Summer Internship Program
"Data Science for Social Good: Education, Health & Interaction"
Location: Moscow, Russia

Program Structure (4 weeks)

Week 1: Fundamentals and Data Immersion

Goal: Intensive introduction to tools and libraries. Familiarization with datasets in selected domains. Project selection.

Key topics:
- ✓ Python for data analysis (Pandas, NumPy)
- ✓ Data visualization (Seaborn, Plotly)
- ✓ Introduction to machine learning (Scikit-learn)
- ✓ Data types: educational datasets, medical indicators, motion sensor data, textual reviews

Project activities: Brainstorming and team formation. Defining goals, metrics, and work plans for each project.

Week 2: Specialization and Data Work

Goal: Deep dive into project-specific methods.

Key topics:
- ✓ Education/Healthcare: Data preprocessing, feature engineering, building and validating classification and regression models.
- ✓ Gesture Recognition: Introduction to computer vision and time series (OpenCV).
- ✓ Sentiment Analysis: Introduction to natural language processing (NLTK, SpaCy).

Project activities: Active exploratory data analysis (EDA), dataset collection (if needed), and creation of baseline models.

Week 3: Model Development and Improvement

Goal: Bringing models to working prototype stage.

Key topics:
- ✓ Advanced ML methods (ensembles, gradient boosting — XGBoost, CatBoost)
- ✓ Gestures: Simple convolutional neural networks (CNN) or recurrent networks (RNN) for sequences
- ✓ Sentiment: Use of pre-trained models (BERT, Word2Vec) for boosting accuracy
- ✓ Basics of MLOps: building a simple web interface with Streamlit

Project activities: Experiments with model architectures, hyperparameter tuning, MVP creation.

Week 4: Visualization, Interpretation, and Presentation

Goal: Building an intuitive interface and preparing for final defense.

Key topics:
- ✓ Storytelling with data
- ✓ Creating interactive dashboards (Streamlit, Plotly Dash)

- ✓ Ethics in Data Science (especially for medicine and education)
- ✓ Preparation for final presentations

Project activities: Final prototype and dashboard refinement, defense rehearsal.

Project Topics

Students will develop applications solving real-world, socially significant problems using modern Data Science tools. Each project aims to produce a working prototype.

Domain 1: Education (EdTech)

Project #1: "Educational Content Recommendation System"
Task: Develop an algorithm that recommends new courses or learning materials based on a student's history and grades.
Tools: Collaborative filtering, content-based filtering.
Outcome: Interactive prototype providing personalized course suggestions.

Project #2: "Student Performance Prediction and Risk Detection"
Task: Build a classification model predicting whether a student is “at risk” of underperformance based on attendance, LMS activity, and prior grades.
Tools: Classification (Logistic Regression, Random Forest), Feature Importance.
Outcome: Teacher dashboard displaying current student status and “alert” signals.

Domain 2: Healthcare

Project #3: "Predicting Disease Risk (based on open medical data)"
Task: Build a binary classification model that estimates the risk of a condition (e.g., cardiovascular disease) using anonymized health metrics such as age, blood pressure, and cholesterol level.
Tools: Classification (XGBoost), feature importance analysis for interpretability.
Outcome: Prototype system that calculates risk and highlights key contributing factors.

Domain 3: Gesture Recognition (Computer Vision)

Project #4: "Hand Gesture Recognition System for Interface Control"
Task: Train a computer vision model to recognize simple static hand gestures (e.g., "OK", "thumbs up", "thumbs down") from webcam input.
Tools: OpenCV for video capture and preprocessing; CNN (Keras/TensorFlow/PyTorch) for image classification.
Outcome: Real-time app recognizing gestures and displaying their names on screen.
Bonus: Bind gestures to actions (e.g., "thumbs up" — move slide in a presentation).

Domain 4: Sentiment Analysis (NLP)

Project #5: "Sentiment Analysis of Course Reviews"
Task: Create an NLP model that classifies course reviews into “positive,” “negative,” or “neutral,” and identifies key student concerns.
Tools: TF-IDF, Word Embeddings, fine-tuning of pre-trained models (e.g., DistilBERT).
Outcome: Streamlit web app that analyzes review text and outputs sentiment and key terms.

Entry Requirements for Students

Education: 3rd–4th year undergraduate students in Computer Science, IT, Mathematics, or Engineering.
Programming: Confident basic level of Python required.
Mathematics: Basic understanding of linear algebra, statistics, and probability theory.

Field Visits (Examples)

IT company developing EdTech or MedTech solutions — observing real-world integration of Data Science products.

University or medical research lab — exploring computer vision or bioinformatics projects in practice.