MINI PROJECT CANVAS

Title (preliminary): Greenspaces and residential segregation in Helsinki

Group members: Petteri Huvio, Ossi Inkiläinen, Jenni Liikanen

Workshop # : 5

MOTIVATION

Project outcomes are aimed for decision makers of different counties in Helsinki, to support zoning and land use.

We are investigating whether the amount of green areas has an effect on segregation in Helsinki.

With this data, zoning can be adjusted to prevent it from causing unnecessary segregation.

LEARNING TASK

(focus on problem definition)

Define the problem setting.

We want to find out how the amount of green areas in relation to built-up areas affects the population of the area.

Is this supervised / unsupervised / other...?
Classification / regression / other...?

We can use either linear regression. Linear regression is a supervised regression method.

What are we planning to learn? E.g. What is the target variable / learning outcome?

DATA COLLECTION

Data sources:

- Greater Helsinki Open Statistical Databases
- Register of public areas in the City of Helsinki (WFS API)

LEARNING APPROACH

(focus on solution implementation)

Which ML/statistical methods seem more relevant for the defined problem setting and why?

We use linear regression.

Which evaluation metrics could be relevant?

Root Mean Squared Error (RMSE) might be suitable evaluation metrics.

Is any special treatment relevant regarding how we choose to split the data or how we cross-validate?

We don't have that much data, so we need to use cross-validation. We need

PREPROCESSING

Preprocessing of data should allow us to easily utilise our data, such that the calculations are possible and there's no rogue information. Data should be preprocessed for our analysis with following steps:

- Select relevant data
- Naming
- Formatting data tables
- Clean up the data
- Categorise data
- Remove or quantify open fields
- Normalise data

Geographic data will be edited in GIS software, so that we get numerical values out of the geographical information.

Table formatted data is handled manually into a similar format with each other to streamline usage.

EXPLORATORY DATA ANALYSIS (EDA)

Look at the data!

What steps are you planning to take towards exploring and understanding better the data you have?

What properties would be meaningful to summarize/visualize in this step?

During EDA we are going to look for example mean, median, range, min and max values and deviation of different variables. We can also create different visualizations of individual variables. We try to understand different variables and how predictive they will be for the end result.

We also create a correlation heatmap of different variables to see if there are variables that have a strong correlation, which might affect linear regression.

COMMUNICATION OF RESULTS

Which type of deliverable will benefit most the end-user? Do we choose to write a blog post, create a website, an app, or other..?

How do we communicate best our results to the predefined target group?

Short description of your interface/workflow (if applicable).

We plan to communicate the results in the form of a blog post or website, depending on how interactive our final presentation of results will be.

We will not only provide visualizations of the data but will write some analysis

VISUALIZATIONS

List any meaningful visualizations you are planning to produce that will be useful to the end user?

Depending on the findings, the visualization may be some of the followings:

- histogram
- scatterplot
- boxplot
- barplot
- lineplot

Are you planning to produce any interactive visualizations?
If so, which types of interactivity might be useful to the end user?

If we have time, we would like to create an interactive map where you can select different areas of the Helsinki region and get information on that area.

DATA PRIVACY AND ETHICAL CONSIDERATIONS

(if applicable)

Are there any fairness constraints that apply to our proposed pipeline?

Is there a need to ask for consent during the data collection process?

Is there a need for data pseudonymization/anonymization?

Any other privacy considerations that come to mind?

We will only use public information that does not contain personal data.

What variables are we using as input?

We will for example use these variables as input:

- the ratio of green areas to the whole area
- the relationship of the built-up area to the whole area
- the ratio of the water area to the whole area
- the length of the coastline
- income and level of education
- Household size
- employment
- migration statistics
- population forecast

to make tests to find the best size for resampling. Probably it is something between 10 and 20 % of the data.

based on our findings in order to make it easier for the reader to see the main issues..

ADDED VALUE 🎁

Is there a possibility for added value from the data we're planning to use?

What is the added value?

How are predictions turned into added value for the end-user?

We want to make it clear how the greenspaces affect the residential segregation in Helsinki. We also hope to gain an insight on how they affect the residential well being.

Considering the optimal balance between greenspaces and built-up areas will enable better town planning.

LEGEND

WEEK 1: Data collection/preprocessing

WEEK 2: EDA & visualizations

WEEKS 3-4: Machine/deep learning

WEEK 5: Fairness & data privacy