# Flood Risk in the UK （Team Dart）

## 1. Data Pre-processing

We determined the transformed value for soil type based on its water absorption capacity.
We divided the zip code, picked the letters showing the district, and used one-hot encode on them.

## 2. Modelling Approaches

Task1
Since the Risk Label column (target) was highly imbalanced, we oversampled the data, which made a very minor improvement.
Because the score based on the given score matrix of a model that treats risk labels as a classification problem is too low (0.72), we treat the problem as a regression problem. We used the score matrix as a criterion for screening models and finally selected the Decision Tree Regressor. After hyperparameter adjustment and dealing with overfitting, we got the best parameter and identified the important features: latitude, longitude, elevation and soil type, and this model does not use historic flooding data.

Task2
We added the households feature in sector_data for training. RandomForestRegression model was selected. To reduce the variance, we try to combine different encodings and whether to use filters. Finally, it is found that using a one-hot encode encoding combination filter (threshold: 10%-90%) has the best effect. This model does not use historic flooding data.

Task3
We decided not to use data scaling because it could have lowered the f1 rating.
We chose the RandomForestClassifier based on its f1 score. While attempting feature selection, we aimed to link the station with easting and northing to have more options. However, due to accuracy concerns and technical challenges, we decided to drop this effort. Ultimately, we found that focusing on features in postcodes_unlabelled.csv and postcodes_labelled.csv datasets produced satisfactory results.

Task4&5
KNN is used for this work. We use accuracy as the evaluation criterion and search for the best parameters. According to the specified formula, we calculated the flood risk. This model does not use historic flooding data.

## 3. Visualization

Our visualisers can be visualized using Folium maps. The notebook that contains the visualization is in the flood_tool folder and it's called "Vizualisation_Graph2".

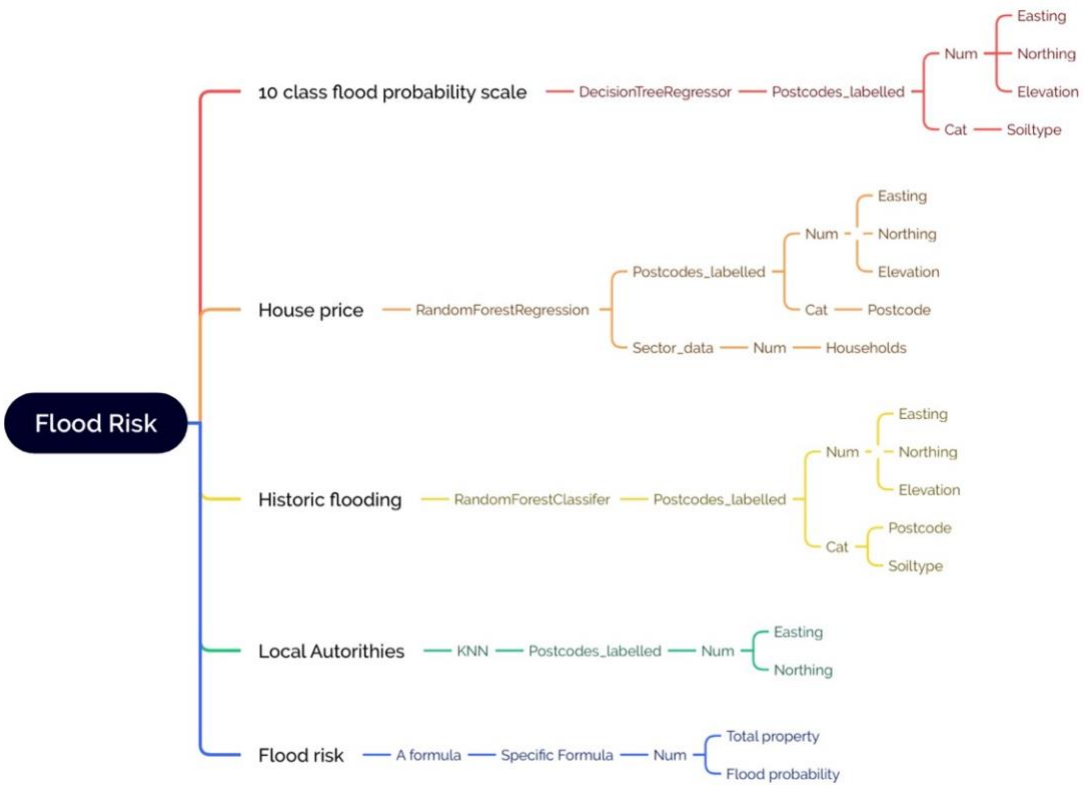
## 4. User Interface

Appropriate scripts, detailed in the package repository documentation, enable the user to run both the risk prediction model and visualisation tools from the command line.
Instruction on how to do so can also be found in documentation.

## 5. Testing

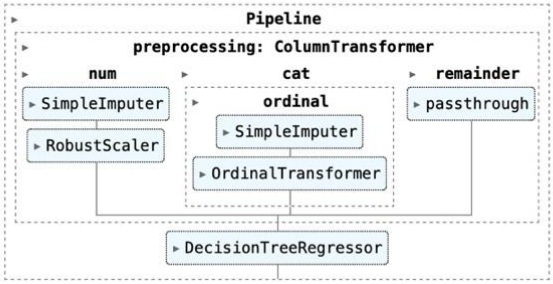
There are test modules to 'pytest' the functions included in tool.py. Instruction on running this test is running the command 'pytest' in the command line.

Appendix 1 (mind map of our work):

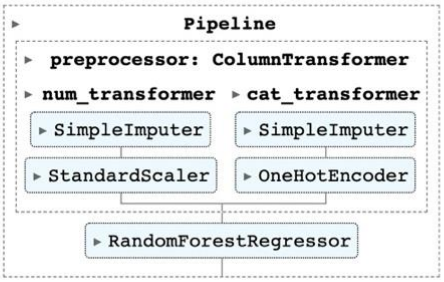


Appendix 2 pipelines' structure:

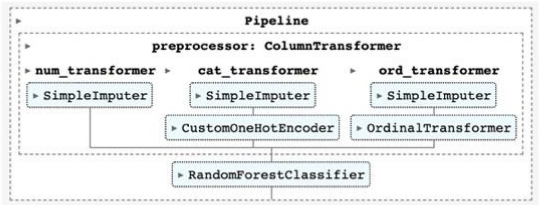
Task1



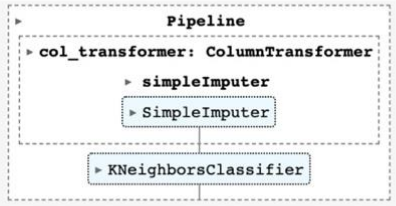

Task2



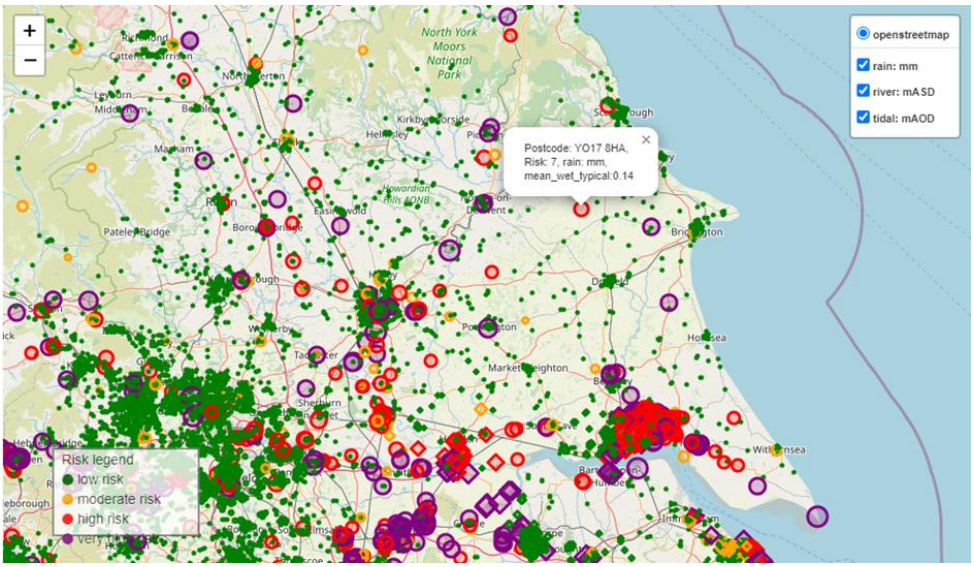

Task3




Task4



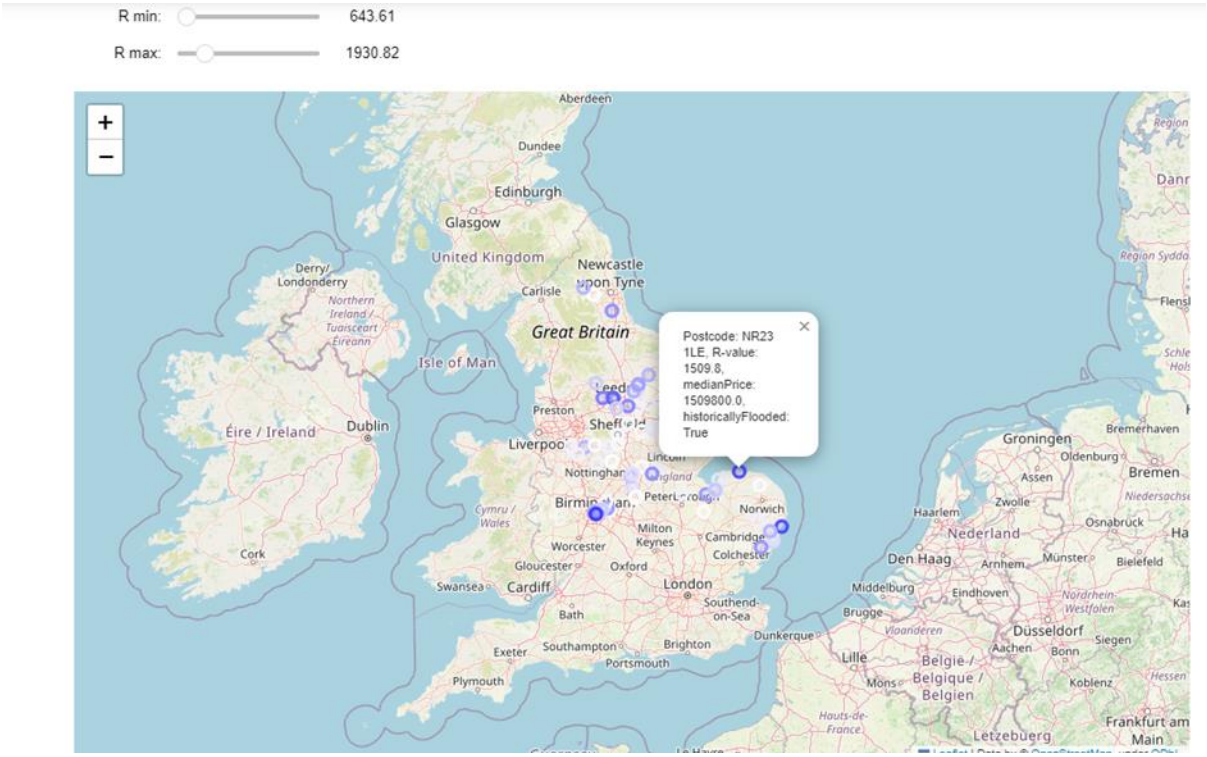
Appendix 3 Visualization for Rain, River and Tidal:




Description: Our different points are displayed according to their category : rain, river and tidal. We can choose which layer to display using the tool at the top right of the map. The colour of our points depends solely on the risk and the size of our points will depend on the risk as well as mean_typ and mean_wet. By clicking on each point, you can access its characteristics.

Rain: circle
river: square
tidal: triangle

Appendix 4 Visualization with the R formula:




Description: We can display our points with a colour that depends on R and varies from white to dark blue. We can choose the range of values we want to display for R, and the colour scale will adjust automatically. By clicking on the points, we obtain the characteristics.