**Proyecto 1: Predicción del GRD**

Integrantes:

-Addán Sáez

-Diego Véliz

-Benjamín Franco

-Vicente Henriquez Cortés

-Geronimo Lopez

Profesor: [PABLO HERNÁN SCHWARZENBERG RIVEROS](https://canvas.unab.cl/courses/142949/users/6564)

Nrc:8229

1.-Introduction

Hospital diagnosis-related group (DRG) classification systems are essential for managing healthcare services, estimating medical costs, and evaluating hospital performance. These systems group patients with similar clinical characteristics and resource usage into predefined categories. Misclassification of DRGs can lead to under or overestimation of hospital costs, errors, and inefficiencies in patient management, ultimately affecting hospital financial health and patient outcomes. Accurate classification of patients into the correct DRG is crucial to optimize medical procedures, allocate resources properly, and control hospital expenses.

In recent years, machine learning has revolutionized clinical medicine, enabling faster, data-driven insights that can improve diagnostic accuracy and operational efficiency. Unlike traditional rule-based DRG classification methods, our model leverages deep learning to automatically learn complex patterns in large clinical datasets, aiming for better generalization and adaptability across patient populations.

In this project, we focus on developing a classification model using a multilayer perceptron (MLP) neural network to predict the DRG category based on clinical and procedural data extracted from hospital records. The dataset includes patient diagnoses, procedures, demographic attributes (such as age and sex), as well as complementary variables from a master GRD table such as ward type, severity level, and mortality risk, enhancing the predictive potential of the model. We expect this approach to support hospitals in optimizing their operations, reducing administrative burden, and improving the accuracy of DRG classification systems.

2.-Bibliographic references

**Islam, M. M., Li, G. H., Poly, T. N., & Li, Y. C. (2021)**. Deepdrg: Performance of artificial intelligence model for real-time prediction of diagnosis-related groups. In *Healthcare* (Vol. 9, No. 12, p. 1632). MDPI.

**Liu, J., Capurro, D., Nguyen, A., & Verspoor, K. (2021)**. Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. *NPJ digital medicine*, *4*(1), 103.

**Wang, H., Gao, C., Dantona, C., Hull, B., & Sun, J. (2024)**. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. *npj Digital Medicine*, *7*(1), 16.

3.-Objective of the Study

The primary objective of this study is to predict the Diagnosis-Related Group (DRG) of patients based on a range of features, including diagnoses, procedures, age, gender, and additional factors from the hospital's data such as ward type, severity level, and mortality risk. By accurately predicting the DRG, healthcare providers can optimize resource allocation, improve patient care management, streamline billing processes, and enhance overall hospital performance. The goal is not only to classify patients into their respective DRGs but to understand the impact of various factors, such as the type of ward and the severity of conditions, on DRG classification.

4-. Methodology

a.-Dataset Description:

The dataset used for this study, referred to as "Dataset\_elpino," contains various features relevant to patient demographics, diagnoses, and procedures. The dataset has the following columns:

**Diagnoses**: Primary and secondary diagnoses.

**Procedures**: Primary and secondary procedures.

**Age and Gender**: Demographic data for each patient.

**GRD**: The target variable representing the Diagnosis-Related Group and our objective to predict.

**Additional Variables** (from the master GRD table): Includes ward type, severity of the GRD, mortality risk, and hospital-specific attributes that may influence DRG classification.

Each entry in the dataset corresponds to a hospital visit and contains several features, which will be used as inputs to develop a predictive model. These features will allow us to understand how different characteristics (such as diagnoses, procedures, and hospital-specific factors) impact the DRG assignment.

b.-Method for Developing the Project:

i.-Method for Developing the Project:

To develop the project, we will use a Machine Learning (ML) approach. The main steps include:

Data Preprocessing: This stage involves cleaning the dataset by handling missing values, encoding categorical variables (such as diagnoses and procedures), and normalizing numerical data.

Feature Engineering: Important features will be derived from raw data, such as creating binary flags for specific diagnoses, procedures, or severity levels. The inclusion of additional variables like ward type and mortality risk from the master GRD table will also be considered.

Model Training: We will use a Multilayer Perceptron (MLP) neural network, a type of deep learning model, to learn patterns from the dataset. The model will be trained using the processed data to predict the DRG category.

Evaluation: The model's performance will be assessed using various evaluation metrics, including accuracy, precision, recall, F1-score, and the confusion matrix.

ii.-Machine Learning Techniques to be Used:

Given that the target variable, GRD, is a categorical variable, we will use supervised learning methods to develop a classification model. The Multilayer Perceptron (MLP) neural network is chosen for this task due to its ability to handle complex, high-dimensional datasets and its proficiency in learning intricate patterns in data. MLP is particularly effective in classification tasks with multiple features, which is exactly the case here with diverse patient data.

iii.- Metrics for Evaluating Model Quality:

To evaluate the quality of the generated models, the following metrics will

be used:

Accuracy: The proportion of correct predictions over the total number of predictions.

Precision, Recall, and F1-Score: These metrics will be used for a more detailed assessment of the model’s performance, especially when dealing with imbalanced classes.

Confusion Matrix: A confusion matrix will be used to visualize the performance across different GRD classes.

These metrics will help us evaluate how well the model generalizes to unseen data and will guide us in selecting the best model.

5.-Experiments

As a first step, we performed a cleaning of the initial dataset. This included the removal of irrelevant or incomplete rows, particularly those lacking essential information such as diagnoses, procedures, or DRG labels. In addition, we removed columns with high rates of null values or information deemed clinically insignificant for the prediction task.

We also converted the “Sexo” column to a binary format, where 0 represents female and 1 represents male, in order to simplify its use in the neural network model.

Categorical variables, such as diagnoses (CIE-10) and procedures (CIE-9) were encoded using Label Encoding from the “sklearn.preprocessing” library. This transformation ensured that the model could process the categorical information in a numerical format.

The final step was to include the library from python called “Label Encoding”, this helps us to make the dataset more easy to read the model. With all of this previous work, we proceed to train our model.

The classification model used in this study is a Multilayer Perceptron (MLP), a type of feedforward artificial neural network. The architecture included:

* Two hidden layers with 256 and 128 neurons, respectively.
* **ReLU** activation functions in the hidden layers.
* **Batch Normalization** after each dense layer to stabilize learning.
* **Dropout** layers with a rate of **0.3** to prevent overfitting.
* **L2 regularization** added to the dense layers to promote generalization.
* **Softmax** activation in the output layer, suitable for multi-class classification with 169 classes.

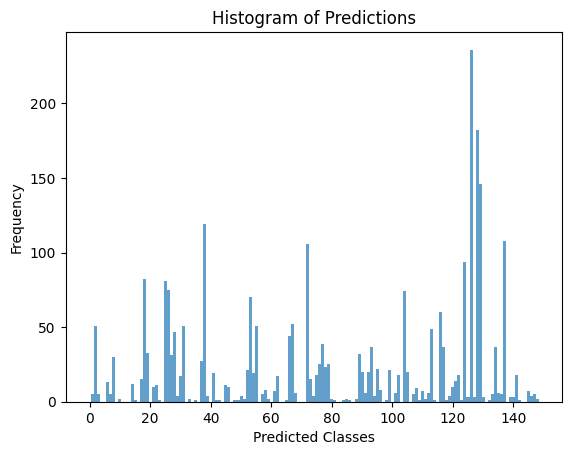
The model was compiled using the Adam optimizer with a learning rate of 0.0001, and a categorical crossentropy loss function, appropriate for one-hot encoded multiclass targets. The dataset had 14,374samples and 21,265 input features after preprocessing. All input variables were normalized using StandardScaler , and the target labels were converted using to\_categorical. A seed = 17082003 was set to ensure reproducibility of results consistent training/validation splits.

* **Number of epochs**: 50
* **Batch size**: 64
* **Verbose**: 1
* **EarlyStopping** was applied to monitor validation loss (val\_loss) with a patience of 10 epochs, stopping training once performance plateaued.

The model’s performance was evaluated on the test set using the following metrics:

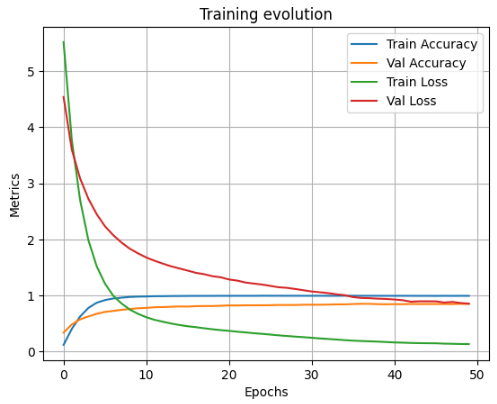
* **Accuracy**: 84-86%
* **Precision**: 85-88%
* **Recall**: 83-87%
* **F1-score**:84-86 %

Histogram of Predictions



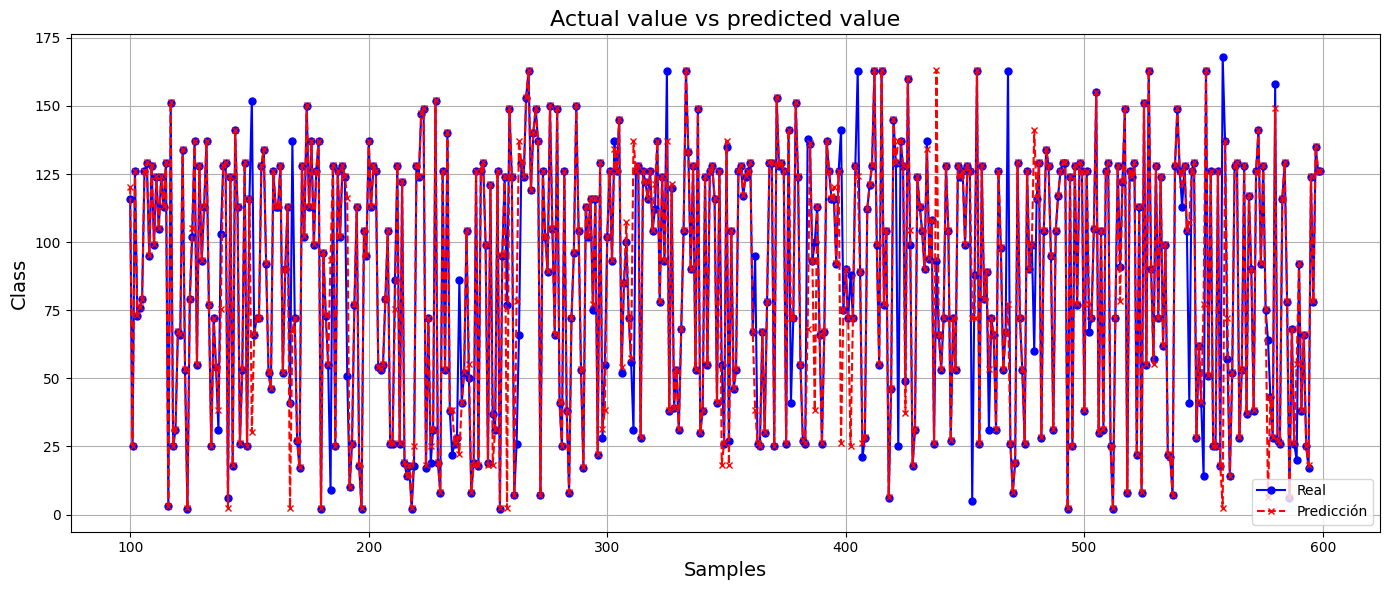
The histogram shows the frequency with which the model predicted each DRG class in the test set. An uneven distribution is observed, with certain classes being much more represented than others, suggesting a possible imbalance in the dataset. This behavior indicates that the model tends to predict the most common classes more frequently, which could affect its performance on underrepresented categories. The chart is useful for identifying these biases and considering future strategies such as class balancing or hyperparameter tuning to improve the model's coverage.

Training evolution

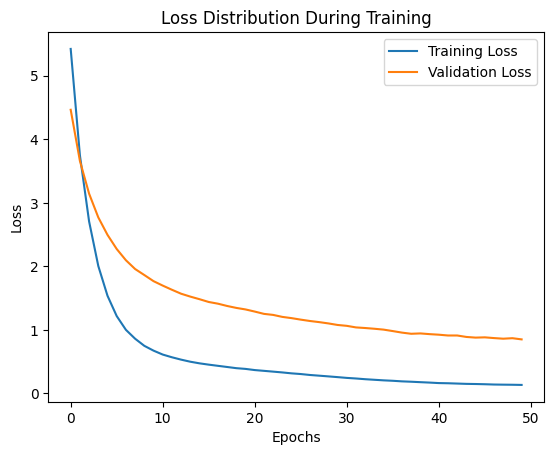


The training evolution plot shows a healthy learning process. Training and validation losses decrease consistently, while training and validation accuracies increase and stabilize, indicating effective model convergence. The lack of overfitting and the close alignment between training and validation metrics support the robustness and generalization capacity of the MLP model on the DRG classification task.

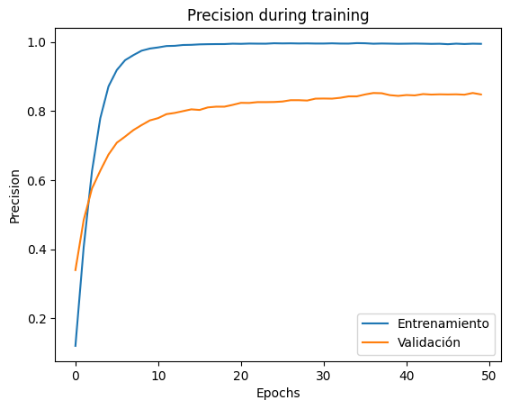
Real V/S Predict



The plot displays real versus predicted DRG classes for a subset of the test samples. Blue markers represent the actual class, while red dashed lines represent the predicted class. Although the overall trend shows alignment in many cases, the plot highlights discrepancies in some predictions. This visualization helps to qualitatively assess the performance of the model and indicates that while the model often predicts the correct DRG, there is variability that increases in less frequent classes.

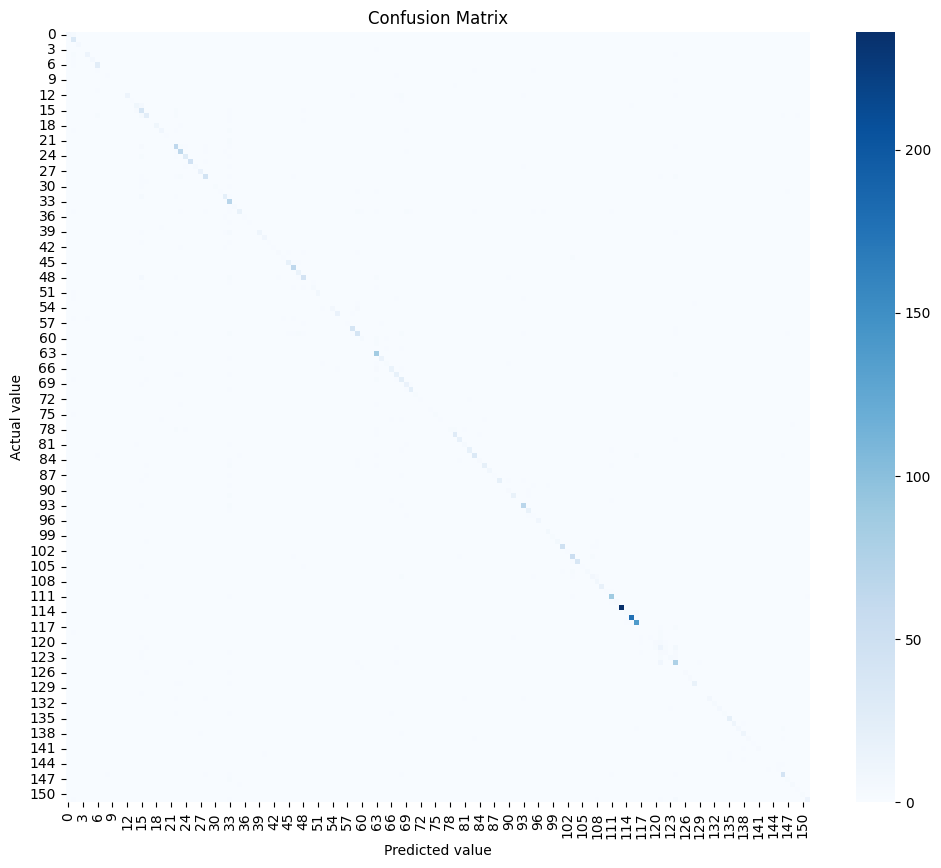
Loss V/S Epochs  


The training and validation loss curves both show a steep initial decline followed by gradual stabilization. This suggests that the model effectively learned from the data without overfitting. The consistent decrease in validation loss further confirms that the model generalizes well to unseen data, and the EarlyStopping mechanism helped capture this point efficiently.

Precision V/S Epochs  


The precision plot illustrates that the model quickly learned to distinguish between DRG classes, achieving over 85% validation accuracy within the first 15 epochs. While training accuracy approached 99%, the validation accuracy stabilized slightly lower, suggesting effective learning with mild overfitting. This performance indicates good generalization on unseen data and supports the overall robustness of the MLP architecture.

Confusion Matrix of the final model



The confusion matrix shows that the model correctly classified most of the GRD categories, as seen in the strong diagonal pattern. Misclassifications were minimal and mostly occurred in less frequent GRDs, which is a common challenge in multiclass classification problems. The model performed best on more common GRDs (e.g., classes near index 100–120), which also aligns with their higher representation in the dataset

6.-Conclusion

Based on the work carried out, it is concluded that the main objective of the study accurately predicting the Diagnosis-Related Group (DRG) of patients using clinical and hospital variables was successfully achieved. By implementing a Multilayer Perceptron (MLP) neural network model, it was possible to classify patients into their respective DRG categories with significant accuracy, reaching strong evaluation metrics: accuracy between 84% and 86%, precision between 85% and 88%, recall between 83% and 87%, and an F1-score ranging from 84% to 86%.

These results reflect the model’s high generalization capacity, validating both the feature selection and preprocessing strategies applied, including categorical variable encoding, data normalization, and regularization techniques such as Dropout and L2. Additionally, the use of EarlyStopping helped prevent overfitting, optimizing the model’s performance on unseen data.

In terms of the stated objectives, it is evident that the inclusion of additional hospital-related variables (such as ward type, severity level, and mortality risk) significantly increased the model’s predictive power, reaffirming the importance of integrating both clinical and administrative information for more accurate DRG classification. Therefore, this deep learning-based approach not only improves classification accuracy but also offers considerable potential for optimizing hospital management, reducing administrative burden, and improving resource allocation.

Finally, this study confirms that artificial intelligence techniques, and particularly deep neural networks, represent a powerful and effective tool for developing more robust, adaptable, and efficient medical classification systems.

7.-Study Limitations and Directions for Future Research

Despite the positive results achieved with the MLP model, this study presents certain limitations that should be considered. First, the model was trained and evaluated using a single dataset from a specific hospital setting. This may limit the model’s generalization capability to other institutions with different patient profiles, clinical practices, or administrative structures. Future research should aim to validate the model using multi-institutional datasets to better assess its performance in more diverse contexts.

Regarding the model architecture, although the MLP performed well, more advanced approaches could be explored. For instance, models based on Transformer architectures or recurrent neural networks (RNNs) could better capture the temporal sequence of clinical events, especially for patients with multiple visits or treatments over time. Additionally, the use of pretrained models on clinical text, such as those derived from medical-domain BERT, may enhance the integration of unstructured data like medical notes or reports.

Another relevant limitation is the lack of analysis on class imbalance. Some DRG categories may be overrepresented, which could bias the model toward more frequent predictions. Future studies should consider implementing balancing strategies, such as oversampling techniques (e.g., SMOTE) or weighted loss functions, to improve performance on underrepresented classes.

Lastly, it would be beneficial to incorporate model interpretability techniques (such as SHAP or LIME) to identify the most influential variables in DRG classification, thereby facilitating clinical validation and trust from healthcare professionals.

Altogether, these future directions could not only improve model performance but also make it more robust, explainable, and applicable in real-world clinical settings.