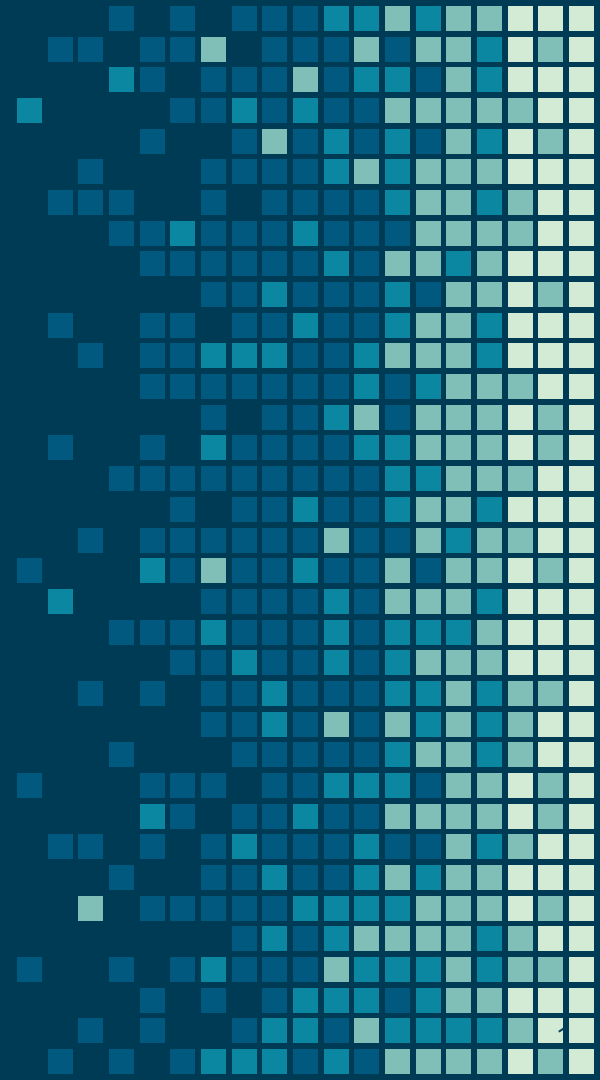


Reporte ejemplo titanic - Option



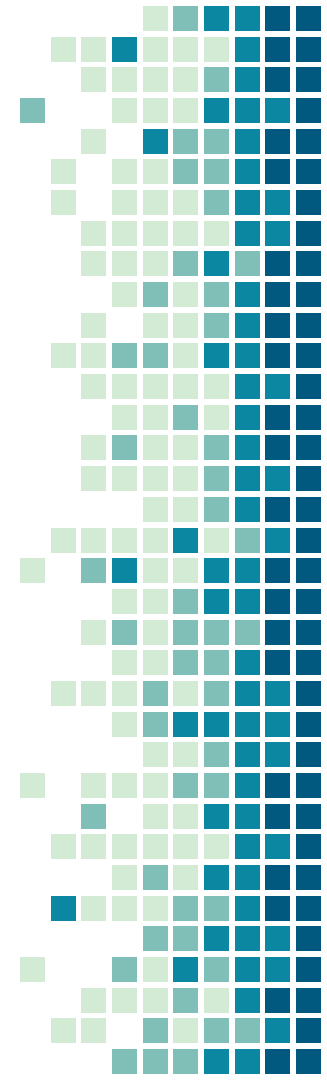
Agenda

Analisis Exploratorio

Procesamiento de los datos

Insights

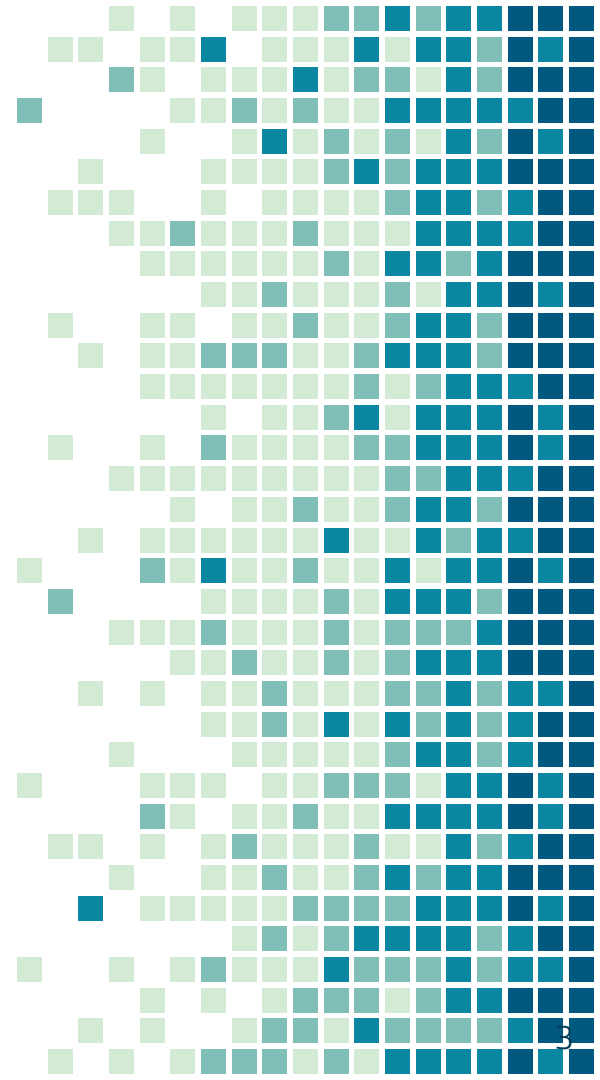
Clasificación



1.

Datos

Analisis exploratorio



Subiendo a Bigquery

Editor de consultas

[OCULTAR EDITOR](#)

```
1 SELECT * FROM `option-bigquery-example.titanic.titanic` LIMIT 1000
```

[Ejecutar consulta](#) [Guardar consulta](#) [Guardar vista](#) [Más](#)

Esta consulta procesará 89,71 KB cuando se ejecute.

Resultados de la consulta

[GUARDAR COMO](#)[EXPLORAR EN DATA STUDIO](#)

Se ha completado la consulta (tiempo transcurrido: 1,275 s; bytes procesados: 89,71 KB)

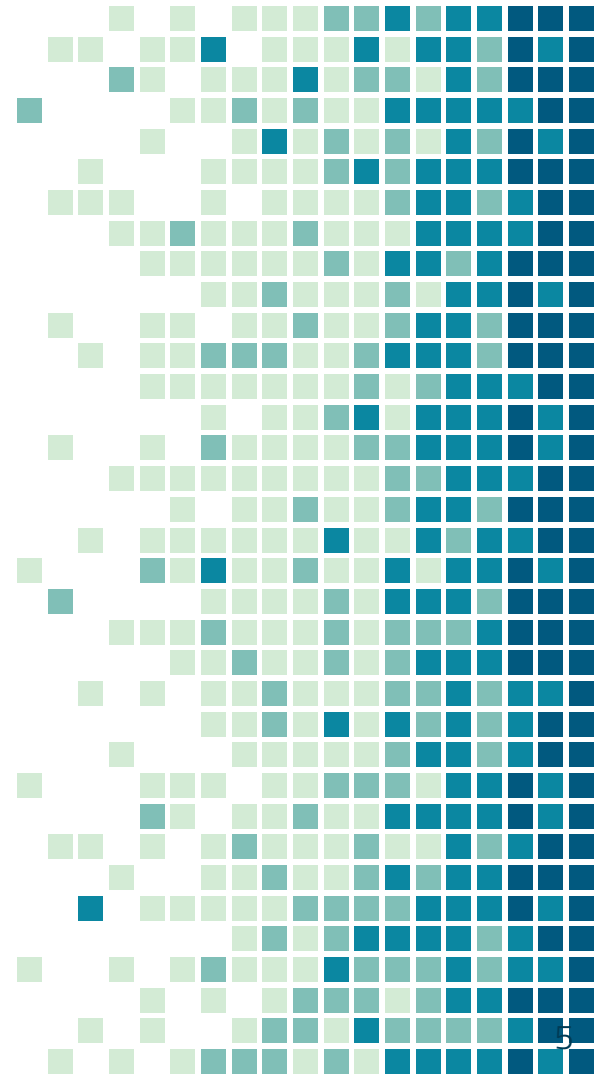
Información de la tarea [Resultados](#) [JSON](#) [Detalles de ejecución](#)

Fila	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	264	0	1	Harrison, Mr. William	male	40.0	0	0	112059	0.0	B94	S
2	634	0	1	Parr, Mr. William Henry Marsh	male	null	0	0	112052	0.0	null	S
3	807	0	1	Andrews, Mr. Thomas Jr	male	39.0	0	0	112050	0.0	A36	S

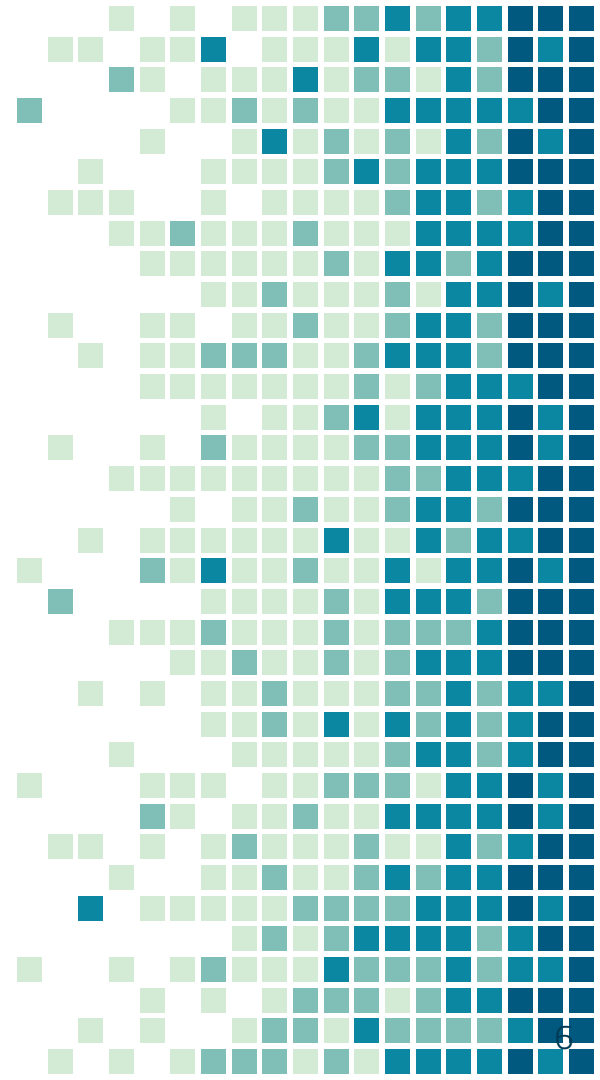
Investigando el dataset

<https://www.kaggle.com/c/titanic/data>

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton



2. Procesamiento



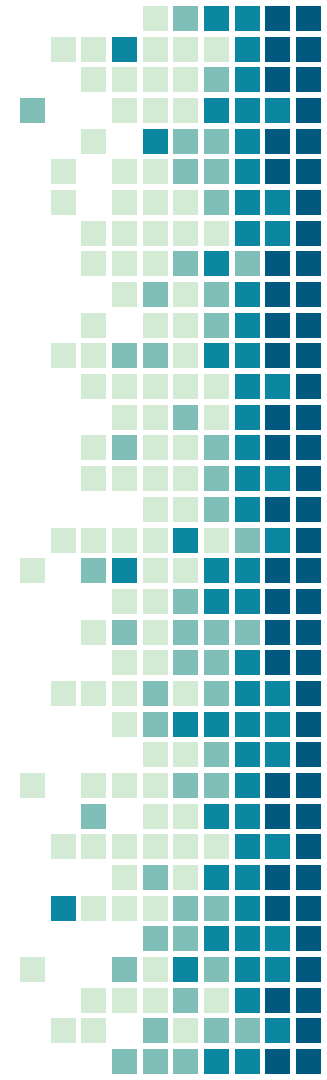
Procesamiento y transformación (2 approaches)

Jupyter Notebook
(Datalab GCP)

Procesamiento,
transformación y
limpieza en python
(mediante Bigquery
API)

Dataprep GCP

Limpieza de los datos a
través de interfaz
interactiva GCP



Etapas

Deleting unnecessary columns

Se eliminan las columnas: Ticket, Passenger Id, Cabin. Ya que no aportan información relevante al problema

Checking missing values

Missing values (%)

-**Embarked: 0.22%**

-**Age: 19.86%**

Solution:

-Eliminar filas NaN en embarked dado que no afecta en el total

-Reemplazar edades con el promedio por sexo y Pclass para ser lo más representativo posible

Coding values

-El sexo se codifica en 0 o 1 para ser interpretado en pasos posteriores.

-Se generan 5 rangos de edad para poder sacar información.

Datalab (Jupyter notebook)

<https://github.com/fbahamonde/optionexample/blob/master/option-titanic.ipynb>



Dataprep

TITANIC FLOW >
titanic ▾
Initial Sample

🔍 ⚙️ ✎️ [Run Job](#)

#	Survived	#	Pclass	RBC	Name	#	Sex	#	Age	#
0-1		1-3		889 Categories		0-1		1-80		0-8
0		3		Leonard, Mr. Lionel	1	36	0			0
0		1		Harrison, Mr. William	1	40	0			0
0		2		Parkes, Mr. Francis "Frank"	1	30.7407070707053	0			0
0		3		Johnson, Mr. William Cahoone Jr	1	19	0			0
0		2		Cunningham, Mr. Alfred Fleming	1	30.7407070707053	0			0
0		2		Campbell, Mr. William	1	30.7407070707053	0			0
0		2		Frost, Mr. Anthony Wood "Archie"	1	30.7407070707053	0			0
0		3		Johnson, Mr. Alfred	1	49	0			0
0		1		Parr, Mr. William Henry Marsh	1	41.28138613861387	0			0
0		2		Watson, Mr. Ennis Hastings	1	30.7407070707053	0			0
0		2		Knight, Mr. Robert J	1	30.7407070707053	0			0
0		1		Andrews, Mr. Thomas Jr	1	39	0			0
0		1		Fry, Mr. Richard	1	41.28138613861387	0			0
0		1		Reuchlin, Jonkheer. John George	1	38	0			0
0		1		Carlsson, Mr. Frans Olof	1	33	0			0
0		3		Burke, Mr. Jeremiah	1	19	0			0
0		3		Hegarty, Miss. Hanora "Nora"	0	18	0			0
0		3		Braund, Mr. Owen Harris	1	22	1			1
0		3		Coxon, Mr. Daniel	1	59	0			0
0		3		Perkin, Mr. John Henry	1	22	0			0
0		3		Lovell, Mr. John Hall ("Henry")	1	20.5	0			0
0		3		Reed, Mr. James George	1	26.50758893280633	0			0
0		3		Dennis, Mr. Samuel	1	22	0			0

9 Columns889 Rows3 Data Types

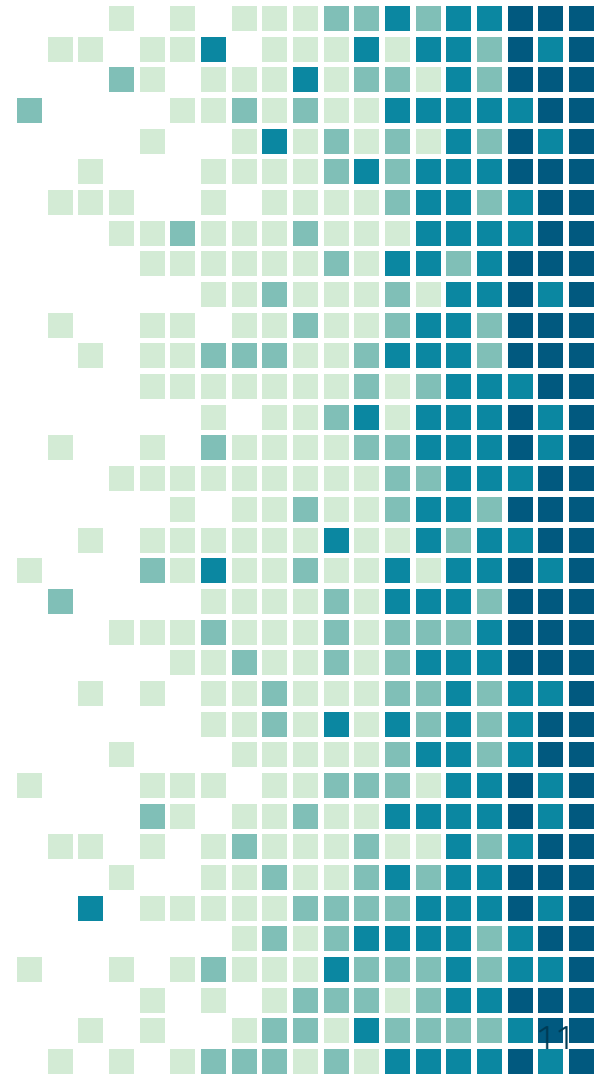
New Step

Recipe

×

```
1 SETTYPE col: Age type: 'Float'
2 DROP col: Ticket action: Drop
3 DROP col: PassengerId action: Drop
4 SETTYPE col: Sex type: 'Integer'
5 SETTYPE col: Sex type: 'String'
6 DROP col: Cabin action: Drop
7 DERIVE type: single value: IF(Sex == 'male',
  1, 0) as: 'Sexcode'
8 DROP col: Sex action: Drop
9 RENAME type: manual mapping: [Sexcode, 'Sex']
10 FILTER type: custom rowType: single row:
  ISMISSING([Embarked]) action: Delete
11 SET col: Age value: IFMISSING($col,
  AVERAGE(Age)) group: Sex, Pclass
12 SETTYPE col: Age type: 'Float'
```

3. Clustering



Etapas

PCA

Se reduce la dimensionalidad para ver la posibilidad de graficar los clústeres a generar

Escoger el número de clusters

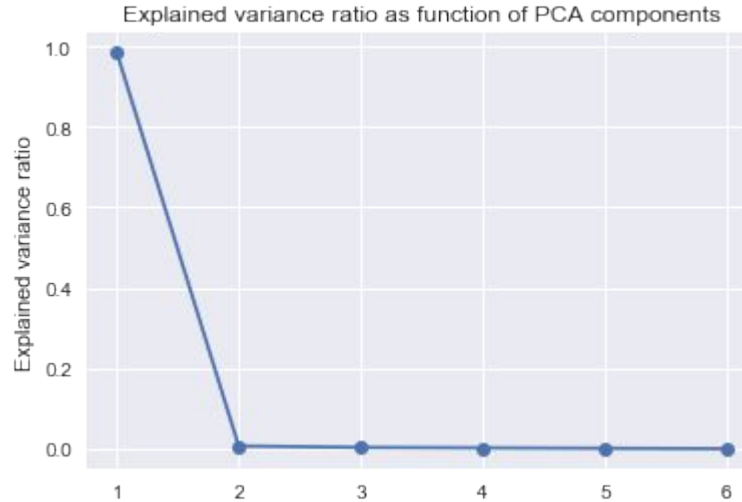
Mediante Elbow Method se escoge la cantidad de clusters

Insights

Se interpretan los resultados para generar información útil

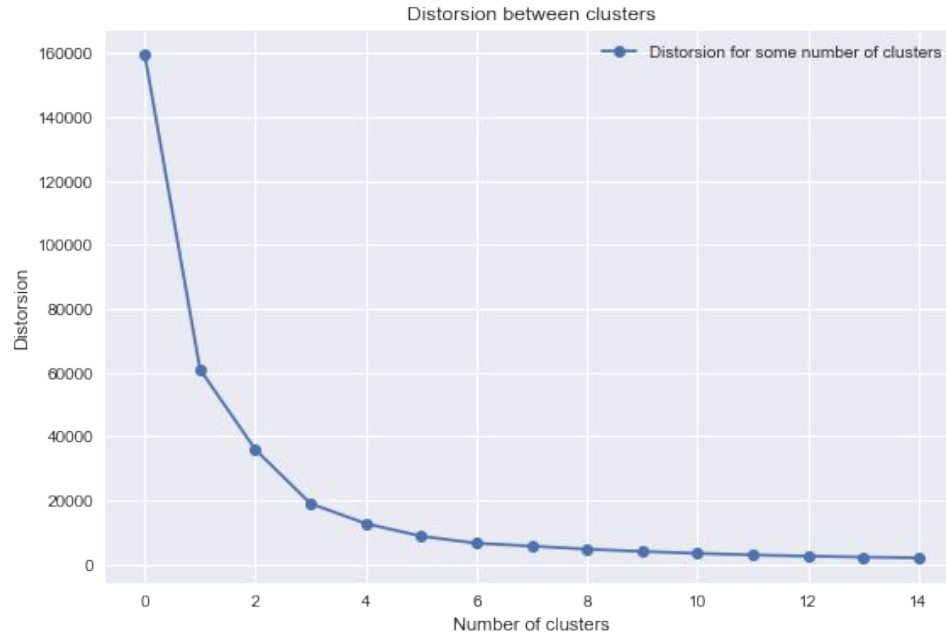
PCA

Al ver que las primeras componentes logran describir la mayor parte del problema, se procede a reducir la dimensión a 2 componentes para su visualización



Elección de cantidad de clusters

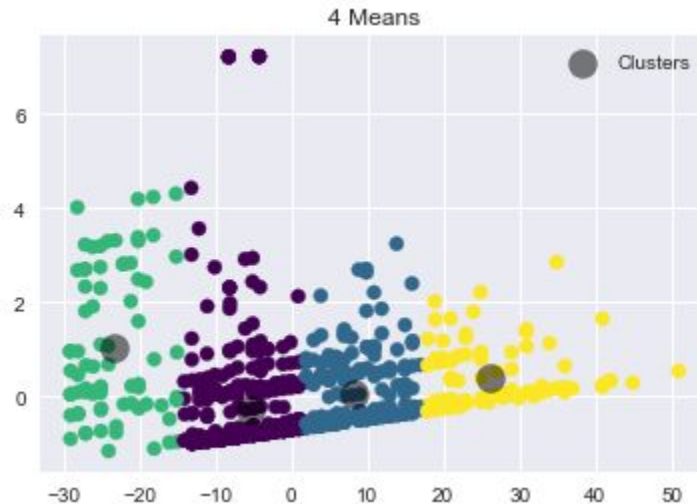
Mediante Elbow Method se procede con la cantidad de clusters. De la Fig. se puede ver que 4 resulta un buen número.



Clusters

Se generan 4 grupos que se encuentran separados principalmente por rangos de edad.

De este análisis se desprende que los grupos etarios con mayor probabilidad de sobrevivir son los menores, independiente de la clase.



4. Insights

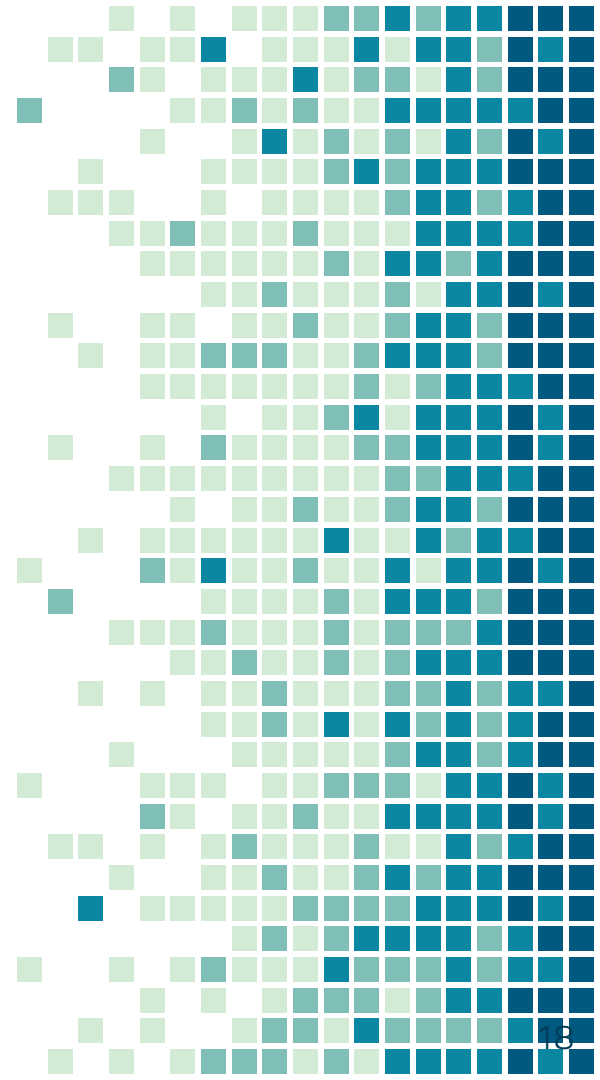


Insights

- * Personas sin familiares tenían mayor probabilidad de sobrevivir
- * Como era de esperarse, en promedio sobrevivió más gente de clase alta (1).
- * En promedio sobrevivieron más mujeres que hombres.
- * Mejores de edad fueron los que más posibilidades tenían de sobrevivir, así mismo, las mejores entre 30 y 60 tienen una altísima probabilidad de sobrevivir, independiente de la clase.
- * Sobre 60 años la probabilidad de supervivencia disminuye a casi 0



5. Clasificación



Resultados modelos

Logistic Regression

Accuracy: 0.74 (+/- 0.16)

Decision Tree

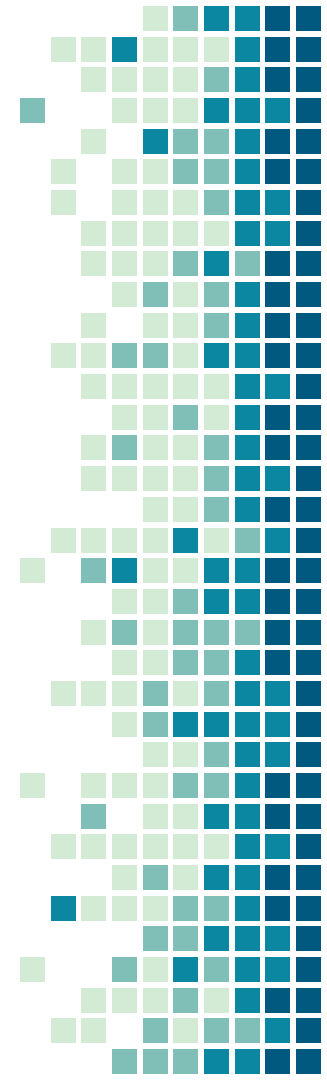
Accuracy: 0.62 (+/- 0.20)

Random Forest

Accuracy: 0.64 (+/- 0.18)

*Utilizando Cross-Validation (5 Folds) y sin optimizar parámetros en los modelos.

**Para la clasificación se codifica Embarked de 0 a 2. Y se reemplazan las columnas asociadas a familiares por la "Alone", que indica si viajó con familiares o solo.



Almacenamiento Bigquery

Para finalizar, se almacenan los datos procesados en Datalab en bigquery. Para su posterior uso o despliegue de insights en Data Studio o similar.



Gracias!

Felipe Bahamonde

felipe.bahamonde.m@ug.uchile.cl