DATA-DRIVEN RISK ASSESSMENT IN INFRASTRUCTURE NETWORKS

ANTÓNIO PEREIRA BARATA[1,2], GERRIT-JAN DE BRUIN[1,2], FRANK TAKES[1], COR VEENMAN[1,3], AND JAAP VAN DEN HERIK[1]

{a.p.pereira.barata, g.j.de.bruin}@liacs.leidenuniv.nl

[1]Leiden Centre of Data Science, Leiden Institute of Advanced Computer Science, Leiden University
[2]Human Environment and Transport Inspectorate, Ministry of Infrastructure and Water Management
[3]Data Science Department, TNO

Human Environment and Transport Inspectorate
Ministry of Infrastructure and Water Management

Universiteit Leiden
Leiden Institute of Advanced Computer Science

ABOUT THE PROJECT

Although awareness around gathering, storing and managing data has increased in the public sector, surprisingly little is done to make use of this data. At the same time, the public sector is under pressure from governing bodies to become more information oriented (data-driven) when it comes to operations, strategic decision making and policy making. This project addresses the aforementioned discrepancy by investigating the possibility of applying novel data science methods to data from the public sector. More in particular, we consider data related to infrastructure and the human environment, i.e., transport over roads and water in the context of liability.

WASTE SHIPMENTS REGULATION

The Waste Shipments Regulation (WSR) comprises the legislation that companies must follow in order to transport waste materials through a EU member state. The legislation requires a notice from any company wanting to transfer waste across Dutch borders, as well as a guarantee fee. Moreover, different waste categories have different fee rates. As a result, some companies might intentionally mislabel their waste. Having domain experts (inspectors) investigate all waste transportation events is not possible: it is impossible to address every case since time and human resources are limited. We want to assist the inspectors in targeting the right transports which should be inspected.

GOAL

To use outlier detection to find transports that strongly deviate from other transports within the reported waste category. This could imply fraudulent behavior from companies, enabling inspectors to take the appropriate action.
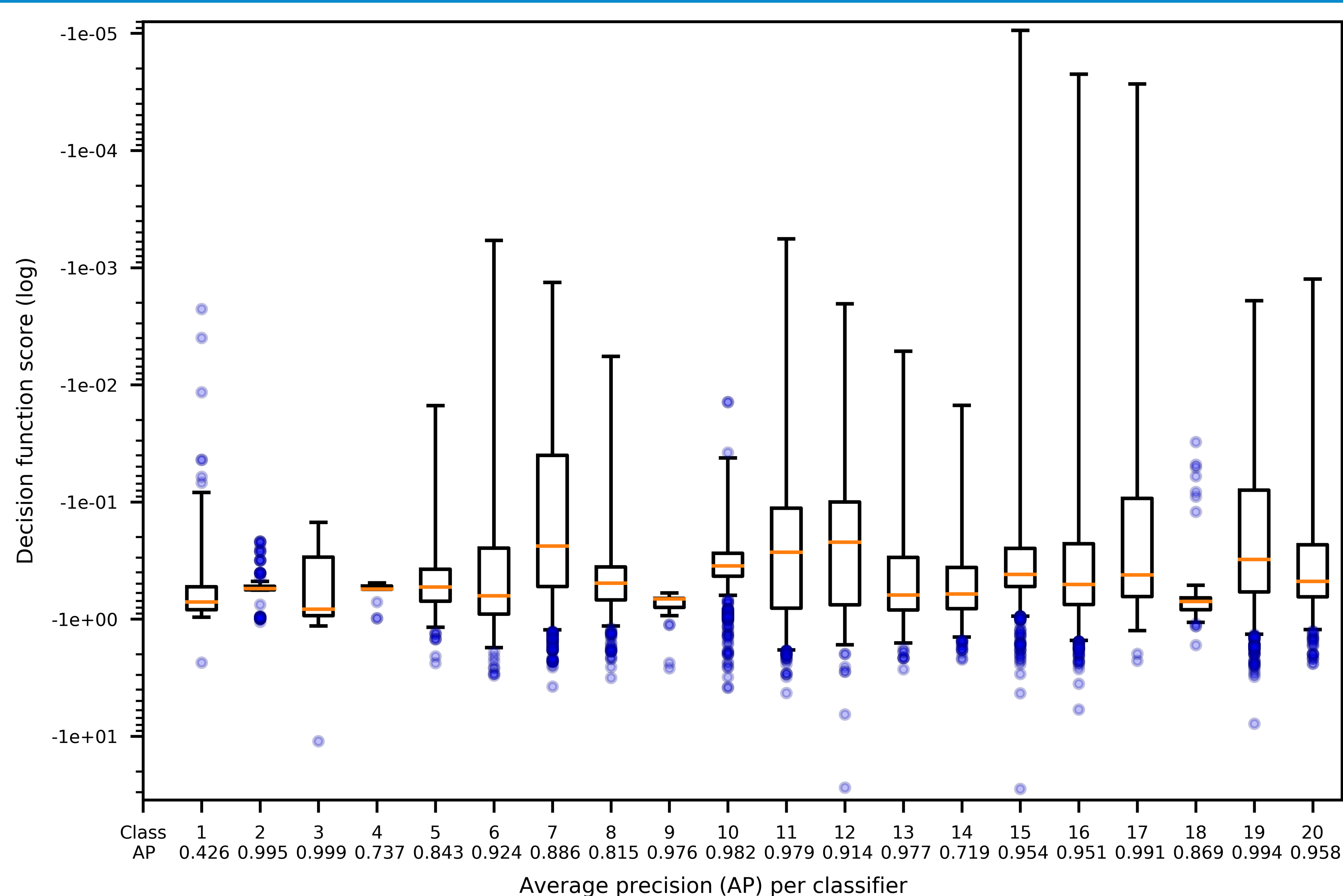
DATA

Example of report variables:
- Waste origin
- Origin company name
- Transport route
- Border location
- Tonnage
- Way of disposal
- Waste type

REAL LIFE SCENARIO

APPROACH

- Supervised learning (target = waste category)
- Linear support vector machine classifier (LSVMC) [1]
- One-vs-rest for the twenty waste categories
- Five-fold train-test split
- Classifier evaluation based on average precision (AP) [2]
- Select classifiers based on performance
- For each selected classifier, false negatives with lowest scores are indicative of possible fraudulent behavior
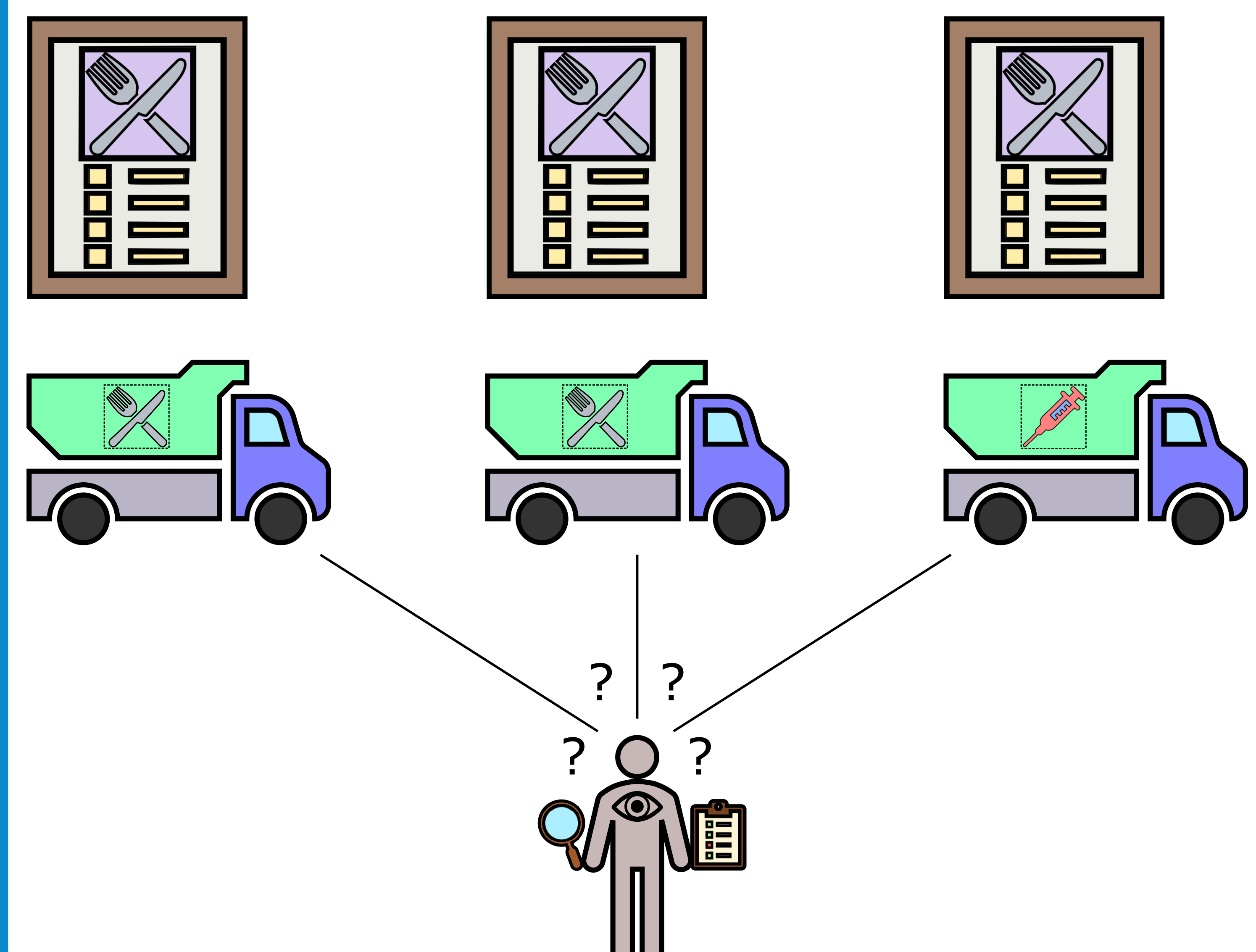
RESULTS

Class	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
AP	0.426	0.995	0.999	0.737	0.843	0.924	0.886	0.815	0.976	0.982	0.979	0.914	0.977	0.719	0.954	0.951	0.991	0.869	0.994	0.958

Average precision (AP) per classifier

FUTURE WORK

We seek to compare our results with actual inspection data while cooperating with domain experts, to improve and/or create methodological approaches that will aid such professionals.

REFERENCES

[1] T. Hastie, R. Tibshirani, and J. Friedman: *The elements of statistical learning.*, Springer, 2 ed., 2009.

[2] T. Saito and M. Remsmeier: *The precision recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets*, PLOS ONE, vol. 10, no. 3, 2015.