THESIS SUPERVISOR'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Stroke mortality prediction**
Author's name:	**Regina Mavrina**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Computer Science
Thesis reviewer:	Ing. Matěj Klíma
Reviewer's department:	Department of Computer Science

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **ordinarily challenging**

How demanding was the assigned project?

Compared to other assignments of master thesis in the software engineering field, I find this one ordinarily challenging. The development part of the work is easy and doesn't require great programming skills. On the other hand, the analytical part of the assignment I find challenging, and the student has to prove her knowledge of statistics and manipulation with data.

Fulfilment of assignment **fulfilled with major objections**

How well does the thesis fulfil the assigned task? Have the primary goals been achieved? Which assigned tasks have been incompletely covered, and which parts of the thesis are overextended? Justify your answer.

The purpose of the assignment was to implement a stroke mortality forecast among the population, using various statistical methods, based on data from the public web platform of open data Kaggle.
The student showed a quick reaction to actual situation when she added a part dedicated to the actual COVID-19 pandemic and its relationship to stroke, although it wasn't part of the assignment.
According to the conclusion in the thesis, the best approaches to stroke prevention are the Law of Large Numbers (LLN) forecasting and the Bayes method. However, it is hard to prove this statement, because the attachment contains only the R code and not the dataset from which it was deduced. Therefore, I am deeply concerned whether it is possible to reproduce the experiment based on the information provided in the thesis, and therefore I am not sure how valid the findings are.

Technical level **E - sufficient.**

Is the thesis technically sound? How well did the student employ expertise in his/her field of study? Does the student explain clearly what he/she has done?

The description of medical data analysis in Chapter 1 and the description of the statistical methods in Chapter 2 is good. It is clear which software the author used to run the experiment and what was the source of the analysed data. However, I am missing some information about the analysis, development, quality assurance, and execution of the software behind the experiment. Afterall it is still a Master thesis in the field of Software Engineering, and those activities should be present in the thesis.
I have some concerns, whether there were always used the correct charts to document some statements (for example, the purpose of the box plot in Fig 1-8, which is mistakenly called a histogram, is unclear to me). Therefore, it seems that the charts could be more described. And the tables too – what is, for example, the meaning of columns ITTER, Value, and Code in Table 4 in section 1.4?).

Formal level and language level, scope of thesis **E - sufficient.**

Are formalisms and notations used properly? Is the thesis organized in a logical way? Is the thesis sufficiently extensive? Is the thesis well-presented? Is the language clear and understandable? Is the English satisfactory?

The readability of the work is made difficult due to a number of formal errors. For example, the headings have different fonts throughout the work, and the font of the body of the text sometimes varies too. Then, some of the sizes of the captions in the graphs are on the edge of readability (Fig. 1-6), some graphs don't have their axis properly named (Fig. 1-

1), some graphs even contain labels in a different language (Fig 1-11 appears to be in Italian). The spacing between paragraphs is inconsistent, and the sizes of the individual code samples are inconsistent too.
Concerning the language**,** sometimes student uses very long sentences, which also makes the readability more difficult. Some of the statements even don't make any sense (page 18: "Stroke is getting younger").

Selection of sources, citation correctness **F - failed.**

Does the thesis make adequate reference to earlier work on the topic? Was the selection of sources adequate? Is the student's original work clearly distinguished from earlier work in the field? Do the bibliographic citations meet the standards?

In the thesis, none of the recommended citation styles was used, and some of the references to the web pages are incorrect (e.g., No. 7, or No. 12 in the Bibliography chapter).
For many of the statements in the thesis, the proper reference is missing (e.g., The last two paragraphs of Section 1.1, or the causes of deaths in the world in Section 1.3, and many more).
I am missing more information about the covid_19 dataset used in Section 1.4 (origin, date, validity).
The number of sources is satisfactory, although it could definitely contain more credible sources.
The biggest objection I have is connected to the originality of the experiment presented in the thesis. Two years ago, there was published a web article called "(Bio)statistics in R," located at https://www.kaggle.com/ruslankl/bio-statistics-in-r-part-1 that contains many similarities, mostly in part called "Healthcare Dataset Stroke Data." In there are located charts that look the same as in the thesis. Sorted in the order of position on the website, the list of basically the same charts is on figures: 2-2, 1-6, 1-7, 1-9. Table 3 contains the same values as on the page, only in different order and precision. From the third part of the article, located at https://www.kaggle.com/ruslankl/bio-statistics-in-r-part-3, there are the same code samples. Namely figures: 3-9, 3-10, 3-11. Figures 3-12, 3-13, 3-14 are from the remaining second part of the article.

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS, SUGGESTED GRADE

Summarize your opinion on the thesis and explain your final grading.

The thesis contains a high number of formal errors. It doesn't contain important parts, which the thesis in the field of software engineering should contain (requirements analysis, selected architecture, design, and testing). The dataset, upon which the conclusion was made, is missing in the attachment. Therefore, I am deeply concerned whether it is possible to reproduce the experiment based on the information provided in the thesis, and how valid are the findings.

I find the similarity of the solution presented in the thesis to the one at the Kaggle web page https://www.kaggle.com/ruslankl/bio-statistics-in-r-part-1 too big. This page is not included in the bibliography, even though some of the charts, tables, and pieces of R code are the same. The list of the similarities is described in the "Selection of sources, citation correctness" section of this report.

The grade that I award for the thesis is **F - failed.**

Date: **19.6.2020** Signature: