**MLCLass, 2 модуль**

**«Машинное обучение с помощью Python»**

Содержание курса  
 **Урок 1. Введение**  
- Введение в машинное обучение, необходимые навыки  
- Задачи классификации, регрессии и кластеризации  
- Краткий обзор Kaggle  
- Открытие соревнования Kaggle Inclass   
- Знакомство с набором данных по автострахованию этого соревнования  
- Статистические распределения, нормализация признаков, приведение к нормальному распределению   
- Деревья решений  
- Применение дерева решений Scikit-learn к набору данных iris и данным из контеста Kaggle Inclass  
- Настройка параметров дерева, кросс-валидация  
  
  
**Урок 2. Обзор средств**  
- Работа с векторами и матрицами в библиотеке NumPy  
- Обзор библиотеки для научных выичслений SciPy  
- Тетрадки Jupyter (IPython) для презентации материала, содержащего код  
- Визуализация данных с Matplotlib  
- Чтение и обработка данных с библиотекой Pandas  
- Решение задачи соревнования Kaggle "Titanic: Learning from Disaster" c помощью Pandas  
- Обзор библиотеки машинного обучения Scikit-learn  
  
**Урок 3. Обучение с учителем. Классификация**  
- Работа с признаками – отбор, преобразование, построение  
- Метрики качества алгоритмов машинного обучения (accuracy, precision, recall, F-score). ROC-кривая, AUC.  
- Метод максимального правдоподобия  
- Логистическая регрессия  
  
**Урок 4. Обучение с учителем. Ансамбли. Переобучение.**  
- Случайный лес (Random Forest)  
- Случайный лес на примере набора данных Titanic  
- Случайный лес на примере набора данных по автострахованию  
- Бустинг (boosting) и бэггинг (bagging)  
- Сравнение бустинга и бэггинга на наборах данных репозитория UCI  
- Стекинг. Демонстрация решения задачи категоризации продуктов Otto (вкратце)  
- Переобучение, кросс-валидация, регуляризация  
- Пример регуляризации для логистической регрессии  
  
**Урок 5. Обучение без учителя**  
- Обзор методов кластеризации, снижения размерности, поиска аномалий в данных  
- Кластеризации городов России по социально-экономическим показателям  
- Сингулярное разложение матрицы  
- Пример снижения размерности изображений, сжатие изображений  
- Снижение размерности как способ визуализации даных  
- Меры качества кластеризации  
  
**Урок 6. Продвинутые методы, API Scikit-learn**  
- Нейронные сети, библиотеки nolearn и Lasagne NN

- Библиотека XGBoost, сравнение с градиентным бустингом в Scikit-learn

- Пример голосования между алгоритмами для повышения качества классификации

- Смешивание (блендинг) алгоритмов на примере задачи Kaggle "Titanic: Learning from Disaster"

- Стекинг. Пример для Titanic  
- Разработка собственного класса Scikit-learn Estimator для задачи по автострахованию Kaggle Inclass. kNN с подобранной метрикой.