MINISTRY OF EDUCATION AND SCIENCE OF REPUBLIC OF KAZAKHSTAN

Al-Farabi Kazakh National University

Artificial intelligence and Big Data

JOURNAL-BOOK

By Externship

STUDENT Irshad Ahmad Oruzgani

Year of Study 2

FACULTY Faculty of Information Technology

SPECIALITY Data Science

служит путевкой на практику, обязательно прилагается к отчету

I. AREA

Student 2 course,

Faculty Information Technology

Full name Irshad Ahmad Oruzgani

Date from 26.05.2025 - 14.06.2025

Dean of the Faculty _________Тұрар Олжас Нұрқонысұлы

Head of Externship: Imanbek Baglan

II. EXTERNSHIP COMPLETION CERTIFICATE

Student: Irshad Ahmad Oruzgani

Place of Enternship :Halyk Academy Lab

Started work: 26.05.2024

Finished work: 14.06.2024

V. THE DESCRIPTION OF STUDENT'S WORK

(by indicating the level of theoretical preparation, quality of work done, labor discipline and disadvantages, if applicable)

Студент Irshad Ahmad Oruzgani за время прохождения внешней практики (externship) проявил себя исключительно с положительной стороны. Он успешно применил теоретические знания в области машинного обучения на практике, анализируя данные по реабилитации пациентов после сердечно-сосудистых заболеваний. Им были использованы инструменты Python, pandas и scikit-learn, а также реализованы модели логистической регрессии, случайного леса и SVC. Студент проявил высокий уровень ответственности, инициативности и интереса к прикладным задачам в сфере здравоохранения.

VI. AN ASSESSMENT OF STUDENT BY THE REPORT AND CERTIFICATION DEPARTMENT ON THE INTERNSHIP

(A brief report is given indicating the advantages and disadvantages, on four-point grading system).

В период прохождения практики Irshad Ahmad Oruzgani продемонстрировал уверенные теоретические и практические знания, проявил себя как квалифицированный, ответственный и дисциплинированный студент. Все поставленные задачи выполнялись в срок, с высокой степенью заинтересованности и вниманием к деталям. Работа была выполнена в полном объёме, охватывая ключевые аспекты применения машинного обучения в медицине. Отчёт по практике содержит необходимую информацию и оформлен в соответствии с установленными требованиями.

III. The list of materials collected during an Externship

1. **Flower Framework Documentation**
   Beutel, D., Topal, T., Mathur, A., Qiu, X., Parcollet, T., Lane, N.D. (2020).
   Flower: A Friendly Federated Learning Framework.
   https://flower.ai/docs/
2. **TensorFlow Federated (TFF) Documentation**
   TensorFlow Federated: Machine Learning on Decentralized Data
   https://www.tensorflow.org/federated
3. **Differential Privacy in TensorFlow**
   Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., & Zhang, L. (2016).
   Deep Learning with Differential Privacy
   https://github.com/tensorflow/privacy
4. **Diabetes Hospital Readmission Dataset**
   Strack, B., DeShazo, J.P., Gennings, C., et al.
   Impact of HbA1c Measurement on Hospital Readmission for Patients with Diabetes
   Available via UCI Machine Learning Repository:
   https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008
5. **Federated Learning Explained | Google AI Blog**
   McMahan, B. et al.
   Communication-Efficient Learning of Deep Networks from Decentralized Data
   https://ai.googleblog.com/2017/04/federated-learning-collaborative.html
6. **Model Evaluation Metrics (Precision, Recall, F1-score)**
   Ken Jee – *Evaluation Metrics Explained Simply*
   https://www.youtube.com/watch?v=85dtiMz9tSo

IV. Conclusion report of student by the results of internship and proposals for improvement

During the externship in Halyk Academy Lab, I worked on predicting heart disease rehabilitation outcomes using machine learning. I applied Logistic Regression, Random Forest, and SVC models, gaining hands-on experience in data preprocessing, model tuning, and evaluation. This practical work strengthened my Python skills and deepened my understanding of applying AI to real-world healthcare problems.

II. RECORDS OF WORKS PERFORMED ON INTERNSHIP

№	date	Summary of the work performed	signatures
	26.05.2025	Started the project by researching open-source healthcare datasets. Chose the Diabetes 130-US Hospitals dataset from Kaggle due to its hospital-wise distribution, feature richness, and permissive license. Downloaded it using Kaggle API.	
	27.05.2025	Performed data cleaning and preprocessing. Dropped irrelevant and missing-value-heavy columns, handled categorical variables using LabelEncoder, and saved the cleaned dataset for future use.	
	28.05.2025	Simulated data silos by splitting the dataset into 5 non-IID subsets based on patient hospital. Each silo was balanced to ~15,000 records.	
	29.05.2025	Wrote an automation script (preprocess_silo.py) to apply uniform preprocessing (encoding, scaling, and null handling) to all silos. Saved processed versions in a new directory.	
	30.05.2025	Set up environments for 5 federated learning frameworks (Flower, TFF, PySyft, OpenFL, Substra) in Conda. Each was tested for compatibility with Python 3.10.	
	02.06.2025	Designed a synchronous client-server architecture using FedAvg. Defined aggregation logic, communication flow, and client participation. Created a visual diagram for clarity.	
	03.06.2025	Selected a shallow MLP model with two hidden layers for binary classification. Built it using TensorFlow and Keras. Tuned loss and optimizer settings for FL compatibility.	
	04.06.2025	Trained the MLP on centralized cleaned data to get a performance benchmark. Achieved 87.11% accuracy and 0.93 F1-score, with high recall (97.11%). Used these for comparison with FL results.	

	05.06.2025	Started federated training using Flower with FedProx strategy. Each of the five clients trained locally for one epoch per round. Accuracy reached 88.77% after the first round.	
	06.06.2025	Added client-side logging to monitor training metrics per round. Logs showed consistent accuracy (~88–89%) across clients, indicating balanced silos. Saved logs for each client.	
	09.06.2025	Implemented centralized evaluation after every FL round using a held-out validation set. Logged performance metrics like loss and accuracy to a JSON file for tracking trends.	
	10.06.2025	Enabled distributed validation: each client tested the aggregated model on local data post-round. Accuracy stayed stable across silos, confirming generalizability and fairness.	
	11.06.2025	Expanded evaluation metrics on both client and server sides. Logged precision, recall, specificity, and F1-score. Found perfect recall but very low specificity, showing a class imbalance issue.	
	12.06.2025	Added client-level Differential Privacy using custom DP-SGD implementation in TensorFlow. Achieved privacy without major performance loss (Accuracy: 88.12% with noise).	
	13.06.2025	Finalized model security by adding L2 norm clipping to client weight updates. This limited leakage risks and aligned with GDPR/medical data standards. Performance remained strong (88.67% accuracy).	