Presentation Script – Health Security Assistance

Slide 1: Title – Health Security Assistance

Good morning/afternoon everyone. Today I present our project "Health Security Assistance", focused on predicting diabetes risk using artificial intelligence models applied to real patient data.
The idea arises from a clear need: to anticipate chronic diseases like diabetes, which have a strong impact both on individuals and on the healthcare system.
This approach aligns with a modern vision of medicine: proactive, personalized, and data-driven.
To this end, we will develop a predictive model based on current patient data to identify new potential clients at high risk of diabetes.

Slide 2: Context

We have real clinical and demographic data that allow us to anticipate health and consumption behaviors.
Our aim is to convert this data into actionable knowledge for marketing, sales, and service development.
Diabetes represents a strategic opportunity for the company due to its high prevalence and chronic nature.

Slide 3: Strategic Approach and Solving the Business Problem

Our strategy is to use this data to build both a Machine Learning model and a neural network to predict the probability of having diabetes.
Afterwards, we will apply the models to the most recent customer data to validate their effectiveness.

- Optimization of commercial and advertising spending, focusing resources where conversion is more likely.
- Development of personalized or preventive products.
- Strengthening the company's **strategic positioning** as a leader in health innovation.
-

Slide 4: EDA

We worked with a **solid, longitudinal, and representative dataset**.

- Training data: 1999–2017 → 15,488 clients.

- Prediction data: 2017–2023 → 6,258 clients.
  These include **clinical variables** (glucose, blood pressure, BMI…) and **demographic variables** (age, sex, race…).
  This temporal split allows us to simulate a **real scenario of future prediction** based on experience.

Slide 5: EDA – Target Variable

The target variable, as you can see, is a binary categorical variable and is imbalanced, since more people are healthy than diabetic.
This imbalance will be key in the development of both ML and neural network models.

Slide 6: EDA – Distributions

We analyzed the distribution of numerical variables. A logarithmic transformation was applied to all of them, achieving an almost normal distribution.

Slides 7: EDA – Group Differences

The most noticeable differences between groups are seen in glycated hemoglobin, BMI, and cholesterol.

Slide 8: EDA – Correlation Matrix

From the correlation matrix, we observe that the variables most correlated with the target are:

- Age
- Glycated hemoglobin
- Blood pressure
- And engineered features

Slide 9: Machine Learning Models

Given the target variable is imbalanced, we used the following metrics:

- Recall
- F1-score
- Average Precision

We tested several classical ML algorithms, and the best performing were:

- Gradient Boosting
- CatBoost
- LightGBM

Slide 10: Hyperparameter Evaluation

We evaluated hyperparameters using random search.
The best results came from LightGBM, particularly achieving 0.72 recall for the minority class.

Slide 11: Test Evaluation (ML)

- Accuracy: 89% → almost 9 out of 10 predictions were correct.
- Precision: 68% → 68% of the predicted positives were truly positive.
- Recall: 72% → 72% of real positives were correctly identified.
- F1 Score: 0.69 → balanced precision and recall.
- AUC-ROC: 0.90 → the model discriminates very well between classes.
- ROC AUC Curve: Area = 0.91, confirming strong performance.

Confusion matrix:

- True negatives (TN): 2,427
- True positives (TP): 327
- False positives (FP): 202
- False negatives (FN): 126

Conclusion: This ML model is robust and balanced. It successfully detects at-risk patients without causing excessive false alarms, which is essential for clinical or commercial decisions.

Slide 12: Deep Neural Network

To address class imbalance:

- **Class 1 (diabetics):** weight increased ×1.2
- **Class 0 (non-diabetics):** weight reduced ×0.9

Architecture:

- Dense layers: 128, 64, 32 (ReLU activations)
- Batch Normalization
- Dropout: 0.4, 0.3
- Output: Dense (1), sigmoid activation

Training:

- Optimizer: Adam
- Loss: binary_crossentropy
- Metrics: accuracy, AUC, recall, precision
- Early stopping: monitored val_auc, patience = 10

- Validation split: 20%, Epochs: 100, Batch size: 32

Slide 13: Test Evaluation (NN)

- **Accuracy:** 81%
- **Precision (class 1):** 43% → more false positives
- **Recall:** 84% → most positive cases detected
- **F1 Score:** 0.56
- **AUC:** 0.90
- **Average Precision:** 0.72

Confusion matrix:

- **TN:** 1,963
- **TP:** 393
- **FP:** 666
- **FN:** 60

Conclusion: The neural network is highly sensitive. It detects most diabetics, even at the cost of more false positives. This is useful when **clinical sensitivity is the priority**.

Slide 14: ML Final Results

 Precision (class 1): 0.66

 Recall (class 1): 0.78

 F1 Score (class 1): 0.72

 Precision (class 0): 0.95

 Recall (class 0): 0.92

 Accuracy: 89%

 Weighted metrics: ~0.89–0.90

 Confusion matrix: TN = 4,718, TP = 861, FP = 438, FN = 241

Slide 15 (NN): Final Results

• Precision **(class 1):** 0.66

• Recall **(class 1):** 0.78

• F**1 Score (class 1):** 0.72

- Precision **(class 0):** 0.95
- Recall **(class 0):** 0.92
- Accuracy**:** 89%
- Weighted **metrics:** ~0.89–0.90
- Confusion **matrix:** TN = 4,718, TP = 861, FP = 438, FN = 241

Closing and Conclusions

This project shows how clinical data and artificial intelligence can become key tools for the early detection of chronic diseases.
It creates value in three key areas:

1. Improved commercial efficiency
2. Direct clinical impact
3. Strategic positioning in digital health