surveillance system in your garden. You need to make sure everything is working as expected and watch for any signs of trouble.
2. **Tools and techniques**:
    - The use of real-time dashboards and automatic alerts can be crucial to keep an eye on model performance. These tools allow you to see at a glance how the model is performing and receive notifications if anything goes outside the normal parameters.
    - Key performance metrics, such as accuracy, recall, and F1-score, should be constantly monitored. In addition, it is important to be aware of changes in input data that could affect model performance.
3. **Responding to change**: Just as in a garden where weather conditions can change, in the world of Machine Learning, data patterns can evolve. Continuous monitoring allows you to identify and respond quickly to these changes, ensuring that your model does not become obsolete.

Updating of models

1. **Maintaining relevance**: Just as plants need to be pruned and nurtured, ML models need to be retrained and updated regularly to maintain their relevance and accuracy.
2. **Updating process**:
    - Updating models may involve collecting new data, readjusting parameters or even completely changing the model approach if the context or objectives have changed significantly.
    - Retraining with more recent data is essential, especially in constantly changing environments where trends and patterns can evolve rapidly.
3. **Post-update evaluation**: Each time a model is updated, it is crucial to evaluate its performance to ensure that the improvements are effective. This involves repeating part of the validation and testing process that was carried out during the initial model development phase.

11. Documentation and reporting

The last step in Big Data project management, and no less important, is documentation and reporting. If we consider the entire project as a long journey, documentation and reporting would be the travel diary and photo album that chronicles what has been done, how it has been done, and the discoveries made along the way. This stage is vital to ensure that the work done is understandable, replicable, and valuable to both the technical team and the stakeholders involved.

Documentation and reporting are essential to close the loop of a Big Data project. Technical documentation ensures that the project is transparent and accessible, while results reports communicate findings and recommendations effectively to a wider audience. Together, they ensure that the value of the project is understood and fully utilized and that the knowledge generated can be a springboard for future efforts and improvements.

Technical documentation

1. **Creating a detailed record**: The technical documentation is like the instruction manual for the entire project. It includes details on the processes used, decisions made, models developed, and any other relevant technical information.
2. **Key components**:
    - **Data description and preprocessing**: Document data sources, cleaning and transformation methods used, and any other data manipulation performed.
    - **Model details**: Include information on model selection, parameters used, training process, testing, and validation results.

- **Code and algorithms**: Maintain a record of the code used, preferably with detailed comments and examples of use.

3. **Importance of documentation**: Clear and complete documentation is essential for future maintenance and updating of the project. It facilitates understanding of the work performed and allows others to replicate or build on the project with ease.

Results Reports

1. **Communicating findings**: Results reports are like telling the story of the project to those who were not on the journey. These reports should present the insights gained, recommendations based on the results, and any important conclusions from the project.
2. **Elements of a good report**:
    - **Executive summary**: A high-level summary of the findings and recommendations, designed to be accessible to those without detailed technical knowledge.
    - **Results and analysis**: A detailed description of the model results, including visualizations and explanations of what these results mean in the context of the problem.
    - **Recommendations and next steps**: Based on the results, offer practical recommendations, and suggest areas for future research or improvements.
3. **Effective communication**: Reports should be clear, concise, and focused on key questions and project objectives. They should be understandable to a diverse audience, from the technical team to non-technical users.

This outline provides a general guideline for Big Data projects aimed at prediction or classification. Depending on the specific needs of the project, some steps may require more emphasis or be adapted.

5.4 Preparation for a future economist with IA

Economists who fail to adapt to the growing influence of AI and data analytics could face significant difficulties in their professional development and employability in an increasingly digitized and data-driven economic environment.

There are many reasons for this, but here are some of the main ones:

1. **Limitation in employment opportunities**: AI and data analytics skills are increasingly in demand. Economists who do not possess them could face limitations in their career options, especially in innovative and technologically advanced sectors.
2. **Disadvantage in research and analysis**: AI allows for handling large data sets and complexities that are difficult to address with traditional methods. Without these

skills, economists could fall behind in terms of accuracy and depth in their research and analysis.

3. **Disconnect with current trends**: The field of economics is evolving with the integration of advanced technologies. Not keeping up with these trends can lead to a disconnect with contemporary practices and approaches.

4. **Reduced competitiveness in the labor market**: In an increasingly competitive and technologically oriented labor market, a lack of AI skills could reduce an economist's competitiveness vis-à-vis more technologically adapted colleagues.

Although responding to these challenges is difficult due to the great technological and other uncertainties that will mark the future in the coming years, here is a decalogue of advice that may be useful for a young economist starting his or her professional career in an environment where artificial intelligence (AI) is becoming increasingly relevant:

1. **Learn the basics of AI and machine learning**: Become familiar with fundamental AI and ML concepts, such as classification and regression algorithms, neural networks, and deep learning.

2. **Understand the application of AI in economics**: Explore how AI is used in economic analysis, from data processing to predicting market trends.

3. **Improve your programming skills**: Acquire knowledge in programming languages used in AI, such as Python or R.

4. **Develop competencies in data analysis**: Learn how to manage and analyze large data sets, an essential skill in the era of big data.

5. **Follow AI trends and developments**: Stay up to date on the latest developments in AI and how they influence the economy.

6. **Participate in projects that use AI**: Look for opportunities to work on projects that apply AI, even if it is in a secondary role.

7. **Collaborate with AI experts**: Work with AI professionals to better understand its application in economic contexts.

8. **Be critical and ethical in the use of AI**: Consider the ethical implications and possible biases in AI models.

9. **Foster a continuous learning mindset**: AI is a rapidly evolving field; it is crucial to maintain a constant learning attitude.

10. **Apply critical and analytical thinking**: Use your economic skills to interpret and question the results obtained by AI models.

To summarize, to facilitate the entry of young economists into an AI-dominated professional environment, the two key ideas are:

Acquire competencies in AI and data analysis: It is essential to become familiar with the basics of AI and machine learning, as well as improve skills in data analysis and programming.

Keep up-to-date and apply critical thinking: Stay informed about advances in AI and its applications in economics and apply a critical and ethical approach to its use to ensure accurate and responsible interpretations.

By following these tips, a young economist can effectively integrate AI into his or her professional practice, staying relevant and competitive in today's changing economic world.

6. Final Reflections

Artificial intelligence (AI) is reshaping the global economic landscape in a way that we are only beginning to understand. Its impact is vast and varied, extending across industries and borders. The economic effects of AI promise to be transformative, redefining operations and strategies in sectors as diverse as manufacturing, services, and finance. This new era of innovation and efficiency is driving significant changes in productivity, labor market dynamics, and industrial structures.

Automation and skill change are crucial aspects of this shift. AI is redefining labor markets by automating routine and predictable tasks, which in turn changes the demand for skills in the workforce. Some jobs disappear, while others emerge, especially those requiring analytical, creative, or interpersonal skills that AI cannot easily replicate. This change poses significant challenges, including the need for job matching and managing economic inequalities.

Parallel to its economic impact, AI is shaping the social future in fundamental ways. Not only is the technology automating tasks, but it is also raising critical questions about ethics, privacy, and the nature of human labor. AI, as it becomes an integral part of our daily lives, forces us to confront its influence on privacy, the biases inherent in algorithms, and the need for proper governance of these systems.

Privacy and data use issues are especially acute in systems that rely on the analysis of large amounts of personal data. In addition, biases in AI are a growing concern, as they can perpetuate and amplify existing inequalities in society. The need for balanced governance and regulation is imperative to balance innovation with the ethical use of AI, involving governments, businesses, and civil society.

AI also has the potential to improve decision-making in politics and public administration but presents risks such as the possibility of manipulation or misuse of the technology for undemocratic purposes.

The AI era offers significant opportunities for economic advancement and improved quality of life. However, it also presents complex and multifaceted challenges that require careful and balanced consideration. How society and governments address these challenges will largely determine the long-term impact of AI on our economic and social lives. Balancing the urgency to innovate with a commitment to uphold ethical and social principles will be crucial to forging a future where technology aligns with human values and well-being.

References

Aguado Sarrió, G. (2015). *Application of machine learning techniques on games* (Doctoral dissertation, Universitat Politècnica de València).

Alonso, Andrés and Carbó, José Manuel, Inteligencia Artificial Y Finanzas: Una Alianza Estratégica (Artificial Intelligence and Finance: A Strategic Alliance) (October 19, 2022). Banco de Espana Occasional Paper No. 2222, Downloable document is in Spanish, 2022, Available at SSRN: https://ssrn.com/abstract=4252710 or http://dx.doi.org/10.2139/ssrn.4252710http://dx.doi.org/10.2139/ssrn.4252710

Alonso, Andrés and Carbó, José Manuel, Inteligencia Artificial Y Finanzas: Una Alianza Estratégica (Artificial Intelligence and Finance: A Strategic Alliance) (October 19, 2022). Banco de Espana Occasional Paper No. 2222, Downloadable document is in Spanish, 2022, Av.

Alvarez, F. (2020). Machine Learning in e-commerce fraud detection applied to banking services. *Science and Technology*, 81-95.

Ameijeiras Sánchez, D., Valdés Suárez, O., & González Diez, H. (2021). Anomaly detection algorithms with deep networks. Review for bank fraud detection. *Cuban Journal of Computer Science*, *15*(4), 244-264.

Badaró, S., Ibañez, L. J., & Agüero, M. J. (2013). Expert systems: fundamentals, methodologies, and applications. *Ciencia y tecnología*, (13), 349-364.

Boden, M. A. (2017). *Artificial intelligence*. Turner.

Borrero-Tigreros, D., & Bedoya-Leiva, O. F. (2020). Predicting credit risk in Colombia using artificial intelligence techniques. *Revista UIS Ingenierías*, *19*(4), 37-52.

Bouza, C., & Santiago, A. (2012). Data mining: decision trees and their application in medical studies. *Mathematical modeling of environmental and health phenomena*, *2*, 64-78.

Camacho, M., Ramallo, S., & Marín, M. R. (2021). Decision trees in economics: an application to house pricing. In *New methods of economic forecasting with massive data* (pp. 61-92). Fundación de las Cajas de Ahorros (FUNCAS).

Capa Benítez, L. B., García Saltos, M. B., Crespo Hurtado, E., Palmero Urquiza, D. E., López Fernández, R., Franco Fadul, M. D. C., & Fadul Franco, J. S. (2017). *Exploratory data analysis with SPSS*. Quito, Universidad Metropolitana.

Ceballos Mina, O. E., & Duque García, C. A. (2022). Econometrics in economics programs: myths and teaching-learning barriers. *Nicolaita Journal of Economic Studies*, *17*(1), 65-82.

Clifton, J., & Laber, E. (2020). Q-learning: Theory and applications. *Annual Review of Statistics and Its Application*, *7*, 279-301.

Cordero-Torres, B. P. (2022). Supervised Learning Algorithms for Sales Projection of Ecuadorian Shrimp with Python Programming Language. *Economics and Business*, *13*(2

Cortina Orts, A. (2019). Ethics of artificial intelligence. In *Annals of the Royal Academy of Moral and Political Sciences (*pp. 379-394). Ministry of Justice.

Corvalán, J. G. (2019). The impact of artificial intelligence at work. *Revista de Direito Econômico e Socioambiental*, *10*(1), 35-51.

Cunningham, P., Cord, M., & Delany, S. J. (2008). Supervised learning. In *Machine learning techniques for multimedia: case studies on organization and retrieval* (pp. 21-49). Berlin, Heidelberg: Springer Berlin Heidelberg.

Dávila Morán, R. C., & Agüero Corzo, E. del C. (2023). Ethical challenges of artificial intelligence: implications for society and economy. Revista Conrado, 19(94), 137-144. https://conrado.ucf.edu.cu/index.php/conrado/article/view/3326.

Desai, A. (2023). Machine Learning for Economics Research: When What and How? *arXiv preprint arXiv:2304.00086*.

Dimonopoli, S. (2022). Data driven economy: come il processo di informatizzazione ha portato ad una nuova economia basata sui big data. Tesi di Laurea in Storia dell'economia e dell'impresa, Luiss Guido Carli, relatore Esposito Guido Tortorella.

García Novoa, C.; Vivel-Búa, M (2022) Studies of the impact of digitalization on the economy. Aranzadi. Civitas, 22 April, 432 pp.

Giraldo Escobar, S. A. (2021). *Algorithmic trading of actions through deep reinforcement learning* (Doctoral dissertation, Universidad Nacional de Colombia).

Hersh, J., & Harding, M. (2018). Big data in economics. *IZA World of Labor*, 2018: 451.

Hoz-Dominguez, E. J., Fontalvo-Herrera, T. J., & Escorcia-Guzman, A. (2019). Creating business profiles for exporting companies using unsupervised learning. *Información tecnológica*, *30*(6), 193-200.

Kotler, P., Setiawan, I., & Setiawan, H. (2022). *Marketing 5.0 Colombia Version: Technology for Humanity*. LID Editorial.

LaPlante, A. (2016). *Architecting data lakes*. O'Reilly Media.

Mamaqi, X., Lope Salvador, V., & Vidal Bordes, J. (2018). Datification, big data, and artificial intelligence in communication and economics. *Datification, big data and artificial intelligence in communication and economics*, 65-82.

Márquez Benavides, L., Moreno Goytia, E. L., & González Ramírez, L. F. (2023). The use of artificial intelligence in an academic environment. *Ciencia Nicolaita*, 89, 244-255.

Mirjalili, V., & Raschka, S. (2020). *Python machine learning*. Marcombo.

Montenegro Meza, M. A., Menchaca Méndez, R., & Menchaca Méndez, R. (2023). A gentle but rigorous introduction to reinforcement learning. *ReCIBE, Electronic Journal of Computing, Informatics, Biomedical and Electronics*, *12*(1), C1-13.

Munárriz, L. Á. (1994). *Fundamentals of artificial intelligence* (Vol. 1). Editum.

Nisbet R., Elder J. & Miner G. (2009). Top 10 Data Mining Mistakes. In *Handbook of Statistical Analysis and Data Mining Applications* (pages 733-754).

Ojeda, S., Pereyra, L. E., & Gontero, S. (2005). Household poverty in Greater Córdoba: application of the logistic regression model. *Journal of Economics and Statistics*, *43*(1), 99-121.

Olarte, E., Panizzi, M. D., & Bertone, R. A. (2018). Market segmentation using data mining techniques in social networks. In *XXIV Congreso Argentino de Ciencias de la Computación (La Plata, 2018)*....

Olguín Gallardo, A. (2018). Relationship between economics and some artificial intelligence paradigms. *Transcender, Accounting and Management*, (7), 26-33.

Oliva Rodríguez, A. (2018). *Development of an image recognition application using Deep Learning with OpenCV* (Doctoral dissertation, Universitat Politècnica de València).

Patel, J., Shah, S., Thakkar, P., & Kotecha, K. (2015). Predicting stock market index using fusion of machine learning techniques. *Expert Systems with Applications*, *42*(4), 2162-2172.

Pérez Verona, I. C., & Arco García, L. (2016). A review on unsupervised learning of distance metrics. *Cuban Journal of Computer Science*, *10*(4), 43-67.

Ponce, P. (2010). *Artificial intelligence: with applications to engineering*. Alpha Editorial.

Quiguiri Daquilema, C. M. (2023). *Contrasting Machine Learning and Econometrics in time series* (Bachelor's thesis, Quito: EPN, 2023.).

Quintía Vidal, P. (2013). *Robots capable of learning and adapting to the environment from their own experiences* (Doctoral dissertation, Universidade de Santiago de Compostela).

Romero Martínez, M., Carmona Ibáñez, P., & Pozuelo Campillo, J. (2021). Utility of Deep Learning in the prediction of business failure at the European level. *Journal of Quantitative Methods for Economics and Business*, *32*, 392-414.

Sánchez Anzola, N. (2015). Support vector machines and artificial neural networks in the prediction of intraday USD/COP spot intraday movement. *ODEON-Observatory of Economics and Numerical Operations*, 9, 115-172.

The Economist (2017). China may match or beat America in AI. The Economist, business section, 21st issue of 2017.

Valenzuela González, G. (2022). Supervised Learning: Methods, Properties and Applications.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.