Facultad de Ingeniería
Ingeniería Civil Informática

Project Module 1

Jara Tripailaf, Benjamín
Silva Peña, Diego

Teacher:
Schwarzenberg Riveros, Pablo
Machine Learning

Santiago of Chile, 2023

Index

Resume

The present report addresses a problem in which a predictor of DRG (Diagnosis Related Groups) must be built from available patient information (symptoms, procedures, sex, and age) by performing an exploratory data analysis, generating the datasets to work with, and subsequently defining metrics to evaluate the models to be used.

Introduction

Currently, the healthcare industry faces a growing demand for medical care, an increase in costs, and increasing pressure to improve the quality of care. In this context, machine learning has established itself as a fundamental tool for processing and analyzing large amounts of data in the field of healthcare. Machine learning is a branch of artificial intelligence that focuses on the development of algorithms and models that can learn from data and improve their accuracy over time.

One of the biggest challenges in the healthcare industry is the accurate prediction of Diagnosis Related Groups (DRGs), which are used to establish the payment rate for medical services provided in a hospital. Accuracy in predicting DRGs is essential to control costs and improve the quality of medical care. In this context, the use of neural networks has become a promising approach to predict patient DRGs. Neural networks are a machine learning model that mimics the functioning of the human brain and allows for the identification of complex patterns in data.

In this report, we present a neural network-based approach to predict DRGs in a hospital using patient data. We will discuss the technical details of this approach, present the results obtained, and compare them with traditional

approaches used in the healthcare industry. Additionally, we will explore the potential of neural networks to improve clinical and administrative decision-making in healthcare. This report aims to provide an overview of the use of neural networks in DRG prediction and its relevance to improving the quality of medical care in the current context.

Development

Similar problems

Machine learning has become an essential tool for solving technological needs, but the healthcare industry uses these technologies to solve management problems and thus strengthen the procedures that lead to patient diagnoses.

We can see this in the healthcare field, where it allowed for the classification of patients who come to the consultation with symptoms of dyspepsia into two groups: those who are very likely to have peptic ulcer disease or gastroesophageal reflux (GER) and those who are very likely to have functional or idiopathic dyspepsia. All thanks to neural networks. (1)

Another example related to the topic of this report is what Amalfi Analytics, located in Barcelona, did, whose purpose is to develop healthcare management support tools using Artificial Intelligence and Machine Learning techniques. One of the tools they use is the use of clusters that identify comorbidity diagnosis groupings by identifying patterns beyond the primary diagnosis or a related diagnosis group (DRG). (2)

Dataset

The dataset in question consists of information on diagnoses, procedures, age, gender, and DRG of each patient.

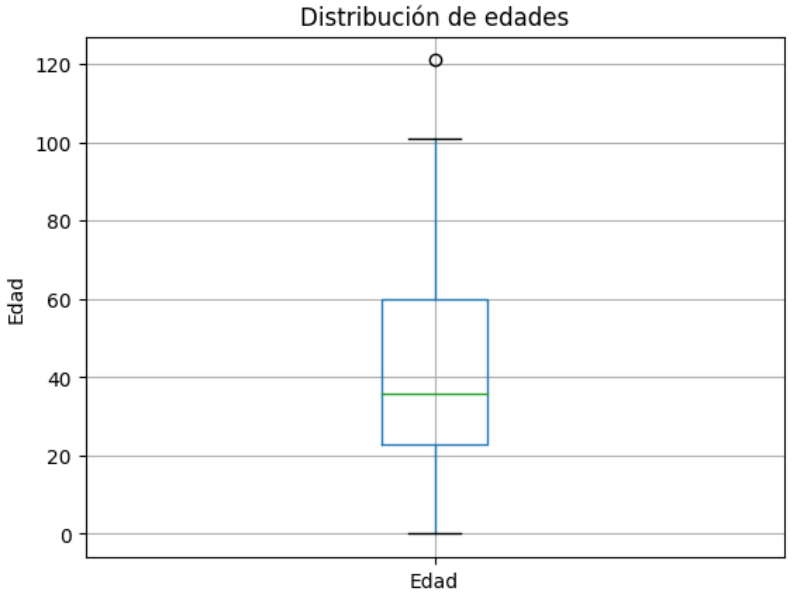
DRG stands for "Diagnosis Related Groups" and is a patient classification system used for the programming and management of hospital care. The system divides patients into groups based on their diagnosis, with the goal of improving the efficiency of medical services and increasing the quality of care. This can help hospitals to more efficiently schedule procedures, allocate resources more effectively, and improve communication among different healthcare professionals. The DRG system is based on classifying patients into groups with similar diagnoses and similar care requirements.

Variable	Type	Description
Gender	qualitative	Gender of the patient
Age	discrete quantative	Age of the patient
	qualitative	Corresponde al diagnóstico principal que presenta el paciente
Diagnóstico Secundario (n)	qualitative	Corresponde a los diagnósticos secundarios que presenta el paciente
Procedimiento Principal	qualitative	Corresponde al procedimiento principal que se le hizo al paciente
Procedimiento Secundario (n)	qualitative	Corresponde a los procedimientos secundarios que se le hicieron al paciente
DRG	qualitative	Corresponds to the dependent variable, which is the Diagnosis Related Group (DRG)

Descriptive statistics of the data

For the exploratory data analysis, the different variables in the dataset were analyzed. The first change made was to rename the columns so they could be easily identified. After this, the text within the variables was removed, leaving only the code, as this is what will be used for the model. This was done for all variables of Diagnoses and DRG. The Sex variable was transformed into a categorical variable, where the value 0 was assigned to Male and 1 to Female. This was done using the Label Encoder method, and finally, all variables were transformed into a factor type, except for age, which is a quantitative variable.

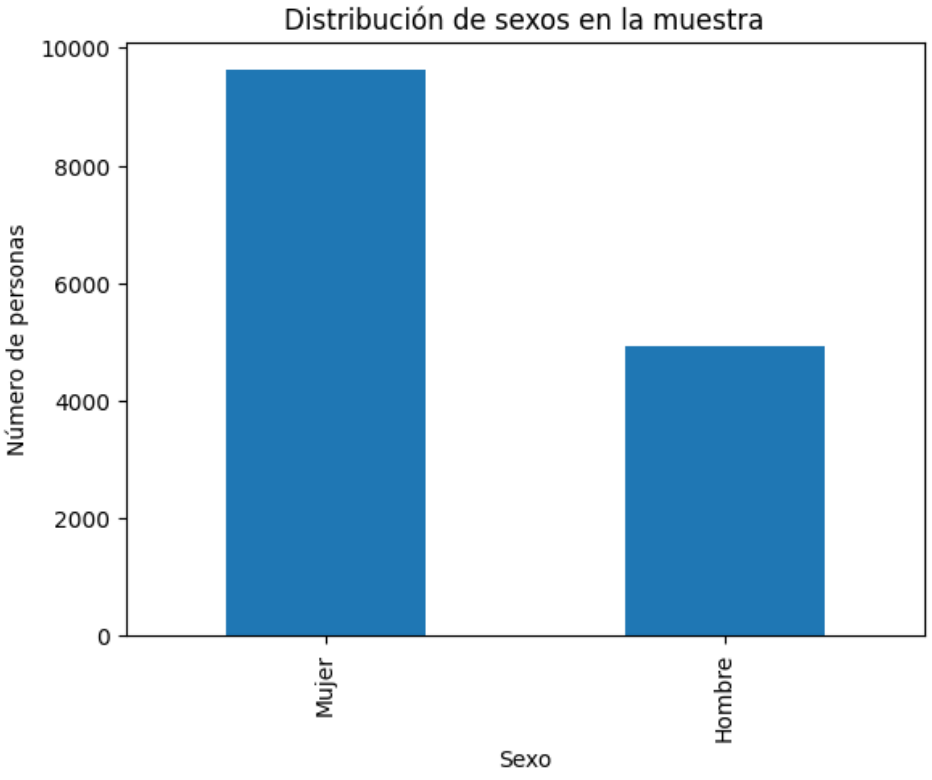
When analyzing the age variable, which is the only quantitative variable, the following measures and a graph were obtained.



```
count    14561.000000
mean        39.426550
std         24.681545
min          0.000000
25%         23.000000
50%         36.000000
75%         60.000000
max        121.000000
Name: Edad, dtype: float64
```

From what we can conclude, the mean age is approximately 39 years old, and the majority of the data is within the age range of approximately 25-60 years. We can also see that there is an age of 120 years, which corresponds to an outlier.

Regarding the sex variable, we have that:




```
count     14561
unique        2
top       Mujer
freq       9617
Name: Sexo, dtype: object
```

There are 9617 women and 4944 men, so there is a wide difference between the sexes. From this, it can be inferred that in the database provided by Hospital El Pino, the majority of their patients correspond to women.

Regarding the main diagnoses, we have that:

```
count     14561
unique     1491
top        O70.0
freq         779
Name: Diag_principal, dtype: object
```

There are 1491 different diagnoses and O70.0 is the most frequent, corresponding to perineal tears that occur during childbirth when the baby's head is too large for the vagina to stretch.

Regarding the main procedures, we have that:

```
count                                          14561
unique                                           528
top       73.59 - PARTO ASISTIDO MANUALMENTE.OTRO
freq                                            1648
Name: Proced_principal, dtype: object
```

There are 528 different procedures and 73.59 is the most repeated, corresponding to assisted delivery.

Regarding the dependent variable DRG, we have that:

```
count     14561
unique      210
top       14610
freq       1220
Name: GRD, dtype: object
```

We can see that there are 210 different DRGs, where the most repeated one is 14610, which corresponds specifically to cesarean delivery, varying between its levels of severity. Due to this, the high difference between the sexes of the patients can be explained, since a large part of the DRGs in the

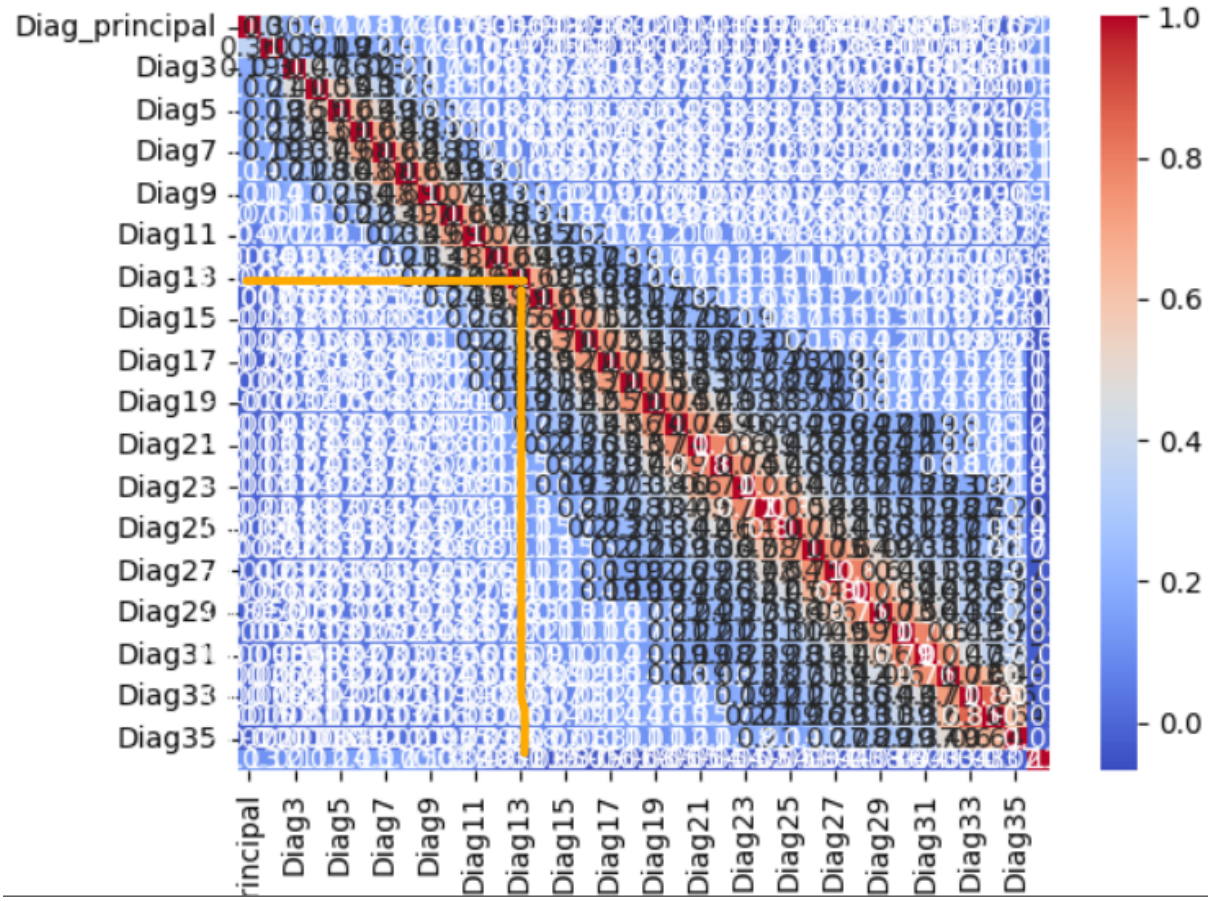
database correspond to women giving birth, whether by vaginal delivery or cesarean section.

Features used for the model

Initially, the variables sex and age were left undisputed, as these two variables are necessary to predict a patient's data, as we need them to calculate the patient's DRG, severity, duration of stay in the hospital, etc. These are crucial data, but the number of diagnoses and procedures is questionable. We need to see how many of these are necessary to obtain good predictions since not all patients will have 35 diagnoses and 30 procedures. Therefore, in this case, less can be more.

It is worth noting that the diagnoses and procedures also need to have their mandatory permanence, as it corresponds to the most important of these.

With this, only the permanence of secondary diagnoses and procedures needs to be analyzed. For this, we used a correlation graph (heatmap) between the different types of diagnoses, which demonstrates the degree of correlation between them.



We can see that up to diagnosis 15, there is a high level of correlation, but beyond 15 the correlation decreases through a darker shade, indicating less correlation. Therefore, only 15 diagnoses were used.

Data partition

Taking this into account, the data sets were defined to be used for training, validating, and testing the model to isolate the random relationships between variables. It is worth noting that it is important to choose an appropriate training set based on the problem, taking into account the proportions relative to the case to avoid overtraining or "overfitting".

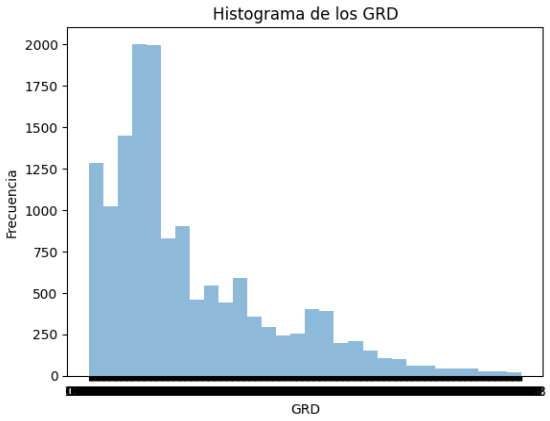
Considering this, it was decided to work with the dataset by dividing the data into 65% for training, 20% for testing, and 15% for validation.

Partición de Datos

```
[13] X = df1.iloc[:, :-1]
     y = df1.iloc[:, -1]
```

Data balance

Al evidenciar un desbalance en las clases se procede a analizar la cantidad de variables por cada clase:




```
Desbalanceo de datos en la columna GRD

(14561, 33)
14610    1220
14612     927
14613     741
07114     501
13416     458
          ...
08420       1
10110       1
10120       1
12111       1
01120       1
Name: GRD, Length: 210, dtype: int64
```

For the "RGD" column, the Pandas function pd.value_counts is used to identify the number of values for each class. From this, it can be concluded that there are 1220 records for the "RGD" 14610, while there is only 1 record for the RGD 08420.

This leads to the conclusion that there is class imbalance, so we proceed to balance it using the "Oversampling" method, which seeks to duplicate samples in the minority classes.

```
from imblearn.over_sampling import RandomOverSampler
from sklearn.datasets import make_classification

# crear una instancia de RandomOverSampler
ros = RandomOverSampler(random_state=0)

# aplicar oversampling a los datos
Xtrain_B, Ytrain_B = ros.fit_resample(Xtrain, Ytrain)
print(f"Tamaño de los datos antes de oversampling: {Xtrain.shape}")
print(f"Tamaño de los datos después de oversampling: {Xtrain_B.shape}")
```

```
Tamaño de los datos antes de oversampling: (8736, 32)
Tamaño de los datos después de oversampling: (152096, 32)
```

When applying the Oversampling method, an increase in data is demonstrated in order to balance the minority classes.

Metric used

Comparison metrics allow us to evaluate the quality of models to choose the best one or to improve the trained model based on these results. Each comparison metric provides a result that is important to analyze, as using different metrics for performance evaluation avoids falling into problems of poor predictions when the model is deployed on unseen data.

This is why the decision is made to use the evaluation metric Accuracy, where the results obtained measure how often the classifier is correct since it is necessary to accurately determine the GRD code for each patient. In this context, applying the formula:

Accuracy = (TP+TN)/Total

From this result, the correct predictions made are concluded as a proportion of all the predictions made

Structures

Initially, we searched for the ideal number of layers with a constant number of neurons, in order to test the number of layers first rather than the number of neurons. The result obtained was that the ideal number of layers was 3, based on the best accuracy obtained in these tests, with a maximum accuracy of 0.7. We defined three models with 3 layers each and 100 epochs, where a grid search was performed to find the appropriate number of neurons for each of these layers.

Model	Neurons	Accuracy
1	80, 240, 120	0.94
2	120, 560, 240	0.96
3	100, 620, 80	0.95

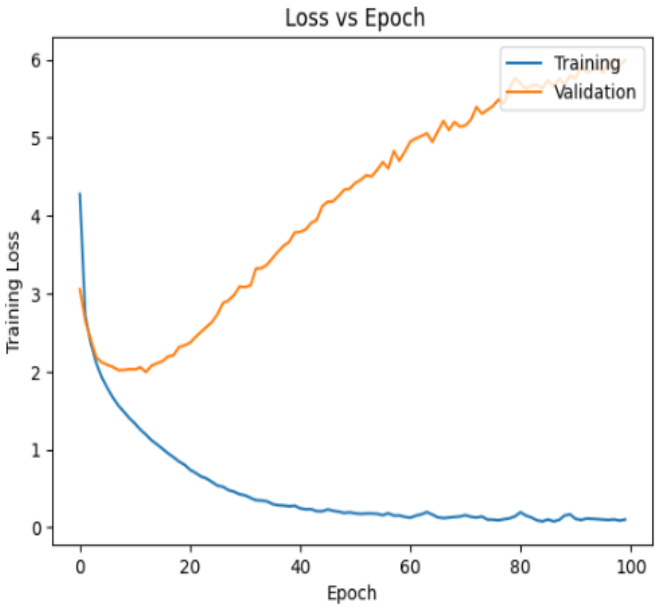
It can be observed that model 2 has the best accuracy.

Results comparison

The technique used to compare the results obtained from the neural model will be Supervised Learning since we have a set of labeled data.

By analyzing the type of label, it is possible to identify that the problem is a classification model where the discrete label corresponds to the "DRG" variable. The results of the neural model were compared using different metrics previously described, comparing the convergence of the graphs, the optimizers through the gradient and the derivative of the activation function in order to perform an analysis on the accuracy of the predictions delivered by the trained model.

Para analizar el posible sobreajuste, se obtuvo este gráfico:



From the analysis, it is observed that due to the increase in loss as the epochs increase, it can be concluded that the model has overfit. This happens when the model may have memorized the training data instead of learning relevant patterns that can be applied to new data.

Conceptual description of neural net

A multiple classification neural network is a mathematical model that attempts to learn patterns in data to make accurate predictions about new data presented to it. In the specific case of predicting a patient's RGDin a hospital, the neural network analyzes input data such as the patient's gender, age, procedures, and diagnoses to determine which DRG category it belongs to.

To enable the neural network to perform this task, pre-labeled data with the corresponding DRG for each patient is needed. During the training process, the neural network adjusts its internal parameters and connections iteratively until its ability to predict the DRGis correct.

Once the neural network has been trained, it can be used to make predictions about new patients in the hospital. When information about a patient is entered, the neural network processes that information through its multiple layers of neurons to produce a prediction of the DRG that best fits the input data.

Alternative of NN model: Decision tree

The decision tree model is used as a comparative model. It consists of each internal node representing a question about a feature of the input data, and each branch that comes out of the node represents a possible answer to the question. The leaf nodes of the tree represent the final decisions or predictions.

During the process of constructing the tree, the best question is sought to divide the data into more homogeneous subsets. This is done by measuring the purity of the resulting subsets using an impurity function such as information gain.

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

arbol = DecisionTreeClassifier()
arbol.fit(Xtrain, Ytrain)

# hacer predicciones en el conjunto de prueba
Ypred = arbol.predict(Xtest)

# calcular la precisión de las predicciones
accuracy = accuracy_score(Ytest, Ypred)
print('Precisión: {:.2f}'.format(accuracy))
```

```
Precisión: 0.81
```

When running the decision tree, an accuracy of 80% is evidenced, which means that 80% of test instances were correctly classified by the model. However, this model does not surpass the accuracy obtained by the neural network model.

Ways of improvement the model

To improve the model, there is a wide range of options:

- Reduce the amount of irrelevant data provided.
- Test different model architectures to optimize hyperparameters and obtain the model with the best configurations.
- Add new data to increase the sample size, which can help improve the model's performance.
- Build two prediction models: one for predicting the DRG of patients with common symptoms (such as childbirth or COVID), and another for those with less common symptoms (such as stroke or chronic diseases). This would prevent all data from being combined in one model, which could lead to imbalanced data and hinder accurate predictions for less common DRGs.

Conclusion

In conclusion, when modeling a neural network, the initial phase of data cleaning is crucial. This involves analyzing correlations, removing unnecessary columns, and so on. Therefore, it is essential to give this phase the attention it deserves, as skipping this step can make it difficult to read the data and train the model satisfactorily.

It is also worth noting that having a large number of columns, as in the dataset studied, can complicate the entire first stage, as it can become cumbersome to work with. Thus, it is important to reduce the number of columns in such situations. Additionally, when dealing with only qualitative variables, as is the case in this study (excluding Age), it can be challenging to obtain meaningful insights from graphs and descriptive statistics. In contrast,

quantitative variables allow for analysis of measures such as the mean, quartiles, and standard deviation.

Finally, it is important to recognize that the modeling process is iterative, and the search for the best model configuration may require repeating the process multiple times. This involves a thorough search for the best model through a trial-and-error approach.

Bibliography

(1) (S/f). Nih.gov. Recuperado el 3 de abril de 2023, de https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7679651/#:~:text=La%20red%20neuronal%20proporciona%20excelentes,enfermedad%20ulcerosa%20p%C3%A9ptica%20o%20RGE

(2) *Aplicación de machine learning en la gestión de las personas con enfermedades crónicas*. (2020, septiembre 23). Hospitecnia. https://hospitecnia.com/tecnologia/inteligencia-artificial/aplicacion-machine-learning-gestion-personas-enfermedades-cronicas/

-Aplicación de machine learning en la gestión de las personas con enfermedades crónicas. (2020, septiembre 23). Hospitecnia. https://hospitecnia.com/tecnologia/inteligencia-artificial/aplicacion-machine-learning-gestion-personas-enfermedades-cronicas/

-Desgarros vaginales durante el parto. (2022, enero 20). Mayo Clinic. https://www.mayoclinic.org/es-es/healthy-lifestyle/labor-and-delivery/multimedia/vaginal-tears/sls-20077129

-Notas de evaluación. (s/f). Astursalud.es. Recuperado el 2 de mayo de 2023, de https://enotas.astursalud.es/-/usos-del-cmbd-los-grupos-relacionados-por-el-diagn%C3%B3stico-grd-ii