****

**75.06/95.58**

**Organización de Datos**

**Primer cuatrimestre de 2018**

**Trabajo Práctico Nº 2:**

**Competencia de Machine Learning**

|  |  |
| --- | --- |
| **Integrantes** | **Padrón** |
| Cozza, Fabrizio Luis | 97.402 |
| López Lecube, Lucio |  |
| Giannotti, Luciano |  |

**Manipulación de los datos**

***CSVs:***

Al momento de juntar los datos se reveló una enorme cantidad de registros para analizar (como también un enorme tamaño de los archivos resultantes), por lo que en un principio se realizó una transformación de los tipos de datos numéricos para reducir el uso de memoria. Sobre este problema también se trabajó con aquellas columnas que no poseían valores numéricos, transformándolas en variables categóricas según correspondiese.

Luego, se transformaron aquellas columnas que tenían un formato determinado (como por ejemplo las fechas) pero no estaban determinados con su formato correspondiente en el dataframe.

A continuación, se verificó que no se produzca la aparición de valores nulos o con NaNs para no molestar a ciertos algoritmos de machine learning y se procedió a eliminar duplicados con cierto criterio (por ejemplo idpostulante e idaviso al mismo tiempo) para no perder información.

A partir del análisis exploratorio de datos se descubrió que en los archivos de avisos y postulaciones del set de datos podría llegar a haber información que se podría recuperar y llegar a ser útil para la parte del aprendizaje, por lo cual se procedió a intentar recuperar la mayor parte posible.

***Training set:***

***Testing set:***

**Algoritmos probados**

*A continuación se muestra una lista de los algoritmos testeados a lo largo del trabajo práctico:*

*Principales:*

* Random Forest (Regressor y Classifier)
* KNN
* Extra Trees (Regressor y Classifier)
* Gradient Boosting (Regressor y Classifier)
* Light GBM (En versiones gradient boosting, random forest y dart)
* XGBoost (Classifier)
* MLP Classifier
* Adaboost

*Secundarios:*

* Linear Regression
* Ridge
* Ridge CV
* Lasso
* Elastic Net
* Lasso Lars
* Orthogonal Matching Pursuit
* Bayesian Ridge
* Huber Regressor

**Análisis de algoritmo*:*** *Algoritmo X*

**Algoritmo final utilizado:** Algoritmo X

**Conclusiones**

**LINK a Github:**