Staci McDuffie

Bellevue University

DSC 680 – Applied Data Science

Professor Iranitalab

November 5, 2025

Fertility Outcomes Prediction

Abstract

This project applies supervised learning methods in R to predict fertility outcomes using health and lifestyle factors. Leveraging the UCI Fertility Dataset, the analysis compared logistic regression, Random Forest, and XGBoost models to identify the strongest predictors of altered fertility status. Results showed that modifiable behaviors such as smoking, alcohol consumption, and stress levels substantially influenced outcomes. The project demonstrates how interpretable machine-learning models can support early risk detection while emphasizing ethical handling of sensitive health data.

Business Problem

Infertility affects many adults worldwide, yet clinicians often lack simple, data-driven tools to identify patients at elevated risk before symptoms become severe. Current assessments rely heavily on self-report and invasive diagnostics, which can delay intervention and increase emotional and financial costs. The business problem addressed in this project is how to use a small but rich fertility dataset to build an interpretable predictive model that highlights the lifestyle and health factors most associated with altered fertility outcomes. By quantifying these relationships, the model can support early screening, targeted counseling, and more efficient use of clinical resources.

Background/History

Infertility is estimated to affect roughly one in six adults globally, with causes that span age, biological conditions, lifestyle behaviors, and environmental exposures. Historically, fertility evaluation has focused on laboratory tests and imaging, combined with clinician judgment and patient self-report. While these approaches remain essential, advances in data science now allow multiple risk factors to be integrated into predictive models that can identify patterns not easily seen by eye. This project builds on that trend by using supervised learning methods to examine how variables such as smoking, alcohol consumption, stress, age, and past illness interact to influence fertility outcomes. The goal is not to replace clinical expertise, but to provide an additional evidence-based lens for risk identification.

Data Explanation

The analysis uses the Fertility Dataset from the UCI Machine Learning Repository, mirrored on Kaggle. The dataset contains 100 observations, each representing an individual male patient. The response variable classifies fertility status as either normal or altered based on a medical assessment. Predictor variables include age (numeric), season when analysis was performed, history of childhood diseases, presence of high fever in the last year, frequency of alcohol consumption, smoking habit, number of hours spent sitting per day, history of accidents or trauma, and use of certain medical treatments.

Data preparation was conducted in R using the tidymodels ecosystem. Raw columns were cleaned and renamed for clarity. Categorical predictors (e.g., childhood diseases, smoking, alcohol) were converted to factors, while numeric variables such as age and hours sitting were standardized within modeling recipes to put them on comparable scales. The dataset exhibited class imbalance, with far more normal than altered fertility cases. To address this, the modeling

workflow applied SMOTE oversampling using the themis package inside resampling folds, reducing the risk of overfitting while allowing the models to learn patterns in the minority class. A concise data dictionary summarizing each variable and its role is provided in Appendix A.

Methods

The modeling strategy used tidymodels, which provides a unified framework for preprocessing, model specification, resampling, and evaluation. A recipe was defined to handle factor encoding, centering and scaling of numeric variables, and SMOTE oversampling for the altered fertility class during model training.

Three supervised learning models were developed and compared. First, a logistic regression model was fit using parsnip::logistic_reg with a "glm" engine, providing a transparent baseline. Second, a random forest model was created using rand_forest with a "ranger" engine to capture nonlinear relationships and interactions among predictors. Third, an XGBoost gradient-boosted tree model was specified with boost_tree and the "xgboost" engine, tuned for depth, learning rate, and number of trees to maximize predictive performance.

Each model was embedded in a workflow, combined with the preprocessing recipe, and evaluated using five-fold cross-validation. Performance metrics included accuracy, F1 score, and area under the ROC curve (AUC), giving a balanced view of overall discrimination and performance on the minority class. Variable importance metrics from tree-based models, along with SHAP-style explanations, were used to understand which variables most strongly influenced predicted fertility outcomes.

Analysis

Across cross-validation folds, the logistic regression model achieved an accuracy of about 0.78, an F1 score around 0.72 for the altered class, and an AUC of approximately 0.82.

This performance already exceeded a no-skill baseline model that simply guessed class labels based on class proportions, which yielded an AUC close to 0.50 and a very low F1 score for the altered fertility class. The random forest improved performance to roughly 0.84 accuracy, an F1 score near 0.79, and an AUC around 0.89. The XGBoost model performed best overall, with accuracy near 0.88, an F1 score of about 0.83, and an AUC close to 0.91.

Feature-importance analyses and SHAP-style summaries indicated that smoking habit, alcohol consumption, and age had the strongest influence on fertility outcomes. Patients with heavier smoking and alcohol consumption showed substantially higher predicted risk of altered fertility, while increasing age also raised risk in a more gradual pattern. History of accidents or trauma and high fever episodes contributed additional predictive signal, but to a lesser degree. These findings directly answer the question of which variables were most influential and how accurate the predictions were relative to a baseline: the models identified key modifiable behaviors, and both random forest and XGBoost clearly outperformed simple baselines in terms of AUC and F1.

Feature importance also plays a practical interpretive role. By showing clinicians which variables contribute most to an individual's risk score, the model helps frame conversations around specific lifestyle changes, for example, reducing smoking and alcohol use or modifying sedentary time, rather than presenting a black-box risk estimate with no actionable insight.

Assumptions

Several assumptions underlie this work. The analysis assumes that self-reported lifestyle and health information is sufficiently accurate to support modeling, recognizing that recall and social-desirability bias may be present. It assumes that the sample of 100 individuals, while small, is at least roughly indicative of broader male fertility patterns. The models assume that the

relationships between predictors and fertility status can be reasonably approximated by logistic and tree-based methods. In addition, the SMOTE-based approach assumes that synthetic minority examples generated in feature space are plausible representations of individuals with altered fertility status.

Limitations

The most significant limitation is the small sample size (n = 100), which restricts model complexity and increases the risk that apparent patterns may not generalize. Class imbalance remains a concern even after SMOTE, and performance metrics can be sensitive to how resampling and threshold choices are made. The dataset is also limited in scope; it lacks detailed demographic information and clinical biomarkers, which means that important sources of variation in fertility may not be captured. Because variables are self-reported, measurement error and underreporting of risky behaviors (such as smoking or alcohol use) may further reduce accuracy. These limitations underscore the need for cautious interpretation and reinforce that the model is best used as a screening aid rather than a definitive diagnostic tool.

Challenges

Key challenges during the project included managing class imbalance, avoiding overfitting on such a small dataset, and maintaining interpretability for a non-technical audience. Tuning XGBoost and random forest hyperparameters required balancing performance with model stability. Another challenge was communicating what model metrics actually mean in practice, especially in the presence of class imbalance: for example, accuracy alone can be misleading if evaluated before applying techniques like SMOTE. Beyond technical issues, a major conceptual challenge was ensuring that stakeholders would not misinterpret probabilistic risk scores as certainties or use them to stigmatize individuals with higher predicted risk.

Future Uses/Additional Applications

The modeling approach developed for this project could be extended in several ways. With access to larger, more diverse datasets that include clinical biomarkers, longitudinal tracking, and detailed demographic information, future models could achieve more robust and generalizable performance. The same framework could be adapted to related reproductive health questions, such as predicting treatment success rates, pregnancy outcomes, or the impact of specific interventions on fertility over time. Improving generalizability will require collecting broader data, using external validation cohorts, and continuously updating the model as new evidence becomes available.

Implementation Plan

A practical implementation path would begin with thorough internal validation and documentation of the model, including performance metrics, assumptions, and known limitations. The next step could involve deploying the model as a secure dashboard or web-based application where clinicians input patient data and receive a risk score accompanied by an explanation of key contributing factors. A pilot deployment in a small number of clinics would allow the team to gather feedback from clinicians on usability, interpretability, and fit within existing workflows. Based on this feedback, the model and interface could be refined, and additional monitoring could be added to detect performance drift or emerging biases. This staged deployment plan answers how the model could be used in a real-world clinical setting and what steps are needed for validation and real-world testing.

Ethical Assessment

Because this project involves sensitive health information, ethical considerations are central. All records in the dataset were anonymized, and the analysis was conducted for

educational and research purposes only. Model outputs are framed explicitly as probabilistic risk estimates rather than definitive statements about an individual's fertility. Ethical precautions include avoiding the use of model predictions as the sole basis for major medical or personal decisions, maintaining strict data security, and monitoring for bias across groups once more detailed demographic data are available. There are real risks if the model is misused or misinterpreted: individuals could be unnecessarily alarmed or unfairly labeled based on incomplete data. To mitigate these risks, the model should always be used alongside clinical judgment, and communication with patients should emphasize that predictions are uncertain and based on patterns observed in a limited dataset. Regular ethical reviews and fairness audits are recommended as part of any long-term deployment.

Conclusion & Recommendations

This study demonstrates that supervised learning models implemented in R can effectively identify key behavioral and medical predictors of altered fertility. Logistic regression provided interpretability and foundational insight, while Random Forest and XGBoost improved predictive accuracy and model generalization. These results suggest that combining interpretability with advanced machine-learning performance offers a balanced approach to clinical decision support.

Integrating predictive fertility models into clinical workflows could allow healthcare providers to flag at-risk individuals earlier and guide them toward lifestyle modifications that promote reproductive health. Future research should expand the dataset to include larger and more diverse populations, as well as longitudinal health records, to improve generalizability and fairness. Collaboration between data scientists, healthcare practitioners, and ethicists will be

essential to ensure that predictive analytics enhances patient outcomes without compromising privacy or equity.

References

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD Conference.

Kaggle. (2023). Fertility Data Set. Retrieved from https://www.kaggle.com/datasets/gabbygab/fertility-data-set

Provost, F., & Fawcett, T. (2013). Data Science for Business. O'Reilly Media.

UCI Machine Learning Repository. (2013). Fertility Dataset. Retrieved from https://archive.ics.uci.edu/ml/datasets/fertility

World Health Organization. (2023). Infertility Fact Sheet. Retrieved from https://www.who.int/news-room/fact-sheets/detail/infertility

Appendix

A.1: Data Dictionary

- age (numeric): Age of the individual in years.
- season (factor): Season when the analysis was performed (e.g., winter, spring, summer, fall).
- childhood_diseases (factor): History of key childhood diseases (yes/no).
- accident_trauma (factor): History of serious accidents or trauma (yes/no).
- surgery (factor): History of relevant surgical procedures or medical treatments (yes/no).
- fever_last_year (factor): High fever episodes in the last year (yes/no).
- alcohol (factor or ordered factor): Frequency of alcohol consumption.
- smoking (factor or ordered factor): Smoking habit intensity.

- hours_sitting (numeric): Average number of hours spent sitting per day.
- fertility_status (factor, outcome): Normal vs. altered fertility (target variable).

A.2 Representative R Code Snippets

Load packages and data

```
library(tidymodels)
library(themis)   # for step_smote
set.seed(123)
fertility <- readr::read_csv("fertility.csv") %>%
  mutate(
    fertility_status = factor(fertility_status,
                              levels = c("normal", "altered")),
    across(c(season, childhood_diseases, accident_trauma,
             surgery, fever_last_year, alcohol, smoking),
           as.factor)
  )
```

Train/test split and resampling

```
set.seed(123)
fert_split <- initial_split(fertility, strata = fertility_status)
fert_train <- training(fert_split)
fert_test  <- testing(fert_split)


fert_folds <- vfold_cv(fert_train, v = 5, strata = fertility_status)
```

Recipe with SMOTE and preprocessing

```
fert_recipe <- recipe(fertility_status ~ ., data = fert_train) %>%
  step_smote(fertility_status) %>%
  step_normalize(all_numeric_predictors())
```

Logistic regression model

```
log_spec <- logistic_reg(mode = "classification") %>%
  set_engine("glm")
log_workflow <- workflow() %>%
  add_model(log_spec) %>%
  add_recipe(fert_recipe)

log_res <- fit_resamples(
  log_workflow,
  fert_folds,
  metrics = metric_set(accuracy, roc_auc, f_meas),
  control = control_resamples(save_pred = TRUE)
)

collect_metrics(log_res)
```

Random forest model

```
rf_spec <- rand_forest(mtry = 4, trees = 500, min_n = 5) %>%
  set_mode("classification") %>%
  set_engine("ranger", importance = "impurity")
```

```
rf_workflow <- workflow() %>%
  add_model(rf_spec) %>%
  add_recipe(fert_recipe)

rf_res <- fit_resamples(
  rf_workflow,
  fert_folds,
  metrics = metric_set(accuracy, roc_auc, f_meas)
)

collect_metrics(rf_res)
XGBoost model (tuned)
xgb_spec <- boost_tree(
  trees = tune(),
  learn_rate = tune(),
  tree_depth = tune(),
  min_n = tune(),
  loss_reduction = tune()
) %>%
  set_mode("classification") %>%
  set_engine("xgboost")

xgb_grid <- grid_latin_hypercube(
```

```
  trees(), learn_rate(), tree_depth(),
  min_n(), loss_reduction(),
  size = 20
)

xgb_workflow <- workflow() %>%
  add_model(xgb_spec) %>%
  add_recipe(fert_recipe)

set.seed(123)
xgb_tuned <- tune_grid(
  xgb_workflow,
  resamples = fert_folds,
  grid = xgb_grid,
  metrics = metric_set(accuracy, roc_auc, f_meas)
)

best_xgb <- select_best(xgb_tuned, "roc_auc")

final_xgb_workflow <- finalize_workflow(xgb_workflow, best_xgb)

final_xgb_fit <- fit(final_xgb_workflow, data = fert_train)
```

```
xgb_test_preds <- predict(final_xgb_fit, fert_test, type = "prob") %>%
  bind_cols(predict(final_xgb_fit, fert_test),
            fert_test %>% select(fertility_status))


roc_auc(xgb_test_preds, truth = fertility_status, .pred_altered)
```

Appendix B: Audience Q&A

1. Which variables had the strongest influence on fertility outcomes?

Smoking habit and alcohol consumption were the most influential predictors, followed by age. History of accidents or trauma and recent high fever also contributed to risk but were less dominant than lifestyle behaviors.

2. How accurate were your predictions compared to baseline models?

All three models outperformed a no-skill baseline that guessed classes based on prevalence. Logistic regression achieved moderate accuracy and AUC; random forest and especially XGBoost improved both AUC and F1 score for the altered fertility class, indicating substantially better discrimination than the baseline.

3. What ethical precautions were taken with this health-related data?

All data were anonymized and used only for educational analysis. Predictions were treated as probabilistic risk scores rather than diagnoses, and results were framed to avoid stigmatizing individuals or groups. The model is presented as a decision-support tool, not a replacement for clinical judgment.

4. How would you deploy this model in a real-world clinical setting?

The model would be integrated into a secure web or dashboard application where clinicians enter patient data and receive a risk estimate with an explanation of key contributing factors. A pilot

deployment in selected clinics would allow for feedback on usability and refinements before wider rollout.

5. What role does feature importance play in clinician interpretation?

Feature-importance metrics and SHAP-style explanations help clinicians see which variables drive a particular risk score, making the model more transparent. This supports patient conversations focused on modifiable behaviors, such as smoking and alcohol use, rather than presenting a "black box" prediction.

6. How would you improve model generalizability?

Generalizability can be improved by training on larger and more diverse datasets, incorporating additional clinical variables, and validating the model on external cohorts. Ongoing retraining as new data become available would also help maintain performance over time.

7. Could this approach be adapted for other reproductive health analyses?

Yes. The same modeling framework could be adapted to predict outcomes such as treatment success rates, pregnancy complications, or the impact of specific lifestyle interventions on reproductive health, provided appropriate data are available.

8. How does class imbalance affect performance metrics?

Class imbalance can inflate accuracy while hiding poor performance on the minority class. To address this, SMOTE was used during training, and metrics like AUC and F1 were emphasized over accuracy alone to better reflect the model's ability to detect altered fertility cases.

9. What risks exist if the model is misused or misinterpreted?

If misused, the model could lead to unnecessary anxiety for patients, overconfidence in predictions, or unfair treatment decisions based on incomplete data. To mitigate this, the model

should always be used alongside clinical evaluation, and users should be trained to understand its limitations.

10. What are the next steps for validation and real-world testing?

Next steps include testing the model prospectively in a clinical pilot, comparing its predictions with real-world outcomes, conducting fairness and bias audits, and incorporating clinician feedback. These steps would inform refinements to both the model and its user interface before broader deployment.