Actividad Final: Análisis de los estilos de cerveza

1. Introducción

La elaboración de cerveza consiste básicamente en la cocción y fermentación del mosto, que es el líquido que se extrae del proceso de remojado de la malta de cebada u otros cereales. Los cambios en el proceso de elaboración y los ingredientes utilizados dan como resultado un gran espectro de cervezas resultantes y su clasificación tiene en cuenta varios factores como sabor, color, graduación alcohólica, e incluso la historia u origen de la receta.
Existen organizaciones de estandarización y métodos precisos para determinar el estilo de una cerveza como lo son los basados en quimiometría y requieren de instrumentos de laboratorio costosos. En la bibliografía existen algunos análisis de clasificación que en general aplican métodos como PCA, LDA, Random Forest, NNA, entre otros, pero las reglas de clasificación no son fáciles de interpretar.
El sitio web Brewer's Friend agrupa un conjunto de herramientas que, entre sus utilidades, se encuentra la de asistir en la elaboración de cerveza artesanal guiando a los usuarios durante todo el proceso. Los usuarios pueden conectar sus instrumentos de elaboración de cerveza a la API del sitio web para realizar el monitoreo en tiempo real del proceso, a la vez que el sitio devuelve información durante este proceso a modo de guía para la mejora del resultado. Además posee algunas componentes de red social como foros de discusión, artículos, productos comerciales, entre otros. Este sitio contiene una base de datos de alrededor de 200.000 recetas publicadas por sus usuarios y la misma será la fuente de información sobre la que se basa este trabajo.

2. Objetivos

Se analizará la clasificación de los estilos de cerveza basándose en tres atributos principales: color, sabor y graduación alcohólica, ya que se asume que son variables que pueden medirse sobre la cerveza preparada y no durante su elaboración. La solución apunta, por un lado, a facilitar la correcta elección del estilo a partir de la obtención del resultado de la elaboración y por el otro, a entender las características de estos atributos que definen cada estilo de cerveza.

3. Materiales y Métodos

Actualmente el sitio Brewer's Friend no permite la descarga masiva de los datos de las recetas subidas por los usuarios, pero en el sitio web Kaggle hay disponible un conjunto de datos de casi 75.000 recetas que fue publicado en 2017 y extraído de esta misma base de datos. Según el autor original de la contribución existen al menos cinco atributos o variables que pueden ser de utilidad: gravedad inicial (OG) y final (FG) del mosto, graduación alcohólica (ABV), índice IBU y color. Los valores OG y FG están directamente relacionados con el valor final de ABV por lo que serán descartados como atributos para la clasificación. Además son valores que no pueden determinarse una vez que el proceso ya finalizó. Las variables que serán tenidas en cuenta para el análisis son:

Color: existen distintas escalas para medir el color de la cerveza. Las unidades empleadas en la base de datos disponible emplea la escala Lovibond, que es un método colorimétrico desarrollado

por Joseph Williams Lovibond en 1883 y consiste en comparar el color de la cerveza con patrones de vidrio coloreado. El rango de esta variable es de 0 a 40 y la unidad es grados Lovibond (°L).
IBU: por sus siglas en inglés, *International Bitterness Units*, es un indicador de la cantidad de alfa-ácidos isomerizados durante la cocción del mosto. A mayor valor de IBU, más amarga es la cerveza. En general la escala abarca el rango de 0 a 100, pero existen cervezas que superan ampliamente este valor.
ABV: por sus siglas en inglés, *Alcohol By Volume*, indica la graduación alcohólica de la cerveza y se expresa en porcentaje de etanol por volumen total. En su mayoría este valor se encuentra alrededor del 5% pero en algunos casos puede superar el 40%.

3.1. Base de datos

El conjunto de datos posee 73.861 filas por 22 columnas, dentro de las cuales, 6 son atributos categóricos y 16 son atributos numéricos. La figura 1 muestra la completitud de las distintas columnas para una muestra del 10% de los datos totales.

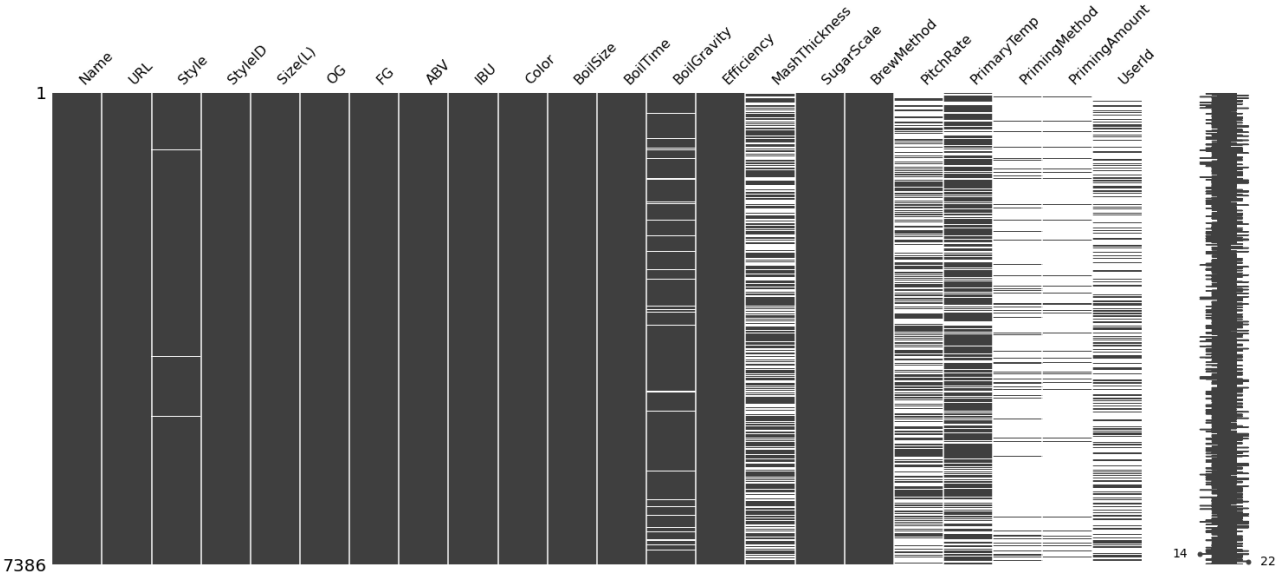

Figura 1. Gráfico de datos faltantes.

El número total de estilos distintos es de 175, de los cuales 100 poseen más de 100 instancias cada uno. La figura 2 muestra la proporción de los 10 estilos más comunes, que abarcan menos de la mitad del dataset. El estilo “American IPA” (Indian Pale Ale) es el más popular de todos y esto puede deberse a que el sitio agrupa mayormente usuarios de Estados Unidos.

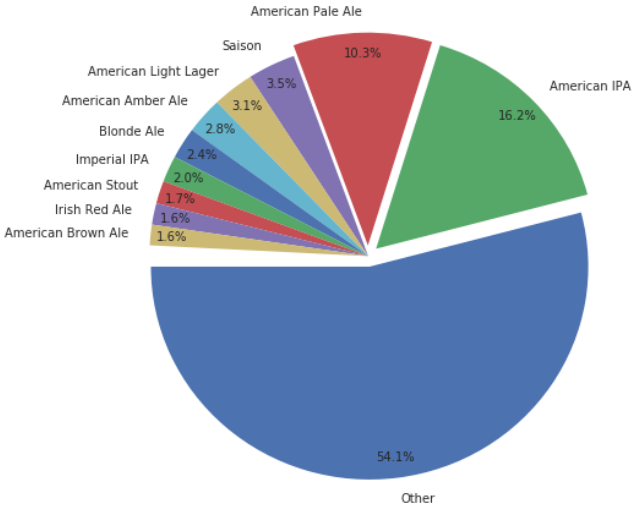

Figura 2. Proporción de estilos.

Con la intención de obtener un primer panorama de cómo están distribuidos los estilos en el espacio ABV, IBU y Color, se aisla un conjunto de 5.000 muestras descartando los valores atípicos (por encima de un desvío estándar σ) para cada estilo individualmente y submuestreando al 25% en forma aleatoria. En la figura 3 puede verse un gráfico de datos dispersos, a la izquierda, en un espacio 3D y a la derecha, la matriz de datos dispersos que se corresponde con las vistas de las tres caras del cubo que define el espacio de variables, donde se observa que algunas clases aparecen agrupadas en regiones separadas.

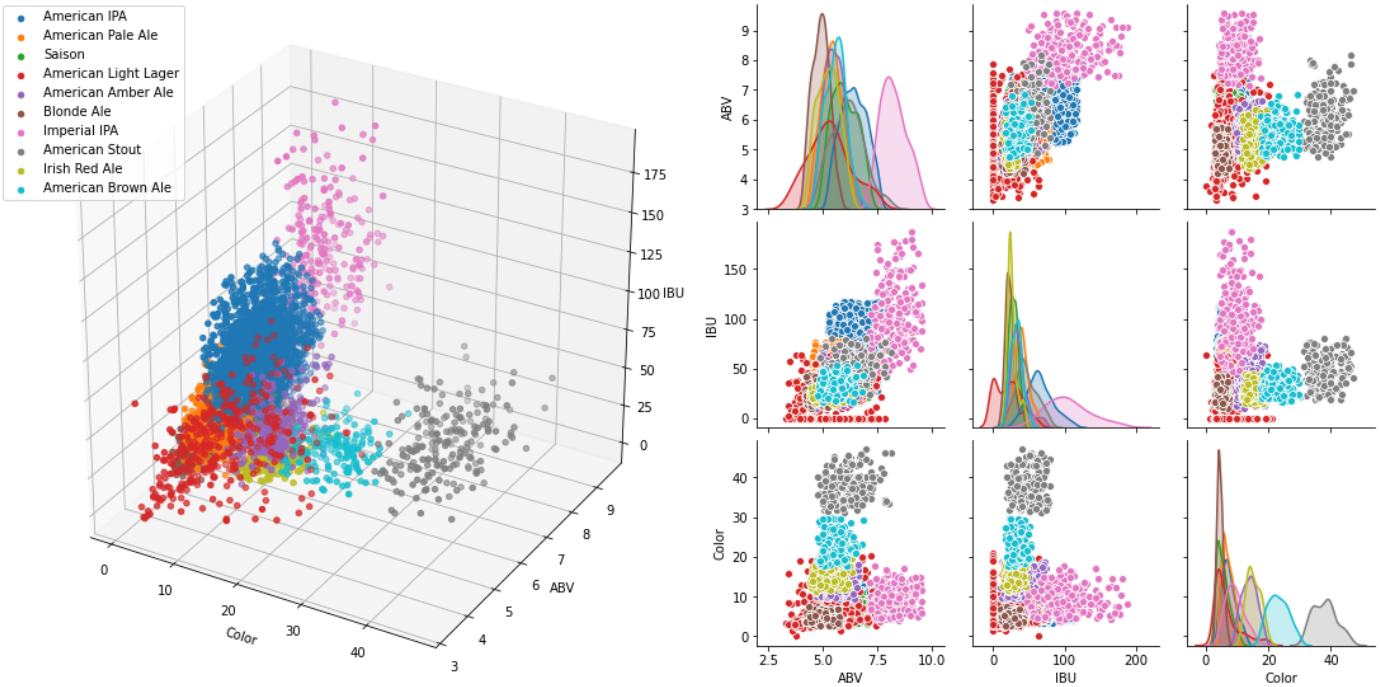

Figura 3. Gráficos de dispersión. Izquierda, tridimensional, derecha, matriz de datos dispersos.

3.2. Análisis de los datos

A continuación se intenta visualizar cómo es la distribución de cada una de estas variables por separado y para eso se emplea un gráfico de violín para el mismo conjunto filtrado que contiene los datos de cada una de las 10 clases principales. El resultado se muesra en la figura 4. En este caso se observa que para algunas clases, la distribución es prácticamente normal o gaussiana y en otros casos presenta dos picos de mayor densidad separados. Para el análisis que sigue a continuación, se asumirá que la distribución de los datos de cada clase es de tipo normal multivariada, es decir, que su función de densidad está dada por

$$f_X = \frac{\exp\left(-\frac{1}{2}(X-u)^T \sum^{-1}(X-u)\right)}{\sqrt{(2\pi)^k \left|\sum\right|}},$$

donde **μ** es la media de la distribución y **Σ** es la matriz de covarianza. Esto se realiza para simplificar el modelo que se describe a continuación.

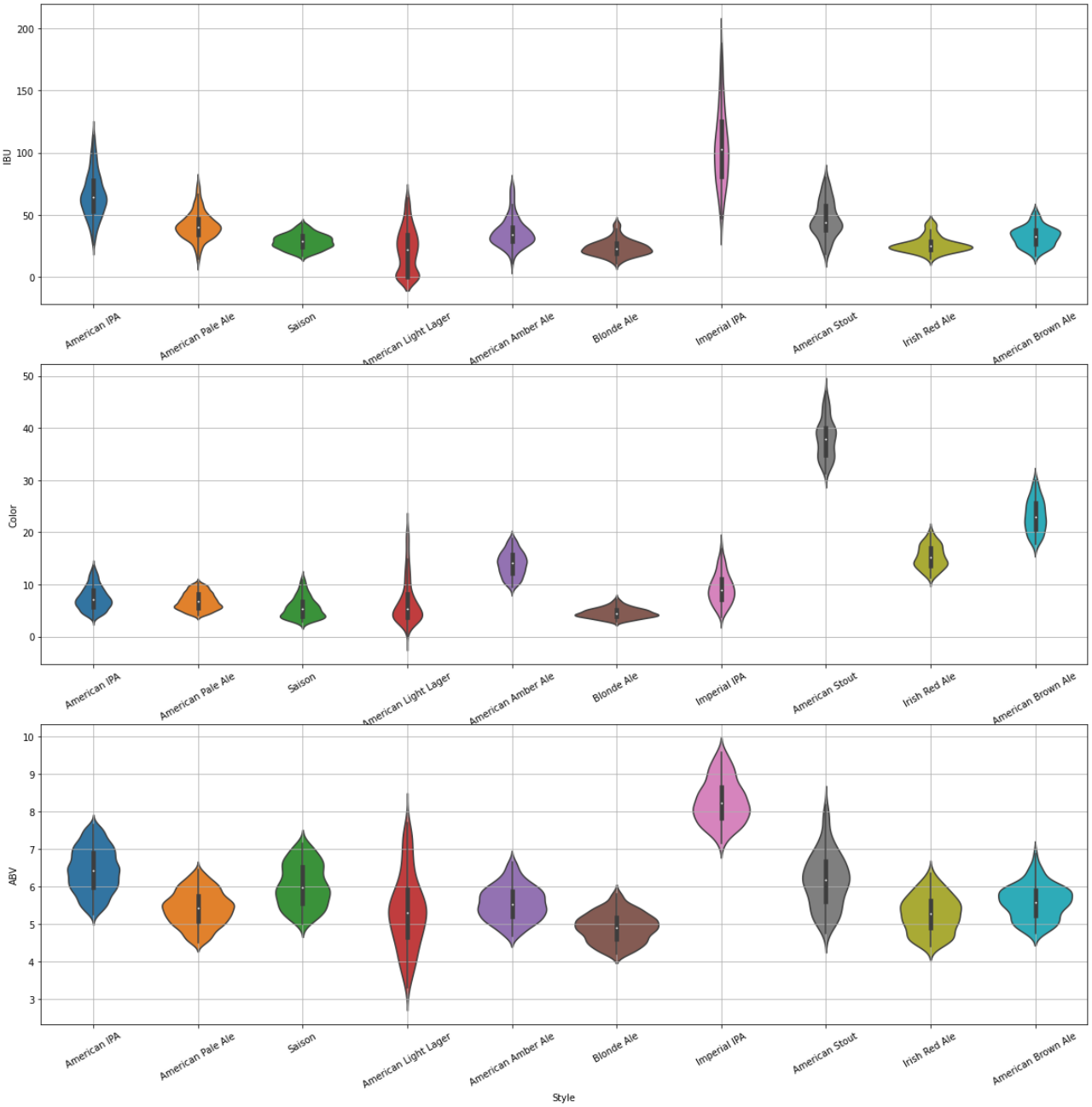

Figura 4. Gráfico de violín de las variables de cada clase.

Asumiendo que la distribución de los datos de cada clase es normal, es posible determinar qué tan cercano o alejado se encuentra un valor de una dada clase mediante la fórmula de Mahalanobis, ya que es más adecuada que emplear distancia euclídea. La distancia entre un punto **x** y la media **μ** de la clase, está dada por

$$d_m(\vec{x},\vec{u}) = \sqrt{(X-u)^T \sum^{-1} (X-u)}.$$

Por lo tanto, contando con la media **μ** y la matriz de covarianza **Σ** de cada clase, es posible estimar qué tan probable es que dicho punto pertenezca a cada clase. Este es el esquema que se utilizará para generar un listado de clases con sus respectivos puntajes de similitud a partir de una tupla de atributos ABV, IBU y Color correspondientes a una receta de cerveza. Esta combinación de valores será denominado “objetivo” a evaluar.

3.3. Diseño de la visualización

El objetivo es contar con un tablero interactivo que permita rápidamente evaluar uno o más combinaciones para los atributos cuantitativos de una cerveza y en forma simultánea visualizar el listado de estilos ordenados de mayor a menor similitud. El usuario debe poder “navegar” el espacio de variables en forma continua y al mismo tiempo ver los estilos de cerveza sugeridos dada la posición actual que se está evaluando. La figura 5 ilustra el diseño tentativo del tablero que permitiría esta funcionalidad.

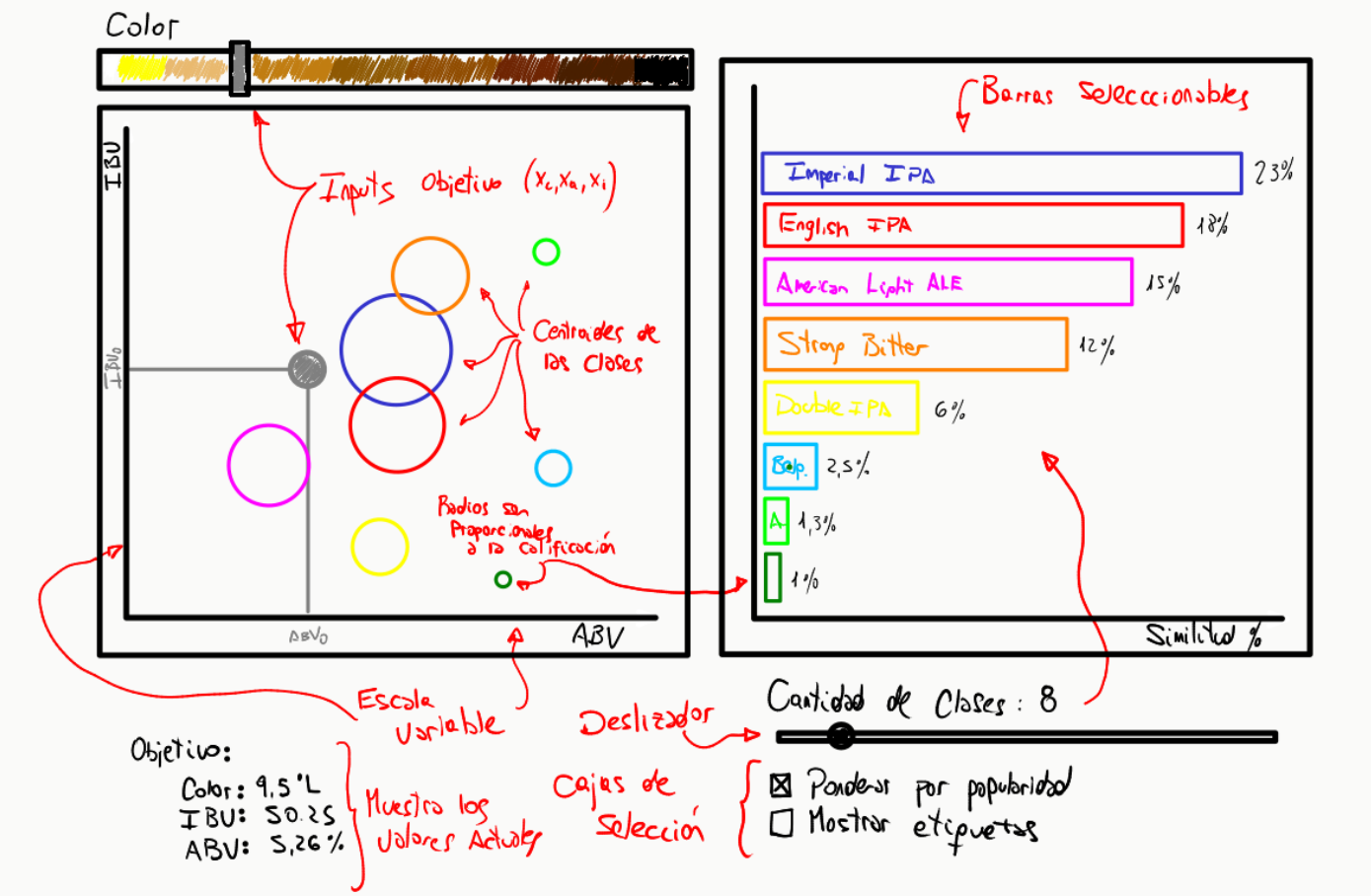

Figura 5. Bosquejo a mano alzada del tablero a implementar.

Resulta intuitivo que las clases más coincidentes con el punto evaluado tengan mayor puntaje, sin embargo, si se emplea como escala la función de distancia de Mahalanobis, la relación es inversamente proporcional. Para plantear un esquema de puntuación, se debe encontrar una función que permita obtener el puntaje de cada clase a partir del valor de la distancia al objetivo calculada. Con este propósito se enunciaron tres condiciones que son las siguientes:

$$1) P_i \propto \frac{1}{d_i},$$
$$2) P_i \in (0,1],$$
$$3) \sum_{i=1}^{N} P_i = 1,$$

donde d_i es la distancia de Mahalanobis entre el objetivo a evaluar y el centroide de la i-ésima clase y P_i es el puntaje correspondiente a dicha clase. La primer condición indica que es necesaria una relación inversamente proporcional entre los puntajes de cada clase y sus distancias al objetivo. La segunda condición permite normalizar el esquema de puntuaciones y la tercer condición cumple la función de asemejar los puntajes de cada clase al concepto de probabilidad estadística. Estos valores pueden expresarse en porcentajes para evitar el uso de muchos decimales.
Una posible expresión que cumple simultáneamente los tres requisitos es la siguiente

$$P_i = \frac{M - d_i}{M \cdot N - S},$$

donde N es la cantidad de clases, M es la distancia entre el objetivo evaluado y la clase más lejana, es decir, la máxima distancia computada y S es la suma de todos los valores de distancia d_i. Esta será la fórmula utilizada para calcular el puntaje de cada clase dado un objetivo a evaluar. El número de clases a considerar será un parámetro de configuración que el usuario pueda seleccionar.

3.4. Modelo unificado de visualización

La figura 6 muestra el modelo de visualización (MUV) para el diseño propuesto. A continuación se listan los estados de los datos que componen el sistema.

Datos Crudos (DC): es el conjunto de datos sin procesar obtenidos de la fuente original.
Datos Abstractos (DA): los DA se obtienen aplicando un procesamiento a los DC. En este caso consiste en agrupar los datos por estilo o clase y calculando para cada una, la media y matriz de covarianza.
Datos a Visualizar (DaV): es un subconjunto de los DA que se filtran a partir de la configuración establecida por el usuario. Las clases se ordenan según la distancia de Mahalanobis al valor objetivo y se filtra la primera parte del conjunto.
Datos Mapeados Visualmente (DMV): se cuenta con dos sustratos espaciales diferentes, uno para visualizar el espacio de variables y la posición del objetivo a evaluar (primario) y otro para ver el ordenamiento de clases o estilos sugeridos (secundario). De los DaV, se emplean los valores promedios (μ) de cada clase para ubicarlos en sus posiciones dentro del sustrato espacial primario y los valores de las puntuaciones de cada clase (d_m) se emplean en el sustrato secundario para determinar las longitudes de las barras horizonales.
Datos Visualizados (DV): consiste en la información que presentan los dos gráficos en simultáneo de la estructura visual del tablero interactivo.

En la figura 6 se puede ver el diagrama de bloques del MUV con los estados de los datos y las principales interacciones.

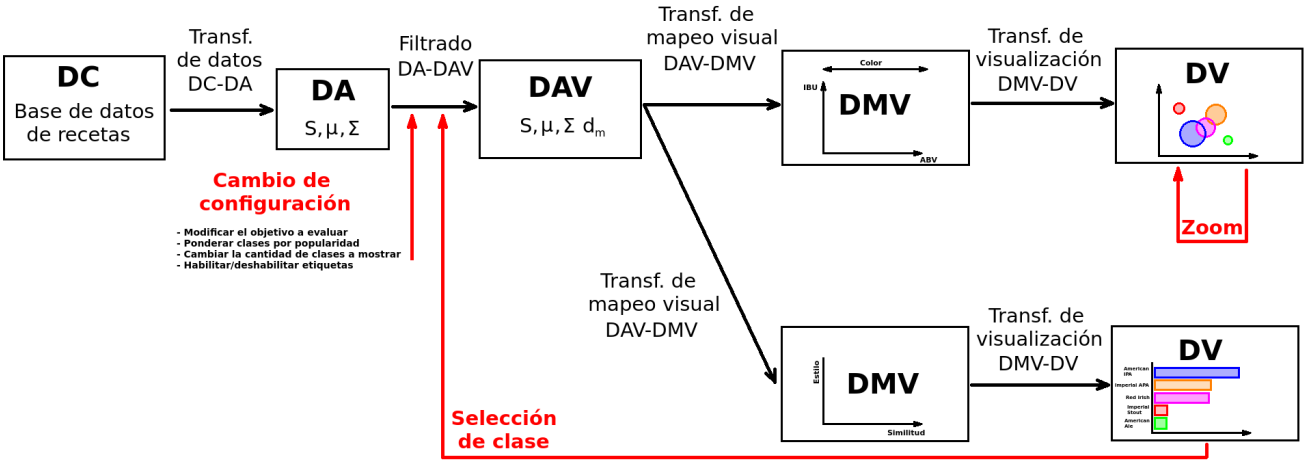

Figura 6. Modelo unificado de visualización.

3.5. Interacciones

Selección del objetivo: la estructura visual primaria contiene dos deslizadores, uno unidimensional para seleccionar la variable de Color y otro bidimensional para desplazarse por el espacio ABV, IBU. A medida que cambia la posición del objetivo, los datos a visualizar se van actualizando constantemente.

Selección de la cantidad de clases a mostrar: el rango de clases a mostrar simultáneamente puede variar de 10 a 60. Cuanto más clases se incluyen, menor es el puntaje que recibe cada una, ya que la suma total debe permanecer constante.

Habilitar/deshabilitar ponderación de clases según popularidad: dependiendo de si se desea que los puntajes de las clases tenga en cuenta el nivel de popularidad de cada una, se dispone de un switch tipo checkbox para controlar esta configuración.

Habilitar/deshabilitar visualización de los nombres de las clases: si dada la selección del objetivo actual existe un gran número de clases superpuestas o agrupadas en un mismo lugar, se dificulta la lectura e identificación de cada clase. En este caso puede ser conveniente deshabilitar las etiquetas de las clases.

Habilitar/deshabilitar asignación de colores según color promedio de la clase: como el espacio de variables tiene tres dimensiones y el sustrato espacial de la estructura primaria sólo dos, no se puede apreciar la profundidad de los centroides de cada clase, lo que correspondería a la componente de Color. Una opción para resolver este inconveniente es asignar los colores de cada clase según sus colores promedio. El mapa de color es limitado en el sentido que no permite diferenciar bien entre clases, sin embargo es posible visualizar cuáles centroides se encontrarían en mayor o menor profundidad dentro de la estructura visual primaria.

Alternar tamaño de centroides proporcional a su puntaje o a su popularidad: por defecto el tamaño de los centroides que se muestran en el sustrato espacial primario es proporcional al puntaje de su clase. Sin embargo es posible alternar esta dependencia para que el radio de cada círculo sea proporcional a la frecuencia de cada clase. De esta manera es posible también comparar la popularidad de cada estilo.

Zoom: para ampliar una dada región de la estructura visual primaria, posicionando el cursor sobre el centro de dicha región y accionando la rueda de scroll del mouse, los límites del espacio se ajustan de tal forma de lograr un efecto de zoom.

Selección de una clase: para ubicar una clase en la estructura visual primaria dada su posición dentro del gráfico de barras horizontal, es posible seleccionar pulsando sobre la barra de cada clase

y en forma simultánea, el círculo correspondiente a esa clase aparecerá resaltado en el espacio ABV, IBU.

3.6. Implementación

Los requerimientos funcionales y no funcionales se determinaron en base al tiempo estimado de desarrollo y las tecnologías disponibles.

- Se debe contar con un tablero interactivo para visualizar gráficamente el espacio de variables y poder desplazar unos cursores de prueba para evaluar distintas recetas de cerveza.
- Los resultados se deben actualizar en forma continua a medida que se modifican los valores del objetivo evaluado.
- Debe ser portable (multiplataforma).
- Debe tener libre acceso.

Para lograr una aplicación portable, accesible e interactiva se realizó la aplicación mediante tecnologías web. Para reducir el tiempo de desarrollo y facilitar la implementación de la interacción entre componentes de la aplicación se utilizó el framework ReactJS. Se agregó Bootstrap para dar estilo a los componentes de la GUI. Los gráficos de barras horizontales se realizaron mediante la librería Highcharts.js y la estructura visual del espacio de variables se realizó dibujando sobre un elemento canvas de HTML5. Con el objetivo de simplificar el despliegue y que la aplicación esté disponible online, se utilizó el servicio gratuito de Heroku.
En la figura 7 se muestra una captura de pantalla del tablero que contiene la aplicación implementada.

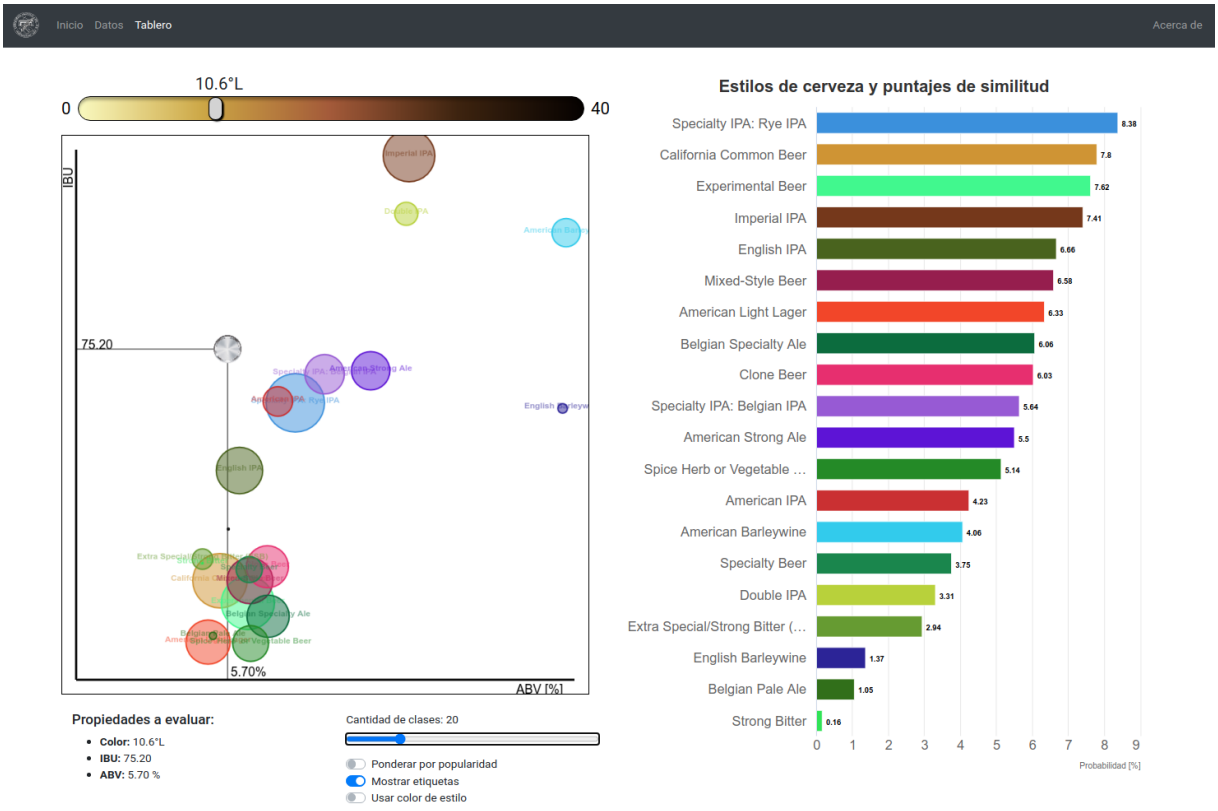

Figura 7. Captura de pantalla de la aplicación.

Resultados y discusión

El tablero desarrollado es relativamente intuitivo si de antemano se conoce el propósito del mismo, con lo cual, para ser liberado como herramienta pública sería necesario acompañarlo con una introducción a los conceptos básicos sobre las propiedades cuantitativas de una cerveza, un tutorial de uso y/o un instructivo.
Se realizó una validación utilizando cervezas comerciales de distintas marcas ingresando los valores que aparecen en los envases y aunque el estilo correcto aparece entre los primeros 10, pocas veces se da un acierto en cuanto al primer estilo sugerido. Esto puede deberse a diferencias entre las propiedades de las recetas artesanales y comerciales o puede haber otras razones que tengan que ver con la distribución estadística de los datos.
Para incrementar la eficacia de la aplicación como herramienta de clasificación, probablemente sea necesario incluir más datos acerca de la relación entre los estilos, por ejemplo incorporando un esquema de organización gerárquico entre los distintos estilos. De esta manera se podría separar los valores de confiabilidad del estilo sugerido en distintos niveles de gerarquía.