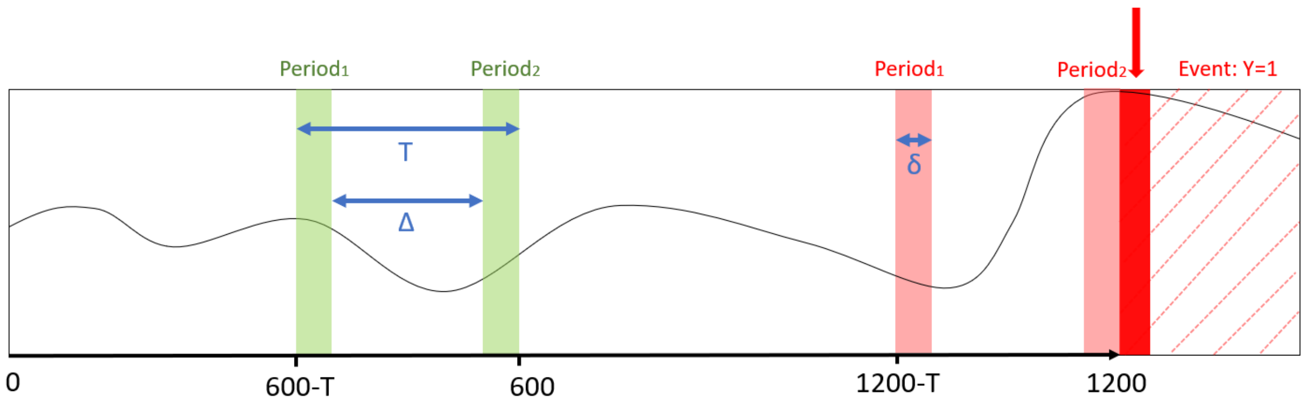

Figure 3: This figure shows how from one time series of duration 20 hours (1200 minutes), we make a cutout to obtain the control sample in green and the case sample in red.

- F2-score is a weighted F-score and used when it is much worse to miss a True Positive than giving a False Positive

$$(1+2^2) \times \frac{precision \times recall}{(2^2 \times precision) + recall}$$

These metrics can be used on the training database, on a test set, and in cross-validation.

Case-crossover APriori Predictive 2 (CAPP2). A complementary approach called Case-crossover APriori Predictive 2 (CAPP2) has been studied in order to improve the quality of prediction. Indeed, in addition to looking for the rules leading to an event of the form

$$\{X^{(1)}_{\{0,0.33\}} = True, X^{(2)}_{\{0.66,1\}} = True\} \implies \{\text{Event} = \text{True}\}$$

we have also looked for the contraposed, which are the rules that do not lead to an event (leading to "Event = False") of the form:

$$\{X^{(4)}_{\{0.33,0.66\}} = True, X^{(3)}_{\{0.33,0.66\}} = True\} \implies \{\text{Event} = \text{False}\}$$

Let us call "Event=True rules" the first rules and "Event=False rules" the second ones. There are different ways to combine these two approaches in order to compute a more robust predictive model. Among them, we could adjust the number of rules proving to be True for each of the two approaches, we could give more weight to the "Event=True rules" for the prediction or give more weight to the "Event=False rules". Cross-validation allows to test, observe and study the behavior of each of these experiments. The decision could be improved by learning the decision combining the two types of rules.

5 APPLICATION

5.1 Data

Numerous sensors were placed at various points in the distillation unit to collect data and monitor the evolution of the system. More than 800 variables were measured, providing information such as the type of input crude, pressures, temperatures, flow rates, valve openings, and chemical measurements. The variables are categorical or continuous and take positive values. These measurements were carried out every minute for 4 months and the identification of flooding events is calculated using a formula involving variables from the outputs of the distillation column and is presented in the data in the form of an additional binary column where a 0 represents a normal operation condition system and 1 represents the flooding event. Each column represents a measured variable and each line describes the system at a specific minute.

In our study, we consider that flooding events are independent of each other. For this reason, we only take into account events that occur at least 20 hours apart from each other. Thus, we identify a total of 38 long time series (of duration 20 hours=1200 minutes) to be studied. We have therefore cut the data into 38 long time series where the last moments correspond to the appearance of the flooding event. In order to build the database $\mathcal{D}_n$ using the case-crossover design, we must have pairs $(X_{t=0,\ldots,T}, Y)$. Therefore, we need to define the duration T of the time series and get samples such that we have couples with labels $Y = 1$ and $Y = 0$ coming from the same long time series. The label $Y = 1$ is simple to obtain because we just have to select the last moments of each of the 38 time series because by definition they all end with a flooding event. For the label $Y = 0$, we had to sample and select a part of the 38 series. Since we assume that the samples are independent, we have to select this period so that it is far enough from the flooding event and under normal operating conditions. With the advice of experts, we decided to select samples at a time distance of 10 hours=600 minutes from the flooding event. This step of selection of periods requires preliminary knowledge of the phenomenon in order to select the periods of "normal" and "abnormal" functioning. In our case, we know that the event is acute and occurs in the hour before the event.

Figure 3 summarizes the principle of the case-crossover design and highlights the data cutout to obtain the control and case of Figure 1 and Figure 2.

Therefore, to learn rules, we have a training database $\mathcal{D}_n = \{((X_{it})_{t=0,\ldots,T}, Y_i), i = 1,\ldots,n\}$ where $n = 76$. 38 samples of $\mathcal{D}_n$ have a label $Y_i = 1$ and 38 samples have a label $Y_i = 0$. The sampling is done every minute and we have 4 hours of measurements for each sample, hence $T = 240$.

5.2 Interpretable Rules found by CAP

In this subsection, we use expert knowledge of the characteristic times of important phenomena to determine certain time parameters such as δ. For the rest of the parameters, we did not want to optimize them too much to avoid overfitting, optimizing the thresholds is an idea to keep in mind if the learning base is large enough.

After preprocessing the data, we computed the Apriori algorithm with the described design with a period duration of $\delta = 60$, 1 hour sampled every minute, and a gap $\Delta = 120$ of 2 hours between period_1 and period_2. We set $min\_support \geq 0.2$ and $min\_len = 2$ and sort the results by confidence and lift. The rules that have been found are shown in Table 3.

Among the rules, we can see the presence of $X^{(1)}$ which is a variable computed from a physical model and used to be, before the random forest model, the variable allowing to determine the appearances of flooding events. Moreover, $X^{(2)}$ is a flow recirculation variable and has been selected by experts as being very likely to explain the flooding appearance.

5.3 CAPP1 Prediction Results

To prevent overfitting and evaluate well the CAPP1 method performance, we decided to do a Leave-Two-Out (LTO). For $j \in \{1, ..., n/2\}$, we take the $(2j-1)^{th}$ and $(2j)^{th}$ element of the database $\mathcal{D}_n$ for testing, such that we have a couple computed from one of the 38 long time series with one element having a label $Y = 0$ and the other $Y = 1$, and we take the $n-2$ other elements of $\mathcal{D}_n$ as a training set. The training set provides data to the Apriori algorithm in order to learn rules using the different metrics we defined. The rules are then sorted by confidence and lift and are ready to be tested. For the testing, as described in subsection 4.5, we predict the two elements in the test set. Finally, we evaluate the prediction by computing the True Positive Rate, the True Negative Rate, and the F2 score and compute the mean of these scores over the 38 tests we have done with our LTO. Thus, in the following, all calculated scores are obtained by cross-validation.

As mentioned in section 4.5, we select the 10 rules with highest confidence and lift, and with two or fewer explanatory variables, then we calculate the quality of the prediction using the defined metrics. We evaluate the predictive performance of the CAPP1 method by a comparison with the one of a random forest (RF) algorithm. The RF is trained with the dataset D_n and takes as input the averages of the input variables over $[T-\delta, T]$, $[T-2\delta, T-\delta], \ldots, [0, \delta]$ and predicts the binary label "there is a flooding at time $T+1$ minute". The results are shown in Table 4.

We could always increase the True Positive Rate by choosing a higher threshold for the minimum support and increasing the number of rules but this will directly affect the True Negative Rate as there is a trade-off between True Positives and False Positives. If our model is more sensitive and often rings an alarm, it will make more errors and then more False Positives.

The results are satisfactory as the True Positive Rate is relatively high and far better than a random prediction without even optimizing our algorithm but are insufficient compared to the random forest algorithm.

5.4 CAPP2 Prediction Results

After several tests, we opted for the following combination: we set $min\_support \geq 0.01$ and sort the results by confidence with a minimum threshold of 0.5. If at least one out of the first 100 "Event=True rules" and less than one out of the first 100 "Event=False rules" is True, we predict that the tested pair leads to a flooding event i.e. $Y = 1$. Otherwise, we predict that the pair does not lead to a flooding event i.e. $Y = 0$. Since a minimum threshold of confidence has been set, the number of rules can be lower but limited to 100. Note that the choice of 100 rules here is empirical and depends on the choice of the minimum support threshold.

CAPP2 has allowed us to improve our prediction results and obtain the scores presented in Table 4.

Algorithm	F2 score	TNR	TPR/recall
Random Forest	0.8127	0.8368	0.8684
CAPP1	0.6991	0.6644	0.8684
CAPP2	**0.9139**	0.9210	0.8947

Table 4: Prediction scores.

These results are promising as the CAPP2 method achieves better scores than RF without optimizing our model with a relatively small dataset and especially with a model that proposes a causal analysis.

6 CONCLUSION AND FUTURE WORK

We have developed a data-driven model based on the case-crossover design and association rule mining for determining the causes of an incident from time series. This approach overcomes two main issues: the lack of interpretability and prediction based on correlations. The understanding of incidents is essential because it would allow to predict in advance their appearance using a causal prediction algorithm and to be able to justify the reliability and confidence contrarily to a black-box algorithm.

The application and study of this approach to our dataset provide conclusive results confirming that the method is promising. This work gives insight to operators working in the refinery with the distillation unit and allows them to understand the mechanisms that trigger the event. The method finds interesting rules and describes associations between variables leading to an event. Among the top rules sorted by confidence, we find the variables that have been suspected to be causal by the experts. The associations make it possible to strengthen them and to add missing information necessary to the understanding of the phenomenon of flooding. In addition, our predictive study has shown that we could build a strong predictive model which could outperform the one actually in production. Indeed, the results on the four-month dataset have confirmed these expectations and there is still a lot of room for improvement.

This method uses expert knowledge to select certain parameters. In the absence of such information, methods to determine these characteristic times must be considered and more failure case data may be needed for this.

Several approaches have been identified for future work. Among them, we could cite the following ideas: instead of choosing two arbitrary quantiles as we did in this work, we could optimize them and adapt their number. We could also deepen the contraposed

rules	support	confidence	lift
$\{X^{(1)}_{\{0,0.33\}} = True, X^{(2)}_{\{0.33,0.66\}} = False\} \implies$ {Event=True}	0.2315789	0.9777778	1.955556
$\{X^{(1)}_{\{0,0.33\}} = True, X^{(3)}_{\{0.33,0.66\}} = False\} \implies$ {Event=True}	0.2236842	0.9770115	1.954023
$\{X^{(1)}_{\{0,0.33\}} = True, X^{(4)}_{\{0,0.33\}} = False\} \implies$ {Event=True}	0.2210526	0.9767442	1.953488
$\{X^{(5)}_{\{0,0.33\}} = True, X^{(2)}_{\{0.33,0.66\}} = False\} \implies$ {Event=True}	0.3118421	0.9753086	1.950617
$\{X^{(1)}_{\{0,0.33\}} = True, X^{(6)}_{\{0.33,0.66\}} = False\} \implies$ {Event=True}	0.2052632	0.9750000	1.950000
$\{X^{(5)}_{\{0,0.33\}} = True, X^{(2)}_{\{0.66,1\}} = False\} \implies$ {Event=True}	0.2552632	0.9748744	1.949749
$\{X^{(1)}_{\{0,0.33\}} = True, X^{(7)}_{\{0.33,0.66\}} = False\} \implies$ {Event=True}	0.2013158	0.9745223	1.949045
$\{X^{(5)}_{\{0,0.33\}} = True, X^{(3)}_{\{0,0.33\}} = False\} \implies$ {Event=True}	0.2355263	0.9728261	1.945652
$\{X^{(1)}_{\{0,0.33\}} = True, X^{(2)}_{\{0,0.33\}} = False\} \implies$ {Event=True}	0.2315789	0.9723757	1.944751
$\{X^{(1)}_{\{0,0.33\}} = True, X^{(8)}_{\{0.33,0.66\}} = False\} \implies$ {Event=True}	0.2315789	0.9723757	1.944751

Table 3: This table displays the rules found by the algorithm sorted by confidence and lift. The support is also shown here.

approach CAPP2, add variables based on mean thresholds, improve our predictive model by aggregating the results over multiple analyses with different Δ and δ and optimize the event detection system.

ACKNOWLEDGMENTS

This work was supported by TotalEnergies and the research laboratory SINCLAIR. We also thank Tobias Kurth, Professor of Public Health and Epidemiology and the Director of the Institute of Public Health at the Charité Universitätsmedizin in Berlin, and Lluvia Ochoa, Researcher in Data Science and Artificial Intelligence at TotalEnergies who provided their insights and expertise.

REFERENCES

[1] H Alkharusi. 2012. Categorical variables in regression analysis: A comparison of dummy and effect coding. *International Journal of Education* 4(2) (2012), 202–210.
[2] Jorge A.M. Azevedo P.J. 2007. Comparing Rule Measures for Predictive Association Rules. *ECML* (2007).
[3] Leo Breiman. 2001. Random Forests. *Machine Learning* 45 (2001), 5–32.
[4] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone. 2017. *Classification And Regression Trees.* 1–358 pages. https://doi.org/10.1201/9781315139470
[5] Clément Bénard, Gérard Biau, Sébastien Da Veiga, and Erwan Scornet. 2021. SIRUS: Stable and Interpretable RUle Set for classification. *Electronic Journal of Statistics* 15 (01 2021), 427–505. https://doi.org/10.1214/20-EJS1792
[6] Yonghong Chen, Govindan Rangarajan, Jianfeng Feng, and Mingzhou Ding. 2004. Analyzing multiple nonlinear time series with extended Granger causality. *Physics Letters A* 324, 1 (2004), 26 – 35. https://doi.org/10.1016/j.physleta.2004.02.032
[7] K. Dembczynski, W. Kotlowski, and R. Slowinski. 2010. ENDER: a statistical framework for boosting decision rules. *Data Mining and Knowledge Discovery* 21 (2010), 52–90.
[8] Doris Entner and Patrik Hoyer. 2010. On Causal Discovery from Time Series Data using FCI. *Proceedings of the 5th European Workshop on Probabilistic Graphical Models, PGM 2010* (09 2010).
[9] Gardner IA et al. Estberg L. 1998. A case-crossover study of intensive racing and training schedules and risk of catastrophic musculoskeletal injury and lay-up in California thoroughbred racehorses. *Prev. Vet. Med* 33 (1998), 159–170.
[10] Maclure M et al. 1993. Triggering of acute myocardial infarction by physical exertion: protection against triggering by regular exertion. *N. Engl. J. Med* 329 (1993), 1677–1683.
[11] Maclure M et al. 1995. Triggering of acute myocardial infarction by episodes of anger. *Circulation* 92 (1995), 1720–1725.
[12] A. Eye, Wolfgang Wiedermann, and I. Koller. 2015. Granger Causality: Linear Regression and Logit Models.
[13] Eibe Frank and Ian Witten. 1998. Generating Accurate Rule Sets Without Global Optimization. *Machine Learning: Proceedings of the Fifteenth International Conference* (06 1998).
[14] C. W. J. Granger. 1969. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* 37, 3 (1969), 424–438. http://www.jstor.org/stable/1912791
[15] T. Hastie, R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. In *Springer Series in Statistics.*
[16] Liang H. Hu M. 2014. A copula approach to assessing Granger causality. *NeuroImage* 100 (2014), 125–134.
[17] T. Imielinski, Arun Swami, and Rajat Agrawal. 1993. Mining Association rules between sets of items in large databases. *ACM SIGMOD* (01 1993), 207–216.
[18] J. R KISTER, H. Z; HAAS. 1990. Predict entrainment flooding on sieve and valve trays. *Chemical engineering progress* (1990).
[19] O. Ludwig, U. Nunes, R. Araújo, L. Schnitman, and H. Lepikson. 2009. Applications of information theory, genetic algorithms, and neural models to predict oil flow. *Communications in Nonlinear Science and Numerical Simulation* 14 (2009), 2870–2885.
[20] M. Maclure. 1991. The Case-Crossover Design: A Method for Studying Transient Effects on the Risk of Acute Events. *American journal of epidemiology* 133 (1991), 144–153. https://doi.org/10.1093/oxfordjournals.aje.a115853
[21] M. Maclure and M. Mittleman. 1997. Cautions about car telephones and collisions. *The New England journal of medicine* 336 7 (1997), 501–2.
[22] Mittleman MA Maclure M. 1997. Cautions about car telephones and collisions. *N. Engl. J. Med.* 336 (1997), 501–502.
[23] Stramaglia S. Marinazzo D., Pellicoro M. 2008. Kernel method for nonlinear Granger causality. *Phys. Rev. Lett.* 100 (2008), 144103.
[24] R. Michalski. 1969. On the quasi-minimal solution of the general covering problem. (01 1969).
[25] Meike Nauta, Doina Bucur, and Christin Seifert. 2019. Causal Discovery with Attention-Based Convolutional Neural Networks. *Machine Learning and Knowledge Extraction* 1, 1 (2019), 312–340. https://doi.org/10.3390/make1010019
[26] J. Pearl. 2000. *Causality: Models, Reasoning and Inference.* Cambridge University Press.
[27] R. Srikant R. Agrawal. 1994. Fast Algorithms for Mining Association Rules, In VLDB Conference. *Mathematical Modelling and Computational Experiments.*
[28] Dilini Rajapaksha, Christoph Bergmeir, and Wray Buntine. 2019. LoRMIkA: Local Rule-based Model Interpretability with k-optimal Associations. arXiv:1908.03840 [cs.LG]
[29] Sebastian Raschka. 2014. *Mlxtend Apriori.* http://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/.
[30] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:1602.04938 [cs.LG]
[31] Jakob Runge, Dino Sejdinovic, and Seth Flaxman. 2017. Detecting causal associations in large nonlinear time series datasets. *Science Advances* 5 (02 2017). https://doi.org/10.1126/sciadv.aau4996
[32] S. Salzberg. 1994. C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning* 16 (1994), 235–240.
[33] Peter Spirtes, Clark Glymour, and Richard Scheines. 2001. *Causation, Prediction, and Search, 2nd Edition.* MIT Press Books, Vol. 1. The MIT Press. https://ideas.repec.org/b/mtp/titles/0262194406.html
[34] Giulia Toti, R. Vilalta, P. Lindner, B. Lefer, Charles Macias, and D. Price. 2016. Analysis of correlation between pediatric asthma exacerbation and exposure to pollutant mixtures with association rule mining. *Artificial intelligence in medicine* 74 (2016), 44–52.
[35] Sholom Weiss and Nitin Indurkhya. 2000. Lightweight Rule Induction. (05 2000).