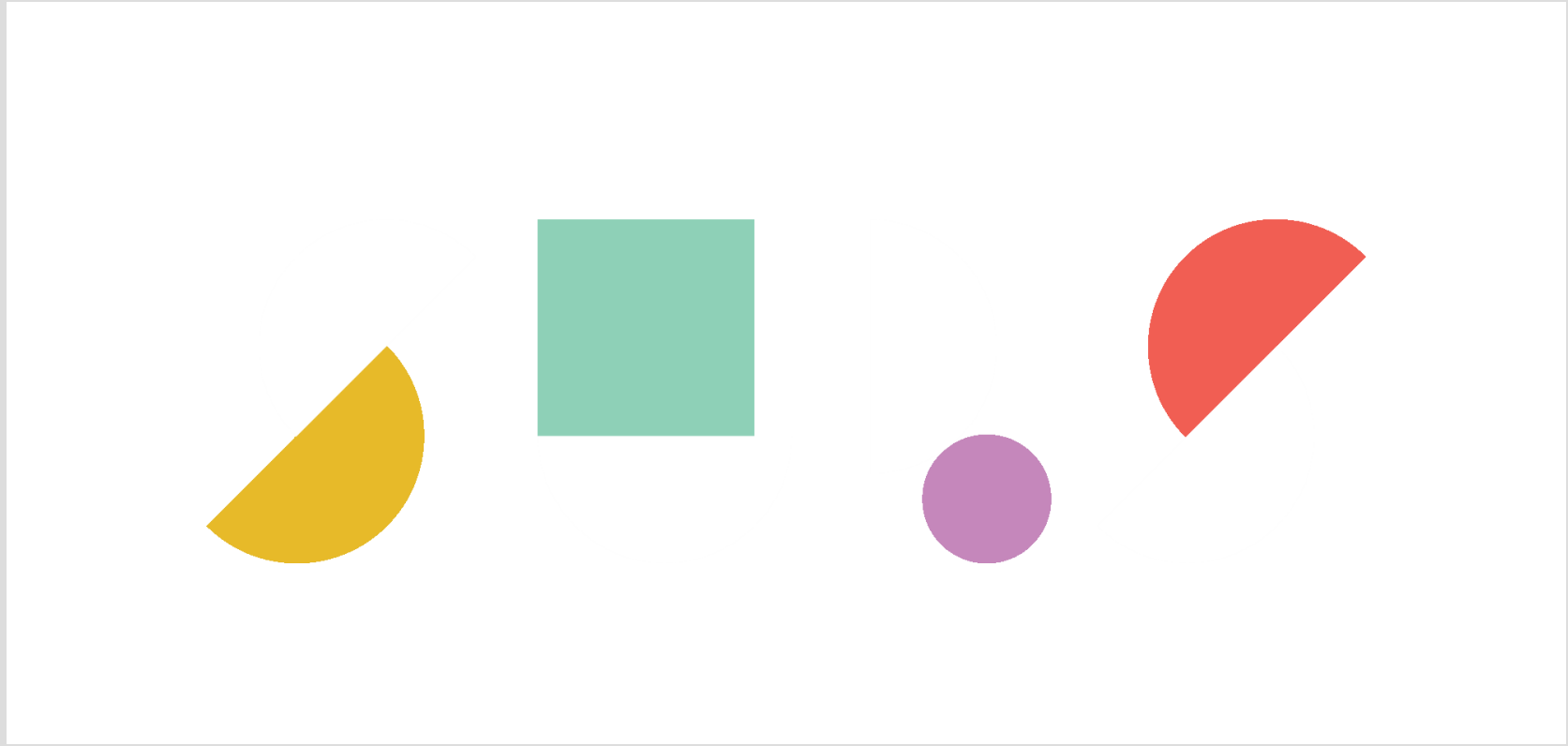
REPUBLIKA SLOVENIJA
STATISTIČNI URAD



Examples of classification at SURS

Črt Grahonja

WP10
26. 9. 2024

Timeline

- 2019:
  - Scraped OJAs → is a text an OJA or not?
  - Classification of computer articles → 17 classes
- 2021:
  - Classification of receipt articles → more than 200 classes
  - Classification of scanned articles → 193 classes
- 2023:
  - Classifying ISCO and NACE from survey answers
- 2024:
  - Scraped OJAs → is specific information present in text (deadline) ← *LLMs*

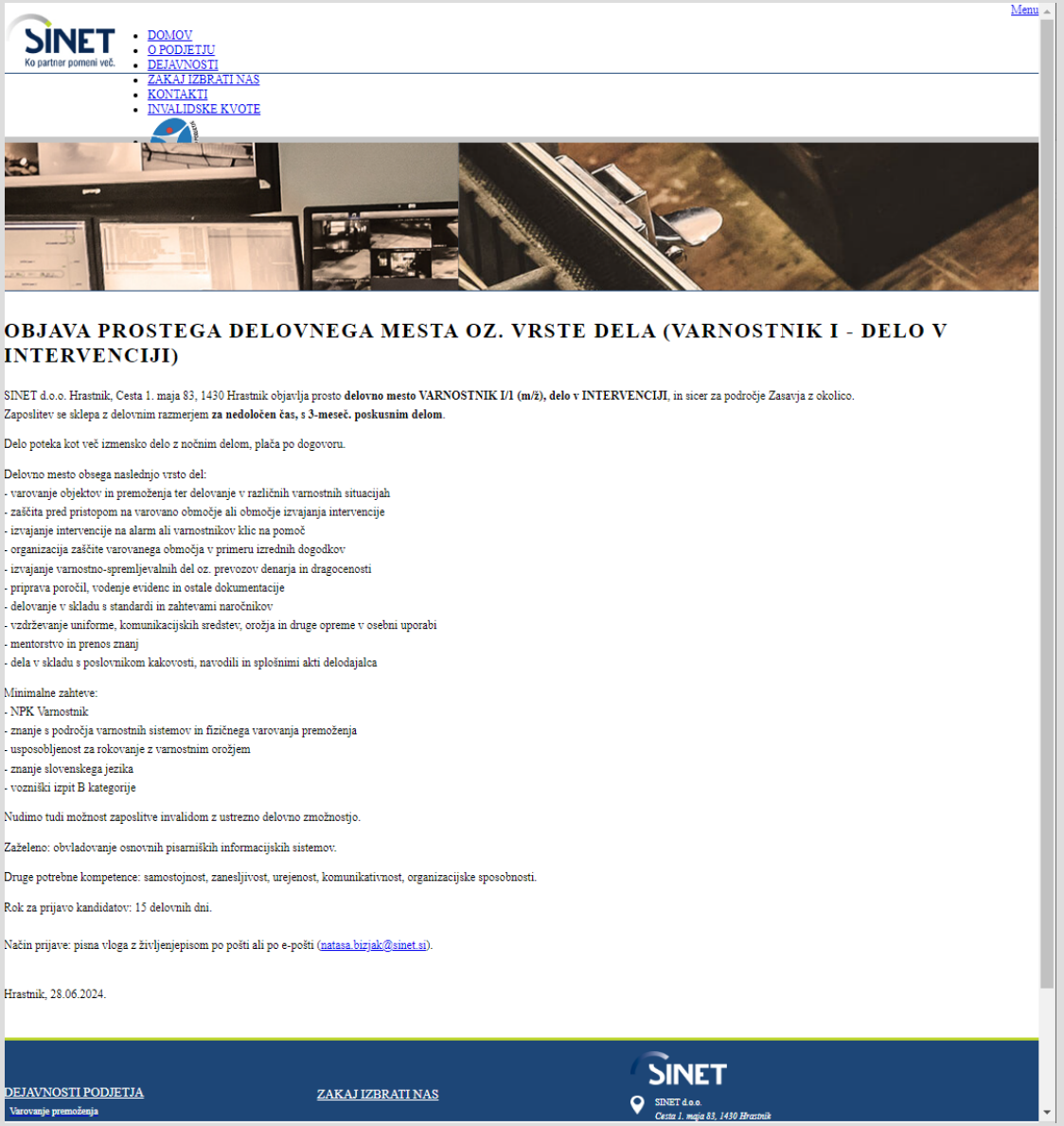
ECOICOP

Classifying OJAs

- Estimation of number of job vacancies:
  - Different models:
    - Linear regression
    - Logistic regression for presence → LinReg and NN
    - Adaboost for presence → LinReg and NN
- Working in a ML toolset (Orange Data Mining)
  - Memory limited → small datasets

OJAs' data

- The collection process:
  - crawl known know companies' pages
  - search for job ad candidates
  - scrape candidates
- Monthly
- Longer texts
- Slovenian (language filter)
  - grammatical cases
  - foreign words
  - *grammatical mistakes*
- 595 texts:
  - 61.8% not OJA, 38.2% OJA



SINET
Ko partner pomeni več.

Menu

- DOMOV
- O PODJETJU
- DEJAVNOSTI
- ZAKAJ IZBRATI NAS
- KONTAKTI
- INVALIDSKE KVOTE

OBJAVA PROSTEGA DELOVNEGA MESTA OZ. VRSTE DELA (VARNOSTNIK I - DELO V INTERVENCIJI)

SINET d.o.o. Hrastnik, Cesta 1. maja 83, 1430 Hrastnik objavlja prosto **delovno mesto VARNOSTNIK I/1 (m/ž), delo v INTERVENCIJI**, in sicer za področje Zasavja z okolico.
Zaposlitev se sklepa z delovnim razmerjem **za nedoločen čas, s 3-mesečnim poskusnim delom**.

Delo poteka kot več izmensko delo z nočnim delom, plača po dogovoru.

Delovno mesto obsega naslednjo vrsto del:
- varovanje objektov in premoženja ter delovanje v različnih varnostnih situacijah
- zaščita pred pristopom na varovano območje ali območje izvajanja intervencije
- izvajanje intervencije na alarm ali varnostnikov klic na pomoč
- organizacija zaščite varovanega območja v primeru izrednih dogodkov
- izvajanje varnostno-spremljevalnih del oz. prevozov denarja in dragocenosti
- priprava poročil, vodenje evidenc in ostale dokumentacije
- delovanje v skladu s standardi in zahtevami naročnikov
- vzdrževanje uniforme, komunikacijskih sredstev, orožja in druge opreme v osebni uporabi
- mentorstvo in prenos znanj
- dela v skladu s poslovnikom kakovosti, navodili in splošnimi akti delodajalca

Minimalne zahteve:
- NPK Varnostnik
- znanje s področja varnostnih sistemov in fizičnega varovanja premoženja
- usposobljenost za rokovanje z varnostnim orožjem
- znanje slovenskega jezika
- vozniški izpit B kategorije

Nudimo tudi možnost zaposlitve invalidom z ustrezno delovno zmožnostjo.

Zaželeno: obvladovanje osnovnih pisarniških informacijskih sistemov.

Druge potrebne kompetence: samostojnost, zanesljivost, urejenost, komunikativnost, organizacijske sposobnosti.

Rok za prijavo kandidatov: 15 delovnih dni.

Način prijave: pisna vloga z življenjepisom po pošti ali po e-pošti (natasa.bizjak@sinet.si).

Hrastnik, 28.06.2024.

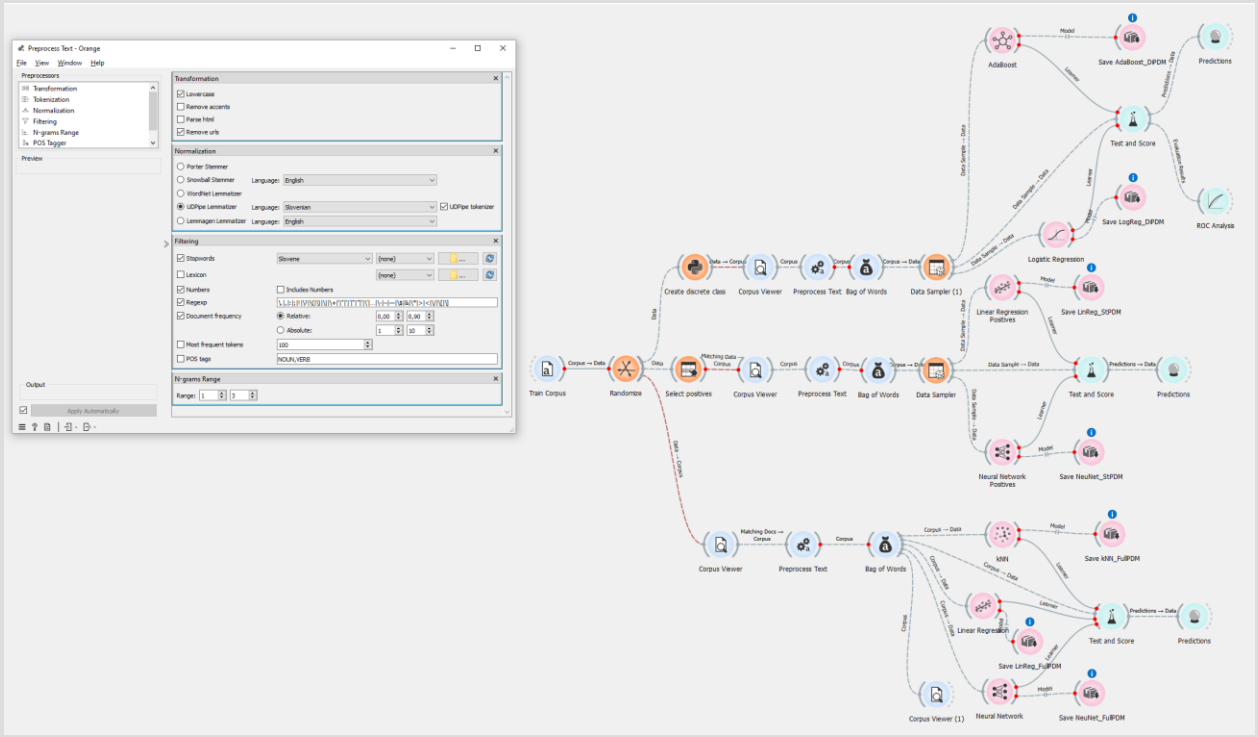
DEJAVNOSTI PODJETJA
Varovanje premoženja

ZAKAJ IZBRATI NAS

SINET
SINET d.o.o.
Cesta 1. maja 83, 1430 Hrastnik

Pre-processing

- Training set: 600 texts
  - 20% for testing
- Text transformations:
  - to lower case
  - removal of URLs
  - lemmatization
  - removal of stopwords
  - removal of punctuation marks
  - removal of words/tokens present in the top 10 percentiles of documents
- Creation of 1-3 n-grams
- BoW

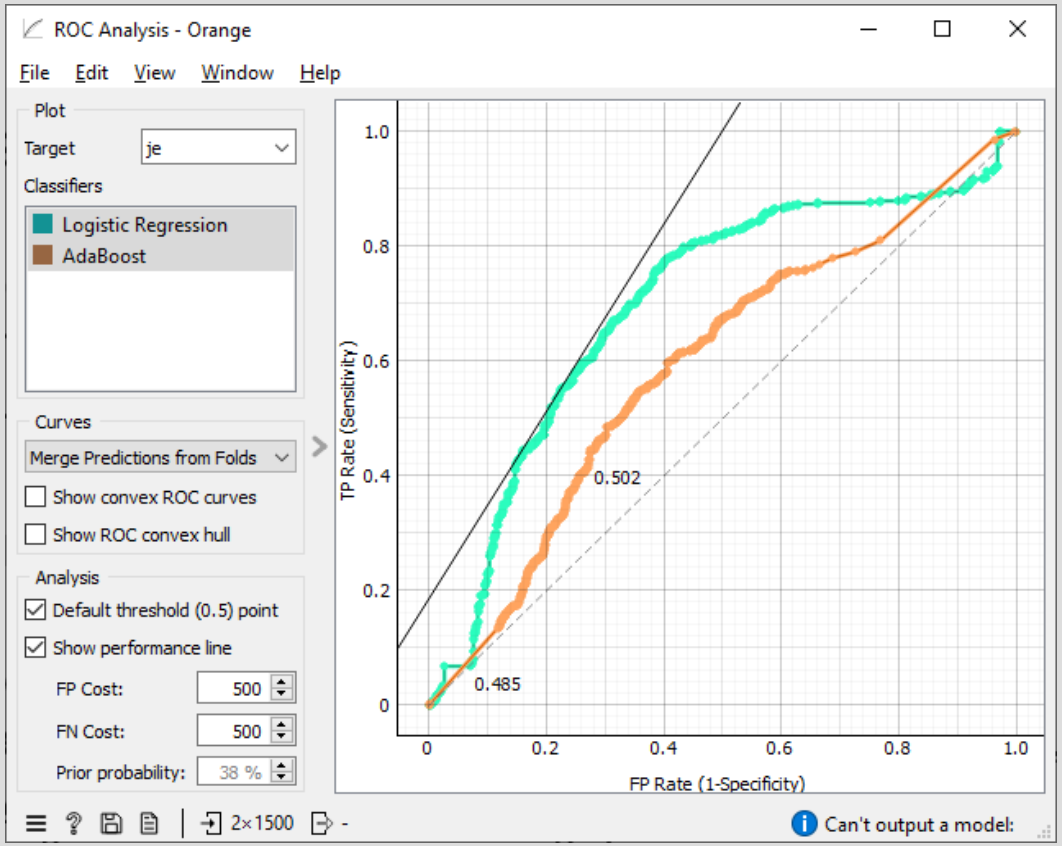


Machine Learning models

- Logistic Regression
  - tested different hyperparameters
    - regularization: none, Lasso, Ridge
    - C parameter: 1, 2, 3, 5, 20, 50, 100, 150, 200, …
    - class distribution: none, balanced
- AdaBoost
  - base estimator: decision trees
  - tested different hyperparameters
    - number of trees: 10, 20, 100, 200, 1000
    - learning rate: 0.001, 0.002, 0.01, 0.02, 0.1, 0.2, 0.3, 0.4, 0.5, …
    - algorithm: SAMME, SAMME.R
    - loss function: linear, square, exponential
- Learning process
  - random sampling 50 times at 75% - 25% split
  - stratified

Training results

- Not too great*
  - Accuracy:
    - Logistic Regression: 61.60%
    - AdaBoost: 61.73%
  - Precision:
    - Logistic Regression: 63.36%
    - AdaBoost: 68.69%
  - Recall:
    - Logistic Regression: 93.37%
    - AdaBoost: 72.74%
  - F1:
    - Logistic Regression: 75.49%
    - AdaBoost: 70.65%



** This results examine positives as non-presence of JV ads*
The distribution is: 37:63 for presence vs non-presence

Additional NER

- In cooperation with IJS
  - Spacy NLP model
- Extracting 3 values:
  - number of OJAs (by extraction of occupation names)
  - publishing dates
  - deadlines
- Training set a JSON with manual extractions of 10.000 texts
  - with literal extracts and interval
- Worked well in 2019
  - dated nowadays


```
{
    "id": 3,
    "content": "iščemo nove sodelavce/sodelavke | integral ‹ › × pošljite povpraševanje info@integ
    "entities": {
        "pubDate": [{
            "text": "",
            "interval": { "start": null, "length": 0}
        }],
        "deadline": [{
            "text": "do 4. 3. 2018",
            "interval": { "start": 1512, "length": 13}
        }],
        "numberOJA": [{
            "text": "najmanj tri voznike/ voznice avtobusa",
            "interval": { "start": 342, "length": 37}
        }]
    }
},
```

```
{
    "texts": [
        {
        "id": 1,
        "orig": "Univ. ali visoka strok. izobr. gradbene smeri ali Univ.
        "entities": {
            "deadline": "Rok za oddajo vlog je 24.6.2020",
            "numberOJA": "o naslednjem prostem delovnem mestu"
            }
        },
        {
        "id": 2,
        "orig": "zrc d.d. trbovlje - o podjetju o podjetju rešitve stori
        "entities": {
            "deadline": "",
            "numberOJA": "nimamo razpisanih prostih delovnih mest"
            }
        },
        {
        "id": 3,
        "orig": "prosta delovna mesta prosta delovna mesta napisal super
        "entities": {
            "deadline": ["objavljeno:    05.06.2015 rok prijave:    30 c
            "numberOJA": ["voznik m/ž", "samostojni električar m/ž"]
            }
        }
    ]
}
```

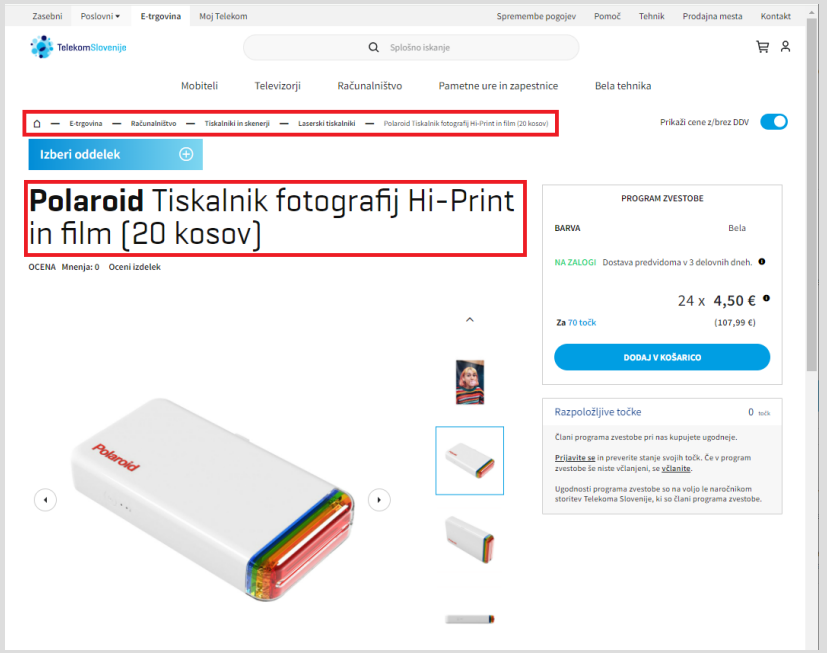
Future work

- See LLM slide at the end

- Update the input domains
  - Improve the training sets
  - Retraining models

Classification of computer components

- *Our biggest success story*
  - *project of 3 years*
- Scraping of computer articles across specific stores' pages
- Relevant information:
  - article name
  - *breadcrumbs*
- 2-stage classification
  - classify to ECOICOP
  - test for validity



Articles' data

- 33 000 examples
  - article URL (used as ID)
  - name of article
  - *breadcrumbs*
  - **both target variables:** ECOICOP, ECOICOP2

Either ECOICOP or 9999999999 ↵

ECOICOP	Type of article	Number of examples
91317293301	CPU	582
91317293302	RAM	1495
91317293303	Motherboard	1010
91317293304	Hard disk	417
91317293305	Power supply	1176
91317293306	Optical units (DVD/RW)	40
91317293307	Graphic cards	1103
91317293308	Speakers	364
91317293309	Case	1691
91317293310	Mouse	1718
91317293311	Keyboard	835
91317293312	SSD disk	867
913172936	Tablet computer	3080
913172960	Portable computer	5891
913272932	PC Monitor	3177
913272935	Portable hard disk	993
913272950	Printer or multifunction device	1927
913372970	Operating system (Microsoft Windows)	77
913372975	Office suite (Microsoft Office)	63
954972980	Printer ink or cartridge	6498

Number of examples in training set

Article classification

C# pipeline – 1st step

1. Introduce data with
   - article URL
   - article name
   - article breadcrumbs
   - 1st target variable – ECOICOP
   - 2nd target variable – ECOICOP2 (binary)
2. Train *1 vs. All Logistic Regression* – 20 ECOICOP classes
   - article name and article breadcrumbs → 1st target variable
   - we also save the winning score

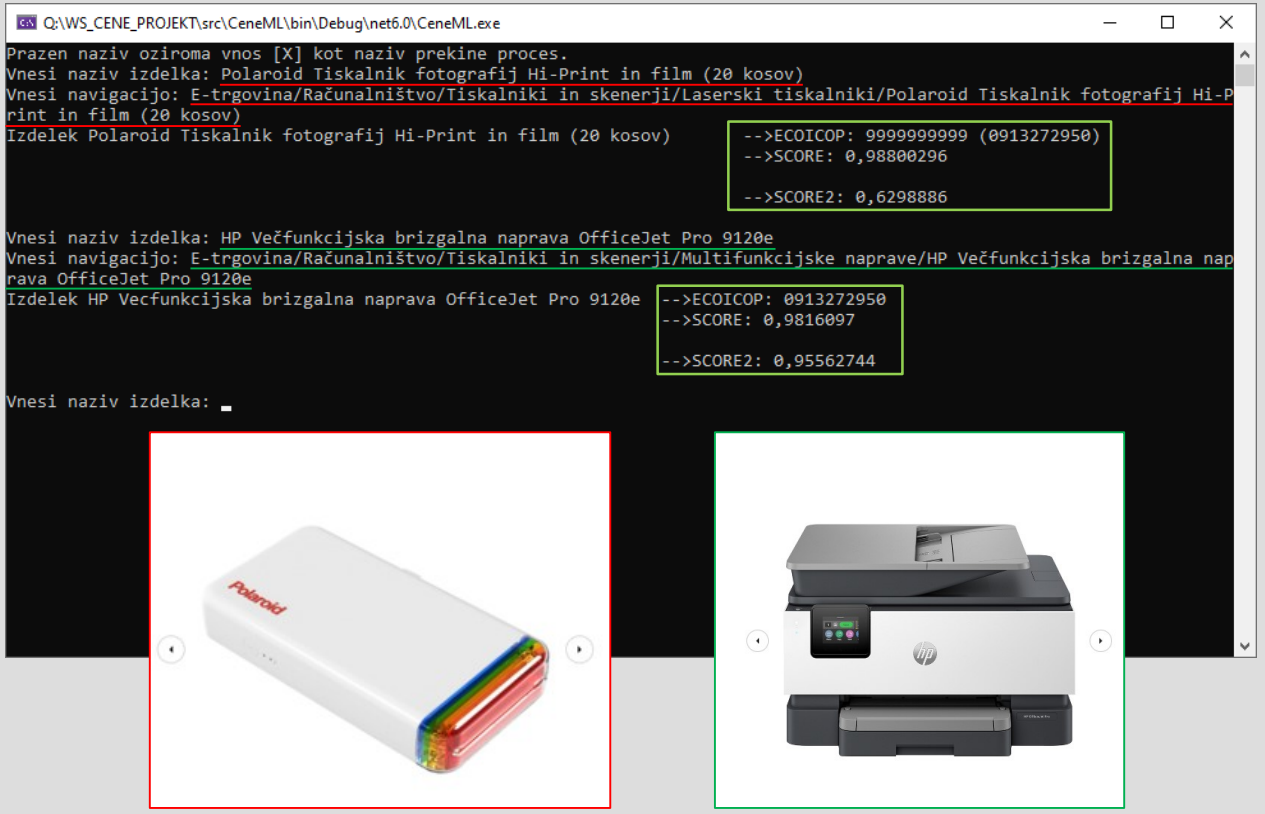
C# pipeline – 2nd step

3. Train 17 additional models
   - On subset of data with relevant class
   - Different model for each class:
     - Logistic regression
     - Maximum Entropy
     - Light Gradient-Boosting Machine
     - Decision Trees/Forest
   - article name and article breadcrumbs → 2nd target variable and score
   - Binary classification: ECOICOP or 9999999999 (*non-valid*)
4. Save Data to Database

- Training with Microsoft.ML tool
- Normalized unigrams and n-grams
- Automatic choice based on Macro-Accuracy
- Many model families
- Many hyper-parameters

In production

- 2-week data classified
  - ~100.000 articles
  - classes used for prediction of the CPI (MARS method)
  - Programmes are prepared for retrain (every 5 years)
    - Domain professionals update relevant classes
    - Filling new/updated training sets



Future work

- Mostly satisfied with how it’s working

- Retraining models to update them
- Maybe adding new classes and eliminating obsolete ones

Classification of OCR articles

- “A lot of work for a mediocre result”
- A large number of classes – 169
  - grouping at higher levels for irrelevant
  - still not all classes
  - very large number of observations
  - a lot of manual work
  - very imbalanced
- Classification of articles on receipts
  - to the highest level possible (7th)
- OCR not perfect
  - mistakes have to be taken into account

Lidl Slovenija d.o.o. k.d.

ARTIKEL	CENA	KOL.	SKUPNO
Artikel	1.35	1.426	1.93
NAK.VREC.JUB. 10 LET	0	1	0.12
SOLATA GENTILE	0.79	2	1.58
ALJAŽEVA SALAM	0.89	2	1.78
KISL. KUMAR.	0.65	3	1.95
TUNA LAST. SOK	0	1	2.49
CVETLIČNI MED	0	1	4.39
KRUHOV ČIPS	1.29	2	2.58
KRUHOV ČIPS	1.29	2	2.58
Total:			19.4
Total po vsoti:			19.4

Training data

- A massive dataset of almost 8.000.000 labelled observations
  - article name ←→ ECOICOP
  - most of it is from scanner data
    - no manpower to create such data otherwise
    - scanner data is fairly close to the actual data
      - future training sets will include more manual data
  - augmented with perturbation to simulate OCR mistakes
  - added additional rare items (03, 09, vaping fluid, …)
  - added problematic pairs (e. g. *zemlja-žemlja*)
  - additional care for special characters (č, š, ž; foreign as well)

Pre-processing and training

- Vectorizer:
  - character (1-3)-gram
- Train-test split: 80%-20%, *stratified*
- Tested different model families:
  - Multiclass Logistic Regression
  - Decision Tree Classifier
  - Random Forest Classifier
  - Support Vector Machines; *not very useful on such big data*
  - Multi-Layer Perceptron Classifier
- Using a Grid Search

Decided on ensemble of MLP + LogReg + DecTree

Not good enough…

- At this point we decided to group articles into higher ECOICOP classes
- Multi-step classification for some very relevant classes
  - milk types, meats, water types, 097 (books, magazines, art supplies …)
  - new training sets
  - new models
    - smaller models: 2-6 subclasses per model
    - 4 second-step models
  - milk and meats: three-steps classification
    - 8 third-step models
- Any score <0.7 → manual check (*human-in-the-loop*)
  - the actual ECOICOP selection: by popular vote *or* by maximum score in the ensemble (the Dec Tree classifier is weighted at 0.5)

The final process

- Using rules we eliminate/classify special cases
- 150.000 articles in 20.000 receipts
  - eliminate empty articles
  - eliminate wrongly detected articles (“total”, “taxes”, “shop points” …)
  - known restaurants’ receipts (from a list of restaurants) → 1111999
  - fuel types → 07221, 07222, 07223, 07224
  - gas cylinders → 0452201

} 198 + 6 = 204 total classes

- Run classifying process
  - anything under the benchmark is marked for manual check
  - run secondary and tertiary steps
    - in case of manual check, will be used as help
- Manual check
  - months of work
  - 11.600 checked

ID	NRACUN	NAZIVARTIKLA	CENA	KOLICINA	SKUPNACENA	COICOPNAZIV	COICOP	COICOP_P	COICOP097_P	COICOP01250_P	zaRocno
52	12	West Blue 20s	4,1	1	4,1	cigarete	0230101	0,999940431			0
53	13	Radenska cl. 1,51	0,75	1	0,75	naravna mineralna voda gazirana	0125001	0,95494666		0,738745104	0
54	13	Donat Mg 11	1,79	1	1,79	naravna mineralna voda gazirana	0125001	0,851740114		0,997055324	0
55	13	Motni sok 11	0,99	1	0,99	OCR_Sok	0121022	0,998864729			0
71	13	RDEČ.GROZD.BREZ PEČK NEP.	3,99	0,928	3,7	grozdje	0116504	0,693576256			1
72	13	Hamburg.bomb. 90g	0,39	6	2,34	OCR_Kruh izdelki	011130133	0,826579174			0
73	13	Žemlja 70g	0,18	1	0,18	OCR_Kruh izdelki	011130133	0,999978886			0
74	13	Žemlja 70g	0,18	1	0,18	OCR_Kruh izdelki	011130133	0,999978886			0
76	14	Faber - Castell	5	1	5	OCR_Drugi majhni potrošni gospodinjski izdelki	0561956	0,441140659			1
107	17	POMARANČNI SOK	1	2	2	OCR_Sok	0121022	0,997409927			0
108	17	NUTELLA 1kg	5,49	1	5,49	čokoladni namazi (nutella)	0118506	0,678105719			1
109	17	BIO SADNI SOK	0,69	1	0,69	OCR_Sok	0121022	0,999690818			0
157	26	TEST NEWGENE HAG SARS-COV-2 ART	14,61	3	43,83	OCR_Papirni in pisarniški material	0974043	0,504819727	0,987141159		1

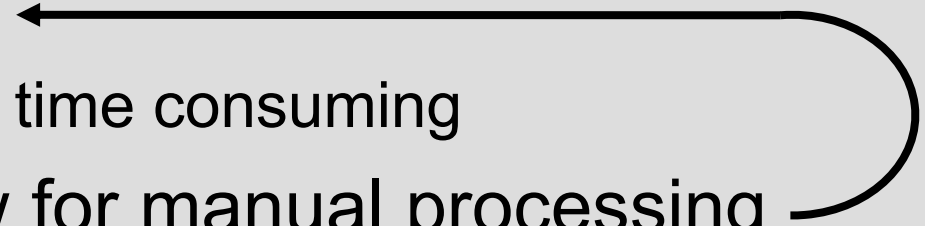
ID	DecisionTreeClassifier	DecisionTreeClassifier_P	LogisticRegression	LogisticRegression_P	MLPClassifier	MLPClassifier_P
52	cigarete	1	cigarete	0,999827714	cigarete	0,999993581
53	OCR_Vode, vode z okusom, gazirane, negazirane	1	OCR_Vode, vode z okusom, gazirane, negazirane	0,996729239	OCR_Vode, vode z okusom, gazirane, negazirane	0,868110741
54	OCR_Kozmetika	1	OCR_Vode, vode z okusom, gazirane, negazirane	0,851740114	OCR_Pripravljena hrana	0,07485858
55	OCR_Sok	1	OCR_Sok	0,999958497	OCR_Sok	0,99663569
71	OCR_Toploteka, pripravljene jedi	1	grozdje	0,999332363	grozdje	0,38782015
72	OCR_Kruh izdelki	1	OCR_Kruh izdelki	0,490279724	OCR_Kruh izdelki	0,989457797
73	OCR_Kruh izdelki	1	OCR_Kruh izdelki	0,999936663	OCR_Kruh izdelki	0,999999995
74	OCR_Kruh izdelki	1	OCR_Kruh izdelki	0,999936663	OCR_Kruh izdelki	0,999999995
76	bonboni	0,130164796	OCR_Drobni čokoladni izdelki	0,182075382	OCR_Drugi majhni potrošni gospodinjski izdelki	0,441140659
107	OCR_Sok	1	OCR_Sok	0,992852148	OCR_Sok	0,999377634
108	OCR_Suho sadje in oreščki	1	čokoladni namazi (nutella)	0,380719788	čokoladni namazi (nutella)	0,97549165
109	OCR_Sok	1	OCR_Sok	0,99988842	OCR_Sok	0,999184034
157	OCR_Testenine	1	bonboni	0,458913763	OCR_Časopisi, revije, razne tiskovine in papirni/papirniški material	0,504819727

Expanding the classes

- An effort to extend the ECOICOP classification
  - Examined all classes that are in [01.1, 01.2, 02.1, 02.2]
    - food, drinks, alcohol, tobacco
    - 192 classes
  - All others become *'not valid'* – 8888801000
- Explanatory variables:
  - article name (as scanned)
  - GTIN

Why?

- Scanned articles from biggest stores
  - ~70 000 articles / month
  - classification by rules and manually
    - often wrong
    - manual part time consuming
  - short window for manual processing



Preparation and training

- Selecting at maximum 5000 observations per class
  - many small classes
  - split: stratified by class
    - validation set: 20%
    - test: 20%
    - training: 60%
- Some pre-processing → Count Tokenizer n-grams (words or characters)
- 2 classifiers:
  - Decision Tree Classifier
  - *1 vs All* Classifier (based on Logistic Regression)
- Testing on combination of hyper-parameters
  - tokens: char vs word
  - trees:
    - Gini vs entropy
    - max depth
  - 1vA:
    - C criterion
    - tolerance

1. DTC pipeline:
   - char vectorizer
   - Gini criterion
   - no depth
2. 1vA pipeline:
   - char vectorizer
   - C = 20
   - tolerance = 0.01

Test results

- 170 00 articles: 1000- / class
- DTC pipeline:

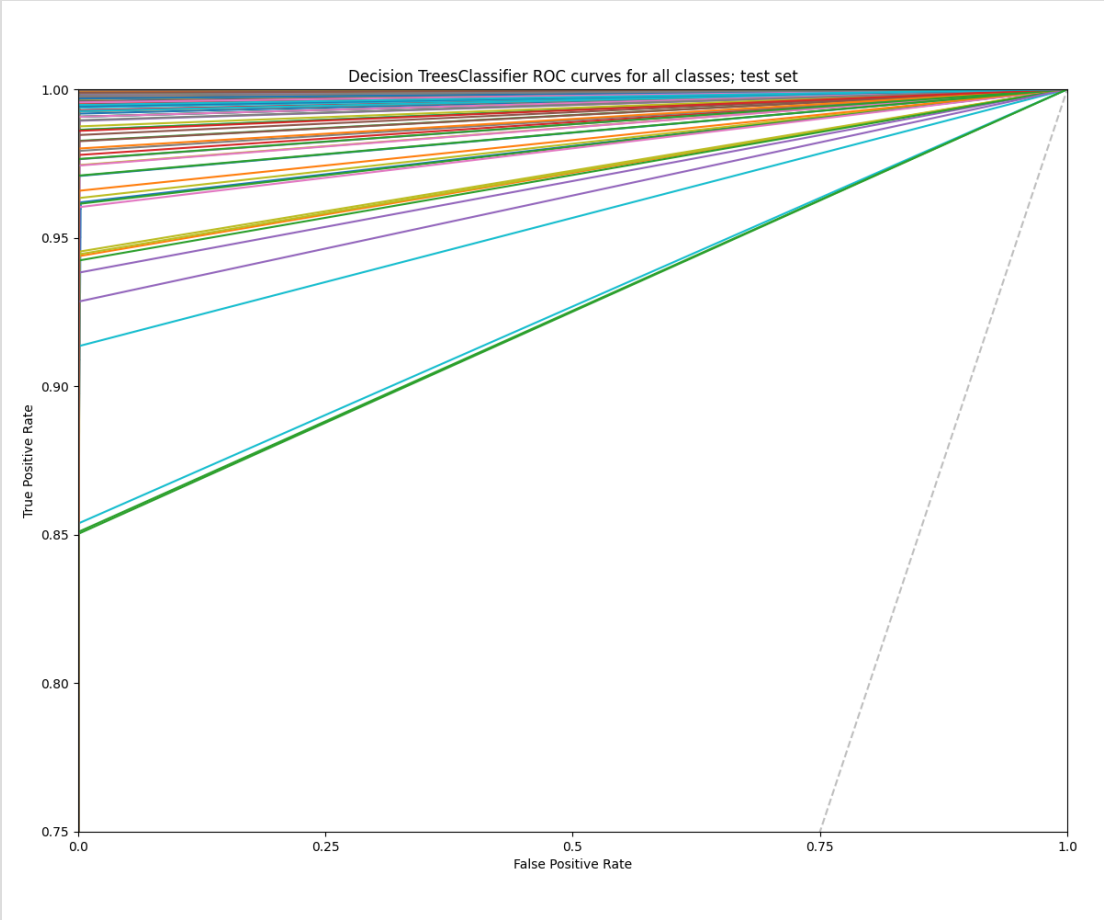
	precision	recall	f1-score	support
accuracy			0.98	171934
macro avg	0.98	0.98	0.98	171934
weighted avg	0.98	0.98	0.98	171934

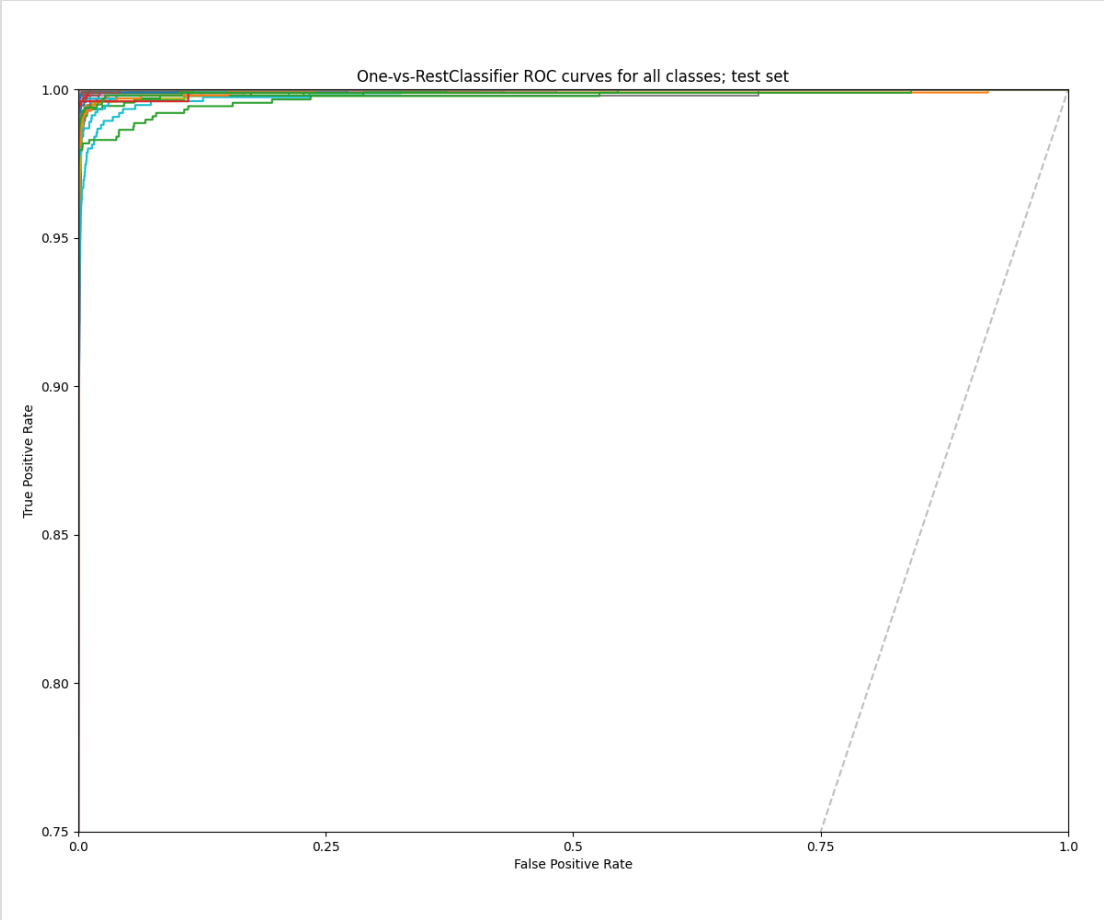
- 1vAC pipeline:

	precision	recall	f1-score	support
accuracy			0.99	171934
macro avg	0.99	0.99	0.99	171934
weighted avg	0.99	0.99	0.99	171934

- Some problematic characteristics:
  - DTC:
    - lowest precision: 0.85 (3)
    - lowest recall: 0.80 (1)
    - lowest F1 score: 0.85 (2)
  - 1vAC:
    - lowest precision: 0.84 (1)
    - lowest recall: 0.82 (1)
    - lowest F1 score: 0.86 (2)
  - lowest support: 76
    - 0220302000 – Heated tobacco, box
  - smallest 11 classes: <500

ROC curves - testing





Validation

- 170 00 articles: 1000- / class
  - 54 000 non-duplicated
- Accuracy:
  - DTC: 95.95 % (98.23 %)
  - 1vAC: 97.53 % (98.73 %)

- Problematic:
  - Non-duplicated classes have really small supports:

0220302000	41	Heated tobacco, box
0117404000	42	Potato puree
0117109000	48	Eggplants
0117119000	51	Leeks
0117118000	55	Lettuce, endive
0114202000	61	Milk, fresh, light
0117106000	64	Cucumbers, fresh
0116122000	66	Cherries
0116120000	67	Apricots
0117107000	68	Beans in pods

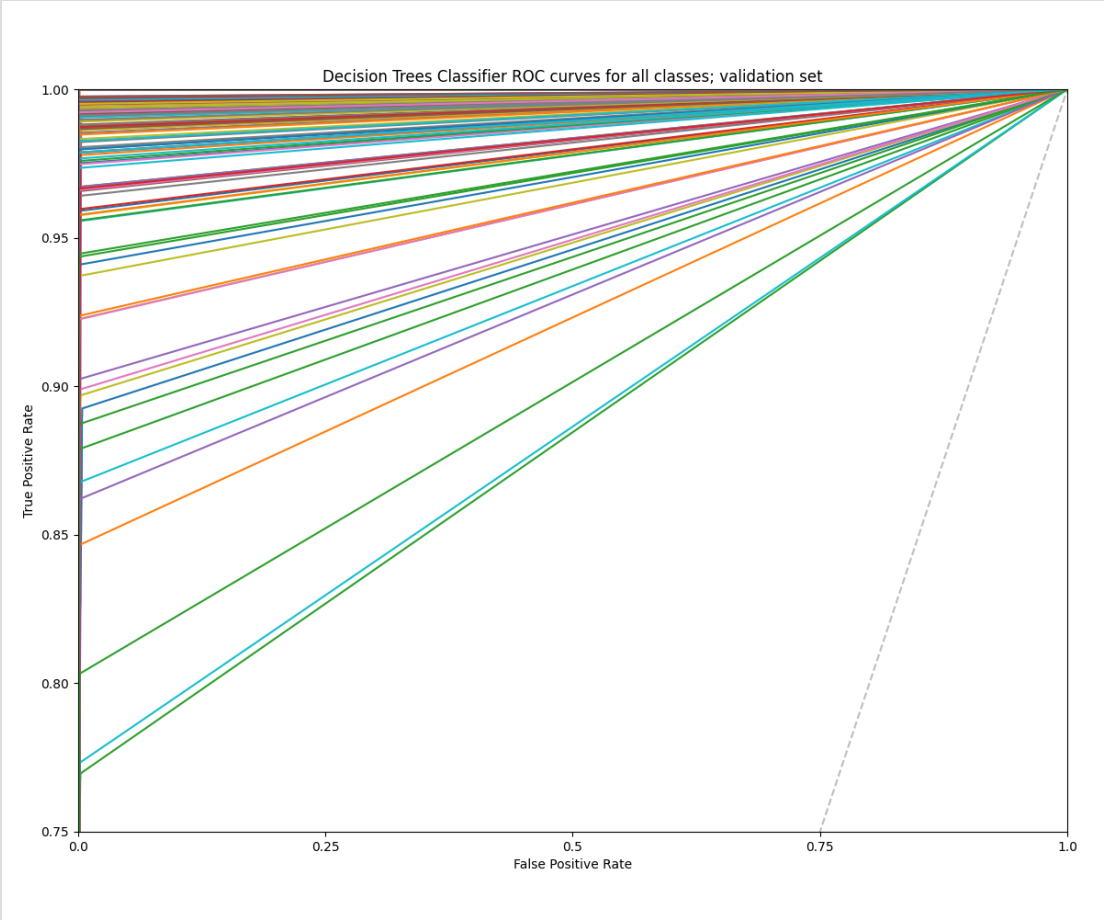
- DTC:

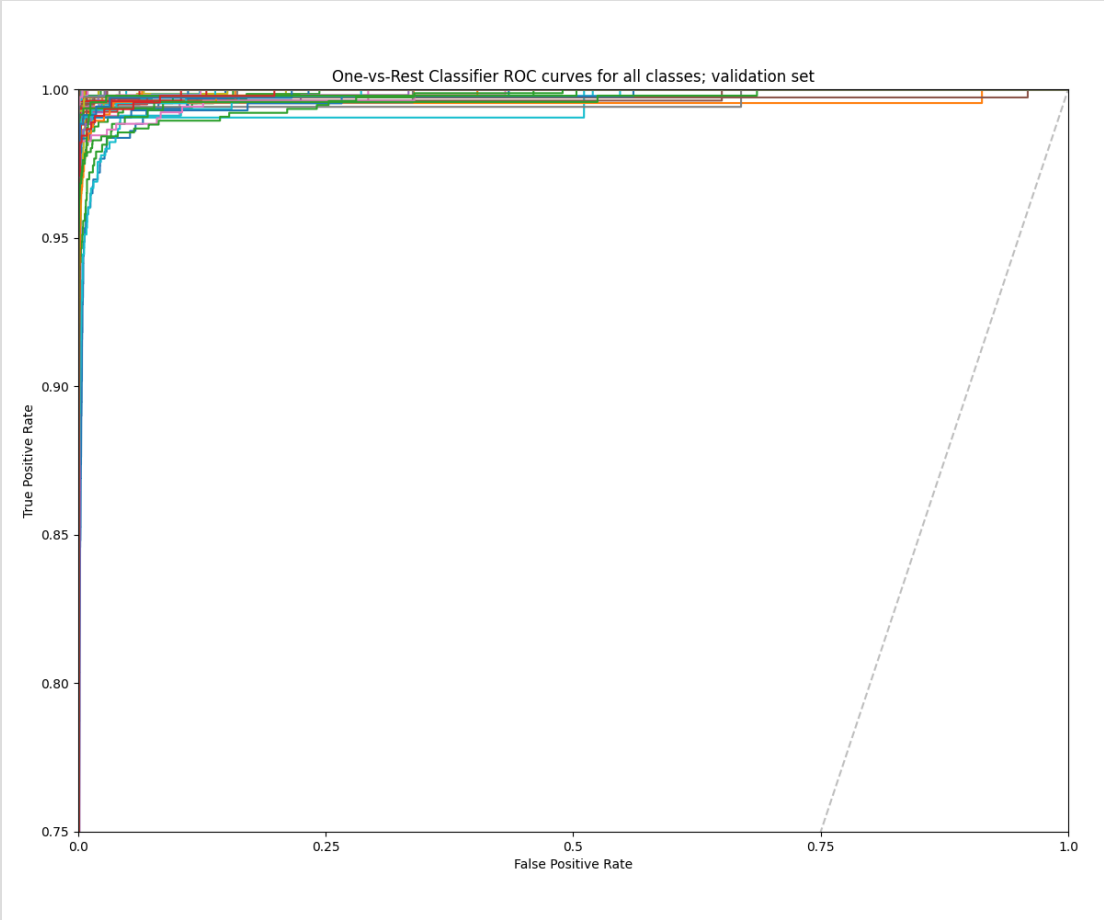
	precision	recall	f1-score	support
accuracy			0.96	54000
macro avg	0.97	0.97	0.97	54000
weighted avg	0.96	0.96	0.96	54000

- 1vRC:

	precision	recall	f1-score	support
accuracy			0.98	54000
macro avg	0.98	0.98	0.98	54000
weighted avg	0.98	0.98	0.98	54000

ROC curves – validation (non-duplicated)





Future work

- A new project dealing with OCR data
  - better OCR for Slovenian data
- New training set currently in preparation
  - a lot more manual work
  - oversampling and undersampling
  - better care about very small classes
- Planning to change the pipeline:
  - steps better organized to classify levels

Classifications of ISCO and NACE

- Data from the Labour force survey
- Classification of answers in the survey
  - occupation
    - very bad quality
    - grammar mistakes
    - no answers/deliberate errors
  - company economic activity
  - name of occupation/ec. activity and its description
- Data is linked to the Statistical Register of Employment
  - what can't be linked is filled manually ← **this part is to be automated**

Extensive pre-processing

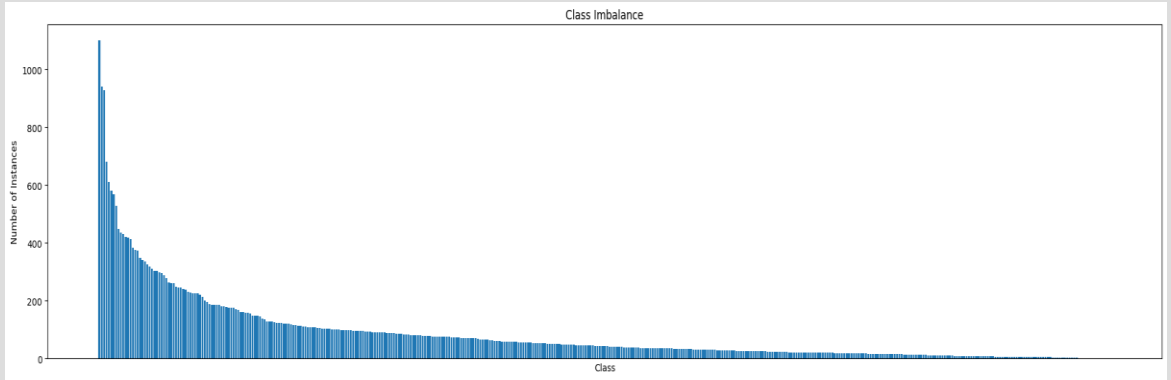
We focused mostly on ISCO

- Special characters
  - different encodings when reading data → a lot of mistakes
  - standardization and cleaning
  - to lower case
- Concatenating names and descriptions
- Count vectorizer, Word2Vec
- Eliminated duplicates
- Eliminated classes with <40 obs.

Analysis of known data

- Only observations with known ISCO codes

	skp1_4	orig_pok11	orig_pok12	opis
0	4120	administratorka	administrativna dela	administratorka administrativna dela
1	2621	kustos	galerijska dejavnost, vodi obiskovalce po muze...	kustos galerijska dejavnost, vodi obiskovalce ...
2	2342	vzgojiteljica predšolskih otrok	vzgojiteljica	vzgojiteljica predšolskih otrok vzgojiteljica
3	3359	svetovalka	pomoč in svetovanje strankam	svetovalka pomoč in svetovanje strankam
4	3113	dežurni električar na operativnemu	izvajanje del za obratovanje in	dežurni električar na operativnemu pogonu




```
skp1_4
5223    1100
3322     942
9329     928
4120     682
4321     611
        ...
8182       6
2641       6
7319       6
7317       6
7514       6
```

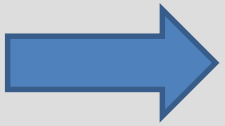
Training

- Using cross-validation
- Different hyperparameters
- Different model families:
  - Logistic regression
  - Decision Tree
  - Random Forest
  - k-Nearest neighbours

- Results not satisfactory: accuracy between 41.38% and 55.83%

The other approach

- Using *huggingface.co* transformers
  - Only on occupation names
  - BertForSequenceClassification, AdamW
    - BertTokenizer
- To lower case
- Removing commas and multiple white spaces
- Removing Slovenian stopwords (nltk)
- Train-test split at 80%:20%



Accuracy still staying low

- Training
  - With Adaptive Moment Estimation with weight decay (AdamW)
  - 5 epochs fine-tuning

```
Epoch 1: Training Loss: 4.3483, Validation Loss: 3.4463, Accuracy: 0.4250
Epoch 2: Training Loss: 3.1925, Validation Loss: 2.8571, Accuracy: 0.4886
Epoch 3: Training Loss: 2.7262, Validation Loss: 2.5532, Accuracy: 0.5372
Epoch 4: Training Loss: 2.4413, Validation Loss: 2.3741, Accuracy: 0.5506
Epoch 5: Training Loss: 2.2419, Validation Loss: 2.2458, Accuracy: 0.5612
```

Future work

- A new training set is being prepared
  - ISCO has been updated recently
  - synthetic examples for small classes
    - using Generative AI
- Trying Bert again, but with descriptions
- Maybe some other models (Zero-shot, but unlikely to work well)

Other generative AI classifications

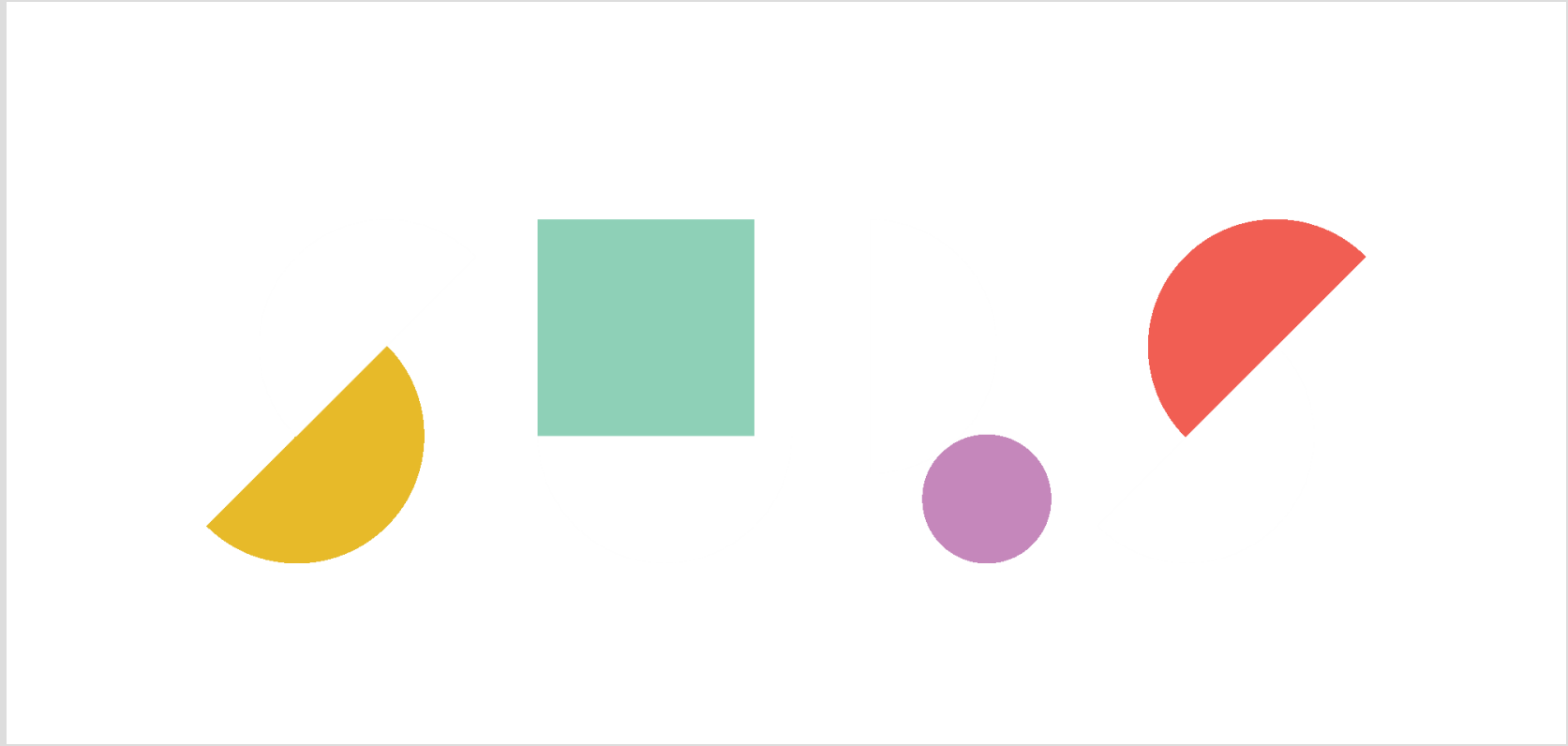
- *OJAs:* collection of specific information
  - publishing dates
  - deadlines
  - number of vacancies advertised

← already in the works

- *Computer equipment scraping:* collecting information about article characteristics and standardization of variables
- *Holiday offers:* clustering of similar offers
- Expand shortened words in receipt and scanned articles

} ideas

REPUBLIKA SLOVENIJA
STATISTIČNI URAD



Thank you for your attention!