AstraZeneca

AstraZeneca Hackathon

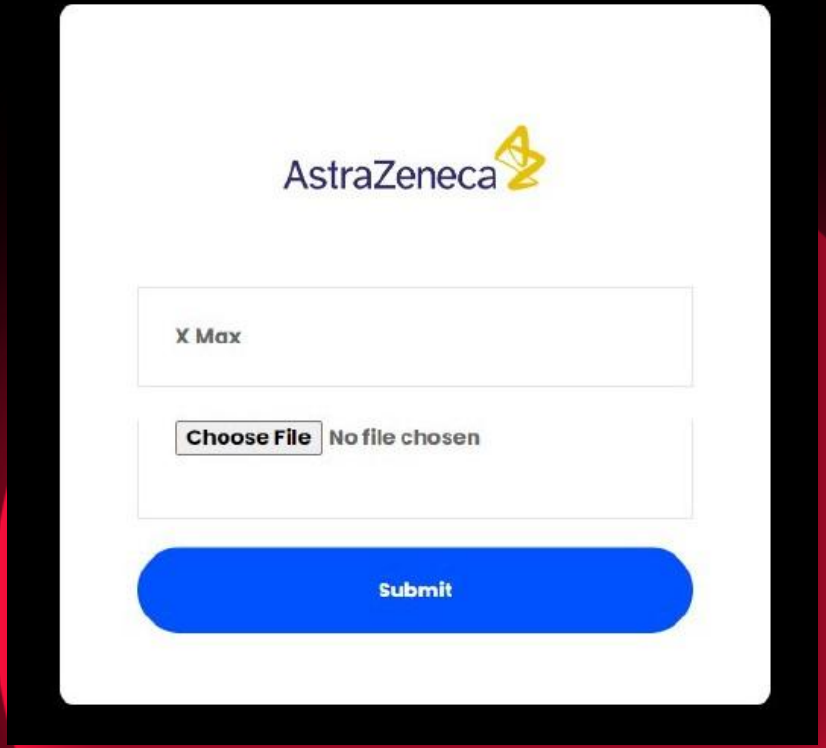
Theme: **Plot Digitizer for KM Plots using Computer Vision and ML**

Team Appendly

- Rajasekar M, Amrita Vishwa Vidyapeetham
- Sai Adarsh S
- Akash C

AstraZeneca

- Reverse Engineer Kaplan Meier curves using Computer Vision and ML.
- Automate Data Extraction.
- Export data points.



Motivation

AstraZeneca

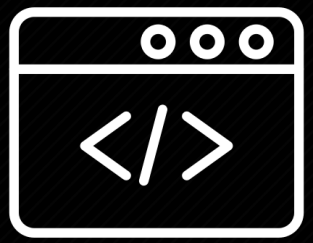
- Difficult, inefficient and time consuming to manually extract data points from a curve.
- No specific algorithm, webapp that caters to the needs.
- The scope of the project is not limited to just Kaplan Meier plots but any 2D line graphs in general.
- Source: https://drive.google.com/file/d/1jPX1pT_wpwc0fyV-bi5bGvEQOvZeYSiq/view

Problem Statement

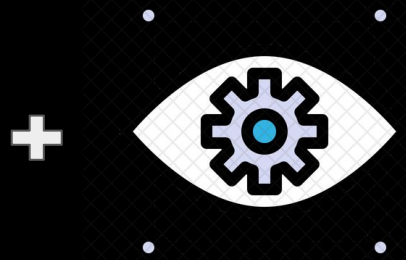
- Inaccuracy, inefficiency, time consumption in extracting probabilistic raw data points from KM curves manually
- Reverse engineer Kaplan-Meier (KM) plots from scientific literature and digitize them into probabilistic raw data points.

Our Solution

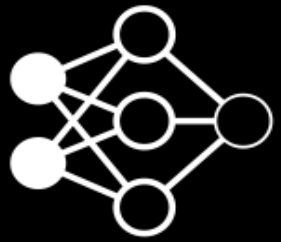
AstraZeneca



- Upload the plot to the web app with max x values and plot image



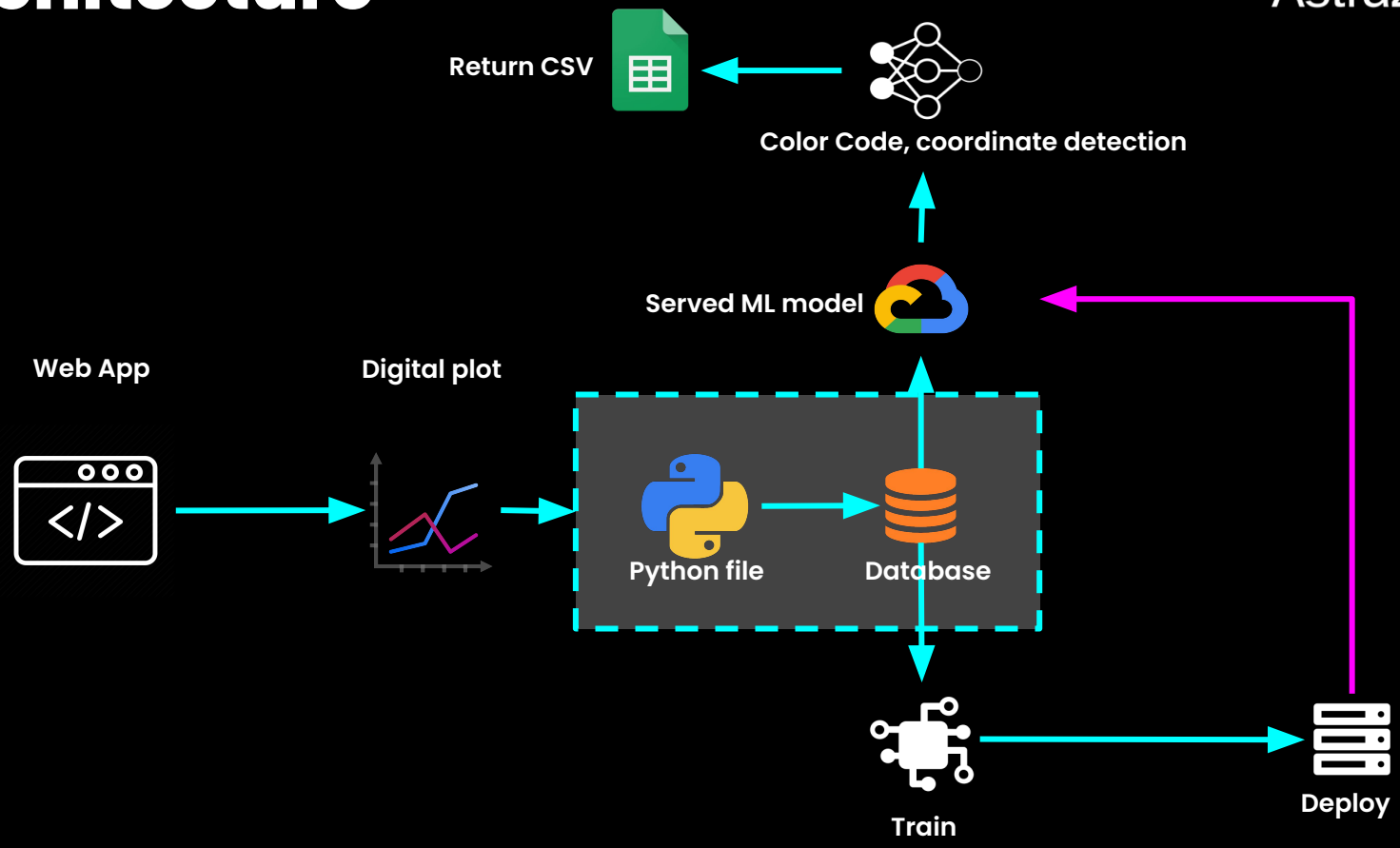
- Use OpenCV with an in-built algorithm tailored to detect the abscissa and ordinate.
- Crop image based on coordinates of the axes.
- Convert image to a PNG file devoid of white background.



- Use Color Thief, Web Colors to identify the colors present in the plot and their respective coordinates.

- Transform coordinates into Y axis probability points and X axis time intervals.
- Append values to a Pandas DataFrame and export file with respect to colors.
- Return CSV file.

Architecture



Tech Stacks

AstraZeneca

Flask
web development,
one drop at a time

scikit
learn

HEROKU

pandas

OpenCV

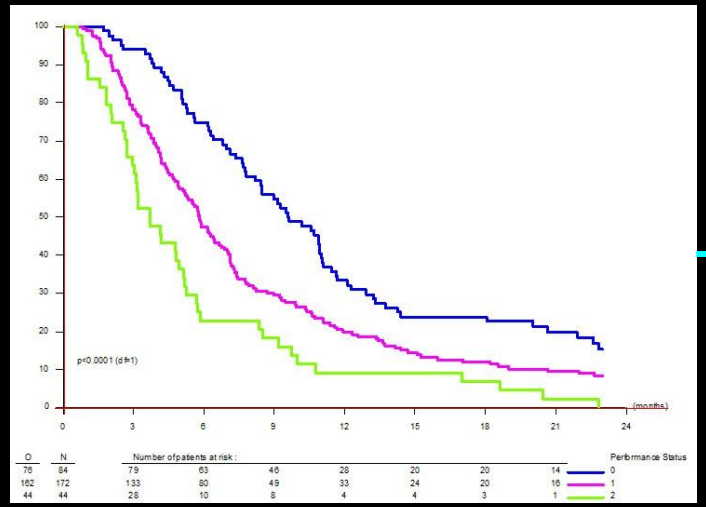
CO

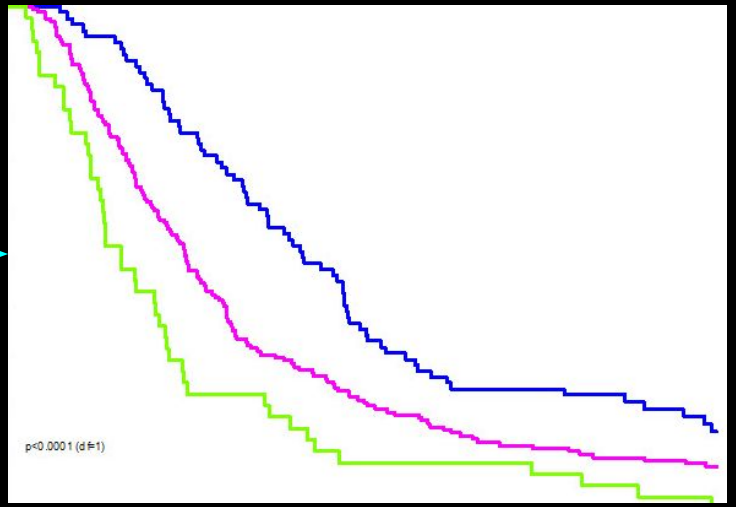
Google
Sheets

Process Visualized (1/4)

- **Live WebApp Demo:**
- **https://astrahacks-appendly.herokuapp.com/**
- **The user then has to upload the KM plot image along with the X axis max value.**
- **This redirects to another route that allows user to download the returned CSV file.**

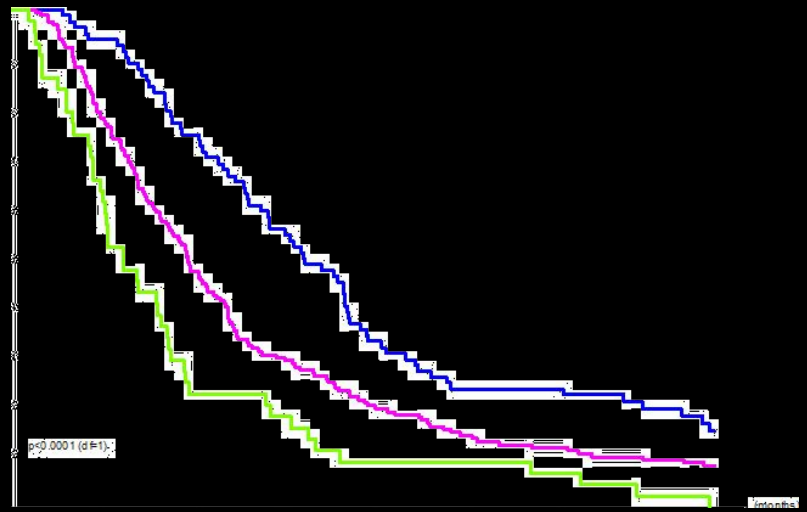
Process Visualized (2/4)





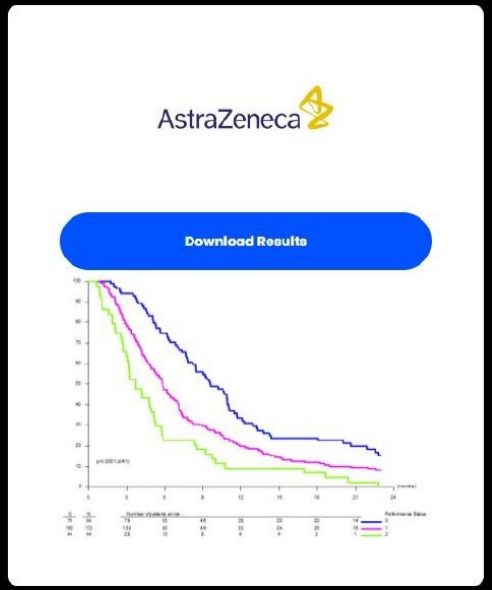
- The graph is cropped based on the red lines superimposed on the axes obtained from a pretrained algorithm which identifies axes from the set of lines detected using OpenCV Hough Line Transform method.

Process Visualized (3/4)



- The cropped image is converted into a png file with the white pixels removed.
- The color codes of the png file are identified using Color Thief and web colors library.
- The coordinates are then associated and segregated based on the color codes.
- The values are added to a keyed list which is then dumped into an excel file.

Process Visualized (4/4)



	A	B
1	X axis value	Y axis value
2	0	0.99
3	0.8	0.97
4	0.8	0.96
5	0.8	0.96
6	0.8	0.95
7	0.8	0.94
8	0.8	0.93
9	0.8	0.93
10	1.6	0.84
11	2.41	0.75
12	3.21	0.56
13	3.21	0.56
14	3.21	0.55
15	3.21	0.54
16	3.21	0.53
17	3.21	0.53
18	3.21	0.52
19	4.01	0.48
20	4.81	0.43
21	4.81	0.42
22	4.81	0.42
23	4.81	0.41

lime | blue | magenta

- The excel file gets downloaded once the user clicks the “Download Results” button.
- The sheets are segregated based on color codes where each color represents a group

Key Features

AstraZeneca

- **Handle Edge Cases:**
  - Get raw data points from digital plot with one click.
  - Data points are automatically segregated based on K color curves and exported to CSV.
  - Extending support to all 2D line graphs.
  - Flask based Web app, API built to ease the process of getting the data points.

- **Find:**
  - Using OpenCV to find axes, X & Y, in a graph.

- **Read:**
  - Detect color code and coordinates of data points on the image and classify them based on colors.

- **Output:**
  - Generate CSV file with coordinates of each curve in a separate sheet with 85 - 90% precision (based on test cases).

- **Add-ons:**
  - A button on the web app to download the data as an excel file

Future Scope

AstraZeneca

- Automate the entire process by using OCR to find the max X value.
- More ways to export data. i.e. csv, JSon.
- Increase accuracy in finding the color codes.
- Train the model with more datasets to improve accuracy when finding the axes.
- Extend the scope of the project to different types of 2D graphs by not just limiting to Kaplan Meier curves.

Quick Summary

- Upload KM plot image to the webapp
- Give the max X value as input
- Axes detection and cropping
- Detect the color codes in the image and their respective coordinates
- Get output as an excel file with high degree of Accuracy

Reference Links

AstraZeneca

- https://towardsdatascience.com/kaplan-meier-curves-c5768e349479

  https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/

- https://drive.google.com/file/d/1jPX1pT_wpwc0fyV-bi5bGvEQOvZeYSiq/view
- https://medcraveonline.com/BBIJ/the-kaplan-meier-estimate-in-survival-analysis.html
- https://www.researchgate.net/publication/272195556_A_Survey_on_Hough_Transform_Theory_Techniques_and_Applications
- https://pypi.org/project/colorthief/

AstraZeneca

Thank you