


Task 2

- **Identifying your business goals**

- **Background-** Our client, city of Tartu, seeks for more efficient use of the smart bikes and public transport during the big events at Raadi. When  was performing in Raadi, the problem was that the public transportation was full of people and all the smart bikes were in use. Because the bikes came in to use this summer, there has not been done much research about them. For our research we are creating a RandomForestClassifier to predict how will the usage of the bikes and public transportation be during the next events and to predict based on bike usage if there was an event somewhere yesterday.
- **Business goals-** We would like to make the city's public transportation and bike sharing service more efficient by showing which routes or lines get congested when a concert is happening in Tartu.
- **Business success criteria -** The project is considered successful if we can find some sort of way to reduce congestion and make the city's transportation smoother during periods of high usage.

- **Assessing your situation**

- **Inventory of resources-** We are going to use the Tartu bike and bus datasets. For data analyzation we will use Python and it's libraries. For Location data we use google api.
- **Requirements, assumptions, and constraints-** We are not allowed to share the data and get rid of it after completing the project. The project should have some useful conclusions about the data and a proposed solution to help fix the problems we may come across. Also, since the project has to be presented with a poster, we have to make the results of our work look good and clear and design the poster nicely.
- **Risks and contingencies-** The fact that we are not very experienced in data mining means that the workflow might be slowed down.

- **Terminology:**
 - SKLEARN** - Python library we use to create a variety of classifiers from which we will choose the best one.
 - Pandas** - Python library we use to analyze data.
 - Matplotlib** - Python library used by us to visualize plots.
 - **Costs and benefits:** Does not affect our project.
- **Defining your data-mining goals**
 - **Data-mining goals** - To create a classifier that predicts events based on the usage of bike data. Also make a map, showing the routes that people take most during events and during the regular day (no events). This map consists of both the bus and the bike routes.
 - **Data-mining success criteria** - Our data-mining is successful if the classifier predicts events with about 75% accuracy and create about 75% accuracy map of the routes used by bikes.

Task 3

- **Gathering data**
 - **Outline data requirements - We need the following data:**
 - a set of bike location data that shows a bikes exact coordinates every 5 or so seconds (time range: 1 day)(amount: more than 5 days worth)(at least 1 day should be during Metallica concert and at least 2 should be from days with no events)(file should be csv)
 - a set of bike data that shows when was a bike taken from which bike stop and where/when was the bike deposited (time range: 1 day)(amount: more than 15 days worth)(file should be csv)
 - a set of bus ticket validation data, that shows when someone with which bus from where and to which direction drove (time range: at least 1 day)(amounts: more

than 15 days worth)(file should be csv)

- a map(in terms of programming) of coordinates for different bike stops (could be a csv file, but sadly is a broken json file)
- **Verify data availability** - We have at our disposal all the needed data of the bike and bus routes during the Metallica concert. Said data (see the section below) is also openable in Python and we can make it into a dataframe. We will also see if there were other big concerts in Tartu on days that we have data for, to make our classifier more accurate.
- **Define selection criteria:**
- **BUS VALIDATION DATA:**
 - 01.07_10.07.csv
 - 11.07_20.07.csv
 - 21.07_30.07.csv
 - 01.08_10.08.csv
 - 11.08_20.08.csv
 - 21.08_30.08.csv
 - USED DATA FIELDS: Liin, Suund, Aeg, Peatus, Peatuse kood, Reisijaid
- **BIKE ROUTES DATA:**
 - routes_201906.csv
 - routes_201907.csv
 - routes_201908.csv
 - routes_201909.csv
 - routes_20190718.csv
 - routes_20190719.csv
 - routes_20190725.csv
 - routes_20190726.csv
 - USED DATA FIELDS: route_code, unlockedat, unlockedatetime, lockedat, lockedatetime, startstationname, endstationname, length
- **BIKE LOCATIONS DATA:**
 - locations_201906_part2.csv
 - locations_201907_part1.csv
 - locations_201907_part2.csv

- locations_201908_part1.csv
- locations_201908_part2.csv
- locations_20190718.csv
- locations_20190719.csv
- locations_20190725.csv
- locations_20190726.csv
- USED DATA FIELDS: route_code, latitude, longitude, coord_date, coord_time

■ COORDINATES MAP:

- 2019_08_28_bicycle_stations_public_and_metallica.json
- USED DATA FIELDS: name, areaCentroid (which as a class has fields latitude and longitude)
- Google map is used to visualize coordinates on the map.

- **REPORT** - We got the starting data privately from our instructor. After looking at the data we realized that data from the day of the Metallica concert was absent. We then found out that there was a public dataset posted on piazza with the needed days' data. While trying to open and look at the different datasets, a few problems came up. The biggest problem was the bike stations locations file, which was in a json format. While trying to parse this file, there were many errors, which were fixed by code from a fellow student. Then, after some googling, we got the parsing to work and converted the json file to a pandas dataframe. The rest of the csv files opened neatly, with no problems so far when converting them to pandas dataframe.

• Describing data:

Routes datasets have about 60 000 lines each, locations datasets have about over 130 000 lines each and the bus datasets have about 240 000 lines each. The datasets satisfy our needs for data, as the needed fields (mostly time and location) are there. There are enough lines in total, if anything there is too much data, but not knowing how much we will need, we have left most of it in here. There is also data from the old bus system (June), which we will not be using, as predicting and mapping an old system along with the new system makes little sense. The most

important part is the data when  concert happened.

BIKE ROUTES DATA:

Used Columns: Route code(long number used as an ID for the data), Unlock date(YYYY-MM-DD), Unlock time(HH:MM:SS), Lock date(YYYY-MM-DD), Lock time(HH:MM:SS), Start station(Station name), End Station(Station name), Ride Length(Kilometres).

BIKE LOCATIONS DATA:

Used Columns: Route code(long number used as an ID for the data), Latitude(Google maps coordinate), Longitude(Google maps coordinate), Coordinate date(YYYY-MM-DD), Coordinate time(HH:MM:SS).

BUS VALIDATION DATA:

Used Columns: Liin (shows which bus line the data is about, details can be searched up from peatus.ee), Suund(shows in which direction the bus is going (from point A to point B or vice-versa, also shows if the bus skipped a stop or not (not needed for our task)), Aeg (HH:MM:SS), Peatus (the vague name of the stop that the bus is at), Peatuse kood (the more specific code for a stop (not equivalent to the name of the stop)), Reisijaid (shows how many tickets were bought in bulk)

COORDINATES MAP:

Used Columns: name(shows the name of the stop, which can be found in the bike routes and bike locations data), areaCentroid.latitude(Google maps coordinate), areaCentroid.longitude(Google maps coordinate)

- **Exploring data** - We are familiar with the datasets and their different features and although the data quality is pretty accurate, there are some minor issues (like file ending getting cut off) with it which hopefully get fixed in the most part. Some of what we have to fix ourselves. A quick theory we had was that during Metallica the lines like 7 (that go from central town to Raadi) get more action than usual. So a quick coding session found out that on the 11th of July there were a total of 1044 registered entries on bus 7 while on the 18th of July (day of the Metallica concert) there were a total of 1288 entries. This does not include the fact that some people may have bought tickets twice, but it supports the idea that people used bus 7 more. Other busses like line


25 didn't see such an increase, as it only increased the riders from 199 to 203.

- **Verifying data quality** - Data is not very well delivered, it has some minor deficits. Some of the data is missing from the end, lines have been left uncompleted. Datasets are in separate files so it was a little bit difficult to find and gather them all. Also, there are some aspects of the data that we do not need, which will have to be cleaned later.

Task 4

Planning your project

List of tasks to accomplish maximum performance in our project:

1. **Getting the data** - For the first task we have to get the right data on which we can build our project. We have received most of the data, but we do not have all the information we need to create our project. We will gather information about different events that have happened around the town on the days that we have smart bikes data about. This will take each member about 3 hours.
2. **Cleaning the data** - Secondly we will clean the data of the information that we do not need. We extract what we need and reformat it how we need it. Then we will integrate our data to make one dataset. This should take each member also about 3 hours.
3. **Analyzing the data** - When the data is cleaned and looks all nice and neat then we will start analyzing the data. In this part of the project we will separate the trash from the good data. Analyzing the data can will take most of our time. Estimated time for each team member is about 5 hours. **It will be done in two separate tasks:**
 - 1) First we will look at the bike's data and see the routes and most used bike stops. Then we will look if there were events near the bike stop or figure out why is the bike stop used the most.  concert data is the one that we will use to train our predictor. All the random bike usages will be removed from the data. As the bike data is larger, time estimation for it is 3 hours.

2) Second subtask is that we will look at the bus data and see where the most people got on the bus and see where the bus was going. Depending also on what ticket is used we can predict if the person went to the event. Bus data analyzing time estimation is 2 hours of the 5 hours.

4. **Creating maps of the routes used** - This is also separated into two different tasks. As this is one of our goals, we want it to be perfect so we will have to integrate it with google maps. So estimated time for this task is about 8 hours per team member.

1) Creating the map of used bike routes.

2) Creating the map of used buses.

5. **Creating the randomforestclassifier** - When we reach this part of the project we start training our randomforestclassifier to start predicting where and when an event would have happened. Then we will see if it is true and then train it again and eventually start using it on the real data to see if it gets the right results there also. As this is also very important in our project, time estimation for this task is about 6 hours each member.
6. **Creating the poster for our project** - As for everything in our project, we want everything to be well made and well presented. So this definitely has to be in comic sans, very colorful and has to be seen from a great distance. Time estimation for this is also a tough 3 hours of giggling for each member of our team.

Methods and tools we plan to use:

- Winrar
- Jupyter notebook
- Piazza
- Excel
- Randomforestclassifier
- Sklearn
- Matplotlib
- Pandas
- Google chrome
- Google maps
- Github

- Facebook(for event data)
- Piletilevi(for event data)
- Google search