Imputation_GUI

Paul Spindler: K12125537
Quang Huy Michael Vu: K01357527

Table of contents:

Goal of our system: 2
Requirements: 2
Use case descriptions: 3
Use case diagram: 3
Indicate which use cases were implemented in the source code: 3
Traceability matrix: 4
Domain model: 4
Architecture diagram: 5
Components description: 5
The design questions: 6
Answers: 6

Goal of our system:

- Gain insights about what kind of predictions the AI made on incomplete tabular data
- Check imputed values

Requirements:

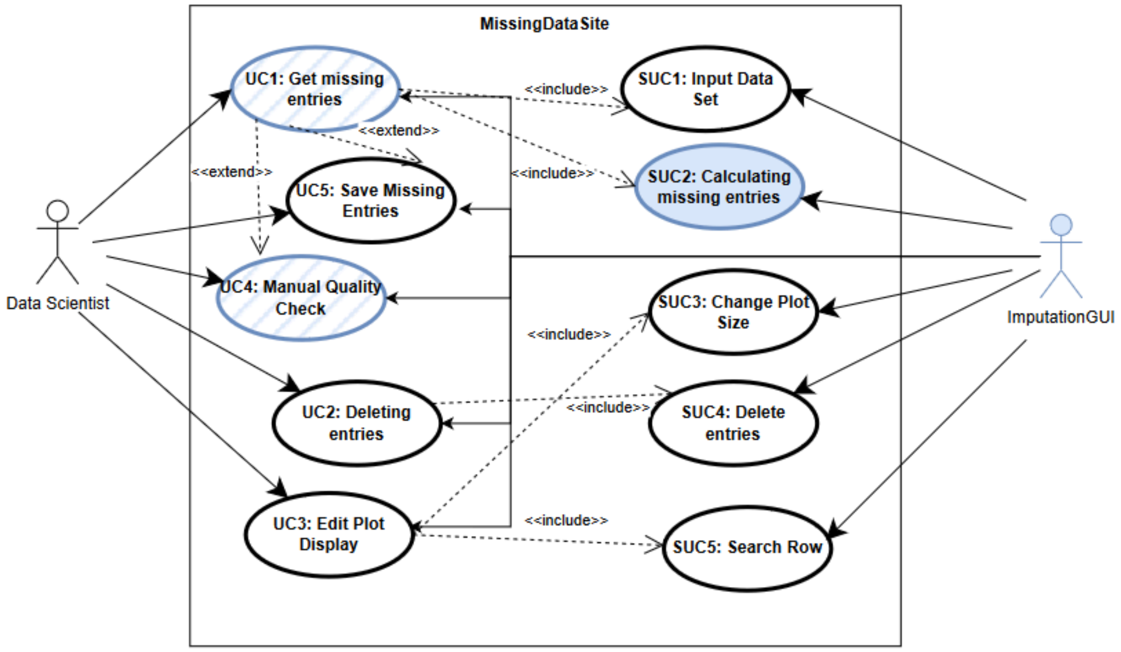
Req0 ImputationGUI shall provide an option to submit a Data Set.
Req1 When a valid Data Set is submitted, ImputationGUI shall calculate missing entries.
Req2 When missing entries are calculated, ImputationGUI shall summarize them in a file.
Req3 When missing entries are summarized, ImputationGUI shall display them to the user.
Req4 When a user changes the plot size that is displayed, ImputationGUI shall display a different plot size.
Req5 When there is a plotted output available, ImputationGUI shall provide the option to search a certain row.
Req6 When the user searches a specific row, ImputationGUI shall provide the row and the next few elements depending on the plot size.
Req7 When there is a plotted output available, ImputationGUI shall provide the option to save the file on the device.
Req8 When there is a plotted output available, ImputationGUI shall provide the option to delete certain entries of the possible missing entries.
Req9 When there is a plotted output available, ImputationGUI shall provide the option to merge the provided Data Set with the missing Data.
Req10 When imputation is finished, ImputationGUI uses Manhattan Distance to compare samples with missing entries to most similar entries.
Req11 When imputation is finished, ImputationGUI evaluates downstream task impact.
Req12 When evaluation of downstream task impact is finished, ImputationGUI displays it for the user's evaluation of correctness.

NfReq0 If the submitted Data Set is bigger than 20.000, ImputationGUI throws an Exception.(Performance)
NfReq1 The ImputationGUI shall limit the comparison of the imputed samples to the 10 most similar samples of data points.
NfReq2 The ImputationGUI shall limit Mean/Mode imputation to less than one second.
NfReq3 The ImputationGUI shall restrain the choice of datasets to tabular data of non-personal information.
NfReq4 The preprocessing (standardizing numerical features, one-hot-encoding categorical features) should not take longer than three seconds.

Use case descriptions:

UC1: Calculating missing entries for a new data set
UC2: User can delete entries by choice
UC3: User can edit the way the plot is displayed on screen
UC4: The User has the Option to check the quality of the provided missing entries set
UC5: User can save the document of the missing entries

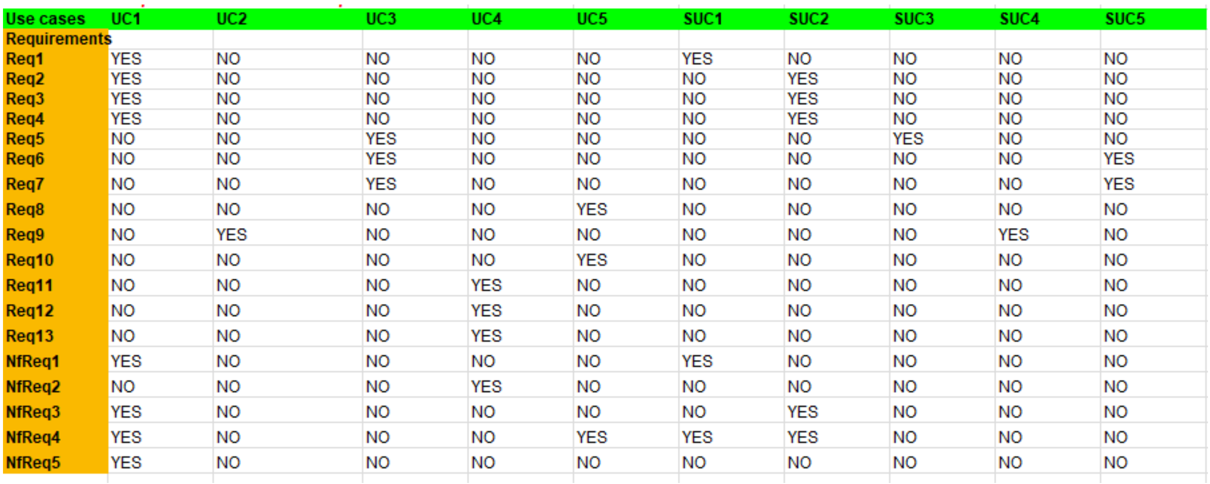
Use case diagram:



Indicate which use cases were implemented in the source code:

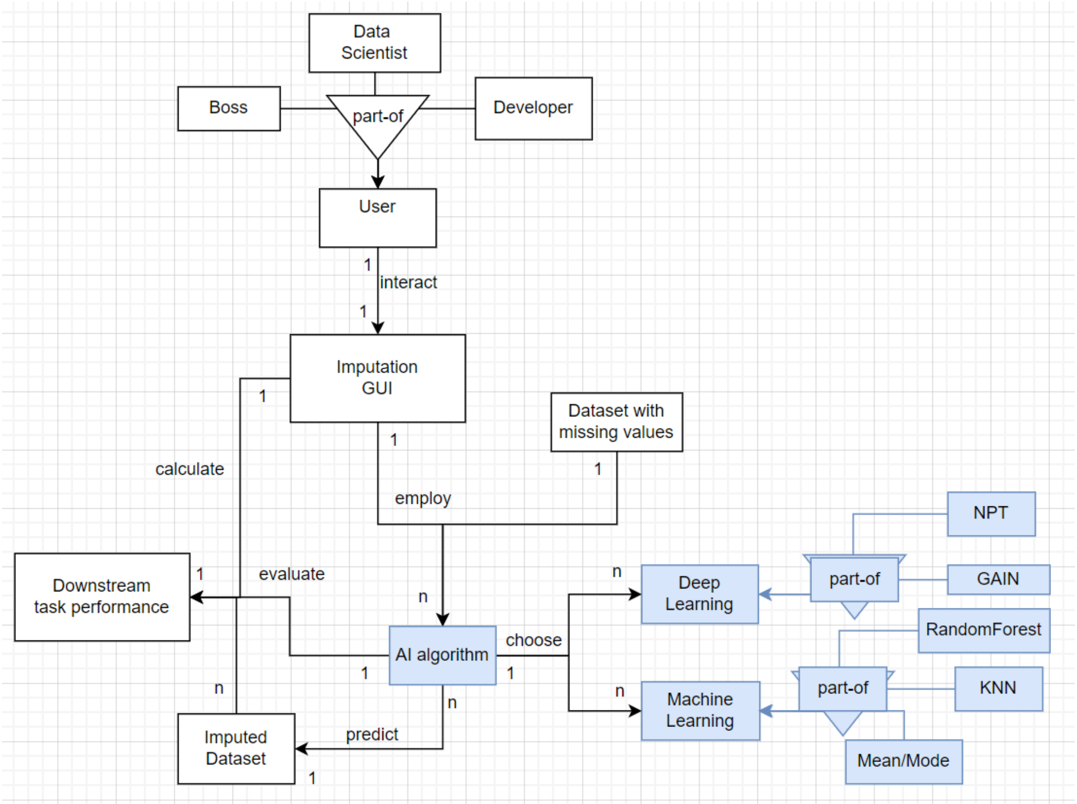
We implemented:
Calculating missing entries for a new set.
User can edit the way the plot is displayed on screen
User can save the document of the missing entries
Users can delete entries by choice -> You can delete entries if you want but its not safed, but you could copy the table manually.

Traceability matrix:

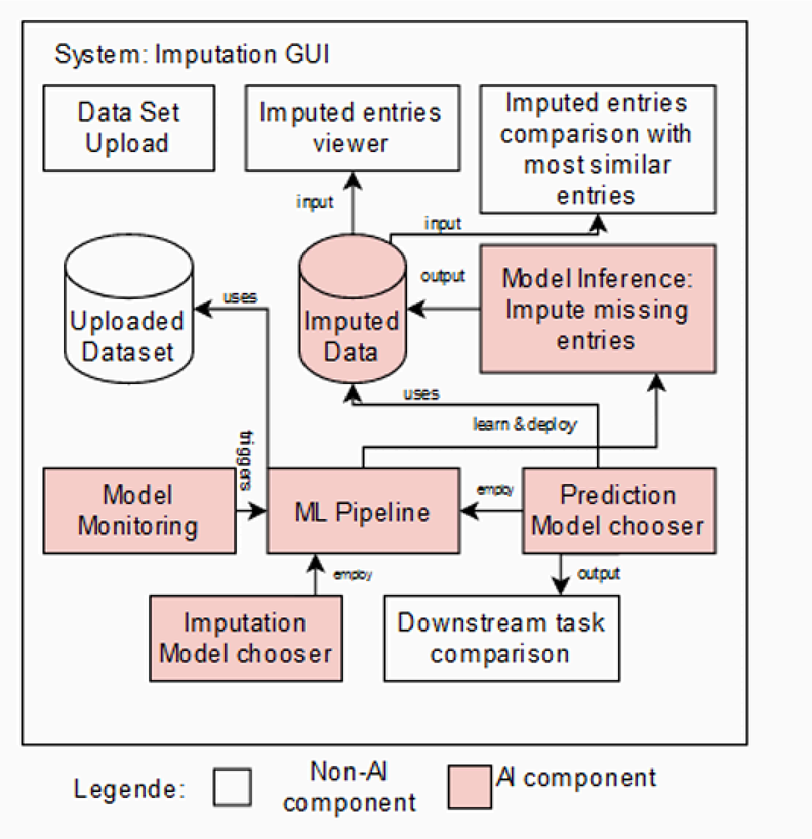


Use cases	UC1	UC2	UC3	UC4	UC5	SUC1	SUC2	SUC3	SUC4	SUC5
Requirements										
Req1	YES	NO	NO	NO	NO	YES	NO	NO	NO	NO
Req2	YES	NO	NO	NO	NO	NO	YES	NO	NO	NO
Req3	YES	NO	NO	NO	NO	NO	YES	NO	NO	NO
Req4	YES	NO	NO	NO	NO	NO	YES	NO	NO	NO
Req5	NO	NO	YES	NO	NO	NO	NO	YES	NO	NO
Req6	NO	NO	YES	NO	NO	NO	NO	NO	NO	YES
Req7	NO	NO	YES	NO	NO	NO	NO	NO	NO	YES
Req8	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO
Req9	NO	YES	NO	NO	NO	NO	NO	NO	YES	NO
Req10	NO	NO	NO	NO	YES	NO	NO	NO	NO	NO
Req11	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO
Req12	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO
Req13	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO
NfReq1	YES	NO	NO	NO	NO	YES	NO	NO	NO	NO
NfReq2	NO	NO	NO	YES	NO	NO	NO	NO	NO	NO
NfReq3	YES	NO	NO	NO	NO	NO	YES	NO	NO	NO
NfReq4	YES	NO	NO	NO	YES	YES	YES	NO	NO	NO
NfReq5	YES	NO	NO	NO	NO	NO	NO	NO	NO	NO

Domain model:



Architecture diagram:



Components description:

- Imputed entries viewer: Displayed the user the imputed values in a GUI.
- Imputed entries comparison with most similar entries: Compare imputed values with the k-most similar samples.
- Model Inference: Impute missing entries, let the selected imputation algorithm calculate its prediction.
- Model Monitoring: if necessary enable the tool “wandb” to monitor the training of deep learning models while machine learning models are not necessary to be monitored. Monitor expected running time of a model.
- ML Pipeline: involves the procedure of preprocessing and training the uploaded dataset Prediction Model chooser: select among the available machine learning and deep learning models to predict targets
- Imputation Model chooser: select among the available machine learning and deep learning models to impute on missing entries
- Downstream task comparison: compare accuracy or regression metric among used imputation models to see if imputed values achieved a better score than mean/mode imputation.
- User Data Set Upload: Lets the user add a data set.

- ImputationGUI: This component handles all the processing and user commands and directs them to the components in charge.

The design questions:

1. Which models should be used for prediction?
2. Which models should be used for imputation?
3. Which metrics for prediction should be used?
4. How many comparison samples should be involved?
5. What kind of dataset limitations are involved?
6. How do we track the inference time?
7. Which tools are used for model monitoring?
8. What kind of datasets are supported?
9. What could be the biggest challenge in our ML pipeline?
10. How can we determine the most critical missing entries to prioritize for prediction?

Answers:

1. Random Forest, XGBoost, CatBoost and NPT
2. SOTA: MissForest, KNN, VAE and NPT
3. Accuracy, R^2, RMSE and MSE
4. maximum 20 samples, by default 10 samples
5. We focus on medium sized datasets with about 20 - 50k samples at first
6. Using the library tqdm with a progress bar to see how long the training and prediction takes.
7. We use wandb instead of tensorboard.
8. We focus mainly on tabular datasets and do not support time series data imputation.
9. Training deep learning methods could take the longest, especially when we also hypertune them.
10. We could use feature importance algorithms to reduce the amount of missing entries