

*2025*

Dkhili Ines

El Guebli Ons

Feki Malek

Chokri Ahmed

Sellemi Oumayma

Bouchakoura Alaa

mACHINE LEARNING PROJECT

BY

**TABLE OF CONTENTS**

# INTRODUCTION

# CRISP DM METHODOLOGY

# BUSINESS UNDERSTANDING

# DATA UNDERSTANDING

* 1. INITIAL DATA
  2. DATA EXPLORATION

# DATA PREPARATION

* 1. DATA CLEANING
  2. DATA IMPUTATION
  3. FEATURE ENCODING
  4. FEATURE ENGINEERING
  5. FEATURE SCALING

# MODELING & EVALUATION

* 1. SUPERVISED LEARNING
     1. K-NEAREST NEIGHBORS
     2. SUPPORT VECTOR MACHINE
     3. DESCISION TREE
     4. LINEAR REGRESSION
     5. XGBOOST REGRESSOR
  2. UNSUPERVISED LEARNING
     1. K-Means

# DEPLOYMENT

# CONSLUSION

# List of Figures

Figure 1: Phases of the CRISP-DM Data Mining Lifecycle…………………………………….6

Figure 2: Feature Names and Types………………………………………………………………………………..8

Figure 3: Boxplot of Numerical Features with Outliers……………………………………….…….8

Figure 4: Frequency Distribution of CustMonVal and Children Features

Figure 5: Missing Value Imputation Using Median and Mode

Figure 6: Imputing Invalid CustMonVal Entries

Figure 7: Label Encoding of Educational Degree (EducDeg)

Figure 8: Feature Engineering: Creating Age and Loyalty Duration

Figure 9: Outlier Removal Using IQR Method on CustMonVal

Figure 10: Outlier Removal in ClaimsRate Using IQR Method

Figure 11: Feature Selection, Train-Test Split, and Standardization

Figure 12: Elbow Method for KNN for Insurance Premium Prediction

Figure 13: KNN MSE and R² Evaluation for Insurance Premium Prediction

Figure 14: Evaluation Results for SVM with Different Kernels

Figure 15: Decision Tree Regression Evaluation Metrics

Figure 16: Linear Regression Evaluation Metrics

Figure 17: XGBoost Evaluation Metrics

Figure 18: Model Export Using Joblib

Figure 19: Input Field Guide

Figure 20: Prediction Form

Figure 21: Output – Predicted Premiums

# List of Tables

Table 1: Business Objectives and Corresponding Data Science Approaches

Table 2: Model Performance Comparison for Insurance Premium Prediction

# INTRODUCTION



Maghrebia Insurance is one of the leading insurance providers in the Maghreb region, offering a wide range of products covering health, automotive, home, and life insurance. Known for its solid reputation and long-standing presence in the market, the company has built a large customer base through traditional insurance models and legacy systems. However, despite its comprehensive service offering, Maghrebia still operates with outdated technologies that limit automation, personalization, and responsiveness.

To meet the growing demands of modern customers and remain competitive in a fast-evolving industry, Maghrebia Insurance is now undergoing a strategic digital transformation. This new phase focuses on integrating artificial intelligence to modernize its services, streamline operations, and offer a more intelligent, customer-centric experience.

As part of this transformation, a 24/7 intelligent chatbot will provide continuous assistance and improve service accessibility. An AI-powered CV matching system will optimize the recruitment process by selecting the best candidate profiles efficiently. Predictive models will forecast customer insurance premiums and estimate total claim amounts, allowing better risk management and pricing strategies. Additionally, a personalized recommendation engine will suggest the most suitable insurance products for each customer based on their preferences and behavior.

By adopting these technologies, Maghrebia Insurance is not only upgrading its digital infrastructure but also positioning itself as an innovative leader in the regional insurance market—agile, proactive, and data-driven.

# CRISP DM METHODOLOGY

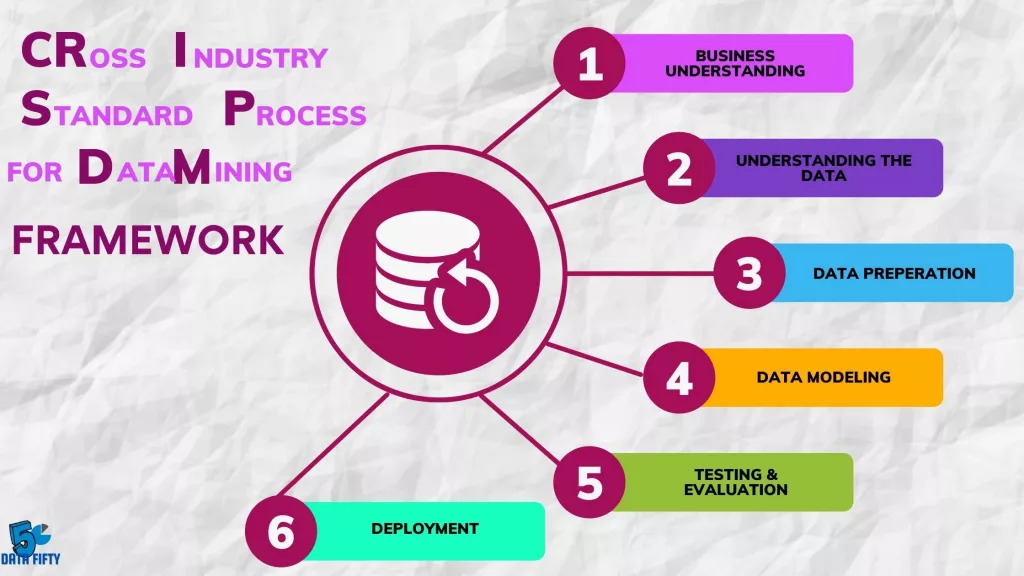


figure1: Phases of the CRISP-DM Data Mining Lifecycle

Many companies undertaking Data Mining projects deploy various methodologies — one of the most well-known is the CRISP methodology.  
CRISP-DM, which stands for *Cross-Industry Standard Process for Data Mining*, is a field-tested method designed to guide data exploration and mining projects.

This method is divided into six key phases, from understanding the business problem to deployment and production:

1. **Understanding the Business Problem**:  
   This step helps define the project's objectives by asking the right business-related questions.
2. **Understanding the Data**:  
   This involves exploring the available data and beginning the initial analysis.
3. **Data Preparation**:  
   This is a cleaning and reorganization phase of the data, often requiring the creation of new calculated values based on existing data.
4. **Modeling**:  
   This step focuses on selecting the appropriate data mining model and its parameters.
5. **Evaluation**:  
   This involves assessing the performance of the model and comparing the results with the project's original goals.
6. **Deployment**:  
   In this final phase, the solution is implemented, and new practices are introduced based on the project's outcomes.

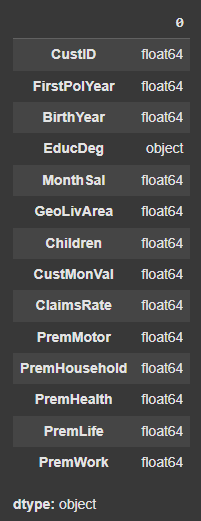
# BUSINESS UNDERSTANDING

table1: Business Objectives and Corresponding Data Science Approaches

|  |  |
| --- | --- |
| Business Objective (BO) | Data Science Objective (DSO) |
| Reduce claim processing time and improve customer experience | **Predict** total claim amounts based on historical and client profile data |
| Optimize pricing strategies to ensure fair and competitive insurance premiums | **Predict** individual client premium amounts using risk and demographic variables |
| Increase hiring efficiency and reduce manual effort in recruitment | **Interpret** CVs using NLP and **match** candidates to job roles automatically (CV Matching) |
| Enhance customer service availability and response time with intelligent automation | **Implement** and **develop** an AI-powered **chatbot** for 24/7 assistance |
| Improve customer retention and reduce churn | **Forecast** churn probability to identify at-risk clients and take proactive action |
| Offer personalized products based on client needs and behavior | **Recommendation System**: **Suggest** tailored insurance offers based on client profile and interaction history |

# DATA UNDERSTANDING

* 1. INITIAL DATA

The datasets used in this project are derived from simulated and structured insurance records, with each business objective relying on a dedicated dataset tailored to its specific analytical needs. These datasets contain relevant information such as customer demographics, contract types, historical claims, monthly premiums, service usage patterns, and loyalty indicators.  
Each use case—such as **claim amount prediction**, **churn detection**, **chatbot development**, and **CV matching**—has its own distinct dataset. However, the **premium prediction** and **product recommendation system** share a common dataset, as both require a deep understanding of customer profiles and financial behavior.

* 1. DATA EXPLORATION

In this stage, we analyzed the dataset used for **insurance premium prediction**, focusing on the structure, distribution, and relationships among variables.

* **Data Types and Structure**

The dataset consists of 33 attributes, mainly numeric (float64) and one categorical feature (EducDeg). The target variables for premium prediction include PremMotor, PremHealth, PremLife, and PremWork.

* **Outlier Detection in Numerical Features**Using box plots, we identified significant outliers in several features, such as:
  + MonthSal (Monthly Salary)
  + CustMonVal (Customer Monetary Value)
  + ClaimsRate

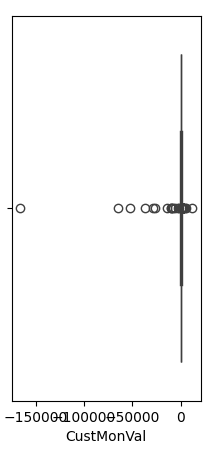
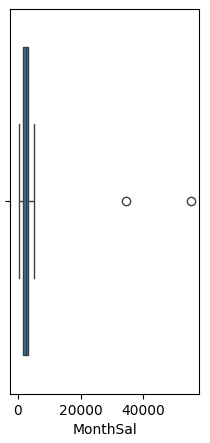
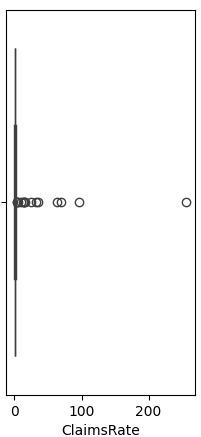
 

figure3: Boxplot of Numerical Features with Outliers

figure2: Feature Names and Types

According to the figure below

* **Children** is binary (0 or 1), showing whether the client has children.
* **CustMonVal** shows a high concentration around certain positive values, but some negative values were also noticed, which could indicate errors or specific financial states.

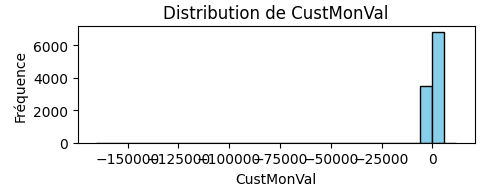
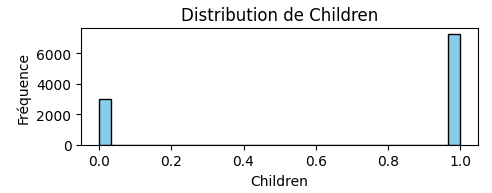


figure4: Frequency Distribution of CustMonVal and Children Features

# DATA PREPARATION

* 1. DATA CLEANING

To prepare the dataset for premium prediction, several key steps were taken:

* **Dropped CustID** (non-informative identifier)
* **Removed rows** with missing values in premium-related columns (Prem\*)
* **Filtered out outliers** in age (<18 or >150), salary (>30,000), and negative CustMonVal
* **Excluded records** where FirstPolYear > 2016 (insufficient history)
  1. DATA IMPUTATION
* **Filled missing values** using median (for numerical) and mode (for categorical)

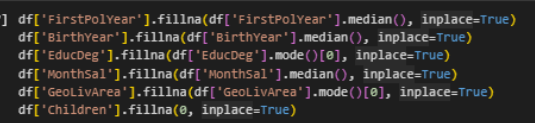


figure5: Missing Value Imputation Using Median and Mode

* **Replaced negative CustMonVal** with the median of positive values

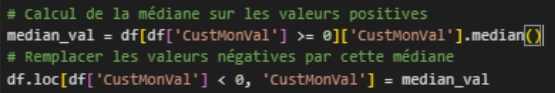


figure6: Imputing Invalid CustMonVal Entries

* 1. FEATURE ENCODING
* **Encoded EducDeg** (label encoding)



figure7: Label Encoding of Educational Degree (EducDeg)

* 1. FEATURE ENGINEERING
* **Created Age and LoyaltyYears**, then dropped BirthYear and FirstPolYear

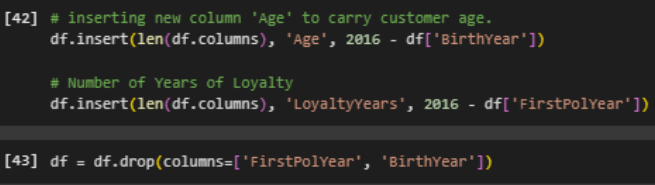


figure8: Feature Engineering: Creating Age and Loyalty Duration

* 1. FEATURE SCALING
* Applied **IQR filtering** for robust outlier handling

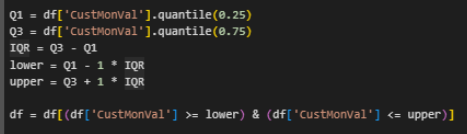


figure9: Outlier Removal Using IQR Method on CustMonVal

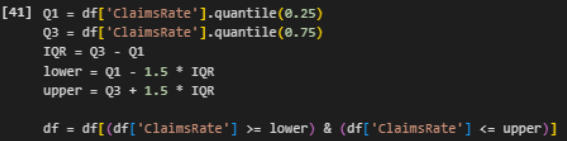


figure10: Outlier Removal in ClaimsRate Using IQR Method

* Used **StandardScaler** for normalization before model training

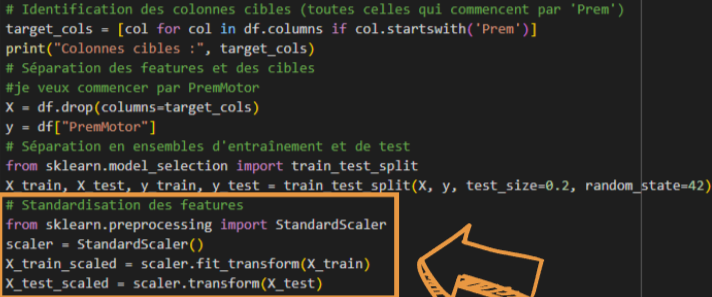


figure11: Feature Selection, Train-Test Split, and Standardization



# MODELING & EVALUATION

* 1. SUPERVISED LEARNING
     1. K-NEAREST NEIGHBORS

In this modeling phase, we used the **K-Nearest Neighbors (KNN) regression algorithm** to predict insurance premiums. The model was applied to the **premium prediction dataset**, after scaling the features to ensure fair distance computation.  
To determine the optimal value of k, we used the **Elbow Method** (see *Figure 12*), plotting the Mean Squared Error (MSE) across different k values. Based on the curve, we selected **k = 3**, which minimized the error without overfitting.

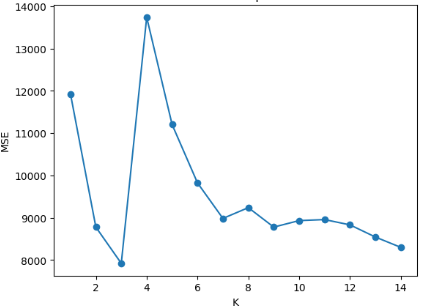


figure12: Elbow methode for knn for Insurance Premium Prediction



After standardizing the features, we trained the model and evaluated it on the test set. As shown in *Figure 13*, the results were: **MSE:** 8300.76, **R² Score:** 0.47

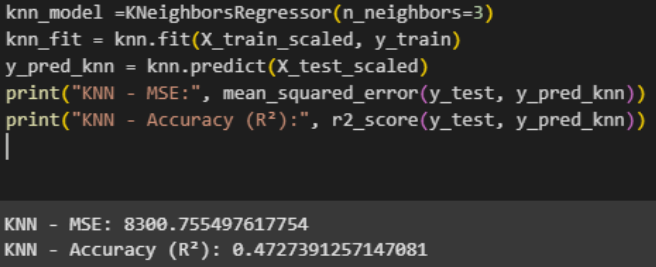


figure13: KNN MSE and R² Evaluation for Insurance Premium Prediction



* + 1. SUPPORT VECTOR MACHINE

We applied **Support Vector Regression (SVR)** to the insurance premium prediction dataset using three different kernels: **linear**, **polynomial**, and **RBF**. SVR is a powerful model that finds a function within a margin of tolerance, making it suitable for complex regression tasks.

After scaling the features, we trained and evaluated each kernel (see *Figure 14*). The results were as follows:

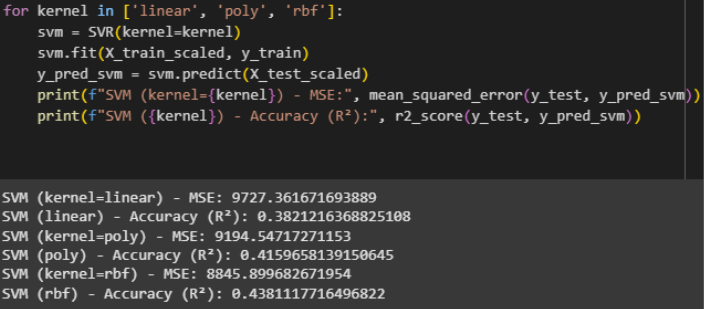


figure14: EVALUATION Results for SVM with Different Kernels for Insurance Premium Prediction



Among the three, the **RBF kernel** performed the best with the lowest error and highest R² score, indicating it captures non-linear relationships better in this dataset.

* + 1. DESCISION TREE

We used a **Decision Tree Regressor** to model the insurance premium prediction task. After training on the scaled dataset, the model achieved the following results (see *Figure 15*):

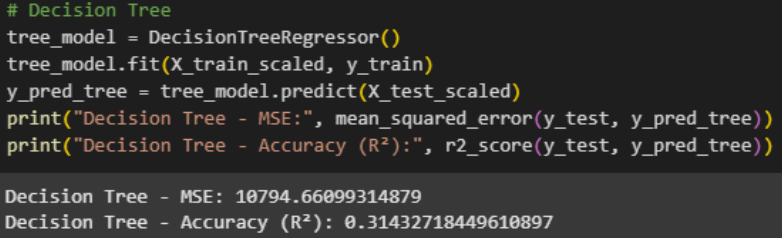


figure15: Decision Tree Regression Evaluation Metrics for Insurance Premium Prediction

These results indicate **low predictive performance**, with the model explaining only **31% of the variance** in the target variable. The high MSE and low R² suggest that the model may be **overfitting the training data** or failing to capture complex patterns, making it less suitable for this regression task without further tuning or regularization.

* + 1. LINEAR REGRESSION

We applied a **Linear Regression model** to predict insurance premiums. After training on the scaled dataset, the model achieved the following results (see Figure 16):

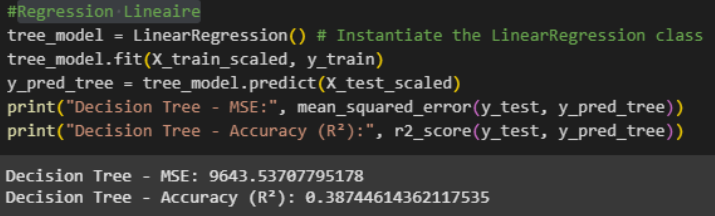
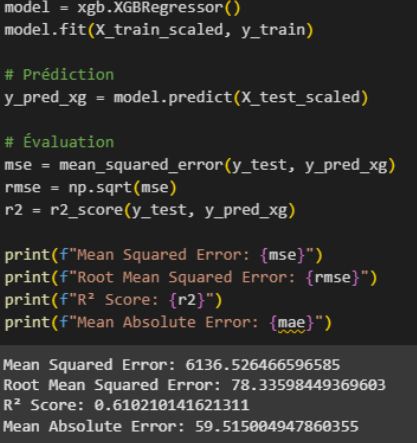


figure16: Linear Regression Evaluation Metrics for Insurance Premium Prediction

These results reflect **limited predictive power**, with the model explaining only **39% of the variance**. This is expected, as **linear regression assumes a linear relationship between features and the target**, which doesn’t hold true in our case. For instance, insurance needs increase both at a **young age** (for coverage) and at an **older age** (for risk), creating a **non-linear relationship** that linear regression cannot capture well—hence the reduced performance.

* + 1. XGBOOST REGRESSOR

We implemented the **XGBoost Regressor**, a powerful ensemble learning algorithm known for handling complex, non-linear relationships and minimizing overfitting. After training on the scaled premium prediction dataset, the model achieved the following results (see Figure 17):



These metrics indicate that **XGBoost delivers the best performance** among all tested models, explaining over **61% of the variance** with the lowest error rates. Its ability to capture intricate patterns in the data makes it the most suitable model for this insurance premium prediction task.

* **Amelioration:**

We implemented the **XGBoost Regressor**, a robust ensemble learning algorithm well-suited for capturing complex, non-linear patterns in data. To further enhance model performance, we applied **RandomizedSearchCV**, an efficient hyperparameter tuning technique that explores a wide range of configurations. This allowed us to identify a better-performing model setup without excessive computational cost.

figure17: XGBoost Evaluation Metrics

for Insurance Premium Prediction

* **Comparative table:**

|  |  |  |  |
| --- | --- | --- | --- |
| Model | MSE | R² Score | Remarks |
| KNN | 8300.76 | 0.47 | Simple baseline, moderate performance |
| Linear Regression | 9643.54 | 0.39 | Struggles with non-linear relationships |
| Decision Tree | 10794.66 | 0.31 | Overfits easily, weak generalization |
| XGBoost | **6136.53** | **0.61** | Best performer; benefits from hyperparameter tuning via RandomizedSearchCV |

table2: Model Performance Comparison for Insurance Premium Prediction

* 1. UNSUPERVISED LEARNING
     1. K-Means

# DEPLOYMENT For deployment, we exported our best-performing model (****XGBoost****) using joblib in. pkl format (Figure 18), allowing it to be integrated seamlessly into the frontend of our insurance prediction plateform

figure18 : Model Export Using joblib forPremium Prediction

To ensure usability for non-technical users, we designed a **user-friendly interface** where clients can input their personal and financial data through a structured form (Figure 20). Each input field is supported by a clear information guide (Figure 19) that explains the meaning and expected format of each field (e.g., salary, education level, geographic region, etc.).

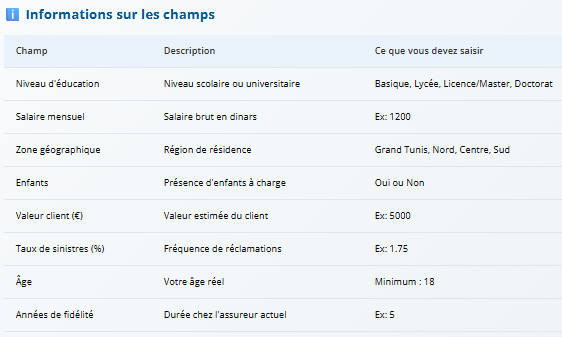
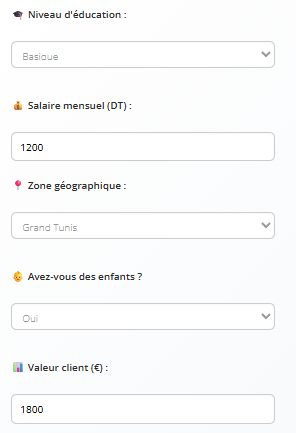


figure19 : Input Field Guide



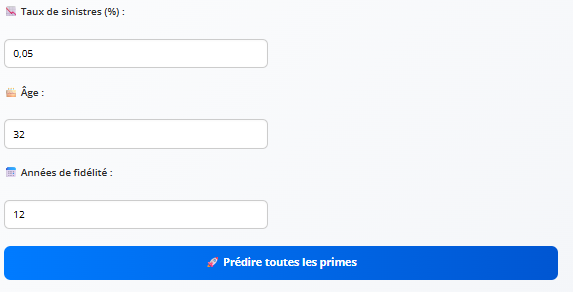


figure20: Prediction Form

Once the form is submitted, the system uses the trained model to predict **insurance premiums** across multiple categories including health, housing, life, automobile, and workplace accident insurance (Figure 21). This setup enables real-time, personalized premium estimation, making the tool both practical and accessible for end users.

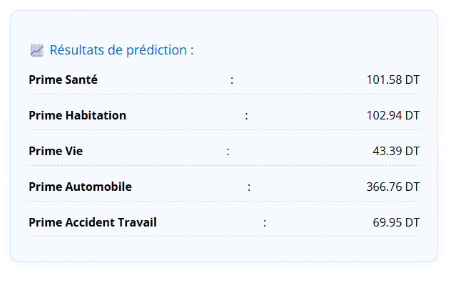


figure21: Output – Predicted Premiums

# CONSLUSION:

This project marked an important step in modernizing the insurance experience through the integration of intelligent, data-driven technologies. By leveraging machine learning and artificial intelligence, we addressed several strategic challenges, including premium prediction, customer profiling, automated CV analysis, and chatbot-based support.

Using a structured approach, we cleaned and prepared real-world data, explored key variables, tested multiple models, and deployed our best-performing solutions into an interactive and user-friendly interface. This end-to-end pipeline enables real-time predictions, automation of processes, and enhanced decision-making for both clients and insurers.

Overall, the project demonstrates how AI can transform traditional insurance services into smarter, more personalized, and more efficient systems — paving the way for broader adoption and future innovation in the sector.

# Bibliography

1. Shearer, C. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing, 5(4), 13–22.

2. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

3. Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.