Taking a Proactive Approach to GDPR with KNIME

Phil Winters
Phil.Winters@knime.com

Summary

The General Data Protection Regulation (GDPR) is the new EU law for the protection of natural persons with regard to the processing of personal data and the free movement of such data. Data Protection Officers and executives will now be liable for noncompliance and are of course key individuals for any GDPR activity. Anyone who uses customer data will also be massively affected in their daily jobs. Those individuals include data scientists, data miners, statisticians, and business executives who deal with customer data for making decisions.

This white paper is designed to help data scientists by highlighting some of the best practices to implement to ensure that we can continue to use customer data in a way that is not only compliant for our organization but useful for the business and even beneficial to the individuals affected. Working examples based on KNIME Analytics platform are provided throughout to back up the concepts introduced.

The workflows used in this whitepaper are available on the KNIME EXAMPLES public server under *50_Applications/34_GDPR_examples*. This workflow group also contains all data required to run the workflows. The data used comes from https://archive.ics.uci.edu/ml/datasets/adult .

This whitepaper highlights some of the best practices for data scientists to ensure that their work is compliant with the GDPR law.

Table of Contents

GDPR Background

On May 25, 2016, the EU passed the world's strongest and most far-reaching law aimed at strengthening citizens' fundamental rights in the digital age. The regulation also tries to facilitate business best practice by unifying rules for companies operating within the EU Digital Single Market. Before that it was up to the individual countries to decide how to implement existing EU laws and recommendations, which added complexity for businesses operating in multiple countries. This new, 88-page General Data Protection Regulation (GDPR) is something that EU member states voted for unanimously: one law for the entire region. And it will apply as of May 25, 2018.

The GDPR will apply to any company, organization or body anywhere in the world that processes the personal data of any EU resident. In simple terms, if you are using personal data about any EU resident, the rules will most likely apply to you and your organization.

In the works since 2012, the GDPR seeks to establish a modern and harmonized data protection framework across the EU. Not surprising of such a complex undertaking, some aspects make for quite alarming reading – particularly the parts about the sky-high fines that can be imposed on persons and organizations in breach of compliance.

Many aspects of the law require careful evaluation and action by organizations and their legal teams and there are many pundits recommending how to move forward if you are a new organization starting from scratch. But the majority of us will already have systems and processes in place that already contain personal data so the "green field" approach will not be suitable - what is required are concrete suggestions about how to get existing systems and processes compliant by May 25, 2018 as well as how to continue moving forward. There are a few overriding themes in the GDPR that center on the forward thinking around the automation required for taking and documenting decisions around customer data and that is where an analytics platform like KNIME plays a major role.

If you are unfamiliar with GDPR, you will quickly want to review the Appendix to focus first on understanding some of the important

terminology used throughout the GDPR to decide whether GDPR affects you.

GDPR and Automated Profiling

There are 173 preambles and 99 articles in the GDPR covering a wide range of topics and requirements and many of those topics – such as establishing a data protection officer, gaining appropriate permissions and handling data breaches – are beyond the scope of this document. But there are specific articles and preambles related to profiling and automated processing that involve personal data as well as the requirements for informing and communicating to the data subjects. It is those articles that we will review here.

- **Article 25 Data Protection by Design and by Default**
  - This is the key article around personal data themselves
  - It requires appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage, and their accessibility.
  - Fundamentally, this means organizations should no longer be capturing personal data "just in case". Instead when personal data are collected, they need to be with a particular purpose in mind, and that purpose must be clearly stated (and permission gained – see next article) before we start. As data scientists, this will require us to think much more in advance of possible business topics and the required personal data, if any, we would need for analysis.
  - But for us data scientists, it does not mean we need to stop exploring. Within the new law, there is also a very important concept introduced around pseudoanonymization, namely:

    Preamble 26: The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This regulation does not therefore

concern the processing of such anonymous information, including for statistical or research purposes.

- **Article 6: Lawfulness of Processing**
  - Provide the capability to explicitly capture and associate one or more permissions with an individual record as per Article 7 defining conditions of consent.
  - Alternately provide the capability to associate a record of lawful processing because of a legal or contractual obligation or a justification.
  - While this is not the job of the data scientist, it WILL be important that we can show what data are being captured and used and to ensure this information is made available to the data protection officer, legal team, or other management team concerning themselves with lawfulness of processing and possibly the necessary permissions.

- **Article 9: Processing of Special Categories of Personal Data.**
  - Provide the capability to mark any identified special category personal data and to restrict access and use of such data.
  - Additionally, provide the capability to associate a record of lawful processing based on one of the 10 exceptions listed in Article 9.
  - For analytics, this will be about documenting whether we have those categories of personal data and, if so, how we are using them so that appropriate permissions or exceptions can be associated with that processing.

- **Article 22: Automated Individual Decision-making, including Profiling**
  - The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.
  - For a data scientist, this means any time we generate a model or a rule based on the personal data of an identifiable person that is then used to automatically categorize or make a decision, then this article comes into effect. Propensity to purchase, next best action, credit score,

channel propensity, recommendation engines and all of the hundreds of topics associated with machine learning and data science that we do with customer data every day will be affected by this Article.

- **Article 12: Transparent Information, Communication and Modalities for the Exercise of the Rights of the Data Subject**
  - There are many articles (Articles 13, 14, 15 to 22 and 34) dealing with data subject requests to know about personal data. For all of these requests, an organization must have the ability to provide the information in an easily understandable form. This applies to all systems storing and using personal data.
  - For a data scientist, this applies to documenting and being able to explain – both internally, to regulatory authorities, and also to the data subjects – exactly which personal data we are using and most importantly how we are using them.

- **Article 15: Right of Access by the Data Subject**
  - The existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject.

As a data scientist, we will need to help make our models and rules as understandable as possible.

The Data Scientist Responsibility and KNIME

All of these articles highlight activities that not only will be of great importance to our organizations but will also be precisely those activities that we as data scientists can tackle to help our organizations to be compliant with the new law. But how do these articles relate to our day-to-day jobs and what do we need to do? To understand that, let us start with a look at a standard analytic process. Whether you follow CRISP, SEMMA, or some other methodology for building analytics workflows, in the end they can all be represented by basic steps to Load, Transform, Evaluate, and Deploy a model or set of rules.

If you think of it, formal “automated individual decision-making” is nothing but our standard process. So, if we are already taking a structured approach to building and storing analytic processes, say using KNIME workflows, then we already have a basis for defining and documenting automation and profiling as stated in Article 22.

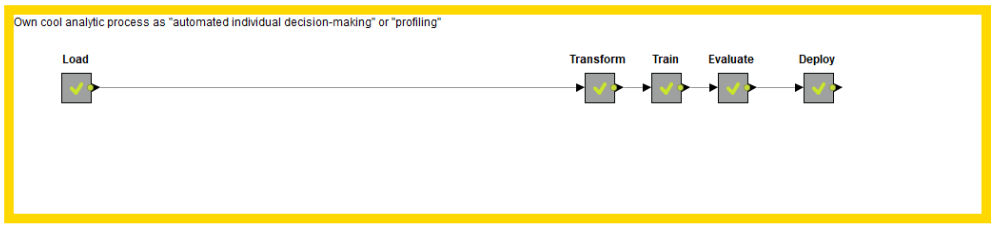


Figure 1.

Taking a structured approach to building and storing analytics processes

Within a standard process, there are concrete places where we can add in functionality to help make decisions and document them to help satisfy the requirements mentioned in the previous Articles and Preambles. They would take the form of 4 concrete sub workflows in our workflow.

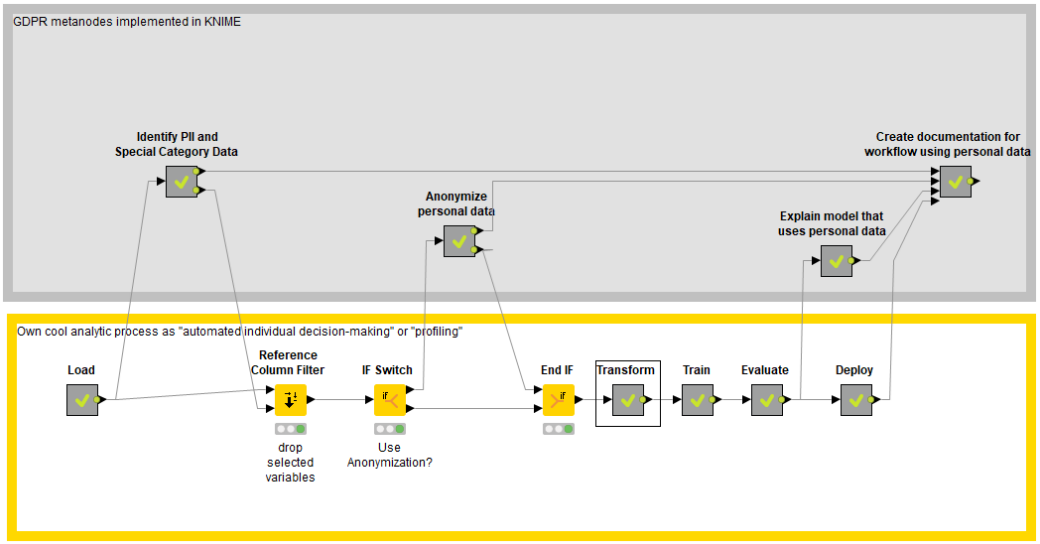


Figure 2.

Adding in functionalities to comply with GDPR

Identify and Flag Personal Data

The first sub-workflow would access our chosen data and would require some interaction and data understanding for completion.

- Does the data contain personal data and, if so, which fields would be used to identify a data subject?
- If the data does contain personal data, are there any fields that should be labeled as discriminatory and, if so, whether they should be:
    - Dropped or

    - Kept with appropriate qualifications
- If any fields have been identified as discriminatory, then run an automated modelling process to determine whether any OTHER fields could be used either singly or in combination to mimic those discriminatory fields. A well-known example here is that certain ZIPCODES in the US are a very accurate indicator of RACE, which is discriminatory.
- This sub-workflow would help us with Articles 25, 6, 9, and 12

Anonymize Personal Data

This sub-workflow would take the previously identified personal identifiers and create an anonymized data set around those identifiers.

- Important would be to document not only accuracy of the model but to run and document steps to ensure there was no straightforward way to reverse anonymized data back to the original data.
- With this sub-process, the decision could then be taken to either use the original data or use the anonymized data for the actual creation of the model or set of rules.
- This sub-workflow would help us with Articles 25 and 6.

Explain Model

Once data have been transformed, trained using a selected machine learning technique, and the evaluation (either automated or manual) has determined that the model is suitable for deployment, then this sub-workflow transforms the selected model into the “best” interpretation and explanation of that model.

- It is common knowledge in the data science community that some models are extremely difficult, if not impossible, to explain in depth (a good example here might be a deep learning model).
- But there is a requirement to explain to the best of our ability, in a clear and understandable way, which fields of personal information play a role in creating the model.
- This sub-workflow would help us with Articles 12 and 15.

Consolidate and Create Documentation

This node pulls together and stores all of the information that has been defined, decided, and created for this particular analytics process. The information will most likely be stored in multiple forms for:

- Internal stakeholders, such as the Data Protection Officer, who will require detailed understanding.
- External compliance organizations, who may require their own form of consolidated reporting.
- And most importantly, a form that can be used to satisfy the requests of a data subject.

This sub-workflow would help us with Articles 12 and 15

By establishing these four sub-processes within our organization and applying them consistently across all our analytic processes, we go a long way in supporting our organizations GDPR compliance requirements.

An Example: GDPR Workflows in KNIME

A concrete example is available on the KNIME public server. For this example, the Adults.CSV file is used. This dataset includes an extract of the 1994 US Census data and is available at https://archive.ics.uci.edu/ml/datasets/adult . The data set includes discriminatory data in the form of race. The data set has been enhanced with "unique identifiers" to simulate personally identifiable data. In addition, two new fields of information have been added. "Union Membership" has been randomly added while "zip code" has been added with a strong correlation to race. These last two, along with the original "race" field, are used to show capabilities of the metanodes.

The four examples relate 1:1 with diagram 2 above, but for demonstration and learning purposes have been implemented as stand-alone examples.

Identify PII and Special Category Data

This metanode requires input from a data scientist who understands the underlying data. It uses Quickforms and JavaScript views to display information about each field of information and to capture the decisions made by the data scientist.

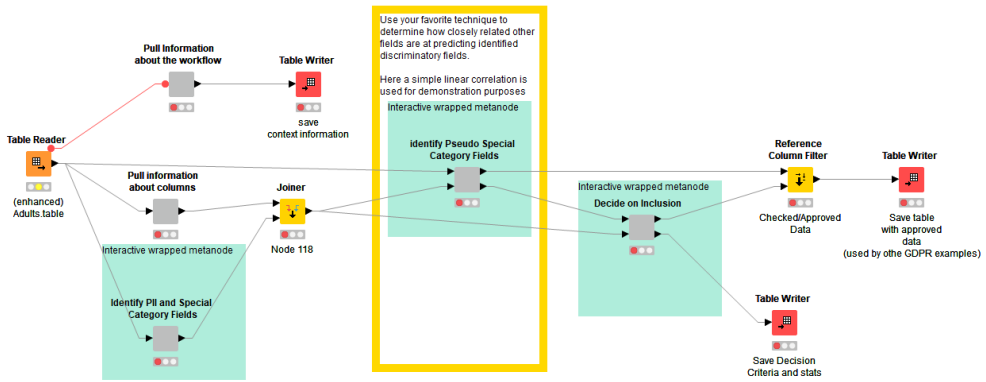


Figure 3.

Identifying PII and special category data using Quickforms and JavaScript Views

The metanode first starts by showing all columns that are available for use along with a sample of data contained within each column. From here, you flag fields of information to either indicate the field is used to identify a unique individual or that the field contains “special categories of data” (i.e.: a field representing a discriminatory field).

In our example, the field “unique ID” would be flagged as a unique identifier and the fields “race” and “Union_Member” would be flagged as discriminatory fields.

Once you have selected the fields, the metanode first checks if any “special category” fields were selected. If they were selected, then a loop is done, setting each “special category” field as the target variable. A method is then run to determine if one or more other fields can model that “special category” field.

This is a classic supervised predictive model where you can choose your own favourite technique, as well as your own measure of accuracy, to determine whether an existing discriminatory variable can be “accurately” modeled by other available data. In the example workflow, a simple linear correlation has been used with a correlation threshold of 0.4. If there is a correlation above 0.4, the corresponding field will be flagged.

In our example, the field “zip code” and “native_country” have been flagged as highly correlated to “race”. We are then presented with the information on all four fields (“race”, “Union_Member” “native_country” and “zip code”) and given the option of keeping or dropping those fields of data as well as to document our decision. In our example, we choose to drop three fields but to keep the field “native_country”.

That choice is made, the fields are dropped, the final data is made available for further use in our analytics process and all of our decisions are recorded in an external table.

Anonymize Personal Data

GDPR emphasizes the use of as little personally identifiable data as possible to complete a task. And with its special emphasis on "pseudo-anonymization" it will always be in our best interest to – whenever possible – use anonymized data.

When anonymization is performed, generally one needs to trade off the ability of the anonymized data to accurately represent the original data with computing resources and elapsed time. Depending on the data sometime well defined anonymized data might need more computing resources (and possibly elapsed time).

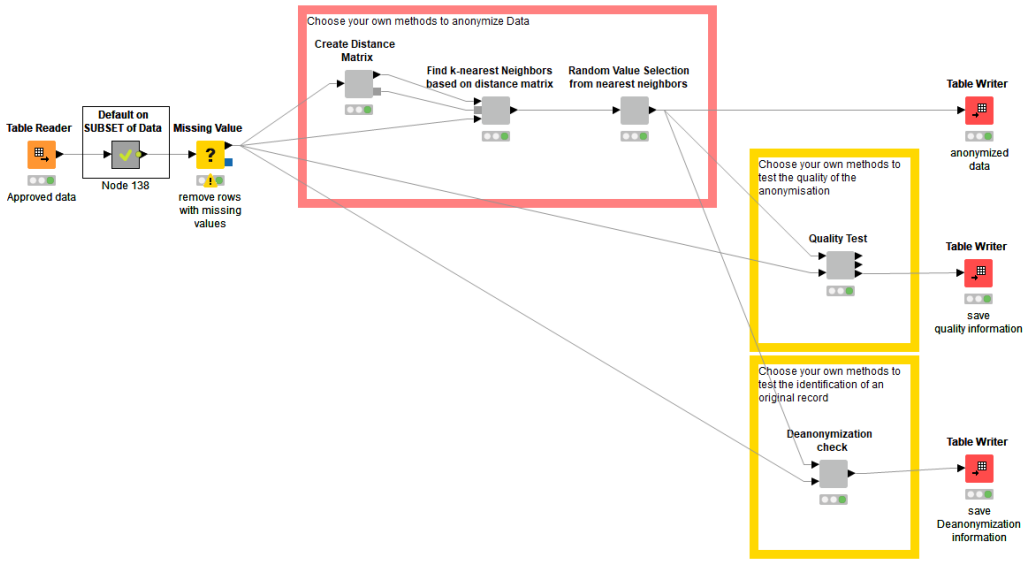


Figure 4.

Sample workflow to anonymize and then test the anonymized data

For our example where we do not have that many fields, we have chosen for illustration purposes to use a distance matrix technique using k-nearest neighbours that finds the two "closest" records for a given record, then randomly creates a new "unique" record from the two closest ones. This technique has the advantage of producing very similar data to the original data. It has the disadvantage that it takes either a lot of memory or a lot of time to compute. It can also have the additional disadvantage that it does not handle extreme outliers well, since the "two closest neighbours" might be more common patterns. Since there are not many extreme outliers in our data, this is acceptable. After creating the

anonymized data, two important tests are performed. First, a quality test is run to compare the quality of the anonymized data to the original data. This is done by running a cross-validation where we train a model and then predict the results first on the anonymized data, then on the original data, and then finally using the model trained on anonymized data to predict on the original data. All accuracies are captured and compared.

Next, we do a de-anonymization test to see if we can use the anonymized data to re-identify the person in the original data.

For our example, the accuracy of the anonymized model is very good (possibly because it clips outliers in finding the nearest neighbours that are more mainstream) and while at the same time, tests to deanonymize do not return any of the original data.

Note that the full example is very computationally intensive. The sample workflow shows the techniques but extremely limits the number of records and columns used.

The full data can be used but on a machine with a small amount of memory it will take some time for the distance matrix to calculate. On a machine with a lot of memory, the elapsed time is of course shorter.

For this example, we have decided that the new anonymized data satisfies our requirements and therefore we make the data available. At the same time, we ensure that all information about the anonymization process is captured and documented.

This is just one example. There are many techniques for anonymizing data including Generalization, Perturbation, Randomization, Masking, Vertex/Edge Clustering, K-anonymity and l-diversity. Each of these techniques has its advantages and disadvantages. Chose the technique that best suits your data, analytics, and compute resource requirements.

Explain Model

In this metanode, we use standard methods to “explain” a model. The law says we must make our best effort to explain what personal data are used in making a decision and do it in terms that a lay person can understand. Obviously if we try to explain the maths or the interactions of each of the fields that will be difficult, if not impossible (as in the case of neural nets or deep learning).

The law does NOT say we have to reveal our model or relative weights or other factors that may be either a company secret or not explainable. So, in this metanode the aim is to simply show fields that play a significant part in the model used in production.

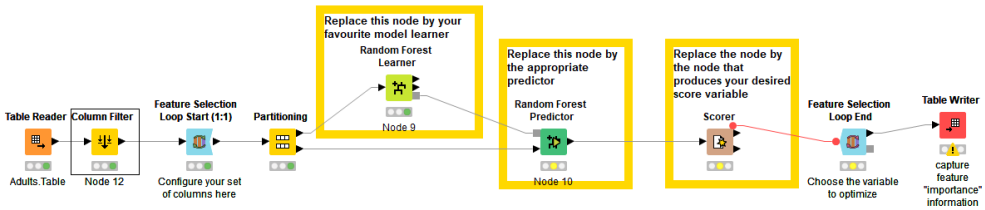


Figure 5.

Workflow for explaining the model

In our example, we have chosen to use a model in production created with Random Forest. To describe what fields are “important” in the model, we use a forward feature selection loop to create a list of relative importance of the different fields. In our case, this creates a list of fields based on “importance” along with the effect on the accuracy of that field on our model. This information is captured.

The choice of method will depend on the machine learning algorithm we have chosen. If we are doing classic supervised learning, we may choose to do a forward feature addition or a backward feature elimination to generate a list of “important features” that were used in the model. If we are doing unsupervised learning, deep learning or other analytic methods then we will want to use appropriate techniques here to also document “important” fields that go into the model. Note that there are even examples for neural networks and deep learning. Possibly not solid enough to defend a theoretical treatise on the subject but very sufficient for our documentation purposes.

Consolidate and Create Documentation

This workflow may be the most important of all since it is where all information is consolidated from the previous workflows and stored for final documentation.

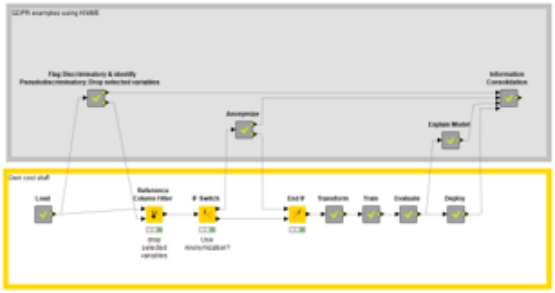
We will need to create different reports for the different stakeholders. Extremely detailed information will be required by our data protection officer and possibly our legal teams internally.

An example might look like this:

Model Name: **Income Category Prediction**

Created By: Phil
Last Updated By: Phil
Data Source: KNIME://LOCAL/GDPR/adults_zipcode_union.table
Workflow: KNIME://LOCAL/Random_Forest_Prediction
Tests Completed: Discrimination_Check
Anonymization
Model Explanation
Documentation

Analytic Process:



Discriminatory Data:

Field Name	GDPR Special Category Type	Pseudodiscriminatory Relationship Description	Action Taken
Zip Code	Pseudo Discriminatory	Correlates to "race" with a value of 0.53	Dropped
race	Discriminatory	Correlates to "Zip Code" with a value of 0.53	Dropped
Union_Member	Discriminatory	None	Dropped
age			
capital_gain			
capital_loss			
education			
education_num			
fnlwgt			
hours-per-week			
income			
marital_status			
native-country			
occupation			
relationship			
sex			
workclass			

Relative Feature Contribution:

Figure 6.

GDPR Compliance Report

The following workflow uses BIRT to create a similar report:

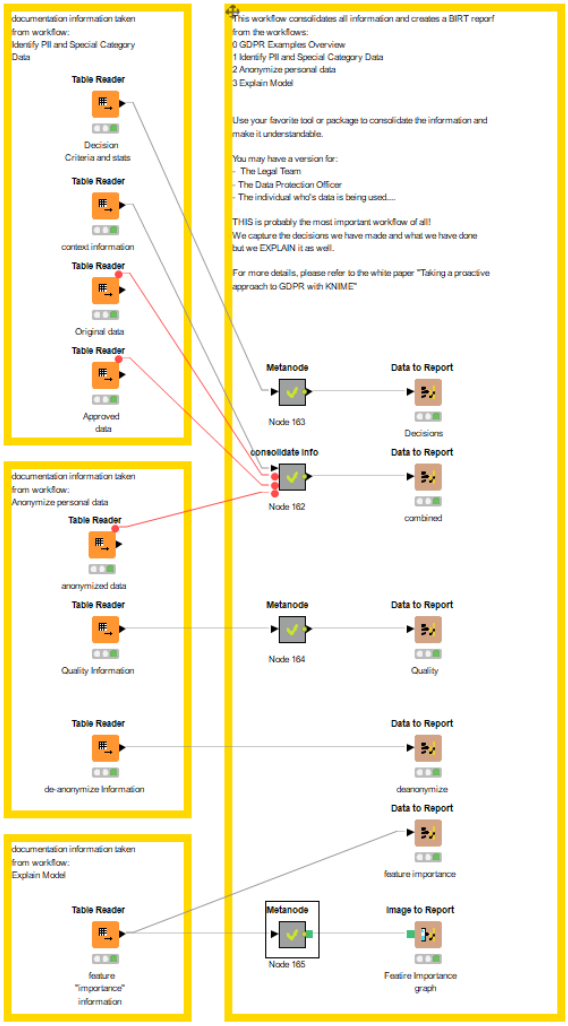


Figure 7.

Workflow to create a sample GDPR report

A different report presenting summarized information may be required by compliance authorities. And of course, the version for the individual data subject will be different again, possibly summarizing at a very high level the information required to fulfil the information request.

In all cases, you will want to use KNIME, possibly along with your favorite reporting tool, to create the required reports as defined by your organization.

Taking the Next Step with KNIME: Collaboration

It is clear there are quite a few possibilities for creating metanodes to support your organizations GDPR activities. Those metanodes can be designed so that they can be used for any analytic process within your organization and – if KNIME Server is installed – accessible and used by all data scientists who work with customer data.

But the opportunity for optimization goes much further. If you are using the KNIME Model Process Factory, you will already have the ideal setup for extending your production modeling and deploying activities by simply linking in your GDPR metanodes into your standard processes.

Conclusion

GDPR is a significant and at times complex law that will affect any of us processing customer data. But with the steps outlined here, you should be able to take significant steps in not only supporting your organizations GDPR compliance activities but also give you the flexibility to do your job as a data scientist.

It is important to understand: most people in our organizations will not have our understanding of the data. We must help our organizations, our management, and our data protection officers with GDPR. Because if we don't take control of these aspects, we will find ourselves with less data, less ability to support our organizations, and less ability to provide an even better customer experience for exactly those people who the law is designed to support – individuals like you and me.

All examples used in this whitepaper are located on the KNIME Public Server under *50_Applications/34_GDPR_examples*. It is hoped that the examples can help get you started faster with helping your organization become GDPR compliant.

Appendix: GDPR Terminology

'Regulation': a legal act of the European Union that becomes on enactment enforceable as law in all member states simultaneously. It becomes enforceable from May 25, 2018 after a two year transition period and, unlike a directive, it does not require any enabling legislation to be passed by national governments and is thus directly binding and applicable. Whether you are affected by this regulation depends on whether you process 'personal data'.

'Data subject': an identified or identifiable natural person who is in the EU or who's behavior takes place within the EU. This applies not only to citizens of those countries but residents of the countries or those doing business or partaking of services within the EU – regardless of their citizenship. Examples of data subjects include but are not limited to consumers, citizens, a business contact, a supplier, or an employee. All are natural persons.

'Personal data': any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural, or social identity of that natural person. If you are a 'controller' or 'processor' of personal data and do so from an EU country, GDPR will apply to you for any data subject, regardless of their physical location. If you are a 'controller' or 'processor' anywhere in the world and you process personal data of a data subject that is a resident in the EU, then GDPR will apply to you. Do not get into thinking there is a distinction between Business to Consumer and Business to Business personal data, legally these are all natural persons!

'Special categories of personal data': personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation. Sometimes this is referred to outside the actual law as "discriminatory data".

'Controller': the natural or legal person, public authority, agency, or other body which, alone or jointly with others, determines the purposes

and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law. In general, if you initiated the collection of personal data either directly or indirectly, then your organization is the 'controller' and liable under GDPR.

'Processor': a natural or legal person, public authority, agency, or other body which processes personal data on behalf of the controller. If you provide a service or system for your clients that has their customer's personal data contained in it, then you are a processor. A controller will want to have certain assurances from you to ensure they are complying with GDPR. The personal data you have about YOUR client contacts makes you the 'controller' of that data.

'Recipient': a natural or legal person, public authority, agency or another body, to which the personal data are disclosed.

'Processing': any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation, or alteration, retrieval, consultation, use, disclosure by transmission, dissemination, or otherwise making available, alignment or combination, restriction, erasure or destruction. That will cover all IT systems that contain personal data, regardless of whether those systems are on your own site, in a cloud or provided by a processor.

'Restriction of processing': the marking of stored personal data with the aim of limiting their processing in the future.

'Profiling': any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyze or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location, or movements. If you are using any sort of rules, machine learning, advanced analytics, or AI in any of your IT systems and if those use personal data, then profiling is being performed.

'Pseudonymisation': the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and

organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person.

‘Consent’: the data subject means any freely given, specific, informed, and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.

Throughout this article wherever possible, the GDPR terminology will be used.