This is Roche

Roche Digital Technology - Madrid

Diego Prado - RDT Section Lead Data & Analytics Chapter
Enrique Caño - RDT Product Manager Content Search

This is Roche

Healthcare environment trends

Significant challenges are shaping healthcare system needs

2/3 of diseases are either still not treated adequately, or not treated at all

1/2 the world lacks access to essential health services and the health equity gap is growing

Global population aged over 65 estimated to **triple** by 2050

In the US alone, by 2050, the number of > 50 year old people with at least one chronic disease is estimated to double to **142m**

1. WHO Aging and Health fact sheet 2022
2. Projecting the chronic disease burden among the adult population in the United States using a multi-state population model 2023

Roche at a glance

Who we are and what we do

128 years
founded in Basel in 1896

A leader in healthcare R&D
with CHF 13.2 billion invested in 2023

3 Nobel prizes and **44 Prix Galien**
since 1974

CHF 60.5 billion*
in Roche Group sales in 2024

45 Roche medicines & 90 diagnostics**
on the WHO List of Essential Medicines & Tests

>100,000
dedicated employees worldwide

>24 million people
treated with our medicines in 2024

30 billion tests
conducted with our Diagnostics products in 2024

*Unless otherwise stated, all growth rates and comparisons to the previous year are at constant exchange rates (CER; average rates 2023) and all total figures quoted are reported in CHF.
** Medicines and tests that have either been developed or acquired by Roche

Our purpose

Doing now what patients need next

Roche

Doing now what patients need next

Innovating new treatments

Demonstrating value

Delivering innovation

Enabling patients

Supporting customers

Early Research & Development

Clinical Development

Manufacturing & Distribution

Access & Commercialisation

Customer Support

Roche Digital Technology

Roche Digital Technology Madrid

Roche Digital Technology (RDT) **Madrid**

Founded in 2003

Technology Hub that supports the entire value chain of Roche

More than 167.000 global users

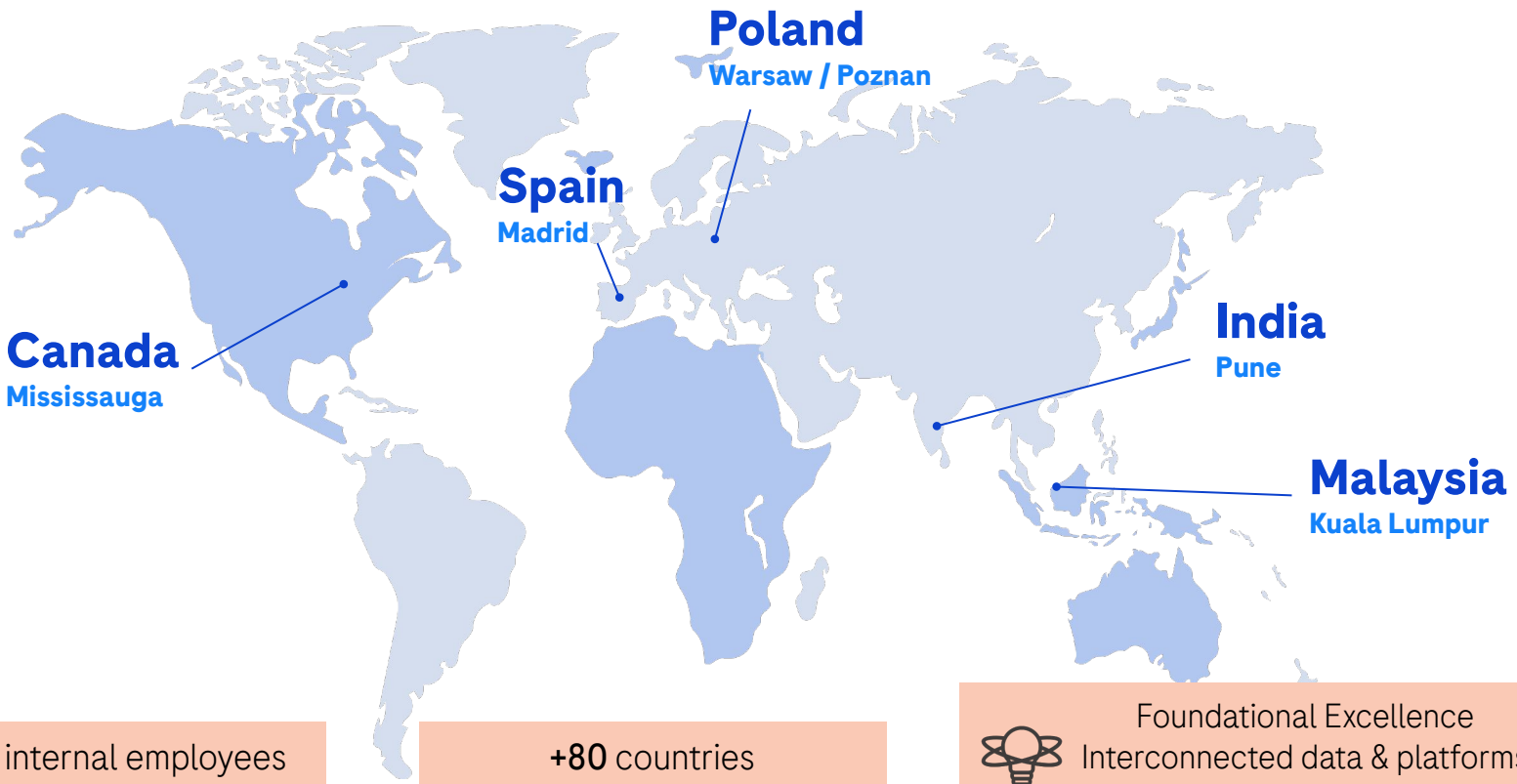
New offices in 2020 - LEED certification

+650 employees
+300 indirect employees

+20 nationalities
39 average age
4 generations

Core areas
- Data & Analytics
- Cybersecurity
- Infrastructure & cloud computing

RDT presence and our Tech Hubs



+4500 internal employees

+80 countries

Foundational Excellence
Interconnected data & platforms
Innovative healthtech

START TECH program

Stories from the frontline: real experiences from our graduates

"The program involved me in innovative, cutting-edge tech projects. My creativity knew no bounds, creating immediate global impact. After the program, I stayed at Roche and I truly expanded my horizons here."

Daniel Hijosa Guzman

DevOps Infrastructure Associate Engineer, Madrid

Now it’s time for action

Show your interest for future opportunities!

Be part of our **START TECH** program for early graduates

Click HERE
to check all the program details

Click HERE
to show your interest for future recruitment*

*currently open at Sant Cugat (Barcelona)!

The Hackathon

Roche Digital Technology Madrid

Roberto Rey Sieiro
Raúl Sánchez Martín

Data & Analytics

Agenda

- Challenge
- Tools
- Recommendations
- Evaluation Criteria
- Awards

Challenge: Diabetes Predictor

Diabetes in numbers

589 m

1 in 9 adults (589 million) suffers from diabetes

Diabetes Atlas 11th Edition | 2025

252m

Over 4 in 10 adults (252 million) with diabetes are unaware of their condition

Diabetes Atlas 11th Edition | 2025

$1 T USD

Global health expenditure has recently surpassed $1 Trillion USD annually

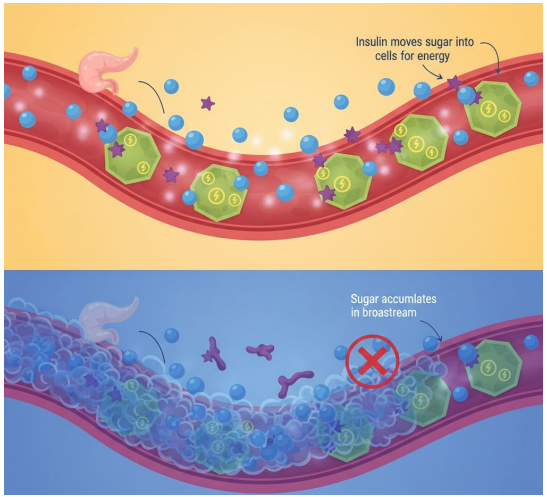
Diabetes Atlas 11th Edition | 2025

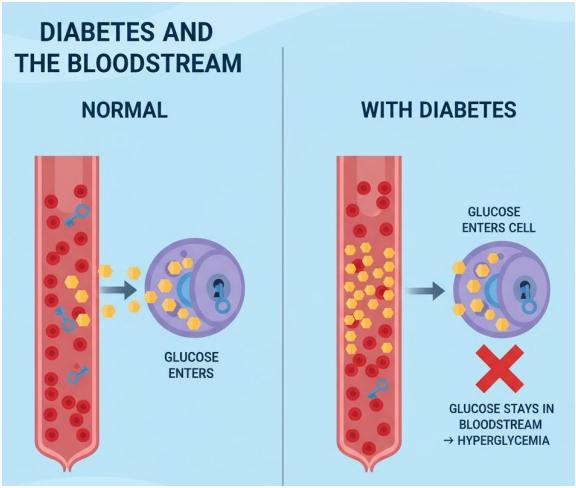
Challenge: Diabetes Predictor

What is Diabetes?

Diabetes is a **chronic metabolic disease** characterized by elevated levels of blood sugar (**hyperglycemia**).

This happens because the body cannot effectively produce or use **insulin**, the hormone required to move sugar (glucose) from the bloodstream into the body's cells for energy





Challenge: Diabetes Predictor

What is Diabetes?

Diabetes is a **chronic metabolic disease** characterized by elevated levels of blood sugar (**hyperglycemia**).

This happens because the body cannot effectively produce or use **insulin**, the hormone required to move sugar (glucose) from the bloodstream into the body's cells for energy

Type 1 (T1D)

Autoimmune destruction of the insulin-producing cells in the pancreas

Absolute lack of insulin

Type 2 (T2D)

Cells become resistant to insulin (Insulin Resistance), and the pancreas cannot produce enough to overcome the resistance

Insulin is ineffective or insufficient to clear glucose from the blood

*Sustained high blood sugar is **toxic**. It progressively **damages the body's vascular system**, leading to severe, often irreversible, complications in the **heart** (heart attacks, strokes), **kidneys** (kidney failure), **eyes** (blindness) or **nerves** (amputations)*

Challenge: Diabetes Predictor

Diagnosis of Diabetes

The diagnosis of diabetes is a **complicated** clinical endeavor because the disease's **silent**, often **asymptomatic** onset allows persistent **hyperglycemia** (high blood sugar) to go **unnoticed**.

Diagnostic Challenge

*To overcome this inherent challenge and confirm persistent hyperglycemia, doctors rely on more than just isolated blood sugar readings. They must **systematically evaluate** a patient's overall risk **profile and metabolic markers** to establish a definitive diagnosis.*

Demographics

- ***Gender***
- ***Age***

Comorbidities

- ***Hypertension***
- ***Heart Disease***

Lifestyle

- ***Smoking History***
- ***BMI***

Markers

- ***HbA1c: % reflects average glucose over 2-3 months***
- ***Random Glucose: mg / dL (snapshot blood sugar reading)***

Challenge: Diabetes Predictor

AI for Clinical Risk Assessment

To finally crack the challenge of silent, asymptomatic onset, **AI-driven software**, leveraging previous patient data, provides a great step forward by automatically analyzing key clinical indicators (gender, age, HbA1c, etc.) and swiftly **augmenting doctors' ability** to identify high-risk individuals.

Medical Note

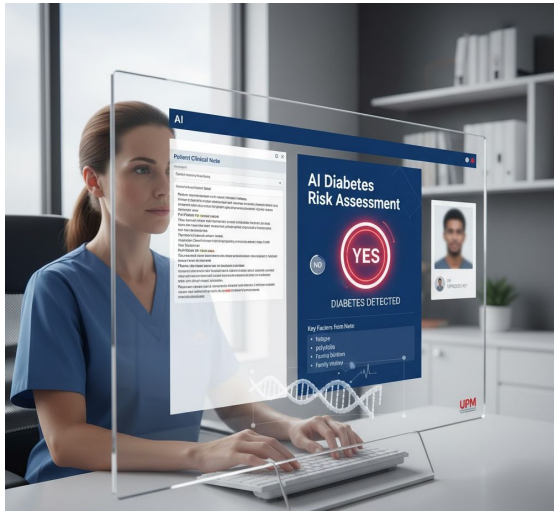
S: 21-year-old male presents for routine evaluation. Reports generally good health, occasional late-night meals, and irregular exercise. No chest pain, dyspnea, or polyuria noted. **Smoking history** is unclear. Family history of **hypertension** in father.

O: BP mildly elevated; **BMI 25.2 kg/m²**. **Random glucose 158 mg/dL**; **HbA1c 5.0 %**. No cardiac murmurs; lungs clear.

A: Elevated random glucose without chronic hyperglycemia on HbA1c. Overweight and hypertensive, placing him at increased metabolic risk.

P: Recommend fasting glucose or OGTT for clarification, blood pressure monitoring, dietary counseling, and regular aerobic activity. Follow-up in 3 months.

AI

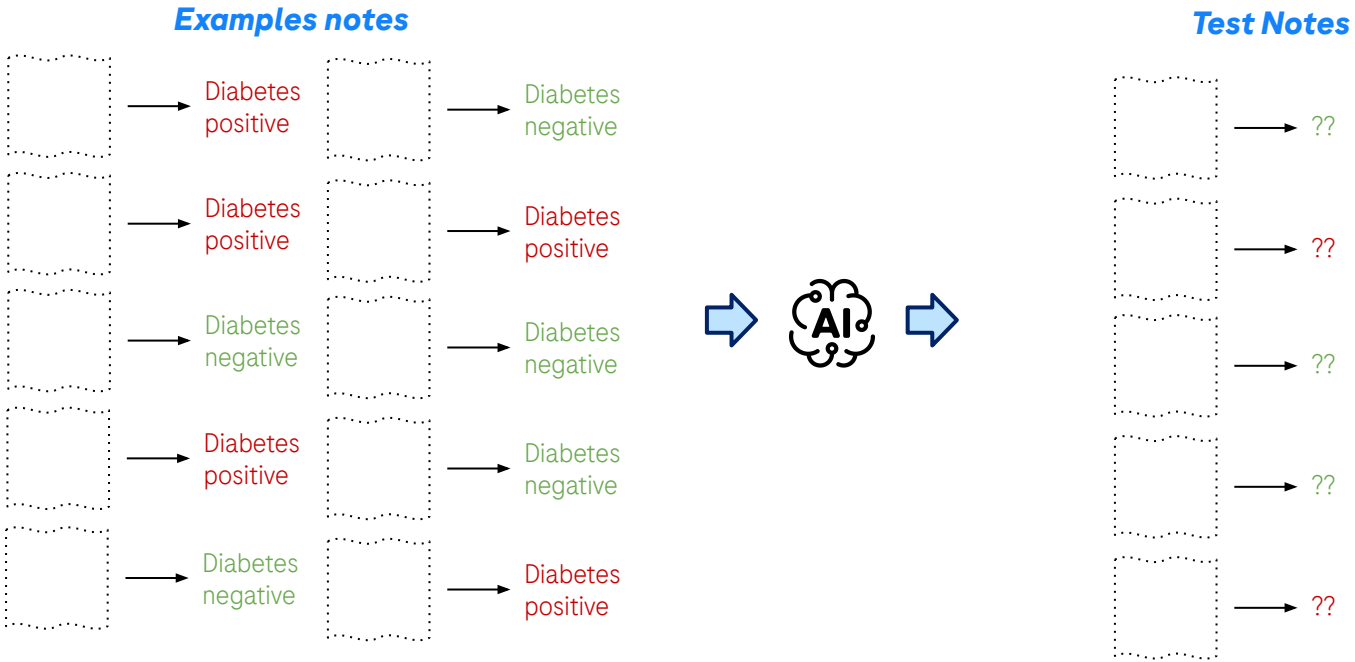


Challenge: Diabetes Predictor

Hackathon Challenge

Code the Cure

Develop an **AI diabetes predictor** that trains on real-world clinical notes and proves its accuracy against new patient data



Tools

Open Source & Commercial Tools

Basic Machine Learning Environments and Libraries			
Category	*Description*	*Open Source Tool*	*Commercial Tool*
Data Manipulation / Cleaning	**Data ingestion (from API)**, feature creation, and data type transformation (for tabular data).	**Pandas** (Link)/ **dplyr** (Link) / **data.table** (Link)	**Microsoft Excel / SQL (Enterprise)** (Excel Link)
ML Algorithms	Establish a robust and interpretable **baseline classification model** (e.g., Logistic Regression, Decision Trees).	**Scikit-learn** (Link) / **R** (Link)	**SAS Viya** (Link)
High-Performance ML	Achieve **maximum predictive accuracy** on structured data using advanced *ensemble* methods.	**XGBoost / LightGBM** (XGBoost Link)	**MATLAB Machine Learning Toolbox** (Link) [1]
Natural Language Processing (NLP)	**Efficient basic text processing**, such as Named Entity Recognition (NER), to extract simple structured variables from clinical notes.	**NLTK (Link) / SpaCy (Link) / tm** (Link) / **Hugging Face Ecosystem** (Link)	**Google Cloud Natural Language API / IBM Watson Discovery (Google NLP Link)**
Statistical Analysis	Rigorous **hypothesis** testing, calculation of confidence intervals, and **statistical explainability** of features.	**Statsmodels** (Link) /**broom** (link)	**IBM SPSS Statistics / Stata** (SPSS Link / Stata Link) [1]
Platforms	**End-to-end environment** for data preparation, model development, and deployment.	**KNIME, H2O.ai**	**Dataiku, Databricks**
Visualization	Presentation of results in **graphical format**.	**Matplotlib, gradio, streamlit, grafana, ggplot2**	**Plotly** (link), **Tableau**

[1] UPM has licenses for these products

[2] One-year free trials for university students in Spain

Tools

Open Source & Commercial Tools

Generative Models and Tools			
Category	*Description*	*Open Source Tool*	*Commercial Tool*
Large Language Models	Feature Extraction (structuring text into variables), Embedding Generation, or Direct Classification through prompting or fine-tuning..	**Raw Models** **LLaMA 3** (Link), **Mistral / Mixtral** (Link), **Gema** (Link) **Model Repository** **Hugging Face** (Link) **Cloud Execution Platforms** **Kaggle** (Link), **Google Colab** (Link) **Local Execution Platforms** -> **Ollama** (Link), **LM Studio** (Link), **GPT4All** (Link) **Free APIs** **Hugging Face** (Link)	**Free Options** **Google AI Studio** (Link, modelo 'gemini-2.5-flash'), **Groq** (Link, 'Developers -> free API key', variedad de modelos) [2] **Paid Options** **OpenAI** (Link) , **Google - Gemini** (Link) , **Claude** (Link), **Cohere** (Link)
Core Frameworks	Construction and training of custom Deep Learning models, or serving as the backend (support) for the Fine-Tuning of LLMs.	**PyTorch** (Link) / **TensorFlow** (Link) / **Keras**	**MATLAB Deep Learning Toolbox** (Link) [1]
Transformers	Access, load, and tokenize pre-trained transformer models (e.g., BERT) to generate text features (embeddings).	**Hugging Face Ecosystem** (Link)	**Amazon SageMaker JumpStart / Google Vertex AI Model Garden** (Vertex AI Link)

[1] UPM has licenses for these products

[2] One-year free trials for university students in Spain

Tools

Open Source & Commercial Tools

IDEs and Development Environments			
Category	*Description*	*Open Source Tool*	*Commercial Tool*
Integrated Development Environment (IDE)	Write, debug, and manage Python/R code locally.	**VS Code** (Core is OS) (Link)/ **Positron** (Link)	**PyCharm Professional** (Link) [2]
Notebooks	Data exploration, visualization, and rapid prototyping by combining code and narrative.	**Jupyter Notebooks / Lab** (Link) / **markdown / quarto** (Link)	**Google Colab (Link) / Databricks Notebooks (Link)**
Version Control	For collaboration, version control, and a reproducible workflow.	**Git** (link) / **Codeberg** (Link)	**GitLab (link) / Azure DevOps (Link) / GitHub (Link)**

[1] UPM has licenses for these products

[2] One-year free trials for university students in Spain

Tools

Recommendation

In case you need to use **Large Language Models** in your solution, we recommend prioritizing free, open-source tools to maximize accessibility and speed. Specifically, utilize online cloud platforms like **Google Colab** (https://colab.research.google.com/) or **Kaggle** (https://www.kaggle.com/) to leverage their free GPU resources. This setup allows you to run **lightweight LLM** models (e.g., 'Qwen/Qwen3-4B-Instruct-2507') found on repositories like **Hugging Face**. You can easily integrate these models and perform local inference (avoiding external API costs) by using the core Python libraries from Hugging Face, such as the transformers library found at: git+https://github.com/huggingface/transformers.git.

*Register in **Kaggle** as soon as possible!*

Evaluation Platform

- Performance evaluation will be done using a **Kaggle Competition**

- Please register in the **competition**:
  https://www.kaggle.com/t/9ae3fe4e1c144f94a699193fabdc1659

- **Leaderboard**:
  50% of test data will be used for **public results**
  50% of test data will be held for **private final evaluation**

- Submissions are evaluated on F1 Score, which provides a balanced measure of precision and recall.
  F1 Score = 2 × (Precision × Recall) / (Precision + Recall)
  Precision = True Positives / (True Positives + False Positives)
  Recall = True Positives / (True Positives + False Negatives)

Evaluation Criteria

The Project: Idea & Execution

⚙️Model Technical Performance (Weight: 40%)

- **Based on Kaggle competition results**
- **Main Metric: F1 Score**. A high F1 Score indicates a well-balanced model.
- **Evaluation with hidden dataset.**
- **Procedure:** Teams will be ranked according to their F1 Score obtained on this hidden dataset.

✅ Process Quality (Weight: 30%)

- **Feature Engineering:** Evaluation of creativity in handling and transforming variables to improve prediction.
- **Modeling:** Selection of the appropriate classification algorithm and optimization of its hyperparameters.
- **Code and Reproducibility:** The code must be clean, documented, and easy to execute.
- **Autonomy:** The team's ability to solve problems independently will be evaluated.

Evaluation Criteria

The Impact: Value & Pitch

🚀 Innovation & Real-World Impact (Weight: 20%)

- **Innovation:** Novelty or creativity in the application of classification techniques or in the data preprocessing approach.
- **Utility and Actionability** (demo): This criterion assesses how well the team transformed the binary predictive model into an usable application by providing clear, context-specific evidence for its prediction, using the clinical notes. Is the project polished, intuitive, and user-friendly? This includes any **Data Visualization**

🎤 Presentation & Pitch (Weight: 10%)

- **Clarity of Presentation:** Fluid and concise explanation of the problem, process, and results, including the interpretation of the F1 Score.
- **Conclusions and *Insights***: Identification of the most important variables and the translation of metrics into **practical recommendations** for the audience or business
- **Q&A:** How well did the team answer questions?

Doing now what patients need next