FIT5145 Data Science Project

Zihan Yin

2024-03-25

Project Description

Domain: Data Science in Himalayan Mountaineering

Topic: Application of Data Science in Improving the Summit Success Rate of Himalayas: Analysing Key Factors for Climbers' Success

As mountain climbing in Nepal and the wider Himalayan region grows in popularity, optimizing mountaineering strategies and enhancing safety awareness have become crucial. This project will use machine learning and statistical analysis to assess key factors influencing summit success and safety. The goal is to provide actionable insights and recommendations that improve success rates and reduce risks, benefiting the global mountaineering community.

Note: The topic is the same as that of Assignment 1, but the content of Assignment 1 is shorten here.

Project Objectives

- Identify and quantify key factors influencing summit success using data science, offering targeted advice for climbers.
- Raise climbers' awareness of potential risks and suggest effective strategies to reduce the risk of injury or death.

Data Science Roles & Responsibilities

1. **Project Manager:** Oversees project planning, coordination, resource management, and stakeholder communication.
2. **Business Analyst:** Gathers requirements from climbers and stakeholders, facilitating communication across teams.
3. **Record Collector & Data Entry Clerk:** Collects, verifies, and digitizes data ensuring its quality and accessibility (R. Salisbury 2017a).
4. **System Architect & Data Engineer:** Designs and implements the infrastructure for data processing and analysis.
5. **Data Analyst & Scientist:** Performs exploratory analyses and builds predictive models to extract insights and make data-driven recommendations.

Business Model

The results of this project may have potential or direct impact on the following areas:

- Mountaineering expeditions in the Himalayan region
- Mountaineering equipment development
- Academic research on high altitude activities
- Insurance product design, etc.

Values & Stakeholders

1. **Climbers and Teams:** Receive customized preparation and safety strategy advice.
2. **Mountaineering Companies:** Improve customer experiences and safety outcomes, adjusting services based on analytical insights.
3. **Academic and Research Institutions:** Gain access to valuable data for studying physiological impacts and environmental factors.
4. **Government and Safety Regulators:** Utilize findings to refine safety regulations and rescue operations.

Project Challenges

1. **Data Integrity:** Address issues related to incomplete or inaccurate data due to various factors such as climber reticence or environmental conditions.
2. **Technical Complexity:** Manage the challenges of complex data processing and model accuracy enhancement.
3. **Terminological Barriers:** Overcome the jargon and specialized knowledge required for effective data interpretation.
4. **Stakeholder Resistance:** Navigate potential resistance from entities that may perceive project outcomes as threats to their operational practices.

Characterising & Analysing Data

Data Characteristics & Softwares

- **Data sources & volume**:

  There' re 4 datasets from The Himalayan Database (R. Salisbury 2017b) and the first 3 of them are relevant to this project.

	Dataset Name	Description	#Rows	#Columns	Size
1	`peaks.csv`	Peaks records	479	25	112 KB

	Dataset Name	Description	#Rows	#Columns	Size
2	`exped.csv`	Expedition records	11,184	66	5773 KB
3	`members.csv`	Member records	85,336	78	34226 KB
4	`refer.csv`	Literature records	15,586	13	1919 KB

- **Data velocity**:

  The data covers all expeditions from 1905 through Spring-Summer 2023 (R. Salisbury 2017c).

Dataset	Change Velocity
`peaks.csv`	Unknown
`exped.csv`	278 expeditions were recorded in 2022.
`members.csv`	3,290 members were recorded in 2022.

- **Data variety**:

Dataset	Feature Types
`peaks.csv`	character, integer, numeric, logical, text, nominal factor
`exped.csv`	character, integer, numeric, logical, text, nominal factor, ordinal factor, datetime
`members.csv`	character, integer, numeric, logical, text, nominal factor, datetime

- **Data veracity**:

  1. Most of the feature names are difficult to understand meanings and need to be renamed.
  2. There' s a lot of redundancy in the database.
  3. Several features have numerous missing values, even close to 100%, requiring a lot of work to process missing values.
  4. Several nominal factor features are inconsistently recorded making them difficult to be analysed, requiring lots of text processing.

     Take the example of the mountaineering company "Arun Treks":

AGENCY
Arun Treks
Arun Treks (Himalayan Guides pmt)
Arun Treks (Jamling)
Arun Treks (Monterosa Treks pmt)
Arun Treks (on Sandy Allan permit)
Arun Treks (Rosa Fernandez pmt)
Arun Treks (Seven Summit Treks pmt)
Arun Treks (Snowy Horizon pmt)
Arun Treks permit

AGENCY
Arun Treks/Ang Nuru
Arun Treks/Thamserku Trekking
Arun Treks/Transcend Adventure
Dream Himalaya Adventures (Dawa from Arun Treks)
Great Escapes Trekking/Arun Treks
Pasang Sherpa via Arun Treks
Summit Nepal Trekking/Arun Treks

- **Software**:

  For data access, the Himalayan Database website provides a `.zip` file containing an executable file `Himal 2.71.exe`, which allows users to query the database using the SQL language (Richard Salisbury 2023). The website also provides online querying of a subset version of the database for users with tablets and smart-phones. The database can also be accessed using a DBMS such as MySQL Workbench. For convenience, we read the 4 `.dbf` files in the `.zip` file, which are the 4 datasets introduced above, directly using R language.

  For data processing, analysing and visualization, we use R or Python because these 4 datasets belong to the category of small data, far from the scale of big data, and R & Python are competent enough.

Data Analysis Methods

Data Pre-processing:

- **Feature Selection and Merging**: Select features directly related to the predictive label and merge certain features to reduce redundancy.
- **Data Type Conversion**: Adjust features to suitable data types to fit model requirements.
- **Missing Value Handling**: Fill in missing values to ensure data completeness.
- **Data Cleaning**: Remove records that are not suitable for the model and those that may cause data joining errors.
- **Data Joining**: Join the `peaks`, `exped`, and `members` datasets into a single main dataset for subsequent analysis.

Reasons for Choosing Random Forest Model:

- **Versatility**: Random forest can adapt to various data types and analysis problems, handling different data characteristics effectively.
- **Automatic Data Issue Handling**: The model can efficiently manage outliers and missing values, suitable for the varied quality of data in this project.
- **Embedding Feature Selection**: The model automatically assesses feature importance during training, helping to reduce overfitting and highlight key factors.

- **Importance Evaluation**: Feature importance is assessed through the decrease in accuracy, visually demonstrating the impact of features on prediction success, aligning with the project's goal to identify key success factors.

Model Parameter Tuning and Performance Evaluation:

- **Parameter Tuning**: Optimize the main parameters of the random forest, such as the number of trees and maximum depth, through grid search techniques to ensure the model is neither overfitting nor underfitting.
- **Performance Evaluation**: Use metrics such as accuracy and the AUC-ROC curve to evaluate the predictive performance of the model, ensuring the reliability and accurate identification of key factors.

Feature Visualization:

- **Visualizing the Relationship Between Key Features and Labels**: Display the relationship between key features identified by the model and summit success rates to intuitively understand the impact of each feature. This step not only enhances the model's interpretability but also facilitates the presentation of analysis results to non-technical stakeholders.

Demonstration

Data Pre-processing:

Conducting the pre-processing steps we describe above, we end up with a main dataset:

- Number of rows: 78,711
- Number of columns: 27
- label feature: `if_success`
- Each row of data describes a record of one expedition by one climber.

Feature	Description
`height`	The height of the peak
`himal`	The mountain range in which the peak is located (Richard Salisbury, Hawley, and Bierling 2021)
`region`	The region in which the peak is located
`mt_climbing_status`	Whether the peak is unclimbed
`first_ascent_year`	The year the peak was first climbed
`if_open`	Whether the peak is on the list of peaks approved by the Government of Nepal for mountaineering expeditions.
`year`	The year in which the climber climbed the peak

Feature	Description
season	The season in which the climber climbed the peak
age	Climber 's age
sex	Climber 's gender
race	Climber 's ethnicity (Sherpa/Tibetan or not)
n_team_members	Number of members in the climber's team
n_camps_above_BC	Number of camps used by the climber's team
length_fixed_rope	Length of fixed ropes used by the climber's team
level_oxygen_usage	Amount level of oxygen used by the climber
level_team_hired_personnel	Proportion of hired guides in the climber's team
if_commercial_route	Whether the route is a commercial route
if_success	Whether the climber was successful in reaching the summit
……	……

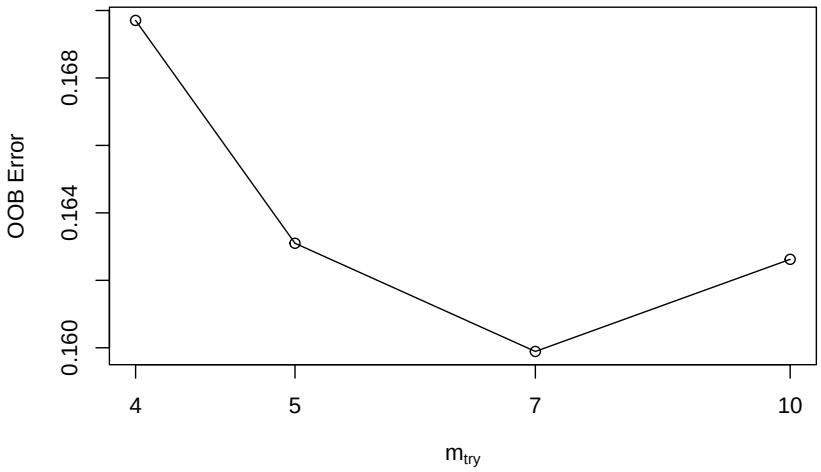
Random Forest Model

Before building the random forest model, let's look at the balance of the labels:

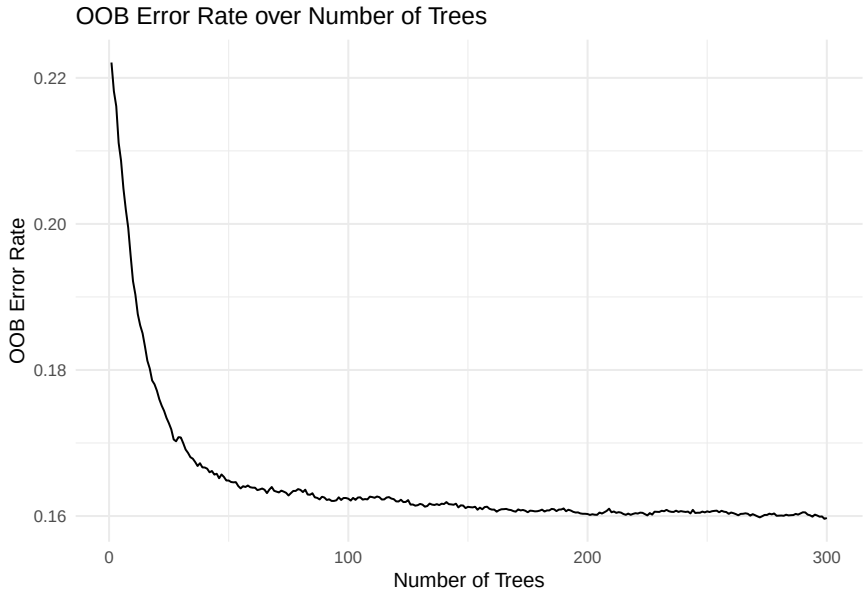
Var1	Freq
FALSE	56.6
TRUE	43.4

The labels are well balanced. Next we split the dataset into a training set and a test set, with 80% and 20% of the data, respectively.

The 2 important hyperparameters in the random forest are the number of trees `ntree`, and the number of randomly selected features in each division of the tree `mtry`. we use `ntree = 300`, and then use the `tuneRF` function to tune `mtry` automatically. As shown below, the optimal `mtry` value recommended by `tuneRF` is 7.



The model is built using `ntree = 300` and `mtry = 7` and an OOB error rate plot is generated. As shown, the lowest OOB error rate occurs when the number of trees is generated to about 200, indicating that there is no need to use `ntree = 300`.

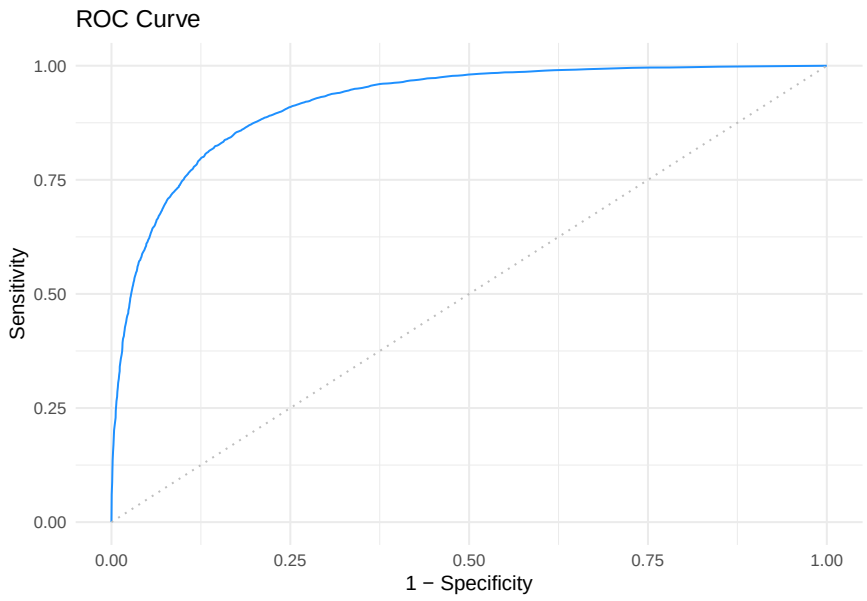


The final random forest model is built using `ntree = 211` and `mtry = 7`. To validate the model performance, prediction is made on the test set and then the following confusion matrix is generated, resulting in an accuracy of 0.8418.

	FALSE	TRUE
FALSE	7739	1253

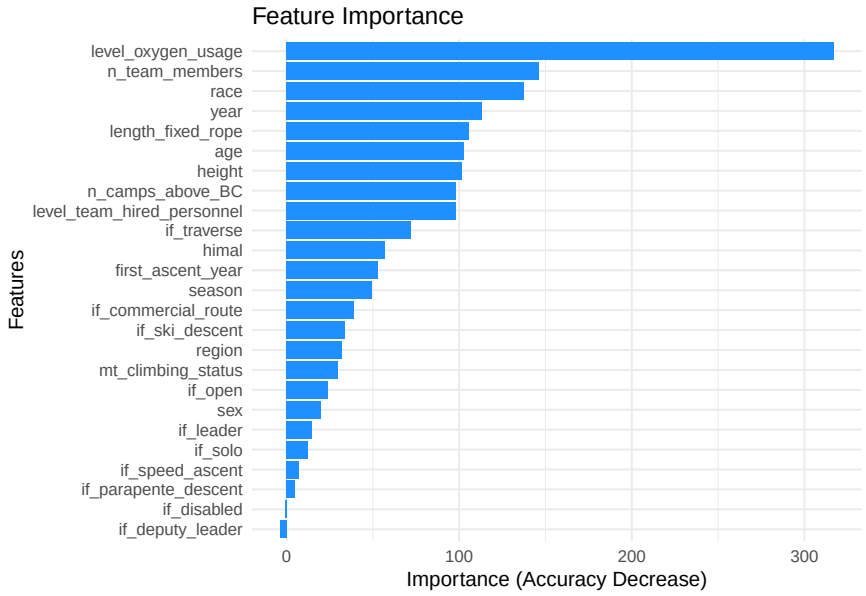
	FALSE	TRUE
TRUE	1238	5513

Evaluate using the more reliable method ROC_AUC. As shown in the plot, the ROC curve of the model is much higher than the completely random classifier and is very close to the upper left corner.



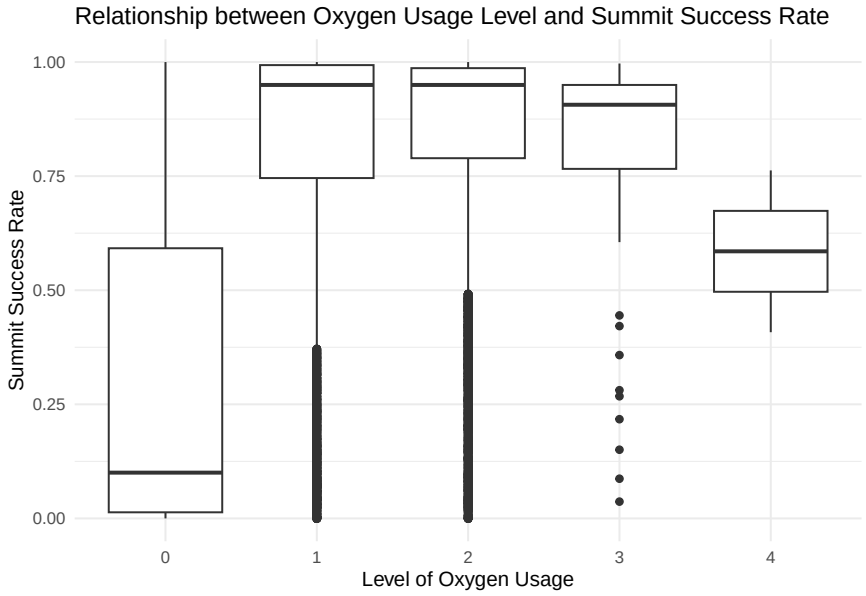
The corresponding AUC value is 0.919. both accuracy and ROC_AUC prove that our random forest model is a reliable and strong classifier.

Next we extract feature importance information from the model and generate the plot. Feature importance here is measured based on the value by which the model accuracy decreases after the features are randomly rearranged, which is more reliable than the default Gini coefficient. As shown in the plot, `level_oxygen_usage` is ranked first in importance, almost twice as high as the second place `n_team_members`.



Visualization

This box plot shows the relationship between different levels of oxygen use and success rates. There's a general tendency for the success rate to increase as the level of oxygen usage increases, especially between levels 1 and 3 where the success rate increases significantly.

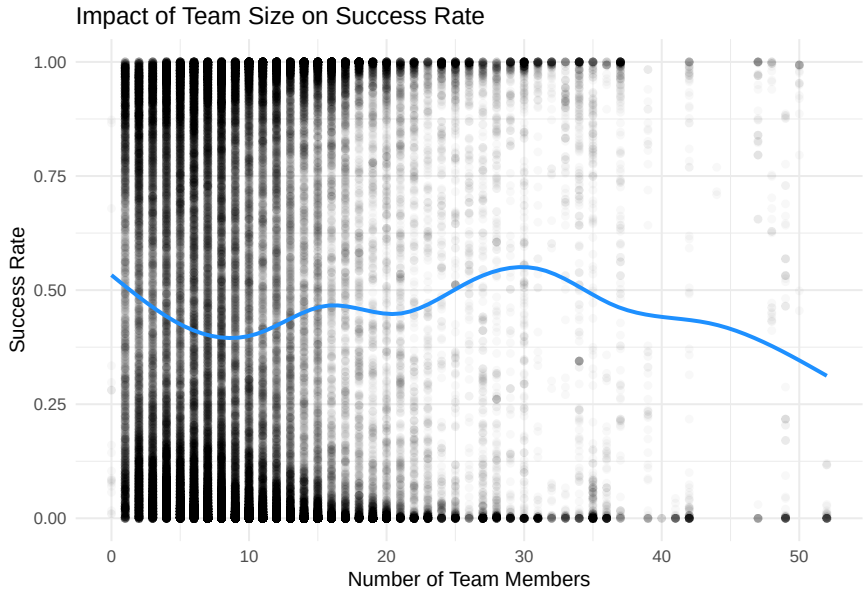


Curiously, the success rate was much lower at level 4. We have two conjectures. One is that most level 4 climbers suffer from acute altitude sickness. Secondly, some climbers who do not have the physical fitness to reach the summit put all their hopes on oxygen.

We believe the use of oxygen directly leads to an increase in the summit rate (Pollard and Murdoch 2006), therefore we recommend:

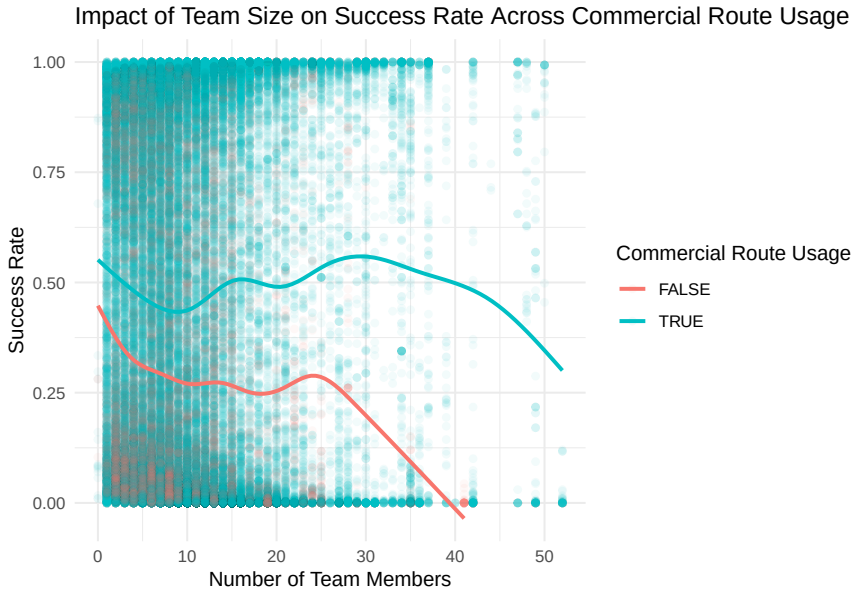
- Climbers who are obsessed with summit success can use oxygen in moderation.
- Equipment manufacturers can develop lighter oxygen equipment to reduce the burden.
- Mountaineering schools can offer courses related to oxygen equipment.
- Researchers studying high altitude activities can consider oxygen as an significant factor.

Focus on the 2nd ranked feature `n_team_members`. This plot shows its direct impact on prediction is less significant than the model importance score.



The Random Forest model may have exploited interactions between `n_team_members` and other features to enhance prediction performance, which is not readily observable in the univariate plot.

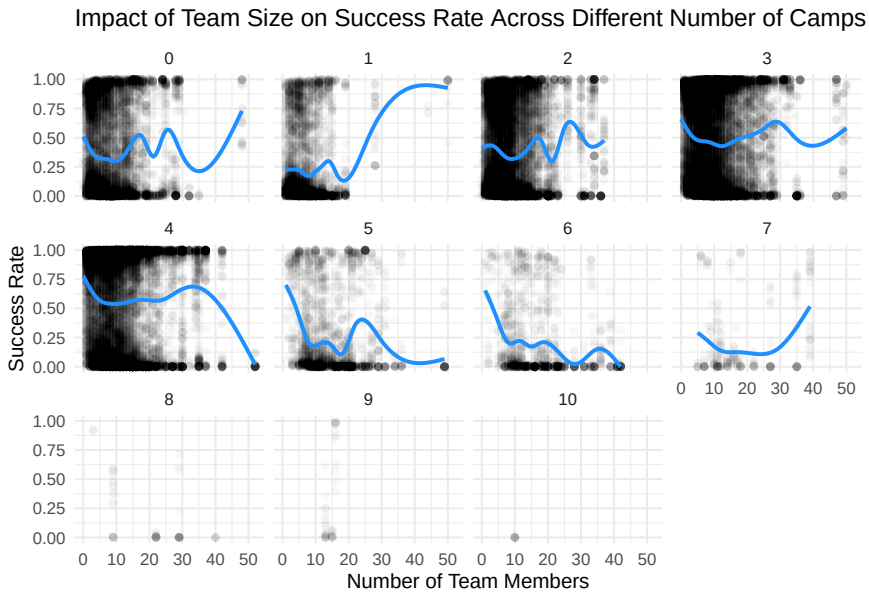
This plot shows the relationship between `n_team_members` and success more clearly based on `if_commercial_route`:



- As the `n_team_members` increases, the success rates on non-commercial routes rise slightly and then gradually decrease. The success rates in the commercial routes follow this trend.
- Commercial routes have overall higher success rates than non-commercial routes.
- For the non-commercial route, the success rates reach a high point when the team size is around 2 to 10, and then gradually decreases, while the commercial route peaks at around 29 team members.

This suggests that a moderate team size is more conducive to success for both commercial and non-commercial routes.

This plot is displayed stratified according to the number of camps:



- For 4, 5, 6 number of camps, the success rates decrease overall as the team size increases.
- For 2, 3 number of camps, success fluctuates a lot.

This difference suggests that the optimal team size varies under different climbing routes, which is an important consideration for planning climbing strategies (Houston and Schoene 2009).

Standard, Data Governance & Management

Standard for Data Science Process

We have adopted a systematic data science process to ensure efficiency and standardization from data collection to model deployment. This process includes data preprocessing, exploratory data analysis, modelling & evaluation, as well as business understanding. We follow the CRISP-DM model to ensure the reliability and effectiveness of our project results.

Data Governance & Management

To ensure compliance with data governance and management, we have established practices concerning data accessibility, security, confidentiality, and potential ethical issues (Tene and Polonetsky 2021):

Data Accessibility and Security:

- All datasets are stored on safe servers, accessible only to authorized project roles.
- Data access is protected with strong passwords.
- Regular data backups are performed to prevent data loss or damage.

Data Confidentiality:

- Climbers' personal information, which could involve personal privacy, is not disclosed to the public, ensuring no personal identities are revealed.
- No details that could identify individuals are disclosed.

Ethical Considerations:

- During the data collection and analysis process, we respect the rights and privacy of all participants.
- We implement a transparent data usage and processing policy to ensure that all stakeholders have a clear understanding of the purposes and methods of data and project results usage.

References

Houston, C. S., and R. B. Schoene. 2009. *Science of Climbing and Mountaineering*. Human Kinetics.

Pollard, A., and D. Murdoch. 2006. *High Altitude Medicine and Biology*. Springer.

Salisbury, R. 2017a. "Team Members." The Himalayan Database.

———. 2017b. "The Expedition Archives of Elizabeth Hawley." The Himalayan Database.

———. 2017c. "The History of the Himalayan Database." The Himalayan Database.

Salisbury, Richard. 2023. "Program Guide for Windows, Himal 2.6." The Himalayan Database; https://www.himalayandatabase.com.

Salisbury, Richard, Elizabeth Hawley, and Billi Bierling. 2021. *The Himalaya by the Numbers: A Statistical Analysis of Mountaineering in the Nepal Himalaya, 1950-2019*. 2nd ed. Ann Arbor, MI: The Himalayan Database.

Tene, O., and J. Polonetsky. 2021. *Big Data Ethics: Balancing Risk and Innovation*. Wiley.