T.C.

TOROS UNIVERSITY

FACULTY OF ENGINEERING

DEPARTMENT OF INDUSTRIAL ENGINEERING

MACHINE LEARNING CONCEPTS-DATA SCIECNCE-CRISP-DM

YUSUF CAN İBİŞOĞLU

195050021

ADVISOR

MEHMET ALİ AKTAŞ

COMPUTER AND SOFTWARE ENGINEERING PROJECT

FEBRUARY 2024

MACHINE LEARNING DATA SCIENCE AND CRISP-DM APPLICATION OF URINE TRACY INFECTOIN CLASSIFIER

ABSTRACT

SUMMARY

In this graduation project, I focused on areas of Machine Learning, Data Science, and CRISP-DM, one of the principles of Data Science. I started by establishing a solid foundation for the fundamental concepts governing these areas and explained the history, evolution, and the mathematical and computational principles upon which these technologies are built.

Subsequently, I addressed an application in the project and conducted detailed research and feature engineering. For instance, I tackled a classification problem with a dataset describing urinary tract infection incidents.

The project provides a step-by-step guide through the process from data acquisition to preprocessing, model training, and evaluation. It presents the results of the classification model, providing insights into its performance and reliability.

In the conclusion sections, the report evaluates the findings of the project, discusses their implications and potential for future research. It also lays the groundwork for advancements in this field by identifying limitations and areas for improvement.

Overall, this project serves as both an educational resource and a case study in the practical application of Data Science. It aims to encourage readers to dive into these fascinating areas and explore their potentials. Whether you are an experienced researcher, a budding technologist, or simply an interested reader, this report offers valuable insights into the world of AI and ML. It emphasizes the transformative power of these technologies and their capacity to bring about progress and innovation.

Keywords: Artificial Intelligence (AI), Machine Learning (ML), Deep Learning, Artificial Neural Networks, SMOTA, Random Forest (RF), Project Management, Analytic Hierarchy Process (AHP), CRISP-DM-, Data Acquisition, Model Training,

MACHINE LEARNING DATA SCIENCE AND CRISP-DM APPLICATION OF URINE TRACY INFECTOIN CLASSIFIER

ÖZET

Bu mezuniyet projesinde, Makine Öğrenimi, Veri Bilimi ve Veri Biliminin prensiplerinden biri olan CRISP-DM alanlarına odaklandım. Bu alanları yöneten temel kavramları sağlam bir temel oluşturarak başladım ve bu teknolojilerin temelleri üzerine kurulu olan tarihini, evrimini ve matematiksel ve hesaplama ilkelerini açıkladım.

Daha sonra, projede bir uygulamaya odaklandım ve detaylı bir araştırma ve özellik mühendisliği yaptım. Örneğin, idrar yolu enfeksiyonu olaylarını tanımlayan bir veri seti ile bir sınıflandırma sorununu ele aldım.

Proje, veri toplamadan ön işleme, model eğitimine ve değerlendirmeye kadar olan süreci adım adım rehberliyor. Sınıflandırma modelinin sonuçlarını sunarak performansı ve güvenilirliğini sağlayan içgörüler sunar.

Sonuç bölümlerinde, projenin bulgularını değerlendirir, sonuçlarının olası etkilerini ve gelecek araştırmalar için potansiyelini tartışır. Ayrıca, sınırlamaları belirleyerek ve iyileştirme alanlarını tanımlayarak bu alandaki ilerlemelerin temellerini atar.

Genel olarak, bu proje hem eğitim kaynağı olarak hem de Veri Biliminin pratik uygulamasında bir vaka çalışması olarak hizmet eder. Bu ilgi çekici alanlara dalma ve potansiyellerini keşfetme konusunda okuyucuları teşvik etmeyi amaçlar. Deneyimli bir araştırmacı, yeni yetenekli bir teknolog veya sadece ilgili bir okuyucu olun, bu rapor yapay zeka ve makine öğrenimi dünyasına değerli içgörüler sunar. Bu teknolojilerin dönüştürücü gücünü ve ilerleme ve yeniliğin sağlanmasındaki kapasitelerini vurgular.

Anahtar kelimeler: Yapay Zeka (YZ), Makine Öğrenimi (MO), Derin Öğrenme, Yapay Sinir Ağları, SMOTA, Rastgele Orman (RF), Proje Yönetimi, Analitik Hiyerarşi Süreci (AHS), CRISP-DM, Veri Toplama, Model Eğitimi.

CONTENTS

PART ONE

1. ARTIFICIAL INTELLIGENCE

PART TWO

2 MACHINE LEARNING

THE THIRD PART

3 ARTIFICIAL NEURAL NETWORKS

THE FOURTH PART

4. DEEP LEARNING

THE FIVTH PART

THE SIXTH PART

THE SEVENTH PART

LIST OF FORMULA

INTRODUCTION

In today's fast-paced tech landscape, AI-driven systems are spearheading revolutionary changes in various industries. At the heart of these shifts are disciplines like artificial neural networks (ANNs) and deep learning (DL). In the era of Industry 4.0, where digitalization and automation are on the rise, ANNs and DL are driving transformative advancements in industrial processes. This thesis explores how integrating ANNs and DL with fundamental sciences impacts industrial applications.

The adoption of ANNs and DL in industry hinges on their strong connection with core sciences. Fields like computer science, mathematics, statistics, and neurobiology lay the groundwork for developing effective algorithms to tackle industrial challenges and manage intricate processes.

Industrial transformation not only boosts efficiency but also addresses critical areas such as cost reduction, energy optimization, and sustainability. ANNs and DL streamline production, making processes faster, more adaptable, and smarter, thus enhancing industry competitiveness and fostering innovation in business models and product designs.

This thesis doesn't just focus on the technical side of industrial tech transformation but also delves into its societal, economic, and ethical dimensions. In the new Industry 4.0 paradigm, the role of ANNs and DL reflects not just technical advancement but also societal shifts. Hence, this thesis aims to provide valuable insights into the evolution of industrial processes and contribute to future tech advancements with greater awareness.

1 ARTIFICIAL INTELLIGENCE

In this section, the concept of artificial intelligence, its application areas, and its history are covered.

1.1. **What is Artificial Intelligence ?**

Artificial intelligence (AI) is a field of study and technology dedicated to imbuing computer systems with capabilities akin to human intelligence. It involves the development of algorithms, software, and hardware designed to execute tasks typically associated with intelligence, including data analysis, learning, decision-making, and problem-solving. AI presents vast opportunities across various sectors, including automation, natural language processing, image recognition, autonomous vehicles, and healthcare. These systems aim to endow computers with human-like cognitive abilities, enabling them to swiftly analyze vast datasets, acquire learning skills, and tackle intricate tasks. Consequently, AI holds immense promise for revolutionizing numerous industries and enhancing human life in the foreseeable future.

1.2. **The History of Artificial Intelligence**

Although the term "artificial intelligence" was coined in 1956, its prominence has surged in recent years due to the proliferation of data, advancements in algorithms, and enhancements in computing power and storage capacity. Early AI research in the 1950s delved into areas like problem-solving and symbolic methods. By the 1960s, the U.S. Department of Defense became intrigued by these studies and initiated projects to emulate basic human reasoning using computers. For instance, the Defense Advanced Research Projects Agency (DARPA) undertook street mapping endeavors during the 1970s. As early as 2003, DARPA developed intelligent personal assistants, predating the widespread adoption of household names like Siri, Alexa, or Cortana.

This groundwork laid the foundation for the automation and logical reasoning capabilities evident in modern computers. It facilitated the development of decision support systems and decision search systems, which are engineered to complement and enhance human capabilities.

1.3.**How Does Artificial Intelligence Work?**

Artificial intelligence works by combining large-scale data with intelligent algorithms and iterative processing. Artificial intelligence with various functions works together with different methods and technologies.

- Machine Learning
- Artificial Neural Network

- Deep Learning Technology
- Cognitive Computing
- Advanced Algorithms
- Graphics Processing Units

2 MACHINE LEARNING

What is machine learning in this section. how to learn about them also includes regression models and classification criteria.

2.1. What is Machine Learning?

Machine Learning (ML) is a type of artificial intelligence (AI) that focuses on creating computer systems that learn from data, providing humanity with superhuman benefits. The broad technical spectrum covered by ML allows software applications to improve their performance over time. It is a scientific field where various algorithms and techniques are developed to enable computers to learn in a manner similar to humans.

Machine learning algorithms are trained to discover relationships and patterns in data. ML can be widely applied in various industries. Machine learning algorithms and computer vision, for instance, are critical components of self-driving vehicles, aiding them in navigating roads safely.

While machine learning is a powerful tool for solving problems, improving business operations, and automating tasks, it is also a complex and challenging technology that requires deep expertise and significant resources. Choosing the right algorithm for a task requires a strong grasp of mathematics and statistics. Training machine learning algorithms often involves large amounts of high-quality data to achieve accurate results. Understanding the outcomes, especially those generated by complex algorithms like deep learning neural networks modeled after the human brain, can be challenging. Additionally, running and tuning ML models can be costly.

2.2. How Machine Learning Works

Machine Learning is a transformative process that empowers machines to learn from specific inputs, holding immense significance across diverse domains. It's crucial to comprehend its capabilities and potential future applications. The journey of Machine Learning commences by supplying training data to a selected algorithm, whether known or unknown, aiding in honing the final model. The quality of training data profoundly impacts the algorithm's efficacy. To assess the algorithm's precision, new input data is introduced, and the resulting predictions are cross-checked against actual outcomes. Should disparities emerge, data scientists iteratively refine the algorithm until the desired accuracy is achieved. This iterative refinement process facilitates continuous enhancement, leading to increasingly

precise predictions over time. Machine Learning encompasses two primary approaches: supervised learning, where models learn from labeled input-output pairs for making predictions, and unsupervised learning, which autonomously uncovers hidden patterns or structures within input data.

2.3.**Supervised Learning and Unsupervised Learning**

Supervised learning is defined as the time when a model is trained on a "Labeled Data Set." Labeled data sets have both input and output parameters. In supervised learning, algorithms learn to map points between inputs and correct outputs. It has labeled data sets for both training and validation.

There are two main categories of supervised learning that are mentioned below

- Regression
- Classification

Unsupervised learning is a type of machine learning technique in which an algorithm discovers patterns and relationships using unlabeled data. The dependent variable is not available in the data set. The goal is to separate observation units based on similar features. The primary objective of unsupervised learning is often to discover hidden patterns, similarities, or clusters in the data; these can then be utilized for various purposes such as data exploration, visualization, dimensionality reduction, and more.

There are two main categories of unsupervised learning that are mentioned below

- Clustering
- Association

2.4 Regression

Regression, on the other hand, is concerned with predicting continuous target variables that represent numerical values. Regression algorithms are supervised algorithms used to find possible relationships between different variables, aiming to understand how independent variables influence the dependent one.

Here are some regression algorithms

Linear Regression

- Simple Linear Regression

- Multiple Linear Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression = rig + lasso

2.4.1 Simple Linear Regression

A simple linear regression model is a method used to predict the dependent variable, explained by a single explanatory (independent) variable, when there is a linear relationship between them. The fundamental goal is to find the linear function that expresses the relationship between the dependent and independent variables, and it is utilized to predict the dependent variable based on the independent variable.

$$\hat{y} = b_0 + b_1 x_i$$

xi: independent variable values.

$b0$ and $b1$: the multiple numbers we need to find from the data set.

$\hat{y}$: estimated values.

formula

2.4.2 Multiple Linear Regression

The primary objective is to find the linear function that expresses the relationship between the dependent and independent variables. The two aims of the research are: 1. To predict the values of the dependent variable through variables identified as influencing the dependent variable. 2. To determine which of the independent variables thought to affect the dependent variable has a greater impact and to attempt to describe the relationship between them.

$Yi=\beta 0+\beta 1Xi1+\beta 2Xi2+\cdots+\beta jXij+\epsilon i$ Formula.2 Multiple Linear Regression

Bi: It is used to control the effects of independent variables in the model.

2.4.3 Ridge regression

Ridge regression is a model fitting method used to analyze any data suffering from multicollinearity. This method performs L2 regularization. The goal is to find coefficients that minimize the sum of squared errors while applying a penalty to these coefficients. When multicollinearity issues arise, ordinary least squares become unbiased but have large variances, leading to predicted values deviating significantly from the true values.

$SSE=\Sigma(yi-\hat{yi})2$ formula 3 ride regression

$SSEL2=\Sigma(yi-\hat{yi})2+\lambda\Sigma\beta 2j$ formula 4 ride regression

λ : lambda parameter

$\Sigma\beta 2j$: Penalize constant

2.4.4 Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) Regression is an alternative biased prediction method to the Least Squares (OLS) approach. In Lasso, coefficients are shrunk towards zero. It can be used for issues related to multicollinearity and overfitting problems.

Lasso Regression is a regression technique where both variable selection and regularization occur simultaneously. Due to its efficiency and speed, it is widely applied in large datasets.

$SSEL1=\Sigma(yi-\hat{yi})2+\lambda\Sigma|\beta j|$ formula 5 lasso regresyon

$\Sigma|\beta j|$: Penalized constant
λ: ayar parametresi

2.4.5 ElasticNet Regression

Linear Elastic Net uses the Python `sklearn.linear_model.ElasticNet` class to predict regularized linear regression models for a dependent variable on one or more independent

variables. The goal is to find coefficients that minimize the sum of squared errors while applying a penalty to these coefficients. ElasticNet combines both L1 and L2 approaches.

$SSEnet$=Σ(yi−$\hat{y}ij$)2+λ1Σβ2j+λ2Σ|βj| formula 6 ElasticNet Regression

2.5 Nonlinear Regression

- K-The Nearest Neighbor KNN
- Support Vector Regression
- Artificial Neural Networks
- CART
- Random Forests
- Gradient Boosting Machines
- XGBoost
- LightGBM
- CatBoost

2.5.1 K-NN

The k-nearest neighbors (KNN) algorithm is a supervised machine learning model used for regression and classification problems, where predictions are made based on the similarities between observations. It is a non-parametric type of learning that can be utilized in both classification and regression problems. Predictions are made based on the similarities between observations.

$$\sqrt{\Sigma(x_i - y_i)^2}$$

Formula.7 K-NN

with euclidean or similar distance calculation, the distance of each we observation is calculated

2.5.2Support vector regression SVR

Support Vector Regression (SVR) is a type of machine learning algorithm used for regression analysis. It is a powerful and flexible modeling technique. The goal of SVR is to find a function that approximates the relationship between input variables and a continuous

target variable while minimizing the prediction error. The objective is to determine a line or curve that can capture the maximum points within a margin range with the least error.

$y=wx+\beta+\epsilon$ Formula 8 Support vector regression(SVR)

The minimization problem :

12||w||2+$C\Sigma\xi i+\xi$\ast Formula 9 Support vector regression(SVR)

Constraints :

$yi-(w$\ast $xi)-\beta$\le $\epsilon+\xi i$

$(w$\ast $xi)+\beta-yi$\le $\epsilon+\xi$\ast i $\xi i,\xi$\ast i\ge 0 i=1,2,3,4,...,m

2.5.3 CART- Classification and Regression Tree

The advanced methods of decision trees are among the highest-performing techniques in terms of prediction accuracy within the scope of artificial learning. The goal of CART (Classification and Regression Trees) is to transform complex structures within a dataset into simple decision structures.Heterogeneous datasets are divided into homogeneous subgroups based on a specified target variable. It is a tendency towards overfitting in the algorithm.

2.5.4 Random Forest

Random Forest is a machine learning algorithm that involves the collective learning and prediction attempts of multiple algorithms or multiple trees. Its ease of use and flexibility have fueled its adoption for both classification and regression problems.

2.5.5 XgBoost

XGBoost (eXtreme Gradient Boosting) is a highly optimized and high-performance version of the Gradient Boosting algorithm with various enhancements. It became a part of our lives with the article "XGBoost: A Scalable Tree Boosting System," published by Tianqi Chen and Carlos Guestrin in 2016. The algorithm's most important features include its ability to achieve high predictive power, mitigate overfitting, handle missing data efficiently, and perform these tasks rapidly.

It can be used with R, Python, Hadoop, Scala, and Julia.

2.5.6 Gradient Boosting Machines GBM

Gradient Boosting is a generalized version that can be easily adapted to both classification and regression problems, serving as an extension of AdaBoost. In Gradient Boosted Decision Trees, we combine many weak learners to find a strong learner. The weak learners here are individual decision trees.

All trees are connected sequentially, and each tree tries to minimize the error of the previous tree. Due to this sequential connection, the learning of boosting algorithms is often slow (can be controlled by the learning rate parameter set by the developer), but it is also highly accurate. In statistical learning, models that learn slowly often perform better.

2.5.7 Light GBM

LightGBM is an algorithm that operates based on histograms. It reduces computation costs by discretizing continuous variables into discrete bins. The training time of decision trees is directly proportional to the calculations performed and, therefore, the number of splits. Thanks to this method, both training time is reduced, and resource usage is decreased.

2.6 Classification Models

Classification is a supervised machine learning method where the model attempts to predict the correct label of a specific input data. It is a modeling type used when the dependent variable in the dataset consists of classes. In classification, the model is fully trained using training data and is then evaluated on test data before being used to make predictions on new, unseen data.

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Support Vector Classification (SVC)
- Artificial Neural Networks
- Classification and Regression Trees (CART)
- Random Forest
- Gradient Boosting Machines
- LightGBM

- CatBoost

2.6.1 Logistic Regression

Logistic regression is a data analysis technique that utilizes mathematics to find relationships between two data factors. It then uses this relationship to predict the value of one of these factors based on the other. The goal is to establish a linear model defining the relationship between dependent and independent variables for classification problems.

Adapted from multiple linear regression to classification problems, logistic regression can be considered as a version subject to slight differences. Predictions typically have a limited number of outcomes, such as yes or no.

$g(x)=ln\pi(x)1-\pi(x)=\beta0+\beta1x1+\beta2x2+...+\beta pxp$

Formula .10 Logistic Regression

2.6.2 CatBoost

CatBoost is an open-source machine learning algorithm based on Gradient Boosting, developed by the Yandex company. It is designed to enhance the performance of Gradient Boosting. Its distinguishing features from other algorithms include a high learning speed, the ability to work with both numerical and categorical as well as text data, GPU support, and visualization options.

2.6.3 Real Life Examples

- Vehicle recognition application from photo
- Linkdn, letgo and gmail reply apps
- Chatbots and personal assistants
- Facebook
- Netflix, amazon and e-commerce recommendation systems
- Spam Blocking
- Sentence completion and smile completion
- Uber
- Fraud prevention studies (Fraud Prevention
- Credit Application evaluation
- Boston dynamics

2.7 Model Training

It is desirable that the model trained in the development of Machine Learning models performs well on data that is new or not encountered before. In order to simulate new/non-encountered data, we divide our existing data into 2 training and test data sets Test-Train

20% of test orj data, 80% of Train orj data

3 ARTIFICIAL NEURAL NETWORKS

This section contains what are artificial neural networks, the relationship of the human brain and its areas of use

3.1What are Artificial Neural Networks?

Artificial Neural Networks (ANNs) represent simplified electronic models inspired by the intricate neural architecture of the human brain. Unlike traditional computing methods, which rely heavily on programmed instructions, ANNs emulate the brain's ability to learn from experiences, paving the way for a more intuitive approach to machine development solutions. By mimicking biological processes, this innovative computational paradigm offers a more elegant system that is less susceptible to performance degradation under heavy loads compared to conventional counterparts.

This bio-inspired approach is poised to revolutionize the computing industry, offering capabilities beyond the reach of current computer systems. While computers excel at tasks involving memory and complex mathematical operations, they often struggle with tasks requiring pattern recognition and extrapolation from past experiences to future actions, abilities inherent even in the simplest animal brains.

Recent advancements in biological research are shedding light on the mechanisms underlying natural cognition, revealing that brains encode information in intricate patterns. These patterns enable us to recognize faces from various angles, highlighting the complexity of information processing in biological systems. This paradigm shift in computation involves leveraging these patterns to tackle new problem domains, eschewing traditional programming in favor of training massively parallel networks. In this domain, terminology diverges from

traditional computing, emphasizing behaviors, responses, self-organization, learning, generalization, and even forgetting as key concepts.

Artificial Neural Networks consist of many cells, and these cells perform complex tasks by working simultaneously.

They have the ability to learn and they can learn with different learning algorithms.

They can produce results (information) for unprecedented outputs.

They can do pattern recognition and classification.

3.2 Analogy to the Brain

The human brain is an extraordinary structure known for its information processing, learning, and decision-making capabilities through complex neural networks and neurons. Artificial neural networks, on the other hand, are artificial intelligence models designed based on this biological foundation, incorporating algorithms in computer systems with the goal of learning and exhibiting intelligent behavior. Understanding the functioning of the human brain serves as an inspirational source when designing and developing artificial neural networks.Artificial neural networks are used to solve complex problems and model learning processes, drawing inspiration from biological neural networks. In this context, the complexity and flexibility of the human brain are considered a source of inspiration for making artificial neural networks more effective and adaptive. This relationship is continuously studied and developed to provide a deeper understanding in both artificial intelligence and neuroscience research.

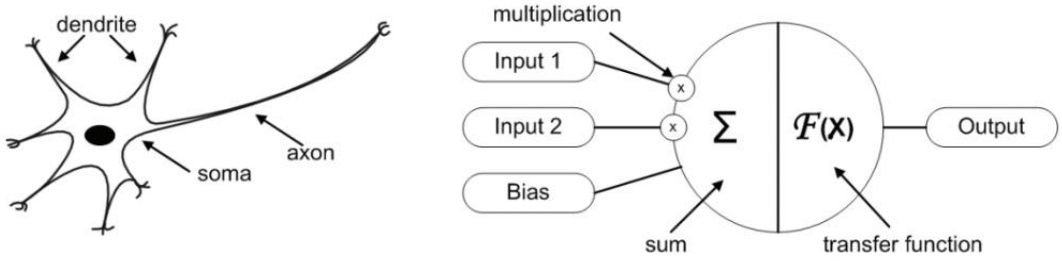


Figure 1

3.3 The History of Artificial Neural Networks

Research on neural networks continued throughout the 1970s and 1980s under the leadership of researchers such as Rumelhart, Hinton and Williams, and thus a method called the backpropagation algorithm was developed that allowed training multilayer neural networks.In the 90s, when interest in neural networks revived with the advent of more powerful computers and the availability of large amounts of data, new architectures such as convolutional neural networks and repetitive neural networks were developed and these

networks were used for a wide variety of tasks, including image recognition, speech recognition and natural language processing.

Figure 2 Artifical Neural Network

- **Inputs:** Inputs are data that come to neurons.
- **Weights:** The information received into the artificial nerve cell is transmitted to the nucleus by multiplying the inputs by the weight of the connections they came from before reaching the nucleus.
- **Combining Function:** The addition function is a function that calculates the net input of an artificial nerve cell by adding up the incoming inputs multiplied by weights and calculating the net input of that cell
- **Activation function:** It is a function that takes the weighted sum of all the inputs in the previous layer and then produces an output value (typically nonlinear) and passes it to the next layer.
- **Outputs:** The value that comes out of the activation function is the output value of the cell. Although each cell has more than one input, it has a single output. This output can be connected to the desired number of cells.

3.4 The Learning Mechanisms of a Neural Network

Neural networks, what needs to be generated in response to input.Depending on the difference between the actual value and the value Deciphered by the network, an error value is calculated and sent back through the system. The error value for each layer of the network is analyzed and used to set the threshold and weights of the next input. In this way, the error continues to be marginally less on each run as the network learns how to analyze the values.

Figure 3

Sigmoid function :The name of the sigmoid function is derived from the "sigma" (Σ) symbol, resembling the function's plotted shape in Cartesian space, particularly the letter "S." This function is particularly useful for the activation (firing) of neurons in an artificial neural network.

Hyperbolic tangent : Sigmoid function can be considered as a slightly modified version that also takes negative (-) values.

RELU: A rectified linear unit (RELU) is a nonlinear function. The ReLU function takes the value 0 for negative inputs, while x takes the value x for positive inputs.

4. DEEP LEARNING

4.1What is the Deep Learning ?

Deep learning an artificial intelligence (AI) method that teaches computers to process data inspired by the human brain. Deep learning models can recognize complex patterns in images, texts, sounds, and other data to produce accurate predictions and predictions.

4.2 Why Deep Learning is Important ?

Artificial intelligence (AI) tries to train computers to think and learn the way humans do.supports many artificial intelligence applications used in daily life products

- Digital assistants
- Voice-activated television remote controls
- Fraud detection
- Automatic face recognition

4.3 Deep learning Tools

- TensorFlow- 2015 google.
- - Keras - François Pollet
- - PyTorch - facebook 2018
- - Scikit-learn - gpu desteği yok
- - OpenCV
- - Caffe
- - Knet.jl

4.4 How Deep Learning is Changing the World

Deep learning is being applied in a wide range of fields from medical diagnosis to self-driving cars. In general, deep learning can be used anywhere that traditional machine learning algorithms have been used in the past. This includes tasks such as image classification, object detection, and speech recognition. However, deep learning is particularly well suited for problems that are difficult to solve using shallow neural networks. As a result, deep learning is being used in many cutting-edge applications such as natural language processing and recommender systems.

Here are a few examples:

- **Self-Driving Cars**: Deep learning is being used to develop self-driving cars. This technology has the potential to revolutionize transportation and make roads safer.

- **Robotics**: Deep learning is being used to develop robots that can interact with humans. This technology has the potential to transform manufacturing and other industries.

- **Medical Applications**: In medicine, deep learning is used to develop diagnostic tools and personalized treatments. This technology has the potential to improve healthcare and save lives.

- **Financial Services**: In finance, deep learning is used to develop fraud detection systems and automate trading. This technology has the potential to transform the financial industry.

- **Virtual Assistants**: Deep learning is being used to develop virtual assistants such as Amazon’s Alexa, Google’s Assistant, and Apple's Siri. These technologies have the potential to transform how we interact with computers.

- **Predictive Analytics**: Deep learning is being used to develop predictive analytics tools. These tools are being used in a wide range of industries to make better decisions about future events.

4.5 Deep Learning Layers

Figure 4 Deep Learning

Observations are transferred to the system in the input layer. The number of nodes in this layer is equal to the number of features that best represent the observation. Values in the input layer can be copied and sent to multiple layers. During the transfer phase, no operation is applied to the values. The hidden layer takes values from the input layer and applies transformation processes by multiplying these values with certain coefficients. There can be multiple nodes in this layer, and by applying specific threshold conditions to these nodes, values can be obtained equal to the number of outputs. Following the hidden layer, there is the output layer, where the system makes predictions based on the obtained value.

4.6 Things that Deep Learning has Achieved

- Almost human-level image classification.
- Almost human-level voice recognition
- Handwriting recognition at an almost Human level
- Autonomous vehicle driving close to human love
- The coming to life of digital assistants such as Google Now and Amazon Alexa
- Development of translation tools
- Synthetic data generation
- ChatGPT...

4.7 Convolutional Neural Network

Convolutional Neural Network (CNN) is a class of deep neural networks used for image recognition problems. It is widely employed by many major companies, including Google and Facebook. Regarding how CNN works, when images are provided as input, they need to be recognized by computers and converted into a processable format. Therefore, the first step is to convert the images into matrix format.

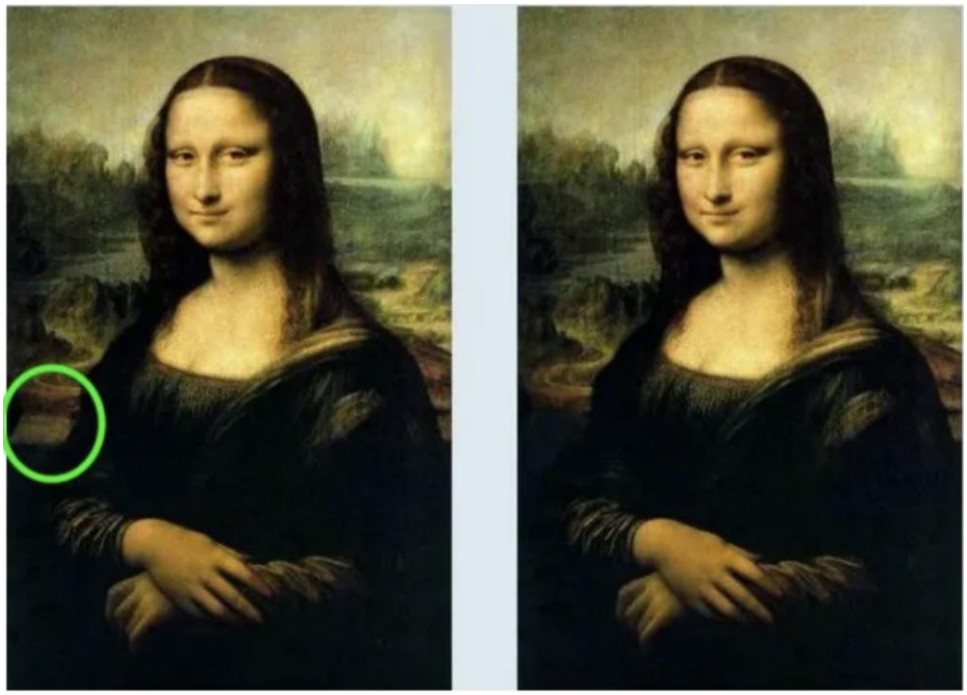


Figure 5 Convolutional Neural Network (CNN)

In Figure 5 , there are 2 paintings of Mona Lisa. At first glance, there is a subtle difference that may not be easily discernible. However, when we provide this structure to machines as a matrix, we enable them to capture even the smallest details. The underlying logic is the same in both the training and prediction stages.

4.8 Deep Learning vs Machine Learning

Deeplearning subset of machine learning. The main difference between machine learning and deep learning is how each algorithm learns and how Decently each type of algorithm uses data.

Deep learning eliminates some of the manual human intervention required by automating most of the feature extraction part of the process. It also earns the title of scalable machine learning, allowing the use of large datasets.

5. Data Science

Data science integrates mathematical and statistical principles, specialized programming, advanced analytics, artificial intelligence (AI), and machine learning techniques alongside domain expertise to reveal actionable insights concealed within an organization's data. These insights serve as guiding factors for decision-making and strategic planning.

The exponential growth in data sources has propelled data science to become one of the most rapidly expanding fields across all industries. Consequently, it's no wonder that the role of a data scientist has been hailed as the "sexiest job of the 21st century" by the Harvard Business Review. Organizations increasingly depend on these professionals to interpret data and offer actionable recommendations aimed at enhancing business outcomes.

The data science lifecycle encompasses diverse roles, tools, and processes, enabling analysts to extract actionable insights effectively. Typically, a data science project progresses through the following stages:

5.1 Data Ingestion

The data science journey kicks off with data ingestion, where diverse sources contribute raw structured and unstructured data through various collection methods. These methods encompass manual data entry, web scraping, and real-time streaming from systems and devices. Data sources range from structured datasets like customer information to unstructured data such as log files, multimedia content, IoT sensors, social media feeds, and beyond.

5.2 Data storage and data processing:

Given the varied formats and structures of data, enterprises must opt for diverse storage systems tailored to the data type they aim to capture. Data management teams play a pivotal

role in establishing standards for data storage and organization, streamlining processes for analytics, machine learning, and deep learning endeavors. This phase encompasses tasks such as data cleansing, deduplication, transformation, and amalgamation, achieved through ETL (extract, transform, load) operations or alternative data integration technologies. Such data preparation is indispensable for ensuring data quality prior to its storage in a data warehouse, data lake, or similar repository.

5.3 Data Analysis

In this phase, data scientists engage in exploratory data analysis to scrutinize biases, patterns, ranges, and distributions of data values. This analytical exploration serves as a catalyst for hypothesis formulation, particularly for a/b testing. Moreover, it aids analysts in assessing the data's suitability for incorporation into modeling endeavors for predictive analytics, machine learning, and/or deep learning applications. Depending on the accuracy of these models, organizations may increasingly rely on these insights to inform business decision-making, thereby fostering greater scalability.

5.4 Communication

Ultimately, findings are articulated through reports and various data visualizations, enhancing their accessibility and comprehensibility for business analysts and other stakeholders. Utilizing a data science programming language like R or Python provides built-in features for crafting visual representations. Alternatively, data scientists may opt for specialized visualization tools to effectively communicate the insights and their business implications.

NOTE : ***Data Science is the doing that mining beneficial/usefull information from the data***

6. CRISP-DM

CRISP-DM is the work flow of the data scientist. Among all the already tried workflows it is the best for the data science. There are 6 stages in this workflow, altough it may always stay

in a curculation.

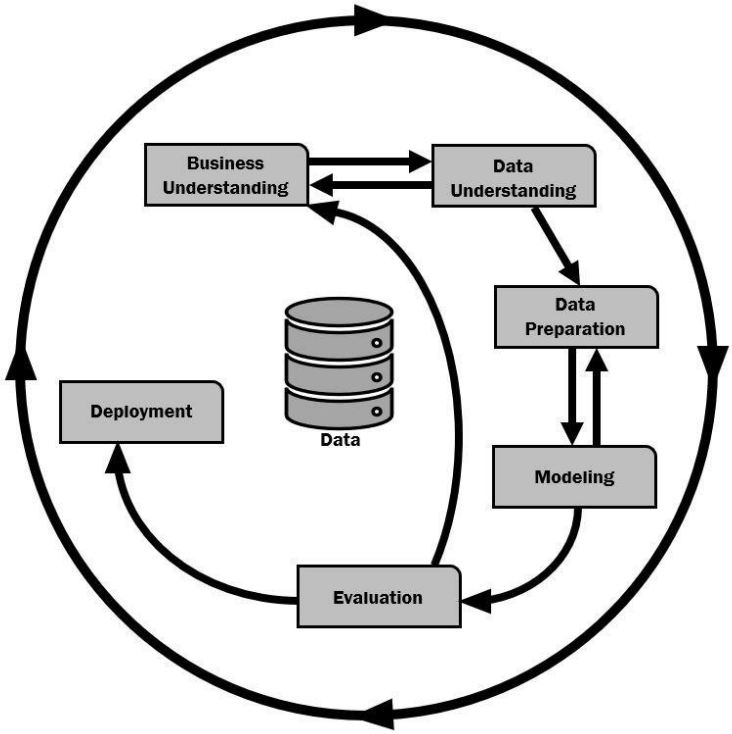


FIGURE 6 (DATA SCIENCE)

These stages are:

- Business Understanding
- Data Understanding
- Data Prepration
- Modelling
- Evulation
- Deployment

6.1 Business Understanding

Every event has it's own properties and it's own attributes. There's always some data sets to be understood but to be able to understand it you should have obtain information about what kind of job you are dealing with. So statments is "What are you people are doing here ?" if there's some one to ask, then ask it! it will save time for you and just listen the parts that important probably, the guy will be telling you all the details(try to ask people what they do they always love to tell whole story even most of the parts it is not the in story itself.)

6.2 Data Understanding

Now that you have familiarized yourself with the task at hand, you will be provided with data sets comprising rows and columns. Although the column headers may appear self-explanatory, it is advisable to exercise caution and not rely solely on their labels. It is prudent to delve deeper into the narrative behind the data. If you were not involved in the data mining process, it is recommended to seek clarification from the individuals responsible for its

collection. At this juncture, you may have gained some familiarity with the dataset; however, it is noteworthy that the data itself may not have yet divulged its essence. Exercise patience and engage in a thorough examination of the dataset from various perspectives. Subsequently, strategic actions must be taken. A data scientist typically conducts a meticulous review of the data's column types, quartiles, standard deviation, identification of outliers, and the presence of missing values (NAN). Upon obtaining a comprehensive understanding of the dataset, it is advisable to visually interpret it for enhanced comprehension. Generating graphical representations facilitates a clearer perception of the dataset's nuances. Subsequently, the identification of the dependent variable and the formulation of predictive hypotheses become feasible tasks as one delves deeper into the dataset.

6.3 Data Prepration

This section is where you do the followings; scale, split, normalization, outlier analyse, missing value analayse, feature engineering(maybe you like to generate some more data) and variable transformation. You are doing most of it due to ML algorithms only able to understand computer language(lineer-numeric values only).

6.4. Modelling

The modeling technique is selected, and depending on the software, the algorithm to be used is determined. Since some techniques, such as neural networks, have specific requirements regarding data structure, the data preparation stage may loop back here. Determining the parameters and tests of the models takes place at this stage.

```
Machine_Learning_Algorithms_Lineer_Regression = [Simple Linear Regression,
Multiple linear regression, Ridge Regressionon, Lasso Regresion,
 Elastic Net Regression(Lasso+Ridge)]

Machine_Learning_Algorithms_notLineer_Regression = [K-nearest neighbor,
 Support Vector Regression, Artifical Neural Network, Random Forest,
 CART(Classification and Regression Trees), GBM(Gradient Boosting Machines),
XGBoost(GBM derivative),LightGBM,CatBoost]

Machine_Learning_Algorithms_Classification=[Lojistik Regression(it type the lineer
 K-nearest neighbor,Support Vector Classifier, CART, Random Forest, GBM, LightGBM,
```

FIGURE 7 (DATA SCIENCE)

6.5 Evaluation

Once the model is created, we need to check how well its performance is. We should measure the performance of our model and compare it with our performance targets to see how it is doing. Generally, in data science projects, the data scientist divides the data into training and

test data. In this step, using test data, it evaluates the success and performance of the model by verifying how true the model created in the previous stage is to reality.

At this stage, let's say you evaluated the performance of your model and the success of your model is 60%, but if your intended success rate is 80%, you can go back to the previous stage and remodel your data set.

6.6 Deployment

After the data of the model evaluated in the application phase, which is the last stage of the methodology, is subjected to different real-time tests and evaluations such as A / B testing, the final information obtained is put live after the approval process.

7 EXAMPLE

You can reach the study from HERE. It is my notebook that I published on kaggle.

7.2 Business Understanding

Classify the patient, if he have the Urinary Tract Infection(UTI) or NOT

In the modern era the UTI is rare disaes but it wasn't like that before. Still there are people suffering from this sickness and we will be using ML-AI technology to make a classification. The Dataset we will be studying on is from 'This dataset was collected from a local clinic in Northern Mindanao, Philippines' and published on Kaggle to conturibute to community.

Data Understanding

- Age (The age of the patient) Note: Some patients are months old, so the age of these patients are preprocessed by dividing it by a hundred) e.g, 8 MONTHS OLD, 8/100 = 0.08
- Gender (The gender of the patient) Note: Either male or female
- Color (urine color)
- Transparency (urine transparency)
- Glucose (glucose is a type of sugar, and its presence in the urine can be an important indicator of certain health conditions)

- Protein (the presence of protein in the urine is one of the parameters examined to assess kidney function and detect potential)
- pH (the pH level measures the acidity or alkalinity of urine)
- Specific Gravity (urine specific gravity is a measure of the concentration of particles in urine compared to water)
- WBC (White Blood Cells) Note: Also known as leukocytes, white blood cells are a crucial part of the immune system
- RBC (Red Blood Cells) Note: RBC are responsible for carrying oxygen throughout the body
- Epithelial Cells (epithelial cells are cells that line the surfaces and cavities of the body, including the urinary tract)
- Mucous Threads (mucous threads are strands of mucus that can be present in urine)
- Amorphous Urates (amorphous urates are non-crystalline formations in the urine that consist of uric acid)
- Bacteria (presence of bacteria in the urine)
- Diagnosis (UTI Diagnosis) Note: Either NEGATIVE or POSITIVE

The first look on the dataset.

Shape and dtypes of our data :

```
dtype: object
```

Missing Value Check

```
**************************************************
```

Handiling the missing value(checked the row without a script and filled according to my own opinion) and changing our target variable in to numerics-labels.

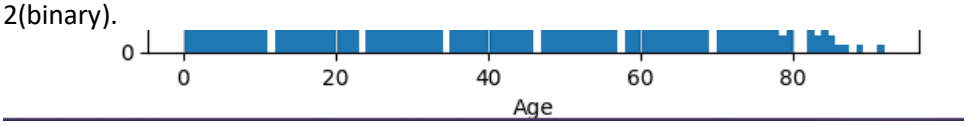
DATA PREPRATİON

In this section;

- First I did Grabbed the collumns according to dtype; categorical, numeric. But If a numeric column does not have more than 10 unique values I took it in to cat_cols. (it is my own script you may check on [here](#) I imported it as yucaib)

- The cat_but_car list means the columns that has high cardinality (no meaning, nothing to mine on this column). Therefore we will be dropping them.
-

- It seems there is 4 numeric but actually pH has 6 unique, Specific Gravity 6 , Diagnosis has 2(binary).
-



Out[24]:

False

- Categoric Collumns summary :
-
- ##############################3
- ##############################3
- ##############################3
- ##############################3
-
-
  - Glucose: NoneSeen + Seen
  - Transparency : Clear and Hazy
  - Color : Light Yellow + Yellow + Dark Yellow(concate rest of them in darker yellow)
  - gender : it will remain still

categoric
e column,
eature

-
-

```
df.loc[(df['Specific Gravity'] == 1.030), 'NEW_GRAVITY']= 'up_normal'
```

-

```
df.loc[(df['Bacteria'] == 'RARE'),'NEW_BACTERİA']= 'Rare'
```

se intervals

- Out[32]:

```
array(['NoneSeen', 'Seen'], dtype=object)
```

-

```
e' if x in ['RARE', 'FEW'] else 'Frequent')
```

-

```
df['NEW_TRANSPERENCY'] = df['Transparency'].apply(lambda x: 'Clear' if x=='CLEAR' else 'notClear')
```

-

AGE X Variabels

- ```
  df.loc[(40<df['Age'])& (df['Gender'] =='MALE')  & (df['pH']<6),'NEW_AGE_GENDER_PH']='SeniormalePH_acidic'
  ```
- ```
  aline
  ```

- ```
  df.drop('WBC', inplace=True,axis =1)
  df.drop('RBC',inplace=True , axis =1)
  ```
- ```
  cat_cols: 12
  cat_but_car: 1
  num_cols: 1
  num_but_cat: 1
  ```
  of them to

LABEL ENCODİNG

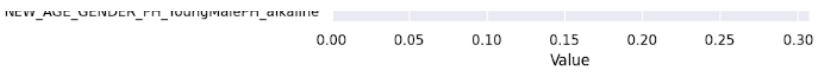
- ```
  df.shape
  ```
  ```
  ut[71]:
  (1436, 14)
  ```
- ```
  In [75]: df.shape
  Out[75]: (1436, 39)
  ```
- ```
  'NEW_AGE_GENDER_PH_YoungMalePH_acidic']
  ```
  at has really Therefore

```
In [81]: df.drop(useless_cols, axis= 1 , inplace= True)
```

STANDARTİZATİON

- ```
  df[num_cols] = scaler.fit_transform(df[num_cols])
  ```

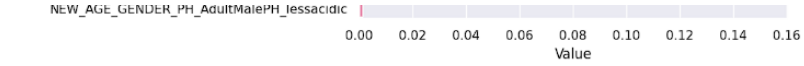
MODELLİNG

- ```
  X_train, X_test, y_train, y_test = train_test_split(X,y,test_size = 0.30, random_state= 17)
  ```
- Modelli data.
  ```
  In [91]: rf_model = RandomForestClassifier(random_state=46).fit(X_train,y_train)
  y_pred = rf_model.predict(X_test)
  ```
- 

  ```

- Precision: 0.0
  F1: 0.0
  Auc: 0.4674

As far: We did; filled missing value, outlier check, label encode, one hot encoding, scaling and modelling. Also , dropped cardinal columns. What we find out is Age is the most important feature among the others. Also to be clear this dataset kinda small and the ratio of Diagnosis (TARGET VARİABLE) is %94 Negative to %6 Positive, So we don't wait that the model to not be in a biased attitude. Let's see how to solve this Problem.

- We will be using SMOTE (Synthetic Minority Over-sampling Technique) due to this problem;
- with smote we can balance the dataset, the target value, DIAGNOSİS is %94 Negative and %6 Positive.
- it will attempts to correct the imbalance in the dataset by augmenting minority class examples with synthetic examples.
- Accuracy: 0.5062034739454094
  Recall: 0.8571428571428571
- Out[184]:
  50.0
- 

- Now it seems the bacteria existence and yellow color is more important to find out if you have UTI or not.

WHAT IF WE DON'T PREPARE ANYTHING

- Now we will be doing only the essantials to fit model and then check the importances.
- Precision: 0.0
- Value

- As a data scientist as my own personal view :
- we find out, importances of the features for;
- feature extraction done data set (Age: 0.30, NEW_TRANSPARENCY: 0.06)
- feature extraction and extended data set(Age: 0.15, New_BACTERİA: 0.14)
- only essentials done dataset, (label encode, scale, drop, fillna) (Age: 0.16, Bacteria_modaret : 0.075)
- SO ; extention and feature engineering is important due to make more general predictions.
- the accuracy scores are so high, probobly these scores are missleading, I think percision score is more accurate.
- In this study, the feature engineering didn't made so much difference.
- As my view ; This dataset is not well Normal Disturubited also it kinda need more explanations
- The UTI, urine enfection, is something rare in society. So it is normal that dataset is not meeting the needs.

## CONCLUSION

In this report, we stated the following: Common Machine Learning Algorithms, Data Science and Data Scientist's Work Flow CRISP-DM and an Data Science example, classification application.

The project provides a step-by-step guide for the entire process, from data acquisition to preprocessing, model training, and evaluation. It presents the results of the classification model, providing insights into its performance and reliability.

## REFERENCES

1. https://arc.net/l/quote/lhzwspej
2. https://www.tensorflow.org/resources/learn-ml?hl=tr
3. -https://www.ibm.com/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks/
4. https://medium.com/operations-management-t%C3%BCrkiye/crisp-dm-metodolojisi-nedir-80158d91622d
5. https://www.tutorialspoint.com/machine_learning/machine_learning_artificial_neural_networks.htm
6. 
7. https://www.forbes.com/sites/kalevleetaru/2019/01/15/why-machine-learning-needs-semantics-not-just-statistics/?sh=1d6d1dc577b5
8. https://www.veribilimiokulu.com/yapay-sinir-agiartificial-neural-network-nedir/
9. https://www.ibm.com/topics/data-science
10. https://www.kaggle.com/code/yucaib/cool-funcs
11. Miuul.com
12. https://medium.com/@y.canfee/a-little-step-to-data-science-3a6916278b9b