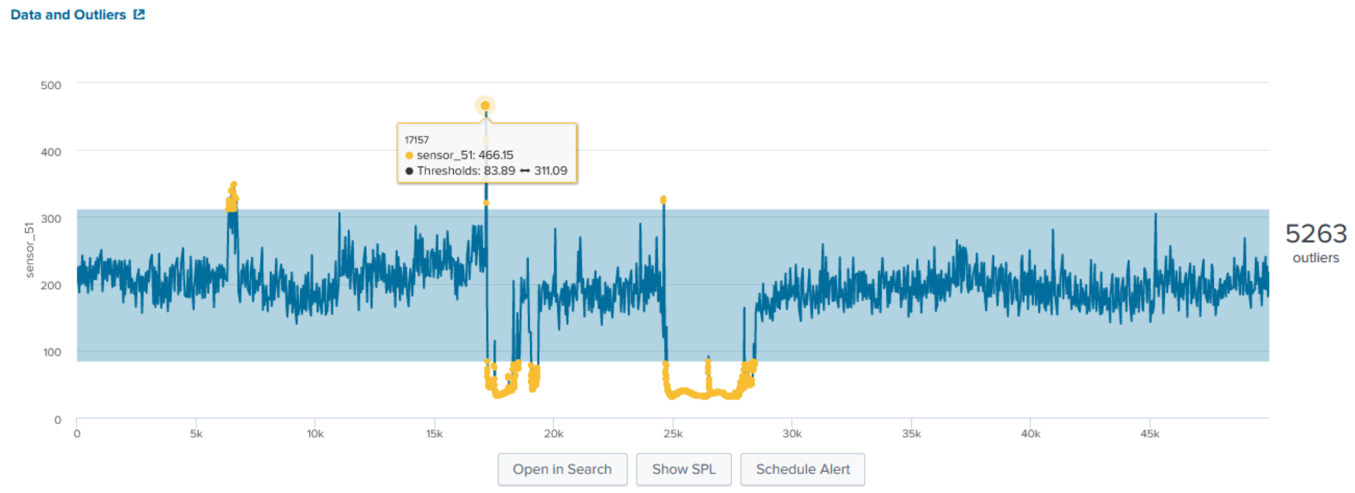
ingeniat
estudios y proyectos SL

DATA ANALYSIS AND MACHINE LEARNING

PROVIDING INSIGHT AND KNOWLEDGE THROUGH DATA WRANGLING



A new paradigm

We are facing a new industrial revolution, “Industry 4.0”, where machines and sensors can connect to your IT infrastructure to provide more profound insight into your business and Key Performance Indicators (KPIs).

With the advent of this new paradigm, systems and monitoring applications are producing enormous amounts of actionable data allowing for cost optimization, prediction of future events, behavior classification, quality control, and a number of other functionalities.

The connection of sensors from remote locations to your local or remote IT infrastructure can be undertaken in a seamless manner through a low cost, energy efficient, and secure Internet of Things (IoT) network.

Business intelligence overviews can be generated, alerts programmed and additional functionality plugged in and actioned based upon the received and analyzed data.

Prediction and forecasting

Machine learning (ML) models can be generated allowing for prediction on most valuable operational parameters.

The process typically involves the definition of a desired goal, the selection of a target class, the training of a ML model to recognize or predict the target and, eventually, the application of that same ML model to new operational data.

An interesting field of application in predictive maintenance considers the prediction of unknown, undesired events, such as equipment failure. Prediction of such unknown disruptive events is typically referred to as “anomaly detection”.

Workflow

Our workflow typically considers the following stages:

- Indexing and visualization. Your data is indexed into our search cluster, and a customized, tailor-made, dashboard with relevant data and KPIs is generated specifically for the source data.

- Sharing and fine-tuning. Remote access to the dashboard is granted.

- Data wrangling, modeling and forecasting. A ML model is generated based upon the processed data, and forecasting algorithms are run to determine the predicted behavior.

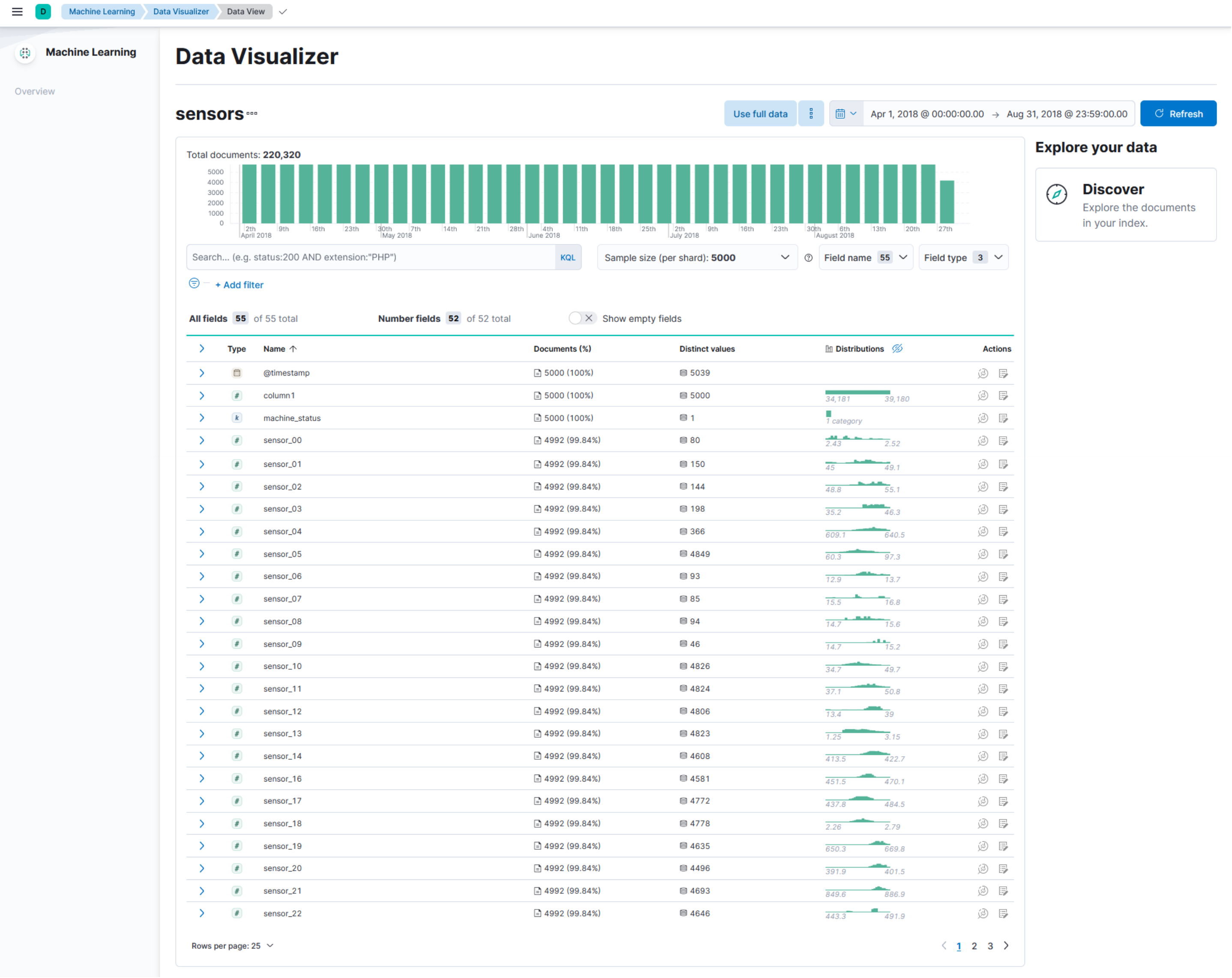
- Forecasting reports. Reports predicting the future behavior are generated and shared to the stakeholders. Where applicable, alarms can be set or actionable advice can be provided.

On a case by case basis, we can also undertake the integration of related sensors or APIs through the implementation of software forwarders which can feed the data straight from the sensors to the indexing cluster.

We can also assist with the local deployment of search clusters or additional infrastructure necessary to undertake data analysis visualization at the customer’s premises.

Indexing and visualization

While data visualization can provide quick insight into relevant metrics, it is hard to scale such insight when presented with different types of charts and graphics.



Data Visualizer

sensors

Apr 1, 2018 @ 00:00:00.00 → Aug 31, 2018 @ 23:59:00.00

Total documents: **220,320**

Search... (e.g. status:200 AND extension:"PHP") KQL

Sample size (per shard): **5000**

Field name 55 · Field type 3

All fields 55 of 55 total **Number fields** 52 of 52 total Show empty fields

Type	Name ↑	Documents (%)	Distinct values	Distributions
📅	@timestamp	5000 (100%)	5039	
#	column1	5000 (100%)	5000	34,181 – 39,180
k	machine_status	5000 (100%)	1	1 category
#	sensor_00	4992 (99.84%)	80	2.43 – 2.52
#	sensor_01	4992 (99.84%)	150	45 – 49.1
#	sensor_02	4992 (99.84%)	144	48.8 – 55.1
#	sensor_03	4992 (99.84%)	198	35.2 – 46.3
#	sensor_04	4992 (99.84%)	366	609.1 – 640.5
#	sensor_05	4992 (99.84%)	4849	60.3 – 97.3
#	sensor_06	4992 (99.84%)	93	12.9 – 13.7
#	sensor_07	4992 (99.84%)	85	15.5 – 16.8
#	sensor_08	4992 (99.84%)	94	14.7 – 15.6
#	sensor_09	4992 (99.84%)	46	14.7 – 15.2
#	sensor_10	4992 (99.84%)	4826	34.7 – 49.7
#	sensor_11	4992 (99.84%)	4824	37.1 – 50.8
#	sensor_12	4992 (99.84%)	4806	13.4 – 39
#	sensor_13	4992 (99.84%)	4823	1.25 – 3.15
#	sensor_14	4992 (99.84%)	4608	413.5 – 422.7
#	sensor_16	4992 (99.84%)	4581	451.5 – 470.1
#	sensor_17	4992 (99.84%)	4772	437.8 – 484.5
#	sensor_18	4992 (99.84%)	4778	2.26 – 2.79
#	sensor_19	4992 (99.84%)	4635	650.3 – 669.8
#	sensor_20	4992 (99.84%)	4496	391.9 – 401.5
#	sensor_21	4992 (99.84%)	4693	849.6 – 886.9
#	sensor_22	4992 (99.84%)	4646	443.3 – 491.9

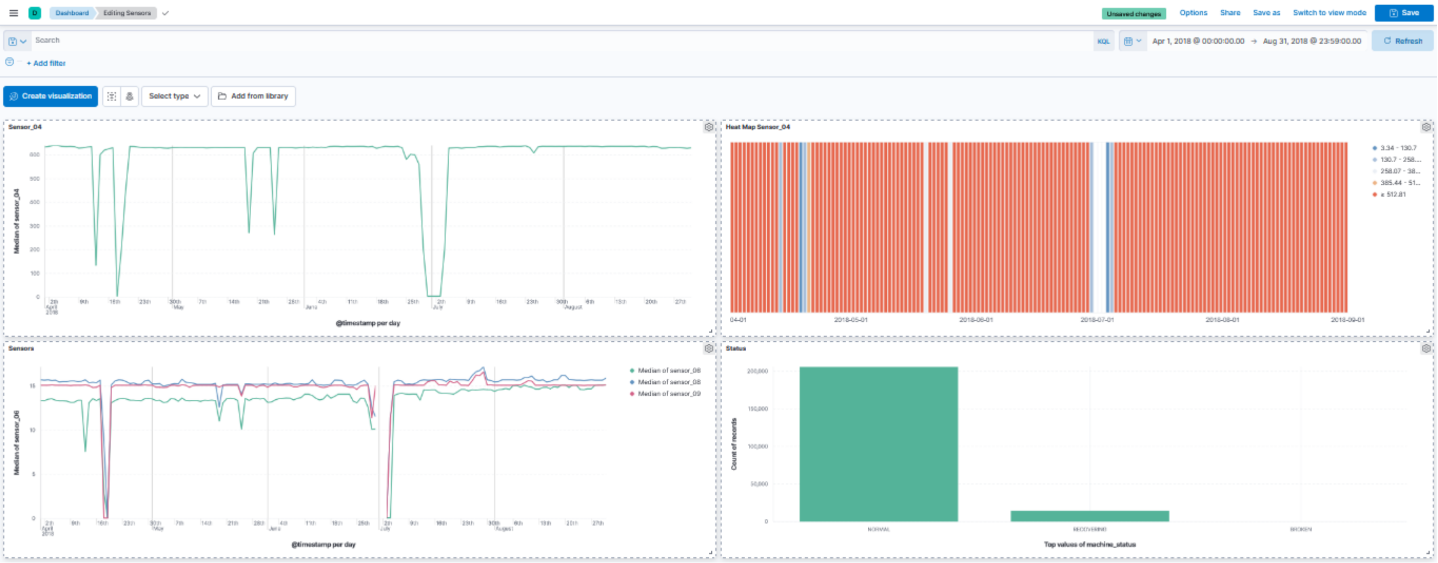
Rows per page: 25 · 1 2 3

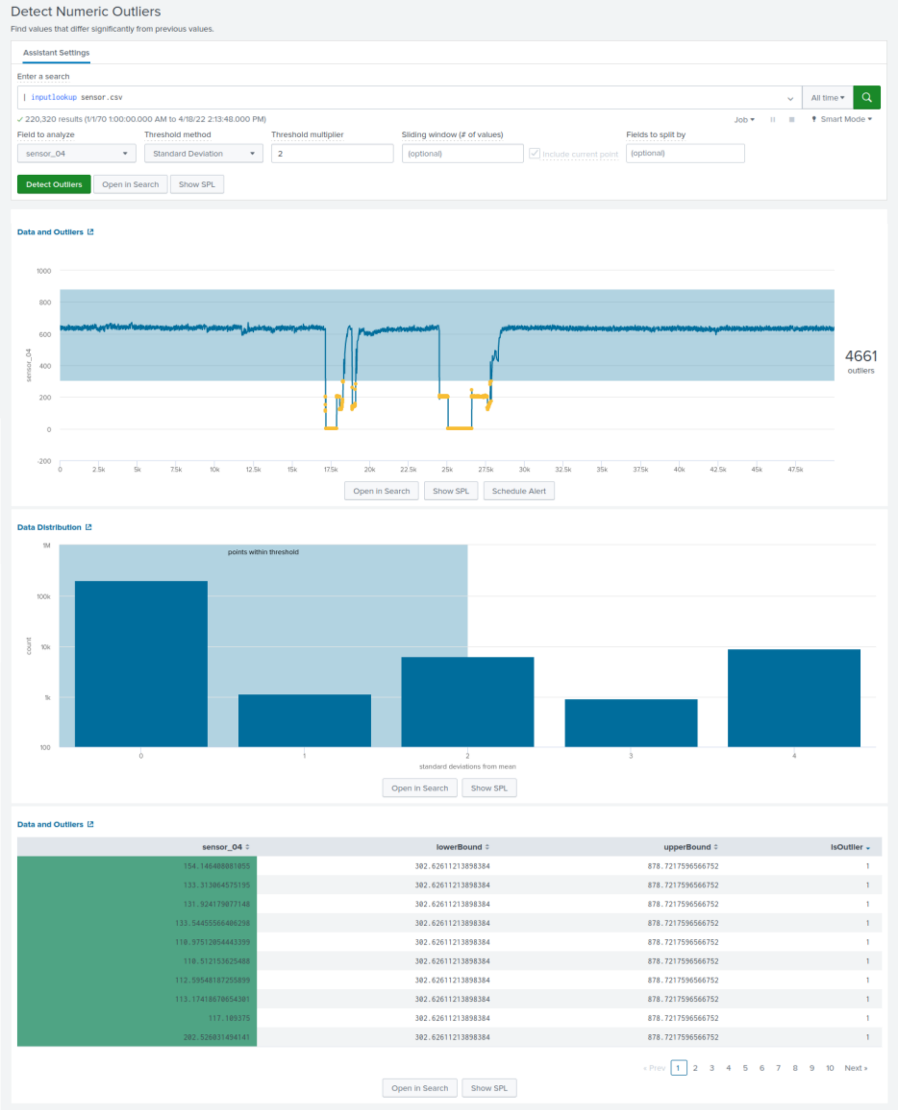
Image above / below

Indexing raw static data, originating from real life sensors installed on rotating equipment, without forwarder or data broker, for processing and machine learning, and dashboard showing real time logging from data broker.

The ability to monitor multiple key performance indicators (KPIs) and metrics is crucial to understanding the operational parameters of equipment as well as the source of potential breakdowns, allowing for streamlining data collection, analysis, and collaboration with dashboards that help users to quickly digest critical information and make smart decisions in real time.

The sample dashboard below shows typical information gathered during operation, with sensor readings being displayed and customized to show not just real data, but processed data (ie, individual or aggregated median values) in real time. This is just a sample, though; more complex tailor-made dashboards can be developed to showcase a wide range of data of interest, as well as KPIs, in real time.





Detect Numeric Outliers

Find values that differ significantly from previous values.

Assistant Settings

Enter a search

```
| inputlookup sensor.csv
```

✓ 220,320 results (1/1/70 1:00:00.000 AM to 4/18/22 2:13:48.000 PM)

Field to analyze: sensor_04

Threshold method: Standard Deviation

Threshold multiplier: 2

Sliding window (# of values): (optional)

☑ Include current point

Fields to split by: (optional)

Data and Outliers

sensor_04	lowerBound	upperBound	IsOutlier
154.146408081055	302.62611213898384	878.7217596566752	1
133.313064575195	302.62611213898384	878.7217596566752	1
131.924179077148	302.62611213898384	878.7217596566752	1
133.54455566406298	302.62611213898384	878.7217596566752	1
110.97512054443399	302.62611213898384	878.7217596566752	1
110.512153625488	302.62611213898384	878.7217596566752	1
112.59548187255899	302.62611213898384	878.7217596566752	1
113.17418670654301	302.62611213898384	878.7217596566752	1
117.109375	302.62611213898384	878.7217596566752	1
202.526031494141	302.62611213898384	878.7217596566752	1

Image above

Detection of numeric outliers by means of unsupervised anomaly detection methods. Data provided by a real life sensor installed on rotating equipment.

Unsupervised anomaly detection

Anomaly detection is a technique which identifies rare observations which are statistically different from the rest of observation, potentially translating into some kind of problem or consequence, such as equipment failure.

Referring to the image on the left, nominal operation can be observed due to the significantly overwhelming number of similar records, whereas two distinct, singular, events can also be observed. Those two events represent a small percentage of anomalous data, while at the same time the anomalies are statistically very different from the normal samples. Those two requirements are typically present in the unsupervised anomaly detection methods, where normal values are clustered using a similarity measure and datapoints which are determined to be far off from the cluster are considered to be anomalies or outliers.

A more complex detection can be undertaken using ML procedures, as long as we have access to labeled datasets containing both normal and anomalous samples. Availability of such data would allow us to construct predictive models which can assist in classifying future data points.

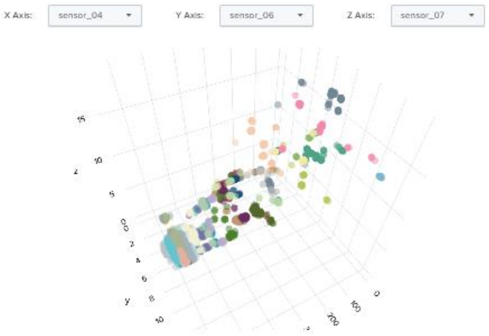
This type of detection is commonly referred to as supervised anomaly detection.

Supervised anomaly detection

Meaningful insight can be extracted from automatic handling and processing of massive amounts of data thanks to anomaly detection, supervised or unsupervised. By continuously monitoring a live environment, detection algorithms can effectively expose themselves to massive amounts of training data, an exposure which would allow them to understand what is "business as usual" and what is otherwise an outlier for an organization's specific environment. In order to be able to gather, clean, structure, analyze and store data (a process often referred to as data wrangling), specific tools and infrastructure capable of processing big volumes of data is required. The more data that is made available to the system, the more accurate the system becomes.

The largest risks that an operation can encounter will almost certainly be unknown, or in other words, something for which preparations in advance could not be made.

The systems need to be as experienced and accurate as possible in order to detect a new unknown that signals the risk.



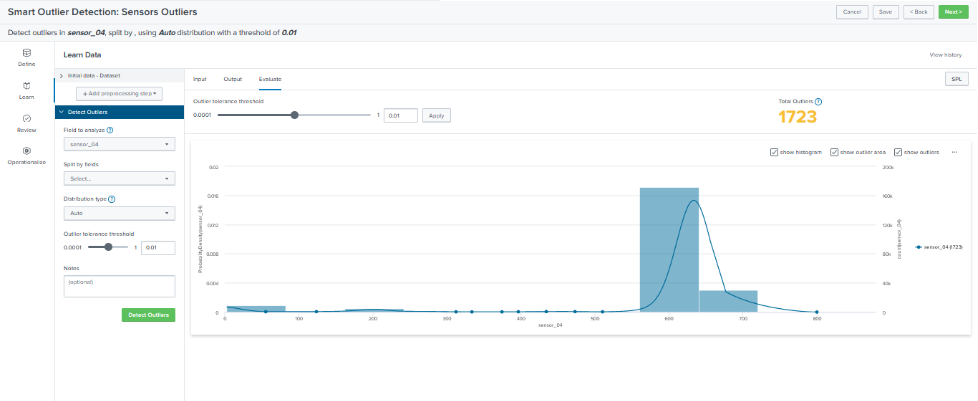
X Axis: sensor_04 ▾ Y Axis: sensor_06 ▾ Z Axis: sensor_07 ▾

Image above

Correlation of data and clustering of sensor readings.

Image below

Classification of outliers during anomaly detection.



Smart Outlier Detection: Sensors Outliers

Detect outliers in ***sensor_04***, split by , using ***Auto*** distribution with a threshold of ***0.01***

Learn Data

Field to analyze: sensor_04

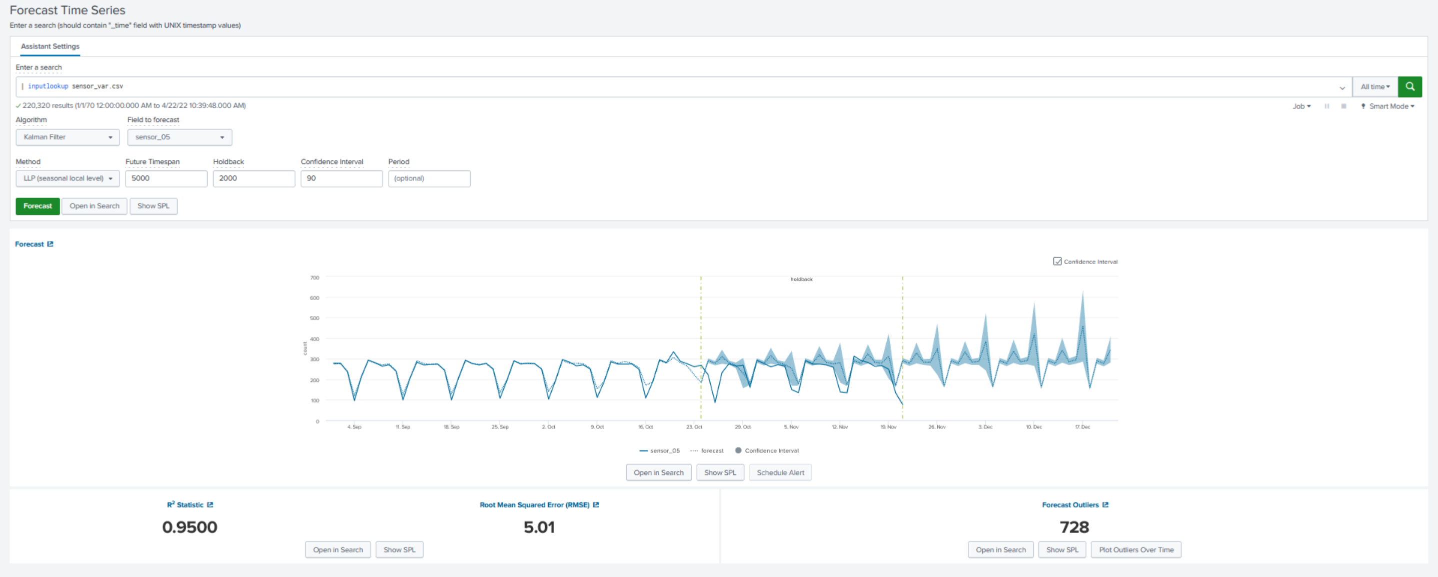
Split by fields: Select...

Distribution type: Auto

Outlier tolerance threshold: 0.0001 – 1: 0.01

Notes: (optional)

Total Outliers: 1723



Forecast Time Series

Enter a search (should contain "_time" field with UNIX timestamp values)

Assistant Settings

Enter a search

```
| inputlookup sensor_var.csv
```

✓ 220,320 results (1/1/70 12:00:00.000 AM to 4/22/22 10:39:48.000 AM)

Algorithm: Kalman Filter

Field to forecast: sensor_05

Method: LLP (seasonal local level)

Future Timespan: 5000

Holdback: 2000

Confidence Interval: 90

Period: (optional)

R² Statistic: 0.9500

Root Mean Squared Error (RMSE): 5.01

Forecast Outliers: 728

Image above

Forecasting of periodic (seasonal) data, utilizing part of the historical time series to validate the model (holdback period) and projecting the forecast into the future with a certain confidence interval. Data provided by a real life sensor installed on rotating equipment.

Predictions and forecasting

Forecasting involves taking models fit on historical data and using them to predict future observations. The underlying intention of time series forecasting is determining how target variables will change in the future by observing historical data from the time perspective, defining patterns, and developing short or long-term predictions, considering the historical patterns.

All the previously, recently, and currently collected data are used as input for time series forecasting where future trends, seasonal changes, irregularities, and such are elaborated based on complex math-driven algorithms.

The image above shows forecasting of a time series with a strong seasonal pattern (seasonality refers to a property of time series that displays periodical patterns repeating at a constant frequency) with some minor residual deviation. The forecasting algorithm is run holding back some historical data which will be used to test the validity and accuracy of the results (central section of the graph), while finally projecting the forecast (rightmost section of the graph).

Not all time series will show such a predictable pattern, but even more complex, random series can be decomposed through the so called Seasonal-Trend decomposition through LOESS (or STL decomposition) allowing for the breakdown of the complex series in more approachable components (seasonal, trends, and residuals).

Predictions can be identically carried out based upon known historical values of other, related, parameters. The random forest regression algorithm, for instance, takes advantage of the 'wisdom of the crowds'; it takes multiple (but different) regression decision trees and makes them 'vote'. Random forest regression then calculates the average of all of the predictions to generate a great estimate of what the expected value for the target parameter should be.

The image below shows the prediction of a range of values of an individual sensor installed on a rotating equipment, based upon historical values of readings taken from a range of sensors also installed on the same equipment, and where correlation, between readings in those sensors and readings in the target sensor, has been known to exist. This procedure is useful in a number of scenarios such as, for instance, predicting the impact of cost variations in material and subcomponents on the final cost of production of a product.

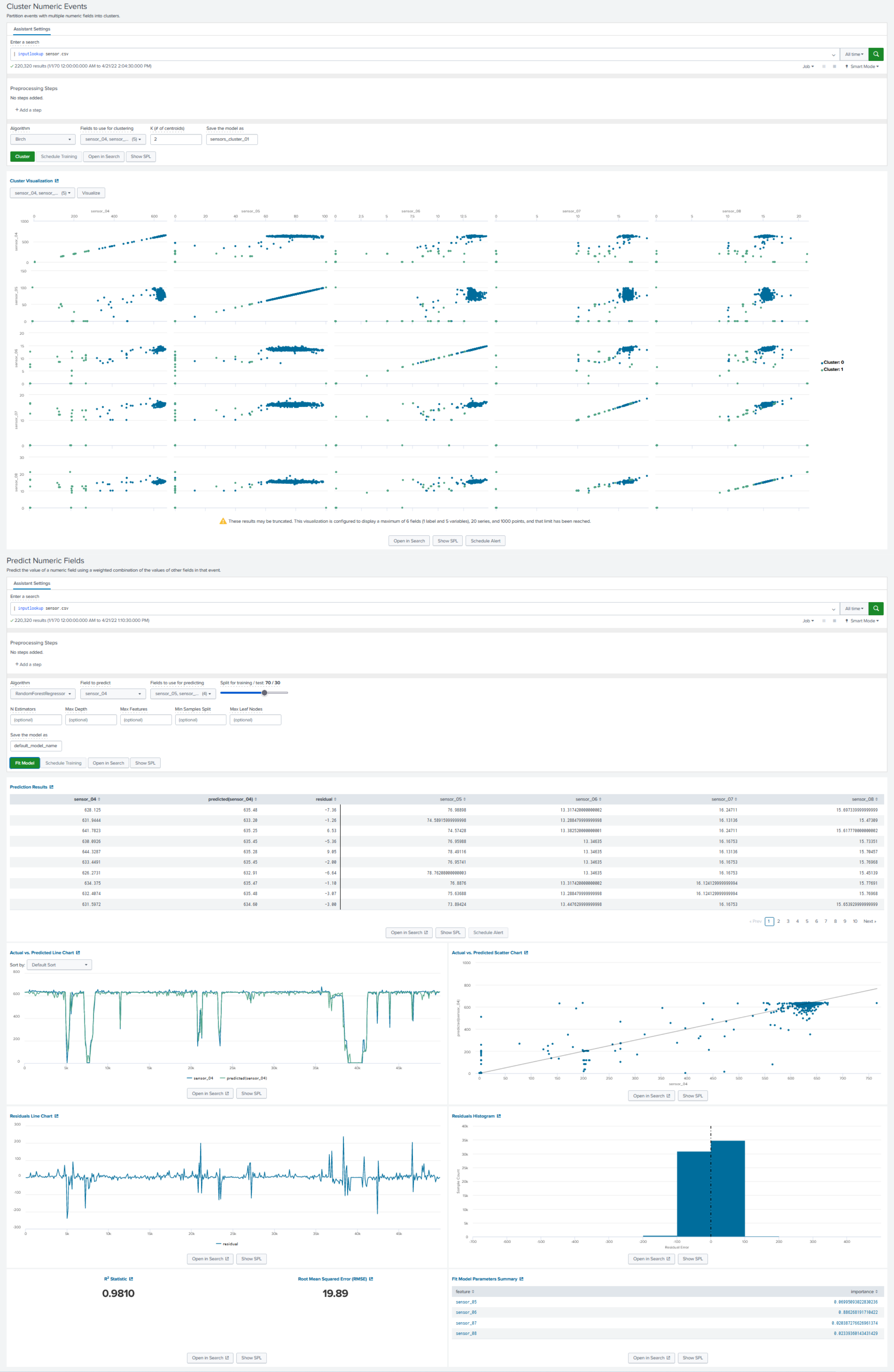
The first chart shows the scatter plot of reading combinations between sensors. Graphs on the diagonal of the chart show perfect self-correlation of a sensor with respect to itself. Graphs in other location show correlation of a sensor with respect to a different one, and it can be observed that correlation is fairly high in all circumstances, with readings grouping tightly in clusters.

Deviations with respect to these clusters are a consequence of anomalous operation of the equipment, yielding outliers in data. Having ascertained proper correlation, certain inputs can in consequence be used in averaged prediction of the desired output, using the random forest regression algorithm, which is shown in the bottom half of the image below.

It can be observed that the predicted model fits the real life results with a significant degree of accuracy. The scatter plot located on the right hand side of the image shows a strong fit, with discrepancies being due to the operational anomalies shown on the line chart located on the left hand side of the image. This is also supported by the spikes which can be observed on the residuals line chart.

Image next

Scatter plot showing clustering of data registers and correlation between readings from different sources, and prediction of target data using input data from other sources by means of the random forest regression algorithm.



Cluster Numeric Events

Partition events with multiple numeric fields into clusters.

Assistant Settings

Enter a search

```
| inputlookup sensor.csv
```

✓ 220,320 results (1/1/70 12:00:00.000 AM to 4/21/22 2:04:30.000 PM)

Preprocessing Steps

No steps added.

+ Add a step

Algorithm: Birch

Fields to use for clustering: sensor_04, sensor_... (5)

K (# of centroids): 2

Save the model as: sensors_cluster_01

⚠ These results may be truncated. This visualization is configured to display a maximum of 6 fields (1 label and 5 variables), 20 series, and 1000 points, and that limit has been reached.

Predict Numeric Fields

Predict the value of a numeric field using a weighted combination of the values of other fields in that event.

Assistant Settings

Enter a search

```
| inputlookup sensor.csv
```

✓ 220,320 results (1/1/70 12:00:00.000 AM to 4/21/22 1:10:30.000 PM)

Preprocessing Steps

No steps added.

+ Add a step

Algorithm: RandomForestRegressor

Field to predict: sensor_04

Fields to use for predicting: sensor_05, sensor_... (4)

Split for training / test: **70 / 30**

N Estimators: (optional)

Max Depth: (optional)

Max Features: (optional)

Min Samples Split: (optional)

Max Leaf Nodes: (optional)

Save the model as: default_model_name

Prediction Results

sensor_04	predicted(sensor_04)	residual	sensor_05	sensor_06	sensor_07	sensor_08
628.125	635.48	-7.36	76.98898	13.317420000000002	16.24711	15.697339999999999
631.9444	633.20	-1.26	74.58915999999998	13.288479999999998	16.13136	15.47309
641.7823	635.25	6.53	74.57428	13.382520000000001	16.24711	15.617770000000002
630.0926	635.45	-5.36	76.95988	13.34635	16.16753	15.73351
644.3287	635.28	9.05	78.49116	13.34635	16.13136	15.70457
633.4491	635.45	-2.00	76.95741	13.34635	16.16753	15.76968
626.2731	632.91	-6.64	78.76208000000003	13.34635	16.16753	15.45139
634.375	635.47	-1.10	76.8876	13.317420000000002	16.124129999999994	15.77691
632.4074	635.48	-3.07	75.63688	13.288479999999998	16.124129999999994	15.76968
631.5972	634.60	-3.00	73.89424	13.447629999999998	16.16753	15.653929999999999

R² Statistic

0.9810

Root Mean Squared Error (RMSE)

19.89

Fit Model Parameters Summary

feature	importance
sensor_05	0.06995093022830236
sensor_06	0.886268191710422
sensor_07	0.020387276626961374
sensor_08	0.02339360143431429

INGENIAT
ESTUDIOS Y PROYECTOS SL

C/VENEZUELA 62-64 ENT.6
15404 FERROL, SPAIN

PHONE
+34 663 700 398

PHONE
+34 644 358 624

EMAIL
CONTACT@INGENIAT.ES

WEBSITE
WWW.INGENIAT.ES