Программа учебной дисциплины
«Введение в Data Science»

Утверждена
Академическим советом ОП
Протокол №_____ от _____._____.20_____

Разработчик	Косолапов Кирилл Вадимович, Департамент больших данных и информационного поиска
Число кредитов	4
Контактная работа (час.)	30
Самостоятельная работа (час.)	122
Курс, Образовательная программа	1 (М) курс, Стратегия развития бизнеса
Формат изучения дисциплины	Без использования онлайн курса

1. Цель, результаты освоения дисциплины и пререквизиты

Цели:

1. The purpose of the discipline of introduction to data science: teaching the basics of working with data.
2. The purpose of the discipline of introduction to statistics
3. The purpose of the discipline of introduction to digital signal processing
4. The purpose of the discipline of introduction to machine learning
5. The purpose of the discipline of introduction to MS Excel and MS Azure
6. The development of critical thinking.

Планируемые результаты обучения (ПРО):

1. Be able to calculate 1) mode, median, average 2) Dispersion, standard deviation
2. Be able to calculate correlation
3. Be able to calculate confidence intervals
4. Be able to test hypotheses using statistical criteria
5. Be able to train machine learning algorithms for the task of classification, clustering and regression
6. Be able to choose a metric for checking the quality of the algorithm

2. Содержание учебной дисциплины

Тема (раздел дисциплины)	Объем в часах	Планируемые результаты обучения (ПРО), подлежащие контролю	Формы контроля

	лк см onl/cp		
Introduction	2 0 15	• Be able to calculate 1) mode, median, average 2) Dispersion, standard deviation	•
Statistics. Distribution.	2 4 15	• Be able to calculate 1) mode, median, average 2) Dispersion, standard deviation	•
Statistics. Confidence intervals and hypothesis testing.	2 4 15	• Be able to calculate confidence intervals	•
Hypothesis testing	4 4 15	• Be able to test hypotheses using statistical criteria	•
Correlation and other data processing methods	2 4 15	• Be able to calculate correlation	•
machine learning basics	4 8 18	• Be able to train machine learning algorithms for the task of classification, clustering and regression • Be able to choose a metric for checking the quality of the algorithm	•
Data Basics	2 4 15	• Be able to train machine learning algorithms for the task of classification, clustering and regression • Be able to choose a metric for checking the quality of the algorithm	•
Часов по видам учебных занятий:	18 28 108		
Итого часов:	154		

Просьба уточнить количество часов контактной работы.В разделе "Общие сведения" указано 30, по таблице получается 46
Просьба уточнить количество часов самостоятельной работы.В разделе "Общие сведения" указано 122, по таблице получается 108.

Содержание разделов дисциплины:

1. **Introduction**
   1.1 Introducing the teacher and course 1.2 Motivating part about data analysis

and ML 1) Introduction to the course (what we will study, how to evaluate) 2) Summary of course items 3) Introductory information on data analysis, examples of use from the industry. 4) Demonstration of data analysis on a "not obvious" statistical example (you can take the example of "weight loss" and statistical significance 5) A few examples where the lack of competent analysis led to adverse consequences (you can tell the story of Bill Gates financing small schools and show the effect of regression to medium on the example of the “heads-tails” experiment)

2. **Statistics. Distribution.**
   2.1 Probability and distribution 2.2 Distribution parameters (mode, median, mean, excess, asymmetry, range, variance, standard deviation) Introductory information on the basics of statistics (distributions, histogram parameters (difference in median, mode and arithmetic mean), significance level, variance, confidence intervals. Examples of analysis of statistics on applied industry problems (distribution of viewers' income).
3. **Statistics. Confidence intervals and hypothesis testing.**
   3.1 The concept of confidence interval 3.2 Calculation of the confidence interval and examples 3.3 Testing hypotheses using a confidence interval 3.4 Calculation of 3 sigma with examples
4. **Hypothesis testing**
   4.1 Significance criteria 4.2 Significance Level 4.3 Student, Criterion, Chi Square, Man Whitney 4.4 Testing hypotheses using criteria 4.5 Calculation of significance level. Verification of the hypothesis for p significance level.
5. **Correlation and other data processing methods**
   5.1 Correlation, autocorrelation 5.2 Spectral region 5.3 Time Series Filters 5.4 Fractals, wavelets, convolution
6. **machine learning basics**
   6.1 What is ML and where is it used 6.2 The tasks of classification, regression, clustering, ranking and forecasting time series 6.3 Basic ML algorithms 6.4 Five Historical Paradigms of ML Development Introductory information about the basics of machine learning (what it is, the task of classification, regression, clustering), the main problems and metrics. Examples of the use of machine learning on applied industry problems (recommendation systems). 7.1 Data Types 7.2 Data preprocessing 7.3 Retraining and regularization 7.4 Quality metrics (completeness, accuracy, f1 measure, roc-auc, confusion matrix) 7.5 Which algorithms are better suited for which tasks
7. **Data Basics**
   8.1 What data can be trusted? 8.2 Basic methods of data manipulation. 8.3 Cognitive bias in data interpretation. We consider the basic techniques of data manipulation (see the books "Statistics and Seals" and "How to Lie Using Statistics") Interactive game: find the manipulation (examples from the books above) Interactive game: “deceive a friend” (task, conduct an experiment in a public opinion poll (you can choose any topic, for example, how many hours a day people watch TV) so as to mislead others) 4) Verification of the additional task (find a data set from the industry and either calculate the reliability of the

hypothesis or build a predictive algorithm). Earning extra points for a task 5) Summing up, rating

3. Оценивание

- **1**, Не блокирующее, Компьютерное тестирование
  Test
- **2**, Не блокирующее, Домашнее задание
  Conducting an experiment showing a random distribution. Conducting a survey of questions about the group (growth, USE score, etc. ..)
- **3**, Не блокирующее, Домашнее задание
  Calculation of distribution characteristics. Calculation of the characteristics of the distribution histogram, construction of the distribution histogram based on the data of task 1
- **4**, Не блокирующее, Домашнее задание
  Conducting an experiment of dependence of one quantity on another. Carrying out 20 measurements of the height of the bounce of the ball from the height of the fall (or from a survey in the group). (not rated)
- **5**, Не блокирующее, Домашнее задание
  Confidence interval calculation.
- **6**, Не блокирующее, Домашнее задание
  The calculation of the significance level. Verification of the hypothesis for p significance level.
- **7**, Не блокирующее, Домашнее задание
  A / B test. Hypothesis testing based on A / B test.
- **8**, Не блокирующее, Домашнее задание
  Calculation of statistics, confidence interval and significance criteria in Excel
- **9**, Не блокирующее, Домашнее задание
  Building a binary classifier. Building a classifier based on passenger data from the Titanic (can be replaced by an industry task).
- **10**, Не блокирующее, Домашнее задание
  The solution of the regression problem. Building a model for predicting the cost of a car (film, book ...).
- **11**, Не блокирующее, Домашнее задание
  The solution to the clustering problem. Clustering data based on the “Iris” task (segmentation of the client base) (not evaluated)
- **12**, Не блокирующее, Домашнее задание
  The solution to the problem of selecting the optimal classification / regression algorithm. Selection of the optimal algorithm for solving the classification / regression problem. (in the format of the competition, the winner will receive additional points)
- **13**, Не блокирующее, Контрольная работа
  Test based on the results of the topic. A test of 10 questions on the material passed.
- **14**, Не блокирующее, Индивидуальная исследовательская работа
  Task 14 Self-analysis of the data set from the Kaggle site, based on the

knowledge gained. The result is a presentation in the “history” format, outlining the work done and conclusions drawn from the data.

Формула округления: Стандартное арифметическое округление

Вид формулы оценивания: Линейная
Формула оценивания:

Total = 0.1 * T13 + 0.45 * Section 1(T1-T8) + 0.25 * Section 2 (T9-T12) + 0.2 * T14
where
• T13 - score for the final test
• Section 1 - arithmetic average for tasks 1–8
• Section 2 - arithmetic average for tasks 9–12
• T14 - score for task 14

4. Примеры оценочных средств

T14 - NON blocking

The analysis of the data set with only the calculation of descriptive statistics is performed and the conclusions are presented in the form of a presentation or essay - 5 points.
An analysis of the data set with only the calculation of descriptive statistics, confidence intervals or correlation is performed and the conclusions are presented in the form of a presentation or essay - 6-7 points.
An analysis of the data set with only the calculation of descriptive statistics, confidence intervals or correlation is performed, at least 1 hypothesis is tested using statistical criteria, and the conclusions are presented in the form of a presentation or essay - 7-8 points.
The data set was analyzed with only the calculation of descriptive statistics, confidence intervals, or correlation, at least 1 hypothesis using statistical criteria was verified, a predictive model based on the machine learning algorithm was constructed, and conclusions in the form of a presentation or essay were presented - 8-10 points.
The assessment may be affected by the correctness of the calculation, the depth of the withdrawal and the general understanding of the actions performed.

Tasks 1-13 are evaluated according to the rule:

Correctly performed calculation -10% of the assessment.
The fact of the assignment is 50% of the assessment.
The answer to additional questions of the teacher and a demonstration of understanding of the material is 40% of the assessment.

5. Ресурсы

5.1. Рекомендуемая основная литература

п/п	Наименование
1	Статистика и котики»: АСТ; Москва; 2018
2	Курс «Построение выводов по данным»
3	Дарелл Хафф. Как лгать при помощи статистики — М.: Альпина Паблишер, 2015.
4	4. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания данных. Петер Фалех.
5	Искусственный интеллект. Современный подход.Стюарт Рассел, Питер Норвиг
6	Математические основы машинного обучения и прогнозирования. Владимир Вьюгин.
7	The Elements of Statistical Learning. The Elements of Statistical Learning. 2003
8	INTRODUCTION TO MACHINE LEARNING. Nils J. Nilsson. 1998
9	Heart Logs: Event Data, Stream Processing, and Data Integration. Jay Kreps. 2014

5.2. Рекомендуемая дополнительная литература

Не требуется

5.3. Программное обеспечение

п/п	Наименование	Условия доступа/скачивания
1	Microsoft Windows 7 Professional RUS Microsoft Windows 8.1 Professional RUS Microsoft Windows 10	*Из внутренней сети университета (договор)*
2	Microsoft Office Professional Plus 2010	*Из внутренней сети университета (договор)*
3	MS Excel	лицензионное по

5.4. Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

п/п	Наименование	Условия доступа/скачивания
	Профессиональные базы данных, информационно-справочные системы	
1	Электронно-библиотечная система Юрайт	URL: https://biblio-online.ru/
2	MS Azure ML Studio	free saas
	Интернет-ресурсы (электронные образовательные ресурсы)	
1	Открытое образование	URL: https://openedu.ru/

5.5. Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

- ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);
- мультимедийный проектор с дистанционным управлением.

Учебные аудитории для семинарских и самостоятельных занятий по дисциплине оснащены ПЭВМ, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.

6. Особенности организации обучения для лиц с ограниченными возможностями здоровья и инвалидов

В случае необходимости, обучающимся из числа лиц с ограниченными возможностями здоровья (по заявлению обучающегося) а для инвалидов также в соответствии с индивидуальной программой реабилитации инвалида, могут предлагаться следующие варианты восприятия учебной информации с учетом их индивидуальных психофизических особенностей, в том числе с применением электронного обучения и дистанционных технологий:

6.1.1. *для лиц с нарушениями зрения:* в печатной форме увеличенным шрифтом; в форме электронного документа; в форме аудиофайла (перевод учебных материалов в аудиоформат); в печатной форме на языке Брайля; индивидуальные консультации с привлечением тифлосурдопереводчика; индивидуальные задания и консультации.

6.1.2. *для лиц с нарушениями слуха:* в печатной форме; в форме электронного документа; видеоматериалы с субтитрами; индивидуальные консультации с привлечением сурдопереводчика; индивидуальные задания и консультации.

6.1.3. *для лиц с нарушениями опорно-двигательного аппарата:* в печатной форме; в форме электронного документа; в форме аудиофайла; индивидуальные задания и консультации.