1 Descripción del problema.

El objetivo de esta competición es predecir de la forma más exacta posible el precio de las casas residenciales en Ames, Iowa, a partir de 79 variables explicativas que describen casi todos los aspectos.

1.1 Variables

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- LotShape: General shape of property
- LandContour: Flatness of the property
- Utilities: Type of utilities available
- LotConfig: Lot configuration
- LandSlope: Slope of property
- Neighborhood: Physical locations within Ames city limits
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- RoofStyle: Type of roof
- RoofMatl: Roof material
- Exterior1st: Exterior covering on house
- Exterior2nd: Exterior covering on house (if more than one material)

- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- ExterQual: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- HeatingQC: Heating quality and condition
- CentralAir: Central air conditioning
- Electrical: Electrical system
- 1stFlrSF: First Floor square feet
- 2ndFlrSF: Second floor square feet
- LowQualFinSF: Low quality finished square feet (all floors)
- GrLivArea: Above grade (ground) living area square feet
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- FullBath: Full bathrooms above grade
- HalfBath: Half baths above grade
- Bedroom: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)

- Functional: Home functionality rating
- Fireplaces: Number of fireplaces
- FireplaceQu: Fireplace quality
- GarageType: Garage location
- GarageYrBlt: Year garage was built
- GarageFinish: Interior finish of the garage
- GarageCars: Size of garage in car capacity
- GarageArea: Size of garage in square feet
- GarageQual: Garage quality
- GarageCond: Garage condition
- PavedDrive: Paved driveway
- WoodDeckSF: Wood deck area in square feet
- OpenPorchSF: Open porch area in square feet
- EnclosedPorch: Enclosed porch area in square feet
- 3SsnPorch: Three season porch area in square feet
- ScreenPorch: Screen porch area in square feet
- PoolArea: Pool area in square feet
- PoolQC: Pool quality
- Fence: Fence quality
- MiscFeature: Miscellaneous feature not covered in other categories
- MiscVal: $Value of miscellaneous feature
- MoSold: Month Sold
- YrSold: Year Sold
- SaleType: Type of sale
- SaleCondition: Condition of sale

1.2 Evaluación

Teoria error logaritmico.

2 Marco teórico regresión.

3 Algoritmos.

Nota: Para la ejecución de los algoritmos los datos nominales se han pasado a numérico

A priori desconocemos que algoritmo se adapta mejor a nuestro problema, es por ello que realizaremos un estudio comparando varios algoritmos. Los algoritmos elegidos son los siguientes:

- Un algoritmo clásico como la regresión lineal.
- Algoritmos "básicos" como KNN y Árboles de clasificación y regresión(CART).
- Otros algoritmos como NeuralNetwork,Gaussian y SVR(para hallar relaciones no lineales)
- Multiclasificadores ya que debido a su fácil paralelización ofrecen una buena escalabilidad. En este caso he seleccionado 3 algoritmos:
  - RandomForest. Basado en crear muchos árboles ligeramente diferentes y obtiene el resultado mediante el voto mayoritario.
  - Boosting. Los árboles se crean de forma lineal teniendo en cuenta los fallos anteriores. Esta técnica permite reducir el sesgo.
  - Bagging.

En la siguiente gráfica se muestran los resultados obtenidos.

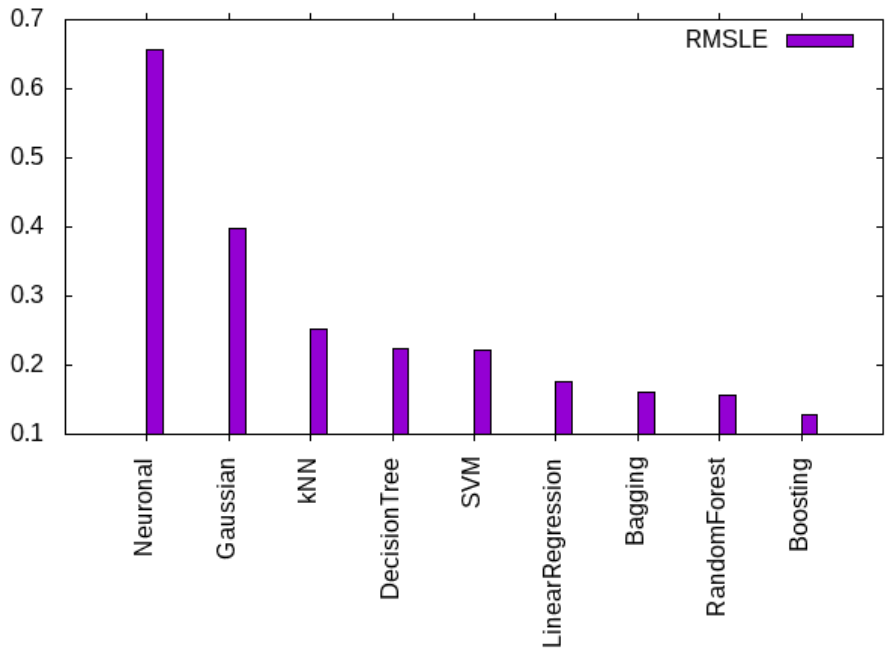

Figure 1: Test de algoritmos.Error.

Se puede observar que los algoritmos que mayor precisión proporcionan en este caso son los multiclasificadores. Por tanto, estudiaremos en profundidad RandomForest y Boosting para mejorar los resultados.

4 Herramientas.

Otra decisión que debemos de tomar es qué herramientas usar para abordar la resolución del problema. Tras un estudio de las bibliotecas disponibles para los distintos lenguajes de programación que domino, seleccioné weka(java), sklearn(python) y R. Para determinar cual de las 3 es la más conveniente en este caso, realizaré pruebas con los diferentes algoritmos y compararemos el tiempo de ejecución y el uso de memoria. Además, se debe tener en cuenta la facilidad para tratar los datos tanto para su lectura, como para la generación de los archivos con las predicciones.

Los resultados obtenidos se muestran en las siguientes gráficas:

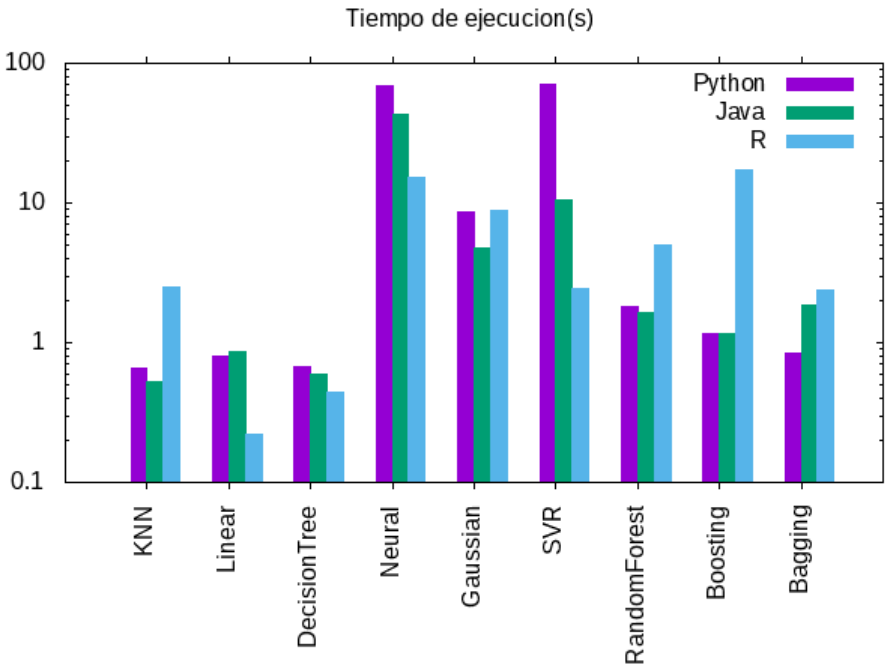

Figure 2: Test de algoritmos.Tiempo de ejecución

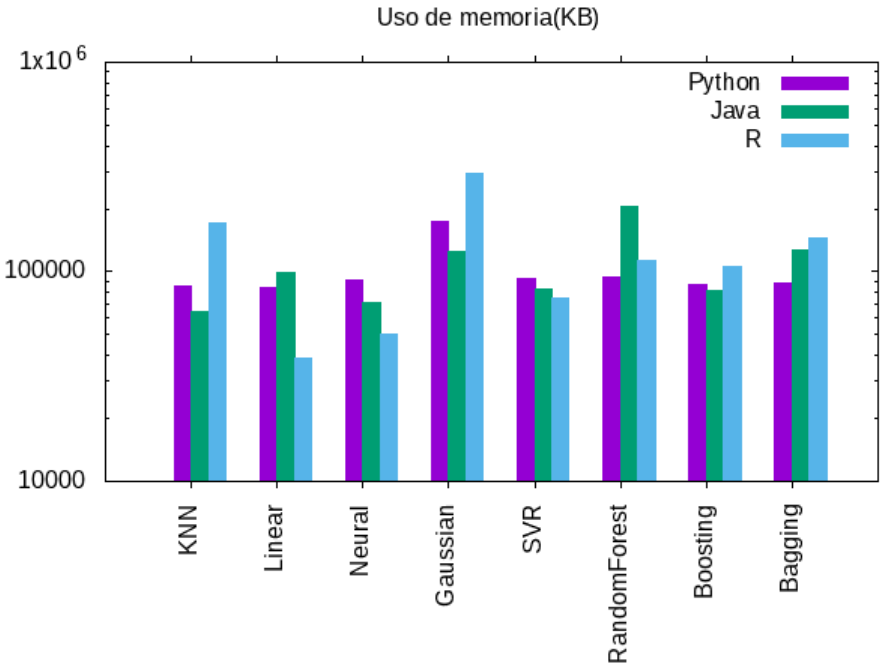

Figure 3: Test de algoritmos.Memoria usada

Observando los resultados podemos concluir que sklearn y weka ofrecen (en media) un rendimiento similar, mientras que R ofrece un mayor rendimiento para los algoritmos básicos pero es inferior en la ejecución de multiclasificadores. Por otro lado, R y sklearn permiten tratar los datos de forma cómoda. Otra cuestión a tener en cuenta, es la buena documentación con la que cuenta sklearn.

En conclusión, el lenguaje a usar será Python ya que permite un manejo de los datos flexible y un código legible propio de este lenguaje. Esta decisión se apoya en que en términos de rendimiento no hay una alternativa "mucho mejor".

5 Preprocesamiento.