Predicting Nutrition Requirements to Boost the Human Immune System

Lecture	Data Science summer semester 2020
Participants	Kadir, Md Abdul (2566465), maktareq@gmail.com Alam, Hasan Md Tusfiqur (2571663), s8haalam@stud.uni-saarland.de Islam, Md Jonybul (2577852), s8mjisla@stud.uni-saarland.de Kosheleva, Elena (2577871), ekosheleva@mpi-inf,mpg.de
Submission date	07-17-20
Chair	Univ.-Prof. Dr.-Ing. Wolfgang Maaß Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken

Executive Summary

The immune system is the defense mechanism of a human being [1]. The immune system fight with viruses (e. g. Coronavirus). Food habit affects the human immune system [2]. So, eating healthy food can increase the recovery rate of CVID- 19 patients. In this project, we explore the relationship between food habits and COVID- 19 effects of the population of a country. Moreover, we predict a healthy diet statistics country wise and visualize the comparison of the predicted and the country's current diet stats. By looking at these stats, a food manufacturer can learn what kind of food they should produce more or should reduce to boost the immunity of the population in a country.

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Contents

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

List of Abbreviation

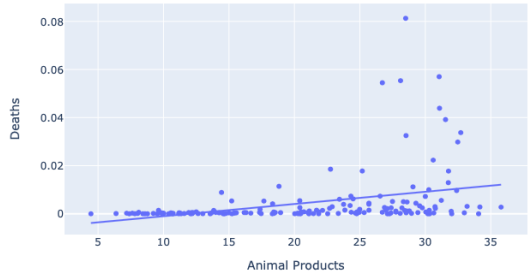
COVID- 19: Coronavirus disease 2019

SVR: Support vector regression

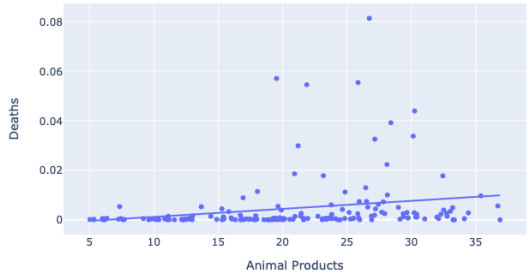
BMI: Body Mass Index

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel:
+49 (0) 681 302 64736

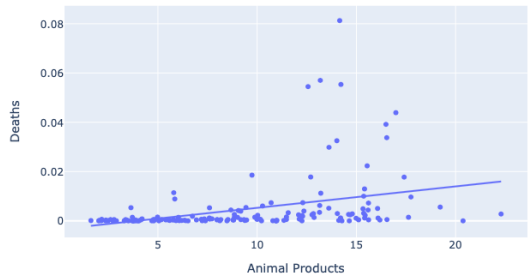
List of Figures



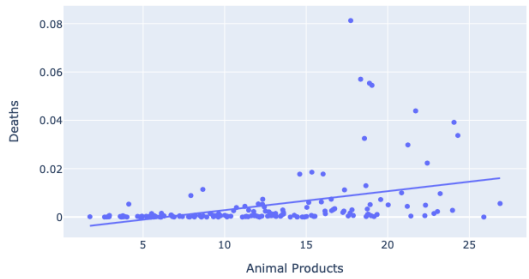
Protein data



Fat data



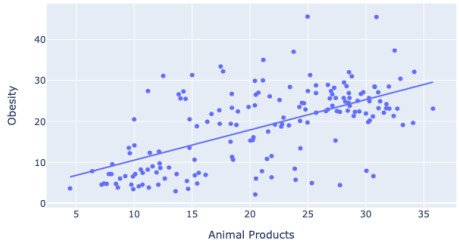
Kcal data



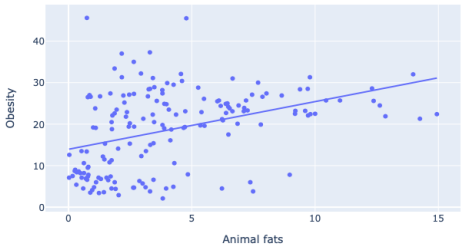
Amount data

Figure 1: Correlation of animal product and death for four kind of data.

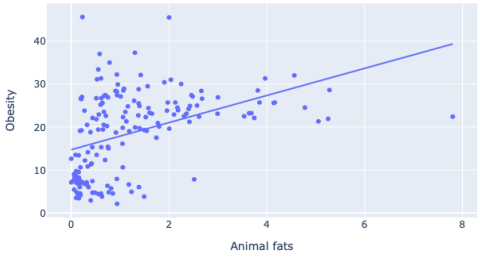
Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736



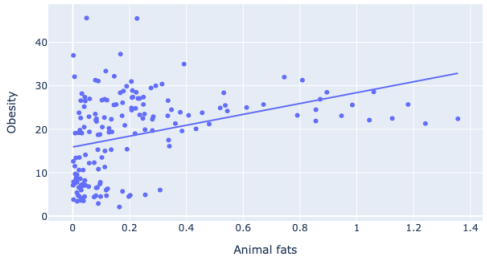
Protein data



Fat data

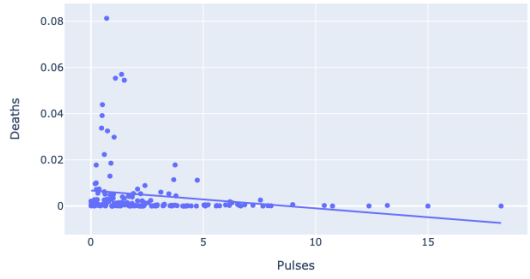


Kcal data

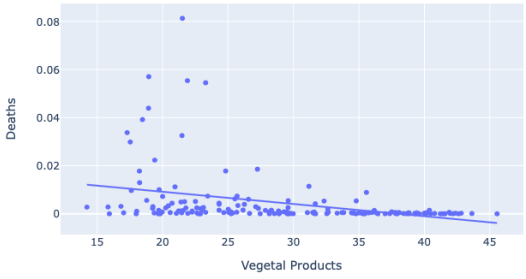


Amount data

Figure 2: Correlation of animal product and obesity for four kind of data



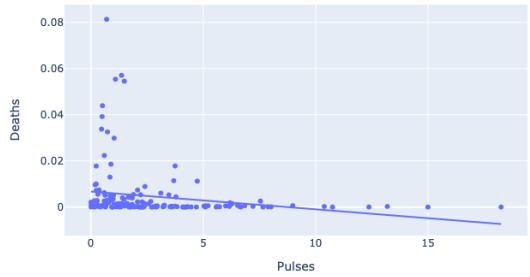
(a) Protein data: Negative correlation



(b) Protein data: Negative correlation

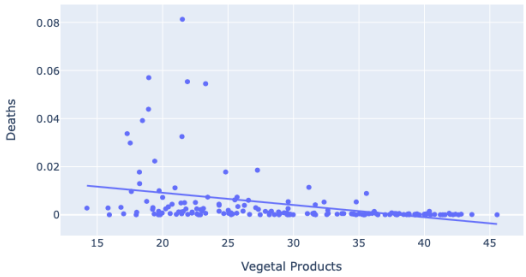
Figure 3: (a) and (b) negative correlation between death and consumption of pulses and vegetal products respectively.

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736



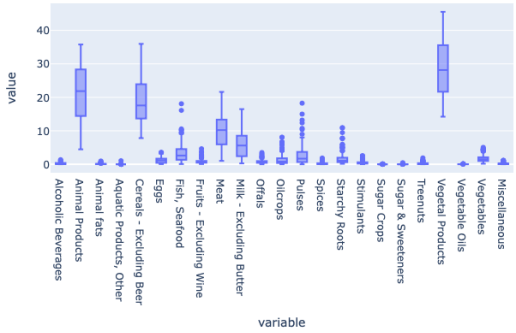
Correlation between Pulses and Deaths is -0.19521348332395022

(a) Protein data: Negative correlation

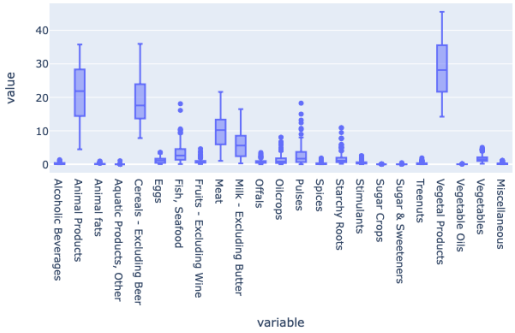


Correlation between Vegetal Products and Deaths is -0.3462739595982403

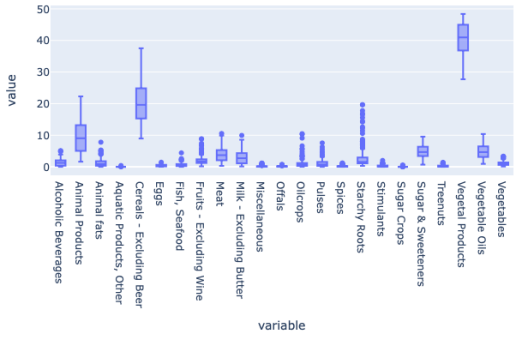
(b) Protein data: Negative correlation



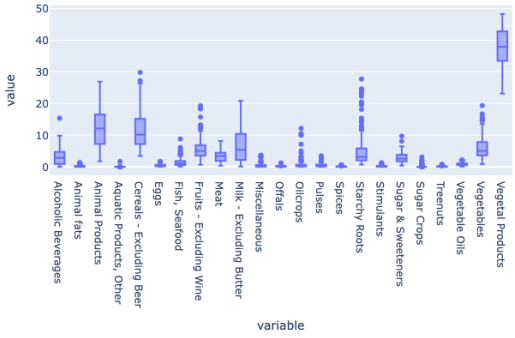
Protein data



Fat data



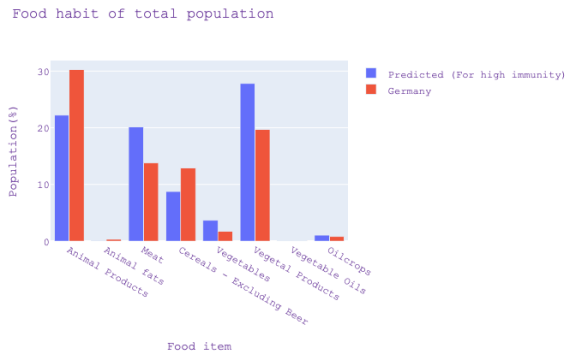
Kcal data



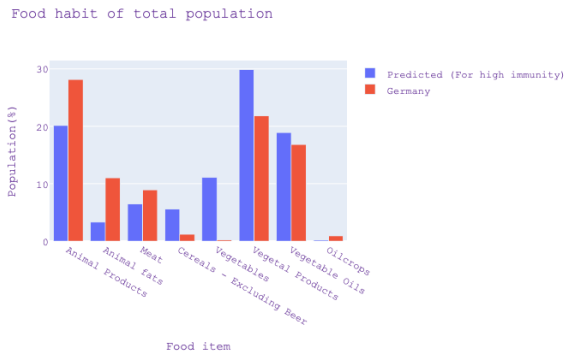
Amount data

Figure 4: Box plot of 4 types of data to recognize outliers.

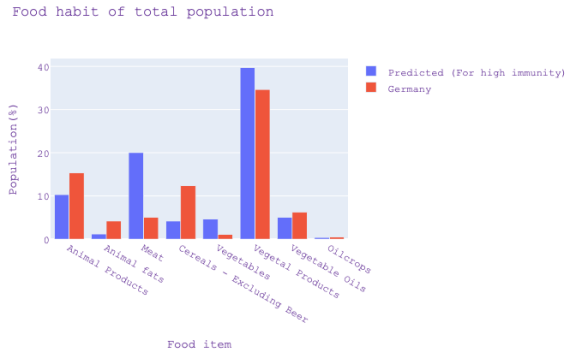
Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736



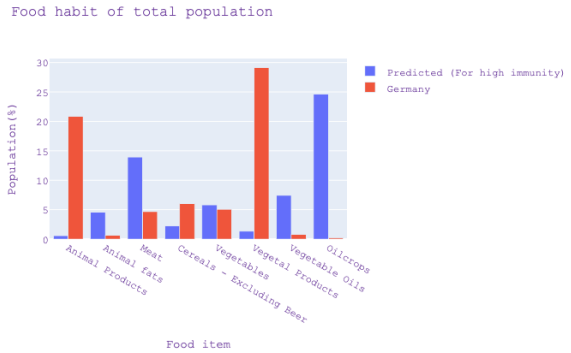
Protein requirement prediction for Germany



Fat requirement prediction for Germany



Kcal prediction for Germany



Amount prediction for Germany

Figure 5: The blue bar shows the ideal consumption of different kinds of food, and the red bar shows the current consumption pattern. To increase the immunity of Germany the red bars must be equalize with the blue bars for corresponding food item. If red bar for a food crosses the blue bar that means food manufactures needs to reduce the production of that kind of food. On the other hand if the red bar is below the blue bar than the manufacturer should increase that kind of food more to match the ideal condition

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

List of Tables

Methods	R^2 Score(Fat)	R^2 Score(Protein)	R^2 Score(Kcal)	R^2 Score(Amount)
Linear Regression	-358.03	-501.27	-351.97	-351.97
Ridge	-0.08	-0.08	0.07	0.07
Lasso	-0.03	-0.10	0.07	0.07
SVR	-0.49	-10.79	0.08	0.08
Decision Tree	-0.01	-0.10	-0.03	-0.03
Random Forest	-0.06	-0.98	0.08	0.08
Neural Network	-0.13	-0.19	-0.30	-0.23

Table 1: R^2 accuracy after 5-fold cross validation for two nutrients consumption, calorie consumption and amount.

Methods	R^2 Score(Fat)	R^2 Score(Protein)	R^2 Score(Kcal)	R^2 Score(Amount)
Linear Regression	-258.79	-81.34	-450.49	-1821.37
Ridge	0.09	0.19	0.15	0.07
Lasso	0.08	0.17	0.16	0.09
SVR	0.15	0.16	0.17	0.06
Decision Tree	0.06	0.15	-0.06	-0.11
Random Forest	0.19	0.23	0.27	0.06
Neural Network	-0.64	-0.69	-0.81	-0.68

Table 2: Prediction on eight output for types of data.

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

1 Introduction

Does the food habit of people affect their immune system? Childs and et al. consider sufficient, and suitable nutrition is important for cells to perform optimally and recover from damages [3]. As we all know that the effect of COVID- 19 is not the same for the countries around the world. For instance, some countries have a high death rate compared to that of others. There are many reasons for this difference. One reason could be the food habit. To analyze the effect of food habits on COVID- 19 recoveries and deaths, we study at what is the percentage of people consuming different kinds of food to get nutrients and calories in different countries and their COVID statistics. We found that there is a correlation between food habit and COVID -19 death and recovery rate.

According to Bousquet et al., dirty has a significant effect on COVID death, among other factors. According to them, some food is antioxidants or may reduce angiotensin-converting enzyme activity, which might be the reason for the low death rate in some countries (e.g. Austria, Baltic States, Czech Republic, Finland, Norway, Poland, and Slovakia) [4]

AbdelMassih and et al. concluded that the average BMI and COVID- 19 death rate has a positive correlation and the correlation factor is 0.43 [5]

Butler and et al. say that the consumption of saturated fats, sugars, and refined carbohydrates (collectively called Western diet, WD) worldwide has effects on obesity and type- 2 diabetes, and as a result, this type of diet could increase the severe COVID-19 pathology and mortality [6].

Though there are many kinds of research on the relation of food habit and fatality of COVID-19 cases, in our knowledge there is no established research on predicting a good diet to increase COVID- 19 recovery for any country.

In our research, we looked at the correlation between food habits and COVID-19 death rate, and also propose a recommendation service for food manufacturer to produce heathy food.

2 Data Set

In our research, we worked on COVID- 19 Healthy Diet Dataset [7] which contains different types of food, world population obesity, and undernourished rate, and global COVID-19 cases count.

There are mainly four parts of the dataset. Each of them represents the percentage of the average consumption of fat quantity, energy intake (kcal), food supply quantity/amount (kg), and protein for different categories of food respectively for 170 nations. Food categories are divided into twenty-three groups (e. g. Meat, Animal fat, Vegetable oil, Vegetable, etc.) Moreover, the dataset also includes up to date confirmed, deaths, recovered, and active COVID- 19 cases and undernourished and obesity percentages of each country.

Before analyzing data we preprocess the data in three steps. The undernourished column has less than expression in some rows, so we consider the ceiling value for that row. There are some missing cells in the dataset, we fill them up by the most frequent value of that type. Also, we scaled the food category data using the standard scaler [8] approach. Standard scaler means standardize features by removing the mean and scaling to unit vari-

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

ance.

We divide the data into 90%:10% training and test data. We used five-fold cross-validation during the parameter optimization of hyper parameters.

3 Procedure and Analysis

Each of the four parts of the data is independent of each other. So we decided to visualize and model them independently.

3.1 Data Visualization

First of all, we visualize the data in many ways. For example, we depict the correlation between all individual food categories, death due to COVID- 19, COVID- 19 recovery rate, undernourishment, and obesity.

In figure 1 (a to e), we see that there is a positive correlation between COVID- 19 death and Animal product consumption in all four parts of data. Similarly, in figure 2 (a to e), we see that there is also a positive correlation between animal fat intake and obesity.

On the other hand, in figure 3 (a to b) pulses and vegetal products have a negative correlation with COVID- 19 death.

This correlation says that there might be a dependency between COVID- 19 death and food habits.

3.2 Model Selection

Before modeling the data for predicting food patterns to boost the immune system we decided to consider the twenty-four food categories as output and COVID- 19 death, recovery, infection and active case, undernourished, and obesity as input.

After deciding the input and output variable, we train four models for four types of data using six regression techniques. We use 5-fold cross-validation to choose the best hyper-parameters and 10% of the data for testing. We trained 7 types of regression algorithms, named: Linear Regression, LASSO, Ridge Regression, Support Vector Regression (SVR), Decision Tree Method, Random Forest and Neural Network. Table 1 represents the performance of all the techniques in the four data sets. If we closely look at R^2 values, we see that for most of the datasets there are very few models that have scores in [0-1] range. That means all model guessing randomly. To overcome the situation, we decided to find out why the model has very bad accuracy. Our educated guess is some of the output variables have outliers and they also don't represent the data distribution properly. As a result the test error increase.

3.3 Tuning the Model

In the next step, we find out which food categories have outliers. Including outliers in training might bias the prediction and decrease test accuracy. We plotted four box plots (figure 4) for the four categories to see which type of food has more outliers. We ignore the feature that has outliers.

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

We fix eight outputs for modeling the system. The outputs are Animal Products, Animal fats, Meat, Cereals-Excluding Beer, Vegetables, Vegetal Products, Vegetable Oils, and Oil crops.

After that, we train the model on fixed output and we found better performance. Table 2 represents the performance of six models on four types of dataset on 8 output variables. In this table, we see most of the regression models perform better than previous cases. Finally, Random forest regression performs better among all regression approaches.

3.4 Prediction

After finalizing the model we interpolate the ideal food habit from the trained model based on a specific input. The specific input has 0% death rate, 100% recovery rate, 0% undernourishment, 0% obesity, and the current active case, an infection rate of a country for which we want to recommend the food.

4 Results and Discussion

After modeling the system we deployed it in a server so that user interactively select a country and compare its' food habit with the predicted ideal food habit that can boost immunity.

In figure 5 we see two four plots for four kinds of food types. Blue bar defines the ideal food consumption of a country to boost immunity. The red bar shows the current food consumption pattern. To boost the immunity one country must equalize their food pattern with the ideal food pattern of that country.

In this research, we could not achieve R^2 close to one. We also tried neural networks to increase performance. However, it fails because of the very small amount of training data. We only have 170 data samples. We also see that some of the output variables do not represent the inherent data distribution and as a result, overall test accuracy decreases. We fix this issue by specifying the output variables. If there is more data we there is a possibility of high accuracy. For example, if the data set contains state wise food consumption statistics the model could have more flexibility to learn the data distribution.

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Bibliography

1. "Does Your Diet Affect Your Immune System?" *WebMD*, WebMD, www.webmd.com/cold-and-flu/qa/does-your-diet-affect-your-immune-system.
2. "Immune System." *Wikipedia*, Wikimedia Foundation, 11 July 2020, en.wikipedia.org/wiki/Immune_system.
3. Childs, Caroline E, et al. *Diet and Immune Function*. 16 Aug. 2019, www.ncbi.nlm.nih.gov/pmc/articles/PMC6723551/.
4. J. Bousquet, C. Akdis, et al. "Is Diet Partly Responsible for Differences in COVID-19 Death Rates between and within Countries?" *Clinical and Translational Allergy*, BioMed Central, 1 Jan. 1970, doi.org/10.1186/s13601-020-00323-0.
5. AbdelMassih, Antoine, et al. "Obese Communities among the Best Predictors of COVID-19-Related Deaths." *Cardiovascular Endocrinology & Metabolism*, U.S. National Library of Medicine, 22 June 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7314342/.
6. Butler, Michael J., and Ruth M. Barrientos. "The Impact of Nutrition on COVID-19 Susceptibility and Long-Term Consequences." *Brain, Behavior, and Immunity*, Academic Press, 18 Apr. 2020, www.sciencedirect.com/science/article/pii/S0889159120305377.
7. Ren, Maria. "COVID-19 Healthy Diet Dataset." *Kaggle*, 13 July 2020, www.kaggle.com/mariaren/covid19-healthy-diet-dataset.
8. Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Appendix

Server link: https://data-science-group3.herokuapp.com

Note: Might take some time to load because it is a free server.

Code Repository:

https://github.com/Mak-Ta-Reque/Nutrition-and-immune-systems

Running local server:

First follow the Installation_guide.pdf and then run the following command from root.

python dash-app/app.py

We can divide our implementation in two parts. The first part is data visualization and exploration and the second part modeling the system

Visualization and Exploration

The *Notebook* directory inside the root contains ***Jupyter notebook*** file for that task. There are two visualization files for visualizing data and plots.

Applying_PCA.ipynb is understanding important features

Regression.ipynb is for experimenting different types of regression model

Modeling the Service

The *dash-app* contains the files for training the model and presenting the data in web browser interactively. The main purpose of the *utils.py* file to process the data and train and test may machine learning models. This script contains *NutritionData,* and *NutritionModel* class and some other utility methods and objects.

The former class is used for pre-processing data and the later class is used for training, testing and prediction models.

The *method_names* contains six the sklearn machine learning objects (Linear regression, Support vector regression, Random forests etc). We interactively train all of this methods.

The *search_four_models* method is used for finding out best hyper parameters of six models on the four datasets and *store_four_modes stores* the best model (.pkl file) in *dash-app/data/models* directory

The *app.py* file is used to visualize the data in web browser (server based system). It calls helper methods to predict the output and produce figures.

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736

Declaration of Authorship

I affirm that I have produced the work independently, that I have not used any aids other than those specified and that I have clearly marked all literal or analogous reproductions as such.

Location, Date Saarbrücken, 07- 17- 20120

Kadir, Md Abdul

Abdul Kadir

Alam, Hasan Md Tusfiqur

Tusfiqur Alam

Islam, Md Jonybul

Jonybul Islam

Kosheleva, Elena

Elena Kosheleva

Chair in Information and Service Systems, Campus A5 4, 66123 Saarbrücken, Germany, Tel: +49 (0) 681 302 64736