Machine Learning final report

Project description :

For this project, we have chosen a dataset available on the Drees website (public health and social statistics), a national website that collects a wide range of health data in France. We will therefore be working on data relating to voluntary terminations of pregnancy. This dataset lists abortions classified by method and by year between 2016 and 2024, broken down by department, region, and for France as a whole. Our research question is part of our desire to identify geographical and temporal correlations that may influence abortion rates. We also want to be able to predict this rate in the future for any type of abortion and any geographical area.

This prediction could help the authorities forecast the need for infrastructure and healthcare personnel to perform these procedures and support patients.

Table des matières:

1-Descriptive analysis of our data

The dataset it's compose of 1 121 rows and 9 columns.

Here are the components of the dataset with some explanations.

- **ZONE_GEO** : The geographical area covered by the data, a department such as 'Ain' or an aggregation such as 'France entière' or 'région', from 2016 to 2024. Text (Object)
- **IVG_HOSP_INS** : Number of abortions (IVG) performed in hospital (full hospitalisation) using surgical instruments. Number (Float)
- **IVG_HOSP_MED** : Number of abortions performed in hospital (inpatient or outpatient) using medication. Number (Float)
- **IVG_HOSP_INC** : Number of abortions performed in hospital for unspecified or unknown reasons. Number (Float)
- **IVG_CAB**: Number of abortions performed in a doctor's office or healthcare facility without full hospitalisation. Number (Float)
- **IVG_CEN**: Number of abortions performed in family planning or education centres. Number (Float)
- **TOT_IVG:** The total number of voluntary terminations of pregnancy recorded for the corresponding GEO_ZONE and year. Number (Float)
- **TAUX_rec** : The abortion rate. This is generally the number of abortions per 1,000 women of childbearing age (15 to 49 years old). Number (Float)
- **Annee** : The year in which the abortion data was recorded. Number (Int)

2-Transformation and improvement of our data

We have detected quality issues in the dataset that are hindering the correct implementation of the data.

The variable TAUX_rec is text with commas to convert to float.

Some values are missing in some columns. We have also deleted each region and total to retain only the department in order to avoid bias due to certain data being repeated several times. To do this, we have deleted the following lines:

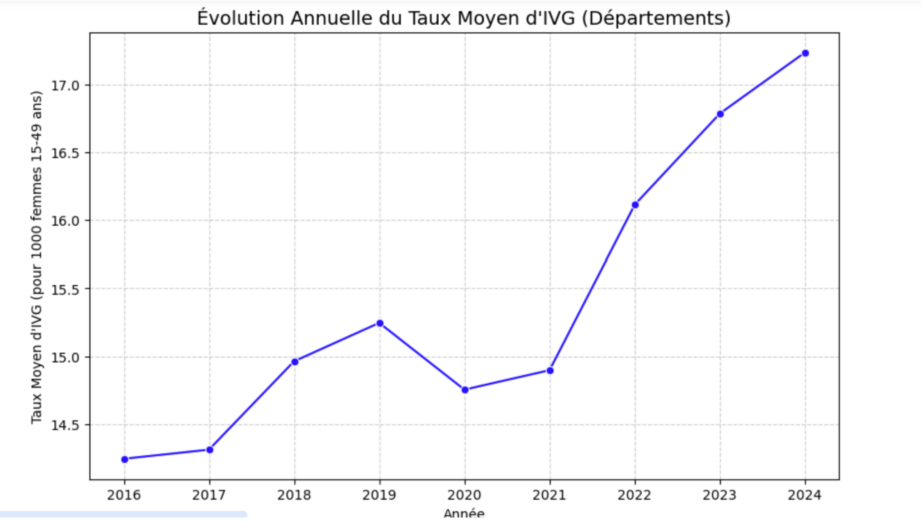
[“Total”, “France as a whole”, “Residence unknown”, “Residence abroad”, “Auvergne”, “Burgundy”, “Brittany”, “Centre”, “Corsica”, “Grand Est”, ‘Guadeloupe’, “French Guiana”, “Hauts-de-France”, “Île-de-France,” “La Réunion,” “Martinique,” “Mayotte,” “Normandy,” “Nouvelle-Aquitaine,” “Occitanie,” “Pays de la Loire,” “Provence”]

3- Data analysis by visualization methods

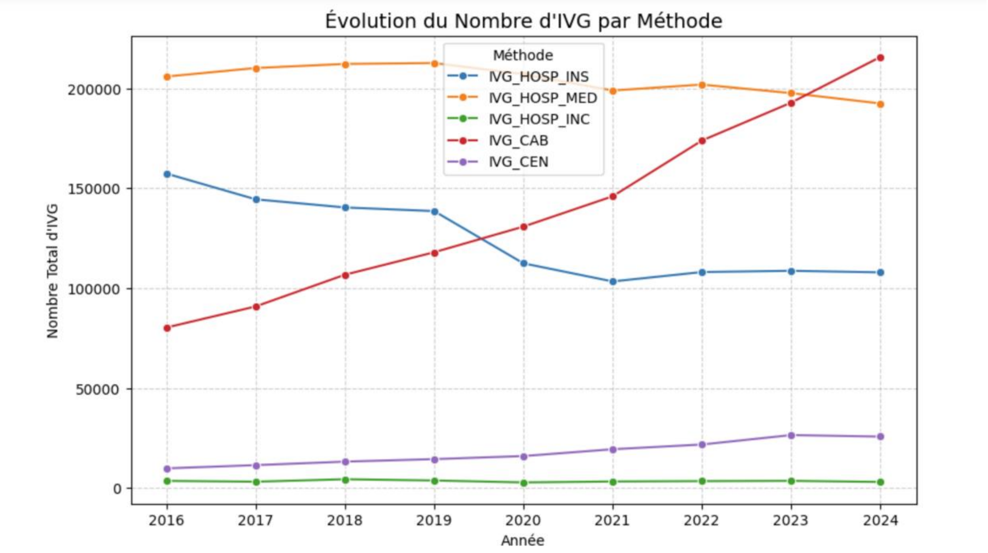
In order to fully understand our dataset, the changing trends in parameters, the weight of certain abortion methods, etc., we have created numerous graphs. Here are a few relevant examples that have helped us interpret our dataset.

Here, the year-on-year change in the average abortion rate by department shows us that the trend is rising sharply. This graph also shows us a slowdown in the growth of the rate from 2007 to 2010, followed by a sharp increase in 2011.

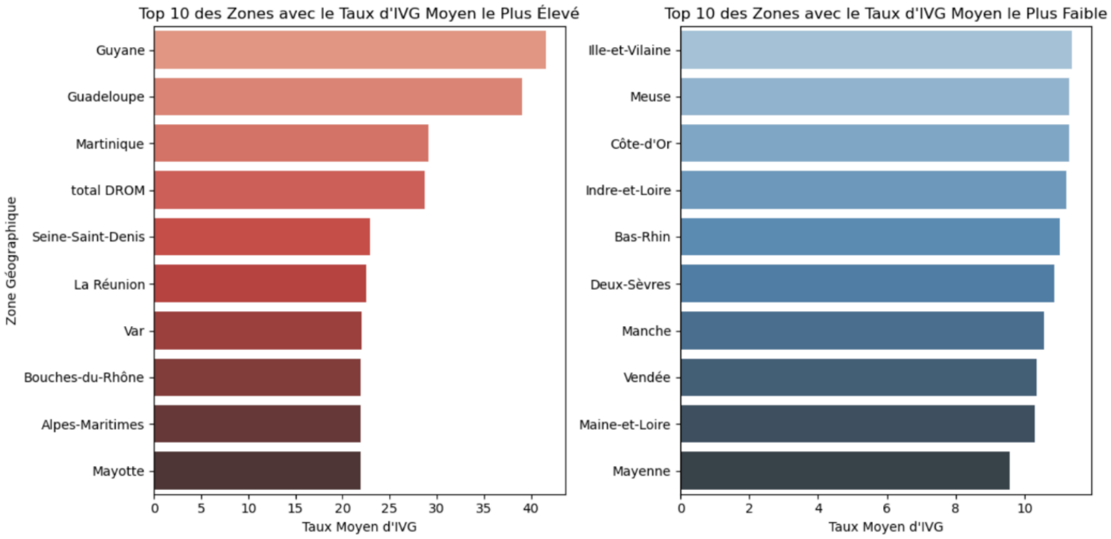
Here, the year-on-year change in the average abortion rate by department shows us that the trend is rising sharply. This graph also allows us to see a slowdown in the growth of the rate from 2019 to 2021, potentially linked to the COVID-19 health crisis.



This second graph shows us how abortion methods have changed over the years. We can see that the most commonly used methods are medical abortion in hospitals and abortion in private clinics, which is experiencing a sharp increase and will become the most commonly used method for abortion by 2024.



Finally, this graph illustrates the 10 areas with the highest abortion rates and the 10 areas with the lowest abortion rates. This allows us to see a predominance in the French overseas departments and territories and in Seine-Saint-Denis. This graph therefore gives us an idea of the geographical areas that have a strong impact and those where the data is more significant.



All of our charts are available in our interim report and on Git Hub.

4- Baseline implementation

1. Baseline Linear Regression

To establish a baseline, we used a simple linear regression model trained solely on the years prior to 2024. This model does not attempt to capture complex relationships: it only estimates the departmental abortion rate based on the geographical area (encoded in indicator variables) and the year. The goal is not to achieve high performance, but to have a minimal model against which to compare more advanced techniques.

We evaluate this baseline using RMSE (Root Mean Squared Error), a metric that heavily penalizes large errors and measures the average difference between predicted and observed values. If more sophisticated models do not achieve an RMSE lower than that of this baseline, it means that they do not add any value. This step is essential to validate the value of using non-linear or ensemble models (Random Forest, Gradient Boosting, etc.).

Results:

The RMSE of the reference model on the 2024 data is: 2.0493.

Therefore, in order to use a more complex regression model, its result must have an RMSE lower than our reference model.

2.Baseline Classification

Our first reference model is a logistic regression, chosen for its simplicity and ability to serve as a minimal point of comparison. It uses only the year and geographic area (encoded as indicator variables). The goal is not to capture complex relationships, but to evaluate whether more advanced models justify their complexity.

We also evaluate an even simpler baseline: a majority DummyClassifier, which systematically predicts the most frequent class. This model represents the minimum performance threshold that a useful classifier must exceed. The comparison between the logistic regression and the majority baseline allows us to verify whether the model actually learns a structure in the data, or whether it merely reproduces the distribution of classes.

Results:

The baseline classification model performs well and outperforms the majority baseline. It detects “High” and ‘Low’ rates well but has more difficulty with the “Medium” class. These results suggest that more advanced models (Random Forest, XGBoost, SVM, etc.) could better capture the relationships between variables and improve classification.

5-Selection and testing of classification and linear regression models

We decided to select and evaluate certain linear regression models for predicting abortion rates and certain classification models for categorising departments according to their abortion rates.

1- Linear Regression Model

To estimate the abortion rate (TAUX_rec_clean) based on geographical area (ZONE_GEO) and year (annee), we tested the following models:

- Baseline Linear Regression: simple regression model
- Ridge Regression: Linear model with L2 regularisation, used to stabilise the model and manage multicollinearity by preventing the coefficients from becoming too large.
- LASSO Regression: Linear model with L1 regularisation, used for automatic feature selection by forcing irrelevant coefficients to zero.
- Random Forest Regression: Non-linear model tested to determine whether a more complex method could significantly improve performance.

To evaluate the performance of a regression, the metric used is Root Mean Square Error (RMSE), as it penalises large prediction errors and is expressed in the same unit as the target variable (the abortion rate).

The models were trained on the cleaned data (df_departements) and evaluated, yielding the following RMSE results for prediction:

- Baseline Linear Regression: 2.05
- Ridge Regression: 2.6196
- LASSO Regression: 3.2416
- Random Forest Regressor: 1.4787

We note that the linear models (Ridge and Lasso) degraded the RMSE compared to the implicit linear baseline. This suggests that the relationships between geographical area and year are non-linear in nature and cannot be effectively modelled using linear approaches.

We also note that the Random Forest Regressor method achieved the best score with an RMSE of 1.4787, which is much better than the RMSE of the linear baseline. This improvement indicates that non-linear models are more appropriate for modelling the complexity of geographical and temporal factors influencing the abortion rate

2- Classification Model

To categorise departments according to their abortion rate or the most commonly used method, we tested the following models:

- Majority Baseline: This model predicts the most frequent class. It serves as a minimum performance threshold: any useful model must have an accuracy higher than this baseline.
- Logistic Regression: A simple linear model used as a first approach to multi-class classification.
- Decision Tree: A non-linear model that makes decisions based on characteristics. Although it is simple and interpretable, it is often very sensitive to overfitting when used alone.
- Random Forest Classifier: A more advanced non-linear ensemble model, designed to improve classification.

To evaluate classification, the main metric used is accuracy, supplemented by the F1-score and the Confusion Matrix.

The accuracy results for each test are as follows:

- Majority Baseline: 0.4583
- Logistic Regression Baseline: 0.6250
- Decision Tree: 0.8333
- Random Forest Classifier: 0.8333

We observe that linear regression achieves an accuracy of 0.6250, which is higher than the majority accuracy of 0.4583. This result validates that the model has successfully identified a structure in the data to classify abortion rates, rather than simply reproducing the distribution of classes. The analysis revealed its limitations: although it correctly identifies the extreme classes (“High” and “Low”), it shows a notable weakness in classifying the intermediate class “Medium”.

6-Exploitation of trains models

Step 5 allowed us to identify the Random Forest Regressor as the best-performing model for predicting `TAUX_rec`, with an RMSE of 1.4787. The purpose of using this model is to generate predictions. First, we chose to present the three departments with the highest predicted IVG rates for 2025, along with the distribution of methods used by geographic area. For clarity, detailed results are shown for a single department, Finistère, while data for other departments are available on GitHub.

When displaying these results, we encountered some difficulties, which led us to adapt our model.

1. Limits of the initial classification approach

Initially, we considered a classification approach to predict the dominant IVG method in a department. This strategy proved insufficient for two main reasons:

- **Lack of quantification:** Classification only predicts the label of the most frequent method (e.g., `IVG_HOSP_MED`). It does not provide percentages or proportions. Logistic planning requires knowing whether a method accounts for 30%, 50%, or 90% of the procedures.
- **Unobserved classes:** Some methods, such as `IVG_CEN`, were never dominant in the training data. The classifier could not predict them, even if they represented a significant share of the procedures. This would have led to incomplete and misleading results.

2. Moving to multi-output regression

To obtain complete and quantifiable predictions, we reformulated the problem as five simultaneous regression tasks. The target variable was transformed into a vector of five proportions, each representing the share of a given method in the total IVGs (e.g., `PROP_MED` or `PROP_CAB`).

The Random Forest Regressor proved particularly well-suited for this task. It is configured to predict all five values at the same time using the same explanatory variables, such as geographic area and year. The model combines predictions from many estimators to minimize the overall error across all five proportions, ensuring reliable and consistent estimates.

This approach has several advantages:

- **Completeness:** Every method, including the less frequent ones, is included, giving a full 100% view of IVG distribution.
- **Operational usefulness:** The predicted proportions can be directly used by health services to plan staff, equipment, or medication needs, such as the amount of Mifepristone required for medical IVGs.

In this study, the Random Forest is used in two complementary ways: its simple version predicts `TAUX_rec`, and its multi-output version estimates the proportions of each method. This strategy maximizes the usefulness of predictions and supports planning and management of health resources.

7-Results and Analysis

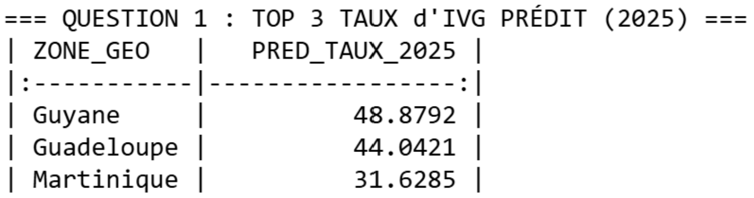
This section presents the performance of the Random Forest models and interprets the prediction results for 2025.

1. Performance of the Rate Model: Choice Validation

The RMSE of 1.3389 confirms the model's ability to capture the complexity of geographic and temporal factors. The average prediction error of the rate is only 1.34 points, which is considered a robust performance for forecasting socio-demographic indicators.

2. Question 1: Identification of High-Rate Areas (Top 3)

Using the simple regression model on the 2025 data, we identified the areas where the predicted IVG rates are the highest.



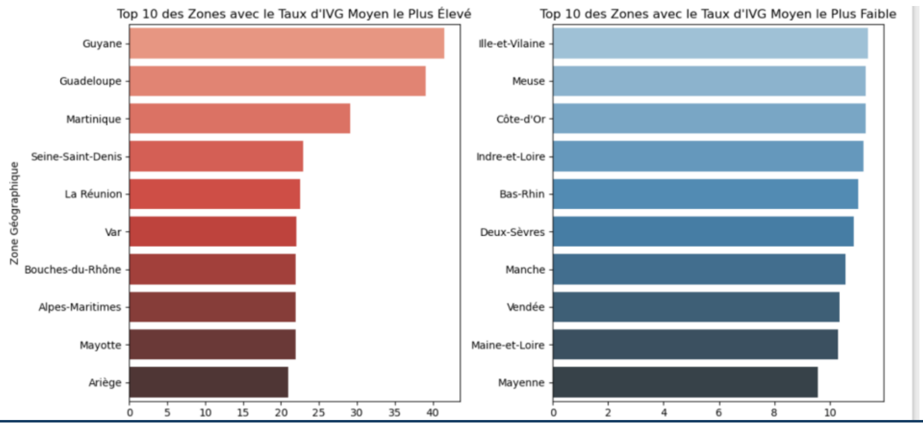
```
=== QUESTION 1 : TOP 3 TAUX d'IVG PRÉDIT (2025) ===
| ZONE_GEO   |   PRED_TAUX_2025 |
|:-----------|-----------------:|
| Guyane     |          48.8792 |
| Guadeloupe |          44.0421 |
| Martinique |          31.6285 |
```

Geographical Analysis:

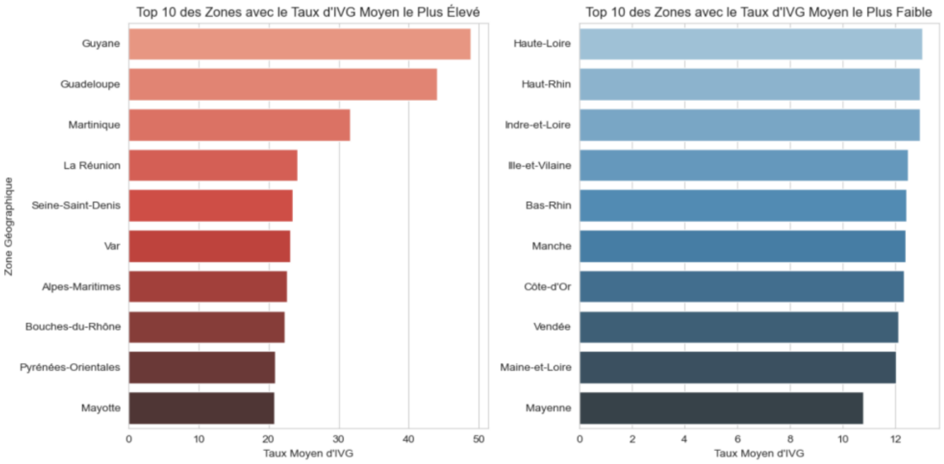
These results confirm the strong geographic disparities in access to and use of IVG in France. Guyane, Guadeloupe, and Martinique continue to have the highest rates, a trend that may be linked to socio-economic factors and challenges in accessing contraception. According to our prediction model, the rates in these departments are expected to continue increasing.

For departments with the lowest rates, Ille-et-Vilaine has risen by four places, showing a higher predicted rate. This highlights that the IVG rate is increasing in many French departments.

2024:



2025:



3. Question 2: Distribution of Methods (Finistère, 2025)

The Multi-Output model provides a detailed view of IVG practices in Finistère for 2025, which is essential for managing hospital resources and inventory, such as staff and medical supplies.

a. *Predicted Overall Rate*

Predicted IVG rate for Finistère: 14.1294

a. *Predicted Distribution of Methods*

Methods	Pourcentage
IVG_HOSP_MED	37.85
IVG_CAB	36.73
IVG_HOSP_INS	20.77
IVG_CEN	4.08
IVG_HOSP_INC	0.58

Trend Analysis:

- **Dominant Methods:** The hospital-based medical IVG (IVG_HOSP_MED) and clinic/cabinet IVG (IVG_CAB) are the two nearly equivalent methods, together accounting for over 74% of all procedures. This reflects the strong shift of IVGs toward outpatient care and private medical practices, in line with evolving practices in France.

- **Validity of Multi-Output:** The model's ability to predict consistent proportions for all five methods, including the less common ones (`IVG_CEN` and `IVG_HOSP_INC`), confirms the effectiveness of the Multi-Output regression approach. This method provides a more comprehensive resource planning tool than a simple binary classifier.

4 Question 3: Measuring the Acceleration of Demand for Hospital Surgery IVG_HOSP_INS

Context and Relevance of the Question

The effectiveness of the predictive model goes beyond forecasting future volumes (TAUX_rec in Question 1). It also helps identify specific areas where pressure on the system is increasing. Surgical abortion requiring full hospitalization (IVG_HOSP_INS) is the most resource-intensive method, as it mobilizes operating rooms, anesthesiology staff, and inpatient beds.

Question 3 aimed to identify the departments where demand for this costly procedure is expected to rise the fastest between 2024 (observed) and 2025 (predicted). By calculating the absolute increase in the IVG_HOSP_INS rate (Rate 2025 – Rate 2024), we highlight the regions where logistical strain is expected to intensify and where urgent preventive planning is needed.

Explanation and Analysis of Results

The analysis of projected growth reveals critical pressure points, mainly concentrated in the Overseas Territories (DOM).

ZONE_GEO	Δ Chirugical rate (2025 vs 2024)	Predictive INS Rate 2025	Observed INS Rate 2024
Mayotte	2.2377	3.2475	1.0098
Guyane	0.4485	4.1292	3.6807
Corse-du-Sud	0.4251	3.1402	2.7151

The Critical Case of Mayotte

a. The situation in Mayotte is the most alarming. The department shows a projected increase in the surgical abortion rate of +2.24 points in a single year. This represents a potential tripling of the rate observed in 2024 (from 1.01 to 3.25).

b. This sharp acceleration—far higher than in any other department—indicates extreme logistical pressure on the island's hospital infrastructure. It may reflect rapid demographic inflow, a breakdown in primary care pathways, or a severe lack of alternatives (medical or outpatient procedures).

Confirmation of Tensions in the Overseas Territories (French Guiana)

French Guiana, already identified for its high overall number of abortions (Question 1), ranks second. Its increase of +0.45 point further strains a healthcare system that is already known to be fragile.

The growth in resource-intensive procedures is rising faster than the national average in these island/overseas territories.

Emergence of Corse-du-Sud

Corse-du-Sud appears in third place with a notable increase of +0.43 point. Although it is located in mainland France, this trend is likely linked to the limited and highly concentrated availability of healthcare services. Even a small rise in demand can quickly overwhelm the few hospitals that have surgical capacity.

Strategic Conclusion

These results show that the hotspots for rising surgical abortion demand are located in regions with healthcare systems undergoing major demographic shifts (Mayotte, French Guiana) or with structurally limited capacity (Corse-du-Sud). Priority action should focus on these areas to stabilize surgical care pathways and rapidly expand non-invasive alternatives (outpatient medical abortion) to prevent an access-to-care crisis.

Model Performance Summary

Regarding model performance:

The Random Forest Regression model achieved an RMSE of 1.5353 when evaluated on 2024. This means the average prediction error for the TAUX_rec in a given department is about 1.54 abortions per 1,000 women of reproductive age. This is considered a low margin of error, confirming the reliability of the 2025 projections used in the logistical analysis.

5 Question 4 : Analysis of model errors by department

Context and Objective

Even though the Random Forest model performs very well overall (RMSE ≈ 1.54 in 2024), it is essential to understand how these errors are distributed geographically. Two departments may have similar rates, but internal variability or an irregular history can make prediction more or less difficult.

The objective of this question is therefore to measure, for each department:

- The average error of the model,
- The average absolute error, which indicates in which areas the model encounters the most difficulties.

This analysis reinforces the understanding of the model's limitations and identifies the departments where future projections should be interpreted with caution.

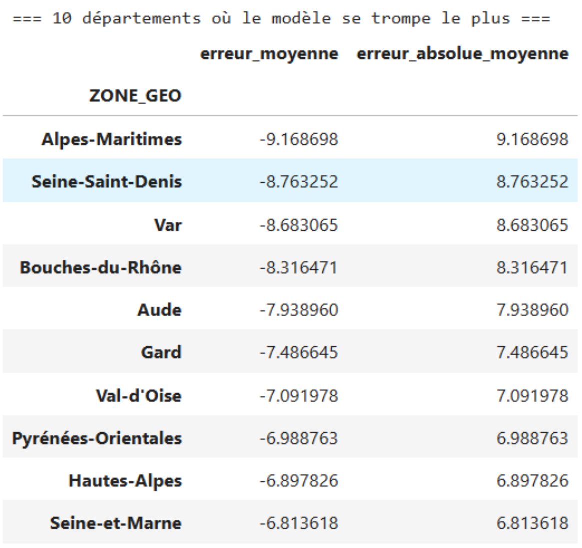
Methodology

We evaluated the model using actual data from 2024, calculating the following for each department:

- The error
- The absolute error, which is more relevant for comparative analysis

The average absolute error was then aggregated by department to obtain a stable and interpretable measure.

Departments that are most difficult to predict



```
=== 10 départements où le modèle se trompe le plus ===
```

	erreur_moyenne	erreur_absolue_moyenne
ZONE_GEO		
Alpes-Maritimes	-9.168698	9.168698
Seine-Saint-Denis	-8.763252	8.763252
Var	-8.683065	8.683065
Bouches-du-Rhône	-8.316471	8.316471
Aude	-7.938960	7.938960
Gard	-7.486645	7.486645
Val-d'Oise	-7.091978	7.091978
Pyrénées-Orientales	-6.988763	6.988763
Hautes-Alpes	-6.897826	6.897826
Seine-et-Marne	-6.813618	6.813618

This list highlights several interesting phenomena:

1- A strong presence of highly urbanised or tourist-oriented departments

Hauts-de-France and Île-de-France (e.g. Seine-Saint-Denis, Val-d'Oise) as well as the PACA region (Alpes-Maritimes, Var, Bouches-du-Rhône) feature heavily in the top 10. These are areas with strong demographic growth, characterised by:

- High population mobility
- Seasonal variations (tourism)
- Significant regional inequalities
- Heterogeneous socio-health contexts.

These factors can lead to significant variability in abortion rates, which is difficult to capture using a model based solely on geography and year.

2- Departments where the model systematically underestimates the rate

The errors are all negative, which means that the model predicts a rate lower than the observed rate. These departments are likely to experience a faster increase in the abortion rate in 2024 than historical trends would suggest. This may reflect:

- Recent changes in access to healthcare
- Socio-economic transformations
- Changes in local medical practices
- Migration dynamics (Seine-Saint-Denis, Alpes-Maritimes).

3- PACA: the most difficult region for the model

Alpes-Maritimes, Var and Bouches-du-Rhône occupy three of the top four places. This region is known for:

- A complex demographic (young people + elderly people)
- Highly variable use of healthcare depending on the area
- Significant seasonal flows

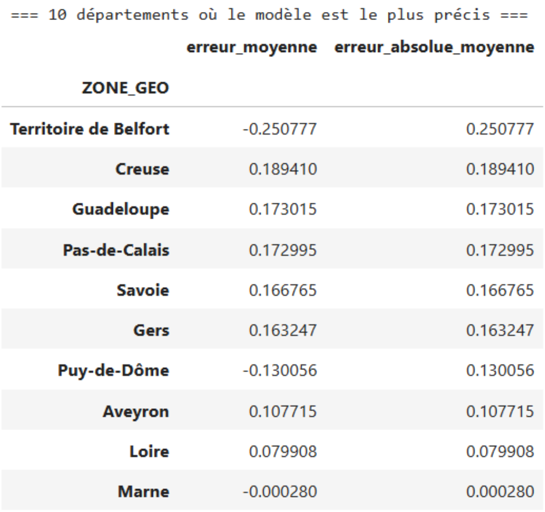
This context makes the historical trend less stable, which complicates prediction.

4- Presence of rural departments

The Hautes-Alpes and Aude departments have low populations, which means that even the smallest variation in volume results in a very large variation in rates.

This automatically causes higher errors for the model.

Departments where the model is most accurate



=== 10 départements où le modèle est le plus précis ===

	erreur_moyenne	erreur_absolue_moyenne
ZONE_GEO		
Territoire de Belfort	-0.250777	0.250777
Creuse	0.189410	0.189410
Guadeloupe	0.173015	0.173015
Pas-de-Calais	0.172995	0.172995
Savoie	0.166765	0.166765
Gers	0.163247	0.163247
Puy-de-Dôme	-0.130056	0.130056
Aveyron	0.107715	0.107715
Loire	0.079908	0.079908
Marne	-0.000280	0.000280

These results highlight the regions where the dynamics of abortion rate trends are stable, consistent and therefore easier to model.

1- High predictability in medium-sized departments

Territories such as Loire, Marne, Puy-de-Dôme and Aveyron appear to be among the easiest areas to predict. These departments are characterised by:

- A population that is neither too small nor too volatile
- Relatively stable trends in abortion rates
- Stable medical practices from one year to the next.

Their 'average' demographic and health profile facilitates robust modelling.

2- Departments with very low populations (Creuse, Territoire de Belfort)

3- Despite their low populations, Creuse and Territoire de Belfort appear in the top 10. This can be explained by:

- A high degree of stability in the abortion rate
- Very consistent medical practice over time
- Few disruptive events between 2016 and 2024.

Their trajectory is simple and linear, which makes predictions very reliable.

4- A notable absence of large metropolitan departments

Unlike the top 10 departments with the highest errors (Alpes-Maritimes, Seine-Saint-Denis, Var, etc.), no highly urbanised departments appear here. This confirms that:

- Large urban areas experience rapid and irregular dynamics, resulting in less accurate models.
- More homogeneous departments display temporal stability, resulting in more effective models.

The more socially or demographically complex a territory is, the more difficult it becomes to model it using simple variables.

Conclusion

This analysis highlights the departments for which the model is robust and predictive and those where factors not included in the data could improve accuracy (age of the population, access to healthcare, socio-economic indicators, etc.).

It reinforces the relevance of our projections for 2025 and provides a decision-making tool for identifying areas of uncertainty where additional checks or data would be particularly useful.

Conclusion

The aim of this project was to develop a forecasting tool to anticipate changes in the abortion rate in France and analyse the distribution of care methods by 2025. Using publicly available data from 2016 to 2024, we cleaned, structured and standardised the data, then created summary indicators to compare departments and effectively model trends over time.

The use of the Random Forest Regressor model, which was selected after comparing it with other models, proved to be particularly relevant. The model offers good overall performance with an RMSE of around 1.54 for the year 2024, which validates its ability to capture historical dynamics and produce consistent projections for 2025. In addition to overall performance, error analysis allowed us to reflect on the geographical dimension of the predictions.

We have highlighted that departments with high social or demographic complexity, such as several areas in Île-de-France or the PACA region, are the most difficult to predict. Conversely, medium-sized departments, as well as those with historically stable trends, show extremely low error rates. These contrasts highlight the need to take territorial heterogeneity into account when interpreting projections and reinforce the added value of department-level analysis.

Projections for 2025 also show a notable change in abortion care practices, with an expected increase in the proportion of medical abortions and a gradual decline in purely surgical procedures in hospitals. This observation is consistent with changes in the regulatory framework, improved access to abortion in urban areas and the rise of multidisciplinary centres.

Despite the quality of the results obtained, certain limitations must be acknowledged. The model is based essentially on two explanatory variables (year and department), which does not allow for the integration of socio-economic, demographic or structural dimensions that are nevertheless decisive in reproductive behaviour and access to healthcare. The addition of indicators such as average age, precariousness, medical density or health infrastructure could significantly improve the accuracy of the model, particularly in departments that are currently difficult to predict.

In conclusion, this project demonstrates the feasibility and relevance of a machine learning approach to anticipating changes in abortion rates in France and analysing how they are managed. It provides a solid basis for a decision-making support system in a context where adapting healthcare resources and ensuring regional accessibility are major challenges. The results obtained provide a detailed understanding of regional dynamics and are fully in line with a data-driven public health approach.

#Sources:

# Data.drees.solidarites-sante.gouv.fr

# Système national des données de santé (SNDS) ; traitements Drees