Course: Introduction to Data Science, Fall 2023

Project: Defluenzer, technical report

Team: Jouko Arko, Martin Häggblom, Micaela Roschier

Link to web-page

1. Introduction

Every year, influenza causes some 290 000 - 560 000 respiratory deaths globally [1]. The Project named “Defluenzer” aimed to provide a model to predict the risk level of getting influenza for elderly people in Helsinki, because in industrialized countries most deaths associated with influenza occur among elderly people [1]. The Reason for choosing the Helsinki region as the geographical region, was that the team was most familiar with that region as a team.

The topic selection and scoping succeeded quite well, since the original plan worked out all in all quite well and no major pivots were needed throughout the project. One item that could have been done differently to improve further the working process and the quality of the outcome, could have been to interview some expert in the field in the beginning of the project e.g. a Finnish Institute for Health and Welfare (THL) expert.

2. Data collection

Two data sets were used to construct the model for predicting the risk of getting influenza:

- Reported monthly cases of influenza in the Uusimaa region in Finland from January 2011 to December 2021 were obtained from THL [2], Hereafter referred to as THL data.
- Observed total rainfall per month and average temperature per month in Kumpula, Helsinki from January 2011 to December 2021 were obtained from Finnish Meteorological Institute [3], hereafter referred to as FMI data.

One of the best decisions our team made during the Data collection process step, was to focus on only two reliable data sets in this project and leave other potential data sets for further later development of the model such as traffic data. The reasoning was that if the project can validate an added value of combining two data sets, then combining additional data sets could potentially bring further value.

Since THL data could only be found as reported monthly cases of influenza, the objective of the project had to be slightly modified from predicting risk of getting influenza on a weekly to a monthly basis. The FMI data would also have been available on a daily basis, but monthly data was chosen so it would be in the same format as the THL data. From this we learned that it would be a good idea in the Data collection process step to take earlier into consideration how to possibly combine different data sets.

3. Preprocessing

The THL data was obtained as a csv file with time and val columns. The time column was in the format yyyy-mm and val column was the number of influenza cases for that specific month. The time column was splitted into separate columns for year and month. The val column contained some empty values, but since the previous and aftercoming month contained low values (< 8) it was assumed that the empty values corresponded to 0 reported influenza cases. Hence the empty values were changed to 0.

The FMI data was also obtained as a csv file. The FMI data did not require any further preprocessing beside selecting the relevant data columns: Vuosi, Kk, Kuukauden sadesumma (mm), and Kuukauden keskilämpötila (degC).

The csv files were easy to preprocess. The teams' decision in the data collection process step to focus on high quality data paid off especially in this preprocessing step. No major reporting items regarding what did not work, possible changes vs. original plan, nor what could have been improved.

4. Exploratory Data Analysis (EDA)

After combining the THL and FMI data into a new data frame it was noticed that the reported influenza incidents differed quite significantly during the corona pandemic Oct 2020 – Jun 2022 compared to similar time periods prior and after the corona pandemic (Figure 1). Due to these abnormal amounts of influenza cases, the previously mentioned time period was removed from the data frame for further analysis.

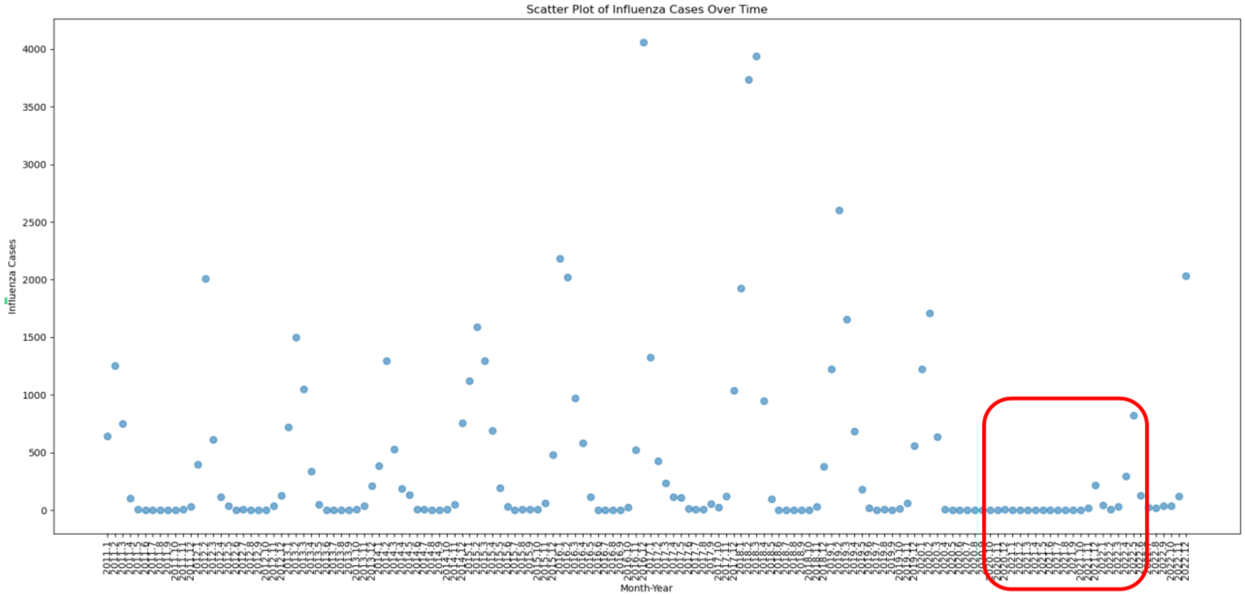


Figure 1. *Scatterplot of reported influenza cases per month. The red shape indicates abnormal amounts of influenza cases compared to other similar time periods.*

Afterwards a graphical EDA was conducted with Matplotlib and Seaborn for the data frame where the time period Oct 2020 – Jun 2022 had been removed (Figure 2).

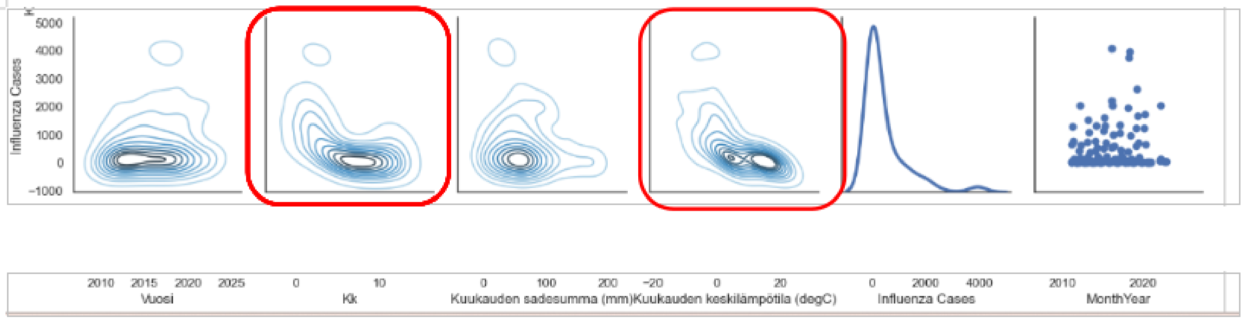


***Figure 2**. Graphical EDA of Influenza cases vs. year, month, monthly rainfall, average temperature, and MonthYear.*

From the graphs it seemed initially that month, and average temperature, affected the amount of reported influenza cases.

The EDA step worked quite well in the project. The abnormal reported cases of influenza differed quite significantly from other historical values during the corona pandemic, hence easy to detect in the EDA. One item that could have been done differently, would have been to proceed earlier to the EDA phase in order to get first validation indications that the data would be useful for the project.

5. Visualizations

The website named Defluenzer makes it possible for the user to choose a date in a calendar and include the temperature of the weather forecast for that day and the website will give the risk of getting infected for that day. The risk level is given as low, medium and high risk. The website will also show one of these smiley faces depending on the risk level.

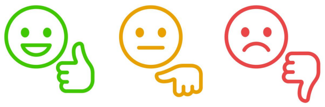
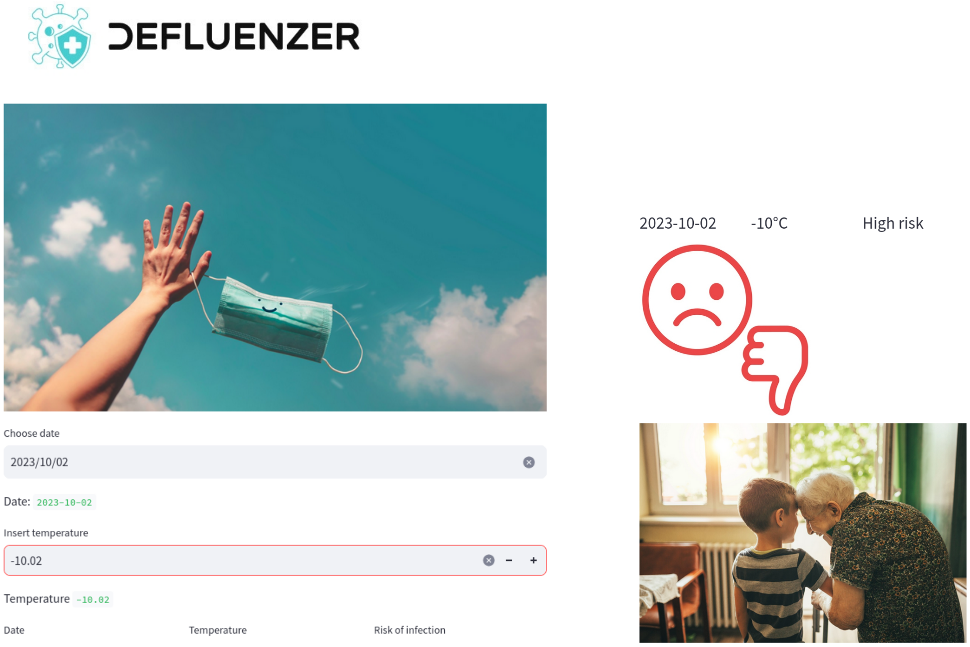


***Figure 3.** Risk levels of getting infected visualization on web-page.*

The Streamlit website code includes the machine learning model and the model will use these input values (date, temperature) for evaluating the risk of getting infected.



DEFLUENZER

Choose date

2023/10/02

Date: 2023-10-02

Insert temperature

-10.02

Temperature -10.02

Date Temperature Risk of infection

2023-10-02 -10°C High risk

Figure 4. *Picture of website's intake of user input (left) and result output (right).*

Even though the group had no previous experience in doing websites, we managed to build a website with Streamlit. That combined with the successful implementation of an easily comprehensive low/med/high risk level, the visualization process step worked really well and could be executed according to the plan.

It would have been better if the user could only choose the date in the calendar and the website would give the prediction. But for this we would need more data and also somehow connect the weather forecast automatically to the code.

Website: https://defluenzer.streamlit.app/

6. Machine learning

80% of the data was reserved for training the supervised linear regression model, and 20% of the data was reserved for the test. Also 70/30 and 50/50 data split of train/test was tested, but that did not improve the score compared to the 80/20 split.

Score train data was 0.39. Score test data was 0.22. The model explains the data quite well, since only 22% of the variability in the outcome data cannot be explained by the the model.

In the beginning of the Machine learning phase, the team spent some time trying to fit the data with a Logistic regression model, despite Logistic regression was not suitable for the data and hence not the project neither. The team could perhaps have tried to understand better the purpose of Logistic regression, prior to attempting to apply it to the data.

7. Communication of results

According to the plan, the team succeeded with providing low/med/high risk of getting influenza in a web-page format.

The initial results were communicated on the 12th of October in the pitching session. The pitching session went quite OK, by following the given course material guidelines for pitching.

Beside the pitching session, the team has presented the project outcome to some elderly people that the team are familiar with. The initial feedback has been positive. The elderly people seem to value a single information source to predict the risk level of getting different diseases.

Overall the communication went quite well, with a nice extra bonus to present the initial results and ideas also to some in the target group.

8. Summary

The project successfully confirmed a correlation between weather data and influenza cases, demonstrating the potential to enhance the understanding of influenza infection risk beyond relying solely on historical reported influenza cases, which was the project's primary objective.

Due to this validation the future development steps would naturally include adding additional data sets to the model, to improve predictability of the influenza infection risk even further. Some potential data sources could be traffic data, event data, and online news content.

These data sources would give further insight in how people and information flows affects the reported influenza cases.

Other natural next steps beside adding more data sources would be to increase the geographical scope from Helsinki to other locations as well.

References:

1. https://www.who.int/news-room/fact-sheets/detail/influenza-(seasonal)#:~:text=There%20are%20around%20a%20billion,650%20000%20respiratory%20deaths%20annually.
2. https://sampo.thl.fi/pivot/prod/fi/epirapo/respinfcare/fact_epirapo_respinfcare

3. https://www.ilmatieteenlaitos.fi/havaintojen-lataus