UNIVERSITÀ DI PISA

**DIPARTIMENTO DI INGEGNERIA DELL'ENERGIA DEI SISTEMI,
DEL TERRITORIO E DELLE COSTRUZIONI**

**RELAZIONE PER IL CONSEGUIMENTO
DELLA LAUREA MAGISTRALE IN INGEGNERIA GESTIONALE**

Knowledge Data Discovery and Predictive Analytics: the case of TIM Digital Support

SINTESI

RELATORI

Prof. Antonella Martini
Dipartimento di Ingegneria dell'Energia dei Sistemi, del Territorio e delle Costruzioni

Dott. Antonio Forti
Team Leader, Elis Consulting&Labs

CANDIDATO

Mariagiulia Garcea
mgiuliagarcea@libero.it

Sessione di Laurea del 02/05/2019
Anno accademico 2017/2018
Consultazione NON consentita

Knowledge Data Discovery and Predictive Analytics: the case of TIM Digital Support
Mariagiulia Garcea

Sommario

Il lavoro di tesi proposto è frutto di un'esperienza lavorativa della durata di cinque mesi, svoltasi all'interno del programma Junior Consulting, presso ELIS *Consulting&Labs* a Roma.

Lo stage aveva come obiettivo lo sviluppo di un progetto per il cliente Telecom Italia, riguardante la sezione *TIM TV & Intrattenimento-Digital Support*. Il progetto si è concentrato sull'analisi dei dati relativi al *TIM BOX* device, lo strumento che consente agli utilizzatori di accedere ad una vasta gamma di contenuti multimediali offerti dalla piattaforma di intrattenimento digitale fornita da TIM. L'ambito del progetto ha riguardato quattro applicazioni della piattaforma: TIMvision, TIMgames, TIMmusic e Sensi Unici. In particolare, il bisogno del cliente era quello di migliorare il monitoraggio dei dati, e creare un modello predittivo. L'analisi condotta aveva due obiettivi principali: il primo riguardava il miglioramento della Dashboard e la mappatura dell'architettura dati che la supportano, mentre il secondo la creazione di un modello predittivo, attraverso tecniche di *Machine Learning*. Al termine del progetto, grazie ai risultati raggiunti, il cliente beneficerà di un'efficiente individuazione dei disservizi della piattaforma e delle loro cause, così come di un aumento della qualità percepita dal cliente finale.

Abstract

This thesis is the result of a five-month educational program with *ELIS Consulting&Labs*, within the *Junior Consulting* project. The stage concerned the build-up of the *TIM Digital Support* project in the section of the so-called *TIM TV & Intrattenimento-Digital Support*. This project focused on data analysis of the *TIM BOX*, a device that allows customers to access the whole multimedia content of the digital entertainment platform provided by *TIM*. Moreover, the analysis regards the apps: *TIMvision*, *TIMgames*, *TIMmusic* and *Sensi Unici*. Particularly, the need of the client was to enhance the data monitoring and create a predictive model. Hence, the analysis had two main aims: the first was the improvement of the Dashboard and the mapping of its data architecture, while the second was the creation of a predictive model using Machine Learning algorithms. In the end, from the analysis' findings, TIM will take advantage of an immediate identification of the inefficiencies and their causes, and a consequent improvement of the quality perceived by customers.

1. Context and goals

This thesis is the result of a five-month educational program carried out from October to April 2019 in *ELIS Consulting&Labs* in Rome, within the *Junior Consulting* project.

The project was assigned by *Telecom Italia Mobile*, a telco company owned by *Telecom Italia S.p.A*, which is the most successful Italian brand in the world in its field. In the last year, TIM enhanced its economic value of about 33%, which led to an improvement of its rank in the *Brand Finance Global 500 (2018)*[1], the well-known classification of the main brands of the entire world. *TIM* counts more than 100 million of customers, and offers not only fixed and mobile telecommunication, but also an on-cloud platform where it is possible to find internet and digital contents for entertainment-video, music, gaming, high technological platforms and IT solutions. The connection with this platform is carried out by a decoder, called *TIM BOX*, connected to the internet and a TV set. The whole system takes advantages from the high quality of *TIM* connection for the access to the platform. Furthermore, to enlarge the offer at disposal of its customers, *TIM* became partner of the major providers of multimedia content such as *Netflix and DAZN*. Nowadays, in Italy there are about 600.000 active decoders that generates a huge amount of data. The scope of this thesis is to analyse data from the TIM apps: *TIMvision*, *TIMgames*, *TIMmusic* and *Sensi Unici*. Indeed, Telecom Italia has a dashboard that refers to decoders data, which monitors many performance indicators related to the devices. However, it is not highly performing since it only shows descriptive charts or Key Performance Indicators (KPI). TIM aims to improve its data analytics, and thus, to - "make efficient and well-informed decisions"-. The goal of the project is to give insights on data from decoders, through the process of Knowledge Discovery Data (KDD)[2]. Hence, the focus of this project is the use of predictive analytics to forecast future events. The application of predictive analytics can change the mind-set of the company, but it also allows to solve problems before their occurrence.

[1] https://www.telecomitalia.com/tit/it/about-us.html

[2] Han Data, J. and Kamber, (2011). Data Mining: Concepts and Techniques

2.Phases, Methodologies and Results

Table 1 below shows the macro-phases of this work, with the main activities, methodologies, results and also the reference thesis paragraphs.

Macro-Phase	Main activity	Methodology	Results	§
1. Literature review and conceptual framework	Research of articles	Search on GOOGLE SCHOLAR and SCOPUS of the paper related to the main topic of the thesis	Definition of the main topic that will set the conceptual framework and choice of 52 articles	1.1
	Selection of articles	Reading the articles and choosing them coherently to the thesis' main topics	Selection of NN paper and definition of the conceptual framework	1.2/1.3
2. Dashboard and Data mapping	Dashboard AS-IS Analysis	Studying and mapping of the Dashboard AS-IS and identification of possible improvements	Mapping of Dashboard AS-IS	3.1
	Identification of business needs	Interviewing the business for the validation of our changing proposal and to clearly identify their needs	Definition of business needs	3.2
	Dashboard TO-BE definition	Insertion of business needs in the new Dashboard and change the KPI or chart defined in the interview	Mapping of Dashboard TO-BE	3.3
	Data catalogue	Studying of the technical documentation and mapping both data flow and data structure from the decoder to the Dashboard	Mapping of data flow and data structure	3.4
3. Predictive analysis	Data selection and Data pre-processing	Definition of the right data source, data migration, and data exploration in order to identify the main variable for the analysis.	Choice of the data source, recognition of the main variables for the ABT	4.1
	Creation of Analytical Base Table (ABT)	Choice of the input and output variables.	Final ABT	4.2
	Identification of correlation	Data exploration of the ABT and identification of possible correlation	Correlation between variables	4.3
	Data prediction	Use of machine learning algorithm to predict the target variable	Final prediction	4.4

Table 1- Phase, Activity, Methodology Results and thesis paragraph

2.1 Literature review and conceptual framework

2.1.1 Research and Selection of papers

This activity aims to build a table of paper in order to develop a conceptual framework used to enhance the knowledge and organize an overview of the main topics used throughout this thesis. The *Table 2* below shows the activities and the results of this first phase:

<table>
<tr><th>Activity</th><th>Methodology and Result</th></tr>
<tr><td>Research of Paper</td><td>In the table below are shown the criteria for the paper research:
<table>
<tr><td>Search word</td><td>Predictive Analytics - Advanced Analytics-Machine learning and Predictive Analytics – Big Data – KDD (Knowledge Data Discovery)</td></tr>
<tr><td>Fields</td><td>Every fields</td></tr>
<tr><td>Location words</td><td>Anywhere in the papers</td></tr>
<tr><td>Document type</td><td>Paper, books</td></tr>
<tr><td>Published date</td><td>From 2010 to 2019</td></tr>
<tr><td>Language</td><td>English</td></tr>
</table>
The output of this research activity was the collection of 52 papers closely related to Predictive Analytics and Knowledge Data Discovery.</td></tr>
<tr><td>Selection of Paper</td><td>Every collected paper has been read carefully and the selection of the relevant articles has been carried out coherently to the main topics. The final papers are 16 and could be divided into 4 categories: Big Data, Data analytics, Predictive Analytics, Machine learning connected to Predictive Analytics.</td></tr>
</table>

Table 2- Results of the Research and Selection of Papers activity

Results

The literature review takes into account all the above-cited categories. Indeed, it started from the concept of Big Data, defining the methodologies to analyse them with Data Analytics. Moreover, it deeply examined Predictive Analytics and, in the end, it explained the relationship between Machine Learning and the latter. In *Table 3* below are shown the results:

Categories	Result
Big Data	As said by Doug Laney in the early 2000, Big Data are a wide number of data both structured and non-structured used to be analysed in order to find patterns, trends or extract value. Gartner, in 2011[3], identify five aspects of the Big Data: Volume (data size), Variety (different forms of data sources), velocity (speed of change), veracity (uncertainty of data) and value (business value). With the gathering of Big Data and Analytics comes out Big Data Analytics in order to extract value from data, find hidden information, identify trend, and thus, make well-informed decision.
Data analytics	The aim of Data analysis, is analyse the relation of the single dependent variable with the construction of model that can give insight on data. All these analyses are necessary to extract value for companies. It was developed a process named Knowledge Discovery Data[4] in order to define a standardized structure in this analysis. This process is composed by five steps: Data Selection, Data Pre-Processing, Data Transformation, Data Mining and Data Evaluation. The analytics could be Descriptive, described by Dr.Michael Wu (2018) as: "the simplest class of analytics, one that allows you to condense big data into smaller, more useful nuggets of information" or Advanced, defined by Gartner[5] as "the analysis of all kinds of data using sophisticated quantitative methods to produce insights that traditional approaches to business intelligence are unlikely to discover". An organization can apply one or more analytical techniques to solve their problems, and the sophistication of the method used is an indicator of "the level of analytics maturity of the organization"[6]. It is important to understand at which stage the analytics maturity of an organization is, in order to fulfill all the gap that will let the company extract all the values from its data[7].
Predictive Analytics	Predictive analytics are models of the Advanced Analytics that allow to predict future events based on past and present data[3]. Nowadays, predictive analytics are widely used in many kinds of company, empowering many advantages: improve data integrity, minimize total number of metrics reviewed, keep management focused on the big picture, improve forecasting and projections, review past present and future perspective in a single metric and find correlations between the variables. In order to take the maximum advantages of predictive analytics it is necessary to collect data in a correct way, gathering structured and non-structured data in a table named ABT, which is the input of the analysis.

[3] Constante, F. and Silva, F. and Herrera, B, and Pereira, A. (2018). Big Data Analytics in IOT: Challenges, Open Research

[4] Hammad, A. and AbouRizk, S. (2014). Knowledge Discovery in Data: A Case Study. Journal of Computer and Communication

[5] Herschel, G. and Linden A. (2015). Magic Quadrant for Advanced Analytics Platforms. Gartner

[6] Reitter, N and List, B. (2013). Analytics Maturity Model. OR/MS Today, Vol. 40. No. 5

[7] Madyon, T. (2017) Four types of big data analytics and examples of their use. Ingram Micro Advisor

Machine learning and Predictive Analytics	Machine learning[8] is most transformative technology of the 21st century. It uses algorithms, that are a sequence of instruction used for converting an input in an output the process that reproduces human intelligence by learning from data or past experience, in order to predict, find patterns on data in analysis. The Supervised learning[9] is the main used methods, 'it *forms their predictions via a learned mapping f(x), which produces an output y for each input x*'. The algorithm of this mapping f are decision trees, decision forests, logistic regression, etc. Machine learning is rapidly increasing guided by practical application, therefore its most driver of expansion is the environment in which it works. Indeed, Machine learning system are improving, *'taking the form of many software that run on a large scale and distributed computing* (such as Spark) *that provide a range of algorithm and services to data analysis'*.

Table 3 – Results of the Literature review and Conceptual Framework activity

2.2 Dashboard and Data mapping

One of the goals of the project is the improving of the Dashboard and the mapping of its data architecture, as to clarify the data structure used as an input for the further analysis.

2.2.1 Dashboard AS-IS, Identification of business needs and Dashboard TO-BE definition

This activity concerned a first mapping of the Dashboard AS-IS, in order to understand the contents of all the views, drown up in a document sent to the business. The goal of this activity was the understanding of all KPIs and graphs shown in the Dashboard, and also the identification of possible changes to apply. Furthermore, thanks to an interview with the business, the Dashboard TO BE was defined, since relevant doubts and other business needs were definitely clarified. Indeed, all the KPIs and graphs were classified into three different categories, based on the action to undertake on them: *Maintain*, *Modify* or *Delete*. Subsequently to the abovementioned activity, the Dashboard TO BE was planned, mapped and presented to the business.

Result

In *Table 4* below are shown the results of the Mapping of Dashboard:

Activity	Result
Dashboard AS-IS Analysis	*ECL89MOM_TIM_Digital_Support_Mapping_Dashboard_AS-IS_v1.0*
Identification of business needs	*ECL89MOM_TIM_Digital_Support_BusinessNeeds_v1.0*
Dashboard TO-BE definition	*ECL89MOM_TIM_Digital_Support_Presentazione_DashboardTOBE_v3.0*

Table 4- Results of the Mapping of the Dashboard activity

2.2.2 Data catalogue

Data Catalogue regards the analysis of technical documents provided by the client, in order to understand the architecture of Dashboard's supportive data. In *Figure 1* is represented the mapping of the structure of data flow from the Decoder to the Dashboard.

[8] El Naqa, I. and Murphy J. (2014). What is machine learning?. Machine Learning in Radiation Oncology, pp 3-11
[9] Mitchell, M. and Jordan I. (2015). Machine learning: Trends, perspectives, and prospects. Science

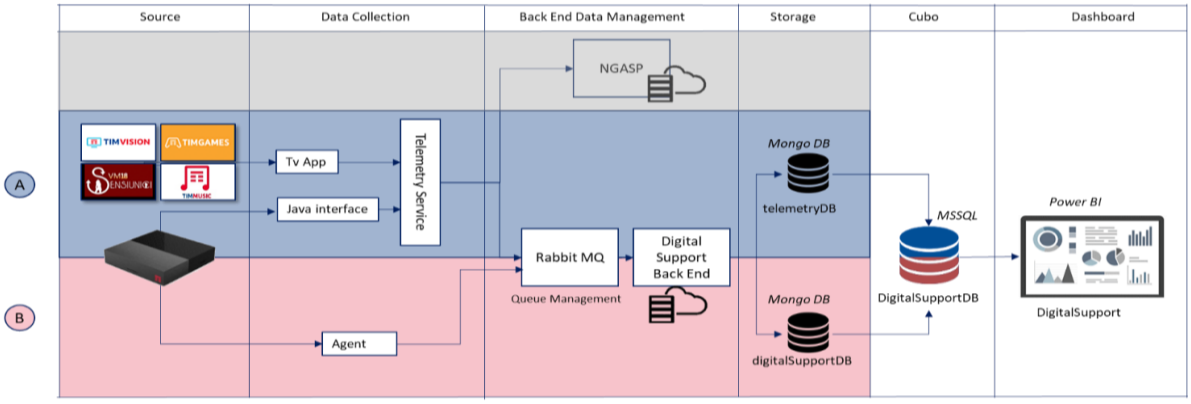

Figure 1 - Data flow and architecture

As it can be seen, the image is divided horizontally into three parts (the grey part is out of scope): the blue one (A), represents data flow of Telemetry, while the lower one (B, in red), shows data flow of Agent. Telemetry data are related to near-real time events, generated by the system or by the end-user, which describe customer's journey. On the other hand, data contained in Agent, catch vital parameters of the decoder (STB) and, particularly, they are generated every 60 seconds. While, from a vertical point of view, data flow is divided into five parts. The first is the source, that is the device, from whom starts the Agent and Telemetry flow. Secondly, there is the part of Data collection in which the process of measuring and gathering information is undertaken, in order to guarantee the integrity of data. The central part is Back End Data Management in which there are servers on cloud that store data. Finally, the fourth part is composed by two different MongoDBs, one for Telemetry and the other for Agent, which get together in a Cubo database (fifth part), the direct data source of the Dashboard. After the study of data structure, all data in Cubo, MongoDB, Agent and Telemetry were mapped. This action is vital for the identification of which data can be used for building up the views in the Dashboard and, principally, to understand where they came from, in order to help the business to develop the new Dashboard. On the other hand, the main goal of the mapping was the comprehension of data structure for the next phase of Predictive Analytics. Indeed, the data mapping was necessary to identify the variables requested by the analysis.

Results

In *Table 5* below are shown the results of the Mapping of Dashboard:

Activity		Result
Data catalogue	Mapping of Agent	*ECL8-9MOM_TIM_Digital support_Agentv1.0*
	Mapping of Telemetry	*ECL8-9MOM_TIM_Digital support_Telemetry_Trapv1.0*
	Mapping of Cubo	*ECL8-9MOM_TIM_Digital support_CUBOv1.0*
	Mapping of MongoDB	*ECL ECL8-9MOM_TIM_Digital support_MONGO_DBv1.0*

Table 5 – Result of the Data Catalogue activity

2.3 Predictive Analytics

The aim of this part is using predictive analytics in order to extract knowledge and helpful insights for the business. The main activities were the ones shown in *Figure 2* below:

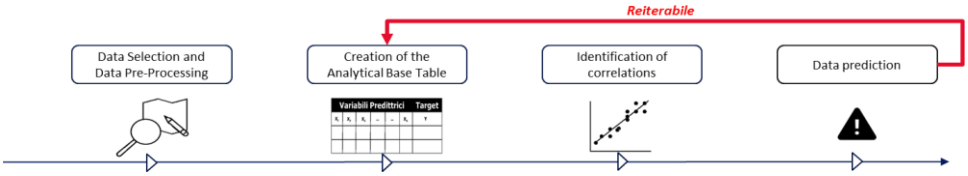

Figure 2 – Main activity of Predictive Analytics phase

It is important to underline that, in the case of a not adequate result, all the process will be reiterated until a new and coherent outcome will be found. In this part there will be shown all the activities of Knowledge Data Discovery's process.

2.3.1 Data selection and Data pre-processing

The aim of this activity is the definition of the variable to fill in the ABT. The process of **Data Selection** regards the choice of the right data source from which it is possible to extract data. Mainly, there were four points of data storage: Agent, Telemetry, MongoDB and Cubo. The choice of the data source is MongoDB since it is the most complete and coherent with the analysis that we wanted to conduct. Indeed, Cubo has aggregated data and its format doesn't allow any prediction. Furthermore, Agent and Telemetry are the roughest data in the flow and for that reason it is very difficult to analyse them.

After having chosen MongoDB as the data source for the analysis, the process of **Data pre-processing** started. The latter is shown in the *Figure 3* below.

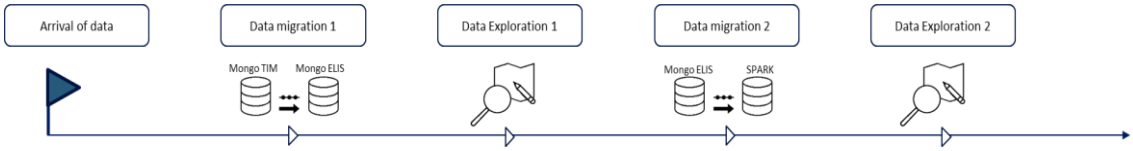

Figure 3 – Process of Data Pre-Processing

Indeed, a first Data Migration process from TIM MongoDB to Elis one started and lasted two days, due to the large amount of data (around 550 giga). The first process of Data Migration allowed us to do a first Data Exploration but, because of the high computational complexity, it wasn't successful. In order to solve the high temporal complexity, it was decided to do a second Data Migration from MongoDB to Spark technology. This infrastructure solved the speed problem that concerns MongoDB, giving advantages in terms of speed related to the optimization of the distributed computing and meanwhile in terms of efficiency, because it is

possible to use Machine learning algorithms directly in Spark. In the second Data Migration it was necessary to select the right data to transfer, because not all were necessary for the analysis. MongoDB is structured into two different databases called DigitalSupportDB and TelemetryDB. The former is the one that stores data from Agent flow, while the latter data from Telemetry flows. Both databases are divided into several collections that contains different kind of data. Hence, five collections were selected: *Device, Device_traps* and *Device_events* from TelemetryDB, while *Informations*, *Dimensions* and *STBs* for DigitalSupportDB. The choice of the collections lied in the kind of data inside that, indeed they contained the minimum number of non-aggregated data useful for the analysis.

After that, due to the large amount of data, a second **Data exploration** started directly in Spark, in order to reduce data dimensions in terms of selected variables and time-interval.

Results

In *Table 6* below are shown the results of this activity in terms of the choice of the Data source, of the collection to use, of the time interval and of the variables.

<table>
<tr><th>Activity</th><th>Result</th></tr>
<tr><td>Data Selection and Data Pre-Processing</td><td>In the table below is shown the Dataset built after these first data analysis on Spark
<table>
<tr><td rowspan="2">Choice of Data Source</td><td colspan="6">MongoDB</td></tr>
<tr><td colspan="3">DigitalSupportDB</td><td colspan="3">TelemetryDB</td></tr>
<tr><td>Choice of the collections</td><td>STBs</td><td>Information</td><td>Dimensions</td><td>Device</td><td>Device_Traps</td><td>Device_events</td></tr>
<tr><td>Choice of the time interval</td><td colspan="6">04-02-2019/17-02-2019</td></tr>
</table>
This Dataset is the input for the construction of the final ABT</td></tr>
</table>

Table 6 – Result of the Data Selection and Pre-Processing activities

2.3.2 Creation of ABT

The Analytical Base Table (ABT) is used to build a predictive model. Data inside this chart are necessary for the identification of the future behaviour of a specific variable, called “Target Variable”. The ABT is mainly composed by three parts: key variables (which uniquely identify every row of the table), predictive variables (which are correlated with the target) and, finally, the target (the aim variable of the prediction). Furthermore, in order to fill in the ABT, it was necessary to solve some relevant issues, as it can be seen in *Table 7* below.

Problems	Solutions
Huge amount of data- around 700 million of rows for each day	Reducing the amount of data analyzed (in order to analyze that directly in Pyton), choosing randomly one day among the time interval previously selected. The day selected was the **13th of February**
The merging of DigitalSupportDB and TelemetryDB was challenging due to: -The difference of transmission frequency between Telemetry and Agent	Creation of two Dataset, one with the union of the Telemetry collections and the other of the Agent one’s. After that, it was undertaken the merging of these two Dataset with the two variables of: -Timestamp (deleting the second values) -SerialNumber

-The difference of primary keys inside the collections of the two Dataset	The result is a Dataset with **585420** entries, that contains all the variables represented in all the collections selected in both Agent and Telemety
The identification of many valueless variables, which have all NaN values	Deleting of NaN values that are not useful for the prediction

Table 7 – Problems and Solution of Data exploration

Moreover, after a first **Data exploration**, it was underlined that the Dataset contained different typologies of values, both numerical and categorical. Each variable was analyzed with graphs in order to understand their patterns and characteristics. Therefore, together with the client, it was decided to set the inefficiencies as the target of the analysis. *Nine inefficiencies* that have an impact on the end-users were found. These are indicated in *Table 8* below. The percentage of inefficiencies on the total of events is **6%** (32917 inefficiencies).

Inefficiencies		
TtFP	BUFFER_UNDERFLOW	TIMGAMES_ERROR
END_BUFFERING	SERVER_ERROR	LONG_BUFFERING
CONTENT_CRIPTED_ERROR	CONTENT_PLAYBACK_ERROR	NETWORK_ERROR

Table 8 - Inefficiencies

These values are in the column 'eventName' in which there are all the events that can occur both in the normal use of the decoder and in case of error. An activity of **Data Transformation** started to convert the values contained in 'eventName' from 'Object' to categorical, creating a new column named 'Inefficiency' with Boolean value: 1 if there is an inefficiency and 0 if there is not. The creation of this new column allowed to classify the predictive model as '*Categorical*' and '*Supervised*'. Moreover, it was necessary to choose the predictive variables as the input for the target one. Started from the TIM's technical documentation it was decided to include in the first analysis all the numerical variables that do not contained NaN values.

Results

In *Table 9* below are shown the results of Creation of ABT activity:

<table>
<tr><th>Activity</th><th>Result</th></tr>
<tr><td>Creation of ABT</td><td>The table below shows the ABT divided into its three parts with its 19 variables. The ABT contained values of the 13th of February 2019.
<table>
<tr><th>Key variables</th><th>Predictive variables</th><th>Target variable</th></tr>
<tr><td>1.SerialNumber
2.Timestamp</td><td>3.model_name
4.bitRatekbps
5.ConnectionType
6.wifiFrequency
7.ramA.Mb
8.storageA.EXTMb
9.wifiSignal
10.storageA.INTMb
11.SNR
12.cpuUsePercent
13.wifiStatus
14.wifiLinkSpeed
15. storage.EXTMb
16.upTimeSec
17.temperatureCpu
18.eventName</td><td>19.Inefficiency</td></tr>
</table></td></tr>
</table>

Table 9 – Results of the Creation of ABT activity

2.3.3 Identification of correlation

The aim of this activity is the recognition of any possible pattern inside the ABT, with the activity of **Data Mining**. The first analysis carried out was the correlation heat map of all the dataset. As it can be seen in *Figure 4* below, the values didn't have a high correlation index.

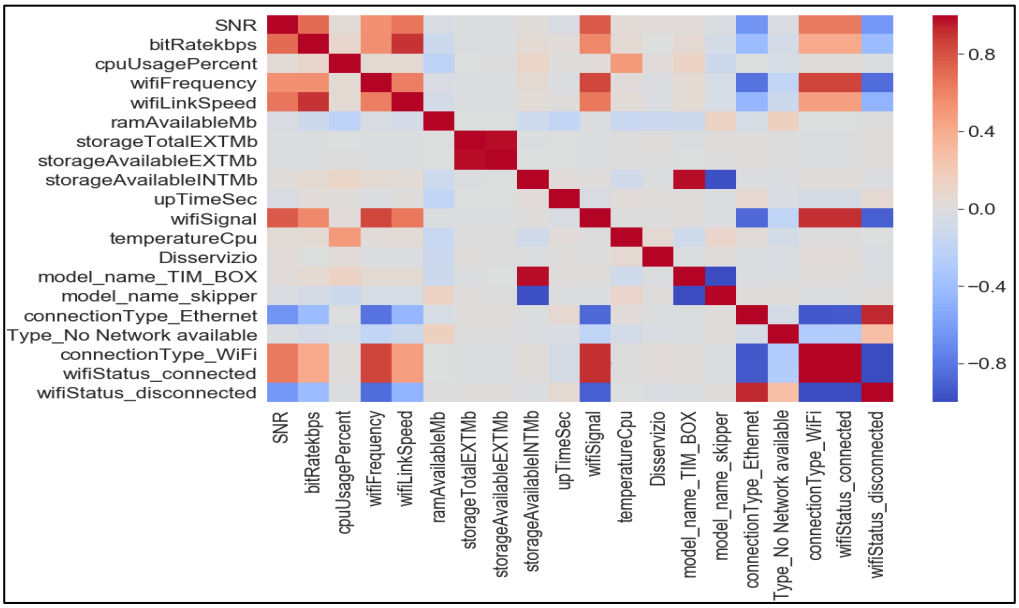

Figure 4 – Correlation Heat Map

Indeed, there are some variable correlated each other's, such as 'SNR' or ''wifiLinkSpeed' with 'bitRatekbps' or 'temperatureCpu' with 'cpuUsagePercentage', but mainly the correlation was very low. Moreover, focusing on the direct correlation with the target column 'Inefficiency', there weren't any strong correlation. Thus, is to point out that 'ramAivaibleMb' had a negative correlation value. Furthermore, it was undertaken another Data Exploration activity in order to represent and analyse the pattern of the more correlated variable.

Results

In *Table 10* below are shown the results of the Identification of correlation activity:

<table>
<tr><th>Activity</th><th>Result</th></tr>
<tr><td>Identification of correlation</td><td>In the table below are represented the correlation values of the more relevant numerical features with the target 'Inefficiency', calculated with the Pearson index.
<table>
<tr><th>Feature</th><th>Correlation</th><th>Feature</th><th>Correlation</th></tr>
<tr><td>SNR</td><td>0.024291</td><td>bitRatekbps</td><td>-0.000404</td></tr>
<tr><td>cpuUsagePercent</td><td>0.029920</td><td>ramAvailableMb</td><td>-0.130919</td></tr>
<tr><td>storageTotalEXTMb</td><td>0.011098</td><td>wifiLinkSpeed</td><td>0.006567</td></tr>
<tr><td>storageAvailableINTMb</td><td>-0.009710</td><td>upTimeSec</td><td>0.014506</td></tr>
<tr><td>wifiFrequency</td><td>-0.02060</td><td>temperatureCpu</td><td>0.054879</td></tr>
<tr><td>storageAvailableINTMb</td><td>-0.009710</td><td>connectionType_Ethernet</td><td>-0.014856</td></tr>
<tr><td>connectionType_No</td><td>-0.029762</td><td>connectionType_WiFi</td><td>0.024568</td></tr>
<tr><td>wifiStatus_connected</td><td>0.024613</td><td>wifiStatus_disconnected</td><td>-0.024613</td></tr>
<tr><td>wifiSignal</td><td>0.003260</td><td>temperatureCpu</td><td>0.054879</td></tr>
</table></td></tr>
</table>

Table 10 – Results of the identification of correlation activity

2.3.5 Data prediction

This final activity focused on the building up of a predictive model with machine learning algorithms. It was decided to use three algorithms: KNN, Random Forest and Decision Tree. The choice was undertaken according to the classification of the predictive model in analysis as '*Categorical*' (Binary) and '*Supervised*'. Therefore, the request of TIM was maximizing both the 'Precision' and the 'Recall' indicators. For these reasons, in the **Evaluation** activity of the model, it will be chosen the machine learning algorithms that maximize these values. For each algorithm it was followed the same process shown in *Table 11* below:

Phases of Machine Learning preparation	Description
Building of a balanced Dataset	Due to the huge differences between the event of Inefficiency (6%) and the one of 'No inefficiency' (94%), two balanced ABTs were built, one with a 1 to 1 ratio and the other with a 1 to 2 ratio (32500 inefficiencies and 65000 no in.)
Final variables used for the prediction	'SNR', bitRatekbps', 'cpuUsagePercent', 'wifiFrequency','wifiLinkSpeed', 'ramAvailableMb', 'storageTotalEXTMb', 'storageAvailableEXTMb', 'storageAvailableINTMb', 'upTimeSec', 'temperatureCpu', 'model_name_TIM_BOX' ,'model_name_skipper', 'wifiStatus_connected', 'wifiStatus_disconnected'
Application of KNN, Random Forest and Decision tree	Training-size=70%, Test-size=30%, random_state=30

Table 11 – Process of machine learning preparation

All the three Machine learning algorithms were set following the above-mentioned activities. In the end, the combination of the **Decision Tree** and the **Random Forest** algorithms with a **1 to 2 ratio** evidenced better results than KNN. In *Figure 5* it is shown the process of the maximization of 'Precision' and 'Recall' with the Features importance with the Decision Tree.

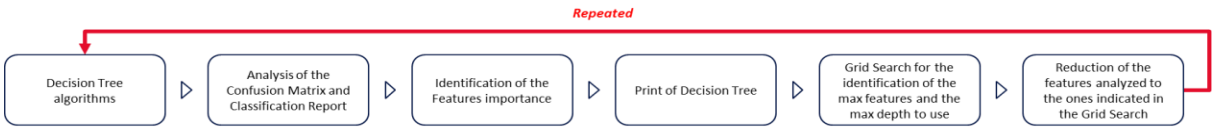

Figure 5 – Process of the feature reduction with Decision Tree

After the definition of the features, these were used in the Random Forest algorithm.

Results

As it is shown in *Figure 6*, the final prediction had a *Precision* of 0.66 for the inefficiencies and *Recall* of 0.87 on the normal events. These values were calculated with a **Random Forest** with 9 features, chosen from the

	precision	recall	f1-score	support
0	0.78	0.87	0.82	19418
1	0.66	0.50	0.57	9832
micro avg	0.75	0.75	0.75	29250
macro avg	0.72	0.69	0.69	29250
weighted avg	0.74	0.75	0.74	29250

Figure 6 – Results of the Random Forest with 1 to 2 ratio

identification of the Features importance: '*ramAvaibleMb', bitRateKbps', 'SNR', 'cpuUsagePercentage', wifiFrequency', 'wifiLinkSpeed','storageAvailableInt', 'upTimeSec', 'temperatureCpu'.*

3. Conclusion and future development

This predictive model allows to prevent the 66% of inefficiencies that influence the end-users. Obviously, it is not as strong as it would be, indeed, due to the huge amount of data, the algorithm refers only to one day since it was impossible for our technology to analyse more than these quantities of data. TIM's future aim is to replicate and validate our model for its complete Data Structure. The benefits of this thesis are:

- The definition of a predictive model that can classify inefficiencies;
- The identification of which are the main variables that influence the inefficiencies in order to solve them.

Appendix A - Process of Features Reduction

The aim of this appendix is to explode the last activity carried out with the **Decision Tree** and the **Random Forest** algorithms. As it was pointed out before, the process for the definition of the features that maximize the Precision and Recall is shown in the figure below.

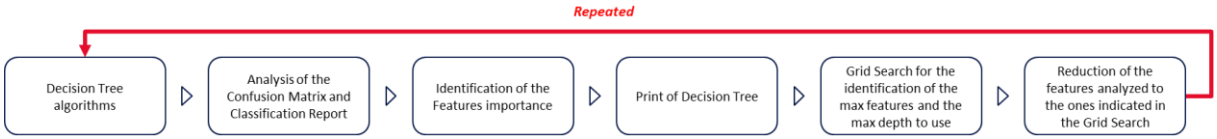

Figure A.1 – Process of the features reduction with Decision Tree

Each phase was undertaken twice until the values of the machine learning were considered satisfying. At first, all the variables in the ABT were included in the analysis, even the features that had low correlation with the target. After the evaluation with the Confusion Matrix and the Classification Report were analysed the Features importance and the result was the one shown in the figure below:

	importance
ramAvailableMb	0.579941
bitRatekbps	0.145011
temperatureCpu	0.083753
wifiFrequency	0.074148
SNR	0.052828
cpuUsagePercent	0.043264
storageAvailableINTMb	0.009840
wifiLinkSpeed	0.006167
upTimeSec	0.005048
storageTotalEXTMb	0.000000
storageAvailableEXTMb	0.000000
model_name_TIM_BOX	0.000000
model_name_skipper	0.000000
wifiStatus_connected	0.000000
wifiStatus_disconnected	0.000000

Figure A.2 – Features importance

After the definition of the feature importance, and the printing of the Decision Tree, the next activity was the Grid Search in order to find the max depth and the max feature that maximize the Recall and the Precision.
The result of this activity was: 11 max features and 6 of max depth.
Indeed, the initial ABT was reduced to the first eleven features highlighted in the 'Features Importance', dropping from the analysis 'wifiStatus_disconnected', 'wifistatus_connected', *'storageTotalEXTMb', 'storageAvailableEXTMb',* 'model_name_skipper' and 'model_name_TIM_BOX'.
The feature selected with the Decision Tree algorithm, were used in the Random Forest and the final result is shown in the table below:

	precision	recall	f1-score	support
0	0.78	0.87	0.82	19418
1	0.66	0.50	0.57	9832
micro avg	0.75	0.75	0.75	29250
macro avg	0.72	0.69	0.69	29250
weighted avg	0.74	0.75	0.74	29250

Figure A.3 – Results of the Random Forest

Appendix B – Choice of the algorithm

All the three Machine learning algorithms were set following the above-mentioned activities. In the Table below are shown the results in terms of Precision and Recall for both the balanced Dataset, for all the algorithms.

Balancing of the Dataset	Algorithms	Precision on 0	Precision on 1	Recall on 0	Recall on 1
1:1	Decision Tree	0,80	0,60	0,42	0,90
	Random Forest	0,72	0,71	0,70	0,73
	KNN	0,68	0,69	0,71	0,66
1:2	Decision Tree	0,72	0,68	0,93	0,28
	Random Forest	0,78	0,66	0,87	0,51
	KNN	0,76	0,68	0,89	0,45

Table C.1 – Choice of the algorithm

As it is evidenced in the table, the Random Forest with a 1 to 2 ratio has the best value both in Precision and Recall.

Appendix C – La mia esperienza

Vorrei descrivere la mia esperienza in questa azidenda con due parole: Competenze e Collaborazione. Avevo sentito parlare di Elis come un'azienda attenta allo sviluppo personale e professionale e per questo motivo ho scelto di intraprendere questo percorso, ma mai mi sarei aspettata di imparare così tanto, grazie soprattutto alla collaborazione tra colleghi.
Lo spirito aziendale che si respira è familiare e consente alle persone di lavorare in sinergia con gli altri, riuscendo in questo modo ad imparare sempre di più da tutte le persone che ci circondano. Ho imparato ad ascoltare ed assimilare più informazioni possibili, a comunicare in maniera efficare, ma soprattutto ho imparato a conoscermi meglio sia a livello personale che lavorativo. Sono inoltre felice di aver acquisito molte conoscenze tecniche grazie al progetto in TIM che ho avuto la fortuna di svolgere, interfacciandomi con il mondo dell'Analisi Dati che non conoscevo, ma che ho scoperto essere molto affascinante e stimolante.
Riconosco che non sempre è stato facile e che spesso il livello tecnico richiesto era molto alto e ho dovuto studiare molto per restare al passo con le attività progettuali. Nonostante ciò, ritengo di essere stata molto fortunata ad aver partecipato a questo progetto e ringrazio le persone che hanno reso possibile tutto ciò.

Fase	Metodologie acquisite	Strumenti acquisiti
Gestione del progetto	Utilizzo di metodologia *Agile* per lo svolgimento del progetto	Trello
Literature review and conceptual framework	Ricerca e studio di articoli scientifici per lo stato dell'arte di Big Data, Predictive Analytics, KDD and Machine learning	
Dashboard and Data mapping	Processo di analisi della Dashboard	PowerBI
	Processo di Mappatura dei Dati	
Predictive Analytics	Estrazione dati e preparazione dell'ABT	MongoDB e Spark
	Applicazione del processo generale di Knowledge Data Discovery per la predizione	Python (librerie Pandas, seaborn e MatplotLib)

Tabella C.1 – Competenze acquisite e strumenti utilizzati nel percorso di tesi