Stroke patients' rehabilitation in Estonia 2010-2020

GitHub repository: https://github.com/K-AMeus/IDSProject

Team members:

1) Kevin Kits
2) Richard Prost
3) Karl-Andreas Meus

Task 2. Business understanding

1. Identifying Your Business Goals

1) Background
    - Stroke is a major public health concern and its impact on individuals and the healthcare system is substantial. In Estonia, between 2010 and 2020 stroke patients underwent many rehabilitation processes. The after-care treatment is not very efficient at the moment and it is also not clear what is the best way to do it. Since rehabilitation is vital for patients' health it is very important to try to make it more efficient and better.
2) Business goals
    - Find out how useful it is for the state and Põlva hospital to keep stroke rehabilitation in the Estonian healthcare system.
    - Find out exactly how efficient is the rehabilitation both in general and by county
    - Highlight regional differences and per capita comparisons.
3) Business success criteria
    - Põlva hospital and/or state can decide whether they should change the treatment process of a specific type (the intensity of psychotherapy for example)

2. Assessing the situation

- Inventory of resources:

  Stroke patients rehabilitation data,

3 BSc students,
1 medical expert,
3 university provided laptops.

- Requirements, assumptions and constraints:

  Not exactly clear what data privacy problems we could face but obviously we have to play it safe. Data should be complete and we might not get access to further/additional data in time as getting this data was quite a process in itself.

- Risks and contingencies

  There could be a social engineering risk factor for our project. As the data is obviously not publicly available and could be sensitive, we have to keep our guard up. Don't know if we can have a backup plan in case that happens.

- Terminology

  Stroke types:
  Hemorrhagic - Due to bleeding into the brain
  Ischemic - Due to cut off blood supply to the brain

  The description of all the attributes are under the task "data understanding"

- Costs and benefits

  Costs: 90h of brainpower
  Benefits: Improved healthcare efficiency, predictive analytics for stroke patients' rehabilitation, personalized medicine/treatment for patients (Ideally)

3. Data-mining goals

1) Data-mining goals.
   - Analyze the differences between active rehabilitation and post-rehabilitation and how they impact patient outcomes.
   - Evaluate whether rehabilitation provides benefits to the Estonian healthcare system.
   - Display the receipt of therapies over time and compare different types of rehabilitation.
   - Highlight regional differences in rehabilitation and conduct a per capita analysis.
   - Create a comprehensive visual representation across years and counties to encompass all significant findings.

- Develop a strategy to handle the presence of numerous zeros in the dataset to ensure that it does not lead to misleading information, especially in scenarios where the median might be zero.

2) Data-mining success criteria
    - Clear identification of differences in outcomes between active and post-rehabilitation.
    - Benefits could be associated with reduced post-rehabilitation needs or lower healthcare costs. Use Kaplan-Meier curves for survival analysis.
    - Significant differences in the timing of therapy receipt should be noticeable. Compare the impact of different types of therapies on patient outcomes.
    - Clearly demonstrate differences in the frequency and quality of rehabilitation across counties.
    - Monitor and handle null values' impact on the analysis.
    - The graphs should be clear and easily interpretable.
    - Highlight key trends and differences across years and counties.

Task 3. Data understanding

- **Gathering data**

    a. Outline data requirements
    We are probably going to need all of the data we can get our hands on. The most important attributes might be: stroke_type, treatment_method, age, therapy_phase. It's possible we find that some other attribute are much more important.

    b. Verify data availability
    The data that we have is not publicly available. This is private data that we got from our medical expert source. As far as we are aware we will get enough data (at this moment we don't have all of the promised data, only a sample).

    c. Define selection criteria
    To narrow down the dataset for our analysis, we will focus on relevant years (2010-2020). Additionally, we may filter out records with missing or inconsistent data, ensuring the reliability of our findings (null attributes).

- **Describing data.**

    The dataset consists of stroke patient rehabilitation records from 2010 to 2020 in Estonia. In the dataset we have attributes: Id (patient identifier), age, sex, county, comorbidity (the simultaneous presence of two or more diseases or medical conditions in a patient), stroke_type (ischemic or

hemorrhagic), dementia, year, treatment_method, therapy_phase_I (active treatment), therapy_phase_II (follow-up treatment), time (time of death), status (alive or dead).

- **Exploring data (the main general data)**

  We will explore our data deeper when we receive it fully. From the current data example we have we can see the attribute data types.
  - **Id -** integer
  - **Age -** integer
  - **Sex -** char ‘f’ or ‘m’
  - **County -** string
  - **Comorbidity -** integer
  - **Stroke_type -** string (ischemic or hemorrhagic)
  - **Dementia -** string (yes or no)
  - **Year -** integer
  - **Treatment_method -** string (surgical or nonsurgical)
  - **therapy_phase_I -** float
  - **therapy_phase_II -** float
  - **Time -** float
  - **Status** - integer (0 or 1)

- **Data inside CSV file “filtreeritud_valim_cleaned” (must used file):**
  - Unnamed - integer (row number from the file before cleaning)
  - Id - String (patient’s ID)
  - Vanus - integer (age of a patient)
  - Sugu - char (gender - M/N)
  - Maakond - string (county from where the patient is from)
  - Pohidiagnoos string (diagnosis by ICD-10 codes; RHK10 in Estonia)
  - Charlson_total_quan - integer (Charlson Comorbidity Index score calculated using a quantified approach)
  - Charlson_total_original integer (The Charlson Comorbidity Index score calculated using original index)
  - Elixhauser_total_vw - integer (Elixhauser Comorbidity Index calculated using the van Walraven (VW) weighting method)
  - Elixhauser_total_swiss - integer (Elixhauser score calculated using a weighting system)
  - Surmakp - date (date of death)
  - Aeg_surm - integer (the number of days from the start of the treatment; stopped counting if a patient died)
  - Staatus_surm - integer (1 if patient is dead and 0 if alive)
  - Arve_nr - integer (invoice number)
  - Aasta - integer (year when the treatment started)

- Aastaid_algusest - float (how much time has passed from 01/01/2010
- Ar_algus - date (start of an active treatment)
- Ar_lopp - date (end of an active treatment)
- Raviasutuse_nimi - string (name of the hospital where a patient was treated)
- Raviasutuse_maakond - string (county of the hospital where a patient was treated)
- Raviasutuse_linn - string (city of the hospital where a patient was treated)
- Ravityyp - string (treatment type)
- Eriala - string (speciality)
- Raviarve_summa - integer (the amount of the medical bill)

The columns for after-care treatment:

1) {treatment_type}_ar - hours of active treatment of a specific type (physiotherapy, logopedia, occupational therapy, psychology)
2) {treatment_type}_1a - hours of active and after-care treatment of a specific type
3) Algsaeg_jarelravi_{treatment_type} - at which day of treatment did the after-care start
4) Algusstaatus_jarelravi_{treatment-type} - whether the patient started that specific treatment or not (1 or 0)
5) Lopuaeg_{treatment_type} - the day when the specific treatment ended
6) Lopustaatus_{treatment_type} - whether the patient ended that specific treatment or not (1 or 0)

The columns for comorbidity

1) Charlson_{abbreviation for the other disease} - comorbidity calculated with the charlson comorbidity index (for example charlson_diab with a value of 1 means that the patient had stroke and has diabetes as well)
2) Elixhauser_{abbreviation for the other disease} the same but with elixhauser index

Columns for general treatment information in active treatment

1) Rhk_ar - the RHK10 indexes for the diseases the patient has
2) Ncsp_kood_ar - the codes for the operations the patient had
3) Ncsp_nimi_ar - the name of the operations tha patient had
4) Aeg_eelnev_dgn - the days from the last diagnoses
5) Aeg_jargnev_dgn - the day for the follow up diagnoses
6) Teenused_ar - (services the patient received)

Task 4. Planning your project

1. **Data Collection (3h/student)**
    a. Figure out what we want to collect
    b. Find a source for data
    c. Get permission to use data
    d. Get data

2. **Data Cleaning (5h/student)**
    a. Figure out our preprocessing.
    b. Clean the data
    c. Handle missing values

3. **Exploratory Data Analysis (5h/student)**
    a. Generate descriptive statistics and visualizations for key attributes.
    b. Conduct correlation analysis between variables.
    c. Find promising patterns
    d. Find anomalies and try to interpret

4. Model Development (5h/student)

a. Choose appropriate models for predicting rehabilitation outcomes.

5. Model Evaluation and Validation (6h/student)

a. Assess model performance using metrics like accuracy, precision, and recall.

6. Documentation and Reporting (6h/student)

a. Summarize findings, insights, and recommendations.
b. Prepare a comprehensive report outlining the entire process.

TOOLS:
1) Python (Pandas, Numpy, Matplotlib, Seaborn)

2) Jupyter notebook, Deepnote

3) Github, Trello, Discord