UNIVERSIDAD POLITÉCNICA DE MADRID

E.T.S. DE INGENIERÍA DE SISTEMAS INFORMÁTICOS

DATA PROCESS PROJECT

DATA SCIENCE MASTER

Project plan: Prediction of the severity of COVID-19 cases among diabetes patients

Authored by: Sheyla Leyva Sánchez, Samuel Salgueiro Sánchez, Mariajose Franco Orozco, Francisco Lozano del Moral

Madrid, November 4, 2024

Project plan: Prediction of the severity of COVID-19 cases among diabetes patients
Authored by: Sheyla Leyva Sánchez, Samuel Salgueiro Sánchez, Mariajose Franco Orozco, Francisco Lozano del Moral
Data Process Project, November 4, 2024

E.T.S. de Ingeniería de Sistemas Informáticos
Campus Sur UPM, Carretera de Valencia (A-3), km. 7
28031, Madrid, España

To cite this work using BibTeX the complete reference is:

```
@mastersthesis{citekey,
title = {Project plan: Prediction of the severity of COVID-19 cases among diabetes patients},
author = {Sheyla Leyva Sánchez, Samuel Salgueiro Sánchez, Mariajose Franco Orozco, Francisco
Lozano del Moral},
school = {E.T.S. de Ingeniería de Sistemas Informáticos},
year = {2024},
month = {11},
type = {Data Process Project}
}
```

Contents

List of Tables

1. Introduction

Context of the Project

The COVID-19 pandemic continues to challenge healthcare systems worldwide, with diabetic patients emerging as one of the most vulnerable groups. Due to compromised immune systems and other underlying conditions, individuals with diabetes are at a significantly higher risk of severe complications, prolonged hospitalization, and mortality when infected with COVID-19 [Fed21]. Moreover, research from Weill Cornell Medicine has demonstrated that COVID-19 can worsen existing diabetes and even trigger new cases by activating immune cells that destroy insulin-producing beta (β) cells [Med24].

This project specifically focuses on diabetic patients aged 40 to 60, a key demographic due to their dual vulnerability and societal role. This age group represents a substantial portion of the working population, meaning that mitigating their risk not only saves lives but also preserves healthcare resources and supports societal productivity. According to the International Diabetes Federation (IDF), the prevalence of diabetes in Spain is **14.8%** among adults aged 20 to 79, equating to approximately 5.1 million people [Dia24]. Additionally, the Spanish National Health Survey 2017 indicates that the prevalence of diabetes increases with age, showing a significant rise starting from 45 years [Min17].

Problem Statement

The primary goal of this project is to develop a Machine Learning (ML) model that can predict which diabetic inpatients are most at risk of developing severe COVID-19 complications. By identifying these high-risk patients early, hospitals can implement timely interventions, reduce ICU admissions, and optimize resource allocation. This approach not only enhances patient outcomes but also helps alleviate the financial burden on healthcare systems by reducing unnecessary treatments and extended hospital stays.

Stakeholders

Hospitals and Healthcare Providers

2. Cost/Benefit Analysis

2.1. Benefits

Diabetic patients who develop severe COVID-19 often require intensive and prolonged treatments, significantly increasing hospital costs. These include the need for prolonged mechanical ventilation, specialized medications to stabilize blood glucose levels, continuous monitoring, and frequent consultations with specialists. These interventions and secondary complications like infections contribute to a substantial financial burden on healthcare systems, as shown in Table 2.1.2.

The specific costs produced by the specialized treatments required for this type of patient can be mitigated through proactive management by implementing a predictive model that can detect which patients are at high risk. This would benefit hospitals by saving a total of 65,287€ (see tables 2.1.2,2.1.1,2.1.3) per patient if these cases are identified in time, in addition to the evident benefits of early risk detection, such as better resource management, reduced use of emergency facilities, and higher survival rates.

Cost Category	Sub-Category	Estimated Cost (€)
Monitorization	Hospital consult	50
	PCR	20
Treatment	COVID-19 pill treatment	1,400
Prevention	Antiviral	2,340
	LaMA[1]	6,090
Maximum Total Cost		**9,900**

Table 2.1.1. Costs associated to preventing a patient from developing severe COVID-19 case. (1) Long-acting Monoclonal Antibodies.

Cost Category	Sub-Category	Estimated Cost/Patient (€)
Hospitalization	Bed[1]	4,746
	ICU[2]	1,300/day
Treatment	Medication	64-3,314
	TACs & X-Rays	230
	MGC	220
	ECMO	39,000
	Assisted Ventilation	56,927
Estimated Total Cost		**75,187 - 114,187**

Table 2.1.2. Current costs associated with the hospitalization of diabetes patients due to COVID-19 infections. (1) Average cost of hospitalization. (2) ICU might be needed for severe cases and the length of stay averages 10 days [Ali+22].

	Prevention Plan
Actual Severe	-65,287€ (TP)
Actual Mild	+9,900€ (FP)

Table 2.1.3. Savings and costs of using a prevention plan.

2.2. Costs

To effectively implement a predictive model for managing high-risk diabetic patients, a strategic allocation of resources is required. Table 2.2.1 outlines the key material costs necessary to bring this project to fruition, ensuring successful deployment and integration. Additionally, significant non-monetary efforts, referred to as immaterial costs, are also required. These include time and effort from healthcare staff and management, as detailed in the subsequent section. A comprehensive breakdown of the budget for model deployment is provided later in Section 4.

Material Costs

Cost Category	**Sub-Category**	**Estimated Cost (€)**
Model Development[1]	Human Resources	50,600-53,800
	Computational Resources	20,000
Model Deployment	Cloud Upkeep	50[2]
	IT Cloud Monitoring	250[3]
	Staff Training + Workshops	13000
First Year Total Cost		**100,200-103,400** [4]

Table 2.2.1. Costs associated with implementing, deploying, and maintaining the model, including salaries and infrastructure, for the first year. (1) It includes API development and the creation of reports and documentation. (2) This is a monthly amount to run predictions of all individuals once. (3) Pay per day as needed. (4) Assuming 64 days of work throughout the year.

Immaterial Costs

Implementing the predictive model involves non-monetary costs that, while not directly tied to financial expenses, require significant time and effort from healthcare staff and management. These costs include:

- **Time Commitment for Training**: Healthcare professionals will need to allo-

cate time away from their regular duties to participate in training sessions on how to use and interpret the model effectively.

- **Effort in Adapting Workflow**: Staff may need to adjust their existing workflows to incorporate the predictive model, which could initially impact productivity.
- **Management Oversight**: Hospital management will have to oversee the integration of the model, dedicating time to ensure that implementation aligns with hospital policies and objectives.

Once the model is finalized and ready for deployment, it is estimated that the integration and adaptation process will take approximately 4 to 6 weeks. During this period, healthcare staff will need to invest around 15-20 hours collectively in training and workflow adjustments, while management will allocate an additional 10-15 hours to oversee and coordinate the implementation.

While the initial investment ranges from €100,200 to €103,400, the long-term financial benefits and improved patient outcomes justify these expenses. By reducing the need for expensive treatments and optimizing resource use, hospitals can achieve significant savings and enhance the quality of care provided to diabetic patients at risk of severe COVID-19 complications. Additionally, the initial time and effort invested by healthcare staff and management will be compensated through lasting benefits, as streamlined workflows and improved patient management practices introduced through the model will lead to sustained efficiency gains. Below is the detailed Return on Investment (ROI) analysis, which quantifies the financial gains relative to the project's initial expenses.

2.3. Return on Investment (ROI)

The Return on Investment (ROI) is a measure used to evaluate the efficiency of an investment or compare the efficiency of multiple investments. It calculates the ratio of the net return on an investment to the initial cost. In this context, the ROI for implementing the predictive model is calculated as follows:

ROI Calculation

$$\text{ROI} = \frac{\text{Total Savings} - \text{Initial Investment}}{\text{Initial Investment}} \times 100$$

Break-Even Point Calculation

To determine the minimum number of severe cases the model needs to prevent to break even, we use the following calculation based on the average savings per severe case prevented:

$$x \times 65,287 = 103,400$$
$$x = \frac{103,400}{65,287} \approx 1.58$$

Since the number of patients must be a whole number, the model needs to successfully prevent severe outcomes for **at least 2 high-risk patients** to break even.

Example ROI Calculation (Worst-Case Scenario)

In the worst-case scenario, if the model only prevents severe cases for 2 high-risk patients per year:

- **Initial Investment**: €103,400 (upper limit of the cost)
- **Estimated Savings per Severe Case Prevented**: €65,287
- **Number of Patients Benefited Annually**: 2

$$\text{Total Savings} = 2 \times 65,287 = 130,574$$
$$\text{ROI} = \frac{130,574 - 103,400}{103,400} \times 100 \approx 26.3\%$$

In this scenario, the ROI is 26.3%, indicating a modest financial gain.

Example ROI Calculation (Projected Scenario)

With a projected performance of preventing severe cases for 10 high-risk patients annually:

- **Initial Investment**: €103,400
- **Estimated Savings per Severe Case Prevented**: €65,287
- **Projected Number of Patients Benefited Annually**: 10

$$\text{Total Savings} = 10 \times 65,287 = 652,870$$
$$\text{ROI} = \frac{652,870 - 103,400}{103,400} \times 100 \approx 531.5\%$$

In this optimal scenario, the ROI reaches 531.5%, demonstrating substantial financial benefits.

Conclusion

Even in the worst-case scenario, the model yields a positive ROI of 26.3%. As the model effectively prevents more severe cases, the ROI increases, reaching up to 531.5% if 10 patients benefit annually. This analysis underscores the significant financial impact and cost-effectiveness of the predictive model in managing high-risk diabetic patients with COVID-19.

3. Framing the Problem as a Data Science Task

3.1. Objective

The objective is to develop a predictive model to forecast severe COVID-19 illnesses in diabetic patients aged 40-60. Severe cases will be defined by specific clinical criteria such as:

1. The need for intubation.
2. Requirement for mechanical ventilation.
3. Hospitalization in general wards.
4. Admission to an Intensive Care Unit (ICU).

The goal of this model is to provide healthcare professionals with a tool that allows for **early identification** of high-risk diabetic patients, enabling **timely interventions**. This will help optimize the allocation of **hospital resources** and improve **patient outcomes**.

The predictions will be based on a combination of **demographic, clinical, and hospital admission data**, which will be processed through a **reproducible and interpretable** ML pipeline.

3.2. Dataset Preparation

3.2.1. Data Integration

To ensure seamless integration of the provided data into our ML pipeline, the process will start with a structured, two-step integration strategy. Firstly, all data will be transformed into a centralized set of tabular datasets. Following this, the datasets will be consolidated into a unified dataset. This process optimizes the data for efficient processing, feature engineering, and modeling.

Data Access and Formats

The hospitals keep all data as files in JSON, Excel, and CSV formats. Because of this, the steps that will be taken are:

- **Tabular (CSV or Excel):** These formats are inherently tabular and therefore do not require transformations to fit into the pipeline.
  - **Estimated Time Required:** 0 days.
- **Hierarchical (JSON):** JSON files will undergo parsing to flatten any nested structures, converting them into a structured, tabular format compatible with our pipeline. Validation will be conducted to ensure data integrity following the transformation.
  - **Estimated Time Required:** 2–3 days.

If any unexpected formats arise, a one-week buffer is recommended to reformat or convert the datasets as needed. This additional time allows for any necessary transformation and validation, ensuring all data is fully compatible with ML pipeline requirements.

Secure and Centralized Access

To ensure efficiency while adhering to strict data protection and ethical guidelines, a centralized, secure access point for all medical data will be established. Ideally, this would be through a compliant cloud-based folder or data warehouse that aligns with regulations outlined in the General Data Protection Regulation (GDPR) [18a] and the Spanish Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales (LOPDGDD) [18b]. This setup minimizes delays in data access and ensures that data is handled in a controlled and legally compliant environment. By implementing these measures, we not only enhance security but also streamline project workflow, allowing our data science team to focus on analysis while respecting privacy and ethical standards.

Data Understanding

To ensure a comprehensive understanding of the data, a detailed documentation to accompany all datasets will be required, which should include:

- **Data Dictionary**: A clear explanation of the variables, their types, ranges, and units of measurement, allowing our data scientists to correctly interpret and utilize the data.
- **Source Documentation**: Information on where the data originates (e.g., hospital systems, clinical trials), which ensures transparency and traceability of the data's sources.

If these documents are not available or incomplete, three **consultative meetings** with medical experts and our data scientists will be added. During these meetings, the necessary information to create the metadata and documentation ourselves will be gathered. This collaborative effort will ensure the data is fully understood and properly prepared for use in the predictive model.

Data Consolidation

1. **Item reconciliation**: Patents can have information across different hospitals. Due to this, the first step will be to check the identifiers to link individual pa-

tients across all datasets. In this case, Spain has two ways to uniquely identify patients based on their Patient identification documents:

- **National Person Identifier**
- Pairs of **Regional Person identifier** and **Region identifier**.

We will make an extra step to minimize the possibility of reconciling different patients that have the same identifiers due to human or machine errors using birth dates. The full names will not be used to not misclassify individuals who changed their name, such as members of the transsexual community.

2. **Feature reconciliation**: We expect similar or equivalent features to appear duplicated across the datasets received. For this, the metadata of each dataset [3.2.1] will be investigated to find and link these features. In this step, units will be checked to maintain a consistent scale.

3. **Consolidation of files**: The equivalences found in the previous two steps will be used to merge all datasets.

3.2.2. Data Assurances for Replicability

For practical replicability, the data used will be representative of the target population and it will be ensured that it reflects the current active COVID-19 strains.

Sampled Population. Data must represent the target population of diabetic patients **aged 40 to 60** in the **territory of Spain**. As such, any samples out of the age range or without the condition will be removed from the data.

To ensure that the samples are representative of the target population, these checks will be carried out:

1. The **geographical distribution** of the samples matches the population distribution in Spain in that age range.

2. The **ethnic and gender distribution** of the samples are similar to the distributions expected in the sampled population.

If there is a statistically substantial deviation from the expected distributions, a meeting with a domain expert will be required to ensure that the data does not undercover any particular population group.

Should this happen, techniques that target sample imbalance (**SMOTE** or **sample weighting**) will be included in the model training to ensure the theoretical results are replicable.

Relevance of Time. COVID-19 mutates rapidly. As such, the data of COVID-19 infections must have been collected **within the past 18 months**.

Should the data be considered insufficient, a domain expert will be contacted to adjust the accepted period for COVID-19 infections.

3.2.3. Ethical Concerns and Data Privacy

To maintain strict patient confidentiality and comply with data protection regulations, we will use a multi-layered approach for handling sensitive data:

1. **Data Collection and Minimization** We will limit the data collected to only what is essential for the analysis. Personal identifiers, such as names or addresses, will be excluded to prevent any potential re-identification.
2. **Data Anonymization Techniques** We will remove or generalize any indirect identifiers, such as birthdates and unique characteristics. By applying generalization (e.g., age ranges) and removing outliers, we ensure that individual identities cannot be traced back.
3. **Encryption and Access Control** All data will be encrypted both in storage and during transmission. Access will be strictly managed: only authorized team members with a need-to-know basis will have access to the data, which will be closely monitored with regular audits.

This approach will safeguard patient data and maintain the highest level of privacy, allowing us to focus on insights while upholding our commitment to data security.

3.3. Approach

This section describes an outline of the machine-learning techniques and methodologies that will be applied to predict severe COVID-19.

Data Pre-processing

To ensure the data is ready for modeling, a series of pre-processing steps will be applied:

1. **Data pruning**:
   (a) **Filtering Data Errors**: Using the metadata [3.2.1] validation will be carried out to eliminate data that is outside the valid ranges (e.g. 256% oxygen saturation).
   (b) **Duplicate Filtering**: Duplicates will be filtered using the primary keys that identify patients to ensure an error was not made in the previous step.
   (c) **Outlier Detection**: Outliers in clinical data will be analyzed using Inter Quantile Range (**IQR**) in the case of quantitative variables and **Chi-Squared tests** for the categorical variables.
   (d) **Completeness Threshold**: Some patients in the dataset are expected to have extremely sparse data. To prune the dataset of these individuals, we will identify the standard amount of data to identify a threshold of data availability per patient based on the number of columns filled. After this, we will prune the patients below the threshold.
2. **Imputing Missing Data**: Basic patterns of missing data will be handled using imputation techniques (e.g., mean/mode imputation for demographic data). For more complex patterns, **Multiple Imputation Chained Equations (MICE)** will be used.
3. **Transformations**:
   (a) **Ordinal Variable Transformations**: Ordinal variables will be transformed into discrete variables to represent the order between them.

(b) **Numerical (Ordinal and Quantitative) Variables**: These variables will be normalized or standardized to ensure uniformity across the dataset.

(c) **Nominal Variable Transformations**: Nominal variables with two possible values will become binary numerical variables. Those that have more than two possible values will be encoded numerically using either One-Hot-Encoding or indices of trainable vector embeddings as part of the input layer of the models.

4. **Class Imbalance**: As we expect a class imbalance (fewer severe cases than non-severe), we will employ methods such as **Synthetic Minority Over-sampling Technique (SMOTE)** to ensure balanced training datasets.

Feature Selection

Selecting the most relevant features is crucial for creating an efficient and interpretable model. In this case, given the nature of the data, we expect a very large number of features, including an unknown amount of redundant or irrelevant features.

To decrease this number to a more manageable set, the following methods will be implemented:

1. **Multivariate Recursive Feature Elimination (RFE)**: This method will rank the features based on their importance and remove less relevant features. We will do this until the number of features is low enough to perform wrapper methods.

2. **Wrapper Methods**: We will a **Random Forest with AdaBoost** model in combination with an annealing-based search strategy to explore the feature space. This technique balances the exploration of new feature combinations with the exploitation of high-performing subsets, ensuring we select the most relevant variables.

Modeling

Multiple ML models will be considered to identify the most effective approach:

- **Logistic Regression**: As a baseline model, logistic regression provides a simple, interpretable method for understanding which factors influence severe outcomes.
- **Random Forest**: This ensemble model combines decision trees to improve accuracy and provide insights into feature importance, allowing clinicians to understand the key drivers of risk.
- **Gradient Boosting Machines (XGBoost)**: For more complex data patterns, XGBoost will iteratively correct errors in predictions, offering high accuracy while managing overfitting.

Evaluation

To assess the model's performance, we will focus on metrics that align with our hospital's needs:

- **Sensitivity (True Positive Rate)**: Given the high stakes of missing a severe case, we will prioritize sensitivity to ensure that we correctly identify at-risk patients.
- **F-1 Score**: This metric emphasizes sensitivity over precision, making it more suitable for cases where false negatives carry a higher risk than false positives.
- **ROC-AUC**: This will provide a broader evaluation of the model's ability to distinguish between severe and non-severe cases across different thresholds.

To ensure robustness, we will use **bias-corrected 10-fold cross-validation**, which will provide a reliable measure of how the model will be performed on unseen data in real-world hospital settings.

4. Work Plan & Detailed Task Breakdown

4.1. Task Breakdown

4.1.1. Data Acquisition and Integration

This section outlines the structured plan and detailed tasks required to develop the predictive model.

4.1.2. Data Acquisition and Integration

1. **Dataset Preparation:**
   (a) **Change Format to Tabular:** Convert the collected data from hospitals into a tabular format to ensure consistency across all records.
   (b) **Filter to Target Population:** Extract and retain only the data relevant to the target population, specifically diabetic patients aged 40-60.
   (c) **Data Consolidation:** Integrate the converted data into a single, comprehensive dataset that is ready for analysis and machine learning applications.

4.1.3. Data Organization and Context Understanding

1. **Data Quality Evaluation:**
   (a) Assess the dataset for completeness and consistency, flagging and resolving any issues with missing or inconsistent records.

(b) Ensure that the dataset accurately represents the target population and includes examples of both severe and non-severe cases. Consult with domain experts to resolve any detected anomalies and apply appropriate weighting if needed to correct imbalances.

2. **Initial Data Analysis:**

   (a) Conduct summary statistics and basic visualizations to gain an understanding of variable distributions and contextual relationships.

3. **Metadata Handling:**

   (a) **Alternative 1:** If complete metadata is available, proceed directly with assessing the relevance of variables to the predictive model.

   (b) **Alternative 2:** If metadata is incomplete, organize three consultative meetings with medical experts to collect the necessary details. This step will add approximately one week to the project timeline.

4.1.4. Data Preprocessing

1. **Data Cleaning:**

   (a) **Handling Missing Values:** Missing values in key clinical and demographic features (e.g., blood glucose levels, oxygen saturation, comorbidity status) will be addressed based on the importance of these variables:

   - **Simple Imputation:** Use median or mode imputation for features where missing values are minimal. This approach is particularly suitable for Logistic Regression, which is sensitive to data inconsistencies.
   - **Advanced Imputation:** Apply Multiple Imputation by Chained Equations (MICE) or k-nearest neighbors imputation for features with more complex missing data patterns, ensuring data completeness for Gradient Boosting and Random Forest models.

   (b) **Outlier Detection and Treatment:** Given the clinical context, identify and treat outliers using methods like the Interquartile Range (IQR) for numerical variables or context-specific thresholds. For example, extremely high or low values in vital signs may need to be verified or capped based on medical guidelines.

2. **Categorical Encoding:**
   (a) **One-Hot Encoding for Logistic Regression:** Transform categorical variables (e.g., gender, comorbidity status) into binary features using one-hot encoding. This ensures that the linear model can interpret the categorical information correctly.
   (b) **Label Encoding for Tree-Based Models:** Use label encoding or ordinal encoding for features where there is an inherent order (if any), making it efficient for Gradient Boosting and Random Forest, which handle categorical features well without requiring one-hot encoding.
3. **Scaling and Normalization:**
   (a) **Standardization for Logistic Regression:** Standardize numerical features (e.g., age, BMI, blood glucose levels) using z-scores to ensure all features are on a comparable scale. This is crucial for the convergence and interpretability of the Logistic Regression model.
   (b) **Normalization for Interpretability (if needed):** While tree-based models like Gradient Boosting and Random Forest are not affected by feature scaling, normalization may still be applied for easier interpretation of variable distributions.
4. **Feature Engineering:**
   (a) **Domain-Specific Features:** Create features based on medical knowledge, such as a composite score indicating the severity of diabetes complications or a derived feature calculating the number of days from symptom onset to hospital admission.
   (b) **Interaction Terms for Logistic Regression:** Consider creating interaction terms between important features (e.g., age and comorbidity status) to improve the performance of the linear model.
   (c) **Feature Selection for Tree-Based Models:** Use feature importance scores from Random Forest or Gradient Boosting to refine the set of features, retaining only those that contribute significantly to the model's performance.
5. **Class Imbalance Handling:**
   (a) **SMOTE (Synthetic Minority Over-sampling Technique):** Generate synthetic samples for the minority class (severe cases) to balance the dataset.

This is particularly useful for improving model performance, especially for Logistic Regression.

(b) **Sample Weighting for Tree-Based Models:** Apply sample weighting to give higher importance to the minority class. Gradient Boosting and Random Forest can handle these weights effectively, ensuring that severe cases are given priority during model training.

4.1.5. Model Development and Training

1. **Model Selection:**

   (a) **Logistic Regression:** Begin with Logistic Regression to establish a baseline model that provides straightforward interpretability and insights into the relationship between variables. This model is ideal for identifying initial risk factors associated with severe COVID-19 outcomes in diabetic patients.

   (b) **Random Forest:** Use Random Forest to capture non-linear relationships and interactions among variables. Its feature importance scores will also help in refining feature selection and understanding key predictors.

   (c) **Gradient Boosting (XGBoost):** Employ XGBoost to achieve higher predictive accuracy by addressing complex patterns in the data. This model is especially effective for imbalanced datasets and provides robust performance through iterative error correction.

2. **Hyperparameter Tuning:**

   (a) **Logistic Regression:** Optimize the regularization parameter (e.g., L1 or L2 penalty) using grid search to prevent overfitting and improve model generalization.

   (b) **Tree-Based Models:** Perform grid search or random search to fine-tune hyperparameters such as the number of trees, maximum depth, learning rate, and subsample ratio for Random Forest and XGBoost. This will ensure the models are well-calibrated for accurate risk prediction.

3. **Model Training:**

   (a) **Data Splitting:** Divide the dataset into training (80%) and testing (20%) sets.

(b) **Cross-Validation:** Perform k-fold cross-validation on the training set to improve the robustness and reliability of the model.

4.1.6. Model Evaluation and Validation

1. **Performance Testing:**

   (a) **Evaluation on Testing Data:** Assess the final models using the testing dataset to gauge their performance in real-world scenarios. This evaluation will help determine the reliability of the model when predicting severe COVID-19 cases among diabetic patients.

2. **Performance Metrics:**

   (a) **Sensitivity (True Positive Rate):** Given the high stakes of missing a severe case, prioritize sensitivity to ensure that high-risk patients are correctly identified. This metric is crucial for the healthcare setting, where early intervention can save lives.

   (b) **F1 Score:** Use the F1 Score to balance precision and recall, especially in handling the class imbalance between severe and non-severe cases.

   (c) **ROC-AUC:** Measure the model's overall ability to discriminate between severe and non-severe cases. A higher ROC-AUC indicates better performance in differentiating risk levels across various thresholds.

3. **Clinical Review:**

   (a) **Collaboration with Healthcare Professionals:** Engage medical experts to review the model's predictions and interpretability. This step ensures that the model's outputs are clinically meaningful and can be integrated effectively into hospital protocols for patient care.

   (b) **Interpretability Tools:** Utilize SHAP (Shapley Additive Explanations) to provide transparent explanations of the model's decisions, facilitating understanding and trust among healthcare providers.

4.1.7. Model Integration and Deployment

1. **API Integration:**

(a) Integrate the predictive model directly with the hospital's Electronic Health Record (EHR) system using secure APIs. This setup will ensure that patient data is automatically processed, and predictions are generated.

2. **User Interface:**

   (a) Develop a simple, user-friendly interface accessible to healthcare professionals. The interface will display prediction results clearly and intuitively, without requiring any advanced technical expertise to operate or interpret the model's output.

4.1.8. Final Reporting and Presentation

1. **Documentation:**

   (a) Develop a **Technical Manual** that provides in-depth information on the model architecture, data handling, and API integration for IT and data science teams.

   (b) Create a **User Manual** designed for healthcare professionals, outlining how to interpret and use the model effectively within the clinical setting.

2. **Launch Event:**

   (a) Organize an opening event to present the functional tool to hospital stakeholders.

   (b) Conduct a live demonstration of the tool, showcasing its user-friendly interface and real-time predictive capabilities.

3. **Adaptation and Training Phase:**

   (a) Following the launch event, begin several weeks dedicated to adapting the tool to the hospital's specific workflow.

4.2. Work Packages

4.2.1. Package 1: Data Acquisition and Integration

- **Tasks:** Data collection, format conversion, data set preparation and consolidation.
- **Duration:** 1 week

4.2.2. Package 2: Data Organization and Context Understanding

- **Tasks:** Data quality evaluation, initial data analysis and metadata handling.
- **Duration**: 2 weeks
  - If metadata is incomplete extend by **1 week**.

4.2.3. Package 3: Data Preprocessing

- **Tasks:** Data cleaning, encoding categorical variables, standardization and normalization of numerical variables, feature engineering and class imbalance handling.
- **Duration:** 3 weeks

4.2.4. Package 4: Model Development and Training

- **Tasks:** Model selection, hyperparameter tunning and model training.
- **Duration:** 1 week.

4.2.5. Package 5: Model Evaluation and Validation

- **Tasks:** Testing the performance of the model and obtaining performance metrics and clinical review
- **Duration:** 1 week.

4.2.6. Package 6: Model Integration and Deployment

- **Tasks:** Integrate the predictive model directly with the hospital's Electronic Health Record System.
- **Duration:** 1 week.

4.2.7. Package 7: Final Reporting and Presentation

- **Tasks:** Develop a technical and user manual that provides the necessary information to interpret and use the model correctly. Organize a launch event and training phase for the usage of the model in hospitals.
- **Duration:** 1 week.

Gantt Chart

The Gantt chart would reflect the alternative paths, highlighting possible extensions in weeks depending on the outcomes of the data acquisition and understanding phases. Milestones and key deadlines will be adjusted accordingly.

Table 4.2.1. Gantt Diagram

Activity	*Weeks*										
	1	2	3	4	5	6	7	8	9	10	11
Data Acquisition and Integration											
Data and Context Understanding											
Data preprocessing											
Model Development and Training											
Model Evaluation and Validation											
Model Integration and Deployment											
Final Reporting and Presentation											

Standard Plan
If metadata is incomplete
(With 1 additional week)

Milestones

- **End of Week 1**: Data acquisition and integration completed.
- **End of Week 3**: Data and context understanding completed.
- **End of Week 6-7**: Data preprocessing completed.
- **End of Week 7-8**: Model development and training completed.
- **End of Week 8-9**: Model evaluation and validation completed.
- **End of Week 9-10**: Model integration and deployment completed.
- **End of Week 10-11**: Final reporting and presentation completed.

4.3. Budget

4.3.1. Human Resources

- **2 Data Scientists**:
  - **Standard Plan (10 weeks)**: 2 * (10 weeks * $1,000/week) = $20,000

  - **Alternative Plan (with 1 additional week)**: 2 * (11 weeks * $1,000/week) = $22,000

- **1 Software Engineer**:
  - **Standard Plan (3 weeks)**: 3 weeks * $1,000/week = $3,000
- **2 Clinical Consultants**:
  - **Standard Plan (6 weeks)**: 2 * (6 weeks * $1,300/week) = $15,600
- **1 Project Manager**:
  - **Standard Plan (10 weeks)**: 10 weeks * $1,200/week = $12,000
  - **Alternative Plan (with 1 additional week)**: 11 weeks * $1,200/week = $13,200

4.3.2. Human Resources Total Costs

- **Standard Plan (10 weeks)**: $50,600
- **Alternative Plan (with 1 additional week)**: $53,800

4.3.3. Computational Resources

- **Servers and Cloud Computing for Development**: $15,000
- **Cloud Maintenance**:
  - **Monthly**: $50
  - **Cloud Engineer (when needed)**: $240/day
- **Software Licenses**: $5,000

4.3.4. Other Costs

- **Staff Training**: $10,000
- **Meetings and Workshops**: $3,000
  - **Additional Meetings** (if needed): Include $2,500 for expert consultations.
  - **Cloud Upkeep** (if needed): Include $240 per day for expert consultations.

Total Estimated Budget

In this total budget it is not considered the possible additional meetings mentioned in Section 4.3.4

- **Standard Plan (10 weeks)**: $83,600
- **Alternative Plan (with 1 additional week)**: $86,800

5. Risk Analysis

In implementing this project, it is essential to identify and mitigate risks that could compromise its success, especially given the high stakes involved in supporting COVID-19 patient outcomes. While every effort will be made to minimize risks, some remain unavoidable due to the complexity and sensitivity of working with medical data and healthcare systems. The following section outlines the most significant risks associated with the project, along with specific strategies for mitigating each, to ensure both data integrity and the reliability of results.

5.1. Data

5.1.1. Data Quality and Availability

Risk Incomplete, missing, or biased data could undermine the model's accuracy, leading to erroneous predictions that may negatively impact clinical decision-making.

Analysis Healthcare data is often fragmented, inconsistent, or contains errors. Given the pandemic context, obtaining uniform data from multiple hospitals can be challenging. Missing values or data bias (such as under-representation of certain demographic groups) could lead to skewed predictions and affect patient outcomes.

Likelihood: *Moderate-High*

- Multiple data sources increase the probability of data quality issues.
- Historical precedent shows data completeness and inconsistencies are a persistent issue in healthcare.

Impact: *Critical*

- Poor data quality directly affects model accuracy and reliability.
- Biased results could disproportionately impact vulnerable populations.
- It could undermine the entire project's utility and trustworthiness.

Risk Score: 12/16 (Critical Priority)

Component	Score
Likelihood	3/4 (High)
Impact	4/4 (Critical)
Final Risk Score	12/16

Table 5.1.1. Risk of Poor Data Quality and Availability

Mitigation :

- **Data Cleaning and Validation**: A thorough data cleaning process will be applied to identify and correct inconsistencies or errors.
- **Imputation Techniques**: Advanced imputation methods such as K-Nearest Neighbors (KNN) will be used to address complex missing data patterns.
- **Multidisciplinary Collaboration**: Close collaboration with medical professionals will ensure the correct interpretation of clinical and demographic data, helping fill in gaps and clarify any ambiguities.

Non-Mitigable Risks Some hidden biases or confounding variables may remain undetected, potentially influencing model results. Additionally, limited access to historical data due to legal or ethical restrictions could affect the model's ability to generalize across different populations.

5.2. Model Risks

5.2.1. Biases and Accuracy Limitations

Risk The model could introduce biases in its predictions or fail to capture critical clinical nuances, leading to suboptimal care decisions for diabetic patients with COVID-19.

Analysis A biased model may overlook high-risk patients, leading to missed opportunities for early intervention, or conversely, misidentify low-risk patients, causing unnecessary use of critical hospital resources.

Likelihood: *High*

- Working with past data, given the evolving nature of COVID-19 and medicine, is bound to limit accuracy.
- Machine learning models commonly struggle with rare but clinically significant cases.

Impact: *Critical*

- Missed high-risk patients could lead to severe health outcomes or mortality.
- False positives could result in unnecessary resource allocation and patient stress.

Risk Score: 16/16 (Critical Priority)

Component	Score
Likelihood	4/4 (Medium)
Impact	4/4 (Significant)
Final Risk Score	16/16

Table 5.2.1. Risk of Model Bias and Limited Accuracy

Mitigation

- **Cross-Validation**: Advanced cross-validation techniques will be used to ensure the model generalizes well to unseen datasets.
- **Continuous Monitoring**: The model will be regularly updated with new data, and its performance will be monitored in real-time to adjust for any deviations in predictions.

Non-Mitigable Risks :
Despite continuous updates, the model may not fully capture the clinical complexity or sudden shifts in disease patterns, such as the emergence of new COVID-19 variants. Predictive models inherently have limitations and 100% accuracy cannot be guaranteed.

5.3. Ethical, Privacy, and Regulatory Compliance Risks

Risk Handling sensitive patient data introduces risks related to data privacy and ensures compliance with regulations such as GDPR (General Data Protection Regulation). A failure to protect patient privacy could result in legal penalties and damage the hospital's reputation.

Analysis This project involves processing highly sensitive data concerning diabetic patients and their health outcomes. Protecting patient privacy is paramount. Any breach could lead to severe legal consequences, undermine patient trust and damage the hospital's reputation.

Likelihood: *Medium (2/4)*

- Existing hospital security protocols and GDPR compliance measures reduce risk.
- Staff training and data handling procedures are already well established.

Impact: *Major* (3/4)

- Privacy breaches would impact hospital reputation and patient trust.
- Regulatory non-compliance could result in significant penalties.
- Project delays possible while addressing compliance issues.

Risk Score: 6/16 (Medium Priority)

Component	Score
Likelihood	2/4 (Medium)
Impact	3/4 (Major)
Final Risk Score	6/16

Table 5.3.1. Risk of Privacy and Regulatory Non-Compliance

Mitigation

- **Regulatory Compliance**: The project will strictly adhere to GDPR and other applicable privacy regulations. Techniques like anonymization and pseudonymization will be employed to ensure patient identities are protected.
- **Informed Consent**: All data used in the project will be covered by appropriate informed consent agreements, with the hospitals being responsible for obtaining and managing these consents. This ensures that patients are aware of how their data is being used, and the hospitals remain liable for compliance with consent requirements.
- **Data Security**: Robust data security measures, including encryption of data both in transit and at rest, will be implemented to safeguard patient information.

Non-Mitigable Risks Even with stringent security measures, there is always a residual risk of data breaches due to advanced cyber threats or human error. These risks, while minimized, cannot be completely eliminated.

5.4. Operational Risks

5.4.1. Model Deployment

Risk While the hospital infrastructure is adequate for the service, altering the implemented system after the project's completion could disrupt functionality. Ensuring system stability is crucial.

Analysis The hospital's existing infrastructure supports the deployment without issue. However, any modifications to the system setup after implementation may affect performance and reliability. Maintaining the system properly and limiting modifications to the designated technician are essential to prevent disruptions.

Likelihood: *Low (1/4)*

- System modifications are typically controlled and well-documented.
- Hospital IT infrastructure follows strict change management protocols.

Impact: *Major (3/4)*

- System disruptions could temporarily affect clinical decision support.
- Recovery procedures would require technical intervention.
- Service interruption might affect multiple departments.

Risk Score: 3/16 (Low Priority)

Component	Score
Likelihood	1/4 (Low)
Impact	3/4 (Major)
Final Risk Score	3/16

Table 5.4.1. Risk of System Modification and Stability Issues

Mitigation

- **System Maintenance**: Ensure that the system is maintained regularly, with clear guidelines specifying that only the designated technician can make modifications or updates.
- **Access Control**: Implement strict access controls to prevent unauthorized changes to the system configuration.

Non-Mitigable Risks Despite precautions, unforeseen circumstances may still necessitate system adjustments, posing a risk to stability. These should be managed carefully under the guidance of the technician.

5.4.2. Human Operation

Risk The effective use of the model depends on hospital staff's ability to use the tool. Insufficient training or resistance to change could hinder adoption and reduce the model's effectiveness.

Analysis Healthcare professionals, while generally adaptable, may face challenges incorporating new tools into their existing workflows. The success of the model heavily depends on proper staff training and acceptance. Staff resistance could stem from time constraints, perceived complexity, or concerns about the tool's reliability. Additionally, varying levels of technical proficiency among staff members might necessitate different approaches to training and support. Without proper adoption, even a technically sound model could fail to deliver its intended benefits in improving patient care.

Likelihood: *Medium-High* (2/4)

- Healthcare staff are accustomed to adopting new methodologies.
- Training programs can be integrated into existing professional development.

Impact: *Moderate* (2/4)

- Reduced adoption would limit the tool's benefits but not cause harm.
- Training issues can be addressed through additional support and documentation.
- Model effectiveness can be gradually improved through user feedback.

Risk Score: 4/16 (Medium Priority)

Component	Score
Likelihood	2/4 (Medium)
Impact	2/4 (Moderate)
Final Risk Score	4/16

Table 5.4.2. Risk of User Adoption and Training Issues

Mitigation

- **Comprehensive Training**: A user manual will be created to guide clinical staff through the use of the tool, and the interface is designed to be simple and intuitive to facilitate ease of use. Any questions or concerns will be addressed during the 4-6 weeks allocated for installation and familiarization with the tool.

Impact\Likelihood	Low	Medium	Medium-High	High
Minor				
Moderate	5.4.1	5.4.2		
Major		5.3		
Critical			5.1.1	5.2.1

Table 5.4.3. Risk Score Matrix.

Non-Mitigable Risks Resistance to change may persist even with extensive training, and human error in using the model cannot be entirely eliminated.

5.5. Other Non-Mitigable Risks

While the project includes comprehensive risk mitigation strategies, some risks remain beyond our control:

- **Virus Evolution**: Unpredictable changes in the virus, such as the emergence of new variants, may affect the relevance and accuracy of the model over time.
- **Socioeconomic and Cultural Factors**: External factors, such as health policies or population behavior can influence outcomes and may not be fully captured by the model.
- **Technological Limitations**: The current state of technology may impose limits on achieving the project's full objectives.
- **External Collaboration**: Success depends partially on the collaboration of hospital staff and other stakeholders, whose engagement and cooperation may vary.

By addressing these risks with targeted mitigation strategies, we aim to minimize the potential impact on the project while maintaining open communication with stakeholders to ensure continuous improvement and problem-solving throughout the project's lifecycle.

6. Viability Analysis

Technological Feasibility

The project is technologically feasible, given the current availability of data, proven machine learning techniques, and the hospital's technical infrastructure.

Clinical, demographic, and hospital admission data for diabetic patients aged 40 to 60 are accessible and ready for use, under the conditions outlined in Section 3.2.2. The proposed machine learning models, such as Logistic Regression, Random Forest and XGBoost are well-documented for their high performance in predictive healthcare applications, ensuring reliable and robust results.

The hospital's existing computational resources are adequate to handle the model's training and deployment requirements. Additionally, the integration of the predictive model into the hospital's Electronic Health Record (EHR) system via APIs is both viable and efficient. This integration aims to deliver predictions with minimal delay, helping healthcare professionals make timely and well-informed clinical decisions.

Financial Feasibility

As detailed in section 2.3 with the **ROI** calculation, the project is financially feasible. Even in the worst-case scenario of the predictions, a positive and significant ROI is reported, indicating that for every euro invested, the returns will exceed the initial cost. Moreover, the total project cost, amounting to X, is considerable but reasonable for a data science initiative in the medical field, especially when taking into account the substantial savings and efficiency improvements anticipated for the hospital system.

Operational Feasibility

Operationally, the project is meticulously designed to integrate seamlessly into the hospital environment and daily workflows. The predictive model will prioritize ease of use, ensuring that medical staff can effectively incorporate it into their routine without disruption. This integration will be achieved via API connections to the hospital's existing systems, allowing secure and efficient access to the model through a user-friendly interface. Additionally, as previously mentioned, comprehensive training will be provided to medical staff to ensure proper usage and interpretation of the model's predictions. To further support seamless operation, a dedicated technician will be available to handle ongoing maintenance and address any technical issues or failures.

Conclusion

The project is feasible and well-suited to address the need for timely and accurate predictions of severe cases. By leveraging existing data and infrastructure, it provides a practical solution that can support medical staff in making informed decisions, ultimately helping to improve patient care and manage hospital resources efficiently.

References

[18a] *General Data Protection Regulation (GDPR)*. 2018. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[18b] *Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales (LOPDGDD)*. 2018. URL: https://www.boe.es/eli/es/lo/2018/12/05/3.

[Ali+22] Yousef Alimohamadi et al. "Hospital Length of Stay for COVID-19 Patients: A Systematic Review and Meta-Analysis". In: *PubMed Central* (2022). Accessed: 2024-11-03. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC9472334/.

[Dia24] Sociedad Española de Diabetes. *España es el segundo país con mayor prevalencia de diabetes de Europa*. 2024. URL: https://www.sediabetes.org/comunicacion/sala-de-prensa/espana-es-el-segundo-pais-con-mayor-prevalencia-de-diabetes-de-europa/.

[Fed21] International Diabetes Federation. *COVID-19 & Diabetes - International Diabetes Federation*. 2021. URL: https://www.idf.org/about-diabetes/covid-19-and-diabetes/.

[Med24] Weill Cornell Medicine. "Study reveals how COVID-19 infection can cause or worsen diabetes". In: *Cornell Chronicle* (2024). URL: https://news.cornell.edu/stories/2024/09/study-reveals-how-covid-19-infection-can-cause-or-worsen-diabetes.

[Min17] Consumo y Bienestar Social Ministerio de Sanidad. *Encuesta Nacional de Salud de España 2017 - Nota Técnica*. 2017. URL: https://www.mscbs.gob.es/estadEstudios/estadisticas/encuestaNacional/encuestaNac2017/ENSE2017_notatecnica.pdf.

Universidad Politécnica de Madrid

ETSI **SISTEMAS INFORMÁTICOS**