U3 MACHINE LEARNING ALGORITHMS

U3.E2 UNSUPERVISED MODELS

Machine Learning Engineer

January 2021, Version 1

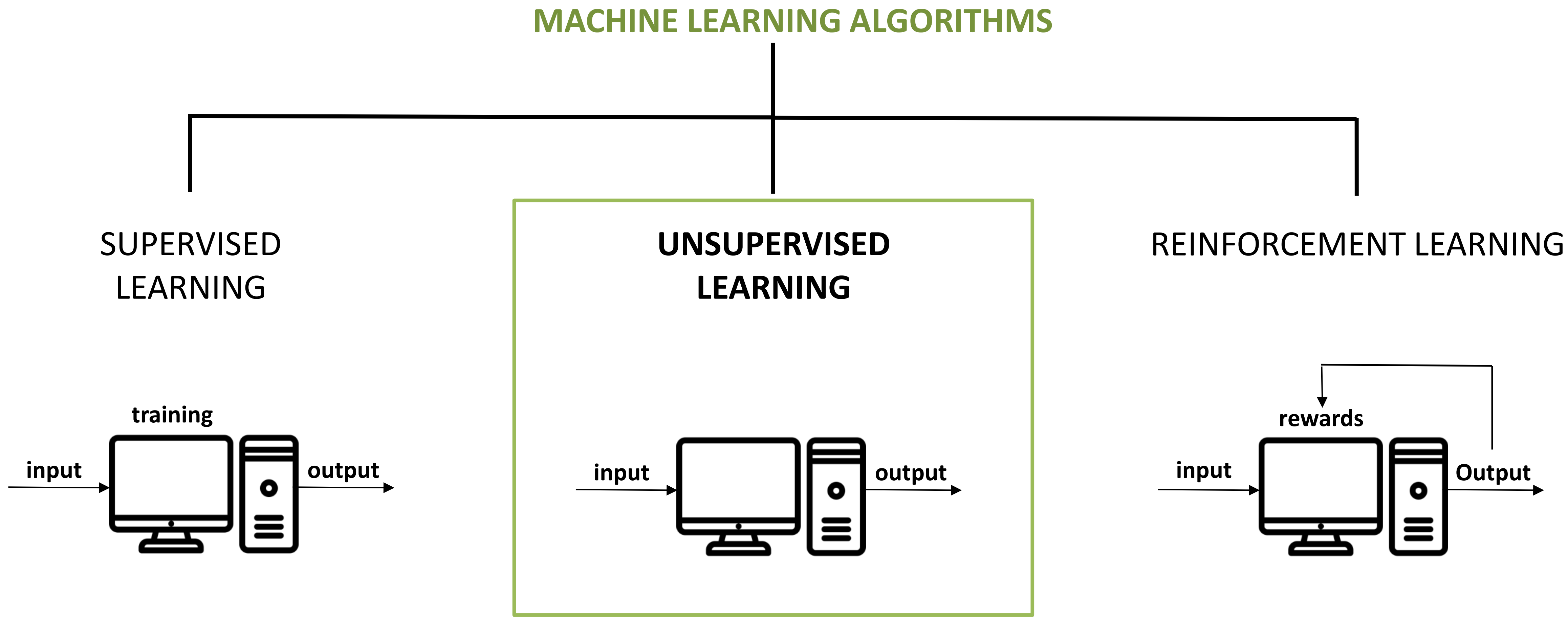
The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

LEARNING OBJECTIVES

The student is able to

MLE.U3.E2.PC1	Define and explain unsupervised learning.
MLE.U3.E2.PC2	Know some of the most commonly used algorithms in unsupervised learning.
MLE.U3.E2.PC3	Know the domain application of unsupervised models.

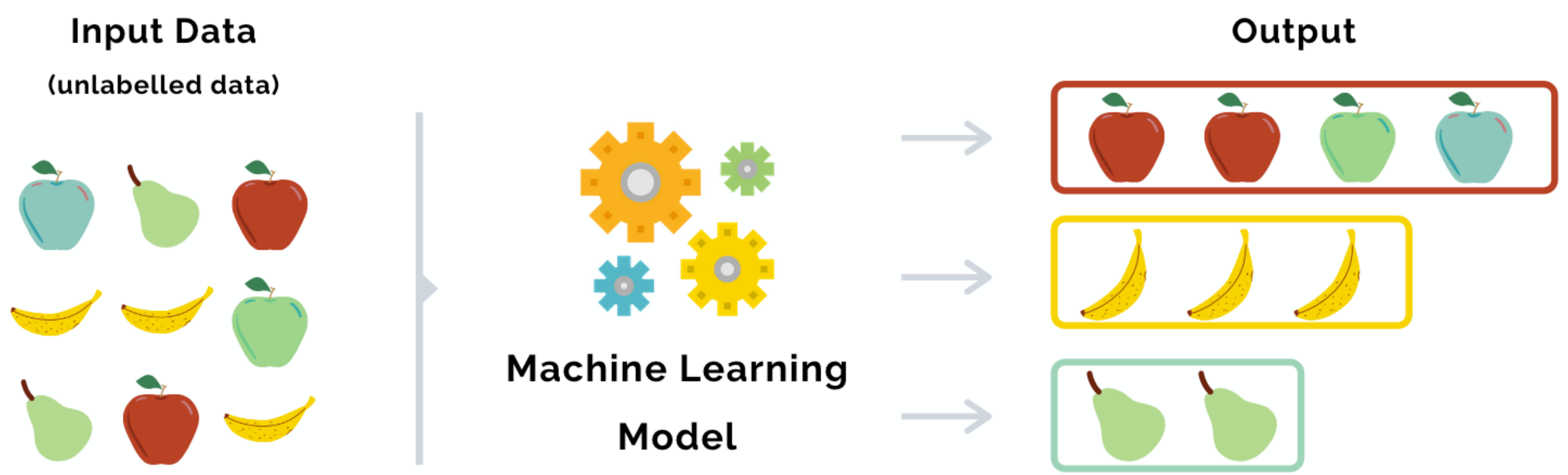
MACHINE LEARNING ALGORITHMS



DRIVES
Development and Research on Innovative Vocational Education Skills

Unsupervised Learning is a class of Machine Learning algorithms that uses them to analyze and cluster unlabeled datasets

In Unsupervised learning the goal is to learn useful structure without labeled classes, optimization criterion, feedback signal, or any other information beyond the raw data



Prime reasons to use Unsupervised Learning:

1. Unsupervised machine learning finds all kind of unknown patterns in data.

2. Unsupervised methods help you to find features which can be useful for categorization.

3. It is taken place in real time, so all the input data to be analyzed and labeled in the presence of learners.

4. It is taken place in real time, so all the input data to be analyzed and labeled in the presence of learners.

DRIVES
Development and Research on Innovative Vocational Education Skills

Unsupervised Learning can be further classified into two categories:

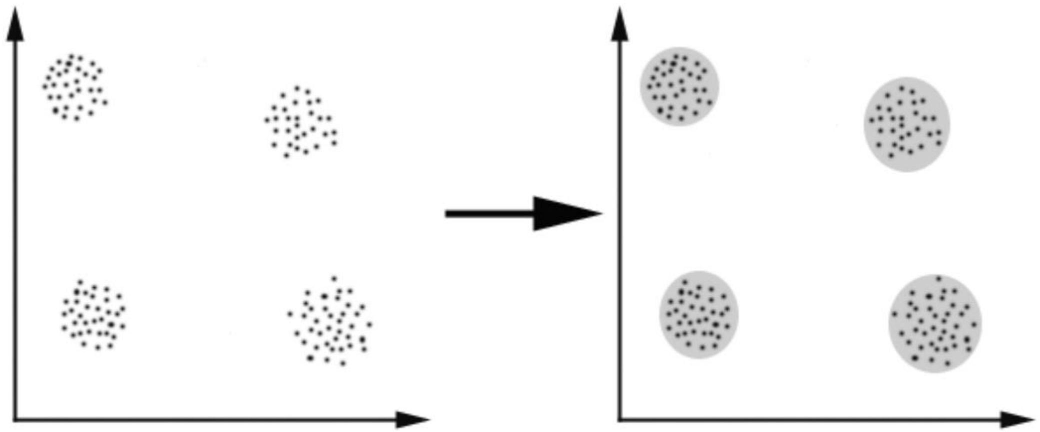
Parametric Unsupervised Learning

- Assumes that sample data comes from a population that follows a probability distribution based on a fixed set of parameters.
- Involves construction of Gaussian Mixture Models and using Expectation-Maximization algorithm to predict the class of the sample in question

Non-Parametric Unsupervised Learning

- The data is grouped into clusters, where each cluster says something about categories and classes present in the data.
- Do not require the modeler to make any assumptions about the distribution of the population, and so are sometimes referred to as a distribution-free method.

A *cluster* is a collection of objects which are "similar" between them and are "dissimilar" to the objects belonging to other clusters.



***Distance-based clustering*.**

Given a set of points, with a notion of distance between points, grouping the points into some number of *clusters*, such that:

- internal (within the cluster) distances should be small
- external (intra-cluster) distances should be large

1 **Exclusive Clustering: K-means**

Most common type of clustering. Each object belongs to na exclusive cluster. Data point belongs to a definite cluster then it could not be included in another cluster.

2 **Overlapping Clustering: Fuzzy C-means**

Uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership.

3 **Hierarchical Clustering: Agglomerative clustering, divisive clustering**

Is based on the union between the two nearest clusters. The beginning condition is realized by setting every data point as a cluster. After a few iterations it reaches the final clusters wanted

4 **Probabilistic Clustering: Mixture of Gaussian models**

Data points are clustered based on the likelihood that they belong to a particular distribution

DRIVES
Development and Research on Innovative Vocational Education Skills

- The main objective of the K-Means algorithm is to minimize the sum of the distances between the points and their grouping centroid.
- It is an iterative algorithm that tries to partition the data set into distinct K subgroups (clusters) without overlapping

The k-means algorithm works as follows:

1. Specify number of K clusters.
2. Initialize the centroids by shuffling the data set first and randomly selecting K data points for the centroids without substitution.
3. Continue iterating until there are no changes to the centroids. i.e., the assignment of data points to clusters is not changing.

DRIVES
Development and Research on Innovative Vocational Education Skills

Advantages

01 Simple, fast to compute

02 Converges to local minimum of within-cluster squared error

03 Since both k and t are small. k-means is considered a linear algorithm.

Disadvantages

01 The user needs to specify k

02 Sensitive to initial centers and outliers

03 Detects spherical clusters

04 Assumes that means can be computed

It is similar in process to the K-Means clustering but it works differently:

1. Choose a number of clusters (K).
2. Assign coefficients randomly to each data point for being in the clusters.
3. Repeat until the algorithm has converged.
4. Compute the centroid for each cluster.

Therefore this algorithm will not overfit the data for clustering like the k-means algorithm it will mark the data point to multiple clusters instead of the one cluster which will be more helpful than giving the point to the one cluster.

DRIVES
Development and Research on Innovative Vocational Education Skills

Advantages

01 Allows a data point to be in multiple clusters

02 More natural representation of the behavior of genes

DRIVES
Development and Research on Innovative Vocational Education Skills

Disadvantages

01 Need to define c (k in K-means), the number of clusters

02 Need to determine membership cutoff value

03 Clusters are sensitive to initial assignment of centroids

04 Fuzzy c-means is not a deterministic algorithm

It can be categorized in two ways:

Agglomerative clustering –
Bottom Up Approach

Divisive clustering –
Top Down Approach

Four different methods are commonly used to measure similarity:

- **Ward's linkage:** States that the distance between two clusters is defined by the increase in the sum of squared after the clusters are merged.
- **Average linkage:** Defined by the mean distance between two points in each cluster
- **Complete (or maximum) linkage:** Defined by the maximum distance between two points in each clustr
- **Single (or minimum) linkage:** Defined by the minimum distance between two points in each cluster

DRIVES
Development and Research on Innovative Vocational Education Skills

- It is the most common type of hierarchical clustering used to group objects in clusters based on their similarity.
- It is also known as AGNES, Agglomerative Nesting.

The AGNES algorithm works as follows:

1. Preparing the data. The data should be a numeric matrix with rows(representing observations) and columns (representing variables).
2. Compute similarity information between each pair of objects in the dataset
3. Using linkage function, groups objects into hierarchical cluster trees.
4. Determines where to cut the hierarchical trees into clusters. Creates partitions of the data

Advantages

01 No assumption of a particular number of clusters

02 May correspond to meaningful taxonomies

03 Easy to implement and gives best result in some cases.

Disadvantages

01 Once a decision is made to combine two clusters, it can't be undone

02 Too slow for large data sets, O(n2 log(n))

03 Based on the type of distance matrix chosen for merging different algorithms can suffer with one or more of sensitive noise, outliers,...

04 No objective function is directly minimized

Top-down clustering requires a method for splitting a cluster that contains the whole data and proceeds by splitting clusters recursively until individual data have been splited into singleton cluster.

1. Considers the entire data as one group
2. Iteratively splits the data into subgroups
3. If the number of a hierarchical clustering algorithm is known, then the process of division stops once the number of clusters is achieved.
4. Else, the process stops when the data can be no more split

DRIVES
Development and Research on Innovative Vocational Education Skills

Advantages

01 More efficient than agglomerative clustering

02 Takes into consideration the global distribution of data when making top-level partitioning decisions.

Disadvantages

01 More complex compared to agglomerative clustering

02 Needs a flat clustering method as “subroutine” to split each cluster

DRIVES
Development and Research on Innovative Vocational Education Skills

- The Gaussian Mixture Model (GMM) is the one of the most commonly used probabilistic clustering methods.

- GMM are classified as mixture models, because they are made up of an unspecified number of probability distribution functions.

Advantages

01 Mixture models are more general than partitioning and fuzzy clustering

02 Clusters can be characterized by a small number of parameters

Disadvantages

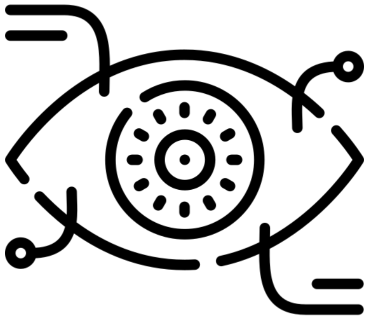
01 Computationally expensive if the number of distributions is large, or the data set contains very few observed data points

02 Need large data sets

03 Hard to estimate the number of clusters

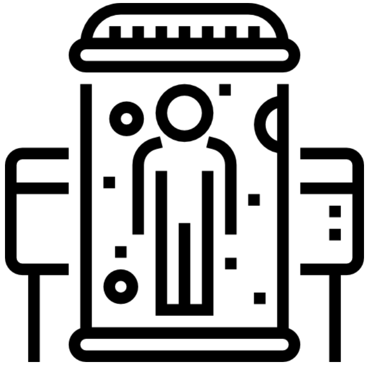
News Sections

Google News uses unsupervised learning to categorize articles on the same story from various online news outlets.



Computer vision

Unsupervised learning algorithms are used for visual perception tasks, such as object recognition.

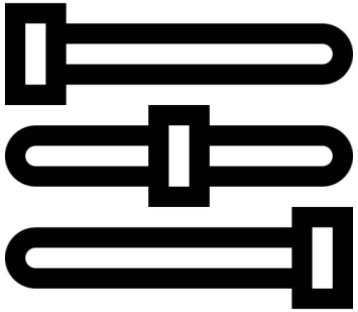


Medical imaging

Unsupervised machine learning provides essential features to medical imaging devices, such as image detection, classification and segmentation

Anomaly detection

Unsupervised learning models can comb through large amounts of data and discover atypical data points within a dataset.



Customer personas

Unsupervised learning allows businesses to build better buyer persona profiles, enabling organizations to align their product messaging more appropriately.

Recommendation Engines

Unsupervised learning can help to discover data trends that can be used to develop more effective.

DRIVES
Development and Research on Innovative Vocational Education Skills

Advantages

01 Less complexity in comparison with supervised learning.

02 No one is required to understand and then to label the data inputs.

03 Takes place in real time such that all the input data to be analyzed and labeled in the presence of learners

04 It is often easier to get unlabeled data

DRIVES
Development and Research on Innovative Vocational Education Skills

Disadvantages

01 You cannot get very specific about the definition of the data sorting and the output.

02 The results of the analysis cannot be ascertained.

03 There is no prior knowledge in the unsupervised method of machine learning.

04 The numbers of classes are also not known. It leads to the inability to ascertain the results generated by the analysisexecute requires

Development and Research on Innovative Vocational Education Skills

- Unsupervised Learning is a class of Machine Learning algorithms that uses them to analyze and cluster unlabeled datasets.
- There are several reasons to use unsupervised learning algorithms: searching for all kinds of unknown patterns in the data.
- Unsupervised Learning can be classified into two categories: Parametric or Non-Parametric
- Unsupervised learning is based on clustering
- There are 4 types os clustering algorithms: Exclusive Clustering, Overlapping Clustering, Hierarchical Clustering, Probabilistic Clustering.
- The most used algorithms are: K-means, Fuzzy C-Means, AGNES
- These algorithms can have several practical applications of which the detation of anomalies and recommendation systems stand out

- Barlow, H. B. (1989). Unsupervised learning. *Neural computation*, *1*(3), 295-311.
- Celebi, M. E., & Aydin, K. (Eds.). (2016). *Unsupervised learning algorithms*. Berlin: Springer International Publishing.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. In *The elements of statistical learning* (pp. 485-585). Springer, New York, NY.
- Hamerly, G., & Elkan, C. (2004). Learning the k in k-means. *Advances in neural information processing systems*, *16*, 281-288.
- Xu, R., & Wunsch, D. (2008). *Clustering* (Vol. 10). John Wiley & Sons.
- Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461.
- Jipkate, B. R., & Gohokar, V. V. (2012). A comparative analysis of fuzzy c-means clustering and k means clustering algorithms. *International Journal Of Computational Engineering Research*, *2*(3), 737-739.
- Zhang, S., Wang, R. S., & Zhang, X. S. (2007). Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and its Applications*, *374*(1), 483-490.

- Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, *32*(3), 241-254.
- Day, W. H., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, *1*(1), 7-24.
- Guénoche, A., Hansen, P., & Jaumard, B. (1991). Efficient algorithms for divisive hierarchical clustering with the diameter criterion. *Journal of classification*, *8*(1), 5-30.
- Sisodia, D., Singh, L., Sisodia, S., & Saxena, K. (2012). Clustering techniques: a brief survey of different clustering algorithms. *International Journal of Latest Trends in Engineering and Technology (IJLTET)*, *1*(3), 82-87.
- Reynolds, D. A. (2009). Gaussian Mixture Models. *Encyclopedia of biometrics*, *741*, 659-663.
- https://people.csail.mit.edu/rameshvs/content/gmm-em.pdf
- https://www.sciencedirect.com/topics/engineering/gaussian-model
- https://brilliant.org/wiki/gaussian-mixture-model/

REFERENCE TO AUTHORS

Diana Ferreira

- PhD student in Biomedical Engineering
- Research Collaborator of the Algoritmi Research Center

iD [0000-0003-2326-2153]

Regina Sousa

- PhD student in Biomedical Engineering
- Research Collaborator of the Algoritmi Research Center

iD [0000-0002-2988-196X]

José Machado

- Associate Professor with Habilitation at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

iD [0000-0003-4121-6169]

REFERENCE TO AUTHORS

António Abelha

- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

0000-0001-6457-0756

Victor Alves

- Assistant Professor at the University of Minho
- Integrated Researcher of the Algoritmi Research Center

0000-0003-1819-7051

REFERENCE TO AUTHORS

This Training Material has been certified according to the rules of **ECQA – European Certification and Qualification Association.**

The Training Material was developed within the international job role committee **“Machine Learning Engineer”:**

UMINHO – University of Minho (https://www.uminho.pt/PT)

The development of the training material was partly funded by the EU under Blueprint Project DRIVES.

Thank you for your attention

DRIVES project is project under **The Blueprint for Sectoral Cooperation on Skills in Automotive Sector**, as part of New Skills Agenda.

The aim of the Blueprint is **to support an overall sectoral strategy and to develop concrete actions to address short and medium term skills needs.**

Follow DRIVES project at:

More information at:

www.project-drives.eu

The Development and Research on Innovative Vocational Educational Skills project (DRIVES) is co-funded by the Erasmus+ Programme of the European Union under the agreement 591988-EPP-1-2017-1-CZ-EPPKA2-SSA-B. The European Commission support for the production of this publication does not constitute endorsement of the contents which reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.