C1: KAGGLE - BIRTHMARK CLASSIFICATION

Oskar Männik and Till Bösch

Task 1. Setting up

The Repository is public on GitHub. Link: https://github.com/OskarMannik/birthmark-classifier

Task 2. Business understanding

Background
Skin cancer is one of the most common cancers worldwide and early detection is the key to effective treatment. The diagnostic process for skin lesions often relies on a visual examination, which can be subjective and depends on the expertise of the dermatologist. By utilising the HAM10000 dataset - a comprehensive collection of dermatoscopic images - this project aims to develop machine learning models to improve diagnostic support and gain insight into risk factors.

Business goals
The main objective of this project is to develop a machine learning model capable of accurately classifying skin lesions based on image data. Beyond classification, the project also aims to gain meaningful insights from patient characteristics such as age or type of lesion to better understand potential risk patterns and initiate preventive measures.

Business success criteria
The project is considered successful if the classification model achieves an accuracy of at least 65 % and thus provides a reliable basis for supporting clinical assessments. Success also includes the detection of patterns in demographic or clinical characteristics.

Assessing your situation
The main thing in the inventory of resources is the dataset itself - Skin Cancer MNIST: HAM10000. It contains 10015 different dermatoscopic images of birthmarks. For the model training we will use Google Colab and if extra computational resources needed, we will contact HPC Center. The code will be written in Python. For the user interface, we will build just a simple HTML page. One requirement that we want to achieve is that the model should classify birthmarks with at least 65% accuracy. We assume that the dataset consists of real-world data and the images are correctly labeled. The main constraint for us is going to be time, as we do not have that much time to develop and test the model perfectly. One risk might be the GPUs. It

will take more time when we have to use HPC Center for the extra computational power. So that is a thing that we have to be prepared for.

Terminology:

- **Lesion**: A region of abnormal tissue.
- **Melanoma**: A serious type of skin cancer.
- **Benign**: Non-cancerous.
- **Histopathology**: Diagnosis and study of diseases of the tissues which involves examining tissues and/or cells under a microscope
- **Actinic keratoses**: A skin disorder that causes rough, scaly patches of skin (pre-cancer)
- **Basal cell carcinoma**: A type of skin cancer that most often develops on areas of skin exposed to the sun
- **Benign keratosis**: A type of benign (non-cancerous) skin tumor or growth
- **Dermatofibroma**: Common overgrowth of the fibrous tissue situated in the dermis
- **Melanocytic nevi**: Medical term for a birthmark (benign)
- **Vascular lesions**: Abnormal growths or malformations in the blood vessels

The costs for our project are minimal, only the time. On the other hand, this is a benefit - we are learning and getting our project done for the studyings. Another benefit will be for healthcare providers by helping them to assess the type of a birthmark.

Data-Mining Goals

This project has two primary data-mining objectives:

- Develop a combined classification model of two different models (images and metadata) that can distinguish between different skin lesion types with a minimum accuracy of 65%.
- Explore the dataset for patterns and correlations, such as identifying age-related risks for specific lesion types, to provide actionable insights that extend beyond classification.

Data-Mining Success Criteria

The success of the data-mining efforts will be measured by:

- Model Performance: Achieving a classification accuracy exceeding 65% across relevant diagnostic categories.
- Insightful Analysis: Identifying statistically relevant trends in the dataset, such as age or lesion type correlations, that can serve as a foundation for further research or clinical applications.

Task 3. Data understanding

Gathering data

Outline of the data requirements
The project requires a dataset of labelled dermoscopic images of skin lesions together with relevant metadata such as patient age, lesion type and diagnostic method. The images must cover the seven lesion types specified in the dataset to ensure that the model can be generalised effectively. Additional attributes such as lesion location and patient demographics are valuable for secondary insights.

Checking data availability
The HAM10000 dataset fulfils these requirements. It contains 10,015 labelled dermoscopic images, each linked to metadata including lesion type, patient age and methods for confirming diagnosis (e.g. histopathology, expert consensus). This dataset from Kaggle is publicly available for academic purposes, ensuring compliance with ethical and data protection standards.

Define selection criteria
Only high quality images with complete metadata are included in the analysis to ensure consistency. Cases without a confirmed diagnosis or with incomplete metadata (e.g. missing age or lesion ID) are excluded. In addition, the focus is on lesion categories that are sufficiently represented in the dataset to allow meaningful model training and validation.

Describing/Exploring data
The HAM10000 dataset contains over 10015 images of skin lesions along with metadata. Cases include a representative collection of all important diagnostic categories in the realm of pigmented lesions: Actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc). Most of the images have dimensions of 600x450 pixels and are in JPG format. They vary in brightness, contrast, and lesion size which might need steps like normalization and augmentation. Metadata consists of person age (age), gender (sex) and the lesion location (localization). The labels and metadata are combined and provided in a separate CSV file where each of the image is linked to its corresponding label and metadata. The distribution of the classes is mildly imbalanced. Melanocytic nevi (nv) is the largest class, accounting 67% of the dataset, which basically makes it imbalanced. The second most common class is melanoma (mel) which is 11% and all the other types are 22% of this dataset. This kind of imbalance is common medicine and we will consider it during data preparation. Most of the patients are 30 to 60 years old with a few younger or older patients. Genders are distributed more or less equally

and the lesions are observed on over the body. There are only 6 missing values in metadata which can be replaced with 0 or mean values.

Verifying Data Quality

The HAM10000 dataset appears well-structured and reliable, but a thorough verification process will be conducted:

1. Completeness: Check for missing values in essential fields, such as lesion type, image data, and patient attributes. Address gaps using imputation or exclusion as necessary.
2. Consistency: Verify that labels and metadata match across all entries, ensuring, for instance, that images correspond correctly to lesion IDs and diagnostic categories.
3. Balance: Assess the distribution of lesion types to ensure adequate representation of each category. Imbalanced classes may require resampling techniques during model training.
4. Image Quality: Inspect image resolution and clarity, excluding any images that are significantly distorted or low-quality.
5. Outliers: Identify and analyze extreme values in numeric metadata (e.g., patient age) to determine if they are errors or valid cases.

Task 4. Planning your project

Task 1: Data Preprocessing and Quality Check

- Description: Clean and prepare the dataset by handling missing values, balancing classes, and verifying consistency.
- Hours: Till: 4, Oskar: 4
- Tools: Python (pandas, NumPy), scikit-learn

Task 2: Exploratory Data Analysis (EDA)

- Description: Analyze and visualize the dataset to identify patterns and guide modeling.
- Hours: Till: 5, Oskar: 5
- Tools: Python (matplotlib, seaborn), Google Colab notebook

Task 3: Model Development

- Description: Train machine learning models for lesion classification and optimize performance.

- Hours: Till: 11, Oskar: 13
- Tools: TensorFlow/Keras/PyTorch, scikit-learn

Task 4: Web Interface Development

- Description: Build a web interface to allow users to upload images for lesion analysis.
- Hours: Till: 8, Oskar: 6
- Tools: Flask, HTML/CSS, JavaScript

Task 5: Final Report and Presentation

- Description: Document findings and create a presentation for results and application.
- Hours: Till: 2, Oskar: 2
- Tools: Microsoft Word, PowerPoint

Total Hours per Team Member:

- Team Member 1: 30 hours
- Team Member 2: 30 hours