**EXAM DOCUMENTATION.**

Dear sigmaritan,

Hope you enjoyed the learning and exploration of the Machine Learning field and you found this field as attractive as we all do. But now it’s time to put all the knowledge accumulated by you into practice. This examination was created with the aim to test your knowledge and to find the level of understanding of the Machine Learning and Data Analysis principles.

The exam will consist of two parts:

1. The exploration of the data set and training of a Machine Learning model on it.
2. The presentation of the insights and results obtained from the work you have done on this data set in a PowerPoint presentation in front of the jury.

The first part should be submitted by loading all the notebooks into a GitHub repository and sending the link to that repository to Vladimir Stojoc by the 27th of March 2022, 00.00.

The second part will be presented individually by each sigmaritans that is examined by the jury consisting of the members of the castor and pollux teams. The date of the presentation will be agreed upon personally with each person and the jury.

**TASK#1.**

In the archive, you have a CSV file. This file is the data that you should use during the exam passing.

In your case, real estate agents have collected data about houses in Iowa, Ames. This dataset includes condition, price and location of the houses in the city of Ames.

The Data Set has the following fields:

* MSSubClass(Type of dwelling involved in the sale identificator)
* MSZoning(General zoning classification of the sale)
* MSZoning(Area of Street connected to the property measured in linear feet)
* Lot Area (Size of the lot measured in square feet)
* Street(Type of road access to the property)
* Alley(Type of alley access to the property)
* Lot Shape
* LandContour(Land flatness level)
* Utilities
* LotConfig(Lot configuration)
* LandSlope(Slope of property)
* Neighborhood
* Condition1(Proximity to various conditions)
* Condition2(Proximity to various conditions (if more than one is present)
* BldgType(Type of dwelling)
* HouseStyle
* OverallQual(Rates the overall material and finish of the house)
* OverallCond(Rates the overall condition of the house)
* YearBuilt
* YearRemodAdd(Remodel date ; same as construction date if no remodeling or additions)
* RoofStyle
* 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)
* ExterCond(Evaluates the present condition of the material on the exterior)
* Foundation
* BsmtQual(Evaluates the height of the basement)
* BsmtCond(Evaluates the general condition of the basement)
* BsmtExposure(Refers to walkout or garden level walls)
* BsmtFinType1(Rating of basement finished area)
* BsmtFinSF1(Type 1 finished square feet)
* BsmtFinType(Rating of basement finished area ; if multiple types)
* BsmtFinSF2: Type 2 finished square feet
* BsmtUnfSF(Unfinished square feet of basement area)
* TotalBsmtSF(Total square feet of basement area)
* 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 ground living area square feet)
* BsmtFullBath(Basement full bathrooms)
* BsmtHalfBath(Basement half bathrooms)
* FullBath(Full bathrooms above ground)
* HalfBath(Half baths above ground)
* Bedroom: Bedrooms above ground (does NOT include basement bedrooms)
* Kitchen(Kitchens above ground)
* KitchenQual(Kitchen quality)
* TotRmsAbvGrd(Total rooms above grade ; does not include bathrooms)
* Functional(Home functionality ;Assume typical unless deductions are warranted)
* 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
* 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
* SaleCondition: Condition of sale
* SalePrice

In this task, you should:

1. Load and analyze the data set, and try to extract insights from the data.
2. Prepare the data to machine learning algorithm ready format,
3. Handle the NaN values in the data set.
4. Apply some post-processing and post-analysis techniques such as Feature Selection, Feature Engineering, or others and see if they can help you.
5. Create a model or a Pipeline that will show the best performance in predicting the target column.

The target column is **SalePrice**. Try to create models that will predict this value.

Some more advice:

* You can split different steps of your work into separated notebooks.
* Commend the code and use the markdown cell type to explain what or why you are doing.
* Explain in a written form the choices and decisions that you made during the processing of the data.
* Use plots, but don’t abuse them.
* Don’t chase a perfect model, better show us why you chose a specific model as the best one.

**TASK#2.**

Your second task is to create and prepare a presentation on the work that you have done. In the presentation that you should explain the work that you have done and why did you make the choices that you made during the analysis and Machine Learning model development. Some more pieces of advice to the presentation are:

* KISS (Keep It Simple Stupid) - keep the presentation as simple as possible.
* Give some context about the field on which your data set is about.
* Try to avoid as much as possible having a lot of text on your slides.
* Try to tell a story.
* Don’t show the whole process, only the most important findings, and decisions.
* Finish with a conclusion.
* Don’t make lots of slides.

Good luck!