Cloud Platforms Final Project – Diabetes prediction model in AWS

Section B | Group 6 | April 3rd, 2024

Approach

Problem Identification

- Understanding the **issue** and **necessary requirements**
- Defining a **business goal** that can be **translated into technical (ML model) goals**

Solution Overview

- Defining **data sources**
- Choosing adequate **machine learning model** & **deploying it** on AWS

AWS Architecture Design

- **Architecture diagram**
- **Data flow** between the different AWS tools we are using

Model Evaluation & Fine-Tuning

- Interpreting the **performance metrics** of the different models
- Suggesting **adjustments to parameters**, if necessary

Conclusion

- **Key takeaways** from the project

- Approach
- Problem Identification
- Solution Overview
- AWS Architecture Design
- Model Evaluation & Fine-Tuning
- Conclusion

Helping the Spanish healthcare system save costs

Primary problem/use case

MINISTERIO
DE SANIDAD

- Direct **healthcare costs of diabetic patients** are **very high**[1]
- **High need** for **reducing** the **number of patients suffering from this condition**
- But: **identified early enough, a diabetes prevention program can be initiated**, dramatically reducing the **risk of developing diabetes**

Objective: build a **diabetes prediction model** that helps **significantly reduce financial burden of diabetes** on the Spanish National Health System

Secondary use cases

Insurance companies, such as...

Allianz

AXA

ZURICH

[1] 6.3-7.4% of the Spanish National Health System expenditure

Main requirements of our solution

1
- The model must demonstrate **high predictive accuracy and reliability**
- It must be able to **generalize well** across the diverse Spanish population

2
- The solution should be **scalable** to accommodate the entire target population across Spain
- It should be **easily accessible** by healthcare providers in various settings

3
- It should be as **cost-effective** as possible
- Therefore, **minimize additional burdens on the healthcare system**

Translating our objective into financial impact

Cost in Spanish healthcare system

- Cost **per person**
- over a **3-year period**
- **Elements**
  - baseline cost
  - diabetes prevention program cost (depending on case)
  - diabetes treatment costs (depending on case)

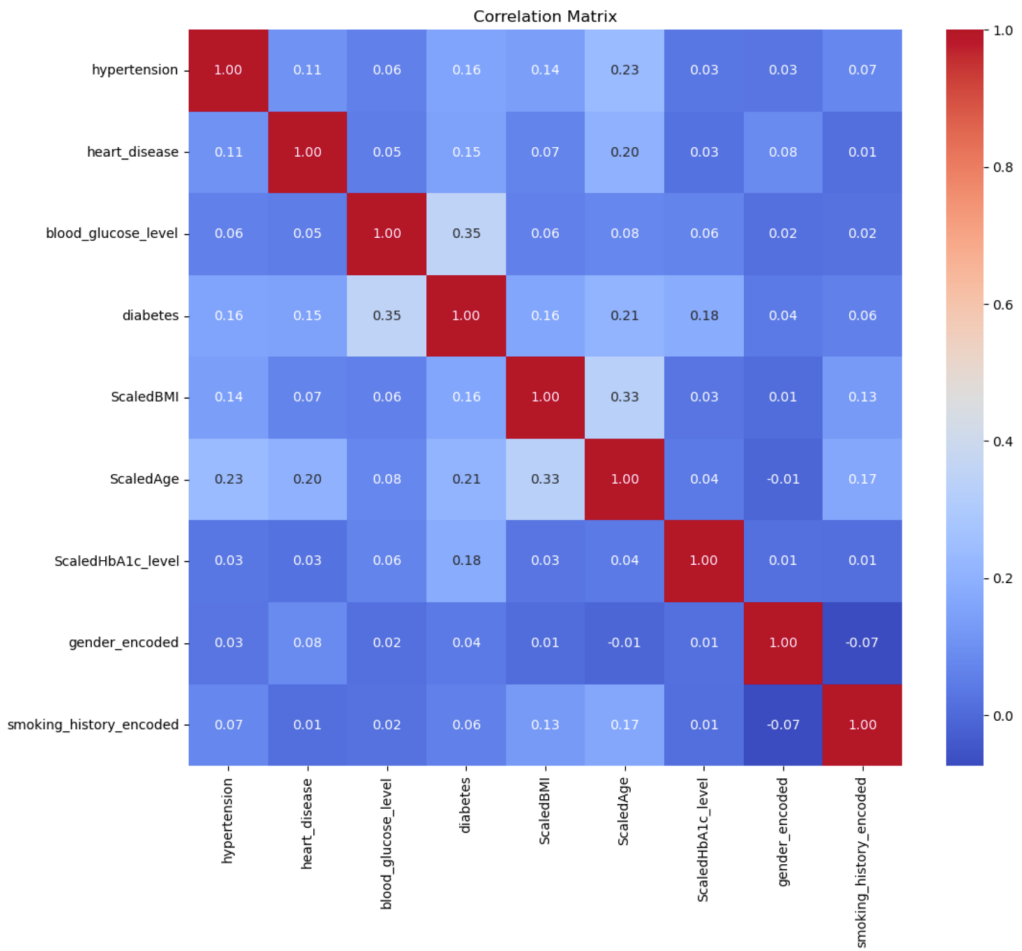
Actual/Detection	Detection "AT RISK"	Detection "NOT AT RISK"
Actual "AT RISK"	- 3,835 €	- 4,522 €
Actual "NOT AT RISK"	- 3,588 €	-3,288 €

Note: To arrive at these numbers, we have employed a calculation model backed by simplifying assumptions. Please find the detailed explanation in the appendix.

Dataset on diabetes and health features

Dataset variables

- **Prediction target:** diabetes (yes/no)
- **Features**
  - hypertension (high blood pressure)
  - heart disease
  - blood glucose level
  - BMI
  - HbA1c level (average blood sugar levels)
  - Gender
  - Smoking history



- Approach
- Problem Identification
- Solution Overview
- AWS Architecture Design
- Model Evaluation & Fine-Tuning
- Conclusion

Main reasons for deploying on AWS

1

- **Scalability of AWS SageMaker** ensures the model can handle the large and variable volumes of health data across Spain

2

- The **pay-as-you-go pricing model** helps manage public health budgets effectively
- This ensures **resources are allocated efficiently** without overspending

3

- AWS provides **robust security measures and compliance capabilities**
- This is **crucial for handling sensitive health data** to comply with regulations

Solution overview

Dataset loaded into S3 bucket

AWS SageMaker notebook

Preprocessing:
- separation into features and target variable; imputation of missing values, scaling of num. variables, one hot encoding of cat. variable; train/val/test split

Model set-up & training (XGBoost):
- launch training job through SageMaker's XGBoost

Tuning of model hyperparameters

Retrieval of model metrics

Model deployment:
- creation of an estimator object by using the best training job
- Deployment to a SageMaker endpoint which can be fed with csv data

Inference generation using the model endpoint

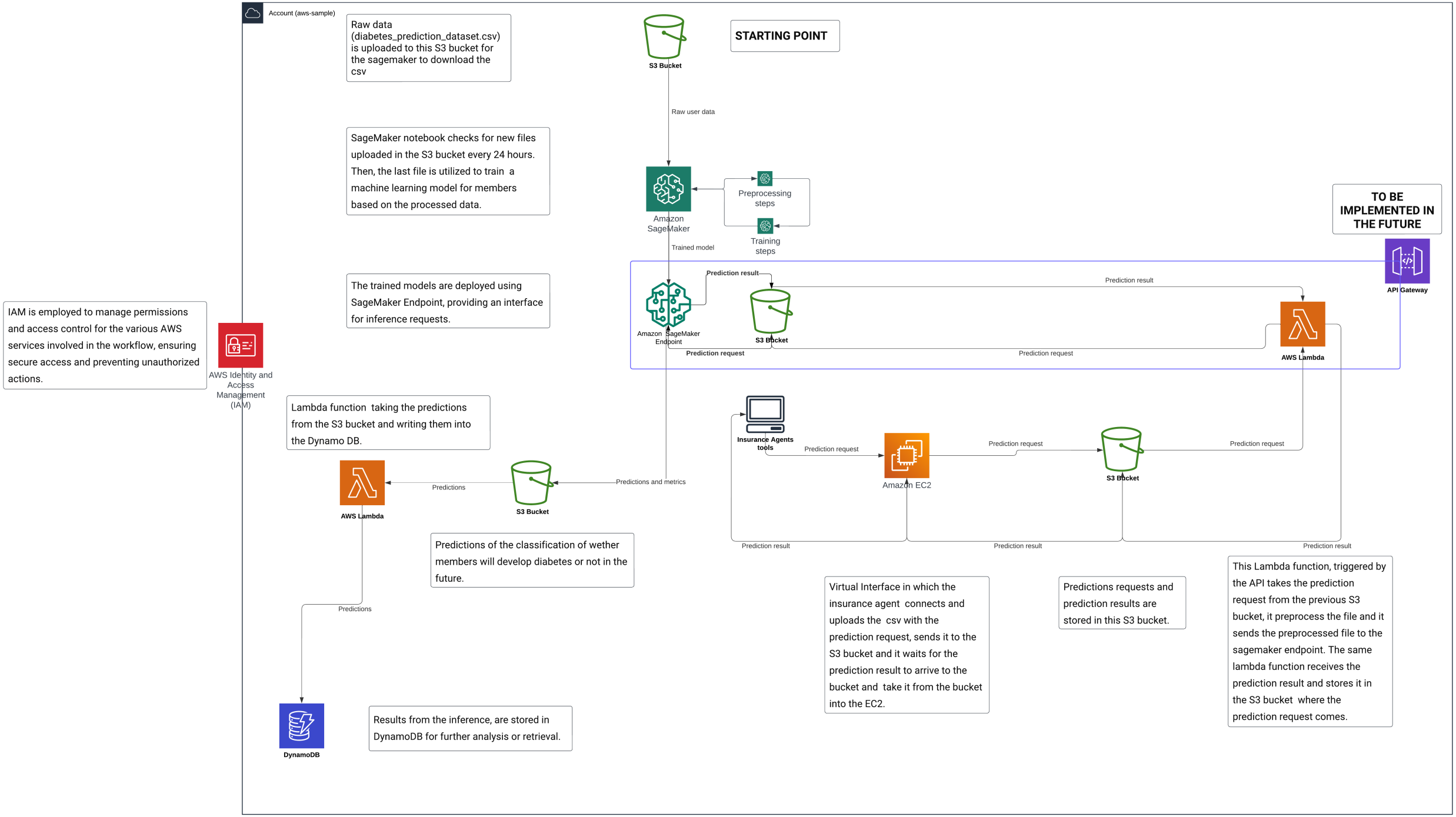
Assessment of generalization performance (against the training set)

Predic-tions loaded into S3 bucket

Lambda function

Dynamo Database to save predictions
- Patient number as identifier
- Storing of prediction (diabetes/ no diabe-tes)

- Approach
- Problem Identification
- Solution Overview
- AWS Architecture Design
- Model Evaluation & Fine-Tuning
- Conclusion


Account (aws-sample)
Raw data (diabetes_prediction_dataset.csv) is uploaded to this S3 bucket for the sagemaker to download the csv
S3 Bucket
STARTING POINT
Raw user data
SageMaker notebook checks for new files uploaded in the S3 bucket every 24 hours. Then, the last file is utilized to train a machine learning model for members based on the processed data.
Amazon SageMaker
Preprocessing steps
Training steps
Trained model
TO BE IMPLEMENTED IN THE FUTURE
API Gateway
The trained models are deployed using SageMaker Endpoint, providing an interface for inference requests.
Prediction result
Amazon SageMaker Endpoint
S3 Bucket
Prediction result
Prediction request
Prediction request
AWS Lambda
IAM is employed to manage permissions and access control for the various AWS services involved in the workflow, ensuring secure access and preventing unauthorized actions.
AWS Identity and Access Management (IAM)
Lambda function taking the predictions from the S3 bucket and writing them into the Dynamo DB.
Insurance Agents tools
Prediction request
Amazon EC2
Prediction request
S3 Bucket
Prediction request
AWS Lambda
Predictions
S3 Bucket
Predictions and metrics
Prediction result
Prediction result
Prediction result
Predictions of the classification of wether members will develop diabetes or not in the future.
Virtual Interface in which the insurance agent connects and uploads the csv with the prediction request, sends it to the S3 bucket and it waits for the prediction result to arrive to the bucket and take it from the bucket into the EC2.
Predictions requests and prediction results are stored in this S3 bucket.
This Lambda function, triggered by the API takes the prediction request from the previous S3 bucket, it preprocess the file and it sends the preprocessed file to the sagemaker endpoint. The same lambda function receives the prediction result and stores it in the S3 bucket where the prediction request comes.
Predictions
DynamoDB
Results from the inference, are stored in DynamoDB for further analysis or retrieval.

Choice of AWS elements

1 S3 buckets

- **Useful** object that can **store** in an **easy** and **intuitive way** the **data**
- Very **useful** to **interact** with **lambda functions**

2 SageMaker

- **Intuitive** and **powerful** terminal to **develop any ML model**
- Enables all steps from **prepro-cessing** to the deployment of the model with the **endpoint interface**

3 Lambda function

- It is the **best tool to trigger information** from other objects and make the information **flow across** the different parts of the **architecture**

4 DynamoDB

- Doing the queries is more **efficient** as application primarily deals with **user profiles and prediction results** & does not require complex joins or transactions
- **Scalability** of this database is a huge advantage

Choice of AWS elements

5 EC2 instance

- Efficient **visual interface** that can be easily **used by third parties** to look at the potential treatments that patients should or should not do

6 API Gateway

- **Server** that **integrates lambda** functions and **http** connections with other AWS services **allowing** for instance **real time inferences**

- Approach
- Problem Identification
- Solution Overview
- AWS Architecture Design
- Model Evaluation & Fine-Tuning
- Conclusion

Model Evaluation & Fine-Tuning

Key metrics before hyperparameter tuning

1
- **train:auc:** 0.978681
- **validation:auc:** 0.976233

2
- **train:logloss:** 0.086373
- **validation:logloss:** 0.091624

3
- **train:error:** 0.027681
- **validation:error:** 0.029700

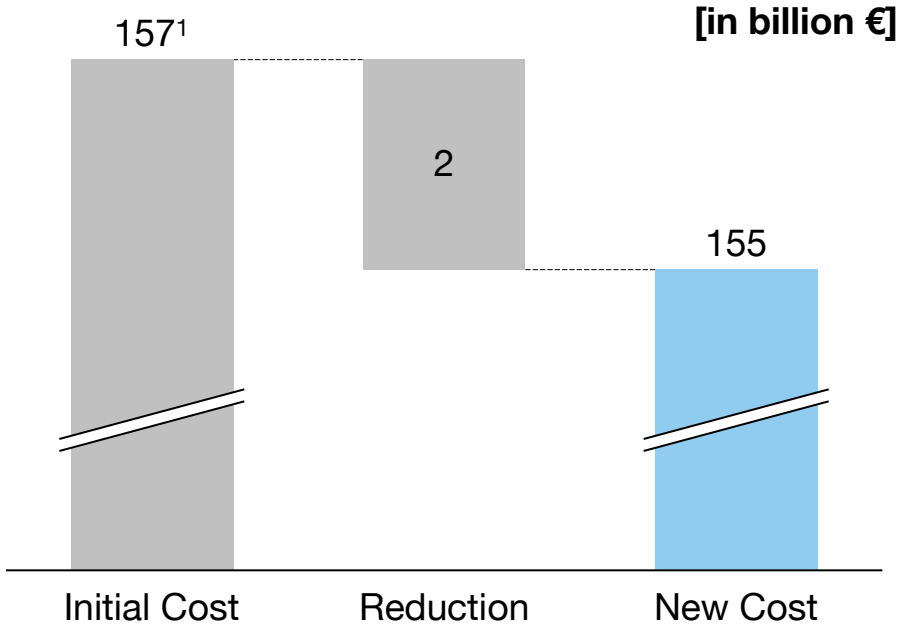
Further suggested improvements

- Overall, these metrics suggest that our **XGBoost model is performing well**
  - no signs of overfitting
  - high AUC values suggest strong classification capability
  - low log loss indicates confidence in predictions
  - low error rates reflect high accuracy
- **Fine-tuning,** which we conducted, only led to **minor performance improvements**
- To achieve **further performance improvements**
  - build separate models for male/female
  - try out different ML models
  - feature engineering

- Approach
- Problem Identification
- Solution Overview
- AWS Architecture Design
- Model Evaluation & Fine-Tuning
- Conclusion

Summary: Business recommendations

Actual/Detection	Detection "AT RISK"	Detection "NOT AT RISK"
Actual "AT RISK"	586	276
Actual "NOT AT RISK"	11	9127



- Generation of **confusion matrix** based on measuring the model's performance against a test set of 10,000 patients
- Results from confusion matrix can be:
  - **combined with financial estimates** per field of confusion matrix
  - **extrapolated** to entirety of Spanish healthcare system (46.5m people)

[1] Based on the assumption that the current percentage of people at risk of developing diabetes would develop diabetes without intervention programs

Summary: Business recommendations

Key Takeaways

- We recommend the Spanish National Health System to deploy our model as specified
- This can enable significant saving potentials as opposed to a normal world without diabetes prevention – the **potential is about EUR 2 bn**
- As we have shown with AWS, the **deployment of such models can be easy and cost-efficient**

Future Considerations

- In the future, it is **key to the success** of the implementation of such a model that **reliable data on all patients is gathered**
    - Doctors and nurses must be sensitized to **check data consistency when recording the data** (e.g., regarding glucose levels)
    - It is **imperative to add the smoking history** from the doctor‘s side as smoking is a feature known to cause diabetes

Appendix

Appendix: Underlying assumptions for quantification

To quantify the financial impact of the diabetes risk prevention, we have made a number of assumptions:

1
- We assume the patients in the data more or less reflect the Spanish population, making it possible to use the data to train a model deployed in the Spanish National Health System

2
- Risk detection will take place based on the data available in our dataset in year 0 (deplyoment of our ML model) for all people part of the Spanish National Health System

3
- If, based on the model, a risk for developing diabetes is detected, the prevention program will take place in year 0
    - In case the prevention program is successful, the respective people will not develop a diabetes condition in the upcoming years
    - In case the prevention program is unsuccessful, these patients will develop diabetes in the following year

4
- In case the risk of developing diabetes is not detected, there is a likelihood of 85% for people to develop diabetes

Appendix: Detailed calculation of the financial impact

	In case of risk detection	Likelihood	Y0	Y1	Y2	Y3	Sum (discounted)[1]
***Detection* at risk & *Actual* at risk**	Program is successful	83%	- 1,165 €	- 865 €	- 865 €	- 865 €	**- 3,835 €** *(weighted for outcome probability)*
	Program is not successful	17%	- 1,165 €	- 1,383 €	- 1.383 €	- 1.383 €	
***Detection* at risk & *Actual* not at risk**	Program is successful	100%	- 1.165 €	- 865 €	- 865 €	- 865 €	**- 3.588 €**
***Detection* not at risk & *Actual* at risk**	Patient develops diabetes	85%	- 865 €	- 1.383 €	- 1.383 €	- 1.383 €	**- 4522 €** *(weighted for outcome probability)*
	Patient does not develop diabetes	15%	- 865 €	- 865 €	- 865 €	- 865 €	
***Detection* not at risk & *Actual* not at risk**			- 865 €	- 865 €	- 865 €	- 865 €	**- 3.460 €**

[1] To more accurately reflect the current value of future payments, we have decided to employ a discount factor of 3.5% in accordance with the current Spanish inflation rate.

Appendix: Financial impact calculation (1)

DETECTION AT RISK AND ACTUAL AT RISK				1	2	3		
		Likelihood	**Y0**	**Y1**	**Y2**	**Y3**	**SUM**	**SUM (WEIGHTED**
Financial Impact	Program works	83%	- 1,165 €	- 865 €	- 865 €	- 865 €		
Financial Impact Discounted		-	*1,165 €*	- *836 €*	- *807 €*	- *780 €*	- 3,588 €	
Financial Impact	Program does not work	17%	- 1,165 €	- 1,383 €	- 1,383 €	- 1,383 €		- **3,835 €**
Financial Impact Discounted		-	*1,165 €*	- *1,336 €*	- *1,291 €*	- *1,247 €*	- 5,040 €	
DETECTION AT RISK AND ACTUAL NOT AT RISK								
Financial Impact		-	1,165 €	- 865 €	- 865 €	- 865 €		
Financial Impact Discounted		-	*1,165 €*	- *836 €*	- *807 €*	- *780 €*	- **3,588 €**	
DETECTION NOT AT RISK AND ACTUAL AT RISK								
Financial Impact	Patient develops diabetes	85%	- 865 €	- 1,383 €	- 1,383 €	- 1,383 €		
Financial Impact Discounted		-	*865 €*	- *1,336 €*	- *1,291 €*	- *1,247 €*	- 4,740 €	
Financial Impact	Patient does not develop diabetes	15%	- 865 €	- 865 €	- 865 €	- 865 €		- **4,522 €**
Financial Impact Discounted		-	*865 €*	- *836 €*	- *807 €*	- *780 €*	- 3,288 €	
DETECTION NOT AT RISK AND ACTUAL NOT AT RISK								
Financial Impact		-	865 €	- 865 €	- 865 €	- 865 €		
Financial Impact Discounted		-	*865 €*	- *836 €*	- *807 €*	- *780 €*	- **3,288 €**	