**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG CƠ SỞ THÀNH PHỐ HỒ CHÍ MINH**

**KHOA QUẢN TRỊ KINH DOANH II**

**--------------------------**



**BÁO CÁO CUỐI KỲ**

**KHO DỮ LIỆU VÀ KHAI PHÁ DỮ LIỆU**

**Đề tài: Xây dựng kho dữ liệu về nhân chủng học. Khai phá dữ liệu này cho mục đích nhận diện chủng tộc.**

**Giảng viên : Nguyễn Ngọc Duy**

**Sinh viên thực hiện:**

Nguyễn Trần Đức Thuận – N19DCCN1

Bùi Thanh Sơn – N19DCCN158

Vũ Cao Kỳ - N19DCCN1

**Lớp:** D19CQCNHT01-N

**TP. Hồ Chí Minh, năm 2023**

1. Giới thiệu đề tài.
2. Khái niệm chủng tộc.

\* Chủng tộc là gì?

Trước kia chủng tộc được coi là một tập hợp các cá thể có những đặc điểm hình thái giống nhau. Ngày nay với việc phát hiện ra “quần thể sinh học” và lí thuyết khu vực địa lý, thì đơn vị cơ bản để nghiên cứu nhân chủng là một tập hợp các quần thể chứ không phải tập hợp các cá thể. Quan niệm này đã mang lại cho định nghĩa về chủng tộc một nội dung mới chính xác và hoàn thiện hơn.

Định nghĩa: “Chủng tộc là một quần thể (hoặc tập hợp các quần thể) đặc trưng bởi những đặc điểm di truyền về hình thái, sinh lí mà nguồn gốc và quá trình hình thành của chúng liên quan đến một vùng địa vực nhất định”.

Như vậy, khái niệm “Chủng tộc” là một khái niệm khoa học phản ánh sự tồn tại khách quan của một thực tại sinh học hoàn toàn xác định. Nhận thức chủng tộc trên cơ sở quần thể (chứ không phải là cá thể) là một bước tiến quan trọng trong lí thuyết nhân chủng và sinh học. Các chủng tộc loài người rất phong phú, những dạng trung gian do hỗn chủng sinh ra ngày càng nhiều, tiến tới làm thay đổi và xoá nhòa ranh giới giữa các chủng tộc đã được hình thành.

1. Các đặc điểm chủng tộc.

Những đặc điểm hình thái riêng vốn có của mỗi chủng tộc tạo thành một tập hợp các đặc trưng cho phép chúng ta căn cứ vào đó để phân biệt chủng tộc này với chủng tộc khác thì gọi là các đặc điểm hay dấu hiệu phân biệt chủng tộc.

Các đặc điểm chủng tộc về mặt phân loại có thể chia ra làm ba loại:

- Loại các đặc điểm mô tả: màu da, màu mắt, màu và kiểu tóc, các hình dạng của mặt, mũi, môi, đầu, ...

- Loại đặc điểm đo đạc (mêtric) như kích thước của đầu, mặt, chiều cao và số đo các đoạn trong thân thể.

- Các loại đặc điểm hóa sinh: nhóm máu, nhóm huyết sắc tố.

Nhưng các đặc điểm chủng tộc lại có giá trị trong sự định chủng không hoàn toàn giống nhau. Một số đặc điểm, có tính chất phổ biến rộng rãi và khả năng bền vững ít thay đổi qua thời gian, có giá trị định chủng cao được gọi là các đặc điểm cơ bản (hay đặc điểm sơ cấp). Đó là màu da, màu tóc, máu mắt, mức độ lông trên thân người, và được dùng để phân loại các “Đại chủng”. Còn những đặc điểm thường bị biến dị khá nhiều trong khoảng mấy nghìn năm lại đây như kích thước mặt, chỉ số đầu,... gọi là những đặc điểm không cơ bản (hay đặc điểm thứ cấp), được dùng để phân loại các “tiểu chủng” hay “nhóm loại hình nhân chủng” là những cấp dưới của đại chủng.

1. Phân loại chủng tộc.

**Bảng phân loại chủng tộc của Cheboksarov (1985)**

|  |  |  |
| --- | --- | --- |
|  | Các đại chủng hoặc các nhóm chuyển tiếp hoặc lai | Các tiểu chủng và các nhóm chuyển tiếp hoặc hỗn chủng. |
| 1 | Đại chủng Xích đạo Phi (Negroid) | Negro, Negrin hay Picme, Boshiman và Hotentot. |
| 2 | Đại chủng Xích đạo Úc (Australoid) | Vedoit, Papua và Melanedien, Australian, Negrito, Ainu. |
| 3 | Hợp nhóm chuyển tiếp giữa các đại chủng Xích đạo và đại chủng Âu | Nhóm Nam Ấn, nhóm Đông Phi, nhóm chuyển tiếp Tây Xu-Đăng, nhóm lai Xu-Đăng, nhóm lai Mỹ và Phi da đen, nhóm lai Nam Phi. |
| 4 | Đại chủng Âu (Europid) | Chủng tộc phía Bắc (da trắng hồng), chủng tộc chuyển tiếp (da hạt dẻ), chủng tộc phía Nam (da nâu). |
| 5 | Hợp nhóm chuyển tiếp giữa Âu và Á | Các nhóm Uran, nhóm Siberia, nhóm lai Trung Á, nhóm lai Siberia, nhóm lai giữa Mỹ và bản địa da đỏ. |
| 6 | Đại chủng Á Mongoloid | Các chủng tộc Bắc Á, chủng tộc địa cực, chủng tộc Đông Á, chủng tộc da đỏ. |
| 7 | Hợp nhóm chuyển tiếp giữa đại chủng Á và Xích đạo | Các nhóm Nam Á, nhóm Nhật Bản, nhóm Micro-Polinedien, nhóm Đông Indonedien, nhóm Magat. |

Cheboksarov nhận định rằng, kết hợp nhiều loại đặc điểm chung trong nghiên cứu thì việc chia người hiện đại Homo sapien ra thành hai ngành Tây (Nêgrôit và Crôpôit) và Đông (Ôxtraloit với Môngôlôit) là hợp lí.

- **Đặc điểm nhân chủng của đại chủng Ôtxtralôit hay thổ dân da đen châu Úc:** Da rất sẫm màu (đen hoặc nâu đen), mắt đen; tóc đen uốn làn sóng; lông trên người rậm rạp, đặc biệt là râu phát triển mạnh; mặt ngắn và hẹp; mũi rộng, lỗ mũi to, sống mũi gầy; môi dày, môi trên vẩu, đầu thuộc loại đầu dài hay rất dài; chiều cao trung bình khoảng 150 cm.

- **Đặc điểm nhân chủng đại chủng Nêgrôit hay người đa đen châu Phi:** Gọi là Nêgrôit vì màu da của đại chủng này đen sẫm. Chữ Nêgrôit có nguồn gốc từ chữ La tinh “niger” nghĩa là đen. Đặc điểm hình thái điển hình của đại chủng Nêgroit là: da đen, tóc xoăn tít, lông trên thân rất ít, trán đứng, gờ trên ổ mắt ít phát triển, cánh mũi rất rộng làm cho mũi bè ngang, sống mũi không gẫy, môi rất dày nhưng hẹp. Ở một số loại hình mông rất phát triển. Về đặc điểm huyết học thì các nhóm máu A1, A2 và R có với tần số cao.

- **Đặc điểm nhân học của đại chủng Orôpôit hay người da trắng châu Âu:** Gọi là đại chủng Orôpôit vì đại chủng này sống tập trung ở châu Âu. Từ Crôpôit có nguồn gốc từ tiếng la tinh nghĩa là châu Âu. Đặc điểm hình thái chủ yếu là: da thay đổi từ màu sáng trắng tới nâu tối; lông trên thân rất phát triển đặc biệt là râu; tóc thường uốn sóng; mặt thường đô ra phía trước đặc biệt là phần giữa mặt; mặt hẹp và dài không vẩu; màu mắt thường xanh, xám hay nâu nhạt, không có nếp mi góc, mũi cao và hẹp; môi mỏng, cằm dài và vểnh, đầu thường là tròn; đặc biệt có núm Carabeli ở răng hàm trên. Nhóm máu giống người Phi, nghĩa là nhóm A1, A2 và R gặp với số cao.

- **Đặc điểm nhân học của đại chủng Mongoloid hay người da vàng châu Á:** Đại chủng Mongoloid về nhiều đặc điểm thường có vị trí trung gian giữa Orôpôit và Otstralôit. Đặc điểm chung là da sáng màu có ánh vàng hoặc ngăm đen; mắt và tóc đen, hình tóc thẳng và cứng; lông trên thân ít phát triển; mặt bệt vì hai xương gò má rất phát triển; mũi rộng trung bình, sống mũi không dô, gốc mũi thấp; môi dày trung bình, hàm trên hơi vẩu; nếp mi mông cổ tỷ lệ cao; đặc biệt có răng cửa hình xẻng là một đặc trưng của đại chủng này. Thường có nhóm máu Diêgô mà không có ở các đại chủng khác. Không có nhóm A2 và rất ít nhóm R. Địa bàn cư trú**:** Đông Á, Nam Á, trung tâm châu Á, Xibêri và châu Mĩ.

\* Các hợp nhóm chuyển tiếp có đặc điểm nhân chủng giữa các đại chủng với nhau nên khó định ra đầy đủ các đặc điểm của nó mà thường biểu hiện rõ trong các nhóm cụ thể.

1. Mục đích của đề tài.

- Dựa vào các đặc điểm dễ nhận biết nhất như màu mắt, màu tóc, màu da, sống mũi, tóc thì ta có thể dự đoán được chủng tộc của họ từ đó biết được tôn giáo, quốc gia, môi trường sống. Sau khi nhận diện chủng tộc, ta có thể tìm ra cách ứng xử, đối đãi, phục vụ phù hợp với nhu cầu, ý thích của từng cá nhân.

1. Công cụ sử dụng
2. Weka là gì ?

- Weka là gì? Weka còn có tên đầy đủ là Waikato Environment for Knowledge Analysis. Đây chính là bộ phần mềm mã nguồn mở được sử dụng miễn phí để khai thác dữ liệu thuộc các dự án nghiên cứu của đại học Waikato, New Zealand.

Weka đã được xây dựng bởi hệ thống ngôn ngữ lập trình Java. Weka thì không ᴄó mã hóa ᴠà ѕử dụng GUI đơn giản. Phần mềm này được tạo ra với mục tiêu là xây dựng nên một công cụ hiện đại với mục đích nhằm phát triển được các kỹ thuật máy học và áp dụng được chúng vào trong các bài toán khai thác dữ liệu ở điều kiện thực tế.

Khi sử dụng phần mềm Weka, bạn ᴄó thể gọi trựᴄ tiếp ᴄáᴄ thuật toán họᴄ máу hoặᴄ là nhập ᴄhúng bằng các mã Jaᴠa. Nó sẽ ᴄung ᴄấp cho bạn một loạt ᴄáᴄ ᴄông ᴄụ như là: trựᴄ quan hóa, tiền хử lý, phân loại, phân ᴄụm…

1. Các tính năng của phần mềm Weka.

Những tính năng siêu việt trong Weka phải kể đến là:

- Mã nguồn mở giúp người dùng dễ dàng tải về và sử dụng

- Hỗ trợ được nhiều các thuật toán máy học (machine learning) và khai phá các nguồn dữ liệu

- Trực quan hóa, dễ dàng xây dựng nên các ứng dụng thực nghiệm

- Do sử dụng JVM nên phần mềm Weka hoàn toàn độc lập với môi trường

Kiến trúc ở trong thư viện phần mềm Weka có hơn 600 class và được tổ chức nên 10 package. Chính vì thế mà người sử dụng có thể dùng được trực tiếp trên phần mềm hoặc là sử dụng những class này để làm bộ thư viện phát triển nên các ứng dụng của riêng mình.

1. Ưu điểm của phần mềm Weka.

Phần mềm mã nguồn mở Weka được bắt đầu phát triển mạnh mẽ vào những năm 1997 và ngày càng được sử dụng phổ biến ở trong nhiều lĩnh vực ứng dụng khác nhau, đặc biệt là trong mục đích giáo dục và nghiên cứu bởi các ưu điểm nổi trội sau đây:

* Tổng hợp được toàn diện các kỹ thuật tiền xử lý cũng như các mô hình hóa dữ liệu.
* Phần mềm này hoàn toàn được sử dụng 1 cách miễn phí theo Giấy phép Công cộng GNU (đây chính là giấy phép cung cấp bản quyền phần mềm miễn phí, đảm bảo giúp cho người dùng được tự do chạy, nghiên cứu và thay đổi hệ thống phần mềm).
* Tất cả các kỹ thuật của phần mềm khai phá dữ liệu Weka đều được dựa trên giả định rằng dữ liệu đã có sẵn ở dưới dạng một tệp phẳng hoặc là quan hệ, trong đó mỗi điểm của dữ liệu sẽ được mô tả bằng một số các thuộc tính cố định (thông thường sẽ là thuộc tính số hoặc thuộc tính danh nghĩa. Và một số các loại thuộc tính khác cũng đã được hỗ trợ).
* Do phần mềm này được viết bởi ngôn ngữ lập trình Java nên Weka đã có tính di động khi mà nó có thể chạy được trên đa số các nền tảng điện toán hiện đại nào đã được thử nghiệm như là: Linux, Windows hoặc là Macintosh.
* Giao diện với đồ họa được thiết kế thông minh, hiện đại giúp người dùng dễ dàng trong quá trình sử dụng.
* Weka còn hỗ trợ cung cấp các quyền truy cập vào hệ thống cơ sở dữ liệu SQL bằng cách sử dụng Java Database Connectivity và nó có thể xử lý và kết quả sẽ được trả về bởi lệnh truy vấn cơ sở dữ liệu.
* Weka còn giúp hỗ trợ được thêm một số nhiệm vụ khai thác dữ liệu tiêu chuẩn, cụ thể hơn đó là giúp xử lý trước dữ liệu, phân cụm/phân loại, hồi quy, trực quan hóa và cuối cùng là lựa chọn tính năng.