Topic #	P01
Domain	ML on Tabular Data
Title	**P01 - Improving the quality of museums data**
Description	Among European countries, Estonia has the most museums per 100,000 inhabitants. You can visit them physically or take a virtual trip via the Estonian museums information system to explore the abundant choice of collections. The Estonian museums information system MuIS is the online gateway to Estonian museums, giving everybody an opportunity to study the collections of the institutions that have joined the system. The information system contains almost 4 million objects, but not all of the objects are labeled ideally. Manual improvement of such a state would take years. Thus, we ask you to help us by creating a smart model to fill in the gaps in the data based on existing descriptive variables. Let's improve the Estonian museum's information system and preserve Estonia's cultural heritage altogether! The goal of this competition is to create a model to predict the type of museum object based on its descriptive variables. Link to the competition: https://www.kaggle.com/t/720d66660aef45a2af496c279664e382
Is data available	Data is already available
Contact person	Kristjan Eljand (**is not ready** to mentor teams)
Organization	ANDMETEADUS OÜ
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium

Topic #	P02
Domain	Computer Vision
Title	**P02 - Object recognition for The Image Bank of Tartu**
Description	Background Tartu or as it is also known as “The City of great thoughts” is the second most populated city in Estonia. It is famous for its university, rich history and aspiration to innovate and make citizens’ lives better. Tartu has its own image bank, which contains 134 000 images divided into folders by general topics. Unfortunately, images are not systematically labeled, which makes searching for the right picture complicated. The accurate prediction model would help Tartu to make the public images searchable and reusable. Goal The goal of this competition is to label the images from the image bank of the city of Tartu. Link to the competition: https://www.kaggle.com/t/4cb972a6bd3e4e5c999a657063cef506
Is data available	Data is already available
Contact person	Kristjan Eljand (**is not ready** to mentor teams)
Organization	ANDMETEADUS OÜ
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium

Topic #	P03
Domain	Natural Language Processing
Title	**P03 - Multi-intent detection in customers' requests**
Description	Background Remember the last time you had to interact with the government? Was it easy, fast, and enjoyable? In Estonia, your answer would most probably be YES! In the government, we tend to put people's needs first and offer public services in the most accessible way. Starting from the early 90s we've chosen the path of e-development and during the last 30 years, there has been a huge leap, which allowed Estonia to become one of the most digitally advanced states in the world. Nowadays we want not only to rely on and cherish our previous success but also to continue moving further in offering top-tier solutions to our citizens. Today we are developing our own virtual assistant Bürokratt which already slowly is and soon will continue to revolutionize the way people interact with a government. Bürokratt allows citizens to contact the government as a whole, erasing borders between different authorities and reducing bureaucracy to the max. When developing Bürokratt our number one priority is user-friendliness. We want to give people an opportunity to solve all the problems with one conversation with Bürokratt. That requires Bürokratt to understand multiple intents, which it is still learning to do. Take part in the competition and help Estonia's virtual assistant Bürokratt to understand multiple intents by creating an ML model for multi-intent detection in the Estonian language. Goal The goal of this competition is to create a model that can detect multiple intents from user queries sent to the Estonian government by its citizens. Link to the competition: https://www.kaggle.com/t/b3f4601206cf49f992f22d6d88607787
Is data available	Data is already available
Contact person	Kristjan Eljand (**is not ready** to mentor teams)
Organization	ANDMETEADUS OÜ
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium/High

Topic #	P04
Domain	ML on Tabular Data
Title	**P04 - Drinking Water Quality Prediction**
Description	Background Despite the fact that most of our planet is covered with water, not more than 3% of this amount is fresh. To make sure that the water is safe to drink the Estonian Health Board has been measuring its quality in more than a thousand water stations across the country thereby making sure that every citizen will get the freshest water right from their tap. To bring water quality measurement to the next level and automate the working process of Estonian water inspectors, we would like to invent a predictive water quality model that would enable us to prioritize the tests or react proactively to the deterioration of the water conditions. Therefore, enhancing the role of scientific and data-driven approaches on a governmental level. Goal The goal of this competition is to create a model that predicts the water quality in Estonian water stations based on the government's open data of the previous measurements. Link to the competition: https://www.kaggle.com/t/2bc38d2ef2ba4269a37374a5ef3a59c3
Is data available	Data is already available
Contact person	Kristjan Eljand (**is not ready** to mentor teams)
Organization	ANDMETEADUS OÜ
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium

Topic #	P05
Domain	Natural Language Processing
Title	**P05 - Skills detection from the text**
Description	Background Have you ever hesitated whilst answering the question “Who do you want to become?”. Statistics show that less than 10% of people know exactly what career path they would like to choose. However not only do workers struggle with finding the right position but HR specialists as well face difficulties when searching for a perfect match. One of the keys to success in choosing the right candidate is having their skills match the task they are going to fulfill. This process can be quite time-consuming and unfeasible for HR specialists due to the ever-changing nature of the underlying documents (professional standards, curricula, syllabuses, surveys, job advertisements, websites). The accurate skill detection model would result in big time saving for analysts, school principals, personnel specialists, career counselors, and for common people planning their own learning and career path. Goal The goal of this competition is to detect skills that correspond to the texts about the forestry/wood industry and metal/machine industry. Link to the competition: https://www.kaggle.com/t/87fa0e15b9cc4698a0c8722ac77eb408
Is data available	Data is already available
Contact person	Kristjan Eljand (**is not ready** to mentor teams)
Organization	ANDMETEADUS OÜ
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium

Topic #	P06
Domain	Computer Vision / Instance Segmentation
Title	**P06 - Rooftop Instance Segmentation**
Description	Our company is working on a project related to building a deep learning model for segmenting and classifying objects from satellite images [A side note from Dima: this company is well-known in Ukraine. They have a good reputation, hence the project although not very broadly described, should be interesting. It might be very challenging, so mind your step.]
Is data available	Data is available online, but it needs to be fetched or scrapped
Contact person	Danylo Bondar (ready to provide substantial mentorship)
Organization	SoftServe
How many teams they are ready to supervise?	**Only 1** team can participate
Approx. Complexity	Hard+

Topic #	P07
Domain	ML on Tabular Data / Feature Selection
Title	**P07 - Learning from small amounts of molecular data**
Description	The project is from the area of nuclear magnetic resonance (NMR) and chemoinformatics, using machine learning. It can be tackled without a specific background in chemoinformatics. It seems that newer molecular machine learning models need large quantities of data [1]. There was a publication recently doing some prediction with low amount of data [2]. This could be used to predict other properties from small sample sizes. I have at least two applications for this. This project would involve looking at the code from [2] (which we have access to) and adopting it for other tasks. Also, a general survey on the question of "how many data are needed" would be part of this work. The main work is to find out the optimal set of features to use for a specific problem. Also, improvements in the architecture might be possible and would be welcome. The project offers the possibility to work with existing models (so there is no need to work from scratch, which might be difficult for students), but offers the possibility to try out various options and understand how they influence the results [1] S. Kuhn, R. M. Borges, F. Venturini and M. Sansotera, "Dataset Size and Machine Learning - Open NMR Databases as a Case Study," 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC), 2022, pp. 1632-1636, doi: 10.1109/COMPSAC54236.2022.00259. [2] Markus Fischer, Benedikt Schwarze, Nikola Ristic, Holger A. Scheidt, Predicting 2H NMR acyl chain order parameters with graph neural networks, Computational Biology and Chemistry, Volume 100, 2022, 107750, ISSN 1476-9271, https://doi.org/10.1016/j.compbiolchem.2022.107750.
Is data available	Data is already available
Contact person	Stefan Kuhn (ready to provide substantial mentorship)
Organization	University of Tartu
How many teams they are ready to supervise?	**Only 2-3** teams can participate There might be a small prize for the best team (TBD).
Approx. Complexity	Medium/Hard

Topic #	P08
Domain	Computer Vision / Reinforcement Learning
Title	**P08 - Teaching 1:10 scale cars to drive in a coty**
Description	Participation in ADL self-driving challenge: https://courses.cs.ut.ee/t/DeltaXSelfDriving/. Ardi made a small presentation in our lecture: https://panopto.ut.ee/Panopto/Pages/Viewer.aspx?id=235741cc-c9c7-4535-8816-af1c00d11a59.
Is data available	Data is not freely available, one might need to perform additional work to get it
Contact person	Ardi Tampuu (ready to provide some mentorship)
Organization	University of Tartu
How many teams they are ready to supervise?	As many teams can participate as would wish There might be a small prize for the best team (TBD).
Approx. Complexity	Hard

Topic #	P09
Domain	Computer Vision
Title	**P09 - Document classification**
Description	Language-agnostic document classification using Siamese networks – we have use case, where we have different type of documents that are in different (might be less spoken) languages. Documents should be classified based on the document type (agreement, diploma, ID, court decision etc). Unfortunately, there are only a few example documents and every such document has some different format
Is data available	Data is available online, but it needs to be fetched or scrapped [Dima: collecting data could be problematic here].
Contact person	Lennart Kitt (ready to provide some mentorship)
Organization	AS SEB Pank
How many teams they are ready to supervise?	**Only 2-3** teams can participate
Approx. Complexity	Hard

Topic #	P10
Domain	Computer Vision
Title	**P10 - Customer type classification**
Description	From time to time we need to understand if payments are done to private individuals or companies. If payments go to other banks, there are no attributes, what type of receiver is counterparty. Though such kind of “flag” could be trained based on a synthetic dataset consisting of names of privates and companies.
Is data available	Data can be artificially generated [Dima: data is not available, which can be a problem].
Contact person	Lennart Kitt (ready to provide some mentorship)
Organization	AS SEB Pank
How many teams they are ready to supervise?	**Only 2-3** teams can participate
Approx. Complexity	Hard

Topic #	P11
Domain	Time Series Analysis
Title	**P11 - Maksekäitur / payment processor**
Description	Ettevõte esitab klientidele arveid. Arvetel on maksetähtaeg. Kliendid tasuvad arveid kes kuidas. Mõni õigel ajal (maksetähtajal), mõni varem, mõni hiljem. Vaja on kliendi ajaloolise maksekäitumise põhjal tekitada mudel, et sellega ennustada, millal konkreetne klient oma järgmise(d) arve(d) võiks tegelikult tasuda. Õhtu lõpuks võiks olla rakendus, mis suudab kõigile hetkel võlas olevatele arvetele üle kogu kliendibaasi ennustada arve laekumise kuupäeva. [Google translate to ENG]: The company issues invoices to customers. Invoices have a due date. Customers pay bills who and how. Some on time (payment deadline), some earlier, some later. It is necessary to create a model based on the customer's historical payment behavior to predict when a particular customer might actually pay their next bill(s). By the end of the evening, there could be an application that can predict the date of receipt of the invoice for all currently outstanding invoices across the entire customer base.
Is data available	Data is already available
Contact person	Kristen Pugi (ready to provide some mentorship)
Organization	Columbus Eesti AS
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Hard

Topic #	P12
Domain	Computer Vision
Title	**P12 - Foxway Trade-In Solution**
Description	We acquire devices from multiple sources, one of which is trade-in. We have run in-store trade-in flow for quite some time now, where we provide technical solution as well as trainings to different retailers, authorized resellers and mobile network operators. With the world changing, more and more trade-ins are done online by the end customer him/herself. Our idea is to provide an omnichannel, easy to use solution for exactly that flow. In addition to being convenient and easy to use, we also want to give the potential customer as accurate pricing as soon as possible - since it is based on the model of the device as well as its condition, we think that the best approach here is using AI/ML to. First of all, the solution must recognize the model of the device, whether it is by analyzing pictures the customer has taken, or simply reading the “About” info from picture provided by the customer. In addition, the condition of the device should be handled automatically aswell, to avoid discrepancies in grading between ourselves and the end customer, to yet again, provide as accurate pricing. The final part is crucial since grading is subjective and compared to our in-store flow, we are unable to train every single prospect to find the defects and grade accordingly. To conclude, the end-customer provides images of the phone he/she wants to trade in, preferrably taken at angles guided by the application itself to ensure everything is covered, and to make it easier and more failsafe to analyze. For building that solution, we already have thousands of images, most of them labeled with what defects they are having.
Is data available	Data is available
Contact person	Marek Rüütli (ready to provide substantial mentorship)
Organization	Foxway OÜ
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium

Topic #	P13
Domain	Time Series Analysis
Title	**P13 - Journey to zero - Predict electricity consumption**
Description	Electricity prices have skyrocketed and consumers around the world are looking for options to reduce electricity costs and environmental footprint. Enefit is one of the largest energy companies in Baltic countries and we would like to help our consumers as much as possible on their journey to zero. Both electricity cost and the environmental footprint could be drastically reduced by forecasting the consumption of the household and optimizing its energy usage (controlling smart energy devices in such a way that minimizes the cost and environmental footprint of the consumption). The goal of this competition is to create an energy consumption prediction model for a single household. Accurate household-level predictions are a critical prerequisite for more sustainable energy usage! The link to the Kaggle competition: https://www.kaggle.com/competitions/predict-electricity-consumption
Is data available	Data is already available
Contact person	Kristjan Eljand (ready to provide some mentorship)
Organization	Eesti Energia
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium/Hard

Topic #	P14
Domain	ML on Tabular Data
Title	**P14 - Classification of Tartu's street network.**
Description	The goal of the project would be to classify the Tartu City Government road dataset into 5 hierarchical categories, based on various parameters - width, speed limit, cover, etc., which we already have in the data layer or which the students themselves identify from orthophotos with image recognition. As a secondary task, which is not however, a mandatory part of the project is to additionally identify cars in informal parking lots if possible (by the roadside). We have the structured base data as spatial data (line layer) with metadata that we share as an extract from a shape file or access to our environment. The result should be in a form that can be linked to our existing data layer.
Is data available	Data is available for internal use, but we can share access to data
Contact person	Tiina Arras (ready to provide substantial mentorship)
Organization	Tartu Linnavalitsus / Tartu City Government
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium

Topic #	P15
Domain	Time Series Analysis (?)
Title	**P15 - Ψ - BOT**
Description	Autonomous intelligent trading agent: the aim is to code a bot that will apply min-max chess algorithm (or any other relevant search algorithms) in the context of trading cryptocurrencies or other financial instruments. Data is freely available and accessible: https://towardsdatascience.com/ https://data.world/datasets/cryptocurrency https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory https://www.cryptodatadownload.com/ https://p.nomics.com/cryptocurrency-bitcoin-api Keeping myself available for a more in-depth discussion.
Is data available	Data is already available
Contact person	Federico Martinazzi (ready to provide some mentorship)
Organization	CHANGEHOLDING OU
How many teams they are ready to supervise?	As many teams can participate as would wish
Approx. Complexity	Medium