

منتدى الدوحة للبيانات

من أجل الابتكار في التنمية المستدامة

22-23 أكتوبر 2024

DOHA DATA FORUM

FOR INNOVATION IN SUSTAINABLE DEVELOPMENT

October 22-23, 2024

Utilizing AI and Machine Learning for Advancing Official Statistics: TurkStat Experience

Doha Second Data Forum

22-23 October 2024

Bilal Kurban, Head of AI & Data Analysis Unit, Turkish Statistical Institute

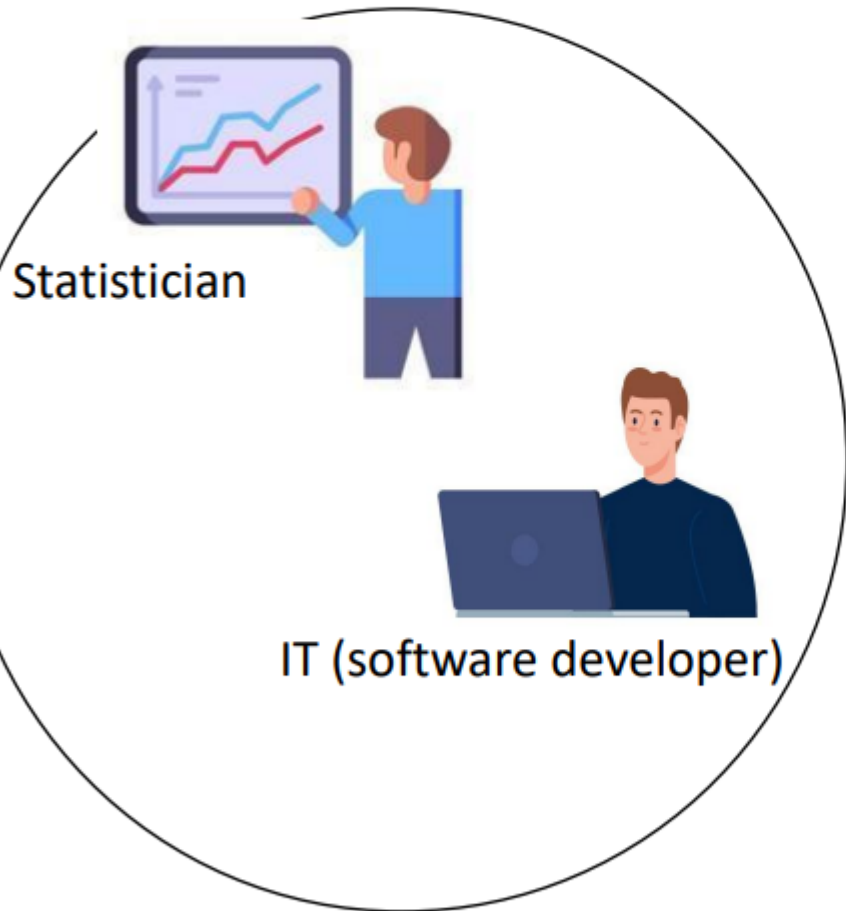
Proportion of failed AI projects

Gartner states that 85% of AI projects fail due to unclear goals and unclear R&D project management processes. Furthermore, 87% of R&D projects never make it to the production phase, while 70% of customers report little or no impact from AI.

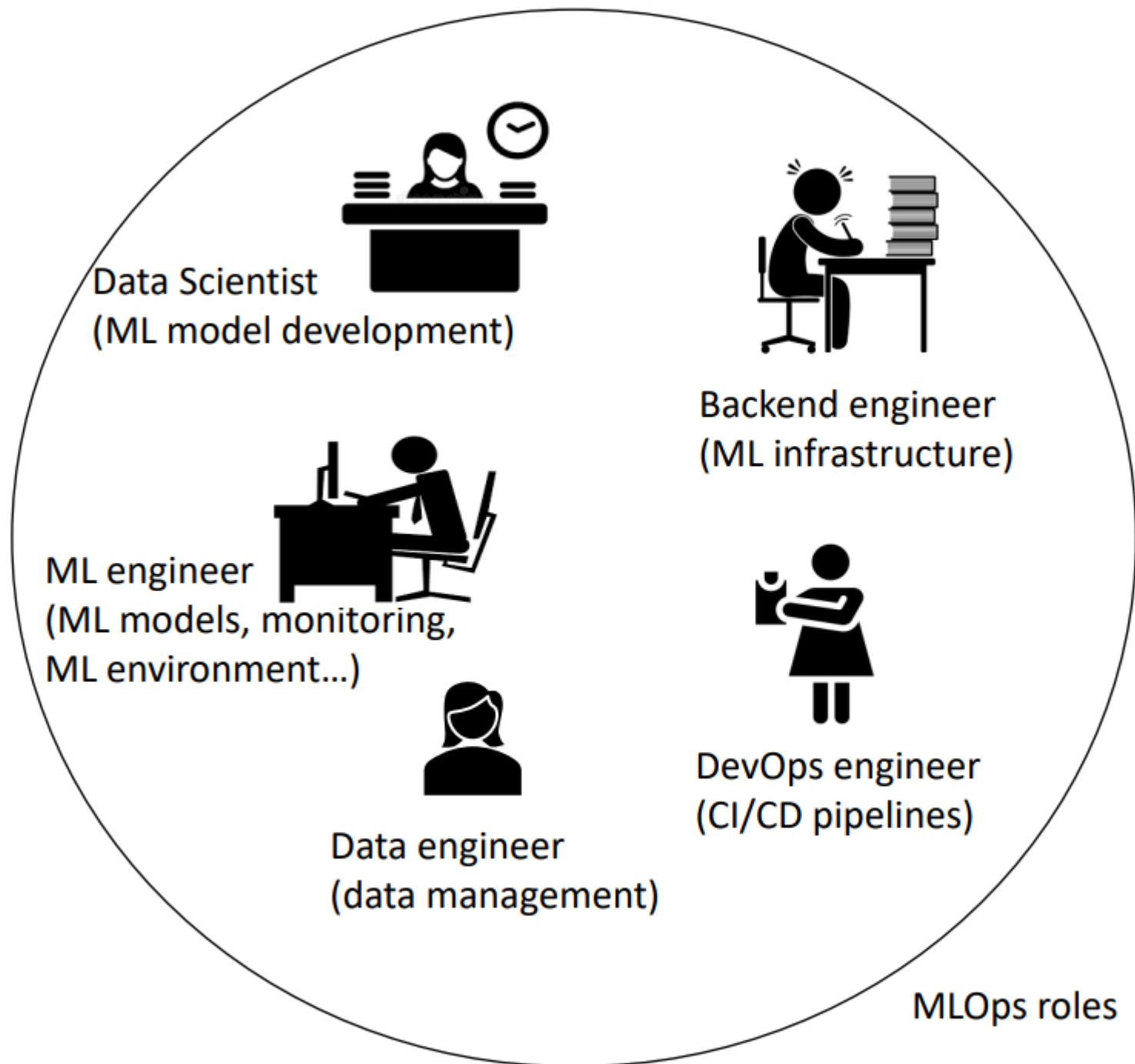
%85



Roles

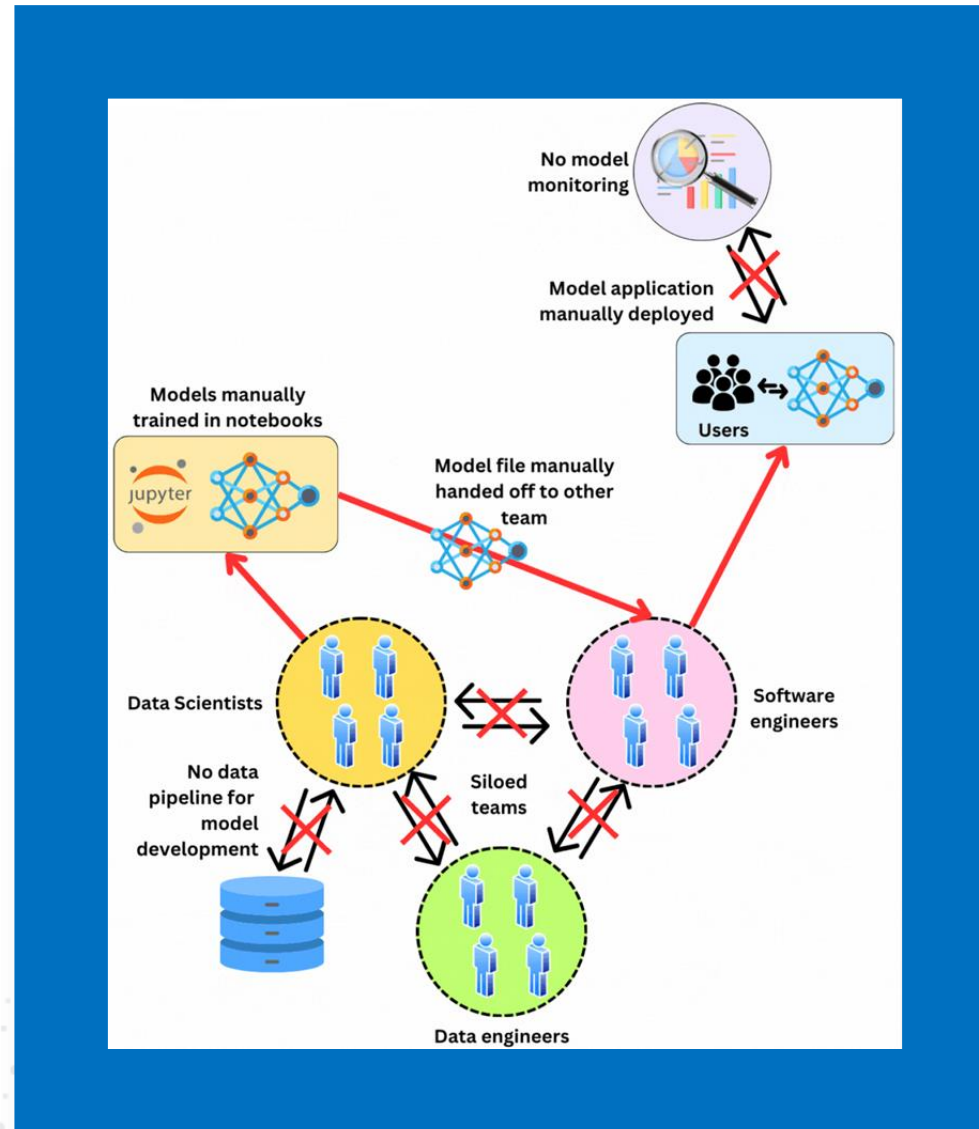


"As-is" state in many NSIs



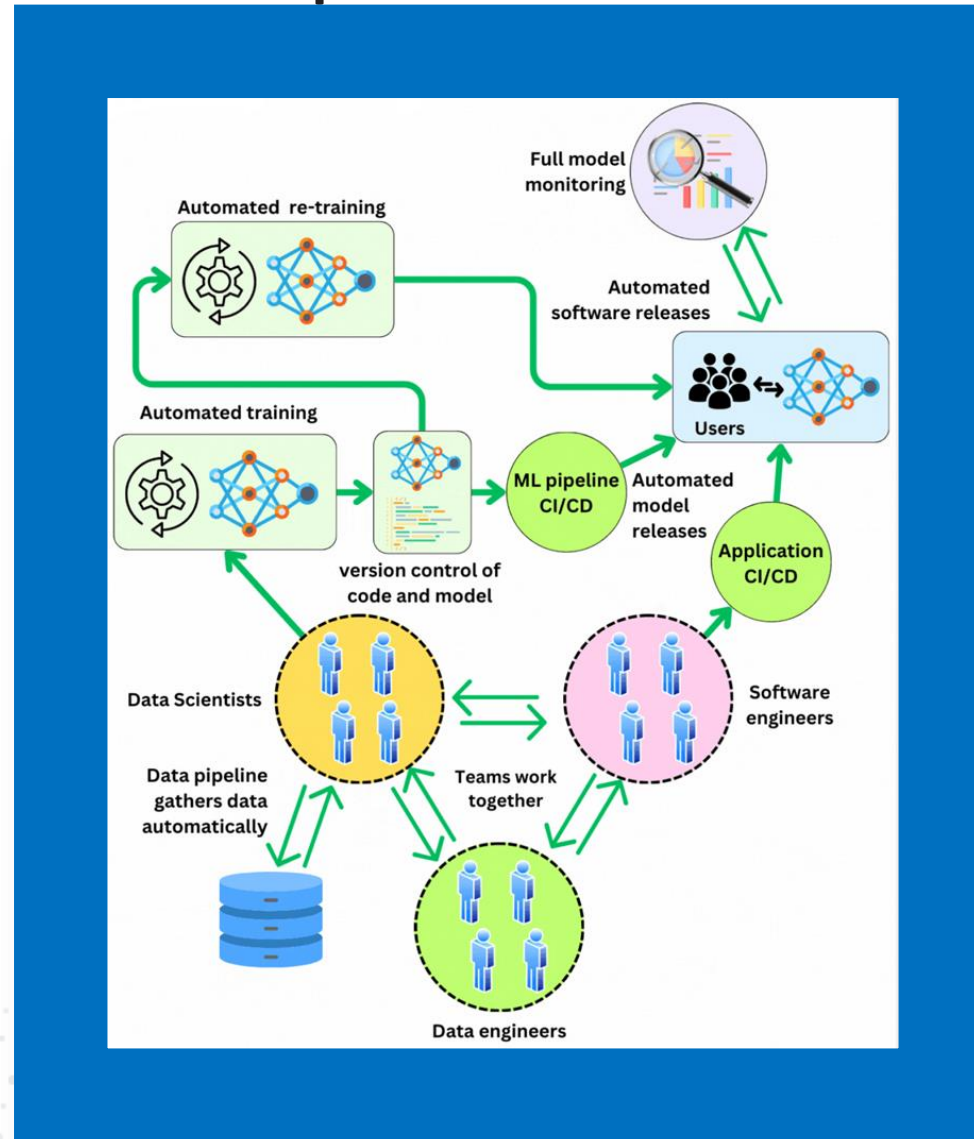
MLOps Maturity Levels

Level 0: Manual process



MLOps Maturity Levels

Level 4: Full MLOps automated re-training



Generic Machine Learning Process Model in TurkStat

Frame problem

1.1 Understand data and problem

1.2 Specify purpose

1.3 Identify expected results

1.4 Prepare working environment

Collect data

2.1 Identify potential data sources and determine the availability of these sources

2.2 Transfer data (from source)

2.3 Explore data

Clean and prepare data

3.1 Clean data (Removing errors, duplicates and imputing etc.)

3.2 Preprocess data (Feature engineering, Scaling, Transforming)

3.3 Store data (Location, Data dictionary, Versioning etc.)

Choose an algorithm

4.1 Experiment models (Performance, Operationalization, Code quality, Reproducibility)

4.2 Evaluate models (Effectiveness, Performance metrics)

4.3 Select model (Selection criteria, taking experimental code and preparing it)

Prepare base model

5.1 Train base model

5.2 Validate base model

5.3 Test base model

5.4 Fine-tune model (Hyperparameter tuning, Ensemble methods)

Launch or serve model

6.1 Get model into production

6.2 Identify reproducibility steps of production model

Monitor model

7.1 Monitor model performance (unit and integration testing)

7.2 Retrain model

7.3 Monitor model deviation (Monitoring changes in data and concept deviation)

Decision Map to Choose an ML Algorithm

Type of Problem

- Classification (predicting classes)
 - Binary Classification
 - Logistic Regression
 - SVM
 - Random Forest
 - Gradient Boosting
- Multi-class Classification
 - Logistic Regression
 - Random Forest
 - Gradient Boosting
 - Neural Networks
- Regression (predicting values)
 - Linear Relationship
 - Linear Regression
 - Ridge Regression
 - Lasso Regression
 - Non-Linear Relationship
 - Decision Trees
 - Random Forest
 - Gradient Boosting
 - Neural Networks
- Unsupervised Learning (Clustering)
 - K-Means
 - Hierarchical Clustering
 - DBSCAN
- Dimensionality Reduction
 - PCA
 - t-SNE (for visualization)
 - Autoencoders (for non-linear reduction)
- Anomaly Detection
 - Isolation Forest
 - One-Class SVM
- Sequential Decision Making (Reinforcement Learning)
 - Q-Learning
 - Deep Q-Networks

Interpretability

- High Interpretability Required
 - Linear Regression
 - Logistic Regression
 - Decision Trees
- Moderate Interpretability
 - Random Forest (can provide feature importance)
 - Naive Bayes
- Low Interpretability Acceptable
 - Deep Learning (Neural Networks)
 - Support Vector Machines

Data Size

- Small Dataset
 - Decision Trees
 - Linear Regression
 - Naive Bayes
- Large Dataset
 - Deep Learning (if computational resources permit)
 - Random Forest
 - Gradient Boosting

Data Complexity

- Linear Relationship Likely
 - Linear Regression
 - Logistic Regression
 - Ridge/Lasso Regression
- Non-Linear Relationship
 - Decision Trees
 - Random Forest
 - Gradient Boosting
 - Neural Networks

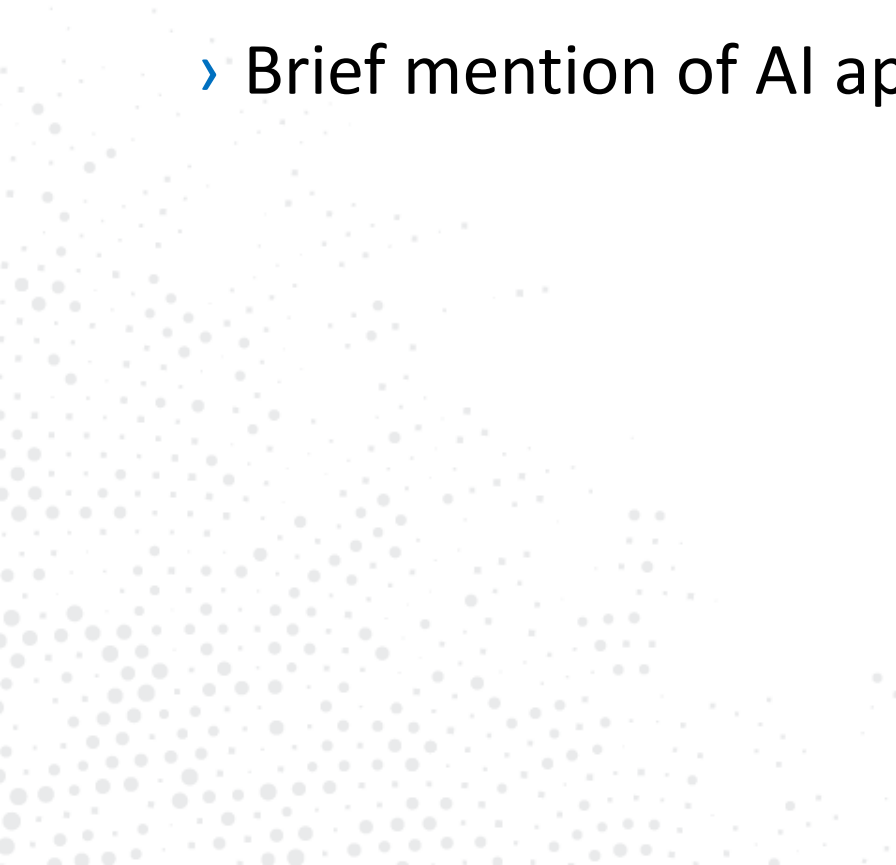
Resource Constraints

- Limited Resources
 - Decision Trees
 - Linear Regression
 - Naive Bayes
- Adequate Resources
 - Random Forest
 - Gradient Boosting
 - Support Vector Machines (SVM)
 - Deep Learning

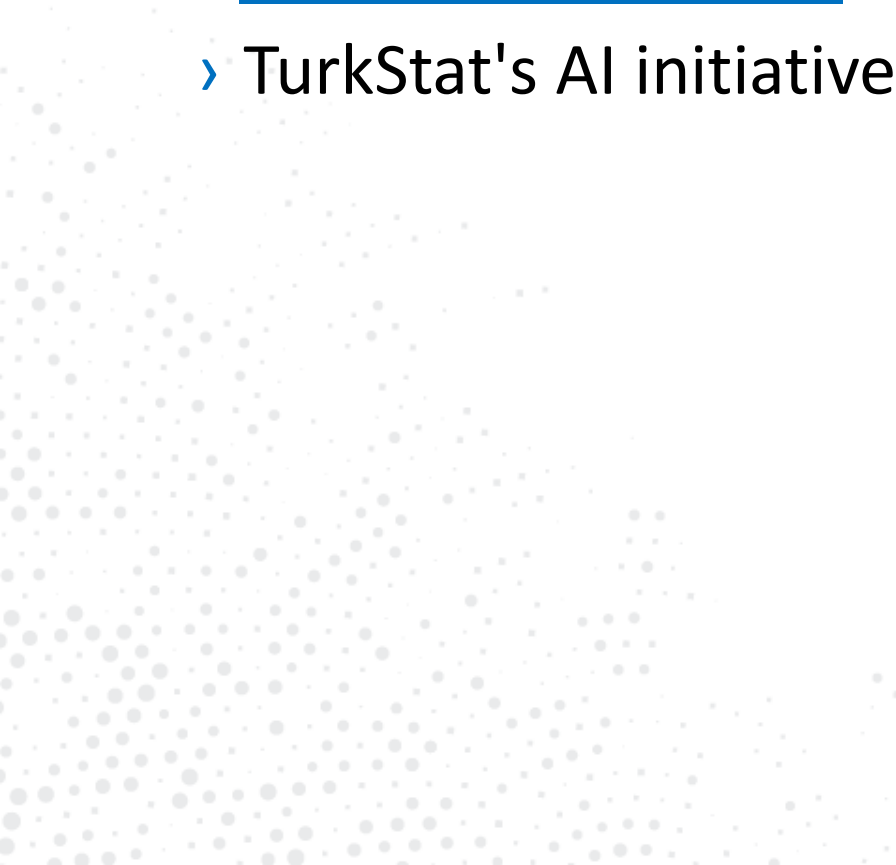
Domain Knowledge

- Strong Domain Knowledge
 - Use domain-specific insights to guide algorithm selection
- Limited Domain Knowledge
 - Start with versatile algorithms like Random Forest or Gradient Boosting

Introduction: AI's Role in Transforming Official Statistics

- › TurkStat's mission in official statistics
 - › Why AI and machine learning are critical for the future of data
 - › Brief mention of AI applications at TurkStat
- 

AI in Official Statistics: The Paradigm Shift

- › AI enhances data collection, processing and analysis.
 - › **New opportunities:** Automation, speed and accuracy.
 - › TurkStat's AI initiatives transforming official statistics.
- 

Web Scraping for Consumer Price Index (CPI) Calculations

- › Web scraping used to collect online prices for CPI.
- › Products include white goods, cars, furniture and more.
- › **Benefits:** Cost reduction, increased data frequency and accuracy.
- › Web scraping contributes 5.1% of CPI data in 2024.

Scanner Data for Automating Price Collection

- › TurkStat's use of scanner data for CPI calculations.
- › Sales data from retail stores: quantity, value, and price.
- › **Benefits:** Accuracy, reduced physical data collection and increased representativeness.
- › Scanner data now accounts for 42.6% of CPI data (as of January 2024).

Machine Learning for Classification

- › TurkStat uses machine learning to automate classification tasks (e.g., COICOP, ISCO-08).
- › **Example:** Job classification using neural networks.
- › **Benefits:** Efficiency, accuracy and scalability.
- › **Challenges:** Data quality, real-world application vs. pre-labeled data.

AI for Labor Market Analysis

- › TurkStat uses AI to improve labor market analysis.
- › Administrative data, web scraping and machine learning for job vacancy classification.
- › Neural networks predict ISCO-08 codes for job vacancies.
- › **Benefits:** Detailed insights into regional labor market trends.
- › Visualization through Business Intelligence (BI) tools.

Speech-to-Text Technology

- › TurkStat uses an offline speech-to-text model for meetings and presentations.
- › Multilingual capabilities, including Turkish.
- › **Benefits:** Efficiency in transcribing meetings and enhancing documentation accuracy.
- › **Applications:** Transcribing, voice-controlled commands, and converting speech into text.

E-Invoice Data and AI Challenges

- › TurkStat's project on predicting HS codes for 7.5 million daily e-invoices.
- › ML models used (e.g., Support Vector Machines, LLMs).
- › **Challenges:** Data inconsistency, labeling issues and real-world application accuracy.
- › Lessons learned and ongoing efforts to improve AI models.
- › Project was cancelled due to quality commitment.

Model	Number of 6-fold HS/CN that the model can predict (a)	The model predicts At least one of the CPAs is correct (b)	% (b/a)	% (b/715)	predicted by the model is correct (c)	% (c/a)	% (c/715)
ChatGPT (manual)	611	418	68%	58%	362	59%	51%
ChatGPT (API)	595	371	62%	52%	326	55%	46%
Claude	539	300	56%	42%	264	49%	37%
ChatGPT4	340	181	53%	25%	169	50%	24%
Gemini	557	225	40%	31%	178	32%	25%
llama	56	0	0%	0%			

Offline & Open Source LLMs

- › Underlying LLMs (gemma & llama) are open source.
- › Able to run offline
- › Used locally without violating data confidentiality.
- › Prompts are not kept in the background
- › Prompts are not used for model retraining.



The screenshot shows the TÜİK Büyük Dil Modelleri interface. It features a sidebar with the TÜİK logo and the text 'TÜRKİŞH STATİSTİKAL ENSTİTÜTE'. The main content area is divided into sections: 'Model & Görev' with a dropdown menu for 'Model seçiniz' (currently showing 'google/gemma-2-2b-it') and 'Görev seçiniz' (currently showing 'Text Generation'); 'Model Ayarları' with checkboxes for 'max_new_tokens değiştir: ?' and 'quantization kullan: ?'. A 'Modeli Çalıştır' button is located at the bottom right of the interface.

TÜİK Büyük Dil Modelleri

Altta çalışan büyük dil modelleri açık kaynak kodlu modellerdir.

Bu modeller çevrimdışı çalışabilme özelliği sayesinde veri gizliliğini ihlal etmeden yerel olarak kullanılabilirlerdir.

İstemler arka planda tutulmamaktadır ve yeniden model eğitimi için kullanılmamaktadır.

Sorgulamak istediğiniz metni aşağıya giriniz (prompt):

İstem

Modeli Çalıştır

Future Directions: AI and Ethics in Official Statistics

- › Importance of ethical considerations in AI applications.
- › Ensuring data privacy and security in AI systems.
- › Addressing biases in AI models and ensuring fairness.
- › Responsible innovation: Balancing automation with transparency.
- › TurkStat's commitment to ethical AI development.

Conclusion: The Future of AI in Official Statistics

- › AI's transformative role in official statistics.
- › TurkStat's innovative applications of AI and machine learning.
- › Challenges and ethical considerations.
- › **Future outlook:** Continued innovation and responsible AI development.
- › Invitation for questions and discussions.

Thanks for listening

› Bilal Kurban – bilal.kurban@tuik.gov.tr