REPORT (Group B6)

Task 1. Setting up

Link to our repository.

Task 2. Business understanding

Identifying our business goals

- *Background*
  - The mobile phone market is highly competitive and constantly evolving. With new models being released all the time, it is essential for businesses to be able to accurately predict the price classes of phones in order to remain profitable. Traditional methods of price prediction, such as using cost-plus pricing, are often inaccurate and do not take into account the many factors that can affect a phone's price class.

- *Business goals*
  - The goal of this project is to develop a machine learning model that can accurately predict the price classes of mobile phones based on their specifications. This model will be used to inform pricing decisions for both new and existing phones.

- *Business success criteria*
  - The success of this project will be measured by the following criteria:
    - The accuracy of the model's price predictions.
    - The time it takes to train the model.
    - How much did we learn in the process.

Assessing our situation

- *Inventory of resources*
  - The following resources are available for this project:
    - A training dataset of mobile phone specs and price classes.
    - A test dataset of only mobile phone specs.
    - A team of two students taking the Introduction to Data Science course.
    - Access to computational resources (personal computers)

- *Requirements, assumptions, and constraints*
  - The following requirements, assumptions, and constraints apply to this project:
    - The model must be able to predict the prices of new phones that are not yet in the market.

        - The deadline for this project is 11th of December 2023.

- *Risks and contingencies*
    - The following risks and contingencies have been identified for this project:
        - The data may not be of sufficient quality to train an accurate model.
        - The model may not be able to generalise to new phones(e.g. For new phones with way better specs than the phones in the training dataset).

- *Terminology*
    - The following terminology is relevant to this project:
        - Machine learning: A field of computer science that allows computers to learn without being explicitly programmed.
        - Mobile phone specifications: The technical characteristics of a mobile phone, such as its processor, RAM, and camera.
        - Price class prediction: The process of forecasting the future price class of a product or service. In our case there are 4 classes: 0 - low cost, 1 - medium cost, 2 - high cost and 3 - very high cost.

- *Costs and benefits*
    - Costs
        - The team members' time.
        - The cost of electricity for charging/powering our laptops/desktop computers.
    - Benefits
        - More accurate model for predicting mobile phone price classes.
        - New knowledge the team members acquire about data science in the process.

Defining our data-mining goals

- Data-mining goals
    - Explore and understand the structure of our datasets, identify and handle missing/bad values.
    - Select multiple machine learning approaches. Train a model that predicts the price range of a mobile phone based on the given features.
    - Evaluate the performance of each trained model and choose the model that demonstrates the highest performance based on the accuracy.
    - Find what features affect the price range the most.

- Data-mining success criteria
    - The accuracy of the model's price classifications.
    - The ability of the model to show how strongly different features affect the price of mobile phones.
    - The insights gained from the model that can be used to inform pricing decisions.

Task 3. Data understanding

- Gathering data
  - Outline data requirements
    - Mobile phone specifications for a variety of mobile phones.
    - Historical price range data for a variety of mobile phones.
    - Maybe some additional data that could provide relevant context or insights.
  - Verify data availability
    - Data is available on kaggle (we took the idea from there).
    - If needed more data is on different mobile phone retail sites. Can be scraped.
  - Define selection criteria
    - The data represents a diverse range of mobile phones.
    - The data is complete with few/none null values.
    - The data is relevant.
    - The data is accurate.
- Describing data
  - Data sources
    - Main data source is Kaggle (https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification).
    - We also have the opportunity to web scrape for some additional data.
  - Data volume
    - We have 2000 rows of train data (with specified price ranges)
    - And 1000 rows of test data (without specified price ranges)
    - Both of the datasets have 21 features. That will change when we modify the data.
  - Data attributes
    - 'Battery_power' - Battery size (mAh)
    - 'Blue' - Has BlueTooth (1 or 0)
    - 'Clock_speed' - Processor clock speed (GHz)
    - 'Dual_sim' - Has dual sim (1 or 0)
    - 'Fc' - Front camera (Mp)
    - 'Four_g' - Has 4G (1 or 0)
    - 'Int_memory' - Storage space (Gb)
    - 'M_dep' - Mobile phone depth (cm)
    - 'Mobile_wt' - Mobile phone width (cm)
    - 'N_cores' - Processor cores (int)
    - 'Pc' - Primary camera (Mp)
    - 'Px_height' - Screen pixel height (int)
    - 'Px_width' - Screen pixel width (int)
    - 'Ram' - Phone's RAM (Gb)
    - 'Sc_h' - Screen height (cm)
    - 'Sc_w' - Screen width (cm)
    - 'Talk_time' - How long can you voice call someone before the battery dies.

        - 'Three_g' - Has 3G (1 or 0)
        - 'Touch_screen' - Has touch screen (1 or 0)
        - 'Wifi' - Has wifi (1 or 0)
        - 'Price_range' - Price range (1 or 2 or 3 or 4)
- Exploring data
    - We analysed the data in our Jupyter notebook found in our repository (Link under Task1)
    - We checked for nulls. Found none.
    - Checked for duplicate values. Found None.
    - Checked if the dataset is balanced. It is perfectly balanced.
    - Generated distributions and boxplots for every feature. Found nothing out of the ordinary.
    - Checked for correlations. Found that features that are heavily correlated are: the screen height and width, screen pixel height and width, 3G and 4G, ram and price range, front camera and primary camera.
- Verifying data quality
    - We chose to turn two features, screen height and width, into one, screen diagonal.
    - Also because every phone that has 4G has also 3G, so it made sense to turn it into one variable network_type where ‘2’ stands for 4G, ‘1’ for 3G and ‘0’ for none.
    - For cameras we took the mean of the front and the primary camera Mp’s.

Task 4. Planning your project

- Make a detailed plan of your project with a list of tasks. There should be at least five tasks. Specify how many hours each team member will contribute to each task.

    - **Fill out the report for homework 10 (First steps of the project)**
        - Sander Põldma: 3h
        - Markkus Koddala: 1h
    - **Set up the project repository.**
        - Sander Põldma: 0h
        - Markkus Koddala: 1h
    - **Gather data** (data from Kaggle, scraping from web)
        - Sander Põldma: 1h
        - Markkus Koddala: 5h
    - **Check the data for nulls, duplicate values, balance etc.**
        - Sander Põldma: 0h
        - Markkus Koddala: 2h
    - **Fix the dataset if needed (features, balance, values etc.).**
        - Sander Põldma: 2h
        - Markkus Koddala: 0h

    - **Data analysis (EDA, feature engineering, preprocessing)**

        - Sander Põldma 10h
        - Markkus Koddala 10h
    - **Select multiple machine learning approaches. Train a model that predicts the price range of a mobile phone based on the given features.**
        - Sander Põldma: 10h
        - Markkus Koddala: 3h
    - **Evaluate the performance of each trained model and choose the model that demonstrates the highest performance based on the accuracy.**
        - Sander Põldma: 6h
        - Markkus Koddala:2h
    - **Find what features affect the price range the most.**
        - Sander Põldma: 2h
        - Markkus Koddala:2h
    - **Analysing and processing data from hinnavaatlus.ee**
        - Sander Põldma 3h
        - Markkus Koddala 3h
    - **Preparations for Poster Session (poster)**
        - Sander Põldma 2h
        - Markkus Koddala 0h
    - **Poster Session**
        - Sander Põldma 2h
        - Markkus Koddala 2h
    - (If we feel that we haven't done enough work, we can develop an easy application or a system that provides an opportunity to evaluate the value range of people's phones.**)**
        - Sander Põldma 10h
        - Markkus Koddala 10h

- List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify.
    - Web scraping
        - Soap
    - Data Preprocessing:
        - Pandas: Data manipulation and cleaning
        - NumPy: Numerical operations on data
        - Scikit-learn: Feature scaling, encoding categorical variables
    - Feature Selection/Engineering:
        - Scikit-learn: Feature selection methods (e.g., SelectKBest, Recursive Feature Elimination)
        - Domain Knowledge: Understanding which mobile features significantly affect the price range
    - Model Selection:

        - Scikit-learn: Various classification algorithms (Decision Trees, Random Forests, SVM, etc.)
        - XGBoost, LightGBM: Gradient boosting methods for improved performance
    - Model Training:
        - Scikit-learn: Model training interfaces
        - Cross-validation: K-fold cross-validation for model evaluation
    - Model Evaluation:
        - Scikit-learn: Metrics for classification evaluation (accuracy_score, precision_score, recall_score, etc.)
        - Confusion matrix visualisation: Matplotlib, Seaborn
    - Tasks:
        - Data Quality: Ensure data collected is reliable, clean, and relevant to the problem.
        - Feature Engineering: Extract meaningful features from the data that can contribute to accurate predictions.
        - Model Selection: Experiment with various algorithms and select the one that best suits the data and problem at hand.
        - Deployment: Focus on creating an intuitive and user-friendly interface for easy access and understanding for users.
        - Monitoring: Regularly monitor the model's performance to detect and address any degradation in prediction accuracy.