



ADVANCED ANALYTICS

Stefano Roselli [s.roselli@cineca.it](mailto:s.roselli@ Cineca.it)



CINECA

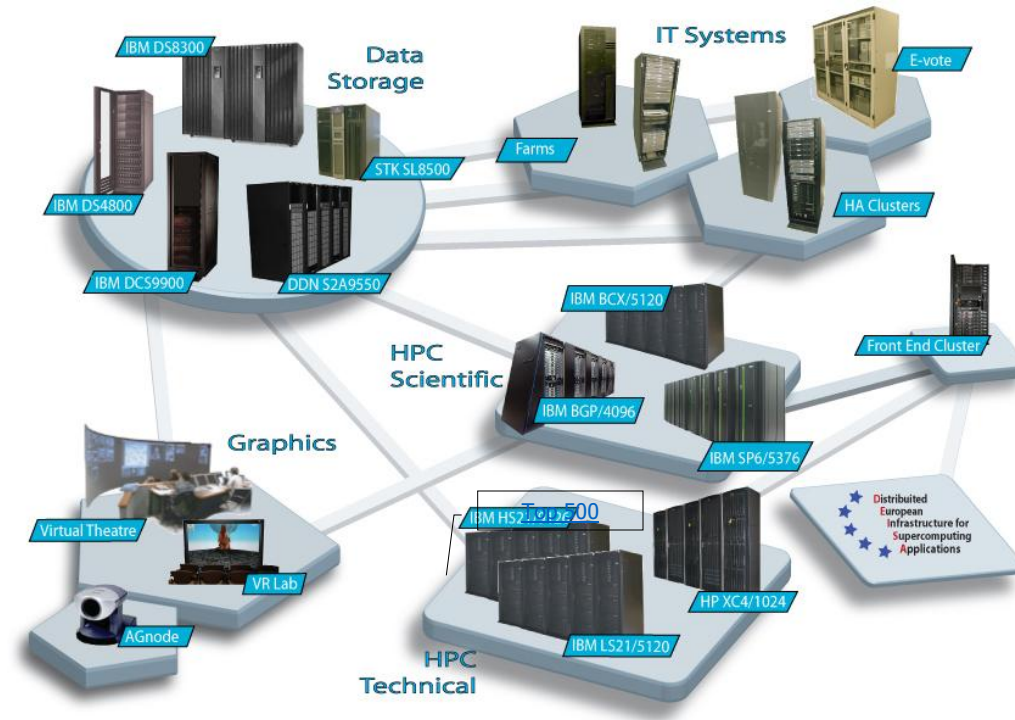
Il Cineca è un Consorzio Interuniversitario senza scopo di lucro al servizio del sistema accademico nazionale istituito nel 1969.

La missione è “promuovere l’utilizzo dei più avanzati sistemi di **elaborazione dell’informazione** a favore della ricerca scientifica e tecnologica, pubblica e privata”, e al trasferimento tecnologico alle imprese e alla Pubblica Amministrazione.

Fanno parte del Consorzio:

- MIUR
- 70 università
- 4 Enti di Ricerca

Circa 900 dipendenti con sedi a Bologna, Milano e Roma



Il Laboratorio Big Data & Analytics

Il Laboratorio di Big Data & Analytics è una iniziativa di CINECA, nel campo della High Performance Analytics per promuovere la sua diffusione e aiutare i decisori aziendali e i professionisti ICT a comprendere le strategie, le potenzialità e le tecnologie dei Big Data e delle tecniche di Data Mining.

PIATTAFORME SOFTWARE:

- IBM Big Insights
- Hortonworks Data Platform

ARCHITETTURE:

- Data Streaming Analysis
- **Large Scale Machine Learning**

TECNOLOGIE:

- Hadoop (HDFS, MapReduce), YARN
- Spark SQL, Hive e HBase
- Storm, Spark Streaming
- Kafka & MQTT
- Spark R e Distributed R
- Librerie: Spark MLLIB, H2O

INFRASTRUTTURA:

HPC IBM NeXtScale server appositamente progettata per i casi di calcolo “data-intensive”:

- 70 nodi IBM NeXtScale con interconnessione a 56 GB/sec
- Intel Ivy Bridge 20 core per nodo, 1480 core in totale
- 128 GB RAM per nodo
- 40 TB SSD locale al nodo, 16 PB di storage in linea



Advanced Analytics

Gli **Advanced Analytics** sono applicazioni informatiche che usano metodi matematici e statistici su sistemi computazionali altamente scalabili per **estrarre valore dai dati**, come trovare schemi ricorrenti (patterns), raggruppamenti (clusters) e relazioni nei dati (rules) per predire futuri comportamenti o scenari, fornendo anche raccomandazioni.

Analisi Predittiva nel CBM

La manutenzione predittiva nella Condition-based maintenance (CBM), si focalizza sull'individuare la probabilità di guasti **prima** che avvengano.

L'applicazione di **machine learning** per predire situazioni di probabili guasti si basa sul *costruire un modello usando dati storici e **addestrarlo** con casi noti, per essere in grado di identificare o classificare situazioni di potenziali guasti e non.* Il modello dovrà essere **validato** usando dati reali di test prima di applicarlo. La validazione fornisce una indicazione (matrice di confusione) sull'attendibilità del modello individuando i veri positivi, i veri negativi, falsi positivi e i falsi negativi.



Machine Learning

Supervised learning

Il sistema apprende da un insieme di esperienze già classificate

Unsupervised learning

Non si hanno casi da cui il sistema può apprendere

Algoritmi Predittivi

Categorical Target Variable:

- Decision Tree
- Random Forest
- Neural Networks
- Support Vector Machines
- K-Nearest Neighbor
- Logistic Regression
- Gradient Boosting Machine

Continuous Target Variable:

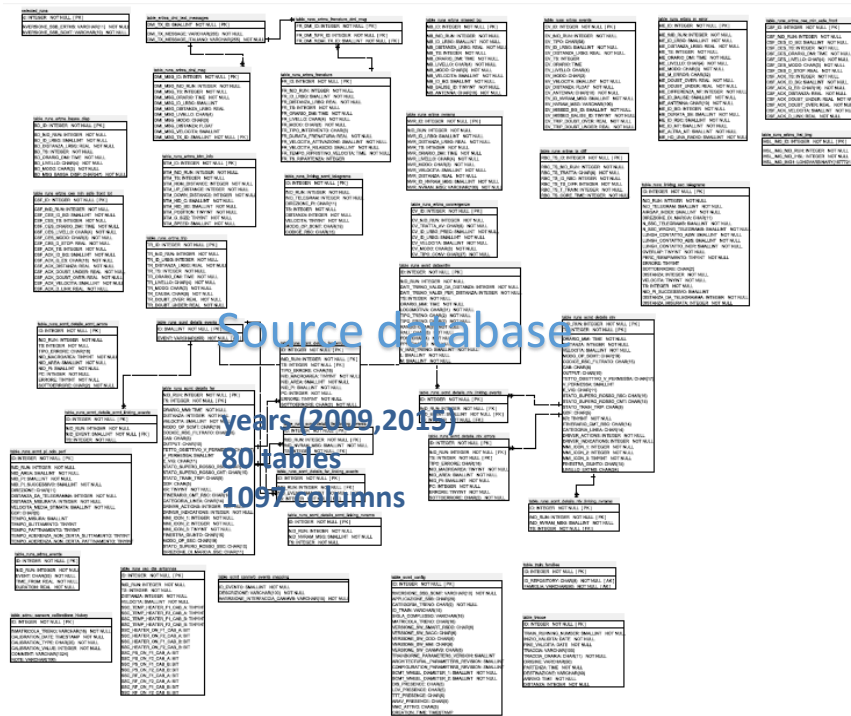
- Linear Regression
- Generalized Linear Model

Algoritmi Descrittivi

- Clustering (K-Means)
- Hidden Markov Chains
- Principal Component Analysis (PCA)
- Self-Organizing Maps (SOM)
- Modelli Causali

Caso Alstom: Data Preparation

- ✓ Individuate le tabelle e le colonne del database utili all'analisi
- ✓ Creato il Master Data con le variabili necessarie all'analisi
- ✓ Creato il Dataset per gli algoritmi di Machine Learning



Source database

years (2009, 2015)

80 tables

1097 columns

[illegible]

Year	Age	Gender	Marital Status	Occupation	Income	Education	Health	Life Expectancy
1997.0	28	M	Married	Healthcare	45000	High School	Good	78.5
1976.0	46	M	Married	Healthcare	45000	High School	Good	78.5
1992.0	33	F	Married	Healthcare	45000	High School	Good	78.5
1989.0	36	M	Married	Healthcare	45000	High School	Good	78.5
1974.0	51	M	Married	Healthcare	45000	High School	Good	78.5

Year	17	18	19	20	21
2014	3780.0	3780.0	3780.0	3780.0	3780.0
2015	3780.0	3780.0	3780.0	3780.0	3780.0
2016	3780.0	3780.0	3780.0	3780.0	3780.0

1	table
---	-------

87 columns

25.000.000 rows

dataset	year	month	week	day	hour	minute	second	microsecond	nanosecond	timezone_offset	dataset	year	month	week	day	hour	minute	second	microsecond	nanosecond	timezone_offset
151575	2015	01	04	01	00	00	00	000000000	000000000	00:00:00	151576	2015	01	04	01	00	01	00	000000000	000000000	00:00:01
151576	2015	01	04	01	00	01	00	000000000	000000000	00:00:01	151577	2015	01	04	01	00	02	00	000000000	000000000	00:00:02
151577	2015	01	04	01	00	02	00	000000000	000000000	00:00:02	151578	2015	01	04	01	00	03	00	000000000	000000000	00:00:03
151578	2015	01	04	01	00	03	00	000000000	000000000	00:00:03	151579	2015	01	04	01	00	04	00	000000000	000000000	00:00:04
151579	2015	01	04	01	00	04	00	000000000	000000000	00:00:04	151580	2015	01	04	01	00	05	00	000000000	000000000	00:00:05
151580	2015	01	04	01	00	05	00	000000000	000000000	00:00:05	151581	2015	01	04	01	00	06	00	000000000	000000000	00:00:06
151581	2015	01	04	01	00	06	00	000000000	000000000	00:00:06	151582	2015	01	04	01	00	07	00	000000000	000000000	00:00:07
151582	2015	01	04	01	00	07	00	000000000	000000000	00:00:07	151583	2015	01	04	01	00	08	00	000000000	000000000	00:00:08
151583	2015	01	04	01	00	08	00	000000000	000000000	00:00:08	151584	2015	01	04	01	00	09	00	000000000	000000000	00:00:09
151584	2015	01	04	01	00	09	00	000000000	000000000	00:00:09	151585	2015	01	04	01	00	10	00	000000000	000000000	00:00:10
151585	2015	01	04	01	00	10	00	000000000	000000000	00:00:10	151586	2015	01	04	01	00	11	00	000000000	000000000	00:00:11
151586	2015	01	04	01	00	11	00	000000000	000000000	00:00:11	151587	2015	01	04	01	00	12	00	000000000	000000000	00:00:12
151587	2015	01	04	01	00	12	00	000000000	000000000	00:00:12	151588	2015	01	04	01	00	13	00	000000000	000000000	00:00:13
151588	2015	01	04	01	00	13	00	000000000	000000000	00:00:13	151589	2015	01	04	01	00	14	00	000000000	000000000	00:00:14
151589	2015	01	04	01	00	14	00	000000000	000000000	00:00:14	151590	2015	01	04	01	00	15	00	000000000	000000000	00:00:15
151590	2015	01	04	01	00	15	00	000000000	000000000	00:00:15	151591	2015	01	04	01	00	16	00	000000000	000000000	00:00:16
151591	2015	01	04	01	00	16	00	000000000	000000000	00:00:16	151592	2015	01	04	01	00	17	00	000000000	000000000	00:00:17
151592	2015	01	04	01	00	17	00	000000000	000000000	00:00:17	151593	2015	01	04	01	00	18	00	000000000	000000000	00:00:18
151593	2015	01	04	01	00	18	00	000000000	000000000	00:00:18	151594	2015	01	04	01	00	19	00	000000000	000000000	00:00:19
151594	2015	01	04	01	00	19	00	000000000	000000000	00:00:19	151595	2015	01	04	01	00	20	00	000000000	000000000	00:00:20
151595	2015	01	04	01	00	20	00	000000000	000000000	00:00:20	151596	2015	01	04	01	00	21	00	000000000	000000000	00:00:21
151596	2015	01	04	01	00	21	00	000000000	000000000	00:00:21	151597	2015	01	04	01	00	22	00	000000000	000000000	00:00:22
151597	2015	01	04	01	00	22	00	000000000	000000000	00:00:22	151598	2015	01	04	01	00	23	00	000000000	000000000	00:00:23
151598	2015	01	04	01	00	23	00	000000000	000000000	00:00:23	151599	2015	01	04	01	00	24	00	000000000	000000000	00:00:24
151599	2015	01	04	01	00	24	00	000000000	000000000	00:00:24	151600	2015	01	04	01	00	25	00	000000000	000000000	00:00:25
151600	2015	01	04	01	00	25	00	000000000	000000000	00:00:25	151601	2015	01	04	01	00	26	00	000000000	000000000	00:00:26
151601	2015	01	04	01	00	26	00	000000000	000000000	00:00:26	151602	2015	01	04	01	00	27	00	000000000	000000000	00:00:27
151602	2015	01	04	01	00	27	00	000000000	000000000	00:00:27	151603	2015	01	04	01	00	28	00	000000000	000000000	00:00:28
151603	2015	01	04	01	00	28	00	000000000	000000000	00:00:28	151604	2015	01	04	01	00	29	00	000000000	000000000	00:00:29
151604	2015	01	04	01	00	29	00	000000000	000000000	00:00:29	151605	2015	01	04	01	00	30	00	000000000	000000000	00:00:30
151605	2015	01	04	01	00	30	00	000000000	000000000	00:00:30	151606	2015	01	04	01	00	31	00	000000000	000000000	00:00:31
151606	2015	01	04	01	00	31	00	000000000	000000000	00:00:31	151607	2015	01	04	01	00	32	00	000000000	000000000	00:00:32
151607	2015	01	04	01	00	32	00	000000000	000000000	00:00:32	151608	2015	01	04	01	00	33	00	000000000	000000000	00:00:33
151608	2015	01	04	01	00	33	00	000000000	000000000	00:00:33	151609	2015	01	04	01	00	34	00	000000000	000000000	00:00:34
151609	2015	01	04	01	00	34	00	000000000	000000000	00:00:34	151610	2015	01	04	01	00	35	00	000000000	000000000	00:00:35
151610	2015	01	04	01	00	35	00	000000000	000000000	00:00:35	151611	2015	01	04	01	00	36	00	000000000	000000000	00:00:36
151611	2015	01	04	01	00	36	00	000000000	000000000	00:00:36	151612	2015	01	04	01	00	37	00	000000000	000000000	00:00:37
151612	2015	01	04	01	00	37	00	000000000	000000000	00:00:37	151613	2015	01	04	01	00	38	00	000000000	000000000	00:00:38
151613	2015	01	04	01	00	38	00	000000000	000000000	00:00:38	151614	2015	01	04	01	00	39	00	000000000	000000000	00:00:39
151614	2015	01	04	01	00	39	00	000000000	000000000	00:00:39	151615	2015	01	04	01	00	40	00	000000000	000000000	00:00:40
151615	2015	01	04	01	00	40	00	000000000	000000000	00:00:40	151616	2015	01	04	01	00	41	00	000000000	000000000	00:00:41
151616	2015	01	04	01	00	41	00	000000000	000000000	00:00:41	151617	2015	01	04	01	00	42	00	000000000	000000000	00:00:42
151617	2015	01	04	01	00	42	00	000000000	000000000	00:00:42	151618	2015	01	04	01	00	43	00	000000000	000000000	00:00:43
151618	2015	01	04	01	00	43	00	000000000	000000000	00:00:43	151619	2015	01	04	01	00	44	00	000000000	000000000	00:00:44
151619	2015	01	04	01	00	44	00	000000000	000000000	00:00:44	151620	2015	01	04	01	00	45	00	000000000	000000000	00:00:45
151620	2015	01	04	01	00	45	00	000000000	000000000	00:00:45	151621	2015	01	04	01	00	46	00	000000000	000000000	00:00:46
151621	2015	01	04	01	00	46	00	000000000	000000000	00:00:46	151622	2015	01	04	01	00	47	00	000000000	000000000	00:00:47
151622	2015	01	04	01	00	47	00	000000000	000000000	00:00:47	151623	2015	01	04	01	00	48	00	000000000	000000000	00:00:48
151623	2015	01	04	01	00	48	00	000000000	000000000	00:00:48	151624	2015	01	04	01	00	49	00	000000000	000000000	00:00:49
151624	2015	01	04	01	00	49	00	000000000	000000000	00:00:49	151625	2015	01	04	01	00	50	00	000000000	000000000	00:00:50
151625	2015	01	04	01	00	50	00	000000000	000000000	00:00:50	151626	2015	01	04	01	00	51	00	000000000	000000000	00:00:51
151626	2015	01	04	01	00	51	00	000000000	000000000	00:00:51	151627	2015	01	04	01	00	52	00	000000000	000000000	00:00:52
151627	2015	01	04	01	00	52	00	000000000	000000000	00:00:52	151628	2015	01	04	01	00	53	00	000000000	000000000	00:00:53
151628	2015	01	04	01	00	53	00	000000000	000000000	00:00:53	151629	2015	01	04	01	00	54	00	000000000	000000000	00:00:54
151629	2015	01	04	01	00	54	00	000000000	000000000	00:00:54	151630	2015	01	04	01	00	55	00	000000000	000000000	00:00:55
151630	2015	01	04	01	00	55	00	000000000	000000000	00:00:55	151631	2015	01	04	01	00	56	00	000000000	000000000	00:00:56
151631	2015	01	04	01	00	56	00	000000000	000000000	00:00:56	151632	2015	01	04	01	00	57	00	000000000	000000000	00:00:57
151632	2015	01	04	01	00	57	00	000000000	000000000	00:00:57	151633	2015	01	04	01	00	58	00	000000000	000000000	00:00:58
151633	2015	01	04	01	00	58															

Year	Country	Population	Age	Sex	Height	Weight	Body Mass Index	Body Fat Percentage	Heart Rate	Blood Pressure	Cholesterol	Glucose	Insulin	Triglycerides	HDL	LDL	Total
2010	USA	308,745,538	38.1	M	1.78	78.2	24.2	15.1	72	120/80	180	100	1.0	150	50	100	150
2011	USA	312,149,937	38.2	M	1.78	78.2	24.2	15.1	72	120/80	180	100	1.0	150	50	100	150
2012	USA	315,491,434	38.3	M	1.78	78.2	24.2	15.1	72	120/80	180	100	1.0	150	50	100	150
2013	USA	318,872,931	38.4	M	1.78	78.2	24.2	15.1	72	120/80	180	100	1.0	150	50	100	150
2014	USA	322,254,428	38.5	M	1.78	78.2	24.2	15.1	72	120/80	180	100	1.0	150	50	100	150
2015	USA	325,635,925	38.6	M	1.78	78.2	24.2	15.1	72	120/80	180	100	1.0	150	50	100	150
2016	USA	329,017,422	38.7	M	1.78	78.2	24.2	15.1	72	120/80	180	100	1.0	150	50	100	150
2017	USA	332,398,919	38.8	M	1.78	78.2	24.2	15.1	72	120/80	180	100	1.0	150	50	100	150
2018	USA	335,780,416	38.9	M	1.78	78.2	24.2	15.1	72	120/80	180	100	1.0	150	50	100	150
2019	USA	339,161,913	39.0	M	1.78	78.2	24.2	15.1	72	120/80	180	100	1.0	150	50	100	150
2020	USA	342,543,410	39.1	M	1.78	78.2	24.2	15.1	72	120/80	180	100	1.0	150	50	100	150

[illegible]

2000 variabel

8.500 rows

Caso Alstom: Analisi Predittiva

Obiettivi

- 1) Valutare in modo automatico se una segnalazione di guasto attivata dal sistema di monitoraggio, sia effettivamente da segnalare all'area di manutenzione

Tecniche di Machine Learning utilizzate

- Decision Tree
- Random Forest
- Neural Networks
- Gradient Boosting Machine



Variabili per ogni evento osservato: tot. 300 var. delle 2.000 disponibili

- logs eventi diagnostica sw AV (Alta Velocità)
- logs eventi diagnostica sw SCMT (Sistema Controllo Marcia Treno)
- eventi rilevanti delle corse AV
- eventi rilevanti delle corse SCMT
- eventi relativi alla odometria

last_run_id	ISSUE	1__CAB_A_ON	1__SERV_TETTO	1__EB_ANTENNA_SWITCH		2__CAB_A_ON	2__SERV_TETTO	2__EB_ANTENNA_SWITCH		3__CAB_A_ON	3__SERV_TETTO	3__EB_ANTENNA_SWITCH	vero/falso
1559453	1	30	50	25		6	41	28		29	51	17	0
1561388	2	17	55	16		23	33	28		12	52	25	1
1561966	1	13	67	11		30	50	26		5	41	27	1
1593270	3	15	67	14		29	45	24		7	45	28	1
1656659	2	16	72	30		27	43	21		9	49	32	0
1656661	2	16	72	32		27	43	21		9	49	32	0
1656676	1	21	72	47		27	43	21		9	49	32	0
1699514	1	19	97	12		22	80	13		20	83	15	1
1704569	1	13	66	15		16	103	14		24	56	15	1
1748299	1	23	78	10		26	80	15		14	40	14	0
1783005	1	32	42	16		17	61	24		15	108	14	0
1817617	1	27	67	13		21	42	11		16	52	10	0
1653170	1	20	35	18		37	66	28		39	32	17	1
1658885	1	23	61	12		18	30	14		43	69	36	1

all'evento

500 Km dall'evento

1000 Km dall'evento

Caso Alstom: risultato ottenuto

Dal risultato ottenuto, emerge la possibilità di ridurre del 25% le false segnalazioni di guasti, che potrebbe tradursi in una riduzione dell'impiego del personale di manutenzione.

Grazie per l'attenzione

Stefano Roselli s.roselli@ Cineca.it