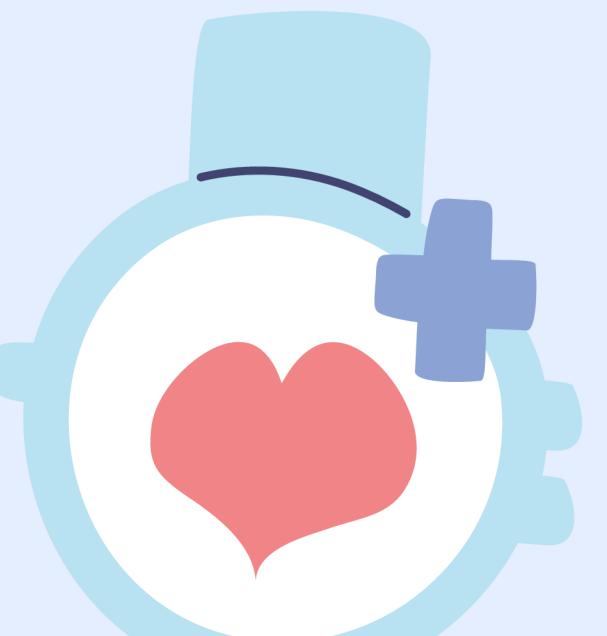


Fibropred: Models predictius

Carla Atienza
Marta Carrión
Alba Figueras
Eva Martínez



Variables eliminades

Severity of telomere shortening - Transform 4

Transplantation date

Pedigree

Death i Cause of death

Pathology pattern Binary i Pathology pattern UIP, probable or CHP

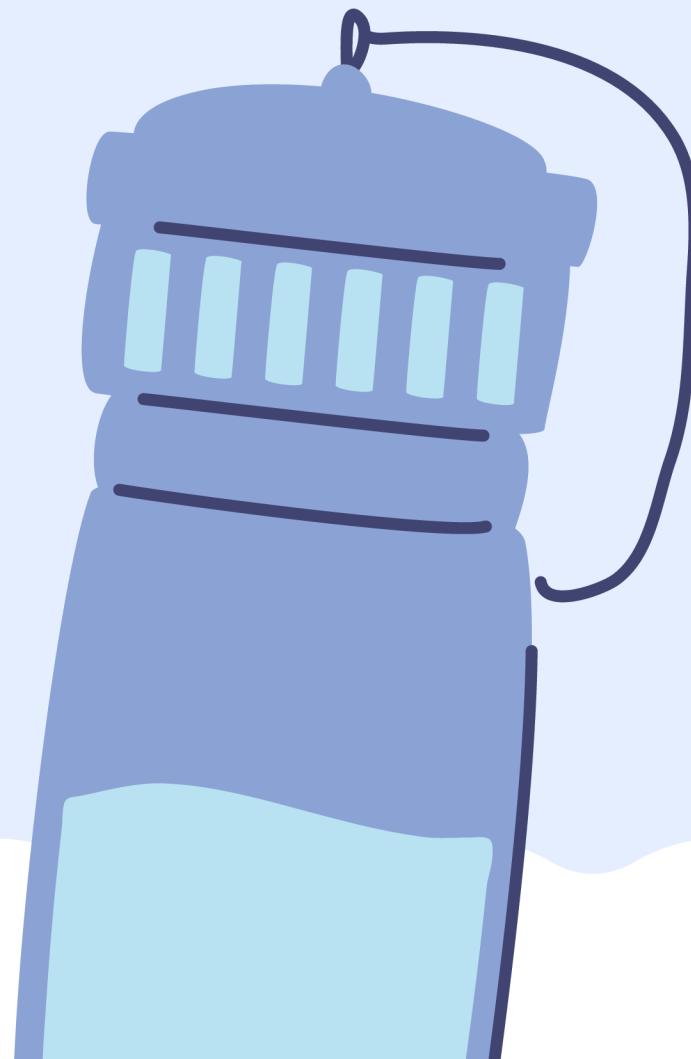
Ja tenim una altra columna del Telòmer i la doctora Molina ens ha dit que podem eliminar aquesta

El model no sabrà llegir dates i, a més, hem de predir si hem de transplantar i no quan, així que no ens serveix la data.

És un identificador de família, però ja tindrem aquesta informació amb les variables de 1st i 2nd degree, i el one-hot encoder funcionarà millor amb variables binàries que amb identificadors

Quan ens vingui un pacient nou, aquestes variables no les sabrem, així que només faran que el model aprengui a utilitzar-les quan realment no les tindrà

Hem considerat que aquesta informació ja se'ns dona a la variable Pathology pattern, la qual conservem



Variables eliminades

Extras AP i Extra

La doctora Molina ens va dir que les podíem eliminar i, a més, afegirien masses variables al fer el one-hot encoding.

COD NUMBER

És un identificador de persona i, per tant, no ens serveix i els models no sabran tractar-lo. A més, hem mirat si una persona apareix més d'un cop (perquè potser aleshores sí que serviria), però només passava una vegada en totes les dades

FVC (L) at diagnosis i FVC (L) 1 year after diagnosis

Ja tenim les variables que calculen el mateix però en percentatge, que és més útil perquè L depèn de cada persona i ens ho va recomanar la doctora Molina

ProgressiveDisease

Hi ha una altra variable que es diu igual amb una minúscula, que és la nostra target, llavors aquesta no cal



Transformacions de variables

Creació de
FVC_difference (%)
i DLCO_difference (%)

$$\begin{array}{r} \text{FVC (\%)} \\ \text{1 year} \\ \text{after diagnosis} \end{array} - \begin{array}{r} \text{FVC (\%)} \\ \text{at diagnosis} \end{array}$$
$$\begin{array}{r} \text{DLCO (\%)} \\ \text{1 year} \\ \text{after diagnosis} \end{array} - \begin{array}{r} \text{DLCO (\%)} \\ \text{at diagnosis} \end{array}$$

Fem la diferència per saber el canvi

Creació de
Mutation_Summary

Ajuntem variables “Genetic mutation studied in patient” i “Mutation type”
Si hi havia mutació però no hi havia nom, posem “No patologica”, sinó el seu nom i sino “No mutació”



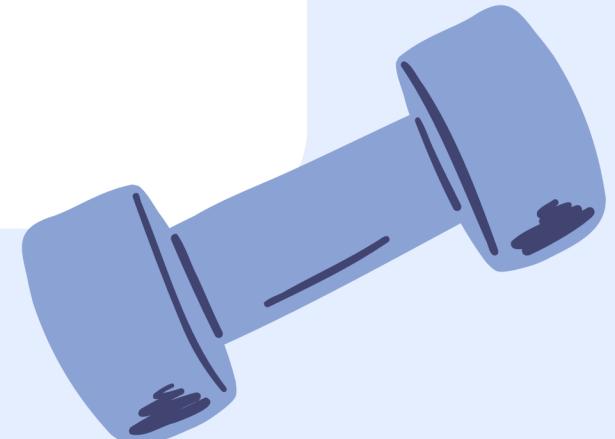
Imputació de missings i noms de variable de missing

Eliminem els 9 pacients que tenen missings a la variable a predir, ja que sinó estaríem influenciant massa al model amb informació que potser no és del tot correcte

Missings a la variable Detail:
'No detalle'

Missings a la variable
Radiological Pattern:
'Indeterminate'

Missings a la variable Detail
on NON UIP:
'No detail'



Imputació de missings i noms de variable de missing

Missings a la variable
Extrapulmonary affection:
Imputem amb k-moda

Missings a la variable Biopsy:
Posem un 0

Missings a la variable
Pathology pattern:
‘Unknown’

Missings a la variable
Treatment:
‘No data’

Imputació de missings i noms de variable de missing

Missings a la variable Type of telomeric extrapulmonary affection:
'Unknown'

Missings a la variable Associated lung cancer:

Imputem amb k-moda

Missings a la variable Other cancer:

Imputem amb k-moda

Missings a la variable Type of neoplasia:

'Unknown'

Imputació de missings i noms de variable de missing

Missings a la variable Hematological abnormality before diagnosis:

‘Unknown’

Missings a la variable Blood count abnormality at diagnosis:

Imputem amb k-moda

Missings a les variables relacionades amb la sang:

Imputem amb k-moda calculant veïns de les variables de la sang

Missings a la variable Liver abnormality before diagnosis:

‘Unknown’

Imputació de missings i noms de variable de missing

Missings a la variable Liver abnormality i Liver disease:

'Unknown'

Missings a la variable Type of liver abnormality i Identified infection:

'None'

Missings a la variable Diagnosis after Biopsy:

Canviem els -9 per Unknown

Missings a la variable RadioWorsening2y:

Canviem els 3 per 1

Imputació de missings i tractament d'outliers

Missings a la variable Type of family history:

'No history'

Missings a la variable More than 1 relative:

Canviem els -9 per 0

Passem els valors de variables categòriques que no surten menys de 5 cops a la variable que significa missing (com Unknown o Indeterminate)
Sinó, el one-hot ens crearà més de 200 variables

'Type of telomeric extrapulmonary affection',
'Type of neoplasia',
'Type of liver abnormality',
'Detail on NON UIP',
'Treatment',
'Type of family history'

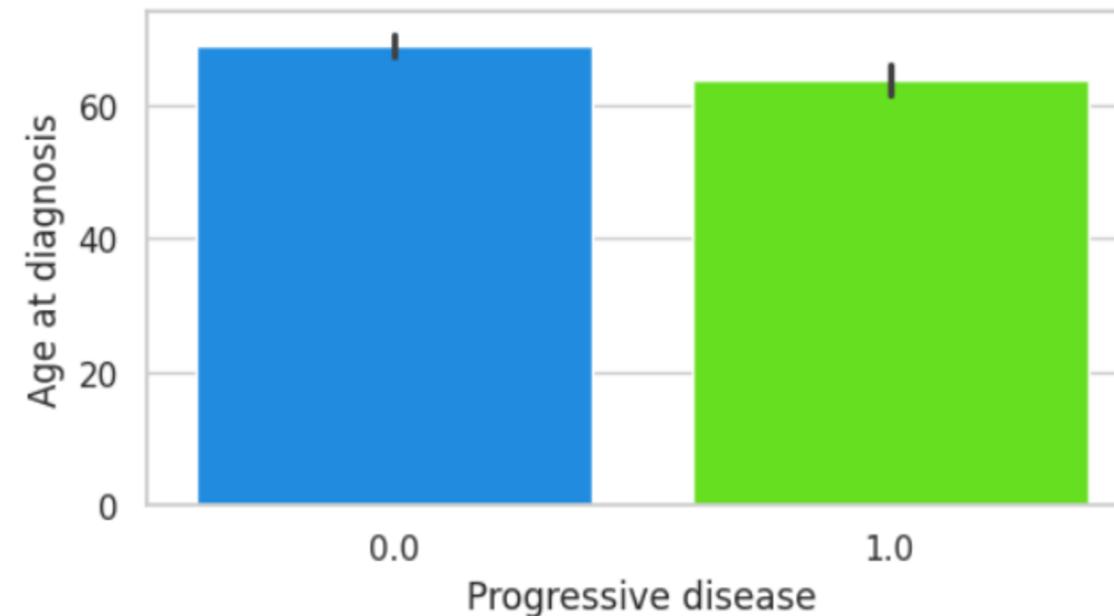
Exemple Detail on NON UIP:

Valores únicos después de la transformación: ['No detail' 'Emphysema' 'Trapping']

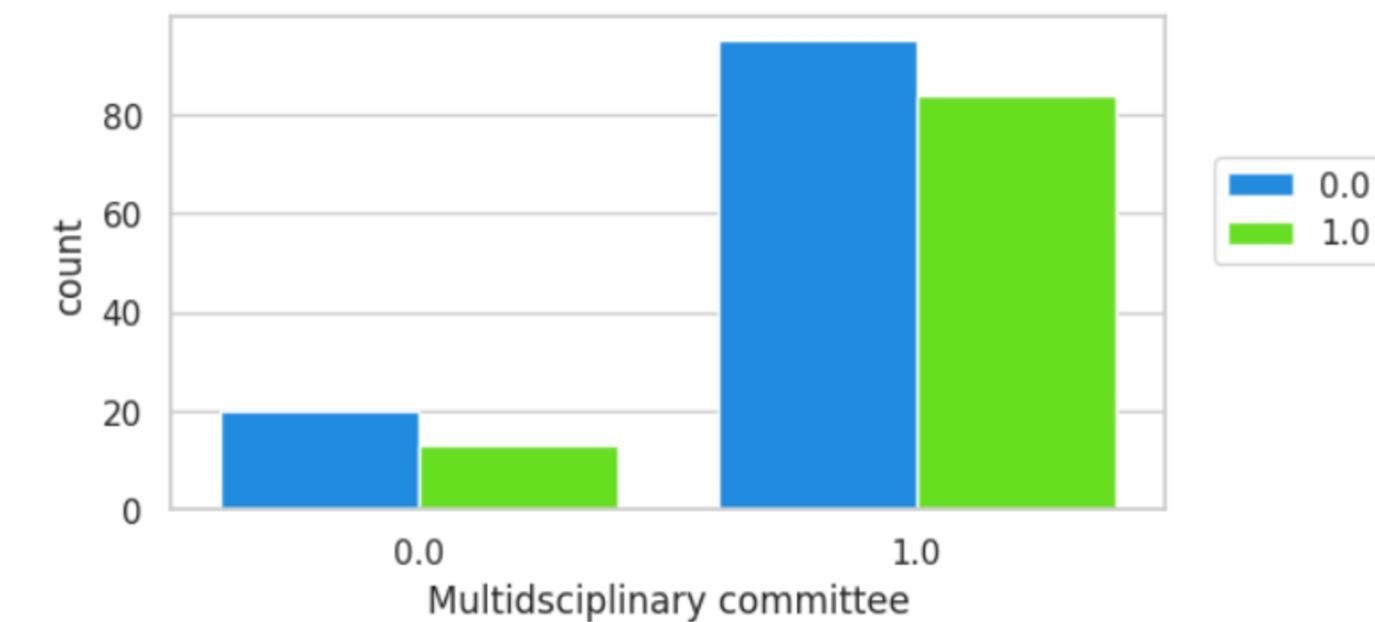
Eliminació de variables que no tenen diferències respecte la target

Hem mirat totes les variables numèriques i categòriques respecte la target per veure si alguna no aportava diferències significatives

Age at diagnosis



Multidisciplinary committee

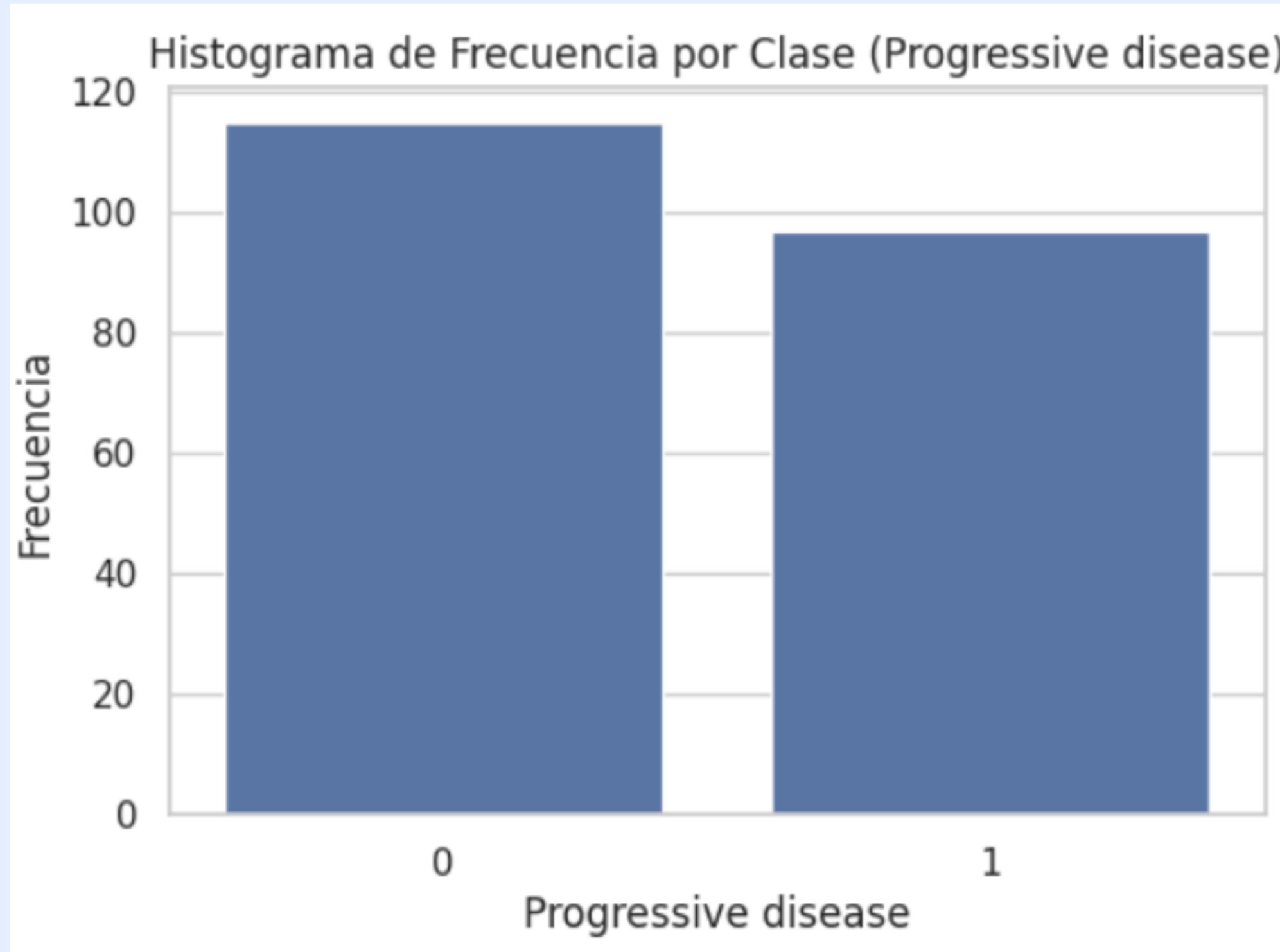


Realment la majoria dels pacients de la database tenen el mateix rang d'edat i, per tant, l'edat no té diferències significatives respecte la target

La majoria de les dades té un 1 a la variable, però no hi ha diferències si ens fixem respecte a la target

Estudi de balanceig de la variable objetciu

Mirem a veure si està balancejat la gent que progressa i la que no,
perquè el model pugui distingir bé entre les dues



Està prou balancejada,
no cal crear dades
sintètiques

Normalització variables numèriques

Utilitzem un MinMaxScaler per normalitzar les numèriques per poder passar-li les dades al model

DataFrame después de la normalización:

```
Age at diagnosis    FVC_difference(%)    DLCO_difference(%)    Sex    \
0                  0.607143              0.287129              0.260255    Male
1                  0.535714              0.193069              0.550486    Male
2                  0.410714              0.460396              0.426491    Male
3                  0.535714              0.000000              0.198938  Female
4                  0.642857              0.451485              0.342010    Male
```

```
FamilialvsSporadic Binary diagnosis    Final diagnosis    TOBACCO    \
0          Familial           No IPF                   3            2
1          Familial           No IPF                   8            2
2          Familial           No IPF                   3            2
3          Familial             IPF                   1            0
4          Familial             IPF                   1            0
```

```
Detail    Comorbidities    ... Necessity of transplantation    \
0  Tobacco-associated      0    ...                                0
1        No detalle         1    ...                                0
2  Tobacco-associated      0    ...                                1
3        Organizing         1    ...                                0
4        No detalle         1    ...                                0
```

MinMaxScaler:

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$



● One-hot encoding de les variables categòriques i

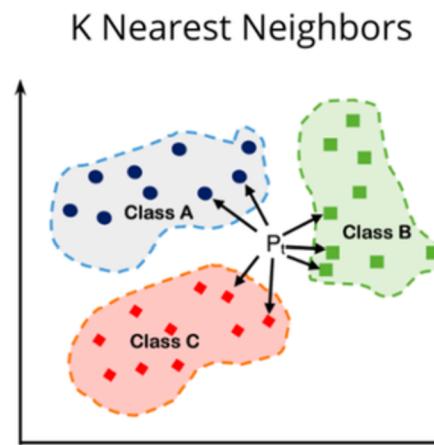
```
Mutation_Summary_No_mutacion  Mutation_Summary_No_patologica \
0                                1
0                                1
0                                1
1                                0
0                                0

Mutation_Summary_PARN \
0
0
0
0
0
```

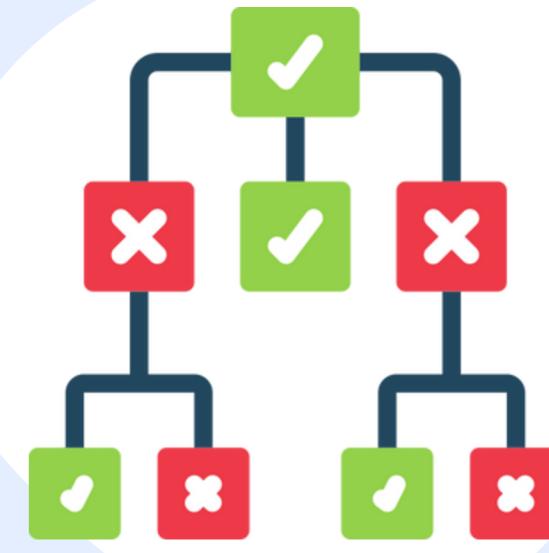
- Split en train i test i
pca per reduir dimensionalitat



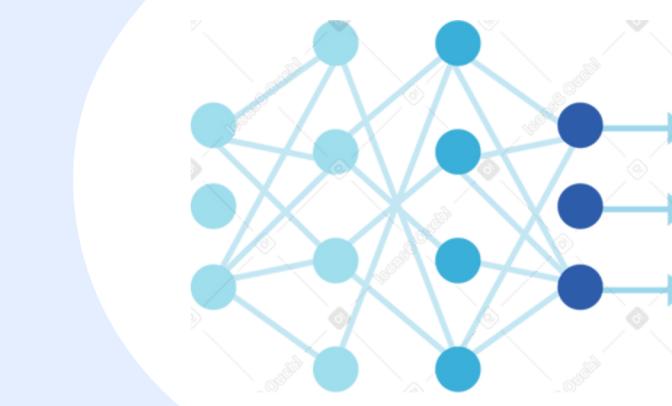
MODELOS ENTRENADOS



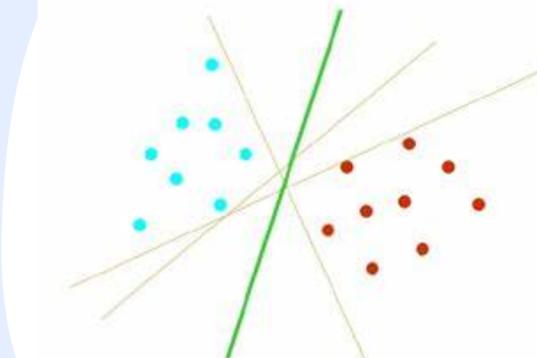
KNN



ARBOL DE
DECISIÓN



RED NEURONAL



SVM

Para cada uno hemos buscado sus mejores hiperparametros con cross-validation

KNN resultados:

1

TRAIN con Cross-Validation:

Accuracy: 0.7045
Precisión: 0.7538
Recall: 0.7045
F1 Score: 0.6893

2

TEST:

Accuracy en prueba: 0.790
Precisión en prueba: 0.784
Recall en prueba: 0.790
F1-score en prueba: 0.783

Arbol de Decisión resultados:

1

TRAIN con Cross-
Validation

fit_time: 0.0191
score_time: 0.0233
test_Accuracy: 0.7807
test_Precision: 0.8008
test_Recall: 0.7807
test_F1 Score: 0.7759

2

TEST:

Accuracy en prueba:
0.6744
Precisión en prueba:
0.7409
Recall en prueba: 0.6744
F1 Score en prueba:
0.6845

Red Neuronal resultados:

1 TRAIN con Cross-Validation:

4/4

12ms/step - accuracy: 0.9963 - loss: 0.3535 -
val_accuracy: 0.8519 - val_loss: 0.6148
Epoch 24/100 4/4

13ms/step - accuracy: 0.9911 - loss: 0.3487 -
val_accuracy: 0.8519 - val_loss: 0.6067
Epoch 25/100 4/4

15ms/step - accuracy: 0.9911 - loss: 0.3471 -
val_accuracy: 0.8519 - val_loss: 0.6020
Epoch 26/100 4/4

14ms/step - accuracy: 1.0000 - loss: 0.3306 -
val_accuracy: 0.8889 - val_loss: 0.6056

2 TEST:

Accuracy en prueba:
0.8837

Precisión en prueba:
0.8462

Recall en prueba: 0.7857

F1 Score en prueba:
0.8148

SVM resultados:



Modelo Final Escojida por sus Grandes Resultados

1

TRAIN con Cross-Validation:

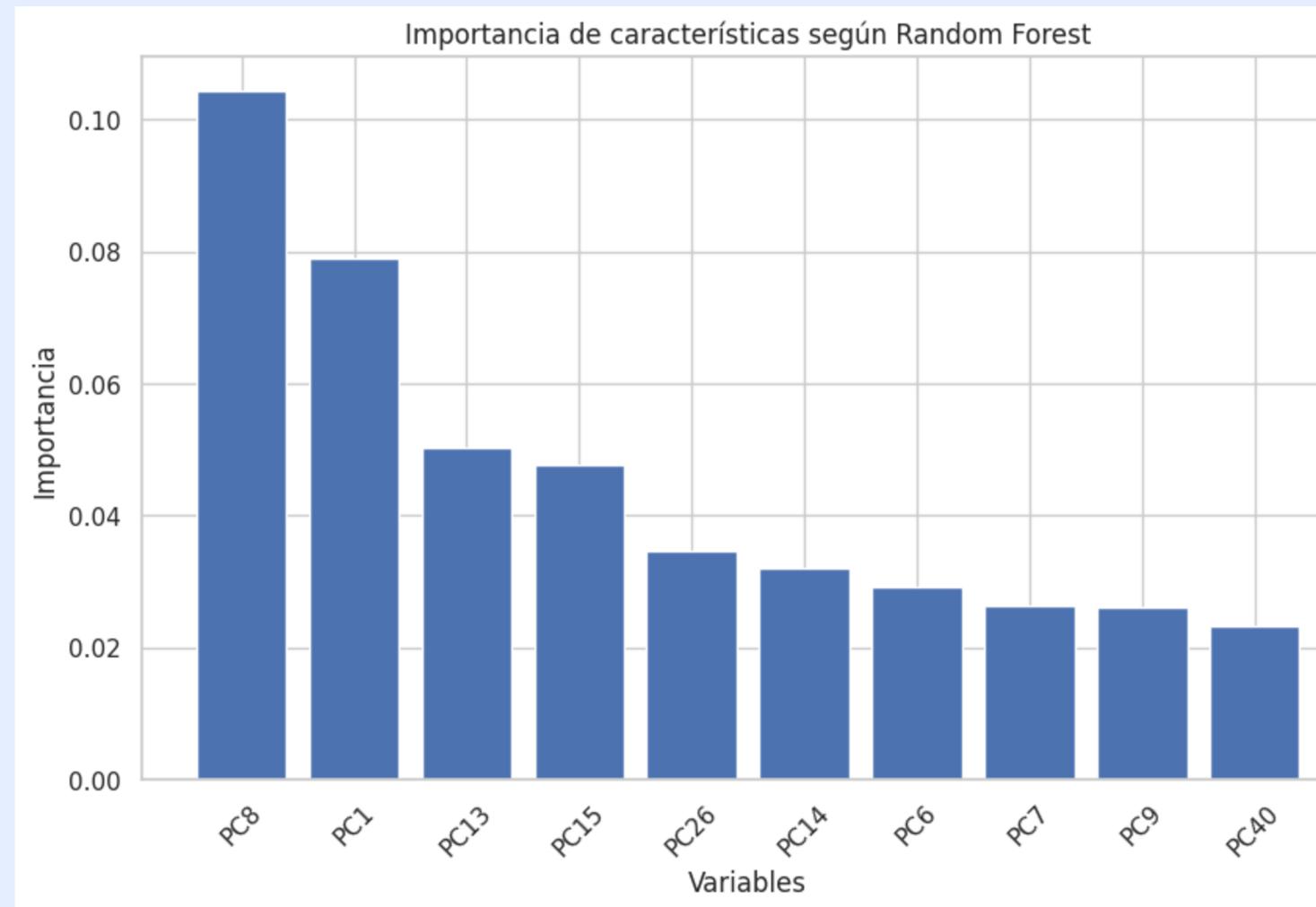
Accuracy: 0.8936
Precisión: 0.9041
Recall: 0.8936
F1 Score: 0.8927

2

TEST:

Accuracy en prueba:
0.9302
Precisión en prueba:
0.9368
Recall en prueba: 0.9302
F1 Score en prueba:
0.9278

Características Más Influyentes en Árbol de Decisión y en SVM



1

PC1 és la combinació d'aquestes variables:

- FamilialvsSporadic_Sporadic
- 1st degree relative_1'
- Severity of telomere shortening_Unknown
- Type of family history_No history'
- Mutation_Summary_No mutacion'

2

PC8 és la combinació d'aquestes variables:

- RadioWorsening2y_1
- Pathology pattern_UIP
- Severity of telomere shortening_1
- Comorbidities_1
- Diagnosis after Biopsy_Unknown']