Data Science in Humanitarian Action

905F3 | Wider Topics in Data Science

UNIVERSITY
OF SUSSEX

Word Counts: 2,710

Candidate Number: [redacted]

Contents:

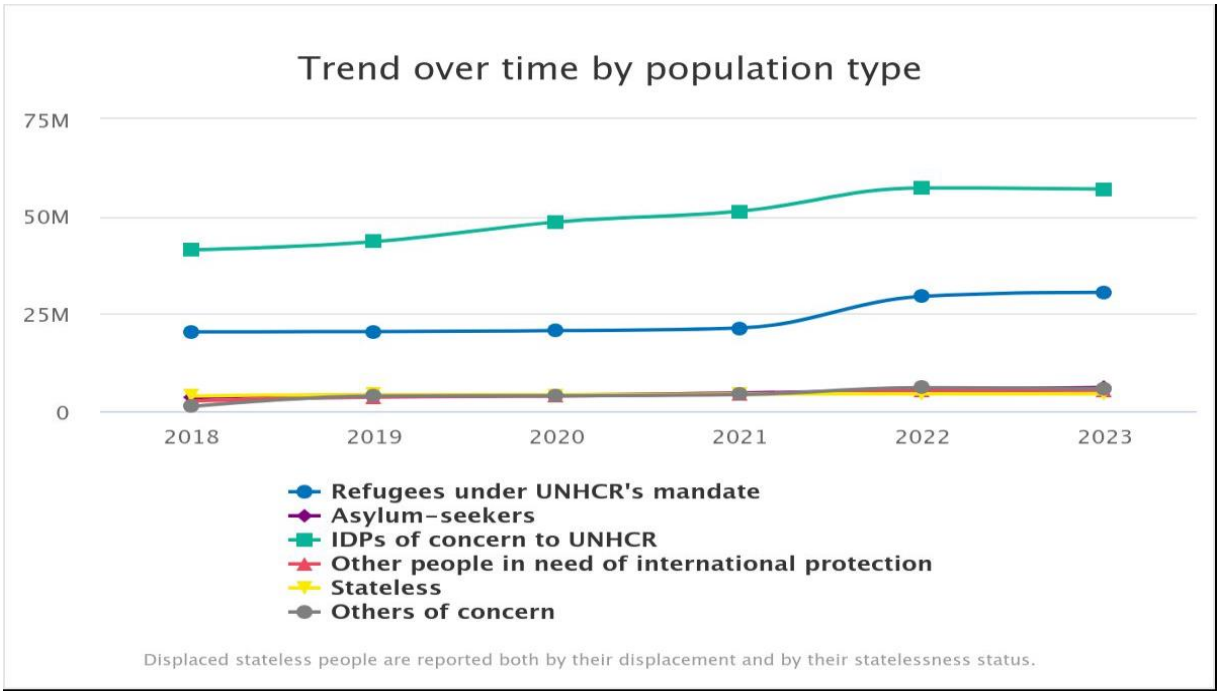
Abstract:

Global humanitarian crises are becoming more serious and complicated, encompassing diverse challenges like war, internal conflict, and climate change, leading to an increase in the number of refugees and immigrants. To solve these complex problems and make a better future, humanitarian actors are now using advanced data science methodologies. This involves predictive analysis to anticipate refugee patterns and flows, enabling proactive preparation and the formulation of effective strategies. This paper will describe in more detail why data science is essential in humanitarian action, especially for refugees and how the actors are utilising data science in real-world cases.

1. Introduction

Global humanitarian crises are constantly occurring, becoming increasingly intense and dangerous recently. We have faced not only ongoing conflicts like the Russian-Ukraine War and the Israel-Palestine conflict, but also several incidents such as the coup in Sudan, the COVID-19 pandemic, and the impacts of climate change. Humanitarian crises include a wide range of cases like armed conflicts, epidemics, famine, natural disasters, and other critical situations, all of which have the strong potential to result in or exacerbate a humanitarian disaster (Humanitarian Coalition, 2015).

Figure 1. Displacement Population Flow (UNHCR, 2018-2023)



Source: UNHCR. (2023). Trend over time by population type, https://www.unhcr.org/refugee-statistics/download/?url=sH5pnE

Many refugees have fled their homes in order to survive during these emergencies. Therefore, they require essential and fundamental assistance such as food, clean water, shelter, education, and medical care. The total number of refugees recognised by the Office of the High Commissioner for Refugees (UNHCR) has constantly increased since 2018 and it has reached approximately 57 million in 2023, up from 41 million.

The United Nations (UN) has made an appeal for $51.5 billion to address the 339 million people's requisitions who need urgent help from the world. Additionally, it is anticipated that the budget will be allocated for global humanitarian crises will exceed $100 billion by 2027 (Humanitarian Practice Network, 2023).

In these contexts, data science is considered an innovative method to assist humanitarian actors. The International Organization for Migrant (IOM) and UNHCR, regarded as representative actors,

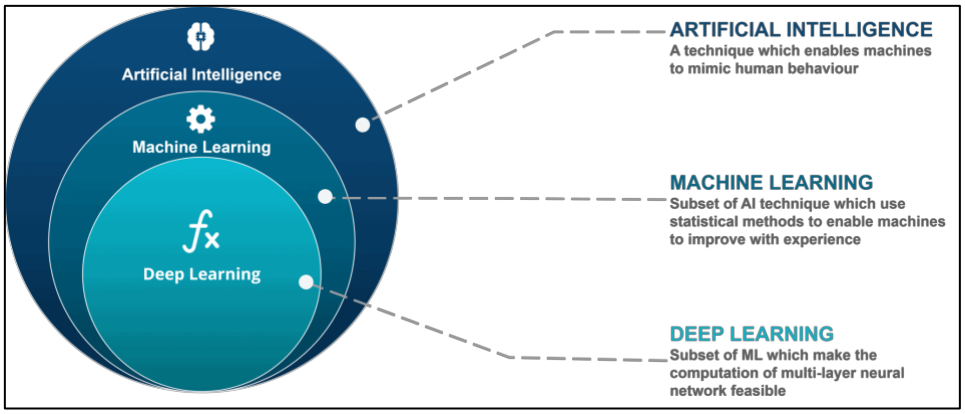
are now employing data science such as Machine Learning and Natural Language Processing (NLP) in their projects and policies. For example, through data science, humanitarian actors and related countries are predicting the flow of refugees. It helps them to develop appropriate strategies and approaches to cope with. Hence, this paper will introduce some cases demonstrating how Data Science is facilitating to solving of humanitarian crises, particularly those involving refugees.

2. Data Science Methodologies in Humanitarian Action

2-1. Machine Learning

Machine Learning (ML) is considered a subset of artificial intelligence that is broadly described as the capability of a machine to replicate intelligent human behaviour (Brown, 2021).

Figure 2. Structure of AI



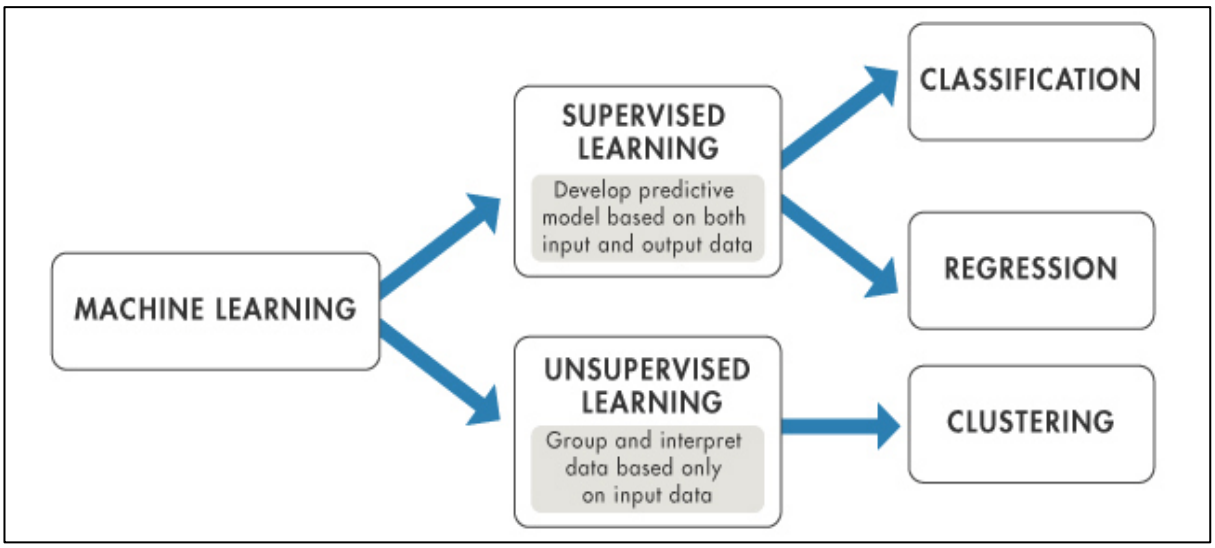
Source: Tondak, akshay (2020). Deep Learning Vs Machine Learning | Know The Difference, https://k21academy.com/datascience-blog/deep-learning/dl-vs-ml/

Machine Learning can learn from data and information without relying on predetermined equations or specific instructions (Tondak, 2020). With the recent accessibility of big data, the significance of machine learning has become increasingly emphasised. Within machine learning, two methodologies known as 'Supervised Learning' and 'Unsupervised Learning' are distinguished.

2-2. Supervised Learning

Supervised Learning relies on training data samples from a dataset where correct classifications are already defined, whereas unsupervised learning utilises training data samples without predefined labels (Sathya and Abraham, 2013). Both supervised and unsupervised learning are used in the humanitarian sector, but supervised learning is typically favoured for regression and classification tasks. This preference stems from the efforts of humanitarian actors to collect data with known labels.

Figure 3. What is Machine Learning?



Source: https://www.mathworks.com/help/stats/machine-learning-in-matlab.html

Regression is employed to measure and identify the relationship between a dependent variable and independent variables believed to spell out it (Ramcharan, 2019). In regression, a variety of techniques can be applied, including Linear Regression, Lasso Regression, Ridge Regression, and Decision Tree Regression, among others.

On the other hand, classification in terms of statistics involves identifying a category to which observations belong based on the features that the observations have. In classification, methods such as Logistic Regression, K-Nearest Neighbors (K-NN), Random Forest, and Decision Tree are often chosen to perform it.

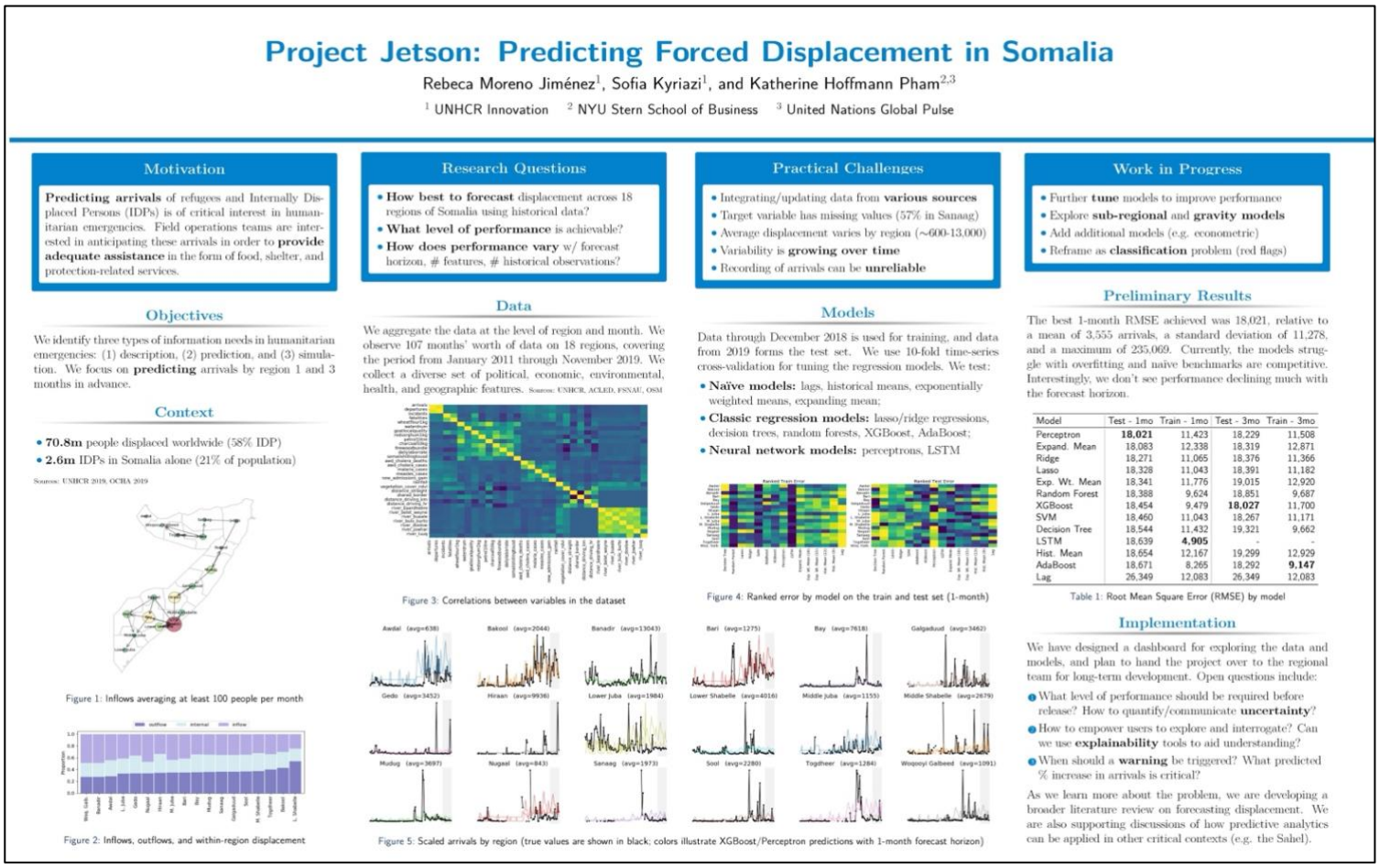
3. Case Studies

In the humanitarian sector, regression is frequently employed to predict the flow and patterns of refugees and migration. Meanwhile, classification is utilised for determining whether an individual is a refugee and to classify refugee laws, providing prompt and meaningful information. This paper aims to provide more detailed cases to enhance understanding of how data science is used in humanitarian action.

3.1. AI Project Jetson: Predicting Migration Patterns During the Somali Conflict

UNHCR's Innovation Service launched an experimental project called 'Project Jetson' in 2017 to explore the potential of data in predicting population movements in Sub-Saharan Africa, particularly in the Horn of Africa (Parater, 2019).

Figure 4. Summary of Project Jetson

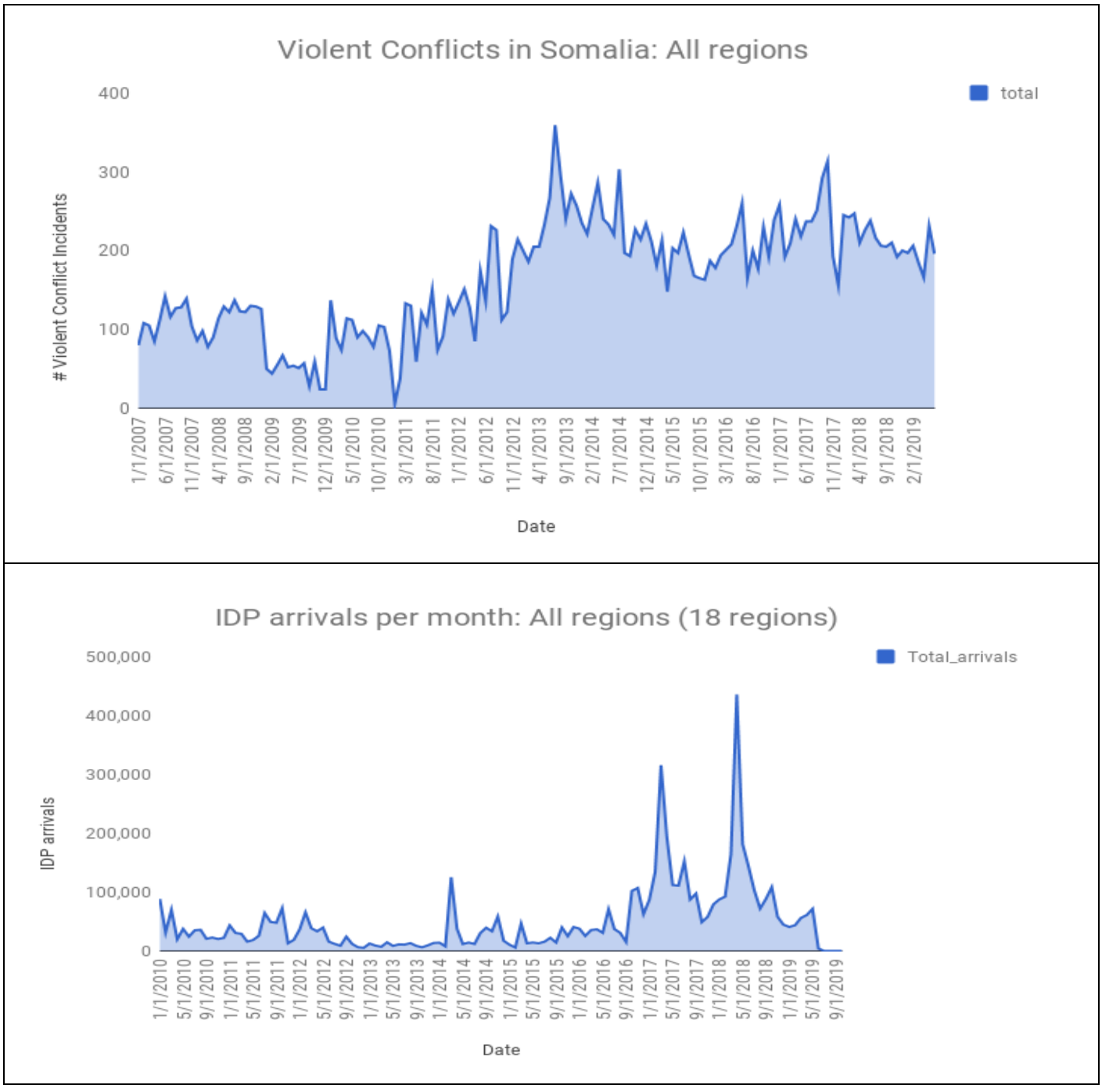


Source: https://github.com/unhcr/Jetson/wiki/4.-Preliminary-Results

Project Jetson is an engine designed to analyse data and generate an estimated number of the displacement of individuals in Somalia through a trained machine-learning model. This project was developed through a data science process. Using supervised learning, Project Jetson uses various factors, both quantitative and qualitative, which can forecast the total amount of internally displaced persons (IDPs) (UNHCR, 2022).

According to UNHCR (2020), they initially collected both quantitative and qualitative data from various datasets such as the UN and the Food and Agriculture Organisation of the United Nations (FAO) data and merged them for this project. Independent variables include violent conflict, climate and weather anomalies such as rain patterns and river levels, and market prices. Some independent variables were transformed into numerical or categorical variables for the machine learning process. On the other hand, 'Forced displacement' was assigned as the dependent variable of this model, focusing on calculating the number of arrivals per region and month (UNHCR, 2020).

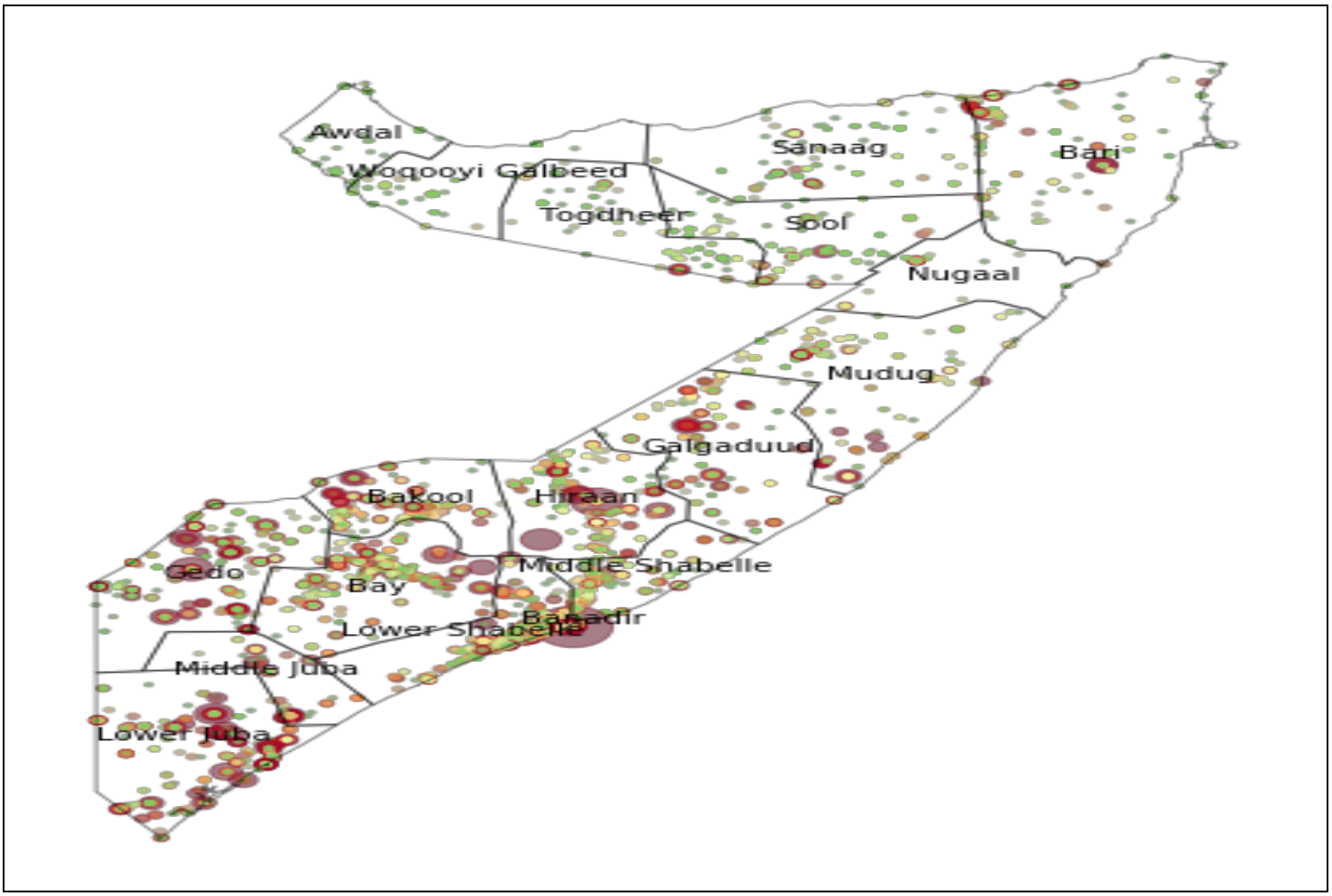
Figure 5. EDA of Somalia Dataset 1



Source: https://github.com/unhcr/Jetson/wiki/2.-Input-Data

Following this, UNHCR cleaned the dataset and conducted Exploratory Data Analysis (EDA) like Figure 5. They examined the total cases of violent incidents in all regions, the number of arrivals of internally displaced people in 18 regions of Somalia per month and so on. Additionally, they visualised and explored the geospatial perspective of these two factors. In Figure 6, the bubbles represent the frequency of violent conflict incidents (UNHCR, 2020).

Figure 6. EDA of Somalia Dataset 2



Source: https://github.com/unhcr/Jetson/wiki/2.-Input-Data

Consequently, UNHCR determined to conduct three experiments to build an AI model 'Project Jetson'. The three experiments are as follows, using R and Python for utilising machine learning, including open-source scripts: predicting forced displacement one month in advance (Experiment 1), extending the prediction horizon to one and three months (Experiment 2), and predicting forced displacement with three months in advanced and giving space for more months (Experiment 3) (UNHCR, 2020).

To build the AI model, UNHCR constructed various regression models, ridge and lasso regressions, multi-layer perceptron, AdaBoost, and XGBoost, as well as decision trees, random forests, and a Long Short-Term Memory (LSTM) neural network. The model was assessed through Mean Square Error (MSE), Root Mean Square Error (RMSE), and others.

Besides, the finalised input data for the model extended to August 2020, covering a total of 116 months of data over 18 regions in Somalia. (UNHCR, 2020).

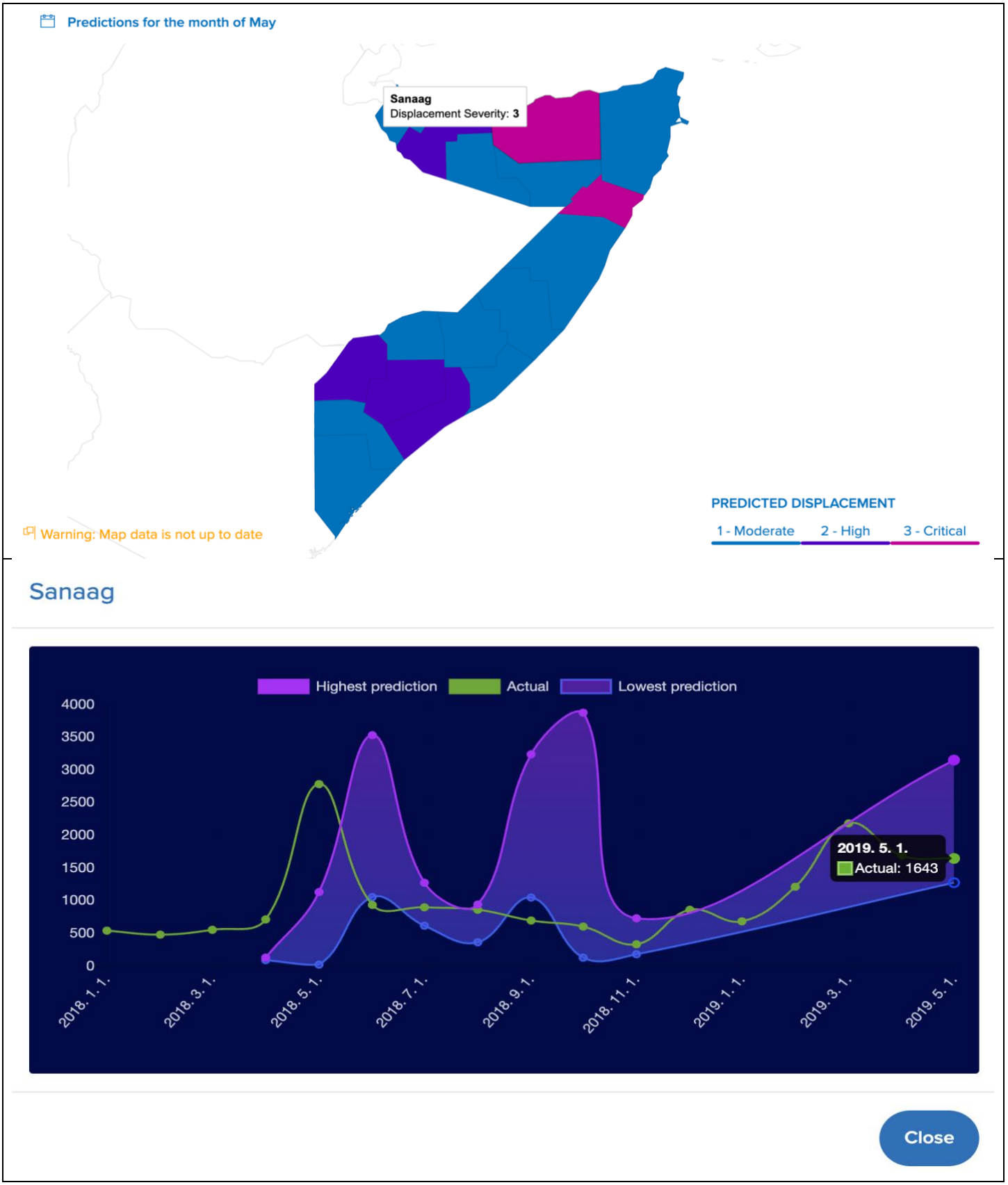
Table 1. Performance of Models

	1-month horizon		3-month horizon	
	Train	Test	Train	Test
Perceptron	6660	6288	6811	6669
Ridge Regression	6152	6389	6709	7015
Expand. Mean	7683	6437	7767	6512
Lasso Regression	6084	6449	6457	7267
Random Forest	5760	6471	5564	6754
Exp. Wt. Mean (23)	7303	6546	7600	6674
LSTM	1377	6555	1449	7195
Exp. Wt. Mean (8)	7195	6601	7794	6702
AdaBoost	5424	6631	4962	7361
Hist. Mean (12)	7428	6632	7739	6774
XGBoost	7239	6965	5849	6749
Decision Tree	6791	7853	6421	6936
12-month lag	9167	8487	9150	8537
1-month lag	8294	8594	-	-

Source: https://github.com/unhcr/Jetson/

As a result, UNHCR opted for the utilisation of machine learning and aimed to continue developing this model, despite the strong performance of naive benchmarks such as Exp. Wt. Mean and 12-month lag. The decision was based on the capability of machine learning to anticipate sudden deviations or significant changes from current trends and its ability to integrate diverse contextual factors to assess their significance (Hoffmann Pham and Luengo-Oroz, 2022). Hoffmann Pham and Luengo-Oroz (2022) stated that ML models can determine whether these factors are essential or just linked to displacement, providing valuable insights for further examination.

Figure 7. Visualisation of Project Jetson



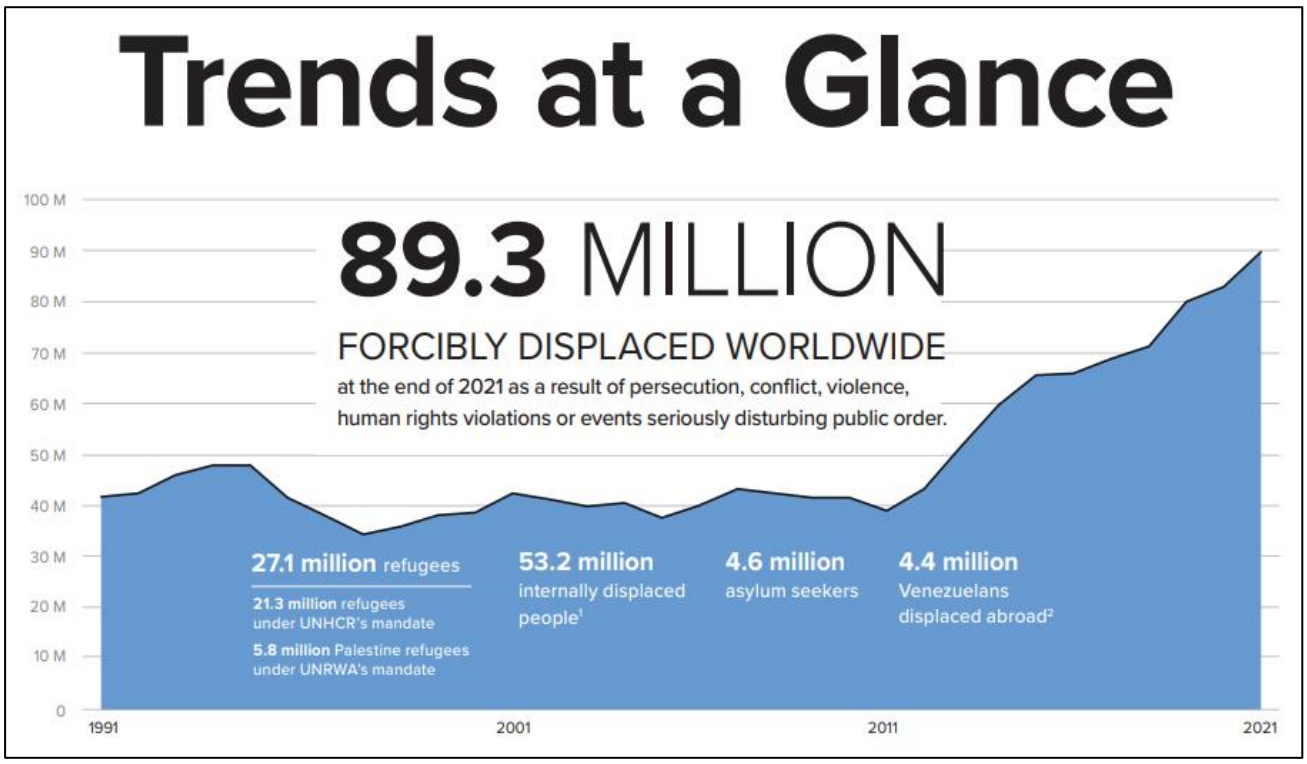
Source: https://jetson.unhcr.org/tech.html

Currently, UNHCR employs this model to forecast the number of IDPs in Somalia as shown in Figure 7, to aid in the procurement of essential assistance such as food, shelter, and protection-related services (UNHCR, 2020). Project Jetson is a pioneering effort to use data science in humanitarian action. UNHCR is actively involved in enhancing this model and exploring its application to other humanitarian crises.

3.2. Refugee Identification System (RIS) in Arizona

In 2021, the global tally of individuals forcibly displaced had reached 89.3 million. Among them, 27.1 million were classified as refugees, while millions of stateless individuals were also included. These stateless individuals have been denied nationality and access to fundamental rights, such as food, education, healthcare services, employment, and freedom of movement until they are recognised as refugees (United Nations, 2021).

Figure 8. Trend of Forcibly Displaced (Worldwide, 2021)



Source: https://www.unhcr.org/media/global-trends-report-2021 (UNHCR, 2021)

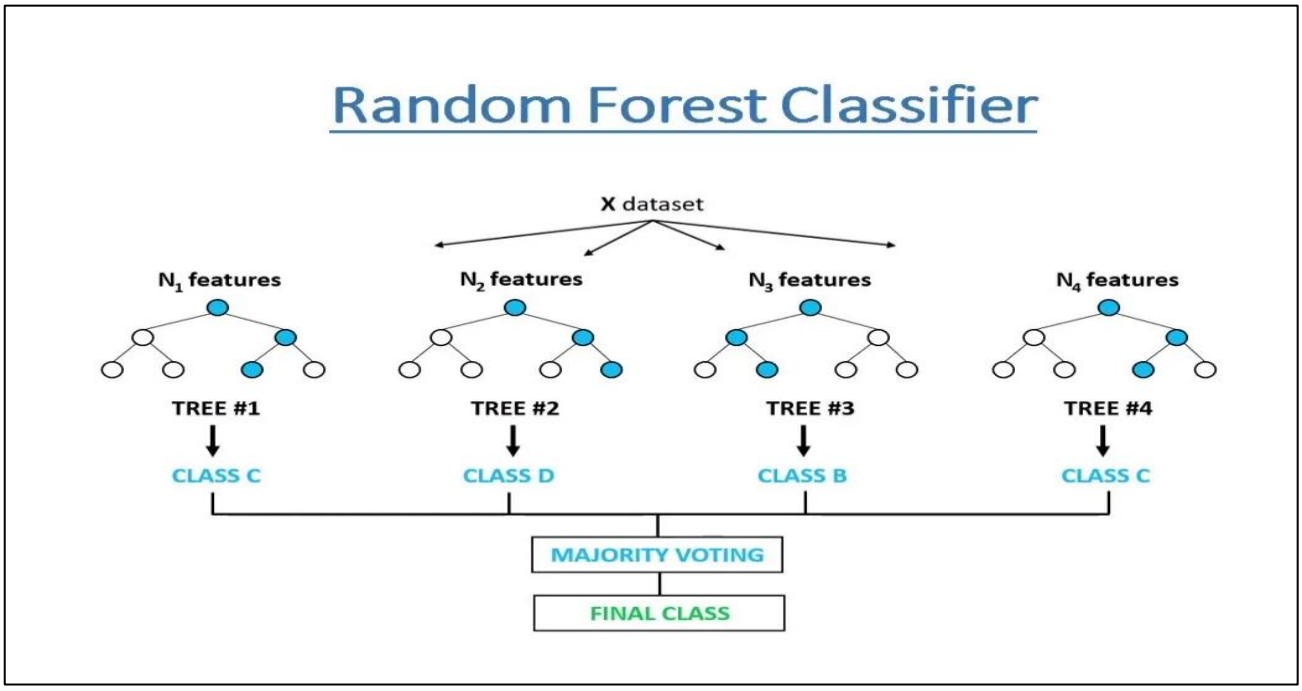
Once stateless individuals are identified as refugees, they can access essential services such as healthcare and education. However, the process of reviewing asylum and confirming refugee status typically takes several months to years. During this time, asylum seekers often struggle to secure fundamental services, especially healthcare.

In this situation, Arizona State University, Valleywise Health, and Creighton University School of Medicine collaborated with several other institutions to develop a Refugee Identification System (RIS) in Arizona. This initiative aims to reduce the overall duration of the refugee identification

process through the use of machine learning (Morrison et al., 2021).

A total of 1,033 people were randomly extracted from the VM Women's Health Center database, spanning the period between May 2020 and December 2020, to develop RIS. The dependent variable, refugee status, included 103 individuals classified as refugees and 930 classified as non-refugees (Morrison et al., 2021). To enhance the model's predictive capabilities, certain features underwent extraction and feature engineering, primarily due to their non-quantitative nature or the absence of exhaustive potential values (Morrison et al., 2021). In total, 26 input features were selected, including age, language rank, and distances to 24 different locations.

Figure 9. Random Forest Classifier



Source: https://medium.com/@mrmaster907/introduction-random-forest-classification-by-example-6983d95c7b91

A random forest is a machine-learning classification method comprised of an ensemble of tree-structured classifiers, each containing nodes. Within a decision tree, purity is measured as the homogeneity of data within a node, indicating how likely data points within the node belong to the same class. Each tree in the random forest contributes a unit vote towards determining the most prevalent class for a given input value and the final class is determined based on the result of the vote (Breiman, 2001).

The random forest classifier model was chosen to perform classification for RIS. In addition, the dataset was split into three different datasets: training, validation, and test data. To address the issue of data imbalance, the research team utilised the Synthetic Minority Oversampling Technique (SMOTE) solely on the training dataset (Morrison et al., 2021). SMOTE operates by generating virtual samples around minority class instances, thereby artificially creating data for the non-

dominant group to handle data imbalance (Chawla et al., 2002). Following this, a grid search was conducted to investigate an ideal combination of hyperparameters, as outlined in Table 2.

Table 2. Configuration of Grid Search for Optimal Hyperparameters

Hyperparameter	Min	Max	Interval
Max features	15	25	1
Min samples per leaf	10	50	5
Min samples to split	2	14	2
Number of estimators	40	200	10

Source: Morrison, M., Nobles, V., Johnson-Agbakwu, C. E., Bailey, C., & Liu, L. (2021). Classifying Refugee Status Using Common Features in EMR. 6.

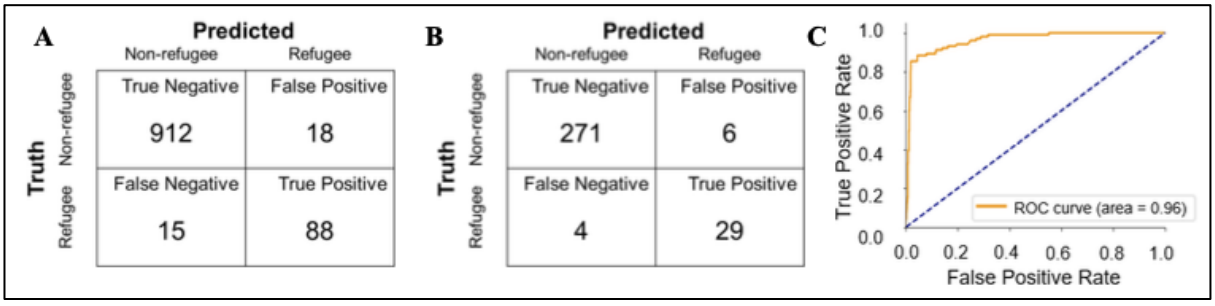
As a result, RIS, based on the random forest model, demonstrated excellent classification results with 97% accuracy under cross-validation. However, it has limitations in that the dataset used only adult female data. To address this limitation, the research team planned to expand their scope to include other major fields of healthcare, such as Pediatrics and Family Medicine.

Table 3. Model Performance Metrics

Metric	Cross-Validation	Holdout Testing
Accuracy	0.97	0.97
Specificity	0.98	0.98
Sensitivity (Recall)	0.85	0.88
Positive Predictive Value (Precision)	0.83	0.83
Negative Predictive Value	0.98	0.99

Source: Morrison, M., Nobles, V., Johnson-Agbakwu, C. E., Bailey, C., & Liu, L. (2021). Classifying Refugee Status Using Common Features in EMR. 6.

Figure 10. Model's Performance



(A) Confusion matrix (Cross-validations) in the train data

(B) Confusion matrix (Holdout test) in the test data **(C)** ROC curve

Source: Morrison, M., Nobles, V., Johnson-Agbakwu, C. E., Bailey, C., & Liu, L. (2021). Classifying Refugee Status Using Common Features in EMR. 6.

Overall, RIS is a machine learning model based on a random forest classification model to identify refugees. Despite not achieving 100% accuracy and being limited to adult female data, the RIS serves as a remarkable example of utilising machine learning for refugee identification. It highlights how data science can assist researchers, public health groups, and healthcare workers in swiftly and accurately identifying refugees using their dataset. This capability is vital for promptly identifying refugees in need of specialised care and support, thereby improving their chance to access healthcare, and ultimately enhancing their long-term health and overall welfare (Morrison et al., 2021).

3.3. Automatic Refugee Law Classification in Refworld

In 1990, UNHCR launched an offline 'The computerised database on Refugee Literature', which is Refworld's predecessor. This database contained 6,500 legal and policy documents about refugees. Later, Refworld transitioned to an online platform, encompassing approximately 76,000 documents related to incidents in countries of origin, documents of policy, and international and national legal frameworks (Refworld, 2024).

Figure 11. Refworld in 1995

Source: https://www.refworld.org/about-refworld/history

The primary objective of Refworld is to facilitate informed decision-making regarding the status of asylum-seekers and refugees. It provides data and information on national and international legal frameworks, policies, laws, and country-specific data to assist UNHCR staff members, government officials, and stakeholders in making determinations concerning refugee and statelessness status (Refworld, 2009).

Initially, navigating the website to find relevant cases and information was time-consuming due to the complexity of asylum decision-making and the counterintuitive layout of the website and functions. Attorneys had to repeatedly search the engine and review numerous documents (UNHCR, 2020). In response to this challenge, UNHCR implemented measures to enhance internal navigation and user intuition. They began summarising cases and tagging decisions with relevant metadata to streamline the process (UNHCR, 2020).

During this manual process, the question arose as to whether it would be feasible to mechanise the extraction of citations and provide the relevant web addresses. Consequently, UNHCR embarked on an initiative to reconstruct the website using artificial intelligence, particularly machine learning. They aimed to leverage AI to streamline the search process within the database, enabling efficient retrieval of entire documents essential to constructing a strong project on behalf of asylum-seekers who require international protection (UNHCR, 2020).

For this purpose, the UNHCR's Refworld team applied for funding from the UNHCR's Innovation Service Innovation Fund which supports humanitarian projects specifically aimed at leveraging artificial intelligence (UNHCR, 2020). Presently, the team is working on this project in collaboration with two external partners specialising in AI. Their focus is on determining the most effective approach to training a machine learning algorithm and how to recognise various legal decisions on Refworld.

As a result of the ongoing experiment with machine learning, UNHCR has achieved significant advancement in key evaluation metrics, particularly in terms of the model's capability to generate relevant and accurately classified results. In certain instances, the precision ranges from 85 to 90 percent (UNHCR, 2020).

According to UNHCR (2020), the aim of the renovation of the website and the AI integration is to streamline the process in various aspects. This entails faster and more precise searches, including the clustering of resembling decisions on a particular topic from different areas. These efforts are aimed at enhancing the effectiveness of the procedure and the quality of legal arguments. (UNHCR, 2020). Furthermore, this initiative will significantly reduce the overall time spent on decision-making regarding the status of asylum-seekers and refugees, ensuring that they receive timely support for their futures.

4. Conclusion

In conclusion, data science serves as an innovative method in numerous humanitarian endeavours, primarily through the application of machine learning. For instance, predictive analysis enables us to understand the migration patterns of refugees and internally displaced persons (IDPs). Additionally, it facilitates the classification of refugee statuses and stateless individuals, along with related cases and laws. Furthermore, not only refugee cases but also climate change responses such as drought detection, and natural disaster risk analysis are using data science as well. Nevertheless, the integration of data science into humanitarian efforts remains relatively limited compared to industries like finance and healthcare.

However, leading humanitarian organisations such as UNHCR and IOM are actively exploring the incorporation of data science into their operations. Notably, both UNHCR Headquarters and certain regional offices, including its offices in Copenhagen and Abidjan, have hired data scientists for various projects. Similarly, IOM has recruited data scientists at its Berlin office to conduct predictive data analysis and mathematical modelling. This trend shows that data science will be a fundamental methodology in humanitarian action in the near future.

Expectations are high for data science to play a significant role in addressing humanitarian crises. Through predictive analysis and modelling, humanitarian action can respond more swiftly and efficiently to crises while establishing a robust project foundation to enhance the success rates of its project. Thus, data science is a vital solution for improving humanitarian action effectiveness and guiding humanity toward a brighter future.

5. References

- Humanitarian Coalition. (2015). What Is a Humanitarian Emergency?, https://www.humanitariancoalition.ca/what-is-a-humanitarian-emergency
- Humanitarian Practice Network. (2023). Humanitarian action is the answer to fewer and fewer of today's humanitarian crises, https://odihpn.org/publication/humanitarian-action-is-the-answer-to-fewer-and-fewer-of-todays-humanitarian-crises/
- UNHCR. (2023). Trend over time by population type, https://www.unhcr.org/refugee-statistics/download/?url=sH5pnE
- Brown, S. (2021). Machine learning, explained, https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained
- Tondak, akshay. (2020). Deep Learning Vs Machine Learning | Know The Difference, https://k21academy.com/datascience-blog/deep-learning/dl-vs-ml/
- Sathya, R. and Abraham, A. (2013). Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification. International Journal of Advanced Research in Artificial Intelligence, 2(2), 34-35. https://doi.org/10.14569/ijarai.2013.020206
- Rodney Ramcharan. (2019). "Regressions: Why Are Economists Obsessed with Them?" Finance and Development (A quarterly magazine of the IMF), https://www.imf.org/external/pubs/ft/fandd/2006/03/basics.htm
- Parater, L. (2019). Jetson: insights into building a predictive analytics platform for displacement. UNHCR Innovation Service. https://medium.com/unhcr-innovation-service/jetson-insights-into-building-a-predictive-analytics-platform-for-displacement-186fb6ddca5b
- UNHCR. (2022). Jetson Stories. https://jetson.unhcr.org/story.html
- UNHCR. (2020). UNHCR Jetson. GitHub. https://github.com/unhcr/Jetson/wiki/2.-Input-Data
- Hoffmann Pham, K. and Luengo-Oroz, M. (2022). Predictive modelling of movements of refugees and internally displaced people: Towards a computational framework. Journal of Ethnic and Migration Studies, 18-27. https://doi.org/10.1080/1369183x.2022.2100546
- Morrison, M., Nobles, V., Johnson-Agbakwu, C. E., Bailey, C., & Liu, L. (2021). Classifying Refugee Status Using Common Features in EMR. medRxiv. https://doi.org/10.1002/cbdv.202200651
- United Nations. (2021). Refugees. United Nations. https://www.un.org/en/global-issues/refugees
- Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 321–357. https://doi.org/10.1613/jair.953
- Breiman, L. (2001) "Random Forests." Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324
- Refworld. (2024). History. https://www.refworld.org/about-refworld/history
- UNHCR. (2020). Giving Legal Teams Better Tools to Represent Asylum Seekers. UNHCR Innovation Service. https://medium.com/unhcr-innovation-service/giving-legal-teams-better-tools-to-represent-asylum-seekers-df7802e815df
- Refworld. (2009). UNHCR's Refworld - Case Law Collection User Guide. https://www.refworld.org/reference/tools/unhcr/2009/en/65317