**Government of Russian Federation**

**Federal State Autonomous Educational Institution of High Professional Education**

**«National Research University Higher School of Economics»**

National Research University
High School of Economics
Faculty of Psychology

# Syllabus for the course
# «Introduction to Data Science»
(Введение в науки о данных)

010400.62 «Applied Mathematics and Informatics», Bachelor of Science

Authors:
Leonid E. Zhukov, professor, lzhukov@hse.ru
Ilya A. Makarov, senior lecturer, iamakarov@hse.ru

Approved by:

Recommended by:

Moscow, 2015

# Introduction to Data Science

**Course Syllabus**

## I. Introduction: Subject and background

***Author, Lecturer:***

Leonid E. Zhukov, Department of Data Analysis and Artificial Intelligence, Professor

***Tutor:***

Ilya A. Makarov, Department of Data Analysis and Artificial Intelligence, Senior Lecturer

***Summary***

Introduction to Data Science (IDS) course is designed as a bachelor-level course anticipating further education at Master Science program “Data Science”. Data Science (DS) is a new, exponentially-growing field, which consists of a set of tools and techniques used to extract useful information from data. Data Science is an interdisciplinary, problem-solving oriented subject that learns to apply scientific techniques to practical problems. The course orients on practical classes and self-study during preparation of datasets and programming of data analysis tasks.

***Prerequisites***

Good mathematical background and programming skills sufficient enough to learn new languages and software are required. Basic knowledge of statistics, linear algebra would be additional plus.

The course has facultative status.

***Aims***

- To develop practical data analysis skills, which can be applied to practical problems.
- To develop fundamental knowledge of concepts underlying data science projects.
- To develop practical skills needed in modern analytics.
- To explain how math and information sciences can contribute to building better algorithms and software.
- To give a hands-on experience with real-world data analysis.
- To develop applied experience with data science software, programming, applications and processes.

## *Background and outline*

Introduction to Data Science (IDS) class is offered as a practical prelude to Data Science Master Science program. Unlike the master-level, offering a great overview of various DS areas and applications, the IDS class is more depth-oriented: a fewer problems and methods will be studied, but to a larger extent.

This course is aimed at providing our students with a solid DS training, which could boost their careers in one of TOP10 mostly required professions in the world. The course is based the most recent DS tools and developments, brought to the students from the author working experience as a director of DS research department in several IT companies.

While the choice of DS, its problems and projects already defines the novelty of this class, we are trying to do our best to provide our students with the most up-to-date learning experience:

- The lectures are taught online – convenient to attend and follow. Using the most current teaching software packages, the students can fully interact with the instructor and classmates, share desktops, share applications, record class videos, take online tests and quizzes.
- The students work with real-world data. Unlike more conservative science classes, we prepare our students to solve real-world problems by working on these problems in the class.
- Independent work is appreciated. The class includes several mini-projects, which each student has to design and implement on its own.
- Analytical skills should evolve during classes. Students will work with noisy data, imperfect practices, human errors, diverse equipment. We teach our students to take data as it is, and to make most efficient use of what's available.
- The following topics will be covered by this introductory course:
    - Data mining
    - Statistics
    - Machine learning
    - Information visualization
    - Network analysis
    - Natural language processing

- Algorithms
- Software engineering
- Databases
- Distributed systems
- Big data

This class topic is new to HSE and Russian universities in general – and this is precisely the void we are trying to fill. DS programs start gaining their momentum in leading universities abroad, which is another reason for HSE to cease the opportunity and to offer a competitive class in this field.

## *Teaching notes*

The lectures are offered online, with class material being rather complex and sometimes unusual. Therefore, full student engagement and interaction with the instructor becomes the key to the class' success. The lecture material is not to be uploaded for public usage.

To keep the students as engaged as possible, we use a combination of teaching tools and methodology:

- Good online teaching software. Constant interaction with the students.
- Class projects. While homeworks are meant to demonstrate the understanding of the current class material, we use small class projects to help students develop their practical DS skills. Students will also search web for proper datasets for some tasks.
- Well-timed interaction during classes and office hours should stand for development and improvement of students' practical skills.

## *Teaching outcomes*

The main outcome of this class is to train a student to do practical DS work. Career-wise, we expect our students to be able to develop into skilled DS researchers or software developers.

After completing the study of the discipline IDS the student should:

- Know basic notions and definitions in data analysis, machine learning.
- Know standard methods of data analysis and information retrieval
- Be able to formulate the problem of knowledge extraction as combinations of data filtration, analysis and exploration methods.
- Be able to translate a real-world problem into mathematical terms.
- Possess main definitions of subject field.
- Possess main software and development tools of data scientist.
- Learn to develop complex analytical reasoning.

After completing the study of the discipline IDS the student should have the following competences:

| **Competence** | **Code** | **Code (UC)** | **Descriptors (indicators of achievement of the result)** | **Educative forms and methods aimed at generation and development of the competence** |
|---|---|---|---|---|
| The ability to reflect developed methods of activity. | SC-1 | SC-M1 | The student is able to reflect developed mathematical methods to DS problems. | Lectures and classes |
| The ability to propose a model to invent and test methods and tools of professional activity | SC-2 | SC-M2 | The student is able to improve and develop research methods of clustering, classification and machine learning. | Classes, labs, home works. |
| Capability of development of new research methods, change of scientific and industrial profile of self-activities | SC-3 | SC-M3 | The student obtain necessary knowledge in DS, which is sufficient to develop new methods on other sciences | Home tasks, paper reviews |
| The ability to describe problems and situations of professional activity in terms of humanitarian, economic and social sciences to solve problems which occur across sciences, in allied professional fields. | PC-5 | IC-M5.3_5.4_5.6_2.4.1 | The student is able to describe real-world problems in terms of DS. | Lectures and tutorials, group discussions, paper reviews. |
| The ability to detect, transmit common goals in the professional | PC-8 | SPC-M3 | The student is able to identify information and mathematical aspects in social | Discussion of paper reviews; cross discipline lectures |

| Competence | Code | Code (UC) | Descriptors (indicators of achievement of the result) | Educative forms and methods aimed at generation and development of the competence |
|---|---|---|---|---|
| and social activities | | | researches; evaluate correctness of the used methods and their applicability in each current situation | |

## *Recommendations to the students*

This class is meant to be interesting, and it's meant to help you unveil a completely new area of human knowledge, supporting the basic course on Data Analysis and Data Mining. It gives the opportunity to learn analytical skills and tools instead of only leveling coding skills. To anyone thinking about taking this class I would suggest the following:

- Take it only if you are interested in learning something new
- Be prepared to work
- Be independent, and look for new, unusual solutions.
- Do not miss/skip classes and homework. First, homework grades will be responsible for the bulk of your class grade. Second, each class is dedicated to a different area, and you do not want to miss any of them.

## II. Schedule

| No | Topic | Total hours | In class hours | | Self-study |
|---|---|---|---|---|---|
| | | | Lectures | Labs | |
| 1 | Introduction to data science | 6 | 1 | 2 | 3 |
| 2 | Exploratory data analysis | 7 | 1 | 3 | 3 |
| 3 | Introduction to machine learning | 7 | 1 | 3 | 3 |
| 4 | Linear regression and regularization | 7 | 1 | 3 | 3 |
| 5 | Model selection and evaluation | 7 | 1 | 3 | 3 |
| 6 | Classification: kNN, decision trees | 7 | 1 | 3 | 3 |
| 7 | Classification: SVM | 7 | 1 | 3 | 3 |
| 8 | Ensemble methods: random forests | 7 | 1 | 3 | 3 |
| 9 | Intro to probability: Naïve Bayes and logistic regression | 7 | 1 | 3 | 3 |
| 10 | Feature engineering and selection | 7 | 1 | 3 | 3 |
| 11 | Clustering: k-means, hierarchical clustering | 5 | 1 | 1 | 3 |
| 12 | Dimensionality reduction: PCA and SVD | 7 | 1 | 3 | 3 |
| 13 | Text mining and information retrieval | 7 | 1 | 3 | 3 |
| 14 | Network Analysis | 7 | 1 | 3 | 3 |
| 15 | Recommender systems | 7 | 1 | 3 | 3 |
| 16 | Relational databases, SQL | 7 | 1 | 3 | 3 |
| 17 | Big data storage and retrieval: noSQL, GraphDB | 7 | 1 | 3 | 3 |
| 18 | Big data distributed computing: map-reduce, spark rdd | 6 | - | 3 | 3 |

| | | | | | |
|---|---|---|---|---|---|
| 19 | Advanced: neural networks and deep learning | 6 | - | 3 | 3 |
| 20 | Generalizing lecture | 4 | 1 | - | 3 |
| 21 | Presentations of final projects | 20 | - | - | 20 |
| | **Total** | 152 | 18 | 54 | 80 |

## III. Assessment

The assessment includes three components:

- Class homework/projects, assigned after each lecture
- Final project

The class grade is computed as 70% of homeworks/projects + 30% of the final project.

In addition to this, student attendance, originality of work and contributions to the class will be taken into account, especially for those with non-zero fractional grade part.

## IV. Reading

### *Recommended:*

1. James, G., Witten, D., Hastie, T., Tibshirani, R. An introduction to statistical learning with applications in R. Springer, 2013.
2. Han, J., Kamber, M., Pei, J. Data mining concepts and techniques. Morgan Kaufmann, 2011.
3. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning, 2nd edition. — Springer, 2009.
4. Murphy, K. Machine Learning: A Probabilistic Perspective. - MIT Press, 2012.

### *Supplementary:*

“Practical Data Science with R”. Nina Zumel, John Mount. Manning, 2014

“Data Science for business”, F. Provost, T Fawcett, 2013

## V. Topics for research work and class projects

- Building recommender system
- Constructing neural network for deep learning
- Statistical data analysis

- Implementation of decision tree model

The syllabus is prepared by Leonid E. Zhukov, Ilya A. Makarov.