# D6 report

Markus Roletsky, Rene Puusepp, Alina Gudkova

## Task 2. Business understanding (0.5 point)

### Ärilised eesmärgid

Leidsime huvitava andmestiku 94110 pildiga 141-st puuviljast, köögiviljast ja pähklist, mida saaks kasutada, et treenida korralik klassifitseerija, mis eristaks neid vilju. Pildid on väga kontrollitud/spetsiifilisel kujul ning see teeb sellise mudeli kasuliku rakendamise küllalt keeruliseks. Pildid on saadud roteeruva aluse peal vilja filmimisest, keskmiselt on 500 pilti vilja kohta. Piltidelt on ära lõigatud taust ning suurus taandatud 100x100 pikslile. Lugedes andmestiku kommentaare jäi koheselt silma trend, inimesed ei saanud selle andmestikuga tulemusi kui mudel pilte  uutest viljadest sai (overfitting?). Ja seda reeglina väga “hea” mudeli korral kui hinnata test hulga põhjal. Siin on mõningad näited:

“The model gained 98% accuracy on training and 92% on validation still gives very wrong predictions on new test images.” -Yash Khandelwal
“Fruit-Prediction-Accuracy 98% and test>92% but not good results after deployment“ -Adyaan Singh
“Keras model with 96% accuracy but wrong prediction” -Jim Hung

Khandelwal sõnastas tegelikult probleemi juure väga hästi: “ I realized this dataset has a flaw. All the images belonging to a particular class is of same type(white background and same piece of fruit)”. Mis tähendab, et variatsiooni sama klassi fotode seas on vähe, eriti arvestades, et kõik ühe klassi pildid on ühest ja samast füüsilisest viljast.
Sealt ka naturaalne küsimus: kuidas me saaksime treenida mudelit paremini toime tulema just uute piltidega?

Meie eesmärk on saavutada paindlikum mudel, ehk katsetada, mida veel annab teha, et saavutada üldistatuse tase, millel oleks praktiline kasutus. Selleks mitte ainult masinõpet kruttides vaid ka andmeid ennast, peamiselt, sest nendest tundub tulenev senini tehtud mudelite nõrkus.

Eesmärk on treenida andmete peal mõni klassikaline mudel ilma suurt modifitseerimist tegemata, ning siis hakata muutma kas treeningandmeid, testandmete esituse kuju, klasside kogust, ja treenimise/testimise osakaalu, mudelit ennast, selleks et saavutada parem tulemus andmestikule võõraste sisendite puhul. Igasugune parandus ja selle lahtimõtestamine on meie kontekstis võit.

# Olud

Projektiks on meil andmestik ise(94110 pilti), kolme tudengi tasuta tööjõud, paar päeva eraldatud vaba aega, 3 arvutit, pilvest eraldatud vahemälu Google colab keskkonnas (3x12gb), TÜ LTAT.02.002 aine materjalid ning interneti avarused. AI kasutus plaanis pole aga Colabi notebooks keskkonda on sisse ehitatud Gemini ning selle baasil ka Google-i Code Generation, sealjuures tasub märkida, et code generation-i kasutusnõuetes on keelatud kasutada AI-d masinõppeks.

Esmane eeldus on, et Code Generation-i piirang kehtib vaid tulu teeniva kasutaja kohta ja mitte õppetööks, jätame esialgu funktsionaalsuse sisse. Nagu ennist mainitud, AI kasutus võib olla piiratud nende enda kasutustingimustest.

Kuna me tegeleme rangelt pilditöötlusega siis on piiratud meie arvutusvõime ning funktsionaalsus ka pilditöötluse valdkonnas, kuna meil pole tasulist automatiseeritud fototöötluse võimekust. Selles valdkonnas säilitame endale paindlikkuse projekti implementeerida uusi teeke või tööriistu kui leiame, et on vaja. Meile allutatud vahemälu on lagi, rohkem juurde osta pole plaanis nii et see võib mõjutada kui suure osa andmetest me tegelikult treenimiseks kasutada saame ja seda omakorda mõjutab kui palju ebaefektiivsusi koodis endas saab olema.

Sellest tulenevalt põhi risk on mälu otsa saamine ja treeningu mahu tagasi tõmbamine, selleks peame veenduma koodi efektiivsuses. Mälu kasutus on mõjutatud ka kindlasti pilditöötluse vahenditest, mida kasutame, mis on esialgu otsustamata.

Töö jooksul viitame treenitud masinõppe mudelile kui lihtsalt “mudel” olenemata mudeli tüübist või treeningu metoodikast. Kuna meil on plaanis kahte mudelit võrrelda siis viitame teisele mudelile kui “kohandatud mudel”. Meie andmestikus on 141 erinevat toiduaine liiki, millele viitame edaspidi kui lihtsalt “vili”, vili võib olla puuvili, köögivili või pähkel. Kuna paar eri liiki võivad kuuluda nt ühe puuvilja alla (nt kaks eri õuna liiki) siis kohtleme neid eraldi viljadena, kuni me ei otsusta klasside arvu kohandada ja vähendada, et paremini mudeliga üldistada.

Kuna tööjõud on tasuta siis saame kulude alla arvestada ainult aega, arvutusvõimet ning tudengite närvirakke. Arvutusvõime on meil sisuliselt fikseeritud ja piiratud, ajakulu saame ise tähtajani korrigeerida, närvide osas oleme säästlikud, kui miski ei tule välja siis läheme ringiga või kasutame üksteise abi. Igasugused deviatsioonid projekti algsest plaanist saavad olema dokumenteeritud.

Paremini üldistatud mudeli loomine mitte ainult ei anna meile akadeemilisel maastikul vajalikud vitamiinid, et aine lõpetada, vaid annab uhkustamise õigused kaggle-is ning piisavalt hea mudeli korral ka reaalse praktilisuse mudeli kasutamiseks viljade klassifitseerimiseks piltidelt. See omakorda avaks võimalused luua programm, mis töötleb lihtsad kaameraga võetud pildid mudelile söödavaks ning annab mugava lahenduse hinnata toiduaine liike. See aitab eristada liike visuaalsete omaduste põhjal, kuna keskmine inimene on seni sunnitud vaid poes silti uskuma, et tegu on mingi kindla liigiga. Mudelil võib olla rakendusi ka toidutööstuses automaatse viljade eraldamise stsenaariumis.

## Andmetöötluse eesmärgid

Peamine eesmärk on leida lahendusi andmestiku kitsaskohtadele ning selle abil parandada andmete rakendusvõimet. Selleks on plaan proovida kaht kuni nelja lahendust, mille kasutamise me otsustame eelnevate efektiivsuse järgi, lahendusi mõtleme mitu, juhuks kui esimesed ei peaks mõjutama täpsust meile olulisel määral.

Lõplikku rakendumisvõimet hindame kahel eri mudelil andmestiku välise pildimaterjaliga, eesmärk on kohandada parema täpsusega mudel just andmestiku väliste andmetega paremini töötamiseks.

# Task 3. Data understanding (1 points)

## Gathering data

**Andmenõuete määratlemine**

Projekti eesmärk on teha pildiklassifikatsioon puuviljade tuvastamiseks. Selleks on vaja andmestikku, mis sisaldab märgistatud puuviljapilte, et treenida ja hinnata masinõppemudelit. Olulised atribuudid on piltide tunnused ja nendele vastavad sildid, mis näitavad puuvilja tüüpi.

**Andmete saadavuse kontrollimine**

Andmestik on avalikult kättesaadav Kaggle'i lehel "Fruits 360 dataset" nime all. See sisaldab ulatuslikku komplekti märgistatud puuviljapilte, mis sobivad klassifikatsioonitöödeks. Andmestik on organiseeritud eraldi kaustadesse treening- ja testimisandmete jaoks, kus iga kaust esindab kindlat puuviljaklassi.

**Valikukriteeriumite määratlemine**

Kasutame kõiki andmestiku 100x100 pilte, et tagada puuviljaklasside tasakaalustatud esindatus treening- ja testandmetes. Katsetame ka võimalusi originaal resolutsioonis piltidega.

## Describing data

Andmestik sisaldab pilte 141 puuviljaklassist, kokku üle 90 000 pildi.

- Treeningandmestik: Ligikaudu 70 000 pilti, mis on jaotatud erinevate puuviljaklasside vahel.
- Testandmestik: Umbes 23 000 pilti, mida kasutatakse mudeli jõudluse hindamiseks.
- Pildi formaat: Kõik pildid on .jpg formaadis.
- Lahutusvõime: Pildid on standardiseeritud suurusega 100x100 pikslit, on ka olemas originaal resolutsioonis pildid, kui keskendume rohkem 100x100 piltidele.

- Sildid: Iga pilt asub kaustas, mis vastab selle klassi nimele, ja toimib seega maatriksina klassifikatsiooni jaoks.

## Exploring data

**Klassijaotus**
Andmestik näib olevat tasakaalus, iga klassi kohta on sarnane arv pilte. See tasakaal tagab erapooletu mudeli treenimise ja hindamise.
**Visuaalne uurimine**
Juhuslik valik pilte näitab puuviljade välimuse mitmekesisust värvi, kuju ja tekstuuri poolest, mis kinnitab andmestiku rikkust. Mõned puuviljad võivad siiski visuaalselt kattuda, mis võib klassifikatsiooni keerulisemaks muuta.
**Dimensioonid**
Kuna pildid on viidud suurusele 100x100 pikslit, on igal pildil 10 000 pikslit, kusjuures igal pikslil on kolm värvikanalit (RGB).

## Verifying data quality

**Täielikkus**
Andmestik on täielik – pole puuduvaid andmeid ega väärtuseid. Iga pilt on seotud vastava klassisildiga.
**Järjepidevus**
Pildid on ühtlase suurusega ja märgistatud. Kaustade struktuur tagab lihtsa seose piltide ja siltide vahel.
**Terviklikkus**
Esmase ülevaatuse käigus ei tuvastatud duplikaat- ega rikutud faile. Kõik failid on juurdepääsetavad ja loetavad.
**Tuvastatud väljakutsed**
Mõned puuviljad võivad välimuselt olla sarnased (nt mõned õunad ja virsikud), mistõttu tuleb valida sobiv mudel ja häälestada hüperparameetreid, et parandada klassifitseerimise täpsust.

Selle protsessi käigus on andmestikku põhjalikult uuritud ja olulisi kvaliteediprobleeme ei tuvastatud. Kõik 141 puuviljaklassi jäävad, et ehitada usaldusväärne klassifikatsioonimudel.

# Task 4

## Plan

**1. Exploring the dataset and preprocessing**

We need to understand the dataset structure and prepare it for training.

- Loading and exploring dataset, verifying file structure, class distribution and image quality.

- Possibly resizing and normalizing images.

- Splitting data into training, validation and test sets.

Estimated effort: 6 hours per team member.

**2. Baseline model development**

We want to train an initial image classification model to assess the feasibility of 50% accuracy to start with. Ideally achieving the accuracy of 80%.

- Närvivõrkude uurimine ja mudeli loomine.

- Training using the original dataset.

- Evaluating performance on the validation set.

- Experimenting with different architectures and hyperparameters, fine-tuning it.

Estimated effort: 10 hours per team member.

**3. Background editing**

Next step is to diversify the dataset by altering image backgrounds.

- Replacing backgrounds with different images.

- Ensuring that modified images retain class-label integrity.

Estimated effort: 3 hours per team member.

**4. Enhanced model training**

After adding new images we want to retrain the model with the updated dataset to improve generalization.

- Integrating the augmented dataset.

- Retraining and fine-tuning the model.

Estimated effort: 8 hours per team member.

**5. Custom test dataset collection**

- Collecting and preprocessing images of local fruits.

- Resizing and normalizing these images.

- Creating a test set for evaluation.

Estimated effort: 6 hours per team member.

**6. Generalization testing**

After that we need to evaluate our model performance on the new custom test dataset.

- Making predictions on the custom test set.

- Calculating performance metrics and identifying failure cases.

- If needed, fine-tuning the model.

Estimated effort: 6 hours per team member.

**7. Report and presentation**

Summarizing findings and presenting the project.

Estimated effort: 6 hours per team member.

## Methods and tools

Programmeerimine: Python (TensorFlow/Keras, NumPy, Pandas)
Visualiseerimine: Matplotlib, TensorBoard
Dokumentatsioon: Microsoft Word, Powerpoint