ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

UNIDAD DE TITULACIÓN

DESARROLLO DE UN MODELO DE RECOMENDACIÓN DE RUTAS PERSONALIZADAS PARA GRUPOS DE CICLISTAS EMPLEANDO SEQUENCE-AWARE RECOMMENDER SYSTEMS

TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL GRADO DE MASTER EN INVESTIGACIÓN EN COMPUTACIÓN

MARCO ESTEBAN AGUIRRE ROJAS
marco.aguirre@epn.edu.ec

DIRECTOR: PhD. LORENA KATHERINE RECALDE CERDA
lorena.recalde@epn.edu.ec

CODIRECTORA: PhD. ROSA DEL CARMEN NAVARRETE RUEDA
rosa.navarrete@epn.edu.ec

Quito, January 2024

APROBACIÓN DEL DIRECTOR

Como director del trabajo de titulación DESARROLLO DE UN MODELO DE RECOMENDACIÓN DE RUTAS PERSONALIZADAS PARA GRUPOS DE CICLISTAS EMPLEANDO SEQUENCE-AWARE RECOMMENDER SYSTEMS desarrollado por Marco Esteban Aguirre Rojas, estudiante de la Maestría de Computación con mención en Sistemas Inteligentes, habiendo supervisado la realización de este trabajo y realizado las correcciones correspondientes, doy por aprobada la redacción final del documento escrito para que prosiga con los trámites correspondientes a la sustentación de la Defensa Oral.

Firmado electrónicamente por:
LORENA KATHERINE
RECALDE CERDA

PhD. Lorena Katherine Recalde Cerda
DIRECTORA DE PROYECTO

Firmado electrónicamente por:
ROSA DEL CARMEN
NAVARRETE RUEDA

PhD. Rosa del Carmen Navarrete Rueda
CODIRECTORA DE PROYECTO

DECLARACIÓN DE AUTORÍA

Yo, Marco Esteban Aguirre Rojas, declaro bajo juramento que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento. La Escuela Politécnica Nacional puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normatividad institucional vigente.

Firmado electrónicamente por:
MARCO ESTEBAN
AGUIRRE ROJAS

Marco Esteban Aguirre Rojas

AGRADECIMIENTOS

Mi gratitud a Lorena Recalde, por su aporte constante como guía, compromiso y recomendaciones al desarrollo del proyecto. Gracias por su carisma humano y fomentar e incentivar a la investigación. Finalmente agradezco su aporte en conocimientos no solo para el desarrollo de este proyecto además en los estudios de la maestría e inicios de un científico de datos en la Escuela Politécnica Nacional. Quiero agradecer Rosa Navarrete por su apertura y compromiso al desarrollo de este proyecto y guía en el desarrollo de la maestría. Valoro las enseñanzas académicas conseguidas en la maestría principalmente impartidas por los profesores durante todas las clases.

A mi familia, especialmente a mis padres gracias , Martha y Guillermo por su apoyo incondicional constante. Gracias a los valores impartidos he podido avanzar en camino de la vida.

Marco Esteban Aguirre Rojas

CONTENT

RESUMEN

En el contexto actual, se observa un crecimiento significativo en la adopción de tecnologías interactivas en el ámbito deportivo, ejemplificado por plataformas como Strava, Garmin Connect y Fitocracy. Este impulso competitivo ha catalizado avances en las características funcionales, resultando en aplicaciones deportivas más accesibles que transforman el panorama sociotécnico de las prácticas deportivas. Sin embargo, a pesar de la presencia generalizada de sistemas de recomendación en sectores como el comercio electrónico, su integración en la industria deportiva sigue siendo limitada.

Este trabajo se centra en la creación de un Sistema de Recomendación de Grupos (GRS) para rutas ciclistas personalizadas, utilizando Sistemas de Recomendación conscientes de la Secuencia. El objetivo principal es mejorar la experiencia del usuario al ofrecer sugerencias de rutas a grupos de ciclistas, promoviendo la exploración de nuevas rutas. Enmarcado en la Investigación en Design Science Research (DSR), el proyecto establece un marco conceptual para el diseño y construcción de artefactos.

Los resultados experimentales revelan una correlación positiva entre la calidad de la ganancia acumulativa normalizada descontada (NDCG) y el tamaño del grupo de usuarios, destacando mejoras en el rendimiento en grupos más grandes debido a la mayor diversidad. El modelo Prod2vec, entrenado con el algoritmo GoSS-Rec, supera consistentemente a las alternativas en métricas de diversidad y novedad, subrayando su potencial para ofrecer recomendaciones diversas y novedosas en grupos de diferentes tamaños. Este enfoque contribuye a la evolución de los sistemas de recomendación en ámbitos no tradicionales como el ciclismo, mejorando la experiencia deportiva para los usuarios.

Keywords: Group Recommender System, Sequence-Aware Recommender Systems, Sequential Prediction, Self-attention, Sports Science

ABSTRACT

In recent times, there has been a notable increase in the adoption of technologies and interactive platforms within the realm of physical fitness and sports. This trend is exemplified by the emergence of products such as Strava, Garmin Connect, and ®Fitocracy. The competitive dynamics within this sector have driven progress in functional features, resulting in the development of more user-friendly applications that reshape the sociotechnical landscape of sports practices. At the same time, despite the widespread prevalence of recommendation systems in sectors like e-commerce and entertainment, their integration into new domains, particularly the sports industry, remains relatively limited.

This study presents the creation of a Group Recommender System (GRS) designed for personalized cycling routes, employing Sequence-Aware Recommender Systems. The main objective is to improve the user experience by providing route suggestions to groups of cyclists, encouraging the discovery of new routes. Anchored in the Design Science Research (DSR) framework, the project lays the foundation for an extensive conceptual framework governing the design and construction of artifacts.

Experimental findings highlight a positive correlation between normalized discounted cumulative gain quality (NDCG) and user group size, underscoring enhanced performance in larger groups due to increased diversity. Additionally, in model training with the GoSS-Rec algorithm, the Prod2vec model consistently outperforms alternatives in diversity and novelty metrics for both small and large groups. This emphasizes the potential of the GoSS-Rec algorithm with Prod2vec to excel in providing diverse and novel recommendations across groups of varying sizes, contributing to the evolution of recommendation systems in non-traditional domains.

Keywords: Group Recommender System, Sequence-Aware Recommender Systems, Sequential Prediction, Self-attention, Sports Science

1 INTRODUCTION

In recent years, several technologies, applications and interactive social networks for fitness and sports have been developed, some of which are marketed as products (e.g., Strava, Garmin Connect, Endomondo, Fitocracy, Runtastic, Map My Ride, My Fitness Pal, and Zwift) [1]. These platforms have gained popularity due to the growing interest in fitness and sports, as well as the integration of technology to monitor training and improve social interactions. Moreover, users often find motivation, support, and a sense of community by connecting with like-minded individuals on these platforms [2].

The combination of fitness and sports platforms with digital technologies results in Online Social Fitness Networks (OSFNs) that have leveraged technology to track and quantify exercise behavior while simultaneously fostering a discursive space for users to engage in social interactions related to fitness. By offering personalized experiences, motivation, and a sense of community, these tools have become valuable assets for individuals striving to lead more healthful and active lives [3]. Among these applications, Strava (OSFN), a prominent fitness platform, has gained immense popularity for its ability to track and analyze various sports activities, from running and cycling to swimming and more. Therefore, Strava has attracted research interest from a variety of scientific domains [4] [5] [6].

The vast amount of data generated by Strava users presents an opportunity to harness this information for developing sophisticated recommender systems (RSs). RSs have proven to be effective in a multitude of domains, from e-commerce to entertainment, by offering personalized recommendations to users based on their preferences and behaviors [7]. In the context of sports, these systems play a pivotal role in personalizing fitness experiences by offering tailored training activities, nutrition plans, rehabilitation routines and product promotions. For instance, [8] explores cycling route recommendations by estimating users' heart rate profiles. Given the analysis of Strava data and user behavior, a route RS has the potential to transform the way cyclists approach their fitness activities in terms of exploration, motivation and overall well-being.

This research studies the synergy between fitness applications, with emphasis on the Strava ecosystem, and the integration of sequential models for route RSs. The development of a group recommender system (GRS) model of personalized routes for groups of cyclists by using Sequence-based Recommender Systems. The main objective is to provide recommendations to a group of cyclists according to the level of work or effort developed by the individual members of the target group. Thus, the groups of cyclists perceive a more pleasant user experience that stimulates them to explore new routes calculated by the model, which is embedded in the GRS. The project is based on an experimental research developed under the DSR (Design Science Research) [9] methodological framework. DSR adapts to the project since it has a comprehensive conceptual framework for the design and construction of this type of solutions called *artifacts* (models, algorithms, instances, etc.).

This project contributes to the evolving landscape of fitness applications, Strava data utilization, and the transformative role of sequential route GRS in shaping personalized fitness journeys. By weaving these elements together, we aim to revolutionize the way fitness enthusiasts engage in physical activities; and thus, unlocking untapped potential for improved performance, sustained motivation and ultimately better overall well-being.

1.1 RESEARCH QUESTION

This work aims to answer the following research question: How can an effective personalized route recommendation model be developed using sequence-aware recommender systems to enhance the experience of cyclist groups, by considering their individual preferences and needs?

1.2 GENERAL OBJECTIVE

Develop a personalized route recommendation model for groups of cyclists by using sequence-aware recommender systems.

1.2.1 Specific Objectives

- ❐ Perform a literature review related to Group Recommender Systems and Sequence-Aware Recommender Systems for route recommendation.
- ❐ Apply data mining with the use of the ®Strava API (application programming interface) to extract and characterize data and establish user profiles and route items as recommendations.
- ❐ Propose a personalized route recommendation model for groups of cyclists using Sequence-Aware Recommender Systems.
- ❐ Evaluate the proposal, through Sequence-Aware Recommender Systems metrics, with respect to other Sequence-Aware models.
- ❐ Present a scientific paper, with the development and evaluation of the model.

1.3 EXPECTED RESULTS

Upon completion of this project, it is expected to obtain a personalized route recommendation model for groups of cyclists using Sequence-Aware Recommender Systems. Thus, user groups will be able to access route recommendations, which are calculated considering the performance profile of the group members and the appropriate sequence of the segments that make up the route.

1.4 THEORETICAL FRAMEWORK

1.4.1 Recommender Systems

Recommender systems (RSs) are software tools and techniques used to find common patterns and relationships, in order to provide suggestions for items that will be useful to a user or group of users [10]. An RS is a specific type of advice-giving or decision support system that guides users in a personalized way to interesting or useful objects in a large space of possible options or that produces such objects as output [11] [12]. Nowadays, the roles of RSs in both industries and academia cannot be overemphasized. The RS techniques have been employed in applications, e-commerce/e-shopping, e-government, e-business, e-learning, e-tourism, e-library, e-resource services, and e-group activities [13]. In recent years, RSs have been a notable surge in interest surrounding, as the following instances demonstrate. RSs play an important role in e-commerce applications and social media, for example, Amazon.com, LinkedIn, Facebook, Tripadvisor, YouTube, Netflix, Spotify, and IMDb. Additionally, numerous media enterprises have begun incorporating and implementing recommender systems as integral components of the services they offer to their subscribers. An illustrative example of this can be observed on the platform Amazon.com, where the website utilizes a recommender system to tailor the online store experience to individual customers [14] [15]. This recommender engine assists users in selecting items to purchase.

The relevance and impact of RSs consider both organizational and business relevance as follows [16]:

- ❒ Increase the number of items sold: This is probably the central role of a commercial RS, aiming to facilitate the sell an additional set of items compared to those usually sold without any kind of recommendation.

- ❒ Sell more diverse items: An additional significant role of an RS is to empower users to discover items that could be challenging to locate without a specific and accurate suggestion.

- ❒ Increase user satisfaction: A well-designed RS can enhance user interaction or satisfaction with the website or application. The combination of precise and efficient suggestions, along with a user-friendly interface, results in recommendations that capture the user's attention and pique their interest.

- ❒ Better understanding of what the user wants: Another significant function of an RS is to delineate the user's preferences.

The reasons mentioned above are some important motivations as to why the system owner introduces an RS and the item providers offer their services/items through the RS platform.

1.4.2 General Recommendation

Early works on recommender systems typically use Collaborative Filtering (CF) [17] recommendation, which is based on user-rated items. Among different Collaborative Filtering (CF) techniques, Matrix Factorization (MF) stands out as the most widely used approach. MF aims to reveal underlying dimensions that capture user preferences as well as item characteristics and estimate interactions by calculating the inner product between user and item embeddings. [18]. In addition, another line of work is based on Content-based (CB) recommendation, which aims at identifying characteristics that are like those a user has preferred in the past and makes recommendations accordingly [19]. Another line of work is item-based neighborhood methods [20]. On the other hand, hybrid RS is the combination of CB and CF, to use characteristics of both approaches through mergers of individual predictions [21].

Recently, deep learning is enjoying massive popularity. Thus, deep learning has been dramatically revolutionizing RS, by changing the recommendation architecture and providing more possibilities to improve its performance (e.g., Precision, Recall, etc.). In addition, there are different deep learning techniques used in RSs, such as Multilayer (MLP), Autoencoder (AE), Convolutional Neural Network (CNN), etc. [22] For example, NeuMF [23] models the interaction with MLP. Wu et al., in [24] present recurrent recommender(RRN), which employes Long Short-Term Memory (LSTM) networks to capture time-related patterns for both users and items. One line of Deep Reinforcement Learning, is studied in [25] by Chen et al. They propose a Top-k off-policy correction for a reinforced recommender system for recommendations at YouTube.

1.4.3 Sequential Recommendation

Sequential dynamics play a crucial role in many contemporary RSs, which aim to capture the context of users' actions based on their recent activities. The main objective of sequential recommender systems is to mine patterns in user interaction sequences. Therefore, a sequential recommendation models the item-item transition from historical user-to-item interactive behaviors and provides suggestions to users in real time. [26]. Sequence recommendation problems are different from the traditional matrix-completion in the general recommendation. Quadrana et al. [27] present a high-level overview of the problem, as shown in Fig1.1. The inputs are an ordered timestamped list of past user actions. The outputs are lists of items whose order in the recommendation list may be relevant. The specific computational tasks are sequential patterns, or they can be co-occurrence patterns. In general, the ordering of the items can be relevant both with respect to the inputs and to the outputs.

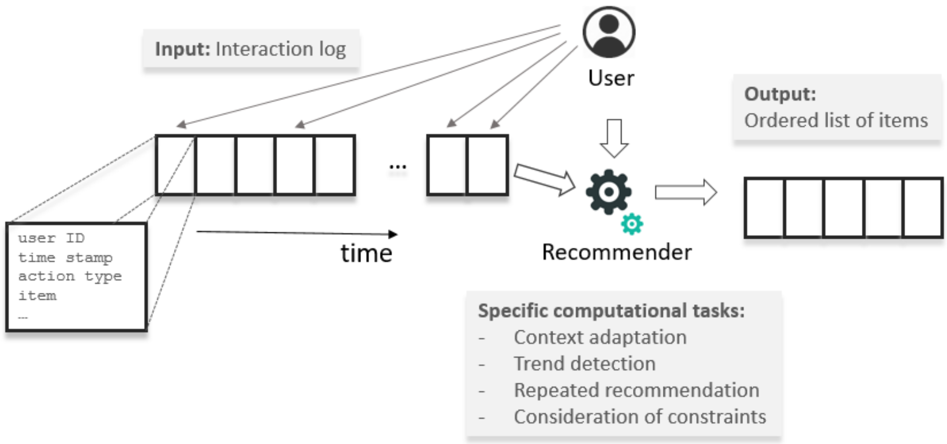


Figure 1.1: High-level overview of sequence-aware recommendation. Source [27].

Nowadays, sequential-based recommender systems are used in different domains of application, [27] such as App recommendation, Query recommendation, e-commerce, and multimedia content transmission, among others. These systems analyze the past behavior of individual users or a community of users as a whole to detect patterns in user interests. Many sequential strategies seek to model item-item transition matrices as a means of capturing sequential patterns among successive items. For instance, Factorazing Personalized Markov Chains (FPMC) [28] fuses a matrix factorization (MC) term and an item-item transition term to capture long-term preferences and short-term transitions respectively. Other methods used to capture transitions are based on Markov Chain (MC) of the first order [29]. In contrast, higher-order MCs consider the connection between the next action and multiple past actions. The user's next action is often influenced by the most recently visited item, which serves a relevant context signal. Methods based on first-order MCs demonstrate impressive performance, particularly when dealing with sparse datasets.

Besides methods that use MCs, there are RNN-based proposals (recurrent neural networks) to model user sequences [30] [31]. For example, GRU4Rec uses RNNs Gated Recurrent Units (GRU) to model click sequences for session-based recommendation [32]. Some of the most recent works explore techniques and mechanisms of so-called Self-Attentive. **Self-Attentive** [33]. For example, the work of Kang and McAuley [34] is notable for introducing SASRec, which is a unidirectional model designed for sequential recommendation.

The model utilizes a causal attention mask and a two-layer transformer decoder, which is essentially a transformer language model. BERT4Rec [35] is the pioneering model that incorporates BERT (Bidirectional Encoder Representations from Transformers) into the field of sequential recommendation. Unlike SASRec, BERT4Rec is a bidirectional model, which allows it to deal with past and future items in the user's behavior sequence. These models represent significant advancements in the sequential recommendation domain and have demonstrated promising results in enhancing recommendation accuracy.

1.4.4 Group Recommender Systems

The rise in social activities and engagements have led to a greater utilization of RSs overall, with a specific focus on group recommender systems (GRSs) [36]. GRSs support users in a group when facing decision-making tasks in scenarios where the recommended items are consumed by groups of users rather than individuals. GRSs aim to assist groups of users in making decisions by suggesting items, products, services, or content that would likely be of interest and relevance to the entire group. To achieve this, the GRS takes into account the preferences and interactions of multiple users within the group [37], and it utilizes algorithms and data analysis techniques to generate recommendations that cater to the group's shared interests and preferences.

1.4.4.1 Basic aggregation functions

Preference Aggregation Functions are mechanisms used to combine individual preferences of group members into a single recommendation or decision. In group recommendation scenarios, different types of aggregation functions are used to approximate the individual preferences of group members because there is no optimal way to aggregate recommendation lists. As mentioned in [38], these functions can be categorized into:

- Majority-based (M) functions, such as Plurality Voting and Borda Count, which determine the winner based on the most frequent or highest-ranked item.

- Consensus-based (C) functions, such as Additive Utilitarian and Average, which take into account the preferences of all group members to determine the winner.

Table 1.1: Aggregation strategies

Strategy	How it works
Plurality/majority voting [M]	Uses ‘first past the post’: repetitively the item with the most votes is chosen.
Average [C]	Averages individual ratings.
Multiplicative [C]	Multiplies individual ratings.
Borda count [M]	Counts points from items’ rankings in the individuals’ preference lists.
Copeland rule [M]	Counts how often an item beats other items (using majority votes) minus how often it loses.
Approval voting [B]	Counts the individuals with ratings for the item above approval threshold.
Least Misery [B]	Takes the minimum of individual ratings.
Most pleasure [B]	Takes the maximum of individual ratings.
Average without Misery [C]	Averages individual ratings, after excluding items with individual ratings below a certain threshold.
Fairness [C]	Items are ranked as if individuals are choosing them in turn.
Most respected person (or Dictatorship)[B]	Uses the rating of the most respected individual.

❒ Borderline (B) functions, such as Approval Voting and Range Voting, that use a threshold or range to determine which items are acceptable or preferred by the group.

The Table 1.1 shows an overview of aggregation strategies of different kinds of aggregation functions taken from social choice theory [36], and their categorization into one of the three aforementioned categories (M, C, and B).

1.4.4.2 Handling Preferences

Before making recommendations, it is essential to understand the preferences of the users. In the context of RSs, preferences are the ordering relation between multiple items that signify the best match for a user from a set of choices [38]. However, acquiring user preferences and interpreting them in a way that leads to relevant item recommendations can be challenging. Most group recommendation applications apply preference elicitation approaches quite similar to those of single-user RSs.

In this scenario, the project is suitable for content-based recommendations. This means that the item(segments) ratings and metadata properties such as categories or labels are explicitly provided to indicate user preferences. The ratings and category preferences reflect the individual preferences of group members.

1.5 STRUCTURE OF THE THESIS

This thesis is structured as follows: in Chapter 2 we present the methodology and define and explain the proposed approach. In Chapter 3 we present the design of experiments, the results obtained and their discussion. In Chapter 4 we provide the conclusions and suggest future work. Finally, we present references, many of them from very recent publications.

2 METHODOLOGY

2.1 DESIGN SCIENCE RESEARCH

RSs provide models for generating suggestions of items that will be of use to a user. These models are designed to support their users in various decision-making processes. They have become one of the most used applications of artificial intelligence due to their essential support in exploring information on the Internet. The development of RSs is a multidisciplinary effort that involves experts from various fields such as artificial intelligence, human-computer interaction, data mining, statistics, decision support systems, marketing, and consumer behavior. Regarding this context and background, the chosen research methodology is Design Science Research (DSR) [39], which can include "constructs, models, methods, or instantiations. In the case of this research, the artifact (final output) would correspond to the recommendation model.

There are several research investigations that integrate DSR in RSs. For example, Arazy et al. [40] present social RSs that leverage users' social relationships to effectively filter and deliver relevant information to users. In [41], the proposed model uses artificial intelligence to gain industry insights, enabling indirect monetization from data. In another study [42] that employes DSR, the development and evaluation of a recommender system integrated into an established computer-supported collaborative learning platform is outlined.

The DSR methodology includes several activities that are typically involved in design science research. These activities are described in [39] and include:

1. **Problem identification and motivation:** This activity involves identifying a problem or opportunity that can be addressed by developing an artifact (model). The researcher should also explain the motivation for why the problem needs to be addressed.

2. **Define objectives for a solution:** The researcher defines the objectives that the ar-

DSR Methodology Phases	Section
Problem identification and motivation	Introduction1
Define objectives for a solution	Objectives1.2
Design and development of the artifact	Proposed Approach 2.2
Demonstration of the artifact	Results
Communication of the research	The research

Table 2.1: DSR Methodology Phases

tifact should achieve in order to solve the identified problem. This involves specifying the requirements and constraints that the artifact must meet.

3. **Design and development of the artifact:** This activity involves designing and developing the artifact that will be used to solve the problem. The artifact can take many forms, such as a software system, a physical device, or a process.

4. **Demonstration of the artifact:** The researcher demonstrates the application of the artifact to solve one or more instances of the problem. This could involve its use in experimentation, simulation, case study, proof, or other appropriate activity.

5. **Evaluation of the artifact:** This activity observes and measures how well the artifact supports a solution to the problem. This requires comparing the intended solution goals with the actual results observed while using the artifact in a demonstration. It requires knowledge of relevant metrics and analysis techniques.

6. **Communication of the research:** The researcher communicates the results of the research to the relevant stakeholders. This could involve scholarly publications professional publications, or other appropriate means of communication.

2.2 PROPOSED APPROACH

Cycling enthusiasts around the world are constantly seeking new and exciting routes to explore, and the advent of digital platforms has revolutionized the way they plan and enjoy their rides. In this digital age, Strava, a popular fitness app for cyclists and athletes, has become a pivotal tool for tracking and sharing cycling activities. To further enhance the cycling experience within the Strava ecosystem, the Group Bike Route Recommendation System, as depicted in Figure 2.1, provides cyclists with highly personalized and engaging route recommendations. This system harnesses the power of user activities, Strava API data mining, ETL processes, group aggregation and profiling, and advanced group recommendation algorithms to provide cyclists with tailor-made route suggestions for group rides. The development of a route recommender system for groups (artifact) using the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology encompasses several phases, including Data Understanding and Data Preparation. In this section, we will delve into each of these phases of project development, offering a comprehensive view of how they synergize to create a system to improve the cycling experience.

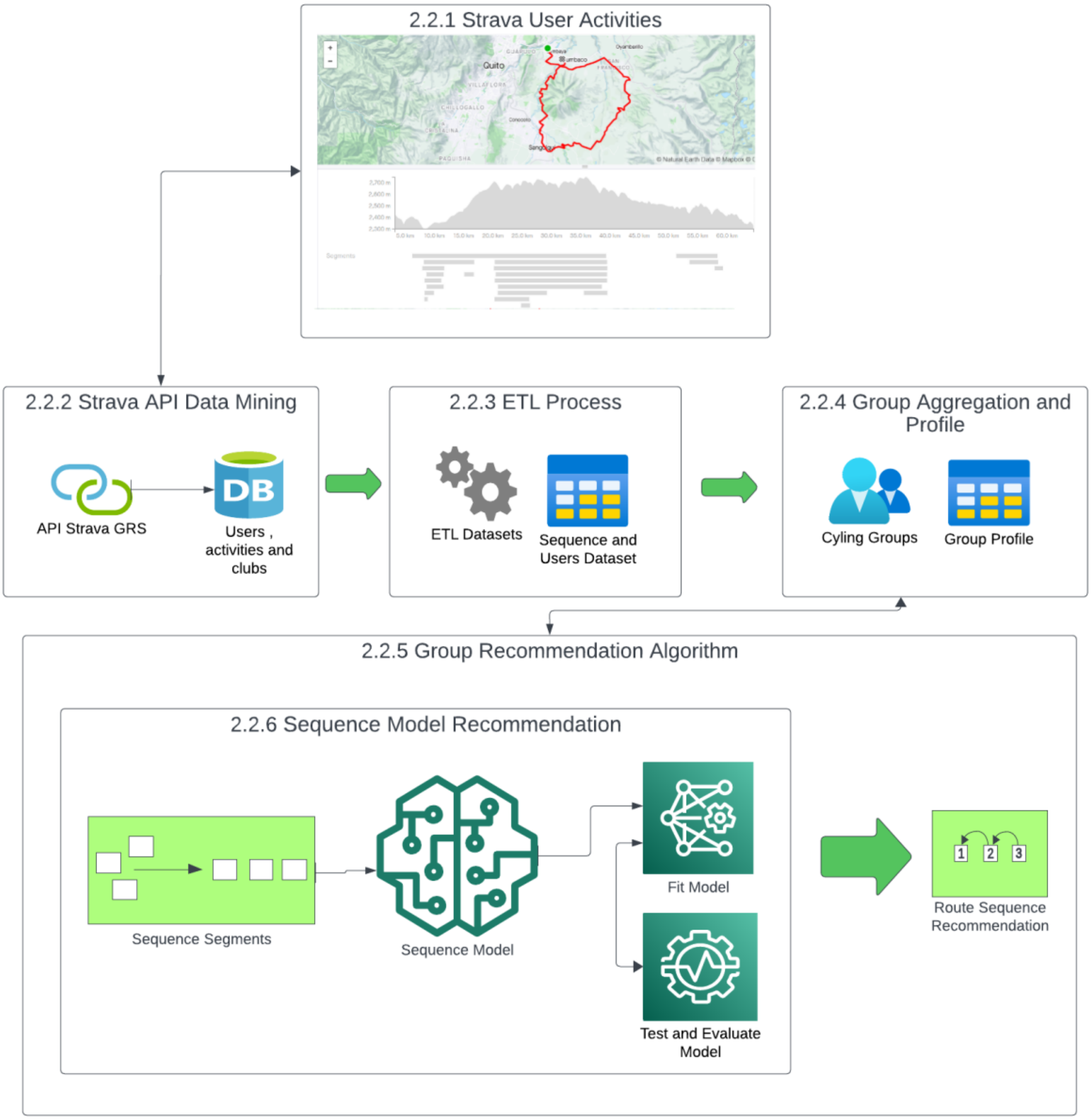


Figure 2.1: Project Description.

2.2.1 Strava User Activities

Strava is a widely used platform where athletes can keep track of and share their fitness activities. Every day, over 100 million athletes from 195 countries use Strava to upload their activities, which include 30 different types of exercises[1]. Therefore, Strava is a rich source of data related to cycling activities. Users record their rides using the Strava mobile app or GPS devices, which generate detailed information including various statistics and metrics, as well as ride data. The Strava cycling activity data specification includes the following key elements:

- **Activity Title:** Users can give their activity a name, which often reflects the type of trip, location or other relevant details.

- **Route Map:** Strava displays an interactive map that traces the exact route that the cyclist took during the activity. This includes details such as starting and ending locations, turns, inclines, declines, and other landmarks along the way.

- **Performance Data:** Strava records essential performance data, such as the total distance covered, elapsed time, average speed, maximum speed, elevation gain, and the highest and lowest altitudes reached.

- **Segments:** If the ride includes popular Strava segments (road or trail sections where cyclists compete), the user's times and rankings on those segments will be displayed.

- **Photos and Comments:** Users can add photos from their ride and comments to share their experiences or anecdotes related to the activity.

- **Tags and Equipment:** Tags can be added to categorize the activity, such as training, group ride, or race. Additionally, users can indicate the type of bicycle they use, whether road, mountain or electric.

- **Advanced Data:** For more dedicated cyclists, Strava offers additional data, such as heart rate, power, and cadence analysis.

[1] https://press.strava.com/articles/strava-releases-year-in-sport-report

2.2.2 Strava API Data Mining

The Strava V3 API[2] is a publicly available interface that enables developers to access Strava extensive dataset. The interface is stable and is currently utilized by Strava mobile applications.

The following is a general guide on how to perform data extraction from the Strava API:

1. **Register as a Developer**: To begin the process, one should establish a Strava account and proceed to create a new application within the My Apps section. This step is instrumental in obtaining the essential credentials required for the utilization of the Strava API. Figure2.2 shows the application *Strava Group Recommendation System* [3] created to obtain data from Strava.

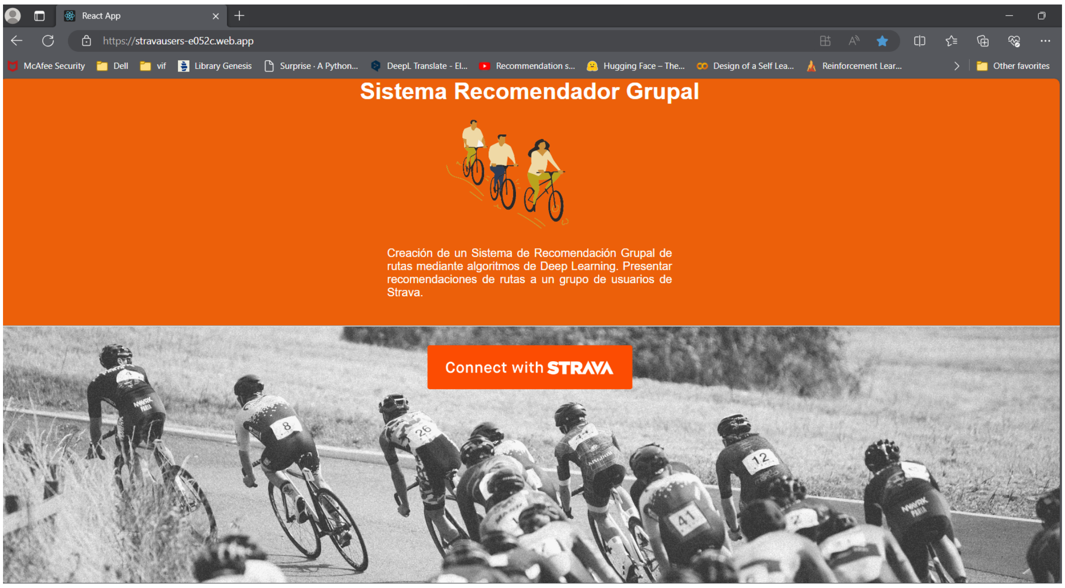


Figure 2.2: Strava Group Recommender System App.

2. **Configuration Application**: The application needs to provide the necessary information, including a description, login URL, and redirect URL. Strava will provide the user with a Client ID and a Client Secret, which are vital to the authentication process when interfacing with the API.

3. **Authentication**: In order to enable users to authenticate applications and authorize

[2] https://developers.strava.com/docs/

[3] https://stravausers-e052c.web.app/

access to their Strava data, one must implement an OAuth2 authentication flow [4].

4. **Make API Requests**: The process involves utilizing the Client ID and Client Secret to acquire an access token after a user has granted authorization to the application. Subsequently, this access token is employed to initiate requests to the Strava API, facilitating the collection of the desired data. Afterward, a wide array of sports-related data, including routes, activities, segments, and more, is accessible via the API for retrieval.

5. **Handle Responses**: The Strava API delivers responses in JSON format. For instance, they can use it to showcase a user's activities, personal statistics, favorite segments, and other relevant information.

2.2.2.1 Strava Dataset

The final dataset obtained from the Strava API includes the following information:

1. **Activity Records:** The dataset will contain records of cycling activities. Each activity entry will include:
   - ❒ id: A appears to be an identifier or unique identifier for a particular record or entry.
   - ❒ user_id: An appears to be an identifier or unique identifier to a user.
   - ❒ start_date: A seems to be a timestamp or date associated with the segment efforts.
   - ❒ segment: A list of segment efforts, and each effort within the list is represented as a dictionary with the following attributes:
     - ✧ start_index: An index representing the starting point of the segment effort.
     - ✧ end_index: An index representing the ending point of the segment effort.
     - ✧ start_date: A timestamp indicating when the segment effort started.
     - ✧ id: An identifier for the segment effort.
2. **Segment Information:** The provided dataset contains information about a segment, which is typically a specific route or portion of a route used for tracking cycling. Information about specific segments can include:

[4] https://developers.strava.com/docs/authentication/

- ❒ id: An identifier for the segment, which is likely unique.
- ❒ distance: The distance of the segment, measured in meters.
- ❒ average_grade: The average grade or incline of the segment, often expressed as a percentage.
- ❒ elevation_high: The highest point of elevation (in meters) along the segment.
- ❒ elevation_low: The lowest point of elevation along the segment, likely measured in meters.
- ❒ total_elevation_gain: The total elevation gain along the segment, often measured in meters.
- ❒ start_latlng and end_latlng: The latitude and longitude coordinates of the starting and ending points of the segment.
- ❒ climb_category: A categorization or classification of the segment based on its difficulty or characteristics.
- ❒ map: A representation of the segment's path or route, possibly in a specific format or encoding.
- ❒ effort_count: The total count of efforts made by athletes on this segment.
- ❒ athlete_count: The total count of athletes who have attempted this segment.
- ❒ star_count: The count of stars or ratings given to this segment by users, often indicating its popularity or quality.

3. **Clubs Information:** This dataset provides basic information about a specific Strava Club. These clubs serve as virtual communities where athletes with similar interests or goals can come together to connect, share their workouts, and engage in various activities. The dataset presents next features:

   - ❒ id: A unique identifier for the Strava Club.
   - ❒ sport_type: The sport_type listed can log various types of activities, such as running, cycling, swimming, etc.
   - ❒ member_count: Indicates the number of members in the club.
   - ❒ id_user: A unique identifier for the user Strava Club.

2.2.3 ETL Process

The Extract, Transform, and Load (ETL) process for the Strava API dataset involves several crucial steps to prepare and structure the data for analysis. Initially, in the extraction phase, data is retrieved from the Strava API through API requests, obtaining raw data related to cycling activities and segment information in JSON format.

In the transformation phase, data cleaning is paramount. This entails identifying and resolving common issues such as missing values, duplicates, and errors within the dataset. Moreover, Segment Efforts are a special case, so it is crucial to ensure that the sequence does not overlap. The comprehensive documentation of all transformations is maintained to ensure a transparent and reproducible record of modifications done. The segment efforts and the management of their overlaps are detailed below.

2.2.3.1 Clear segments sequence in Activity Record Dataset

A sequence activity dataset with overlapping segments represents a collection of activities or events, each defined by a start index and an end index. Overlapping segments in this context mean that there are activities within the dataset that share some common index intervals. This can happen when one segment starts before another one ends or when multiple segments occur concurrently within the same index range. Figure 2.3 shows an example of overlapping $\{s_1, s_2, s_3, ...s_{15}\}$ segments.

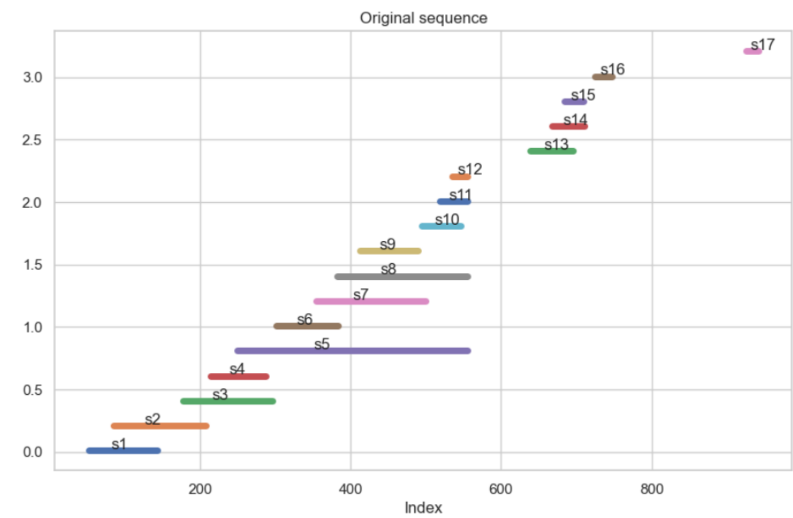


Figure 2.3: Original sequence segments.

The problem of overlapping segments, often referred to as the Interval Scheduling Problem, is a classic scheduling problem encountered in various real-world scenarios. In this problem, you are given a set of segments, each with a start and finish index, and the objective is to select a maximum number of non-overlapping segments. Algorithm 1 applies a greedy strategy to find a set-segments that maximizes the utilization of a resource or minimizes conflicts.

Algorithm 1 Greedy Interval Sequence

$Let\, segments = \{s_0, s_1, s_2 ... s_{n-1}\}$ ▷ denote set of n segments
$non\_overlapped \leftarrow \emptyset$
$non\_overlappedt[0] \leftarrow segments[0]$
$segments\_end\_index \leftarrow s_0\{end\_index\}$
for $i \leftarrow 1\, to\, n-1$ **do**
    $start\_index \leftarrow s_i\{start\_index\}$
    $end\_index \leftarrow s_i\{end\_index\}$
    if $start\_index \geq segments\_end\_index$ **then**
        append $segments[i]$ to $non\_overlapped$
        $segments\_end\_index \leftarrow end\_index$
    end if
end for
return $non\_overlapped$

The greedy approach works because selecting the activity with the earliest completion rate frees up the resource for as many other activities as possible. This strategy maximizes the utilization of the resource and is proven to provide an optimal solution for this problem.

In the provided code, the sequence_segment function implements the greedy Algorithm 1, ensuring that it selects non-overlapping segments with the earliest finish time from the list of segments. The result of implementing the greedy algorithm in the previous example is shown in Figure 2.4, where it can be seen that the chosen segments are not overlapped. In other words, a sequence of non-overlapping segments has been obtained.

2.2.4 Group Aggregation and Profile

In Boratto [11], introduces different types of groups: *established groups* (shared and long-term common interests), *occasional groups* (common aim in a particular moment), *random groups* (do not have anything in common), and *automatically identified groups* (where individuals with similar preferences have to be grouped). In the context of Strava Groups, we work with *established groups*, as members come together due to shared and enduring inter-

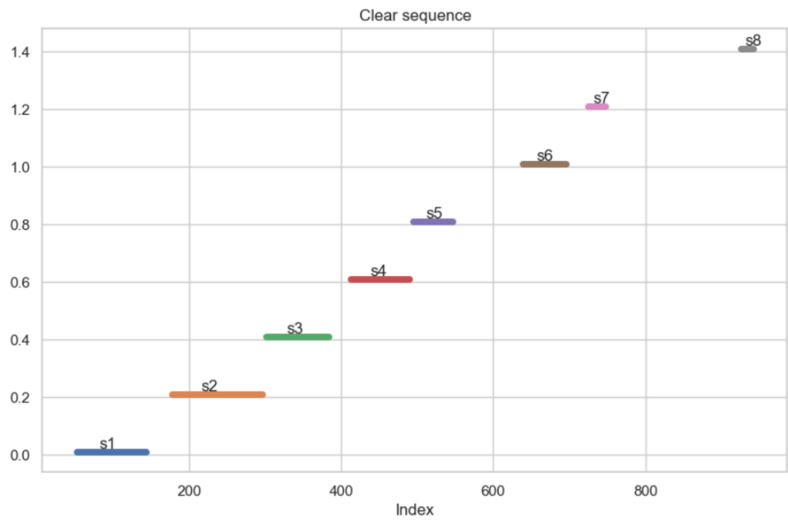


Figure 2.4: Final clean sequence segments.

ests, such as running, cycling, or swimming. These groups typically revolve around specific activities, such as local cycling clubs or training teams.

2.2.5 Group Recommendation Algorithm

In this subsection, we first describe key preliminary definitions and the problem statement. Next, we describe the sequence model variants that are sequentially ordered user-item interaction logs in the recommendation process, and self-attention based sequential model. We describe each component of our proposed models in the following sections.

2.2.5.1 Preliminaries and Problem Statement

Definition 1 Group users: denoted as g_k, we are given a set of users $U = \{u_1, u_{s2}, u_3..., u_m\}$, where each $g_k \subseteq U$ and k satisfies the condition $k \geq 2$, that all user in $g_k = \{u_1, u_{s2}, u_3..., u_k\}$ receives the same recommendations.

Definition 2 Segment: denoted as s, is defined as a uniquely identified location, which represents a specific route or portion of a route used for cyclists. A sequence represents route denoted as P, which is a set of segments u, $P = \{s_1, s_2, s_3..., s_p\}$ that the user travels sequentially.

Definition 3 Riding Segment: User ride segment is a quadruple $v_u^t = (s_u^t, l_u^t, t, u)$ that represents a user u riding segment s_u^t at location l_u^t at time stamp t.

Definition 4 Route activity: A user visit sequence is a set of activities visited by the user u, represented by $V_u = \{v^u_{t1}, v^u_{t2}, v^u_{t3}, ...v^u_{tn}\}$. The historical route sequences of all users in a dataset are represented as $V = \{V_{u1}, V_{u2}, V_{u3}, ..., V_{um}\}$, where m signifies the total count of users.

Definition 5 User segment Interests: Given the set of segments, denoted as $S = \{s_1, s_2, s_3, ..., s_p\}$, each segment, is characterized by attributes of interest observed in the segment dataset, such as $s_1 = (effort\_count, athlete\_count, star\_count)$.

Problem Statement: Given the information of all Strava user routes (visited segments V), the output of our proposed model is to recommend the sequence n segments to the user group g_k on the basis of the last activities of the group members, where that sequence forms the route to be performed by the group g_k.

2.2.5.2 Group-oriented Segment Sequence Algorithm

The Group-oriented Segment Sequence (GoSS-Rec) algorithm is designed for recommending a sequence of segments to a user group, taking into account historical routes (segment sequences) and a pre-trained recommendation model. The Algorithm 2 Group-oriented Segment Sequence (GoSS-Rec) starts by defining the input parameters, such as the historical route sequences for user group g_k and the recommendation model $Model$. The algorithm initializes key variables and calculates an initial recommendation Rs_0, which corresponds to the first segment to visit. It then iterates through the segments, generating recommendations and forming a cycling route path Cp by joining the segments. The final recommended sequence of segments $s0 - sn - 1, Rs_n$, is returned, along with the cycling route path Cp. This algorithm offers a formal and structured approach to leverage historical data and a recommendation model to provide a tailored sequence of segments for the user group.

2.2.6 Sequence Model Recommendation

- ❒ **PopRec:** It is the simplest baseline that ranks items according to their popularity judged by the number of interactions (i.e., number of associated actions, views, purchases, or interactions, the items have received).

- ❒ **MDP [29]:** MDP-based recommender systems can adapt to changing user prefer-

Algorithm 2 Group-oriented Segment Sequence (GoSS-Rec)

Require: V_U: Historical route sequences of all users.
Require: g_k: Group of users to recommend.
Require: $Model\,\{n, v_u^{t1}\}$: pre-train model sequence recommendation.

Define n, in base historical g_k users.
Initialize $Rs = \{selected\_segments\}$ recommended sequence of segments for g_k.
Initialize $Cp = \{finalrouterecommendation\}$.
$Rs_0 = segments\_end\_index = v_u^{t1} \leftarrow$ the end_index of the first segment.
$Rs_n = Model\,\{n, Rs_0\}$ generate sequence recommendation.
for i from 1 to $n-1$ **do**
    Append segment Rs_i to Cp
    Append segment Join segment Rs_i to Rs_{i+1} to Cp
end for
Return Rs_n as the recommended sequence of segments for group g_k, and Cp is cycling route path.

ences and evolving item catalogs by continuously updating their policies based on user interactions.

❒ **FPMC [43]:** It is a hybrid recommender system that combines Matrix Factorization with first-order Markov, which capture long-term preferences and dynamic transitions respectively.

❒ **Prod2vec [44]:** Recommendations are created by returning the k-nearest neighbors of the last items in the user profile, whose relevance is weighted using a simple exponential decay, where the last item in the user profile is the most relevant one, and the first item the least relevant.

❒ **KNN [44]:** The method considers the last element of a given session, and then returns as recommendations those elements that are most similar to it in terms of their co-occurrence in other sessions.

❒ **SASREC [34]:** It is a sequential next-item recommendation method based on the left-to-right transformer architecture. This strategy employs the multi-head self-attention mechanism to capture users' sequential behaviors and interactions.

2.2.7 Evaluation Metrics

The choice of offline evaluation [45] for this recommendation project is crucial since the recommender is not yet deployed, and there is currently no real-time user interaction. By utilizing historical or pre-existing datasets, we can simulate user behavior and assess the system's effectiveness before its deployment in production environments. This preliminary evaluation allows for parameter adjustments, correction of potential biases, and enhancement of recommendation diversity. Moreover, without the need for real-time interaction, it reduces complexity and associated costs, facilitating the comparison of different algorithms and model optimization.

2.2.7.1 Sequence Metric

We adopt common Top-N metrics, including *Normalized Discounted Cumulative Gain (NDCG)*. Equation 2.1 assumes a decreasing user interest with a logarithmic curve, paying more attention to possible relevant items in the middle of the list. This approach, widely used in the literature [46] [35], is reflected in the *Normalized Discounted Cumulative Gain* at K (NDCG@K):

$$\text{NDCG@}K = \frac{\text{DCG@}K}{\text{IDCG@}K} \tag{2.1}$$

where *Discounted Cumulative Gain (DCG@K)* accomplishes this by discounting the reward for items in lower ranks:

$$\text{DCG@}K = \sum_{i \in \{i | \text{rank}_u(i) \leq K\}} \frac{y_{u,i}}{\log_2(\text{rank}_u(i) + 1)} \tag{2.2}$$

and $y_{u,i}$ is either a binary label or a relevance score (such as a rating). *IDCG@K* is the ideal discounted cumulative gain; that is, the discounted cumulative gain that would have been achieved via an optimal ranking function ${rank_u}^{optu}(i)$ [47].

2.2.7.2 Novelty and Diversity

Other measures of quality beyond accuracy, including novelty and diversity, are considered in [48]. Novelty can be broadly defined as the distinction between current and previous experiences, while diversity pertains to the internal variations within components of an experience. Introducing novelty and diversity as key elements in achieving the desired outcome implies adopting a broader viewpoint in addressing the recommendation problem. This perspective focuses on the ultimate utility of recommendations, rather than solely emphasizing a single aspect such as accuracy.

Novelty: It is typically defined as the complement of the popularity of an item, and in the recommendation it expresses items that the user did not know. A variant is to define novelty as $-logp(i)$, which gives the self-information of an item i [49]. In comparison to a simple complement of popularity, this approach places more importance on very rare items. In this context Equation 2.3, the novelty of a recommendation list is given by the aggregation of individual novelties of each item in the list [50].

$$\mathrm{Novelty}(R) = \frac{\sum_{i \in R} - \log_2 p(i)}{|R|} \tag{2.3}$$

Where $p(i) = \frac{|\{u \in U, r_{u,i} \neq \bullet\}|}{|U|} \cdot |U|$ is the number of users who have voted for item i.

Diversity: It is defined as the opposite of similarity in the case of items that are of the items. The frequently considered diversity metric and the first to be proposed in the area is the so-called average intra-list distance—or intra-list diversity. The Equation 2.4 ILD (by Smyth and McClave [51]) suggests measuring the diversity of a recommendation list $R(|R| > 1)$ as the average distance between pairs of items in the list:

$$\mathrm{ILD} = \frac{1}{|R|(|R|-1)} \sum_{i \in R} \sum_{j \in R} d(i,j) \tag{2.4}$$

The dist function can be measured in several ways depending on the scope, by means of the Jaccard complement or Cosine similarity.

2.3 SUMMARY

In this chapter we discussed the DSR methodology used in the development of the project. We identified the Strava datasets and their data characteristics. We have addressed the preliminaries and problem statement for the sequential recommendation segments of a user group. Finally, we defined the GoSS-Rec recommendation algorithm in detail and the evaluation metrics to be used in the small and large group experiments explained in the next chapter.

3 RESULTS AND DISCUSSION

This chapter presents the results of the dataset specific to activities, clubs, and segments in the city of Quito, Ecuador. Additionally, the design of experiments, the results of the selected quality measures and the results comparing the different sequence models are presented.

3.1 RESULT: STRAVA DATASET

3.1.1 Activity Records

The Activity Records dataset contains information about cycling activities and segment sequences. Additionally, this dataset is specific to activities in the city of Quito, Ecuador. The dataset comprises 4 columns and 18,441 rows. The columns are defined as follows, id: Integer data type, user_id: Integer data type, start_date: Integer data type, segment: a dictionary containing start_index (integer data type) and end_index (integer data type) An example is show in Table 3.1. Each row in the dataset represents a record of a cycling activity, and describes the activity identifier, user identifier, start date of the activity, and segment information, which includes start and end indices in the dataset.

The provided results offer descriptive statistics about the lengths of sequences in the dataset. With a total count of 18,441 sequences, the average length is approximately 11 segments, with a standard deviation of around 9.49, indicating notable variability in sequence lengths. The minimum sequence length is 2, while the 25th percentile reveals that a quarter of the se-

id	user_id	start_date	segment
3	1	1678891235	{’start_index’: 80, ’end_index’: 193, ’start_date’: 1673824420.0, ’id’: 731}, {’start_index’: 193, ’end_index’: 229, ’start_date’: 1673824848.0, ’id’: 298}, {’start_index’: 249, ’end_index’: 888, ’start_date’: 1673825062.0, ’id’: 458}, {’start_index’: 900, ’end_index’: 985, ’start_date’: 1673827658.0, ’id’: 3176}

Table 3.1: Activity Dataset Example

quences have a length of 5 or less. The median (50th percentile) is 9, suggesting that half of the sequences are 9 or less. Moving to the 75th percentile, three-quarters of the sequences are 14 or less. The maximum sequence length is 169. In summary, these statistics provide insights into the distribution of sequence lengths, as shown in Figure 3.1, highlighting both central tendencies and the extent of variability in the dataset.

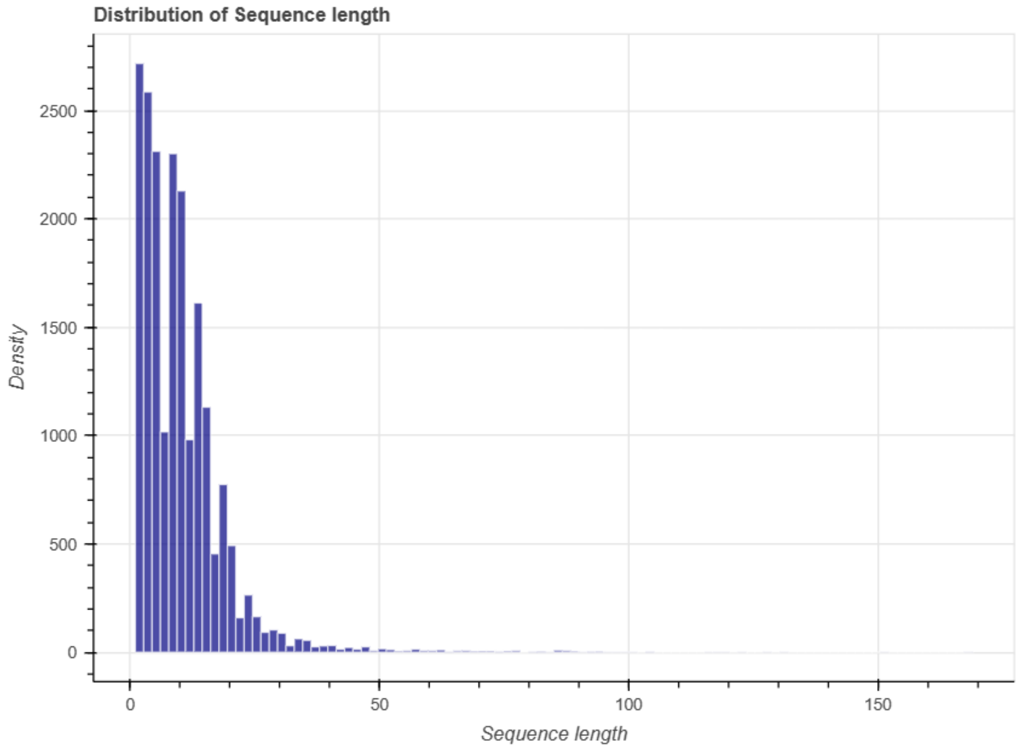


Figure 3.1: Distribution of sequence lengths

3.1.2 Segment Information

In this dataset, a comprehensive overview of 5817 segments is presented, each uniquely identified. Additionally, this dataset is specific to segments in the city of Quito, Ecuador. The numerical features provide valuable insights about the nature of the segments, encompassing distances traveled, average gradients, elevation metrics, climb categories, total elevation gains, and various counts related to athlete engagement and recognition (efforts, participants, and stars). Geographical coordinates of start and end points allow for potential spatial analyses. Visualization, such as histograms and maps are powerful aids in uncovering numerical and spatial trends.

Figure 3.2 shows the Segments dataset. The visual representation highlights a significant concentration of segments within the metropolitan district of Quito and its environs. Actually, this geographic focus adds crucial contextual information for the interpretation of the data presented. The criteria for categorizing climbs on Strava include a minimum average gradient of 3.0%, a segment distance of at least 300 meters, and a calculated value derived from the climb's length and grade exceeding 8,000. Then, so for example the category 0 segments in red are flat segments or descents.

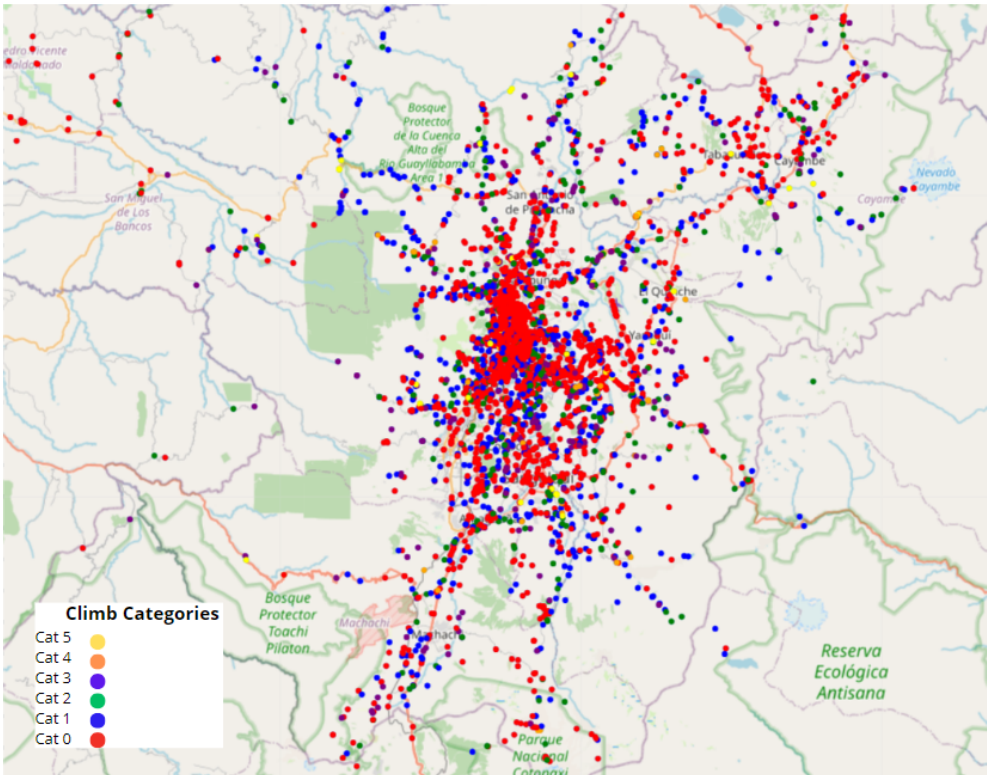


Figure 3.2: Segment Climb Categories.

3.1.3 Clubs

This dataset comprises 1431 entries with five columns: id, sport_type, member_count, id_user, and new_id. The member_count column denotes the number of members, with a mean of 90030.91 and a standard deviation of 598380.1. The range spans from 1 to 4847928. The sport column categorizes activities into cycling, running, and triathlon.

There are 152 groups of varying sizes and compositions. These groups are characterized as occasional groups given that these members share a common objective during specific moments of engaging in sports. That is to say, these groups represent gatherings driven by a common purpose, such as participating in a certain sporting activity. The diversity in their sizes and compositions underscores the dynamic nature of these cohorts, highlighting the multifaceted nature of collective engagement in sports at different points in time. In Figure 3.3, the groups are visually depicted, showcasing their different sizes ranging from 2 users to as many as 23 users. This visual representation provides a clear overview of the distribution and diversity of these occasional sports groups.

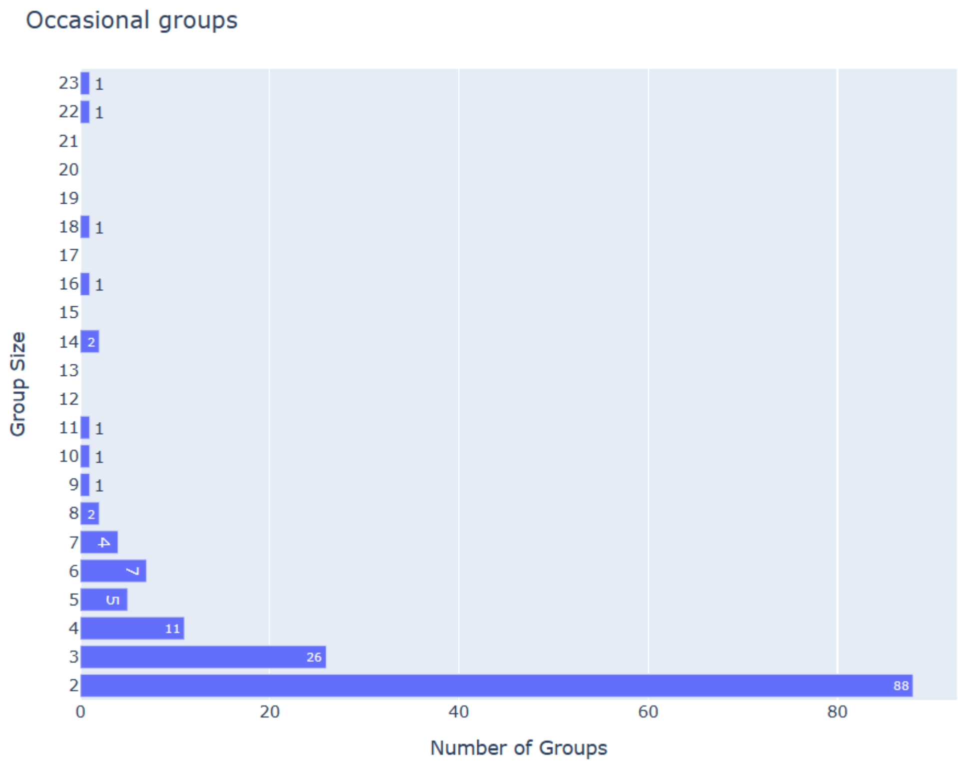


Figure 3.3: Strava clubs group detection

3.2 DESIGN OF EXPERIMENTS

This section shows the design of experiments. The experiments focus on: a) finding the optimal parameters for the proposed method, and b) comparing the result of the models with the proposed method. The subsection describes the environment designed to perform all the experiments and presents the subsection presents the evaluation measures for the recommendations. Finally, we explain the experiments carried out to test the proposed model.

Method	Parameters
PopRec	Popularity Recommender doesn't have any hyper-parameter.
MDP	Mixed Markov Chain minimum order =1, maximum order =1.
FPMC	latent factors =2, number of training epochs 5.
Prod2vec	2 minimum item frequency, size of embeddings =5, window=5, exponential decay=0.9.
KNN	k=10, similarity jaccard.
SASREC	2 self-attention blocks, learning rate 0.001, batch size 128, dropout 0.2.

Table 3.2: Parameters used in the experiments

3.2.1 Experimental setup

In this section, the application of recommendations to the models outlined in Section 2.2.6 is showcased through offline evaluation. By employing 152 groups and taking, for instance, a group with an average segment of 10, which represents the mean segment per user activity in the dataset, these examples provide insights into the performance of the models and recommender algorithms. They demonstrate the ability of the system to generate relevant suggestions based on historical data. This offline evaluation sheds light on the effectiveness of the models and algorithms in delivering group recommendations.

Users have been grouped according to the club they belong to, although users belonging to more clubs in common are more likely to be grouped together. In addition, we used the Average Strategy as an aggregation strategy to sum the individual ratings into a group rating. To conduct the experiments, we divided the activities into test and training sets.

Some of these recommendation methods require different parameters to work. We have configured and performed parameter tuning to maximize the quality of the recommendation for the data set. Table 3.2 contains the adjusted parameters for each model.

We will evaluate these methods using the previously defined quality measures. The experiment has been conducted for two group sizes and mean segments of group:

a) Size 2 to 5 users, to test the performance of each recommendation method on small groups of users.

b) Size from 6 to 23 users, to test the performance of each recommendation method on a large group of users.

c) All groups with an average of 10 segments per activity, representing the average segment per user activity in the data set.

EXPERIMENT	Metric	PopRec	MDP	FPMC	Prod2vec	KNN	SASREC
	NDCG@10	0.6324	0.6399	0.5838	0.765	**0.7816**	0.6962
Group size 2-5	**Novelty**	0.6598	1.3951	1.1516	**2.8971**	1.3048	0.7459
	Diversity	0.0976	**0.2123**	0.196	0.1694	0.1785	0.1785
	NDCG@10	0.7802	0.7626	0.7112	**0.8375**	0.7743	0.7818
Group size 6-23	**Novelty**	0.6756	1.1405	0.9855	**2.3056**	1.1636	0.7014
	Diversity	0.0928	0.1839	**0.2304**	0.1491	0.1671	0.1671
	NDCG@10	0.6538	0.6576	0.6022	0.7755	**0.7805**	0.7246
All groups n=10	**Novelty**	0.6621	1.3582	1.1275	**2.8115**	1.2844	0.7394
	Diversity	0.0969	**0.2082**	0.201	0.1664	0.1768	0.1768

Table 3.3: Results of recommendation to groups.

3.3 RESULTS AND DISCUSSION

Section 3.3.1 shows the prediction and recommendation results obtained using the proposed method, with each model being compared, and discusses the main strengths and drawbacks observed.

3.3.1 Comparison of results

Table 3.3 corresponds to small user groups (groups containing 2 to 5 users), large user groups (groups containing 6 to 23 users), and finally the results of the metrics for all groups with an average of 10 segments per activity. The most relevant conclusions obtained are as follows:

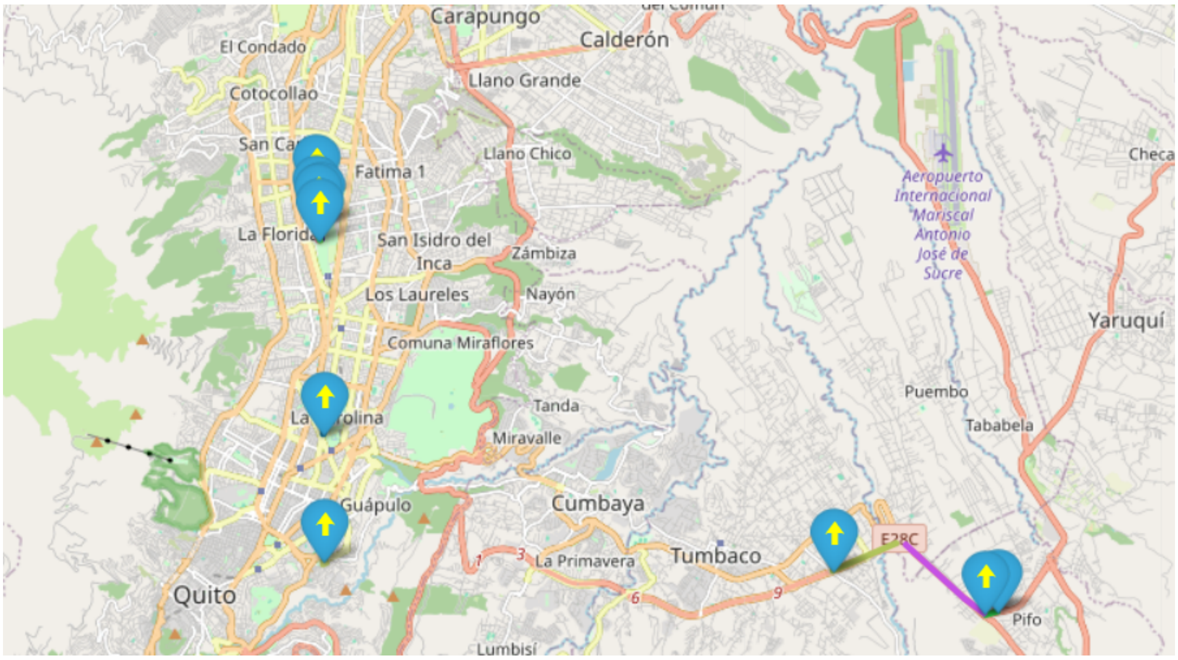
- In general, the quality of the Normalized Discounted Cumulative Discounted Gain (NDCG) increases as the groups are larger than 6 users. This is the expected result, due to the higher diversity of users in smaller groups and higher number of user groups smaller than 6 users per group.

- The result of applying the GoSS-Rec algorithm 2 and evaluating the results in the designed experiments shows the Prod2vec model with better performance in diversity and novelty metrics in small and large groups.

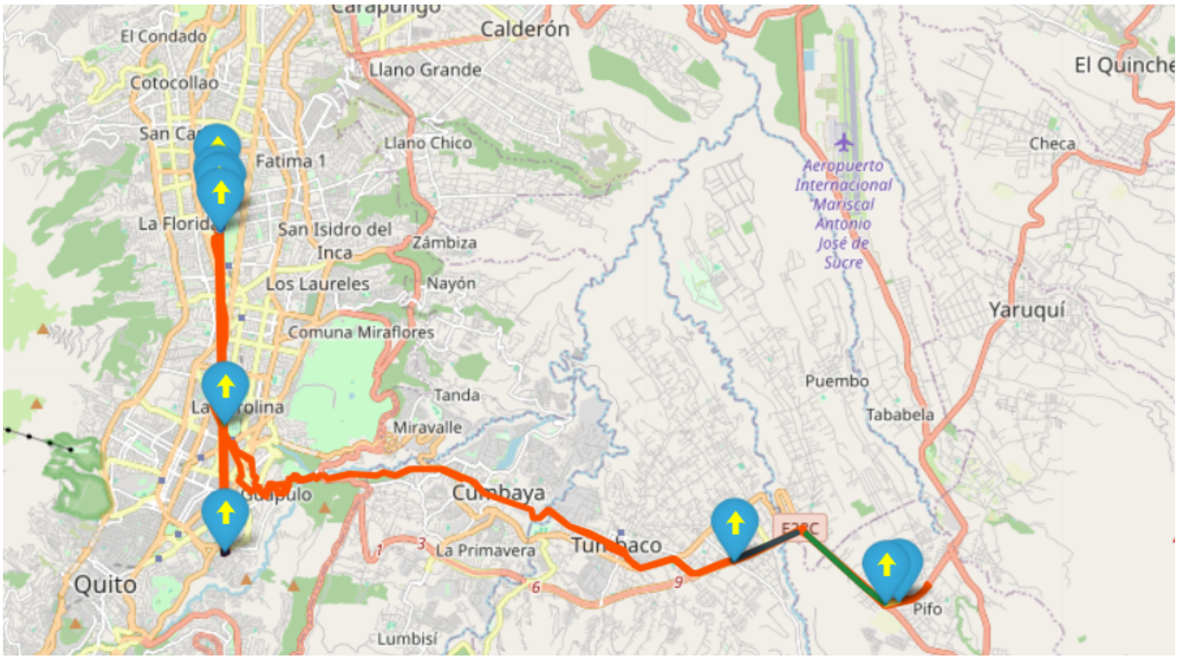
- Analyzing the experimental results in the offline evaluation. It is observed that Prod2vec has greater novelty and diversity for the different group sizes 6-23. However, SASREC achieves superior NDCG@10 for all groups, underscoring its strength in the overall effectiveness of the recommendations.

3.3.2 Results final route example generated

We present a toy example of the application of the GoSS-Rec algorithm with the trained models for the parameters defined in Table 3.2. Figure 3.4a presents the set of segments generated by the PopRec model, while Figure 3.4b presents the final route generated where the segments already form a final route. Figure 3.5a presents the recommendations that are generated by ordering the elements by their transition probability to be the next one, given the group profile, and then 3.5b presents the final route generated where the segments already form a final route. The result of segment recommendation with the FPMC model is presented in Figure 3.6a and the total generated route in Figure 3.6b. Figure 3.7a is the result of the recommendations with the Prod2vec model, where the recommendations that are generated by returning the k nearest neighbors of the last elements in the user profile, whose relevance is weighted by a simple exponential decay and then Figure 3.7b presents the final path. The KNN recommendation model that compares the entire current session with previous sessions in the training data to determine the elements to be recommended is presented in Figure 3.8a, and the total path constructed with the segments generated using the recommender is shown in Figure 3.8b. The result of the recommendation with the KNN model for the recommendation of the next segments is presented in the Figure 3.9a and the complete route generated is presented below 3.9b.

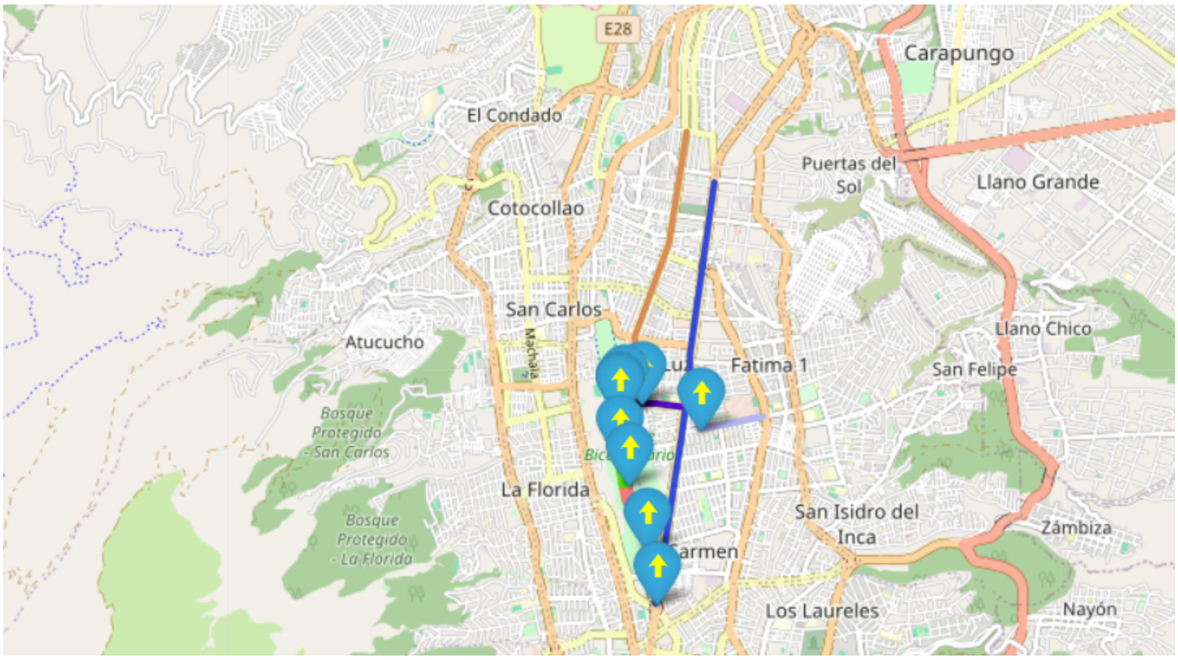


(a) PopRec Results n=10

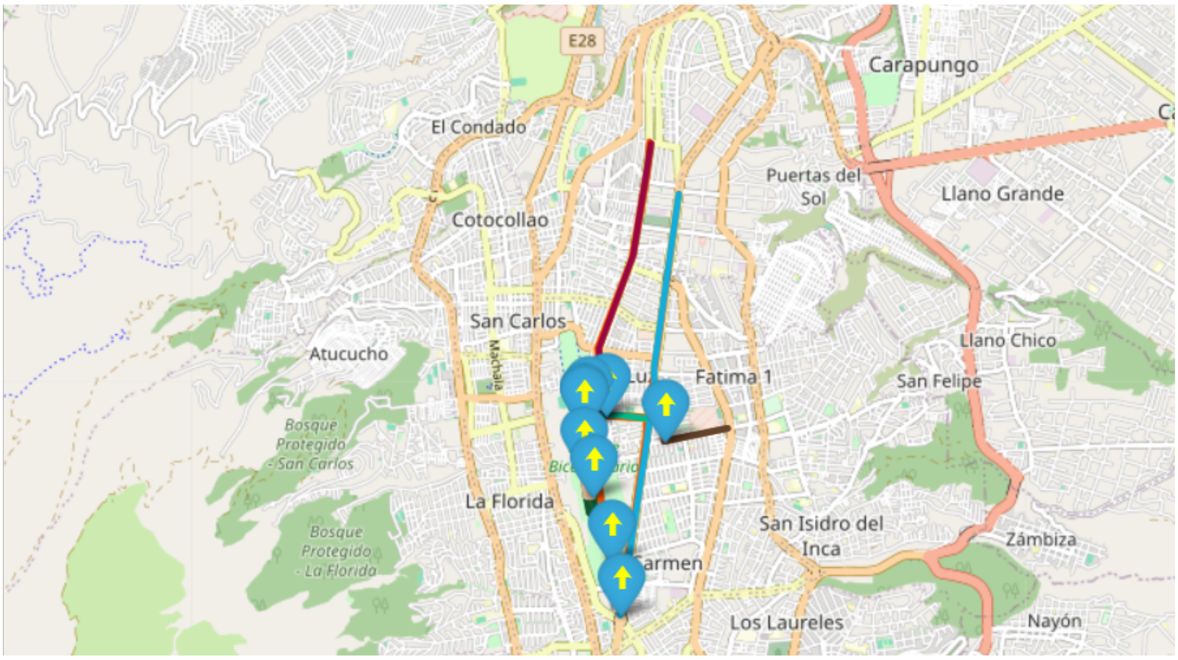


(b) PopRec Results n=10 route

Figure 3.4: PopRec Results

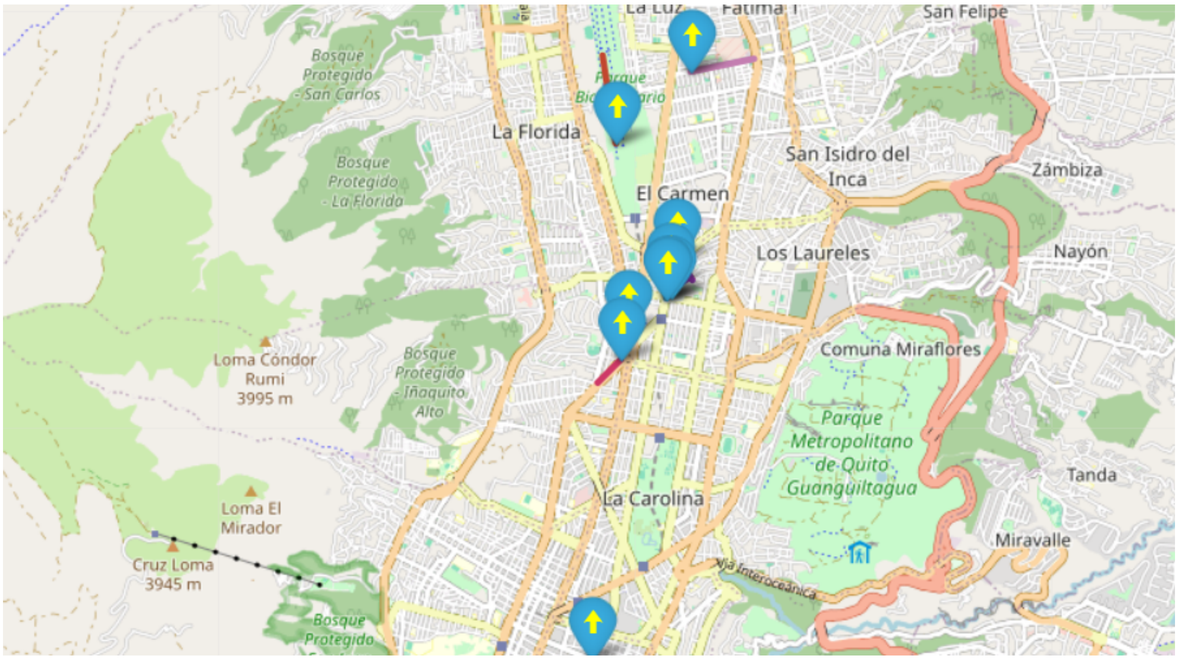


(a) MDP Results n=10

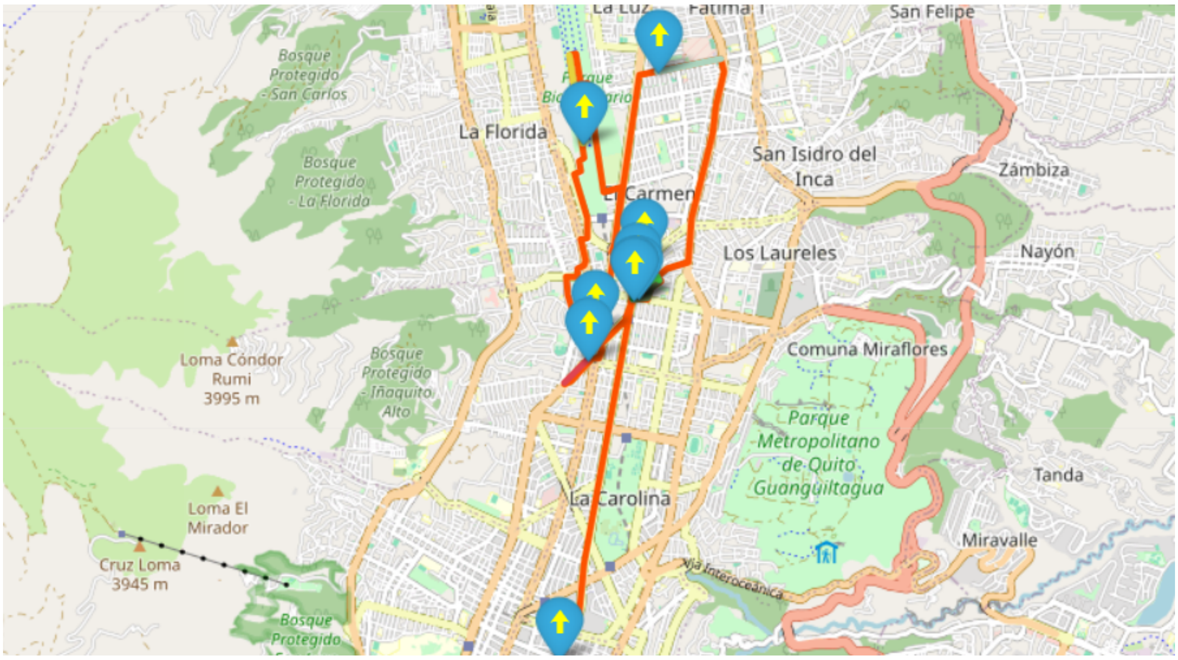


(b) MDP Results n=10 route

Figure 3.5: MDP Results

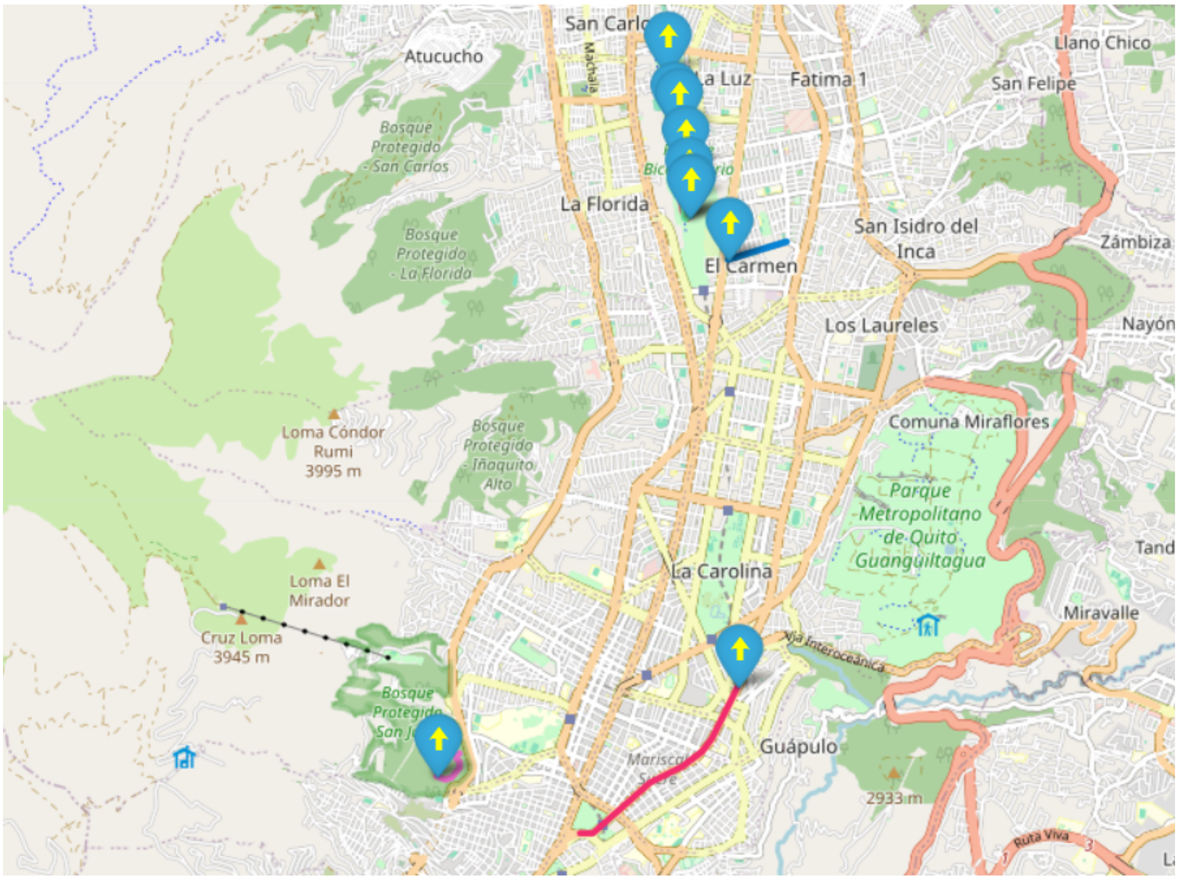


(a) FPMC Results n=10

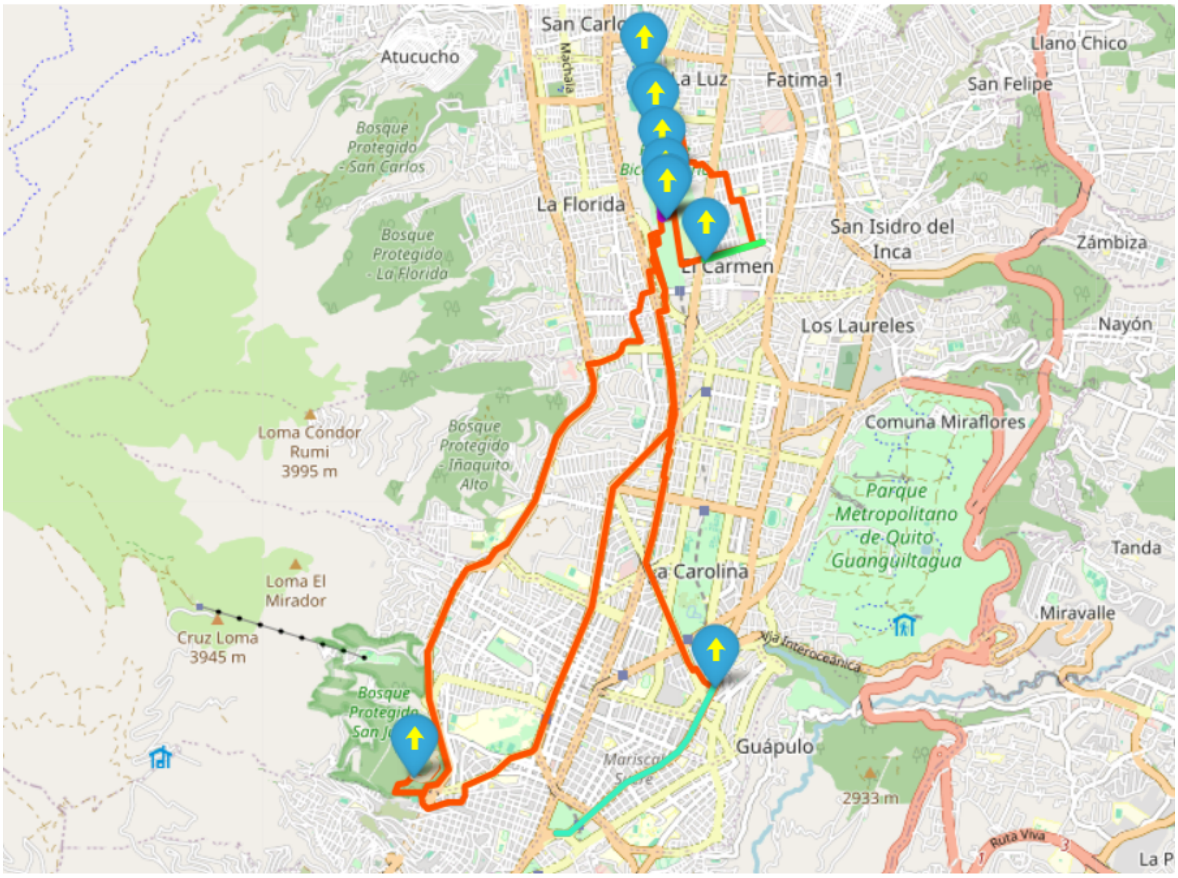


(b) FPMC Results n=10 route

Figure 3.6: FPMC Results

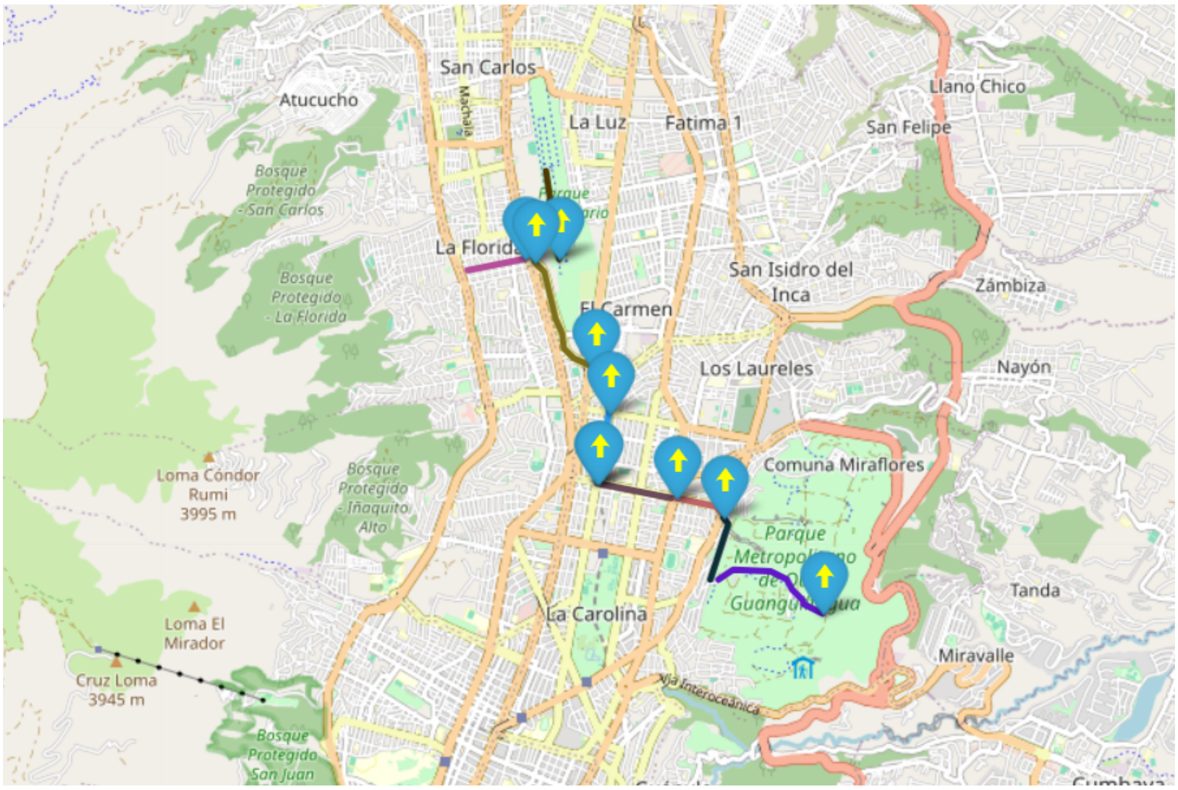


(a) Prod2vec Results n=10

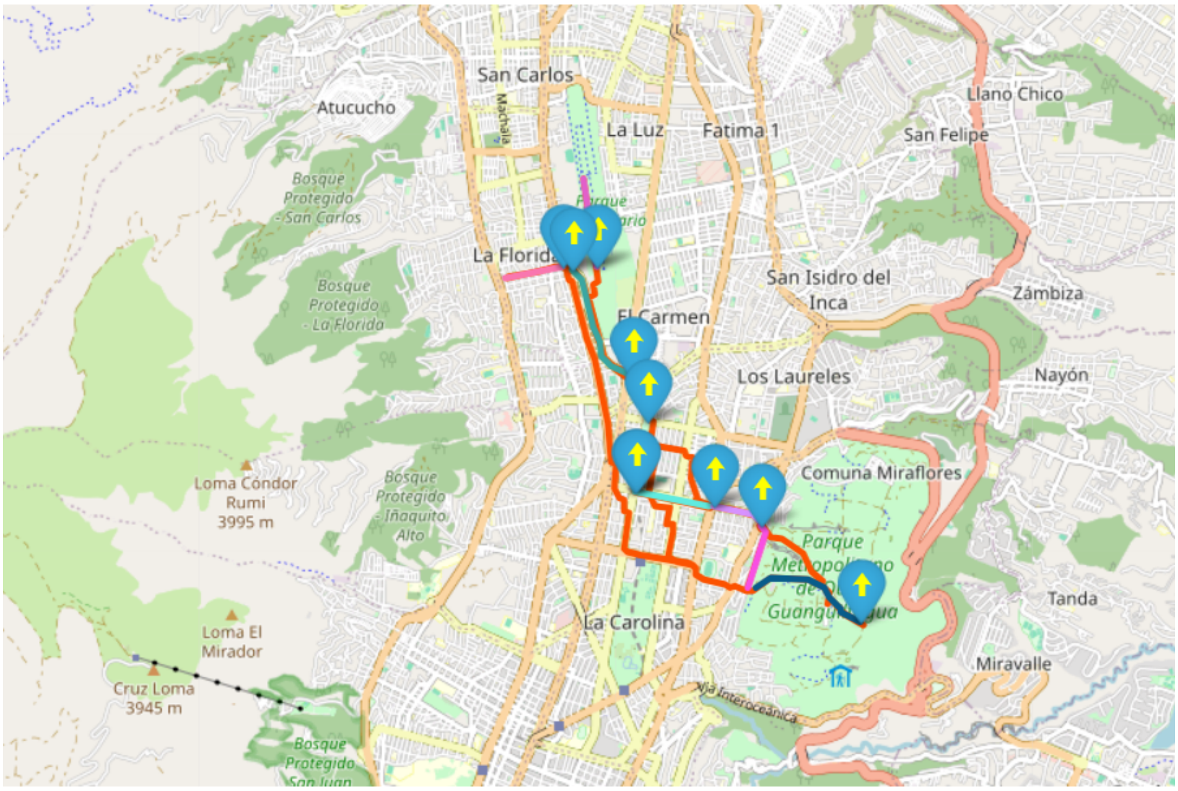


(b) Prod2vec Results n=10 route

Figure 3.7: Prod2vec Results

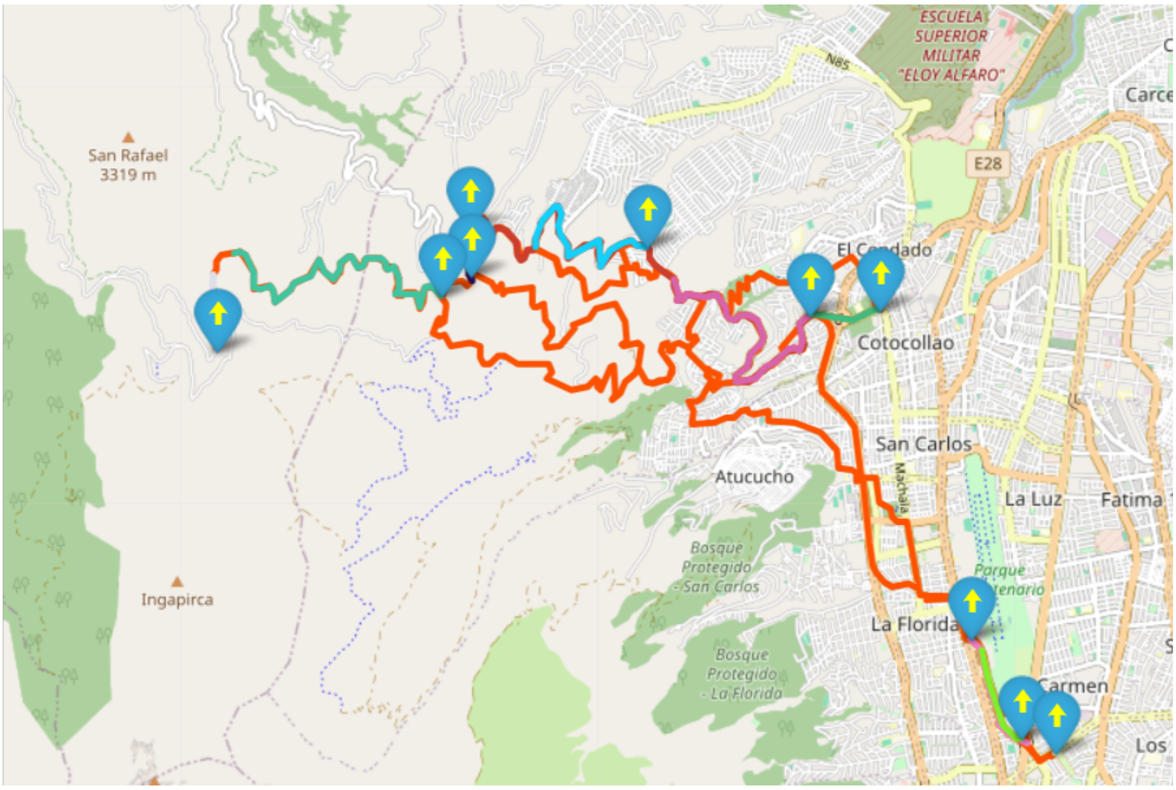


(a) KNN Results n=10

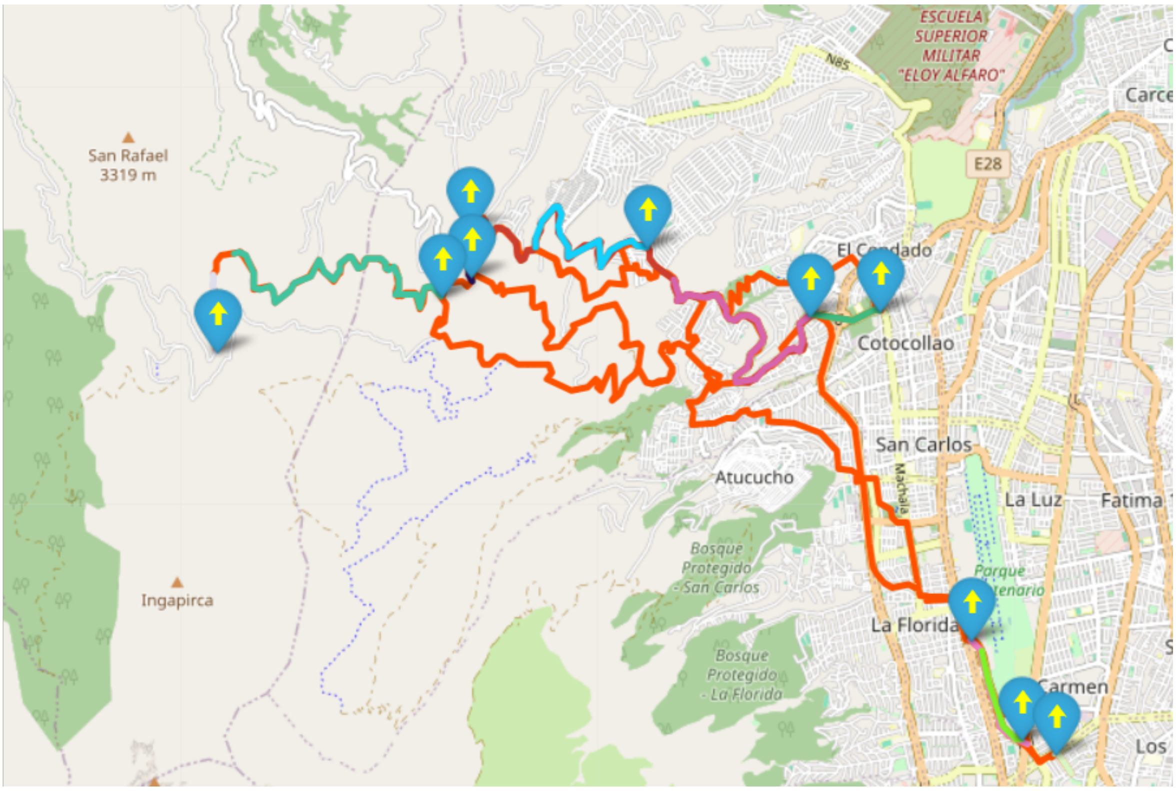


(b) KNN Results n=10 route

Figure 3.8: KNN Results



(a) SASREC Results n=10



(b) SASREC Results n=10 route

Figure 3.9: SASREC Results

4 CONCLUSIONS AND FUTURE WORK

This chapter concludes the research efforts towards GRSs based on sequential recommender algorithms. First, it highlights the contributions of this thesis: 1) it provides a dataset where users, activities, segments, clubs are highlighted as useful information for further research. 2) we also propose an algorithm that generates group recommendations through a new measure of number of segments per activity for groups of users, segments and clubs. At the end of this chapter, future work is proposed based on the following.

4.1 CONCLUSIONS

The primary aim of this project was to craft a sophisticated and personalized route recommendation model explicitly tailored for groups of cyclists. The chosen methodology focused on integrating sequence-aware recommender systems to enrich the cycling experience through intelligently curated route suggestions.

In accomplishing the foundational steps, an exhaustive literature review was conducted to explore the intricacies of GRSs and Sequence-Aware Recommender Systems within the context of route recommendations. This critical analysis provided a nuanced understanding of existing methodologies, setting the stage for subsequent phases of the project.

The following step involved the application of data mining techniques, facilitated by the Strava API, to extract and characterize pertinent data. Beyond the extraction of relevant information, this process laid the foundation for user profiling and characterization of route items, essential components to generate personalized recommendations that addressed the diverse preferences of cycling groups.

Drawing from the insights gained through the literature review and data mining efforts, a novel personalized route recommendation model rooted in Sequence-Aware Recommender Systems was introduced as the next step. This innovative model, crafted to account for the sequential nature of cycling activities, promised to deliver refined and context-aware suggestions, aligning closely with the diverse needs of group cyclists.

The Strava dataset curated from the Strava API encompasses a rich and comprehensive repository of information, which is crucial for the development and evaluation of personalized route recommendation models. The dataset, structured around cycling activities, segments, and Strava Clubs, provides intricate details such as user and segment identifiers, activity timestamps, geographical coordinates, and various performance metrics. This wealth of information, ranging from individual efforts to broader club dynamics, not only enables a granular understanding of cyclists' behaviors but also opens avenues for exploring the diverse characteristics of cycling segments. The dataset's inclusion of effort counts, athlete participation, elevation data, and user ratings enriches the analytical potential, laying a robust foundation for deriving meaningful insights and enhancing the precision of future route recommendation algorithms.

The results highlight a positive correlation between NDCG and user group size, revealing an improvement in performance as groups exceed 6 users. This observation implies that the NDCG metric is more effective on larger user cohorts. The rationale for this phenomenon is related to the greater diversity of larger groups, compared to smaller ones, and the prevalence of a considerable number of user groups with fewer than 6 members.

When training the models and using GoSS-Rec algorithm and evaluating the results in the designed experiments, the Prod2vec model demonstrates better performance on diversity and novelty metrics in both small and large groups. This suggests that, at least in the context of the designed experiments, the GoSS-Rec algorithm with Prod2vec outperforms other alternatives in terms of diversity and novelty metrics in groups of different sizes.

The successful execution of these foundational elements underscores the project's significance as a noteworthy contribution to the field of personalized route recommendation systems. The integrated approach, combining literature insights, data-driven techniques, model innovation, rigorous evaluation, and scholarly dissemination, positions the project at the forefront of advancements in recommending cycling routes for groups.

4.2 FUTURE WORK

Deployment of Application for Strava Users: The deployment of an application enabling Strava users to plan their routes based on generated recommendations represents a significant avenue for future work. This initiative not only enhances user engagement by providing a valuable planning tool but also facilitates the collection of a larger volume of new data. This, in turn, sets the stage for conducting real-time evaluations and analyses. The integration of user-generated route plans with the recommendation system creates a dynamic feedback loop, allowing for continuous improvement and refinement of the personalized route recommendation model.

Measurement of User Engagement and Effectiveness: The deployment of the application opens up opportunities for measuring various aspects, including user engagement and the effectiveness of the recommendation model. Through analytics and user-interaction data, it becomes possible to evaluate how frequently users engage with the route planning feature, providing insights into the application's overall utility. Additionally, user feedback on the recommended routes can be systematically gathered to assess the accuracy and relevance of the suggestions. This data-driven approach to measurement creates a foundation for ongoing optimization and refinement of the personalized route recommendation model, ensuring its alignment with the evolving needs and preferences of Strava users.

5 REFERENCES

[1] D. J. Rivers, “Strava as a discursive field of practice: Technological affordances and mediated cycling motivations,” *Discourse, Context Media*, vol. 34, p. 100 345, 2020, ISSN: 2211-6958. DOI: https://doi.org/10.1016/j.dcm.2019.100345. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2211695819301576.

[2] D. Lupton, S. Pink, C. Heyes LaBond, S. Sumartojo, *et al.*, “Digital traces in context: Personal data contexts, data sense, and self-tracking cycling,” 2018.

[3] S. Carter, J. Green, and E. Speed, “Digital technologies and the biomedicalisation of everyday activities: The case of walking and cycling,” *Sociology Compass*, vol. 12, no. 4, e12572, 2018, e12572 SOCO-1284.R2. DOI: https://doi.org/10.1111/soc4.12572. eprint: https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/soc4.12572. [Online]. Available: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/soc4.12572.

[4] J. Couture, “Reflections from the ‘strava-sphere’: Kudos, community, and (self-)surveillance on a social network for athletes,” *Qualitative Research in Sport, Exercise and Health*, vol. 13, no. 1, pp. 184–200, 2021. DOI: 10.1080/2159676X.2020.1836514.

[5] D. P. McArthur and J. Hong, “Visualising where commuting cyclists travel using crowd-sourced data,” *Journal of Transport Geography*, vol. 74, pp. 233–241, 2019, ISSN: 0966-6923. DOI: https://doi.org/10.1016/j.jtrangeo.2018.11.018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0966692317307445.

[6] G. Bruinvels, E. Goldsmith, R. Blagrove, *et al.*, “Prevalence and frequency of menstrual cycle symptoms are associated with availability to train and compete: A study of 6812 exercising women recruited using the strava exercise app,” *British journal of sports medicine*, vol. 55, no. 8, pp. 438–443, 2021.

[7] D. Jannach and M. Zanker, “Value and impact of recommender systems,” in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. New York, NY:

Springer US, 2022, pp. 519–546, ISBN: 978-1-0716-2197-4. DOI: 10.1007/978-1-0716-2197-4_14. [Online]. Available: https://doi.org/10.1007/978-1-0716-2197-4_14.

[8] J. Ni, L. Muhlstein, and J. McAuley, "Modeling heart rate and activity data for personalized fitness recommendation," in *The World Wide Web Conference*, ser. WWW '19, San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 1343–1353, ISBN: 9781450366748. DOI: 10.1145/3308558.3313643. [Online]. Available: https://doi.org/10.1145/3308558.3313643.

[9] A. Hevner, S. Chatterjee, A. Hevner, and S. Chatterjee, "Design science research in information systems," *Design research in information systems: theory and practice*, pp. 9–22, 2010.

[10] F. Ricci, L. Rokach, and B. Shapira, "Recommender systems: Techniques, applications, and challenges," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. New York, NY: Springer US, 2022, pp. 1–35, ISBN: 978-1-0716-2197-4. DOI: 10.1007/978-1-0716-2197-4_1. [Online]. Available: https://doi.org/10.1007/978-1-0716-2197-4_1.

[11] A. Felfernig, L. Boratto, M. Stettinger, and M. Tkalčič, "Decision tasks and basic algorithms," in *Group Recommender Systems : An Introduction*. Cham: Springer International Publishing, 2018, pp. 3–26, ISBN: 978-3-319-75067-5. DOI: 10.1007/978-3-319-75067-5_1. [Online]. Available: https://doi.org/10.1007/978-3-319-75067-5_1.

[12] A. Felfernig, G. Friedrich, and L. Schmidt-Thieme, "Guest editors' introduction: Recommender systems," *IEEE Intelligent systems*, vol. 22, no. 3, pp. 18–21, 2007.

[13] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: A survey," *Decision Support Systems*, vol. 74, pp. 12–32, 2015, ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2015.03.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167923615000627.

[14] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," *IEEE Internet Computing*, vol. 7, no. 1, pp. 76–80, 2003. DOI: 10.1109/MIC.2003.1167344.

[15] B. Smith and G. Linden, "Two decades of recommender systems at amazon.com," *IEEE Internet Computing*, vol. 21, no. 3, pp. 12–18, 2017. DOI: 10.1109/MIC.2017.72.

[16] D. Jannach and M. Zanker, "Value and impact of recommender systems," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. New York, NY: Springer US, 2022, pp. 519–546, ISBN: 978-1-0716-2197-4. DOI: 10.1007/978-1-0716-2197-4_14. [Online]. Available: https://doi.org/10.1007/978-1-0716-2197-4_14.

[17] Y. Koren and R. Bell, "Advances in collaborative filtering," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. Boston, MA: Springer US, 2015, pp. 77–118, ISBN: 978-1-4899-7637-6. DOI: 10.1007/978-1-4899-7637-6_3. [Online]. Available: https://doi.org/10.1007/978-1-4899-7637-6_3.

[18] Y. Koren, S. Rendle, and R. Bell, "Advances in collaborative filtering," *Recommender systems handbook*, pp. 91–142, 2021.

[19] S. S. Khanal, P. Prasad, A. Alsadoon, and A. Maag, "A systematic review: Machine learning based recommendation systems for e-learning," *Education and Information Technologies*, vol. 25, pp. 2635–2664, 2020.

[20] Y. Koren, "Factorization meets the neighborhood: A multifaceted collaborative filtering model," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 426–434.

[21] E. Çano and M. Morisio, "Hybrid recommender systems: A systematic literature review," *Intelligent data analysis*, vol. 21, no. 6, pp. 1487–1524, 2017.

[22] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," *ACM computing surveys (CSUR)*, vol. 52, no. 1, pp. 1–38, 2019.

[23] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T.-S. Chua, "Neural collaborative filtering," in *Proceedings of the 26th international conference on world wide web*, 2017, pp. 173–182.

[24] Y. Wu, C. DuBois, A. X. Zheng, and M. Ester, "Collaborative denoising auto-encoders for top-n recommender systems," in *Proceedings of the ninth ACM international conference on web search and data mining*, 2016, pp. 153–162.

[25] M. Chen, A. Beutel, P. Covington, S. Jain, F. Belletti, and E. H. Chi, "Top-k off-policy correction for a reinforce recommender system," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019, pp. 456–464.

[26] H. Zhang, E. Yuan, W. Guo, *et al.*, "Disentangling past-future modeling in sequential recommendation via dual networks," in *Proceedings of the 31st ACM International Conference on Information Knowledge Management*, ser. CIKM '22, Atlanta, GA, USA: Association for Computing Machinery, 2022, pp. 2549–2558, ISBN: 9781450392365. DOI: 10.1145/3511808.3557289. [Online]. Available: https://doi.org/10.1145/3511808.3557289.

[27] M. Quadrana, P. Cremonesi, and D. Jannach, "Sequence-aware recommender systems," *ACM Comput. Surv.*, vol. 51, no. 4, Jul. 2018, ISSN: 0360-0300. DOI: 10.1145/3190616. [Online]. Available: https://doi.org/10.1145/3190616.

[28] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, "Factorizing personalized markov chains for next-basket recommendation," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10, Raleigh, North Carolina, USA: Association for Computing Machinery, 2010, pp. 811–820, ISBN: 9781605587998. DOI: 10.1145/1772690.1772773. [Online]. Available: https://doi.org/10.1145/1772690.1772773.

[29] G. Shani, D. Heckerman, R. I. Brafman, and C. Boutilier, "An mdp-based recommender system.," *Journal of Machine Learning Research*, vol. 6, no. 9, 2005.

[30] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," *arXiv preprint arXiv:1511.06939*, 2015.

[31] A. Beutel, P. Covington, S. Jain, *et al.*, "Latent cross: Making use of context in recurrent recommender systems," ser. WSDM '18, Marina Del Rey, CA, USA: Association for Computing Machinery, 2018, pp. 46–54, ISBN: 9781450355810. DOI: 10.1145/3159652.3159727. [Online]. Available: https://doi.org/10.1145/3159652.3159727.

[32] Y. K. Tan, X. Xu, and Y. Liu, "Improved recurrent neural networks for session-based recommendations," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, ser. DLRS 2016, Boston, MA, USA: Association for Computing Machinery, 2016, pp. 17–22, ISBN: 9781450347952. DOI: 10.1145/2988450.2988452. [Online]. Available: https://doi.org/10.1145/2988450.2988452.

[33] J. Yu, H. Yin, X. Xia, T. Chen, J. Li, and Z. Huang, "Self-supervised learning for recommender systems: A survey," *IEEE Transactions on Knowledge and Data Engineering*, 2023.

[34] W.-C. Kang and J. McAuley, "Self-attentive sequential recommendation," in *2018 IEEE international conference on data mining (ICDM)*, IEEE, 2018, pp. 197–206.

[35] F. Sun, J. Liu, J. Wu, *et al.*, "Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, ser. CIKM '19, Beijing, China: Association for Computing Machinery, 2019, pp. 1441–1450, ISBN: 9781450369763. DOI: 10.1145/3357384.3357895. [Online]. Available: https://doi.org/10.1145/3357384.3357895.

[36] J. Masthoff and A. Delić, "Group recommender systems: Beyond preference aggregation," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. New York, NY: Springer US, 2022, pp. 381–420, ISBN: 978-1-0716-2197-4. DOI: 10.1007/978-1-0716-2197-4_10. [Online]. Available: https://doi.org/10.1007/978-1-0716-2197-4_10.

[37] J. Masthoff, "Group recommender systems: Aggregation, satisfaction and group attributes," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. Boston, MA: Springer US, 2015, pp. 743–776, ISBN: 978-1-4899-7637-6. DOI: 10.1007/978-1-4899-7637-6_22. [Online]. Available: https://doi.org/10.1007/978-1-4899-7637-6_22.

[38] A. Felfernig, L. Boratto, M. Stettinger, and M. Tkalčič, "Algorithms for group recommendation," in *Group Recommender Systems : An Introduction*. Cham: Springer International Publishing, 2018, pp. 27–58, ISBN: 978-3-319-75067-5. DOI: 10.1007/978-3-319-75067-5_2. [Online]. Available: https://doi.org/10.1007/978-3-319-75067-5_2.

[39] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.

[40] O. Arazy, N. Kumar, and B. Shapira, "A theory-driven design framework for social recommender systems," *Journal of the Association for Information Systems*, vol. 11, no. 9, p. 2, 2010.

[41] P. Hanafizadeh, M. Barkhordari Firouzabadi, and K. M. Vu, "Insight monetization intermediary platform using recommender systems," *Electronic Markets*, vol. 31, no. 2, pp. 269–293, 2021.

[42] E. Eryilmaz, B. Thoms, K.-H. Lee, and M. de Castro, “Development of a reading material recommender system based on design science research approach,” 2019.

[43] S. Rendle, C. Freudenthaler, and L. Schmidt-Thieme, “Factorizing personalized markov chains for next-basket recommendation,” in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW ’10, Raleigh, North Carolina, USA: Association for Computing Machinery, 2010, pp. 811–820, ISBN: 9781605587998. DOI: 10.1145/1772690.1772773. [Online]. Available: https://doi.org/10.1145/1772690.1772773.

[44] M. Quadrana, D. Jannach, and P. Cremonesi, “Tutorial: Sequence-aware recommender systems,” in *Companion Proceedings of The 2019 World Wide Web Conference*, ser. WWW ’19, San Francisco, USA: Association for Computing Machinery, 2019, p. 1316, ISBN: 9781450366755. DOI: 10.1145/3308560.3320091. [Online]. Available: https://doi.org/10.1145/3308560.3320091.

[45] A. Gunawardana, G. Shani, and S. Yogev, “Evaluating recommender systems,” in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. New York, NY: Springer US, 2022, pp. 547–601, ISBN: 978-1-0716-2197-4. DOI: 10.1007/978-1-0716-2197-4_15. [Online]. Available: https://doi.org/10.1007/978-1-0716-2197-4_15.

[46] R. He, W.-C. Kang, and J. McAuley, “Translation-based recommendation,” in *Proceedings of the Eleventh ACM Conference on Recommender Systems*, ser. RecSys ’17, Como, Italy: Association for Computing Machinery, 2017, pp. 161–169, ISBN: 9781450346528. DOI: 10.1145/3109859.3109882. [Online]. Available: https://doi.org/10.1145/3109859.3109882.

[47] J. McAuley, *Personalized machine learning*. Cambridge University Press, 2022.

[48] P. Castells, N. Hurley, and S. Vargas, “Novelty and diversity in recommender systems,” in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. New York, NY: Springer US, 2022, pp. 603–646, ISBN: 978-1-0716-2197-4. DOI: 10.1007/978-1-0716-2197-4_16. [Online]. Available: https://doi.org/10.1007/978-1-0716-2197-4_16.

[49] T. Zhou, Z. Kuscsik, J.-G. Liu, M. Medo, J. R. Wakeling, and Y.-C. Zhang, “Solving the apparent diversity-accuracy dilemma of recommender systems,” *Proceedings of the National Academy of Sciences*, vol. 107, no. 10, pp. 4511–4515, 2010.

[50] M. Kaminskas and D. Bridge, “Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems,” *ACM Trans. Interact. Intell. Syst.*, vol. 7, no. 1, Dec. 2016, ISSN: 2160-6455. DOI: 10.1145/2926720. [Online]. Available: https://doi.org/10.1145/2926720.

[51] B. Smyth and P. McClave, “Similarity vs. diversity,” in *International Conference on Case-Based Reasoning*, 2001. [Online]. Available: https://api.semanticscholar.org/CorpusID:32908534.

6 ANEXOS

You can find the repository on GitHub: https://github.com/meaguirre3/Strava_GRS.git.