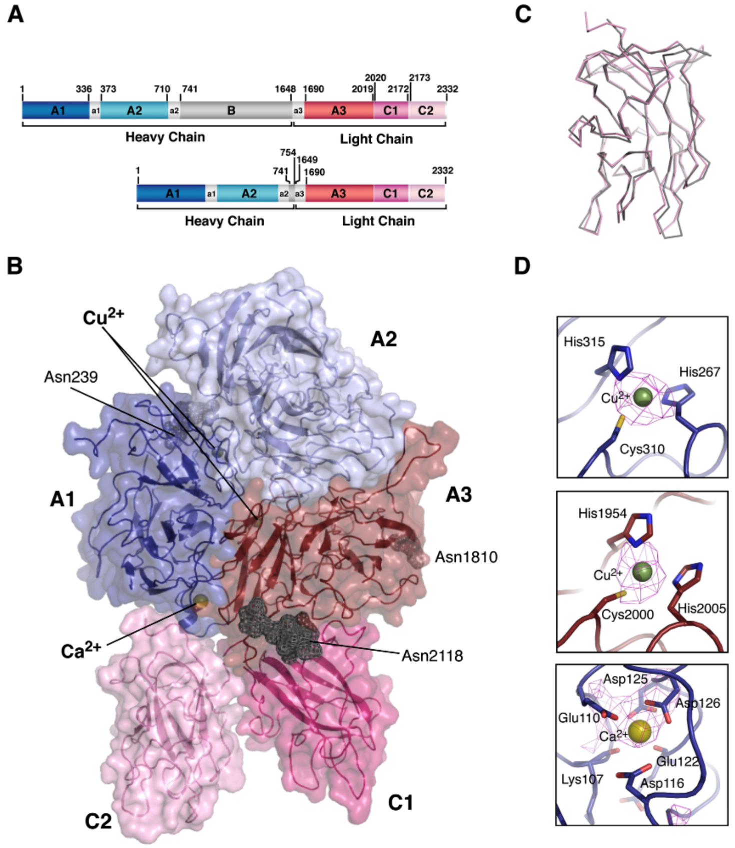

Figure 1: Figure from [17] depicts the structure of FVIII, which is composed of 2332 amino acids divided into domains: A1, A2, A3, B (not shown in the figure), C1, C2. The three primary domains, A, B and C, each play a significant role in the protein's structure and function. The A domain is crucial for FVIII activation. Although the B domain is not depicted in the figure and is less understood, is the largest of the three main domains. Lastly, the C domain is suggested to influence another factor, the von Willebrand factor, which also contributes to the blood coagulation cascade [16].

2. Description of Hemophilia Characteristics and Need for Diagnostic Improvement

According to data from the annual report of the World Federation of Hemophilia (WFH), approximately 250,000 people worldwide live with hemophilia, characterizing it as a rare genetic disease.

Hemophilia, a condition linked to the X chromosome, predominantly affects men due to genetic inheritance. Women, having two X chromosomes, rarely develop the disease, except in rare cases of mutations in both chromosomes. This condition compromises blood clotting and can be caused by inherited mutations (70% of cases) or spontaneous mutations (30% of cases), not necessarily related to family history [20].

There are two main types of hemophilia: A, characterized by a deficiency of clotting factor VIII (FVIII), and B, associated with a deficiency of clotting factor IX (FIX) [6]. Hemophilia A represents the majority of cases, with about 80 to 85% incidence, while Hemophilia B covers the remaining cases. These data are confirmed by the WFH report, which in 2020 reported about 80% of patients with Hemophilia A, approximately 15% with Hemophilia B, and the remaining 5% unidentified.

The classification of hemophilia severity is based on the level of clotting factor activity, dividing cases into mild, moderate, or severe. Severe cases presents significant risks of potentially fatal spontaneous bleeding, requiring constant care and specific treatment [14].

The conventional treatment of hemophilia is done through the replacement of deficient clotting factors, varying according to the severity of the cases. However, the high costs associated with treatment, especially for severe cases, highlight the importance of an accurate and effective diagnosis [9].

The diagnosis of hemophilia involves analysis of family history, symptoms, and laboratory tests to measure the levels of clotting factors. However, it is crucial to improve this process, seeking less invasive and safer methods, especially for severe cases. The implementation of alternative techniques, such as replacing invasive tests with less intrusive methods to diagnose the severity of hemophilia, can prevent complications in severe patients [1].

Considering that the majority of hemophilia cases correspond to type A, it is crucial to develop innovative and less invasive diagnostic approaches to identify and assess the severity of this form of the disease. To achieve this goal, it is essential to explore the FVIII protein (Figure 1) in detail, as it plays a fundamental role in the cause of Hemophilia A.

3. Dataset Description and Data Analysis

This section describes the dataset used in this research and also shows the data analysis of its features.

3.1. Dataset Description

The dataset contains 443 records, each representing a point mutation in the FVIII protein, which describes the position of the mutation and the amino acids before and after the mutation. Among the 21 features, there are information about genetic characteristics of the mutation, structural characteristics of the protein at the mutation position, and data about the interaction network of the residues of this protein. In addition, there is a variable that defines the severity of Hemophilia A.

The preparation of this database involved two steps: 1) verification and cleaning of null data, resulting in the removal of 28 records; 2) standardization of textual data, eliminating unnecessary white spaces and unifying the representation of the letters to allow correct identification and differentiation of categories. At the end of this process, the number of records in the base was reduced from 443 to 415.

It is essential to understand the meaning of each feature for the interpretation of the results. The features can be grouped into:

a) Genetic characteristics of the mutation in the FVIII protein

- `AAHGVS` (numeric): Position of the amino acid where the mutation occurred, considering the complete form of the protein.
- `AALegacy` (numeric): Position of the amino acid where the mutation occurred, considering the mature form of the protein.
- `aaOne` (categorical): Original amino acid before the mutation.
- `aaTwo` (categorical): New amino acid after the mutation.
- `AAdist` (numeric): Distance between the variables `aaOne` and `aaTwo`, defined by the distance matrix available in the Supplementary Data of the article by [15].

b) Structural characteristics of the FVIII protein at the mutation site

- `Domain` (categorical): Domain of the protein where the mutation occurred.
- `psi` (numeric): Torsion angle around the Cα-C.
- `phi` (numeric): Torsion angle around the N-Cα.
- `bfactor` (numeric): Debye-Waller factor.
- `areaSAS` (numeric): Solvent accessible surface area.
- `areaSES` (numeric): Solvent excluded surface area.
- `kdHydrophobicity` (numeric): Hydrophobicity of the amino acid.
- `ConsurfDB` (numeric): Degree of conservation of the amino acid.

c) Characteristics of the interaction network of residues of the FVIII protein at the mutation site

- `degree` (numeric): Number of connections of an amino acid in the network.
- `betweenness` (numeric): Number of shortest paths between all pairs of amino acids that pass through an amino acid.
- `closeness` (numeric): Average distance between an amino acid and all other amino acids in the network.
- `burts` (numeric): Measure that represents the dependence of an amino acid on others.
- `pr` (numeric): Measures the importance of an amino acid in the network.
- `auth` (numeric): Measures the relevance of an amino acid in the network.
- `kcore` (numeric): Subgraph where all amino acids are connected to at least k other amino acids in the subgraph.

d) Severity of Hemophilia A according to the level of FVIII

- `CalculatedSeverity` (categorical): Label of the severity of Hemophilia A defined by the levels of FVIII: severe (FVIII:C < 1), moderate (1 >= FVIII:C <= 5) and mild (FVIII:C > 5).

For a deeper understanding of the process of creating this dataset and the features contained in it, readers are encouraged to consult the article by [15] for more detailed explanations.

3.2. Data Analysis

Data analysis was performed in three stages: univariate, bivariate, and multivariate analysis.

3.2.1. Univariate Analysis

The univariate analysis aimed to understand the variables individually. The categorical variables were analyzed for the number of categories to understand the balance of the database. The numerical variables were analyzed in relation to their distributions, applying the Shapiro statistical test to verify the normality of the distributions and creating boxplots to identify outliers and understand their dispersions.

Labels

The analysis of the labels of the response variable (`CalculatedSeverity`) through a bar graph, found a slight imbalance in the database. The distribution of the labels is