UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

GESTIÓN Y ANÁLISIS AUTOMÁTICO DE SIMULACIONES CLÍNICAS

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN

ÁLVARO IGNACIO LEÓN SÁNCHEZ

PROFESOR GUÍA:
MAURICIO CERDA VILLABLANCA

MIEMBROS DE LA COMISIÓN:
JÉRÉMY BARBAY LEFEVRE
JOSÉ SAAVEDRA RONDO

SANTIAGO DE CHILE
2020

RESUMEN DE LA MEMORIA PARA OPTAR
AL TÍTULO DE INGENIERO CIVIL EN COMPUTACIÓN
POR: **ÁLVARO IGNACIO LEÓN SÁNCHEZ**
FECHA: 2020
PROF. GUÍA: Mauricio Cerda Villablanca

GESTIÓN Y ANÁLISIS AUTOMÁTICO DE SIMULACIONES CLÍNICAS

La simulación clínica es una estrategia didáctica y evaluativa considerada fundamental dentro de los currículum en carreras de salud debido a su gran utilidad en el desarrollo de las competencias profesionales del alumno. Esta estrategia impulsa el cambio de paradigma en el cual se deja de centrar la educación en la enseñanza y pasa a centrar su atención en el aprendizaje.

El Centro de Habilidades Clínicas es un espacio de la Facultad de Medicina de la Universidad de Chile dirigido a experiencias de aprendizaje, enseñanza y evaluación en el área de salud, donde se enfatiza la simulación clínica en beneficio del proceso de enseñanza y aprendizaje de habilidades clínicas y quirúrgicas, con el fin de contribuir a la formación del equipo de salud, acorde al propósito de la Facultad de Medicina.

El centro cuenta con la plataforma llamada **Kamaleon**, creada por alumnos del DCC en el curso Proyecto de Software. Esta plataforma permitiría coordinar y registrar las distintas actividades de simulación clínica realizadas en el centro, pero no se utilizo debido a errores y dificultades técnicas para su operación.

Este documento describe las mejoras realizadas sobre esta plataforma, con el objetivo de incrementar la retroalimentación de los estudiantes en carreras de la salud y disminuir las tareas de gestión y evaluaciones repetitivas asociadas.

En particular, se adaptó la plataforma para permitir la gestión de actividades en tiempo real y por videoconferencia, como también se hizo la integración del registro clínico electrónico (utilizado para evaluar a los estudiantes), permitiendo así tener un ingreso de los datos centralizado y estandarizado para todas las actividades realizadas por el CHC. Estas mejoras fueron subidas a la plataforma en producción y se validaron con usuarios reales al final de la etapa de desarrollo, obteniendo una puntuación promedio de `88.9 %` en la encuesta de usabilidad *SUS*.

Este informe también describe el desarrollo de un clasificador del rendimiento de los estudiantes en el registro clínico electrónico y la secciones temporales relevantes de los videos, mejoras que no fueron incluidas en la plataforma en producción por inconsistencia en los datos y falta de datos respectivamente.

Utilizando un regresor para clasificar el desempeño en los diagnósticos clínicos, se alcanzó un `MAE: 0.9` y `MSE: 0.027`, y clasificando cada fotograma de los videos para generar las secciones temporales, se alcanzó un `Accuracy: 0.98` y `F1-score: 0.98`.

Tabla de Contenido

Índice de Tablas

Índice de Ilustraciones

Capítulo 1

Introducción

1.1. Simulación clínica

La simulación clínica es una estrategia didáctica y evaluativa considerada con mucha importancia dentro de los currículum en carreras de salud tales como medicina, enfermería, obstetricia, kinesiología, terapia ocupacional, fonoaudiología y nutrición, debido a su utilidad en el desarrollo de las competencias profesionales del alumno. Esta estrategia impulsa el cambio de paradigma en el cual se deja de centrar la educación en la enseñanza y pasa a centrar su atención en el aprendizaje.

Las experiencias prácticas son fundamentales en la formación médica y la simulación se ha convertido en una excelente manera de realizar dichas actividades prácticas debido a sus múltiples ventajas. El paciente simulado/estandarizado (PS) actualmente ha evolucionado como una útil alternativa para proveer experiencias clínicas, tanto formativas como sumativas. En este contexto, el Examen Clínico Objetivo Estructurado (ECOE), Encuentros Clínicos con Pacientes Simulados/Estandarizados (ECliPSE)[Anexo A] y otras actividades simuladas permiten desarrollar fielmente los roles del paciente simulado en la representación de caso, evaluación y retroalimentación.

Las experiencias de aprendizaje basadas en la simulación deben incluir una sesión de *debriefing* (Retroalimentación grupal posterior a una actividad) dirigida a promover el pensamiento reflexivo, ya que el aprendizaje del estudiante depende de la integración de la experiencia y la reflexión.

La reflexión es la consideración consciente del significado y la implicación de una acción, que incluye la asimilación de conocimientos, habilidades y actitudes con conocimientos preexistentes. Esta reflexión puede conducir a nuevas interpretaciones por parte del alumno. El pensamiento reflexivo no ocurre automáticamente, pero se puede enseñar; requiere tiempo, participación activa en una experiencia realista y orientación por parte de un facilitador efectivo [6].

Las habilidades del *debriefer* son importantes para garantizar el mejor aprendizaje posible, ya que aprender sin orientación podría llevar al alumno a transferir negativamente un error a su práctica sin darse cuenta de que había sido una mala práctica, repetir errores, centrarse solo en lo negativo o desarrollar fijaciones. La investigación proporciona evidencia de que el proceso de información es el componente más importante de una experiencia de aprendizaje basada en simulación [6, 4].

1.2. Centro de Habilidades Clínicas

El Centro de Habilidades Clínicas[1] (CHC) es un espacio de la Facultad de Medicina de la Universidad de Chile dirigido a experiencias de aprendizaje, enseñanza y evaluación en el área de salud. En el CHC se enfatiza la simulación clínica, en beneficio del proceso de enseñanza y aprendizaje de habilidades clínicas y quirúrgicas, con el fin de contribuir a la formación del equipo de salud, acorde al propósito de la Facultad de Medicina.

Ubicado en el Campus Occidente de la Universidad de Chile, la unidad cuenta con la infraestructura necesaria para aplicar técnicas de simulación clínica. En los escenarios clínicos se dispone de un espacio semejante a un centro asistencial con 6 boxes y 7 salas de atención clínica, dotadas de vidrios-espejos y sistemas de audio y video, para supervisión directa y remota.

Figura 1.1: Escenarios clínicos del CHC durante el desarrollo de un ECliPSE

De esta manera, los estudiantes de pregrado de las ocho escuelas de la Facultad de Medicina pueden acceder a situaciones clínicas intencionadas, seguras y estandarizadas simulando a las que enfrentarán en su vida profesional.

Estas actividades implican un enorme esfuerzo de coordinación y movilización de recursos, dado el gran número de estudiantes que atenderán pacientes, recibirán retroalimentación y participaran de sesiones de *debriefing* en forma estandarizada.

[1] http://chc.med.uchile.cl

El CHC cuenta con la plataforma llamada **Kamaleon**, la cual fue creada por alumnos del DCC en el curso Proyecto de Software el segundo semestre del año 2018.

Esta se desarrolló con el objetivo de permitir coordinar y registrar de forma centralizada las actividades de simulación clínica realizadas en el centro, permitiendo que usuarios puedan acceder a los videos correspondientes a la interacciones y permitiendo al paciente simulado de evaluar el encuentro con los alumnos.

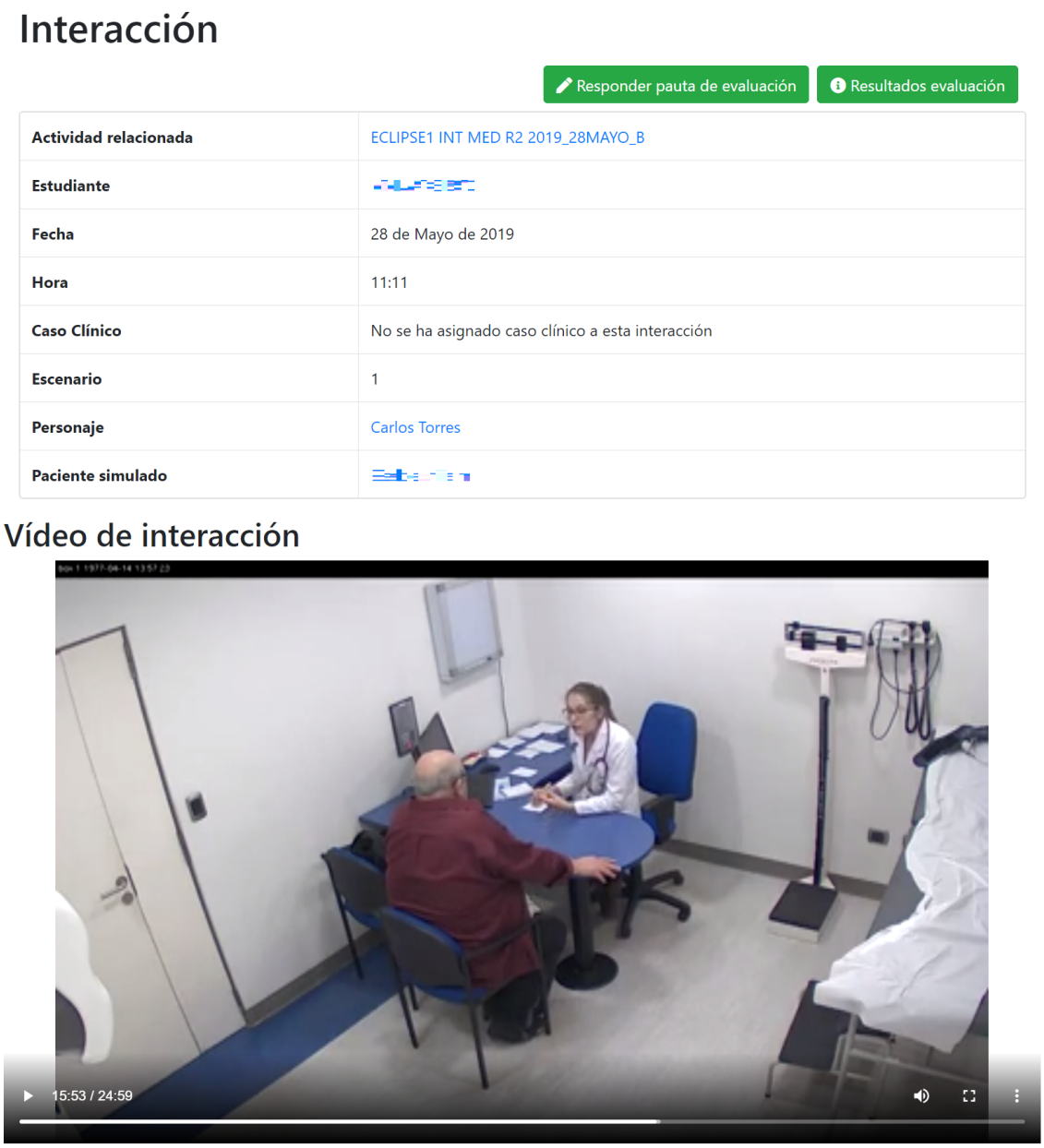


Figura 1.2: Vista original de una interacción en la plataforma

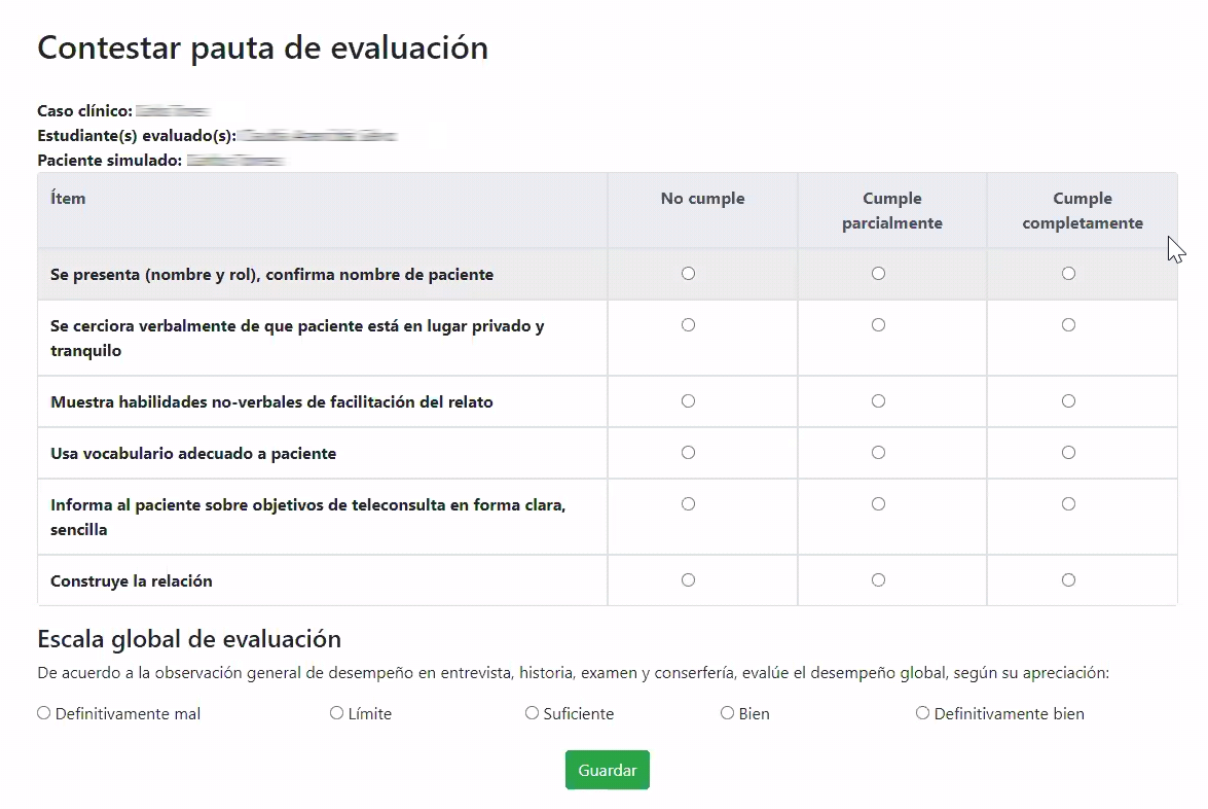


Figura 1.3: Vista de la evaluación del paciente simulado del encuentro con los alumnos

Esta plataforma permitiría gestionar las distintas actividades de simulación clínica realizadas en el centro, pero no se utilizó debido a errores y dificultades técnicas para su operación. Pese a esto, el trabajo realizado representó una base sólida para el cumplimiento de su objetivo.

1.3. Objetivos

El trabajo realizado se centró en el desarrollo de mejoras en la plataforma ***Kamaleon***, con el objetivo de incrementar la retroalimentación de los estudiantes en carreras de la salud de la Universidad de Chile y también disminuir las tareas de gestión y evaluaciones repetitivas asociadas.

Mientras se exploraron métodos para automatizar las tareas de evaluación, el objetivo del trabajo se extendió a la exploración y mejor entendimiento de los datos existentes de años anteriores, y dio paso a mejorar los la organización y composición de los elementos fundamentales de la evaluación de los estudiantes, en particular, una mejor definición de los casos clínicos y sus pautas de evaluación asociadas.

1.3.1. Objetivo principal

Mejorar la retroalimentación de los participantes de las simulaciones clínicas desarrolladas en el CHC, extendiendo las funcionalidades de la plataforma Kamaleon.

1.3.2. Objetivos específicos

Para lograr el éxito del proyecto, se plantearon 5 objetivos específicos a desarrollar:

1. **Detección y corrección de errores:** Corrección de errores en la plataforma para asegurar su correcto funcionamiento.

2. **Mejora del flujo de información:** Unificar el ingreso de datos de las actividades realizadas en el CHC, con principal énfasis en actividades de tipo *ECliPSE*.

3. **Mejora de las interfaces de usuario:** Creación de nuevas vistas y mejorarás de vistas existentes, en función de mejorar la retroalimentación de los usuarios.

4. **Clasificación automática del rendimiento de los estudiantes:** Clasificar el rendimiento de los estudiantes de forma automática en las respuestas del registro clínico electrónico y así facilitar la evaluación.

5. **Análisis automático de los videos:** Extracción de información relevante a partir de los videos para dar acceso rápido a esta información y facilitar la evaluación.

Capítulo 2

Marco Teórico

2.1. Soluciones existentes

Existen múltiples herramientas en el mercado dedicadas a la grabación de video para debriefing, permitiendo anotaciones en el video y una posterior evaluación. Entre las herramientas destacamos:

- **B-Line Medical - SimCapture**[1]: "Solución en la nube eficaz y rentable para registrar, informar y realizar un seguimiento del aprendizaje de simulación".
- **CAE Healthcare - LearningSpace**[2]: "Solución de gestión de centros de simulación para escuelas de medicina, programas de educación de enfermería, programas de simulación de salud aliados y hospitalarios".
- **IVS - Simulation**[3]: "Solución de grabación para simulaciones".

Estas están diseñadas específicamente para las simulaciones clínicas, aunque existen muchas mas (por ejemplo Codimg [4]) las cuales son aplicadas a variadas áreas de aprendizaje que se vean beneficiadas por el debriefing, como lo son simulaciones de aviación y de entrevistas.

Estas herramientas no se ajustan a las necesidades del CHC debido a que, entre otras cosas, estas utilizan:

- Licencias propietarias, haciendo que el software no sea adaptable a las necesidades del centro
- Hardware propietario, que restringe el uso de cámaras, micrófonos y/o servidores a utilizar y no son compatibles con el hardware que se dispone actualmente
- Licencias de alto costo y/o de suscripción. El alto costo se debe a que las herramientas proporcionan muchas características tales como almacenamiento en la nube o captura de múltiples fuentes de información, muchas de las cuales no son necesarias para el CHC. Se hizo una cotización del servicio *SimCapture* por ser uno de los más populares, quienes nos indicaron que el costo del servicio para un centro como el CHC es de aproximadamente 25 Millones de pesos por sala, el cual es demasiado alto para el centro considerando que este cuenta con 6 boxes y 7 salas de atención clínica

[1] https://www.blinemedical.com/medicalschools.html
[2] https://caehealthcare.com/learningspace
[3] https://ipivs.com/solutions/simulation
[4] https://codimg.com/en/

2.2. Plataforma actual y metodologías de desarrollo

2.2.1. Arquitectura de la plataforma

Kamaleon tiene un diseño Modelo–Vista–Controlador con componentes modulares con documentación adecuada, por lo que se decidió modificar y extender las funcionalidades de esta en vez de crear una nueva plataforma.

La plataforma tiene 3 componentes principales:

- **Web App**: Plataforma web hecha en Django. Su principal objetivo es tener la lógica de negocio para satisfacer las necesidades tecnológicas del CHC.
- **NVR API**: API REST hecha en Django. Su principal objetivo es facilitar la comunicación entre *Web App* y el *NVR* en el tema de los videos del CHC.
- **NVR**: Grabador de vídeo en red. Su principal objetivo es grabar los videos de las actividades del CHC.

Las mejoras desarrolladas en la plataforma se hicieron sobre el componente *Web App* en el framework Django en el lenguaje *Python3*. El servidor en producción despliega los servicios en `Docker` en 2 contenedores, los cuales se componen por:

- **Web App**:
  - `Django 2.1`: Framework
  - `MySQL 5.5`: Base de datos (Modelo ER inicial en la figura 2.2)
  - `Nginx`: Servidor web/proxy
  - `Gunicorn`: Servidor HTTP para Python WSGI
- **NVR API**:
  - `Django 2.1`: API REST para gestionar llamadas y ejecutar comandos compilados de *Hikvision SDK* (código de fuente en *C++*), el cual permite obtener los videos del NVR de las distintas cámaras en los box del CHC

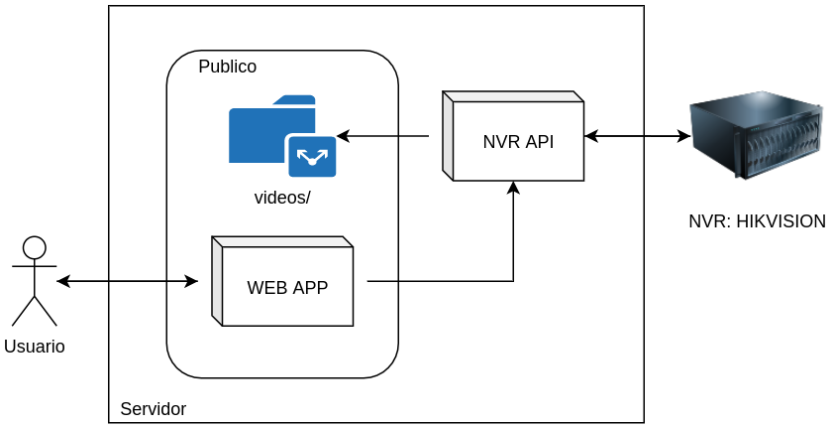


Figura 2.1: Comunicación entre los servicios

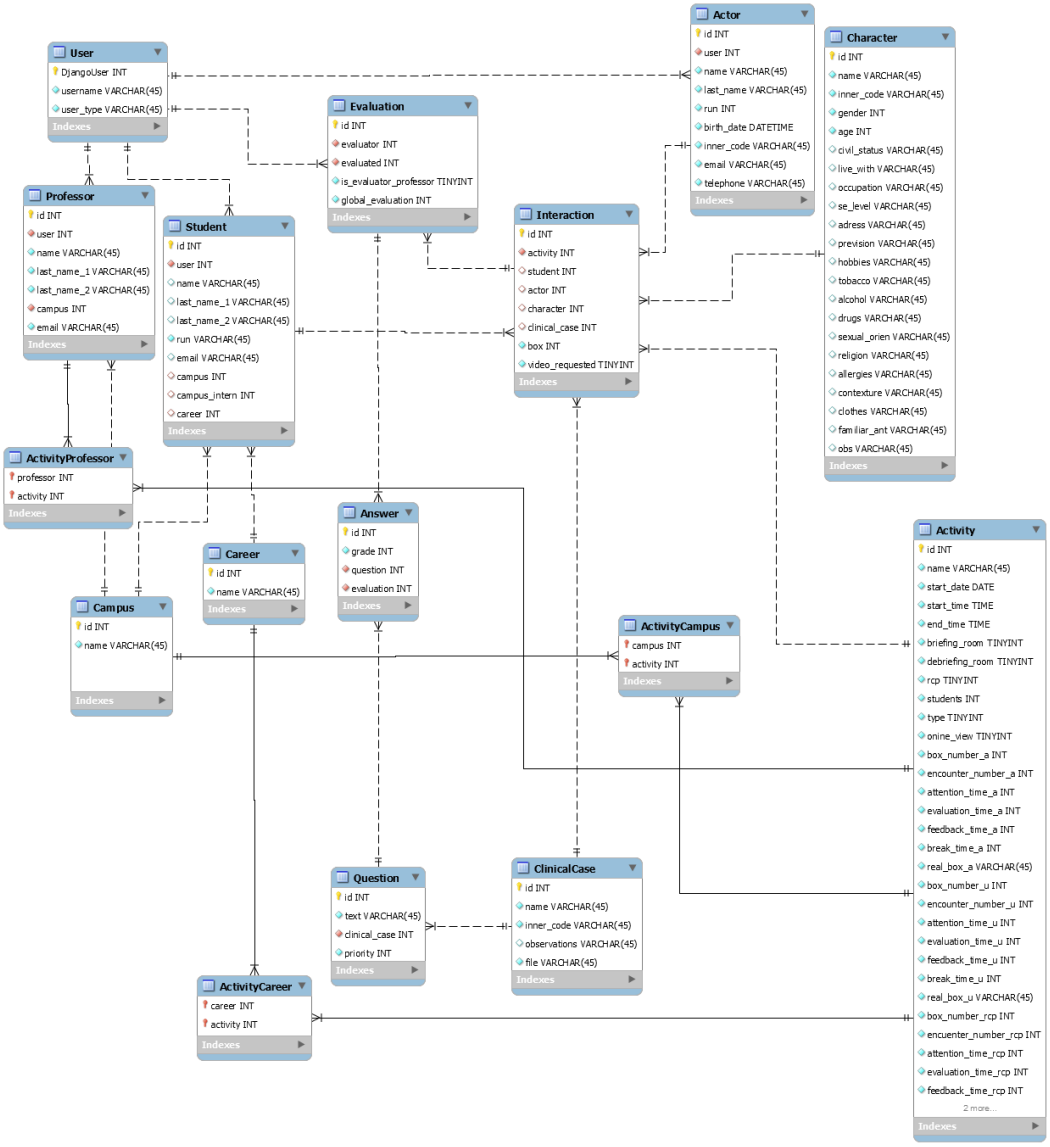


Figura 2.2: Modelo Entidad-Relación de la base de datos inicial

El diseño de la base de datos no tuvo modificaciones mayores con las mejoras realizadas en la plataforma. De la figura 2.2, las entidades principales que se tratan en este trabajo son:

- **Interaction**: Encuentro clínico entre un paciente simulado y un alumno
- **Activity**: Corresponde a una actividad de simulación clínica en el CHC, permite organizar los distintos encuentros clínicos (*Interaction*) que se realizarán
- **User**: Tabla de los usuarios, tiene una relación 1-1 con las tablas de cada tipo de usuario: *Professor*, *Student* y *Actor*
- **Character**: Personajes interpretados por los pacientes simulados (*Actor*)
- **Evaluation**: Tabla de la evaluación del encuentro del paciente simulado, junto a la tabla *Question* permite crear formularios personalizables para cada caso clínico (*ClinicalCase*)

2.2.2. Metodologías de desarrollo

Se utilizaron metodologías ágiles para el desarrollo de las tareas, con reuniones periódicas para revisar el avance, realizar análisis de riesgo y determinar las próximas tareas a realizar [1].

Siguiendo esta metodología, se divide el desarrollo en pequeños incrementos que minimizan la cantidad de planificación y de diseño iniciales. El trabajo se realiza en iteraciones, o *sprints*, de periodos de tiempo que duran alrededor de cuatro semanas. Cada iteración implica trabajar en todas las funciones: planificación, análisis, diseño, codificación, pruebas unitarias y pruebas de aceptación. Al final de la iteración, se muestra un producto funcional a las partes interesadas y se planifica las próximas tareas a desarrollar en el siguiente sprint. Esto minimiza el riesgo general y permite que el producto se adapte rápidamente a los cambios.

Es posible que una iteración no agregue suficiente funcionalidad para garantizar el lanzamiento de la plataforma, pero el objetivo es tener una versión disponible (con errores mínimos) al final de cada iteración. Es posible que se requieran varias iteraciones para nuevas funciones. El software que funciona es la principal medida del progreso.

2.2.3. Tecnologías consideradas

Mientras que para el desarrollo del componente web se decidió continuar las tecnologías ya utilizadas, existen múltiples alternativas para:

- Extracción y preprocesado de los datos (video y texto)
- Entrenamiento, validación y despliegue en producción de modelos predictivos

Entre los lenguajes más populares que cumplen los requisitos tenemos a *R* y *Python*, ambos complementados con poderosas librerías. Se eligió usar *Python* para tener una integración más simple con el componente web de la plataforma en producción y por estar más familiarizado con el lenguaje y sus librerías en general.

Para facilitar la creación, pruebas y despliegue de los modelos predictivos se usó un framework de aprendizaje automático, existen múltiples alternativas que cumplen los requisitos planteados, con licencias open-source o libres y que permiten usarse con el lenguaje *Python*, entre las más populares encontramos:

- TensorFlow
- Keras
- PyTorch
- Caffe
- Scikit-learn

Se eligió *TensorFlow 2.1* [5] por ser una de las más robustas, usadas tanto para investigación como en producción y por poseer actualizaciones constantes junto una comunidad y soporte activo.

Se utilizaron múltiples librerías de Python para la extracción y preprocesado de datos, estas permitieron probar distintas maneras de abordar el desarrollo sin que ninguna impida el uso de otras. Entre estas vale la pena mencionar:

- Pandas: Extracción y preprocesado de texto [6]
- NumPy: Operaciones entre matrices [7]
- Scikit-learn: Cálculo de vectores (tf-idf [8]) y reducción de dimensionalidad (LSA [9])
- OpenCV: Extracción de imágenes de videos [10]

[5] https://tensorflow.org/
[6] https://pandas.pydata.org/
[7] https://numpy.org/
[8] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html
[9] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html
[10] https://opencv.org/

2.3. Metodologías de análisis automático

Para la clasificación automática del rendimiento de los estudiantes en respuestas de texto libre y para la extracción de secciones temporales a partir de los videos, se utilizaron distintas metodologías de análisis automático y evaluación correspondiente, descritas a continuación.

2.3.1. Análisis de lenguaje natural

El registro clínico electrónico se evalúa con una pauta diferente para cada caso clínico y en particular se evalúa al alumno en función de los diagnósticos y las indicaciones, correspondientes a respuestas de texto libre.
Las pautas de evaluación de un caso constan de 2 secciones para el diagnóstico y para las indicaciones, que podemos simplificar como: frases o palabras que debe contener la respuesta para recibir puntaje parcial o total, y frases o palabras que descuentan puntaje o lo anulan por completo (no siempre se especifica). Para los diagnósticos estas palabras hacen referencia directamente al caso clínico, usualmente el puntaje perfecto corresponde al nombre del caso clínico más otras afecciones relacionadas a este.

A modo de ejemplo, se listan respuestas de los estudiantes y la pauta de evaluación respectiva para el caso clínico "*EPOC*":

- **Pauta:**
  - Diagnósticos (Puntuación Máxima: 3):
    - *TABAQUISMO/FUMADOR* (+1 punto)
    - *EPOC Gold 2 VEF 1 moderado* (+2 puntos)
    - *EPOC (solo)* (+1 punto como puntuación parcial)
    - *Gold 2 /VEF 1 Moderado* (+1 punto como puntuación parcial)
  - Indicaciones (Puntuación Máxima: 8):
    - *NO FUMAR* (+2 puntos)
    - *VACUNA INFLUENZA* (+2 puntos)
    - *BD PERMANENTE ACCION LARGA (SALMETE/IPRAT)* (+2 puntos)
    - *BD SINTOMATICO ACCION CORTA (SALB/IPRAT)* (+1 punto)
    - *CON AEROCAMARA* (+1 punto)
- **Respuestas de estudiantes:**
  - Diagnósticos:
    - *EPOC* (nota: 50 %)
    - *EPOC. Tabaquismo* (nota: 75 %)
    - *EPOC Etapa B. Tabaquismo IPA: 40.* (nota: 100 %)
  - Indicaciones:
    - *Salbutamol según requerimiento. O ipatropio* (nota: 30.6 %)
    - *Cese habito tabáquico. Salbutamol SOS. Salmeterol 2 puff c/8. Control en 1 mes* (nota: 66.7 %)
    - *Se incia tratamiento con LABA y salbutamol SOS. Se refuerza necisadad de dejar de fumar y se inicia plan conjunto para hacerlo. Se indica necesidad de vacunacion en inverno para prevenir exacerbaciones.* (nota: 88.9 %)

Inicialmente se pensó usar algoritmos de Word Embeddings tales como *Fasttext*, *GloVe* o *Word2Vec* y generar una calificación mediante un clasificador supervisado o un regresor. Sin embargo, se optó por usar una aproximación *bag-of-words* para representar los textos y luego un regresor basado en un MLP (*multilayer perceptron*) para obtener las predicciones. De acuerdo a eso, se revisarán en esa sección los principales algoritmos asociados.

Los textos de cada respuesta se representan en base a la frecuencia de términos usando el modelo *tf-idf* (*Term frequency – Inverse document frequency*).

TF-IDF: Normalmente, la ponderación tf-idf se compone de dos términos: el primero calcula la frecuencia de término normalizada (TF), también conocida como el número de veces que aparece una palabra en un documento, dividido por el número total de palabras en ese documento; el segundo término es la Frecuencia Inversa de Documentos (IDF), calculada como el logaritmo del número de documentos en el corpus dividido por el número de documentos donde aparece el término específico.

TF: (*Term frequency*) Mide la frecuencia con la que aparece un término en un documento. Dado que cada documento tiene una longitud diferente, es posible que un término aparezca muchas más veces en documentos largos que en documentos más cortos. Por lo tanto, la frecuencia de los términos a menudo se divide por la longitud del documento (también conocido como el número total de términos en el documento) como una forma de normalización.

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

IDF: (*Inverse document frequency*) Mide la importancia de un término. Al calcular TF, todos los términos se consideran igualmente importantes. Sin embargo, se sabe que ciertos términos, como 'es', 'de' y 'eso', pueden aparecer muchas veces pero tienen poca importancia. Por lo tanto, debemos sopesar los términos frecuentes mientras aumentamos los raros.

$$IDF_i = \log \frac{|D|}{|d : t_i \in d|}$$

Debido al frecuente uso de abreviaciones se decidió usar *n-gramas de caracteres* en vez de palabras completas, permitiendo llegar a mejores predicciones.
Esto conlleva la creación de vectores descriptores de alta dimensionalidad y con información redundante, lo cual implicó que fuera necesario mucha más memoria, más espacio en disco y especialmente más tiempo de entrenamiento y predicción.

Por este motivo se decidió reducir de la dimensión de estos vectores utilizando *Latent semantic analysis (LSA)*[7].

LSA: El análisis semántico latente es una técnica en el procesamiento del lenguaje natural, para analizar las relaciones entre un conjunto de documentos y los términos que contienen mediante la producción de un conjunto de conceptos relacionados con los documentos y términos.
LSA asume que las palabras que tienen un significado cercano aparecerán en fragmentos de texto similares. Se construye una matriz que contiene recuentos de palabras por documento (bag-of-words) a partir de una gran parte de texto y se utiliza la *descomposición de valor singular (singular value decomposition o SVD)*[11] para reducir el número de filas conservando la estructura de similitud entre columnas. Luego, los documentos se comparan tomando el coseno del ángulo entre los dos vectores (o el producto escalar entre las normalizaciones de los dos vectores) formado por dos columnas cualesquiera. Los valores cercanos a 1 representan documentos muy similares, mientras que los valores cercanos a 0 representan documentos muy diferentes.

[11] https://en.wikipedia.org/wiki/Singular_value_decomposition

2.3.2. Análisis de imágenes

La clases que se desean extraer de los videos son:

- **Espera:** Cuando el estudiante espera solo en el box, antes y después de interactuar con el paciente simulado.
- **Conversación:** Cuando el estudiante y el paciente simulado hablan, ya sea una entrevista o consejería al paciente.
- **Lavado de manos:** Cuando el estudiante se lava las manos, es esperable que se haga antes y después del examen físico.
- **Examen físico:** Cuando el estudiante realiza un examen físico al paciente simulado, usualmente en la camilla.

(a) Espera (b) Conversación

(c) Lavado de manos (d) Examen físico

Figura 2.3: Fotogramas de cada una de las clases

Existen múltiples formas para extraer las secciones temporales a partir de los videos, pero desde un principio se decidió utilizar una red neuronal convolucional (*Convolutional neural network* o *CNN*) para clasificar los fotogramas de los videos y con esto crear las secciones temporales.
Debido a que las cámaras en cada box se encuentran ubicadas en el mismo lugar y en la misma dirección, es posible hacer predicciones usando los fotogramas de los videos sin necesidad de usar información del audio o de la temporalidad de las imágenes.

CNN: Las redes neuronales convolucionales son una clase de redes neuronales profundas que han demostrado un alto rendimiento en la clasificación de imágenes. Dada a la naturaleza de las convoluciones dentro de estas redes, estas son aptas para poder aprender a clasificar todo tipo de datos donde estos estén distribuidos de una forma continua a lo largo del mapa de entrada, y a su vez sean estadísticamente similares en cualquier lugar del mapa de entrada. Por esta razón, son especialmente eficaces para clasificar imágenes, por ejemplo para el auto-etiquetado de imágenes.
Las CNN consisten en múltiples capas de filtros convolucionales de una o más dimensiones. Como redes de clasificación, al principio se encuentra la fase de extracción de características, compuesta de neuronas convolucionales y de reducción de muestreo. Al final de la red se encuentran neuronas de perceptrón completamente conectadas para realizar la clasificación final sobre las características extraídas. La fase de extracción de características se asemeja al proceso estimulante en las células de la corteza visual. Esta fase se compone de capas alternas de neuronas convolucionales y neuronas de reducción de muestreo. Según progresan los datos a lo largo de esta fase, se disminuye su dimensionalidad, siendo las neuronas en capas lejanas mucho menos sensibles a perturbaciones en los datos de entrada, pero al mismo tiempo siendo estas activadas por características cada vez más complejas.

2.3.3. Métricas de evaluación

Para evaluar la clasificación automática del rendimiento de los estudiantes en las respuestas del RCE, se utilizará el *error absoluto medio (MAE)* con respecto a las evaluaciones ya disponibles de los estudiantes.

MAE: (Mean absolute error) El error absoluto medio es una medida de la diferencia entre dos variables continuas. Considerando dos series de datos (unos calculados y otros observados) relativos a un mismo fenómeno, el error absoluto medio sirve para cuantificar la precisión de una técnica de predicción comparando por ejemplo los valores predichos frente a los observado.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i|$$

MSE: (Mean squared error) El error cuadrático medio de un estimador mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. El MSE es una función de riesgo, correspondiente al valor esperado de la pérdida del error al cuadrado o pérdida cuadrática. La diferencia se produce debido a la aleatoriedad o porque el estimador no tiene en cuenta la información que podría producir una estimación más precisa.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2$$

Para validar las secciones temporales predichas por la CNN, se validará la fidelidad y el uso provechoso de los datos extraídos con miembros del CHC y otros usuarios finales. En particular, se comparará la segmentación automática del tiempo del video con respecto a una segmentación manual generada por expertos y se comparará usando métricas de intersección de conjuntos, como por ejemplo *DICE* [12]. Para validar las predicciones de cada uno de los fotogramas se usaron las métricas *accuracy*, *precision*, *recall* y *F1-score*. Las definiciones de estas métricas están en el contexto de clasificación.

Accuracy: Es la proporción de predicciones correctas entre el número total de casos examinados.

Precision: Es el ratio entre el número de documentos relevantes recuperados entre el número de documentos recuperados.

$$precision = \frac{|\{documentos\ relevantes\} \cap \{documentos\ recuperados\}|}{|\{documentos\ recuperados\}|}$$

Recall: Expresa la proporción de documentos relevantes recuperados, comparado con el total de los documentos que son relevantes existentes, con total independencia de que éstos se recuperen o no.

$$recall = \frac{|\{documentos\ relevantes\} \cap \{documentos\ recuperados\}|}{|\{documentos\ relevantes\}|}$$

F1-score: Es la medida de precisión que tiene un test. Se emplea en la determinación de un valor único ponderado de la precisión y la exhaustividad (*precision* y *recall*).

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

[12] https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient

Capítulo 3

Problemas

Junto a la contraparte del proyecto se pudieron identificar distintos problemas a solucionar a través de mejoras en la plataforma **Kamaleon**, con el objetivo de incrementar la retroalimentación de los estudiantes y también disminuir las tareas de gestión y evaluaciones repetitivas asociadas. Los problemas se separan en 3 ejes principales:

1. **Gestión y registro de actividades:** La plataforma no se podía utilizar debido a errores, no permitía usarse durante las actividades y no unificaba todos los datos asociados a las actividades realizadas
2. **Evaluación del registro clínico electrónico:** Las respuestas del registro clínico electrónico pueden ser extensas, numerosas y toman tiempo en evaluar
3. **Evaluación de habilidades de entrevista:** Los videos de las interacciones contienen información relevante para la evaluación de las habilidades de entrevista, que no es posible analizar eficientemente sin ver horas de videos

3.1. Gestión y registro de actividades

En el CHC no se utilizaba ningún sistema centralizado ni estandarizado para la coordinación de las actividades, como tampoco para el registro detallado de dichas actividades. Las actividades se organizaban principalmente a través de email, los registros se hacían en archivos Excel y Word, y las evaluaciones se respondían usando Google Forms. Como no se usaba un sistema centralizado, no existía un manera práctica de vincular todos los datos relevantes de una actividad.

La plataforma Kamaleon es un primer acercamiento a resolver este problema, permitiendo además acceder a videos de las interacciones y que el paciente simulado evalué los encuentros.

Como se mencionó anteriormente, esta no pudo utilizarse debido a errores técnicos. Es por esto que surge la necesidad de corregir estos errores para así poder aprovechar todo el potencial de la plataforma.

Además de estos errores, se observó junto a los miembros del CHC que varias características podrían mejorarse para facilitar su uso y presentar la información de mejor manera.

En particular se manifestó que los estudiantes no pueden acceder a las actividades del día actual, como tampoco reciben toda la información relevante que podrían, para tener

simulaciones clínicas más realistas y posteriormente tener una reflexión más íntegra de las actividades realizadas.

3.2. Registro clínico electrónico

3.2.1. Descripción

El registro clínico electrónico (RCE) corresponde a un registro formal que contiene datos clínicos del paciente simulado como son anamnesis, examen físico, informe de resultados de exámenes, hipótesis diagnósticas e indicaciones.

Los estudiantes lo deben completar para cada interacción, pero no podían revisar sus propias respuestas después de enviarlas, ni ver la calificación obtenida y ni la pauta asociada. Los estudiantes sólo recibían una nota promedio de la actividad algún tiempo después a través de la plataforma U-Cursos, lo que limita la posibilidad de reflexionar sobre su desempeño, como también la posibilidad de reclamar la nota obtenida si considera que no fue apropiada.

Este registro es evaluado por el profesor según la pauta de evaluación dada por el caso clínico interpretado por el paciente simulado.

Este formulario se completaba utilizando formularios de Google Forms y luego el docente utilizaba la hoja de cálculo resultante de este formulario para evaluarlo. Los estudiantes no pueden acceder a esta hoja de cálculo ya que revelaría las respuestas de todos los estudiantes durante la actividad.

Otra de las desventajas de esta solución es la descentralización de la información, lo cual implica que el estudiante debía responder en múltiples ocasiones datos personales y de la interacción, tales como su nombre, sede, tipo de atención (consulta o urgencia) y el nombre del paciente, lo cual llevaba a la duplicación de datos ya que toda esta información debería estar registrada solo una vez (en la plataforma Kamaleon). Además, el uso de la hoja de cálculo resultante perjudica al profesor a la hora de evaluar las respuestas, ya que este debía asociar manualmente el caso clínico (cada uno con una distinta pauta de evaluación) al nombre ficticio de los pacientes simulados en la respuesta del estudiante antes de poder hacer la revisión.

Existen varios motivos por los cual el formulario no fue integrado a la plataforma desde un comienzo, principalmente porque esta no fue diseñada para ser utilizada durante las actividades por los estudiantes, sino que se filtraban de forma que los estudiantes sólo podían acceder a actividades ya finalizadas uno o más días antes. Otro motivo es que los componentes del formulario no estaban completamente definidos por parte de los miembros del CHC cuando se creó la plataforma.

3.2.2. Evaluación

Las respuestas del registro clínico electrónico pueden ser extensas y toma tiempo para evaluar, especialmente considerando el número de interacciones que se deben evaluar en cada actividad.

Se analizó la forma en que los docentes evalúan las respuestas de los estudiantes y el uso de las pautas asociadas, y se llegó a la conclusión que revisar una respuesta era una tarea bastante repetitiva y potencialmente automatizable. En términos simples, la revisión se reduce a encontrar ciertas palabras o frases de la pauta en la respuesta del alumno, para así sumar o restar cierto puntaje.

Pese a esto, el hecho que la revisión se hiciera en la hoja de cálculo para muchas interacciones a la vez aumentaba las posibilidades de error, especialmente considerando que es el mismo docente que debía asociar manualmente la pauta a cada respuestas, lo que además toma tiempo.

Con el fin de disminuir las tareas de evaluación repetitivas asociadas a las actividades, se propuso utilizar las respuestas redactadas por los estudiantes para clasificar de forma automática su rendimiento, calificación que el docente podrá aprobar o modificar si lo desea.

3.3. Evaluación de habilidades de entrevista

El ECLIPSE da la oportunidad de generar diversas oportunidades de observación y evaluación de habilidades clínicas. Las interacciones se pueden observar directamente por vidrio espejo, registro de video en vivo y ahora a través de la plataforma, lo que es útil en la evaluación de habilidades comunicacionales y de entrevista clínica, con pautas aplicadas por pacientes simulados y docentes.

La plataforma permite crear formularios personalizables que los pacientes simulados contestaran después de cada interacción, estas preguntas son formuladas de acuerdo con la pauta CAT (Communication Assessment Tool), MIRS (Master Interview Rating Scale) y pautas de observación de anamnesis y examen físico.

Después de las actividades los alumnos podrán acceder a estos resultados junto a cada video a través de la plataforma, permitiendo así reflexionar sobre su desempeño y sacar mejor provecho de la experiencia.

Ya que estos videos contienen información sumamente relevante para el aprendizaje y la evaluación, y que para los ECliPSEs existen una serie de acciones a realizar que son estándares y deben desarrollarse en un tiempo finito, se propuso realizar un análisis automático de los videos, el cual rescataría datos clave de las interacciones y permitiría encontrar rápidamente distintas partes de un video como también el poder analizar información de múltiples interacciones sin tener que ver cada uno de los videos.

Capítulo 4

Solución

4.1. Detección y corrección de errores

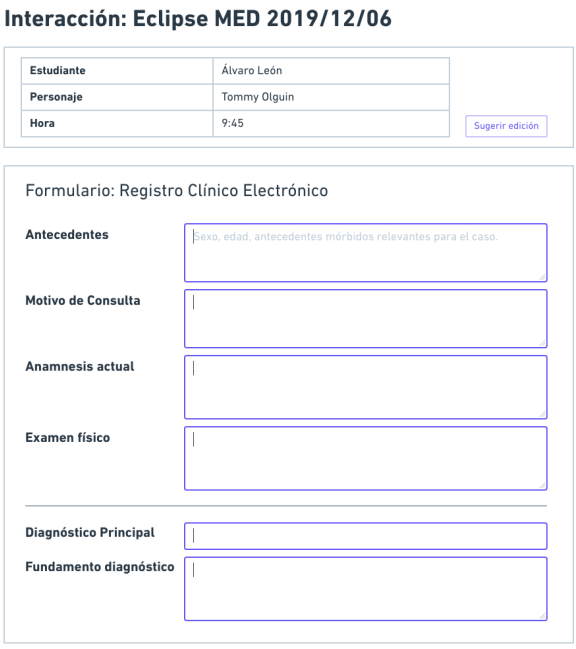
En primer lugar se corrigieron los errores que impedían el uso de la plataforma. Estos estaban asociados principalmente a problemas con la configuración en Docker y se hicieron evidentes cuando se reinicio el servidor.
Se detectaron y corrigieron los siguientes errores para asegurar el correcto funcionamiento del Kamaleon:

- Se corrige error en la solicitud del video de una interacción, ya que no estaba funcionando la solicitud a una API que prepara el video para de actividad. El problema estaba en la configuración del componente NVR API (Ver 2.1) en Docker, cuando se intentaba levantar la aplicación se intentaba instalar dependencias inexistentes e innecesarias.
- Se corrige la subida y lectura de PDFs asociados a los distintos casos clínicos, el problema estaba en los permisos que tenía la carpeta contenedora en el servidor.

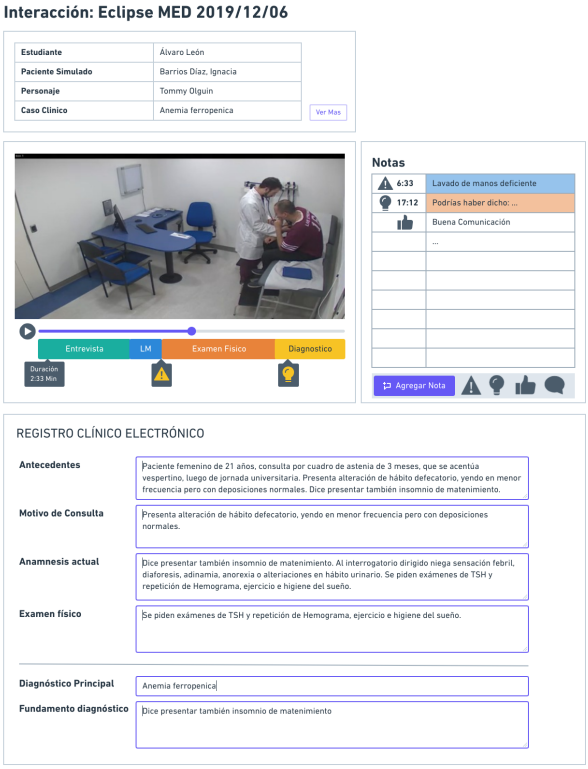
4.2. Mejora de las interfaces de usuario y del flujo de información

Creación de mejorarás y de nuevas vistas existentes, con énfasis en unificar el ingreso de datos asociados a las actividades realizadas, gestionar las actividades en tiempo real y tener una mejor visibilidad de los datos relevantes para cada usuario.

Antes de empezar la etapa de desarrollo, se propusieron distintos bosquejos de interfaces de usuario, los cuales fueron presentados a miembros del CHC junto a cuáles serían las nuevas funcionalidades a desarrollar. Estos bosquejos corresponden a la vista de una interacción para un estudiante y para el docente (Ver Figura 4.1).



(a) Vista del estudiante de una interacción en progreso



(b) Vista del docente o administrador de una interacción finalizada

Figura 4.1: Bosquejos propuesto para las vistas de interacción

4.2.1. Integración del Registro Clínico Electrónico

Como se mencionó anteriormente, el registro clínico electrónico es un registro formal que los estudiantes realizan para cada interacción, el cual contiene datos clínicos del paciente simulado como son anamnesis, examen físico, informe de resultados de exámenes, hipótesis diagnósticas e indicaciones.

En primer lugar se definieron los campos y el formato que tendría este formulario.
Los campos usados en años anteriores en Google Forms (omitiendo los datos del alumno y de la actividad) son:

- Antecedentes del paciente
- Motivo de consulta
- Anamnesis actual
- Examen físico
- Diagnósticos
- Indicaciones

Los campos para diagnósticos e indicaciones se encontraban en dos configuraciones: primero se usaba un solo campo de texto para cada uno, pero después se optó por dividir la respuesta por prioridad, es decir, existían campos para diagnóstico e indicación principales y otros 4 campos para otros posibles diagnósticos e indicaciones.

Se decidió que el formulario en la plataforma debía tener un solo campo de texto para las indicaciones y permitir ingresar 1 o más diagnósticos ordenados por prioridad.

Para ingresar los diagnósticos, se usó código en *Javascript* que ya se utilizaba en la plataforma para crear las pautas de evaluación de los pacientes simulados, el cual permite:

- Crear el número deseado de respuestas de texto, ordenadas por prioridad
- Editar y eliminar repuestas
- Cambiar la prioridad de las respuestas arrastrándolas con el mouse
- Este además conserva el estilo general del sitio en *Bootstrap*

Cuando se presentó el nuevo formulario en la plataforma a los miembros del CHC, se manifestó la utilidad de integrar a la vista y al formulario:

- **Antecedentes del paciente/caso:** Corresponde a información clínica sobre el paciente simulado, que puede incluir datos generales del paciente, atenciones previas y resultados de distintos exámenes. Esta información se sube en formato PDF y se presenta junto al registro clínico electrónico, permitiendo a los alumnos tener una simulación más realista para pacientes con atenciones previas.
- **Datos Significativos/Relevantes:** Con toda la información que tendrán disponible los alumnos, se decide reemplazar el campo *Antecedentes del paciente* por *Datos Significativos*, ya que no es importante compilar todos los antecedentes del paciente, sino que basta con listar los datos significativos que resuman lo más importante y que permitan justificar los diagnósticos planteados. Al igual que los diagnóstico, se permite ingresar 1 o más datos ordenados por prioridad.

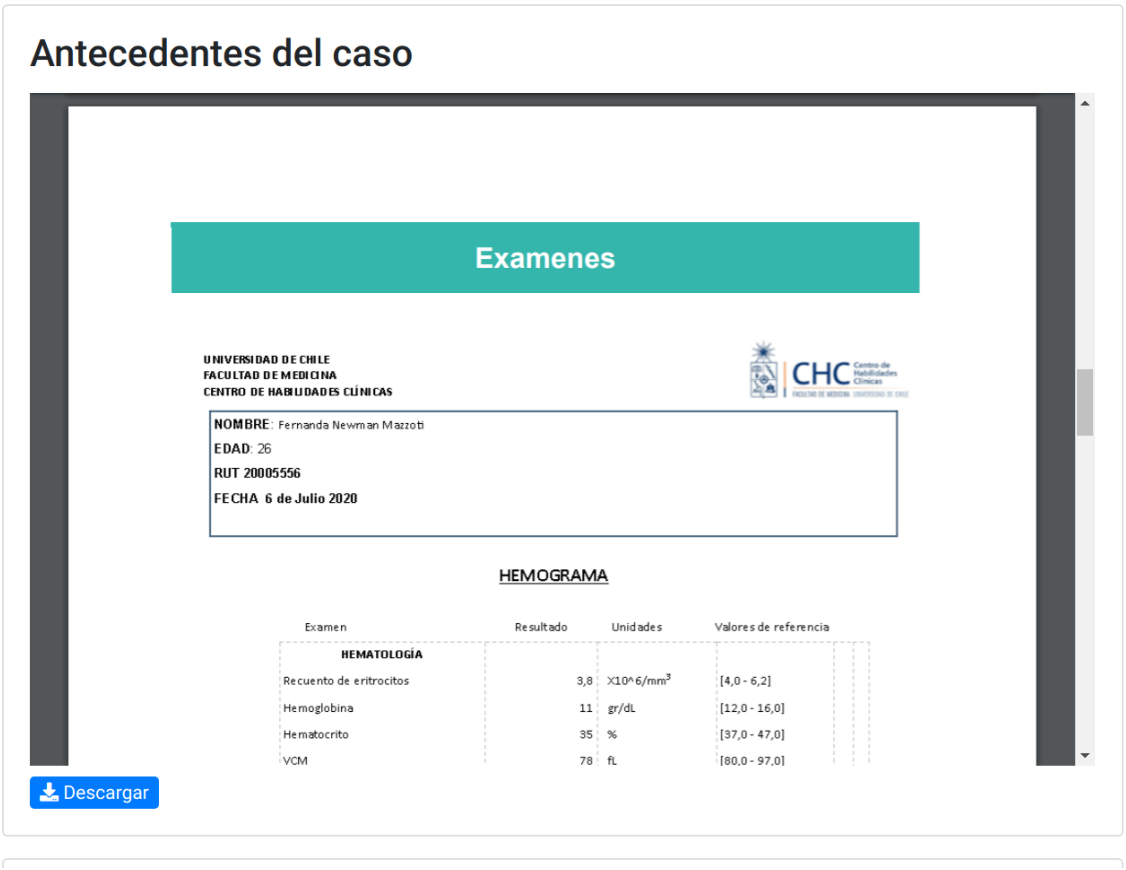


Figura 4.2: Antecedentes del caso

Para asociar los datos significativos que justifican a uno o más diagnósticos, se modificó el campo de diagnósticos para que en cada uno de los diagnósticos planteados muestren una lista de *checkbox* correspondientes a los datos significativos anotados, que el alumno deberá marcar para hacer la asociación.

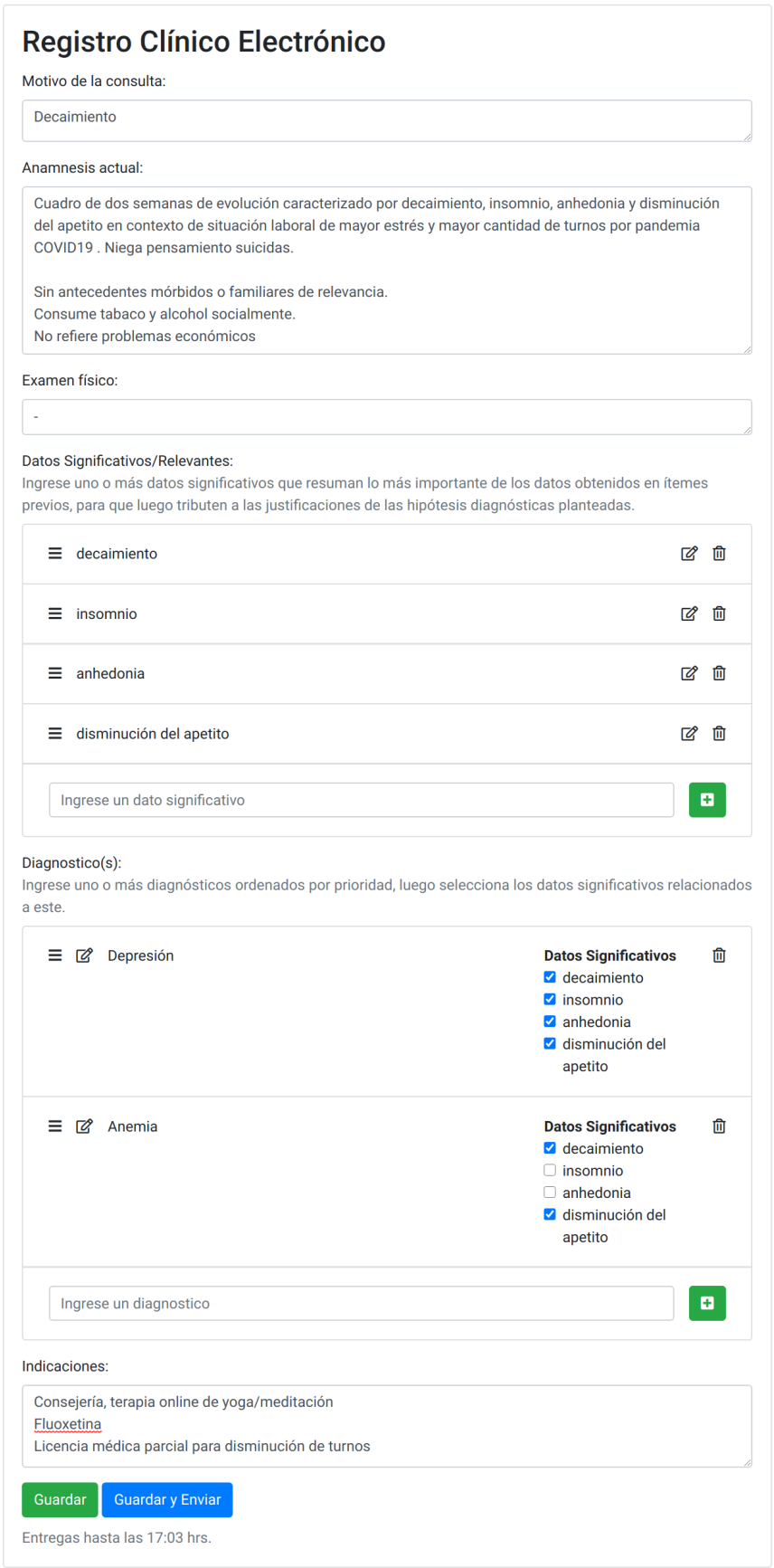


Figura 4.3: Registro clínico electrónico completado

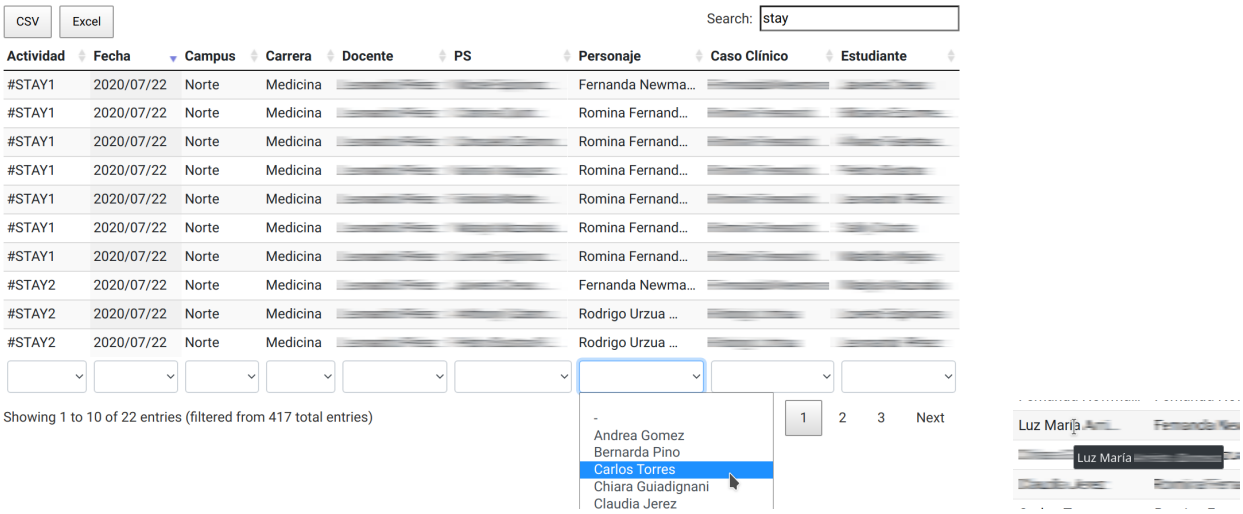
4.2.2. Datos Históricos

Se crearon nuevas vistas que permiten a los usuarios ver en tablas un resumen de todas las interacciones asociadas a él (todas las existentes para los administradores).

En el backend se filtran los datos relevantes para el usuario y en el frontend utilizó la librería *DataTables*[1] para visualizar esta información. Esta librería permite:

- Ordenar por columna
- Filtrar filas tal que contengan un elemento dado en una columna
- Filtrar filas tal que contengan un *string* dado en cualquiera de las columnas
- Exportar esta información filtrada y ordenada en formato Excel o CSV

Ya que son muchos los parámetros asociados a cada interacción, se decidió que algunas de las columnas solo serán visibles en los archivos exportados mientras que en la tabla solo se mostrarían columnas seleccionadas. Por esta misma razón se limita el número de caracteres a mostrar de cada celda, pero se podrá ver el contenido completo en los filtros, a través de un tooltip y en los archivos exportados.



(a) Filtros en la tabla (b) Tooltip

Figura 4.4: Tabla de datos históricos

El archivo exportado incluye datos generales de la actividad y las respuestas del registro clínico electrónico escrito por el estudiante, permitiendo así tener una fuente centralizada y estandarizada de los datos.

[1] https://datatables.net/

4.2.3. Actividades por Videoconferencia

La contingencia por el COVID-19 resultó en una creciente popularidad de la telemedicina, y ya que no se pudieron realizar ninguna de las actividades en el CHC como se habían planificado, se acordó adaptar la plataforma para la gestión de actividades por videoconferencia.

Para este nuevo tipo de actividades se establecieron reglas que modifican la plataforma a nivel de interfaz y reglas del nuevo flujo para cargar los videos a la plataforma.

En las interfaces se limitan las opciones para la creación y modificación de actividades a ser tipo Consulta y no de Urgencia, por lo tanto se bloquean las opciones asociadas a urgencias cuando la actividad es por videoconferencia.

No se obtiene el vídeo de la interacción a través de la API del NVR (Ver 2.1), sino que permite a los pacientes simulados subir los videos provenientes de videoconferencias (en particular, se usa la herramienta de videoconferencias *Zoom*).
Para realizar este tipo de actividades, se debe tener abierto en el computador de todos los usuarios la plataforma en el navegador y la aplicación Zoom.

Antes de realizar pruebas con usuarios reales, ya se habían realizado pruebas de atenciones simuladas remota con pacientes simulados, por lo cual miembros del CHC y algunos pacientes simulados ya se encontraban familiarizados con el uso de Zoom Rooms para formar consultas y luego reunir los participantes en una sola sala para el debriefing.

Cuando se realizaron las pruebas de validación, los participantes recibieron instrucciones sobre cómo usar el Kamaleon junto con Zoom para grabar videos de cada interacción (cuándo comenzar, pausar o detener la grabación), y los pacientes simulados no tuvieron problema en seguir estas instrucciones para generar un video por cada encuentro.

4.2.4. Gestión de actividades en tiempo real

Se adecuaron las vistas de cada uno de los roles de usuario para permitir la gestión de actividades en tiempo real.

Primero se definieron distintos estados para las actividades e interacciones:

- Las actividades se agrupan según su fecha de inicio en actividades de Hoy, Futuras, y Pasadas.
- Las interacciones tendrán un estado según la hora y fecha:
  - **Inactiva:** Si la interacción y/o actividad aun no empieza
  - **En Proceso:** Periodo durante la actividad, desde la hora de inicio de la interacción hasta que finaliza. Este tiempo tiene como duración la suma de los minutos de atención y evaluación, determinados al crear la actividad
  - **Finalizado:** Después de pasado el tiempo de atención y evaluación

Cada usuario visualizará las actividades e interacciones de distinta manera según el estado de las interacciones. En síntesis:

- **Administradores:** Tienen acceso a una vista general para administrar la actividad en todo momento, permite ver el estado de todas las interacciones durante actividades y hacer cambios pertinentes en todo momento, también podrá ingresar a una interacción y hacer los cambios que desee.
- **Docentes:** Tienen permisos similares a los administradores para gestionar actividades existentes, aunque solo pueden acceder a actividades donde se registró como docente a cargo.
- **Pacientes simulados:**
  - Se bloquean o ocultan ciertos elementos principalmente porque no son relevantes para estos usuarios, tales como el RCE
  - Pueden contestar la pauta de evaluación asociada a cada interacción (esta es diferente según el caso clínico que está interpretando.)
- **Estudiantes:** Tienen los permisos más restrictivos:
  - Solo pueden acceder a actividades pasadas o la del día actual
  - En la actividad del día, solo podrán acceder a la interacción que se encuentren finalizada o en proceso.
  - Solo pueden responder el RCE en interacciones en progreso
  - No podrán visualizar el caso clínico del paciente simulado hasta que la interacción finalice y la actividad sea marcada como *publicada* por el administrador o docente a cargo. Esto con el fin de evitar que los alumnos sepan el diagnóstico correcto de los pacientes simulados y puedan compartirse esta información

4.3. Clasificación automática del rendimiento de los estudiantes

Se clasifica el rendimiento de los estudiantes de forma automática en las respuestas del registro clínico electrónico. Esto va acompañado de la exploración de los datos existentes y de una mejor definición de las pautas de evaluación.

El RCE se evalúa con una pauta diferente según el caso clínico del paciente simulado, en particular se evalúa en función de los diagnósticos y de las indicaciones.
Se califican los diagnósticos y las indicaciones por separado, pero estas notas se promedian y constituyen una de las distintas evaluaciones que recibe el estudiante en cada actividad. Con el fin de dar un mejor feedback es importante que el alumno conozca la nota por separado.

Las pautas de evaluación de un caso constan de 2 secciones para el diagnóstico y para las indicaciones, que podemos simplificar como: frases o palabras que debe contener la respuesta para recibir puntaje parcial o total, y frases o palabras que descuentan puntaje o lo anulan por completo (no siempre se especifica).
Para los diagnósticos estas palabras hacen referencia directamente al caso clínico, usualmente el puntaje perfecto corresponde al nombre del caso clínico más otras afecciones asociadas a este.

4.3.1. Descripción del dataset

Se tienen respuestas de los RCEs de los años 2018 y 2019 en formato Excel, dando un total de **2454** respuestas. Los parámetros de interés de cada respuesta son:

- **Diagnósticos:** Una o más columnas de diagnósticos, escritas por el estudiante en texto libre
- **Indicaciones:** Una o más columnas de indicaciones, escritas por el estudiante en texto libre
- **CasoID:** Número único correspondiente a cada Caso Clínico, escrito por el docente
- **Porc DG:** Calificación del diagnóstico de 0 % a 100 %, escrito por el docente
- **Porc IND:** Calificación del indicaciones de 0 % a 100 %, escrito por el docente

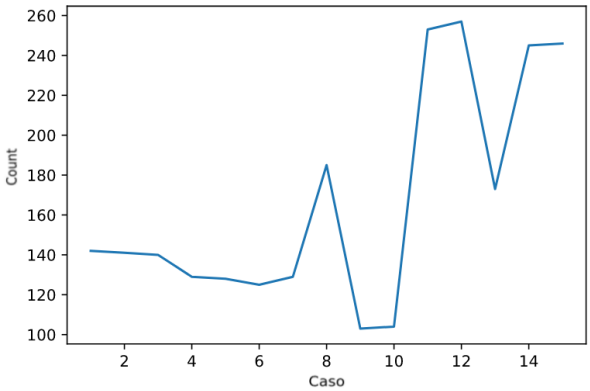
Primero que nada se debió estandarizar los CasoID, ya que en algunas ocasiones se usaba un mismo número para diferentes casos. Se definieron **15** casos clínicos distintos.

También se debió estandarizar las columnas de respuestas: cuando se tenían múltiples columnas, estas se concatenaron para así tener un input estandarizado de todos los datos disponibles.
Se hizo la separación de columnas con el carácter **&**, ya que no se encuentra en ninguna de las respuestas. Por otro lado, cuando se tienen múltiples respuestas en una sola columna, se insertó el carácter **&** para reemplazar los separadores más comunes usados por los alumnos, tales como ':', ',', ';', '\n', ' - ', ' + ', '1.- ', '2) ', etc.

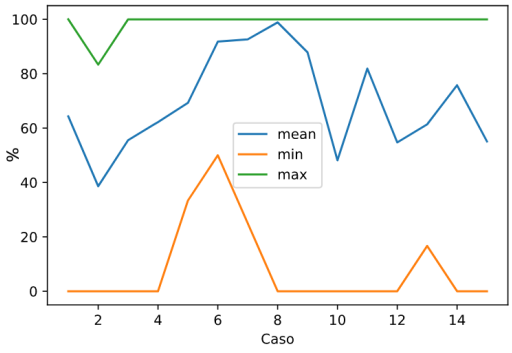
CasoID	Caso Clínico
1	EPOC GOLD 2
2	Neumonía
3	Asma – Rinitis alérgica
4	Neumonía adquirida en comunidad ATS I
5	Sepsis de foco urinario
6	Síndrome Coronario Agudo con/sin SDST
7	Cetoacidosis diabética
8	Hipotiroidismo primario
9	Lumbago mecánico
10	Síndrome metabólico
11	Anemia ferropénica
12	Enfermedad renal crónica
13	Várices esofágicas – Daño hepático crónico
14	Crisis hipertensiva
15	Diarrea disentérica en vías de prolongación

Tabla 4.1: Casos Clínicos

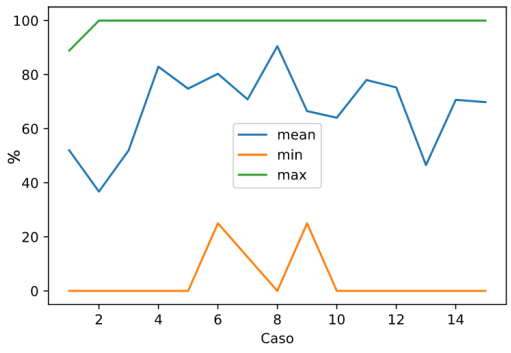
La cantidad de respuestas por cada caso clínico es dispareja, como también la distribución de calificaciones por caso clínicos:



(a) Cantidad de respuestas



(b) Distribución de notas de Diagnósticos



(c) Distribución de notas de Indicaciones

Figura 4.5: Distribución de los datos en función de cada caso clínico

4.3.2. Clasificación de diagnósticos

Se usó el modelo *bag-of-words* para representar los textos correspondientes a las distintas respuestas. Utilizando la frecuencia de cada palabra se logró tener un mejor entendimiento de las respuestas, mejorar la limpieza de los textos y también encontrar inconsistencias en la revisión.
En el Anexo B se encuentran los textos preprocesados agrupados por la frecuencia de palabras, organizados por el caso clínico y nota asociada.

Se generaron clasificaciones mediante un regresor [2], el cual tiene como input el vector de frecuencias de palabras concatenado con la representación *One-hot Encoding* del CasoID.

Se probaron distintas variaciones de esta estrategia y preprocesado de los textos, tal como el uso de *stop-words*, uso de distintos rangos de *n-gramas*, uso de *n-gramas de caracteres* (n-gramas con substrings de palabras), reducción de dimensionalidad, entre otros.

Buscando minimizar el *error absoluto medio (MAE)* de las predicciones de las notas (valores reales en el rango `[0.0, 1.0]`), se llegó a la siguiente configuración:

1. Se crean vectores con el modelo *Term frequency – Inverse document frequency* (*tf-idf*) [3], usando *n-gramas de caracteres* con largos en el rango `[2, 16]`, lo cual nos retorna vectores de dimensión `>50000`.
2. Se reduce la dimensionalidad del vector usando *Latent Semantic Analysis* (*LSA*) [7] y se obtiene una representación vectorial de dimensión `500`. Esto permite entrenar y obtener predicciones más rápidamente que usando los n-gramas de caracteres directamente.
3. Los modelos *tf-idf* y *LSA* ajustados se guardan en memoria, permitiendo crear la representación vectorial de las respuestas rápidamente para hacer predicciones en producción. Se espera que el uso de n-gramas de caracteres y LSA entreguen un mejor vector descriptor cuando se ingresen palabras que no están en el dataset (como se observó en la validación), ya que las palabras pueden ser muy variadas debido a abreviaciones o faltas de ortografía
4. Concatenar el vector con la representación *One-hot Encoding* del CasoID.
5. Se generaron predicciones con una red neuronal utilizando un modelo secuencial con 6 capas ocultas densamente conectadas, un vector de entrada de dimensión `515` y una capa de salida que devuelve un único valor continuo. Se utiliza `ReLU (Rectifier Linear Unit)` como función de activación. Para el entrenamiento:
   - Se usó el 10 % del dataset para validar (261 respuestas de validación)
   - Se incluyeron capas de regularización *(Dropout)* y *Early-Stopping* para evitar el sobreajuste *(overfitting)*, deteniendo el entrenamiento después de aproximadamente `120 epochs`
   - Se usaron los hiperparametros **Loss:** `MAE` y **Optimizer:** `Adam (Adaptive Moment Estimation Optimizer)`

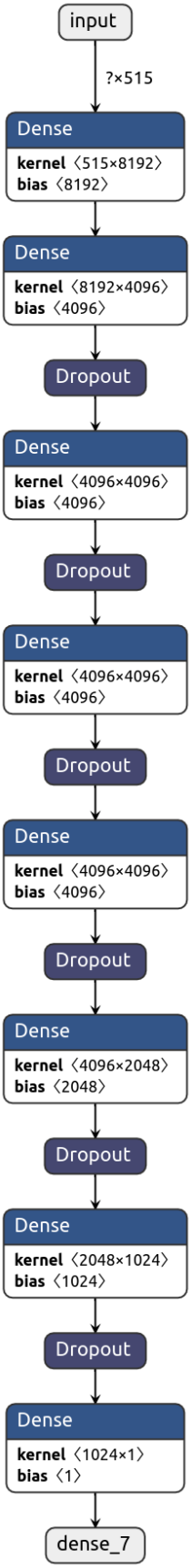
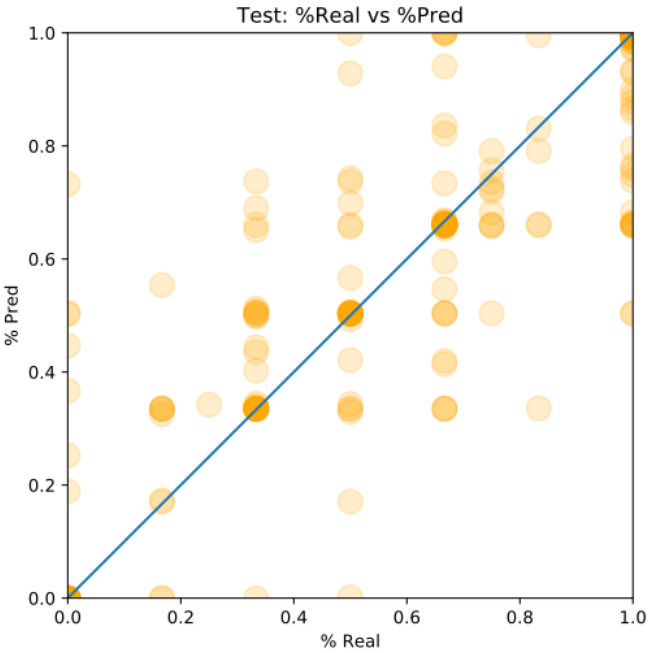
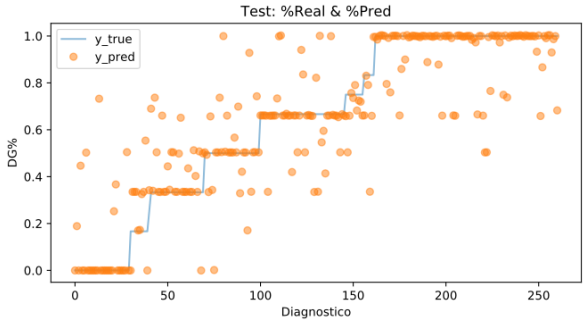


Figura 4.6: Arquitectura de la red neuronal para la clasificación de diagnósticos clínicos



(a) Real vs predicción



(b) Real y predicción

(c) Distribución del error

Figura 4.7: Evaluación (261 respuestas de validación)

De esta manera se llegó a un MAE: 0.9, MSE: 0.027.
En los anexos (Figura B.1 y B.2) se encuentra la evaluación separada para cada uno de los casos clínicos, y en el capitulo 5.2 (Validación de la clasificación automática del rendimiento de los estudiantes) se analiza con más detalle los resultados obtenidos.

Se optó por esta estrategia tradicional en vez de algoritmos de Word Embeddings tales como *Fasttext*, *GloVe* o *Word2Vec* debido a que:

- El dataset no es lo suficientemente grande como para que los algoritmos de Word Embeddings aprendan las peculiaridades del dominio
- El lenguaje en las respuestas es altamente técnico y se utilizan muchas abreviaciones, lo cual restringe el uso de algoritmos de Word Embeddings pre-entrenados
- Es esperable que oraciones y palabras de las respuestas se repitan y además coincidan con las pautas de cada caso clínico

El uso del modelo *tf-idf* para describir los textos facilitó la interpretación y análisis de los vectores (*n-gramas*) correspondientes a las respuestas, debido a la similitud de la revisión de las pautas con el aprendizaje del regresor, en particular facilita:

- Encontrar similitudes entre las respuestas de los estudiantes e identificar las respuestas más comunes para cada caso clínico
- Encontrar irregularidades e inconsistencias en la revisión del RCE (respuestas similares con distinta nota)
- Identificar casos clínicos con variaciones en las respuestas y con pautas de evaluación ambiguas

Podemos respaldar el uso de este método por la similitud del problema a resolver con el estudio "*TF-IDF vs word embeddings for morbidity identification in clinical notes: An initial study*"[3], donde se comparan distintos métodos de Deep Learning y Word Embeddings para identificar 16 tipos de morbilidad dentro de descripciones textuales de registros clínicos. En este estudio se emplearon las técnicas de aprendizaje en procesamiento del lenguaje natural *GloVe* y *Word2Vec* junto a un modelo *Bidirectional Long-Short Term Memory (LSTM)* y se comparó el rendimiento con el modelo *tf-idf* utilizando *Support Vector Machine (SVM)* y *Multilayer perceptron (MLP)*.
En este estudio, se evaluó el desempeño de las distintas estrategias usando el `F1-score`:

- Tf-Idf / Support Vector Machine (`F1-score promedio: 98.47`)
- Tf-Idf / Multilayer Perceptron (`F1-score promedio: 97.72`)
- GloVe Pre-entrenado (`F1-score promedio: 91.21`)
- Word2Vec Pre-entrenado (`F1-score promedio: 77.64`)
- Word2Vec Entrenado en el dominio (`F1-score promedio: 56.20`)

Se observó como el uso del modelo *tf-idf* supera a todos los enfoques de aprendizaje profundo (en cada uno de los casos) y se concluye que la razón de esto son las de características específicas que aparecen solo para una de las categorías que el modelo *tf-idf* describe.

4.3.3. Indicaciones

Las respuestas de los estudiantes en las indicaciones son mucho más abiertas a interpretación comparado con los diagnósticos (Anexo B). Estas expresan instrucciones en lenguaje natural como también recetan medicamentos usando sus nombres comerciales y con distintas unidades de medidas y formatos.

Al notar que la estrategia usada para los diagnósticos no estaba dando buenos resultados para predecir la calificación en las indicaciones, se decidió centrar el esfuerzo en la predicción de diagnósticos, donde los resultados eran más prometedores.

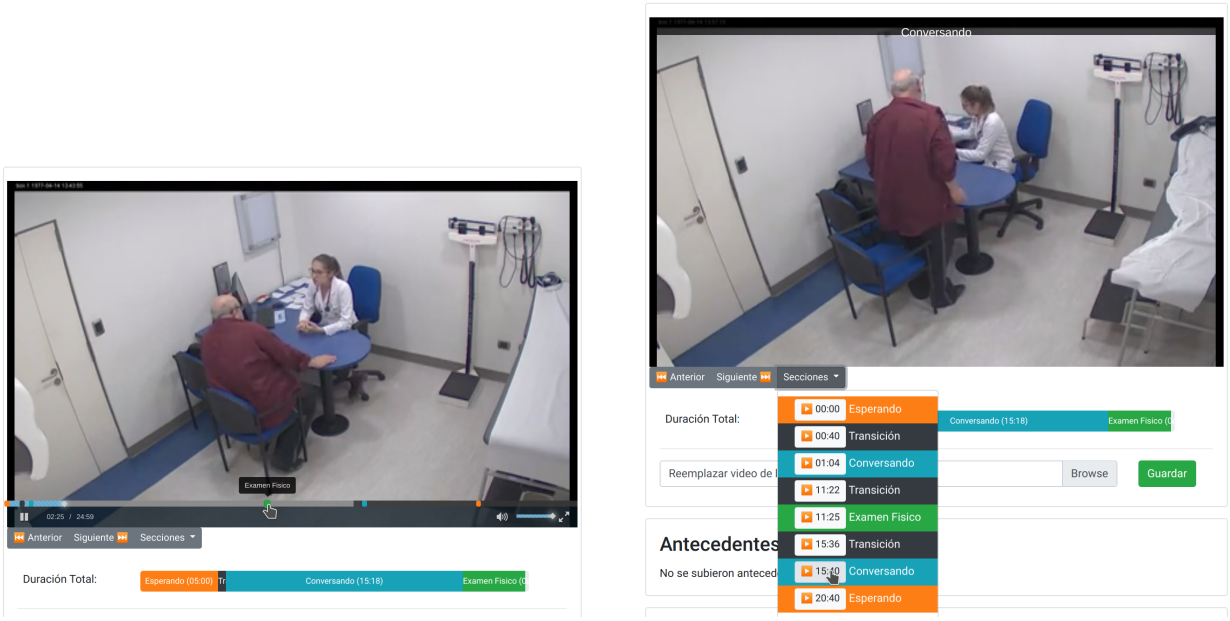
Para lograr mejores resultados se requeriría de un dataset mas completo y técnicas más avanzadas de procesamiento de lenguaje natural.

4.4. Análisis automático de los videos

Se mostró interés en generar secciones temporales en los videos y mostrar la duración de cada una. Las acciones estandarizadas para las actividades de tipo ECliPSE (detalle de estas actividades en el Anexo A) que se desean extraer de los videos (y el nombre de la clase a clasificar) son:

- **Espera (`waiting`):** Cuando el estudiante espera solo en el box, antes y después de interactuar con el paciente simulado.
- **Conversación (`talking`):** Cuando el estudiante y el paciente simulado hablan ya sea una entrevista o consejería (diagnóstico) al paciente. Se pretendía clasificar por entrevista y por diagnóstico por separado, pero resulta ser innecesario ya que ambas corresponden a los participantes hablando y solo se diferencian en que la entrevista ocurre antes del examen físico, mientras que el diagnóstico ocurre después.
- **Lavado de manos (`washing_hands`):** Cuando el estudiante se lava las manos, es esperable que se haga antes y después del examen físico, y que tenga una duración adecuada de al menos 30 segundos.
- **Examen físico (`physical_exam`):** Cuando el estudiante realiza un examen físico al paciente simulado, usualmente en la camilla.

Para mostrar esta información, se diseñó un reproductor de video usando el framework *Video.js*[2] y el plugin *videojs-markers*[3], acompañado de otros controles en la interfaz de las interacciones.



(a) Marcas de tiempo (b) Menú desplegable

Figura 4.8: Video de interacción con marcas de tiempo

[2] https://videojs.com/
[3] http://www.sampingchuang.com/videojs-markers

Se tenían guardados en el servidor 20 videos, pero solo 9 estaban correctamente recortados a la duración de cada interacción. Estos videos corresponden a un total de `3.58` horas, equivalente a `386632` fotogramas.

Se utilizó una red neuronal convolucional (CNN) la cual se entrenó y validó usando cada cuadro de imagen con los videos existentes en la plataforma.

Se utilizaron varias configuraciones de entrenamiento usando 8 videos para entrenamiento y 1 para validación. Los videos de validación seleccionados corresponden a interacciones con distintas combinaciones de estudiantes y pacientes simulados para asegurarnos que la CNN permita generalizar al predecir. Se muestran los resultados de validación obtenidos con uno de estos videos, en el cual se realizan todas las clases de acciones a predecir.

Experimentando con distintas arquitecturas de redes convolucionales, se llegó a los resultados deseados usando 4 capas para la extracción de características usando `filtros (kernels)` de `3x3`, `ReLU` y `max-pooling` de `2x2`, mientras que para la clasificación se usaron 4 capas densamente conectadas usando `ReLU` como función de activación exeptuando la última capa donde se usó `softmax`.
Los hiperparametros utilizados en el entrenamiento de la CNN fueron **Optimizer:** `Adam (Adaptive Moment Estimation Optimizer)` y **Loss:** `Sparse categorical crossentropy`. Se incluyeron capas de regularización y *Early-Stopping* para evitar el sobreajuste.

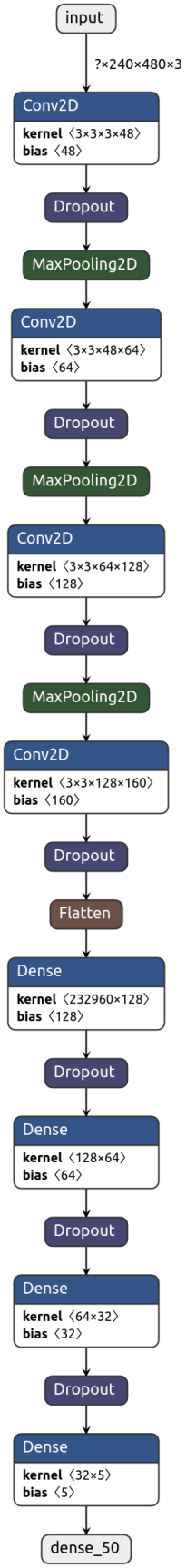
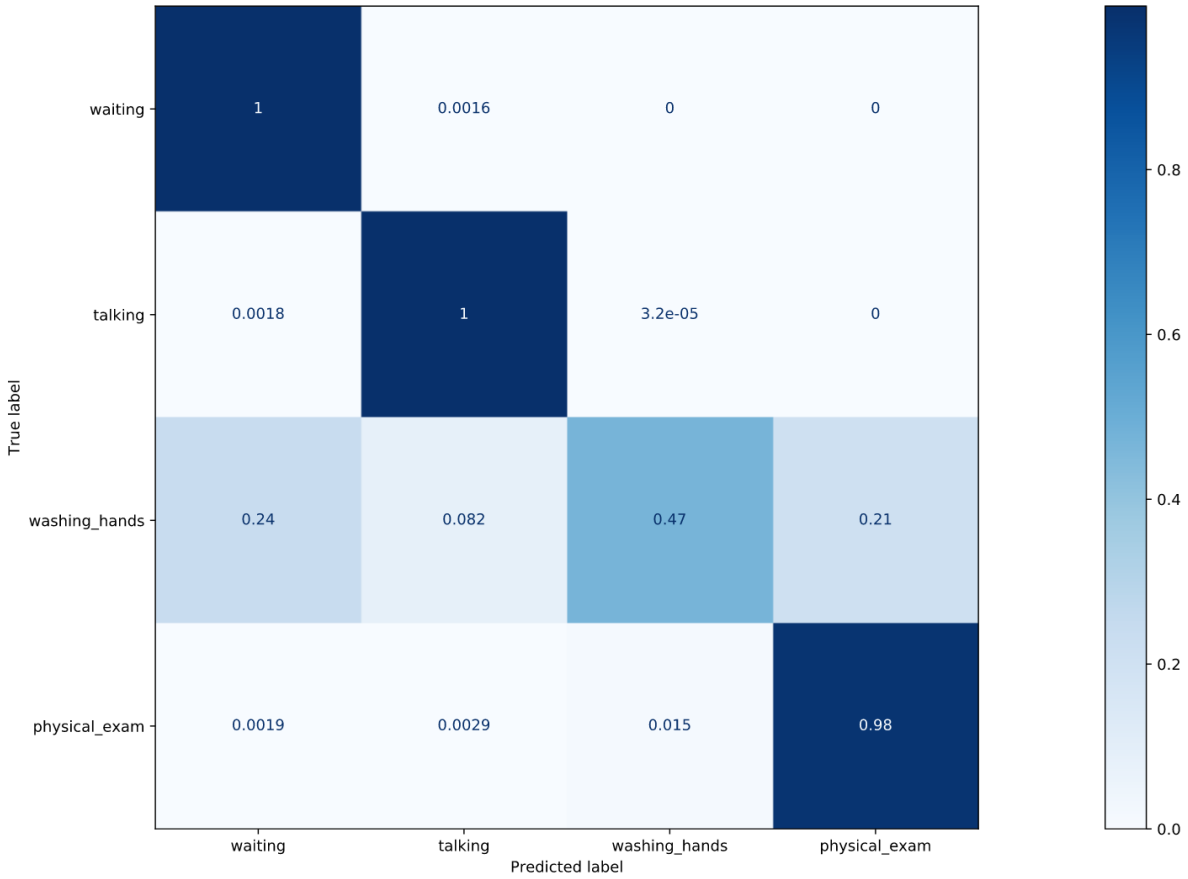


Figura 4.9: Arquitectura de la CNN

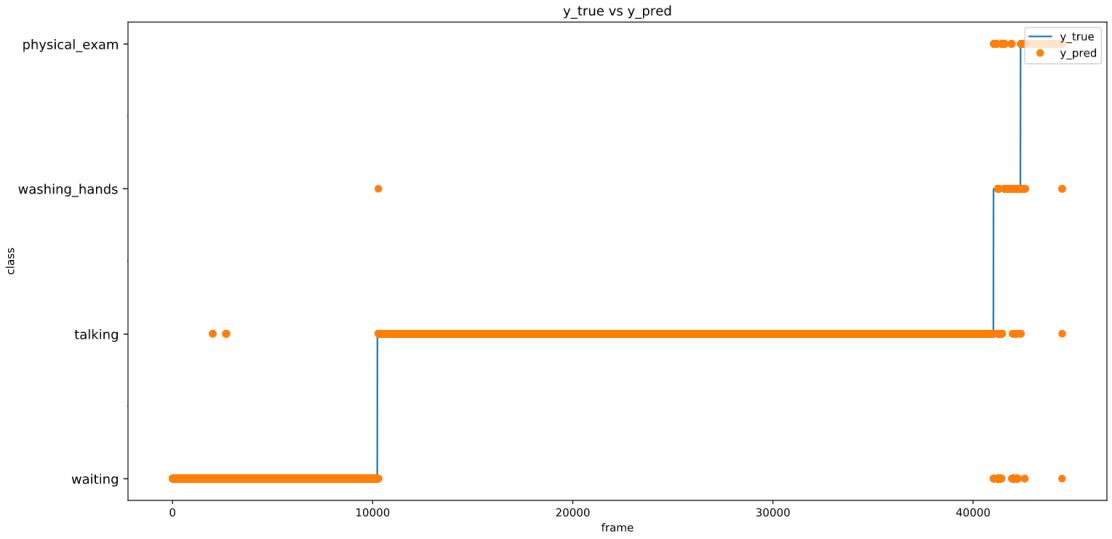
Con este clasificador llegamos a un **Accuracy: 0.98** y **F1-score: 0.98** promedio (**F1-score: 0.88** promediando el score de cada clase), más detalles de los resultados en la tabla 4.2 y en el capitulo 5.3 (Validación del análisis automático de los videos) se analiza con más detalle los resultados obtenidos.

	precision	**recall**	**f1-score**	**support**
waiting	0.96	1.00	0.98	10232
talking	1.00	1.00	1.00	30781
washing_hands	0.95	0.47	0.63	1351
physical_exam	0.88	0.98	0.93	2099
accuracy			0.98	44463
macro avg	0.95	0.86	0.88	44463
weighted avg	0.98	0.98	0.98	44463

Tabla 4.2: Métricas resultantes al evaluar el clasificador



(a) Matriz de confusión normalizada



(b) Clase real y predicha de cada cuadro del video, permite reconocer *outliers*

Figura 4.10: Evaluación con un video

Capítulo 5

Validación

Se validó el correcto funcionamiento de la plataforma a través de pruebas funcionales, estructurales y de integración para asegurar el correcto funcionamiento de la plataforma [5].

Además, se utilizaron metodologías ágiles para el desarrollo de las tareas, con reuniones periódicas con miembros del CHC que permitieron validar constantemente las mejoras en la plataforma.

Inicialmente se planificó evaluar el funcionamiento de la plataforma usando las múltiples actividades que se realizan en el centro durante el semestre, pero debido a la contingencia mundial por el COVID-19 y posterior paro en la Facultad de Medicina de la Universidad de Chile, esto no fue posible.

5.1. Validación de interfaces

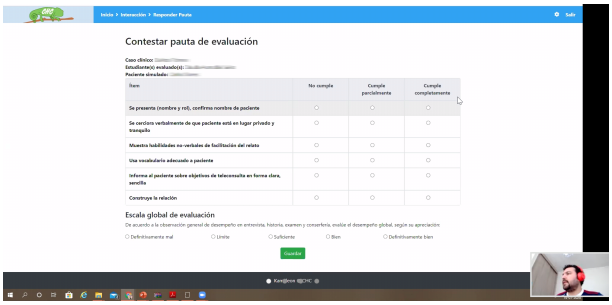
Durante el semestre, la validación de las interfaces solo se pudo realizar con miembros del CHC, quienes tendrían roles de administrador y docente en la plataforma, quedando pendiente la validación por parte de estudiantes y pacientes simulados.
Se valida constantemente con los miembros del CHC que la plataforma realmente facilite la gestión de las actividades y que los cambios que se realicen en las interfaces mejoren la usabilidad, pero no se hizo una validación formal de usabilidad hasta el final de la etapa de desarrollo.

5.1.1. Actividades de prueba

Como se mencionó anteriormente, se adaptó la plataforma para la gestión de actividades por videoconferencia, permitiendo que al final de la etapa de desarrollo se organizaron 2 actividades remotas con usuarios reales.

Los usuarios de estas actividades fueron pacientes simulados y alumnos, quienes permitieron validar las nuevas funcionalidades de la plataforma como también evaluar su usabilidad, ya que para ellos fue la primera vez que interactuaban con la plataforma. También permitieron a los administradores y docentes usar la plataforma en un escenario realista.
Estas actividades se llamaron **STAY** (por Simulated Telemedicine Attention sYstem), la primera de estas correspondió a un piloto con 8 actores del CHC, donde 4 hacían el papel de alumno y 4 de paciente simulado (su rol real en la plataforma). Cada uno de los pacientes

simulados tenía asignado un casos clínico distinto y se hicieron 2 rotaciones, es decir, cada alumno atendió a 2 pacientes simulados, dando un total de 8 interacciones (Figura 5.1).



(a) Inducción a los participantes

(b) Vista de administrador de la actividad en progreso

Figura 5.1: Piloto STAY

La actividad tuvo una buena recepción por parte de los participantes, aunque se consideraron mejoras menores a la usabilidad de la plataforma que no fueron vistas anteriormente. El principal inconveniente fue que en varias ocasiones los estudiantes olvidaron de guardar las respuestas del registro clínico electrónico hasta después de que el plazo de entrega había finalizado.
Esto se corrigió para la siguiente actividad, dando mayor visibilidad a este plazo y con mensajes de recordatorio.

Para la 2da actividad se invitó a estudiantes voluntarios de 5to y 6to año de Medicina, donde participaron un total de 7 estudiantes y 12 pacientes simulados, en un total de 14 encuentros clínicos. Esta actividad también tuvo una buena recepción y los alumnos no tuvieron problemas en desarrollar extensas respuestas en el registro clínico electrónico dentro del plazo.

(a) Mensaje recordatorio

(b) Invitación a estudiantes de pregrado

Figura 5.2: Prepación STAY

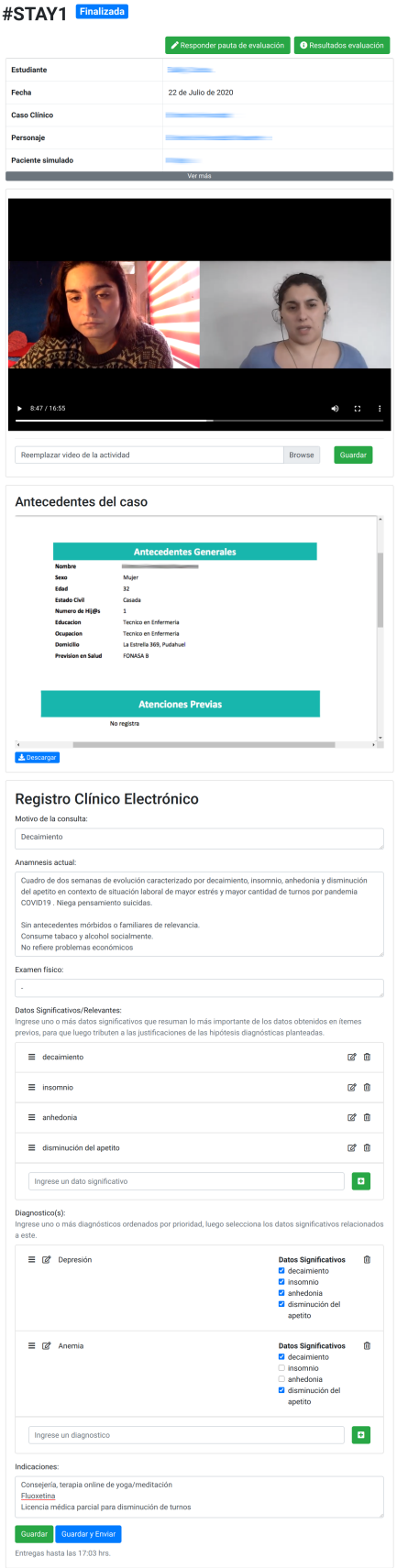


Figura 5.3: STAY: Ejemplo de una interaccíon finalizada

5.1.2. Encuesta de usabilidad

Se validó la plataforma utilizando la escala SUS (System Usability Scale), la cual consiste en una encuesta de usabilidad donde se le asigna un puntaje entre 1 y 5 a cada pregunta.
Para interpretar los resultados, se evalúan por separado los puntajes de preguntas pares e impares, ya que las preguntas pares señalan una apreciación negativa de la usabilidad de la plataforma, mientras que las impares señalan una apreciación positiva.
En el caso de las preguntas pares, el valor asignado es 5 menos el puntaje de la pregunta, en el caso de las impares simplemente se le resta 1 a cada pregunta, el resultado de ambas se suma y se multiplica por 2.5, este resultado final que puede ir de 0 a 100 representa la usabilidad de la plataforma o aplicación.

Actividad	**Rol de Usuario**	**Puntuación**
Piloto STAY	Administrador	95
Piloto STAY	Administrador	93
Piloto STAY	Administrador	75
Piloto STAY	Alumno (PS)	95
Piloto STAY	Alumno (PS)	65
Piloto STAY	Docente	73
Piloto STAY	Paciente Simulado	93
Piloto STAY	Paciente Simulado	93
Piloto STAY	Paciente Simulado	90
Piloto STAY	Paciente Simulado	85
STAY	Alumno	93
STAY	Alumno	90
STAY	Paciente Simulado	100
STAY	Paciente Simulado	100
STAY	Paciente Simulado	98
STAY	Paciente Simulado	95
STAY	Paciente Simulado	78

Tabla 5.1: Resultados de la encuesta SUS

La tabla completa de preguntas y respuestas con su respectiva escala y puntajes se puede consultar en el Anexo C.

La pregunta impar con menor puntaje promedio obtenido fue "*Imagino que la mayoría de la gente aprendería a usar esta plataforma en forma muy rápida*"(P7), mientras que la pregunta par con mayor puntaje promedio fue "*Creo que necesitaría ayuda de una persona con conocimientos técnicos para usar esta plataforma*"(P4). Se puede inferir que los usuarios sienten que la mayor deficiencia en la usabilidad del sitio está en la falta de interfaces intuitivas y falta de instrucciones integradas en la plataforma, lo cual no permitirían a nuevos usuarios utilizarla apropiadamente sin las instrucciones dadas por los miembros del CHC.
Un resultado mayor a 68 puede ser considerado mejor que la media, y considerando que la puntuación promedio de las encuestas fue de 88.9 y que solo 1 usuario de los 17 arrojó una puntuación peor que la media, se considera que se la plataforma ofrece una buena usabilidad a sus usuarios.

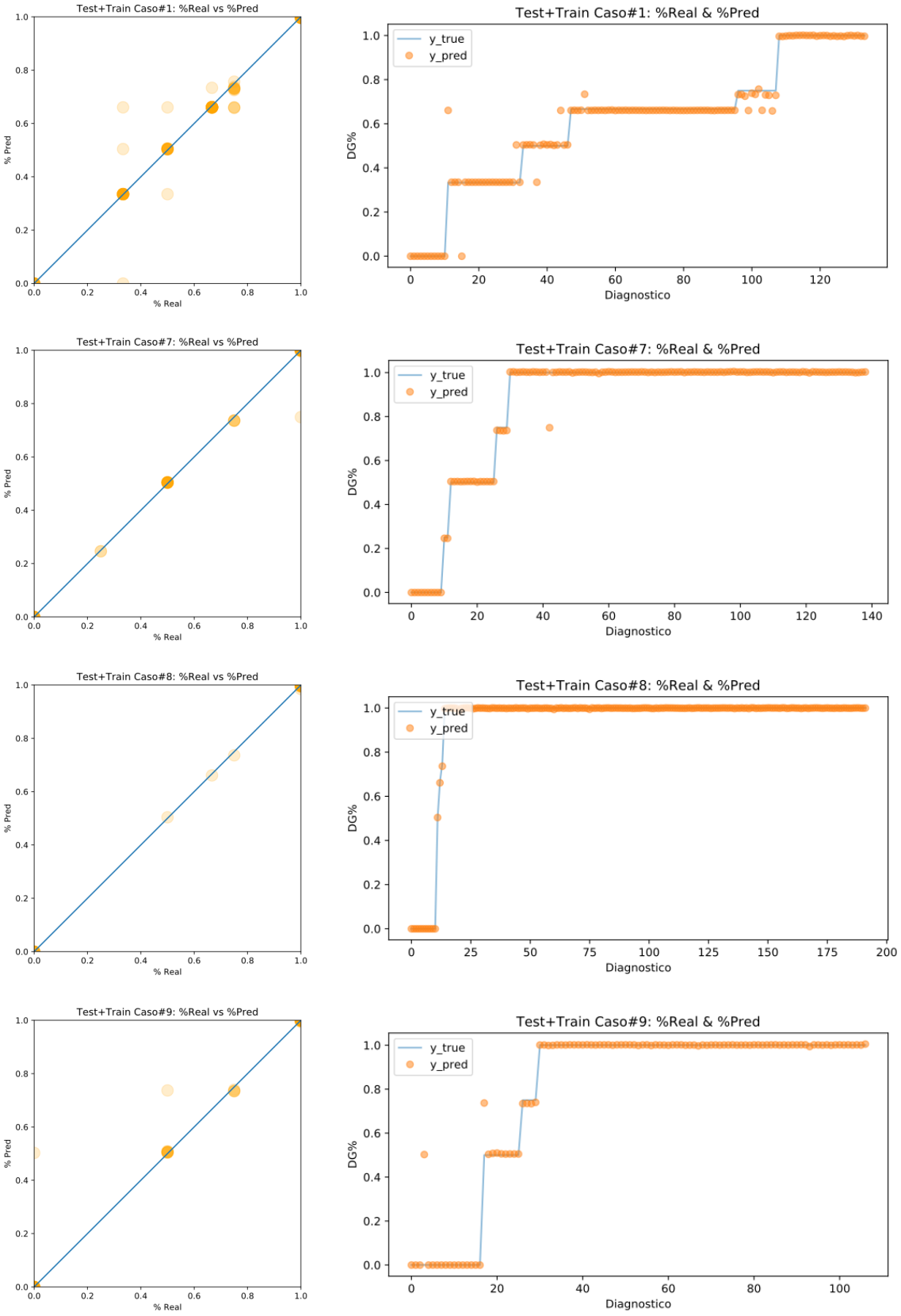
5.2. Clasificación automática del rendimiento de los estudiantes

Las predicciones de los diagnósticos dieron resultados bastante prometedores para algunos de los casos clínicos, aunque no para otros. Con el 10 % de las respuestas para validación, se llegó a un `MAE: 0.9` y `MSE: 0.027`.

Caso	**#**	**<5 %**	**<10 %**	**<20 %**	**<30 %**
(8) Hipotiroidismo primario	22	100.0 %	100.0 %	100.0 %	100.0 %
(7) Cetoacidosis diabética	13	92.3 %	92.3 %	92.3 %	100.0 %
(9) Lumbago mecánico	14	85.7 %	85.7 %	85.7 %	92.9 %
(1) EPOC	10	50.0 %	80.0 %	90.0 %	90.0 %
(6) SCA c/s SDST	16	75.0 %	75.0 %	87.5 %	87.5 %
(4) Neumonía adquirida ...	13	61.5 %	69.2 %	84.6 %	92.3 %
(15) Diarrea disenterica ...	22	68.2 %	68.2 %	77.3 %	81.8 %
(14) Crisis hipertensiva	31	64.5 %	67.7 %	87.1 %	90.3 %
(5) Sepsis de foco urinario	12	66.7 %	66.7 %	75.0 %	83.3 %
(13) Várices esofágicas ...	14	64.3 %	64.3 %	71.4 %	85.7 %
(3) Asma – Rinitis alérgica	19	63.2 %	63.2 %	94.7 %	94.7 %
(12) Enfermedad renal crónica	30	40.0 %	56.7 %	73.3 %	86.7 %
(2) Neumonía ...	14	35.7 %	50.0 %	78.6 %	78.6 %
(11) Anemia ferropenica	21	47.6 %	47.6 %	61.9 %	61.9 %
(10) Síndrome metabólico	10	40.0 %	40.0 %	90.0 %	90.0 %

Tabla 5.2: Porcentaje de predicciones que tienen una diferencia con el valor real menor a 5 %, 10 %, 20 % y 30 % (Respuestas de validación)

Se obtuvieron buenos resultados para los casos clínicos *#8, #7, #9 y #1* en distintas pruebas, ya que en estos casos no existen mayores ambigüedades ni variantes en las respuestas que limiten la capacidad de predicción. En estos casos el error de predicción es menor al `80 %` considerando una diferencia `<10 %` como aceptable (ver tabla 5.2). Esta consistencia en las respuestas de los alumnos para estos casos clínicos se puede observar en la tabla B.1 en los anexos, exceptuando el caso #7 donde habían diferencias que no permitieron hacer la agrupación.



(a) Real vs predicción

(b) Real y predicción

Figura 5.4: Evaluación de los casos clínicos *#1, #7, #8 y #9* con todas las respuestas del dataset

Sin embargo, no es posible confiar en los resultados para todos los casos clínicos, debido a la distribución de notas en las respuestas disponibles y porque las pautas de algunos casos clínicos pueden tener variaciones en diferentes actividades, y por lo tanto hay información que no está explícita en el dataset que es importante para poder hacer las evaluaciones y predicciones correctamente.
Estas variaciones se hicieron evidentes cuando se revisaron las respuestas que arrojaban las peores predicciones con los miembros del CHC y se observaron respuestas donde un caso clínico se veía vinculado a otro. Un ejemplo de esto es el caso clínico #6: *SCA c/s SDST* o *Síndrome Coronario Agudo con/sin supradesnivel del ST*: pese a que se observa un bajo error en las predicciones de este caso clínico, la distinción entre tener o no *SDST* es importante para la corrección pero no está explícita en el dataset. Por este motivo, no es posible que la red neuronal pueda predecir distintamente para ambos casos con los datos disponibles sin antes asignar distintos códigos para ambos casos.

5.3. Análisis automático de los videos

Se esperaba que se generaran alrededor de 360 horas de video cada semana (30 horas de video por semana en cada uno de los 12 box) pero ya que no se realizaron actividades presenciales durante el semestre, el entrenamiento y validación solo se pudo hacer con los 9 videos completos que se encontraban cargados en el servidor.

Con un **Accuracy:** `0.98` y **F1-score:** `0.98` en promedio (**F1-score:** `0.88` promediando el score de cada clase) y observando la matriz de confusión (Figura 4.10), consideramos que tenemos buenas predicciones excepto por la clasificación de la clase `washing_hands` (`F1-score: 0.63`), ya que es clasificada erróneamente con mayor frecuencia especialmente por la clase `waiting`. Esto puede explicarse dado que la clase `washing_hands` tiene muy pocos ejemplos de entrenamiento y validación ya que usualmente los alumnos se lavan las manos en menos de 30 segundos.

Con estas predicciones se crearon las marcas temporales de los videos, las cuales fueron revisadas y aprobadas por el CHC, permitiéndonos concluir que es perfectamente factible usar esta estrategia para clasificar los fotogramas de los videos. Ya que no se hicieron objeciones con las marcas temporales predichas, se omitió el cálculo de métricas de intersección de conjuntos.

Pese a estos resultados y que se hiciera la integración de estas marcas temporales en la base de datos y la interfaz de la plataforma (reproductor de video y datos históricos), se detuvo el desarrollo de esta funcionalidad ya que no sería de utilidad en el tiempo próximo y se dio prioridad al desarrollo de la integración de las actividades por videoconferencia.

Capítulo 6

Conclusiones y trabajo futuro

En el presente trabajo de título, se describe las mejoras realizadas en la plataforma Kamaleon con el objetivo de incrementar la retroalimentación de los estudiantes en carreras de la salud y disminuir las tareas de gestión y evaluaciones repetitivas asociadas.

Con todo lo descrito a lo largo de este informe, es posible notar que la principal ventaja que entregan las mejoras realizadas sobre la plataforma fue la unificación del ingreso de datos asociados a las actividades realizadas y la gestión de las actividades en tiempo real, con énfasis en la adaptación de la plataforma para la gestión de actividades por videoconferencia.

Estas mejoras en las interfaces y en el flujo de la información, que pudieron ser validadas adecuadamente vía una encuesta de usabilidad SUS (con un puntaje promedio de 88.9 %), permitirán que la plataforma pueda ser utilizada incluso cuando los estudiantes no puedan estar físicamente en el centro, lo que será de gran utilidad ante la incertidumbre causada por el COVID-19 para los miembros del CHC como para los estudiantes del área de salud, ya que permitirá dar una preparación realista a los estudiantes ante el escenario actual.

No se pudo concretar todo el desarrollo planificado de las mejoras, donde se evaluaría el funcionamiento de la plataforma constantemente usando las múltiples actividades que se realizan en el centro durante el semestre. Además, estas actividades permitirían almacenar datos correspondientes a videos de las interacciones y respuestas de los estudiantes en el registro clínico electrónico.
Pese a esto, el trabajo realizado demuestra la factibilidad de evaluar diagnósticos de forma automática, siempre y cuando se identifiquen las variantes de los casos clínicos. De igual forma, es factible evaluar los comportamientos en los videos dados los resultados obtenidos.

La plataforma permite estandarizar los distintos casos clínicos ya que evidencia las ambigüedades, y junto a más respuestas estandarizadas del registro clínico electrónico, se espera que en un futuro se pueda utilizar el trabajo desarrollado para clasificar el rendimiento de los estudiantes y crear mejores predicciones. De igual forma, se espera que en un futuro se guarden más vídeos correspondientes a las actividades presenciales en el centro, lo cual permitiría utilizar el trabajo desarrollado para la clasificación en los videos y crear mejores predicciones.

Bibliografía

[1] Manuel Brhel, Hendrik Meth, Alexander Maedche, and Karl Werder. Exploring principles of user-centered agile software development: A literature review. *Information and Software Technology*, 61:163–181, May 2015. doi: 10.1016/j.infsof.2015.01.004. URL http://dx.doi.org/10.1016/j.infsof.2015.01.004.

[2] Charu C. Aggarwal and ChengXiang Zhai. *A Survey of Text Classification Algorithms*, pages 163–222. Springer US, Boston, MA, 2012. ISBN 978-1-4614-3223-4. doi: 10.1007/978-1-4614-3223-4_6. URL https://doi.org/10.1007/978-1-4614-3223-4_6.

[3] Danilo Dessì, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. TF-IDF vs Word Embeddings for Morbidity Identification in Clinical Notes: An Initial Study. In *CEUR Workshop Proceedings*. CEUR-WS, Apr 2020. URL http://ceur-ws.org/Vol-2596/paper1.pdf.

[4] Ivette Motola, Luke A. Devine, Hyun Soo Chung, John E. Sullivan, and S. Barry Issenberg. Simulation in healthcare education: A best evidence practical guide. AMEE Guide No. 82. *Medical Teacher*, 35(10):e1511–e1530, 2013. doi: 10.3109/0142159X.2013.818632. URL https://doi.org/10.3109/0142159X.2013.818632.

[5] Jkří Ranňik. *Software engineering: the current practice*. Chapman & Hall, 2012.

[6] Sharon Decker, Mary Fey, Stephanie Sideras, Sandra Caballero, Leland Rockstraw, Teri Boese, Ashley E. Franklin, Donna Gloe, Lori Lioce, and Carol R. Sando. Standards of Best Practice: Simulation Standard VI: The Debriefing Process. *Clinical Simulation in Nursing*, 9(6):S26–S29, Jun 2013. doi: 10.1016/j.ecns.2013.04.008. URL http://dx.doi.org/10.1016/j.ecns.2013.04.008.

[7] J. Zeng, J. Ge, Y. Zhou, Y. Feng, C. Li, Z. Li, and B. Luo. Statutes Recommendation Based on Text Similarity. In *2017 14th Web Information Systems and Applications Conference (WISA)*, pages 201–204, Nov 2017. doi: 10.1109/WISA.2017.52. URL https://doi.org/10.1109/WISA.2017.52.

Anexo A

ECliPSE

El ECliPSE es una metodología de simulación basada en los principios de objetividad y estandarización, desarrollada en el CHC a partir del modelo de ECOE y el modelo de Observación de Examen de Caso Largo con paciente real. Corresponde a un modelo de simulación de alta complejidad, buscando recrear de la manera más fidedigna posible el rol de médico que deberá asumir el estudiante en cada uno de los encuentros clínicos con pacientes estandarizados que tendrá oportunidad de atender en el transcurso de su formación.

El propósito de la actividad es observar, analizar, discutir, estimular la auto evaluación y retroalimentar el encuentro clínico del estudiante con sus pacientes en un entorno simulado, ayudándose de herramientas evaluativas como rúbricas y pautas estandarizadas tanto de inducción, desarrollo, retroalimentación y debriefing. En este modelo de encuentros clínicos, el paciente simulado-estandarizado rota por los boxes de atención, tal como lo haría un paciente real en atención abierta de medicina.
En el CHC se dispone de 12 boxes para escenarios simultáneos, con visión y registro de audio-video. Cada estudiante tiene dos pacientes en agenda para atender sucesivamente. Usualmente los tiempos destinados son:

- **Atención de paciente (entrevista, examen, consejería):** 25 minutos
- **Registro clínico electrónico:** 10 minutos
- **Retroalimentación del PS:** 5 minutos
- **Total por escenario (encuentro clínico con PS):** 40 minutos
- **Total actividad (dos escenarios):** 1 hora y 20 minutos

Anexo B

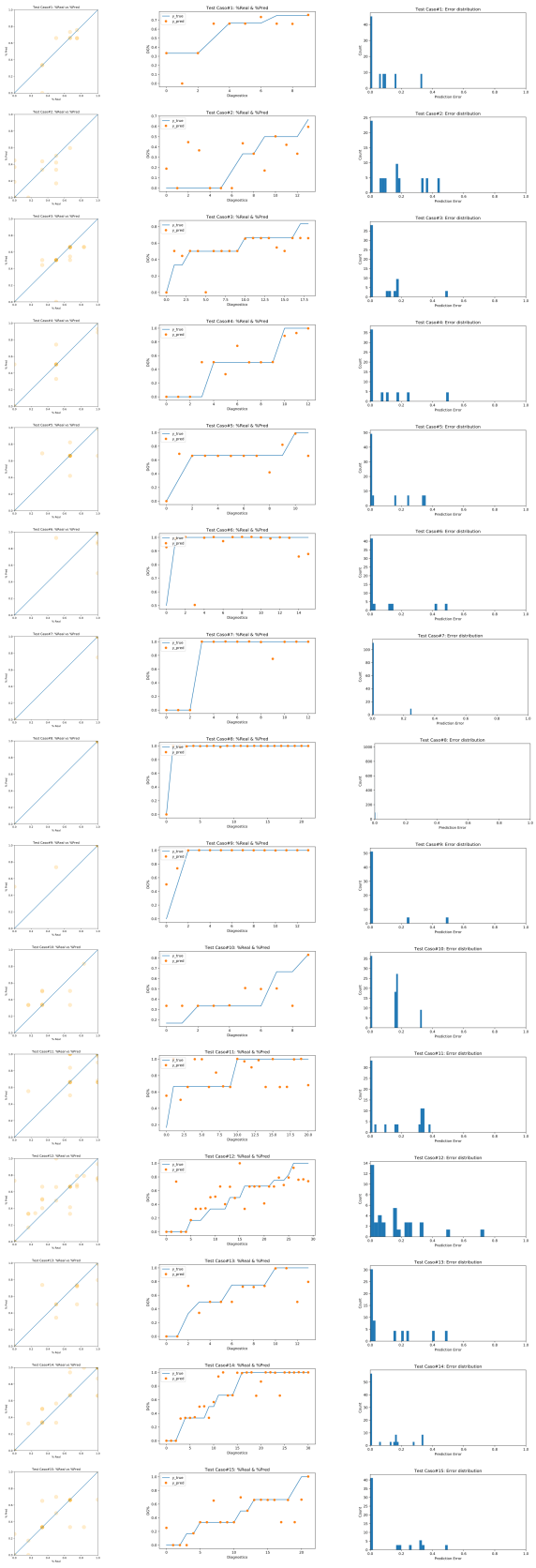
Diagnósticos

Caso	Nota	Diagnóstico
1	33 %	epoc | asma |
		epoc | cancer pulmonar |
		epoc | obesidad |
	50 %	epoc |
	67 %	epoc | tabaquismo activo | obesidad |
		epoc gold 1 grupo d | tabaquismo activo |
		epoc | tabaquismo activo |
		epoc | tabquismo activo |
		epoc | tabaquismo cronico |
		epoc gold a | tabaquismo |
		epoc | tabaquismo |
	75 %	epoc gold 2 |
		epoc gold 2 b |
		epoc moderado |
		epoc | tabaquismo |
	100 %	epoc b | tabaquismo activo |
2	50 %	neumonia por pneumocystis jiroveci |
3	0 %	asma |
	33 %	asma |
	50 %	rinitis alergica | asma bronquial |
		rinitis alergica | asma |
		rinitis alergica | obs asma |
		asma alergica rinitis alergica |
		asma mal controlada | rinitis alergica |
		asma | rinitis alergica |
		asma mal controlado |
		asma no controlada |
	67 %	asma alergica no controlada | rinitis alergica |
		asma alergica | rinitis alergica |
		asma alergico | rinitis alergica |
		asma alergico | rinitis alergica | |

		asma atopica | rinitis alergica |
		asma con mal control | rinitis alergica |
		asma mal controlada | rinitis alergica |
		asma no controlada | rinitis alergica |
	83 %	asma alergica no controlada | rinitis alergica |
4	50 %	neumonia adquirida en la comunidad |
		nac leve |
		nac |
	100 %	neumonia adquirida en la comunidad | clase 1 |
		nac i |
5	67 %	pielonefritis aguda | sd febril |
		pielonefritis aguda |
		itu alta no complicada |
		pna no complicada |
		pna |
	100 %	sepsis de foco urinario obs pna |
6	100 %	iam con sdst |
8	100 %	hipotiroidismo clinico | anemia normo | normo |
		hipotiroidismo clinico |
		hipotiroidismo | obs depresion |
		hipotiroidismo | dislipidemia |
		hipotiroidismo |
		hipotiroidismo primario |
9	0 %	lumbociatica |
	100 %	lumbago mecanico | tabaquismo activo |
		lumbago mecanico agudo |
		1 | lumbago mecanico | sin banderas rojas |
		lumbago mecanico sin banderas rojas |
		lumbago mecanico |
		lumbago mecanico | policonsumo |
10	33 %	sd metabolico | esteatosis hepatica |
		sd metabolico |
11	67 %	anemia ferropenica | obs hemorragia digestiva alta |
		anemia ferropenica | intolerancia a sulfato ferroso |
	100 %	anemia ferropenica |
		anemia ferropenica | hipermenorrea |
		1 | anemia ferropenica moderada |
		anemia ferropenica moderada |
		anemia ferropenica | obs metrorragia |
		anemia ferropriva moderada | metrorragia |
		anemia | microcitica | hipocromica | ferropenica |
12	25 %	hipertension arterial | dislipidemia | tabaquismo activo |
		hipertension arterial | dislipidemia |

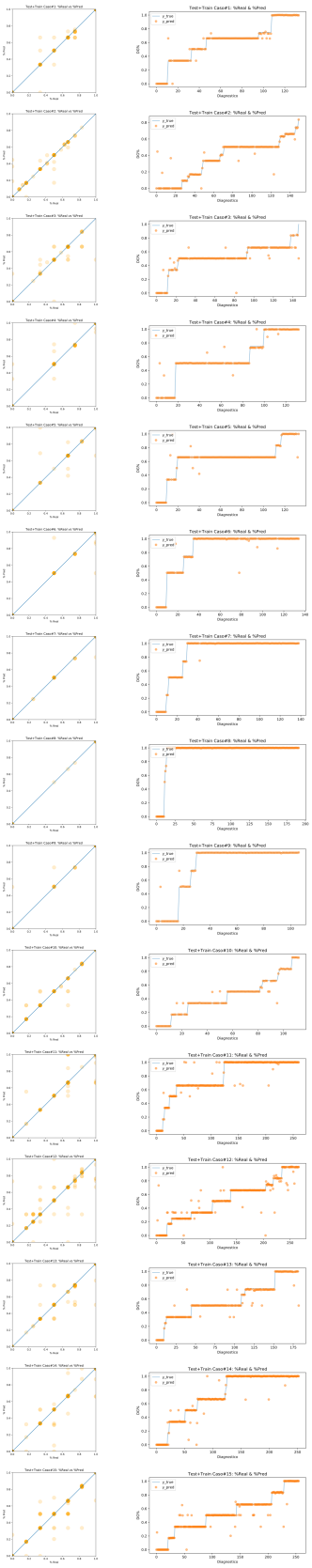
		hta en tratamiento | dislipidemia en tratamiento |
		dislipidemia | hta |
		hta | dlp | sobrepeso |
	67 %	sd metabolico | tabaquismo |
	100 %	sd metabolico | hta | dlp | tabaquismo |
13	50 %	daño hepatico cronico | varices esofagicas |
		daño hepatico cronico |
		dhc en estudio |
	75 %	daño hepatico cronico | varices esofagicas pequeñas |
		daño hepatico cronico | varices esofagicas |
		dhc de etiologia no precisada | varices esofagicas pequeñas |
14	67 %	emergencia hipertensiva | hipertension arterial |
		emergencia hipertensiva | hemorragia subaracnoidea |
	100 %	emergencia hipertensiva acv ¿hemorragico? |
		emergencia hipertensiva | cefalea con signos de alarma |
		emergencia hipertensiva obs encefalopatia hipertensiva vs ave hemorragico |
		emergencia hipertensiva | cefalea con banderas rojas |
		cefalea con banderas rojas obs hemorragia subaracnoidea emergencia hta |
		crisis hipertensiva | obs emergencia hipertensiva |
		emergencia hipertensiva (encefalopatia hipertensiva) |
		emergencia hipertensiva | hemorragia subaracnoidea |
15	33 %	diarrea disenterica en vias de prolongacion | deshidratacion leve |
		disenteria | diarrea en vias de prolongacion |
		diarrea en vias de prolongacion | sd disenterico |
		sd diarreico en vias de prolongacion | obs sd disenterico |
		sd diarreico en vias de prolongacion | sd disenterico |
	50 %	disenteria | diarrea en vias de prolongacion |
		diarrea en vias de prolongacion | disenteria |
	67 %	diarrea disenterica en vias de prolongacion | deshidratacion leve |
		diarrea en vias de prolongacion | disenteria |
		sd disenterico | diarrea en vias de prolongacion |
		sd diarreico prolongado | disenteria |
		sd disenterico | obs enfermedad inflamatoria intestinal |

Tabla B.1: Diagnósticos agrupados por la frecuencia de palabras, el caracter \| representa la separación de distintas respuestas. Se reemplazaron caracteres especiales y mayúsculas. Solo se incluyen las respuestas que aparecen más de una vez.



(a) Real y predicción (b) Real vs predicción (c) Distribución del error

Figura B.1: Evaluación por caso clínico (261 respuestas de validación)



(a) Real vs predicción

(b) Real y predicción

Figura B.2: Evaluación por caso clínico (todas las respuestas del dataset)

Anexo C

Escala de usabilidad de sistemas

La escala SUS (System Usability Scale) consiste en una encuesta de usabilidad donde se le asigna un puntaje entre 1 a 5 a cada una de las siguientes preguntas:

1. Creo que usaría esta plataforma frecuentemente
2. Encuentro esta plataforma innecesariamente compleja
3. Creo que la plataforma fue fácil de usar
4. Creo que necesitaría ayuda de una persona con conocimientos técnicos para usar esta plataforma
5. Las funciones de esta plataforma están bien integradas
6. Creo que la plataforma es muy inconsistente
7. Imagino que la mayoría de la gente aprendería a usar esta plataforma en forma muy rápida
8. Encuentro que la plataforma es muy difícil de usar
9. Me siento confiado al usar esta plataforma
10. Necesité aprender muchas cosas antes de ser capaz de usar esta plataforma

Actividad	**Rol de Usuario**	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**Puntuación**
Piloto STAY	Administrador	5	1	5	1	4	1	4	1	5	1	95
Piloto STAY	Administrador	5	1	4	2	4	1	5	1	5	1	93
Piloto STAY	Administrador	3	1	5	1	3	3	3	1	3	1	75
Piloto STAY	Alumno	4	1	5	1	4	1	5	1	5	1	95
Piloto STAY	Alumno	4	2	3	4	4	2	4	2	4	3	65
Piloto STAY	Docente	5	2	4	4	4	2	5	2	4	3	73
Piloto STAY	Paciente Simulado	5	1	5	3	5	1	5	1	4	1	93
Piloto STAY	Paciente Simulado	5	1	5	2	5	1	4	1	4	1	93
Piloto STAY	Paciente Simulado	5	1	5	3	5	1	4	1	5	2	90
Piloto STAY	Paciente Simulado	5	1	5	4	5	1	4	1	4	2	85
STAY	Alumno	5	1	5	1	5	1	4	1	3	1	93
STAY	Alumno	5	1	5	2	4	1	5	1	4	2	90
STAY	Paciente Simulado	5	1	5	1	5	1	5	1	5	1	100
STAY	Paciente Simulado	5	1	5	1	5	1	5	1	5	1	100
STAY	Paciente Simulado	5	1	5	1	5	1	4	1	5	1	98
STAY	Paciente Simulado	5	1	5	1	5	1	3	1	5	1	95
STAY	Paciente Simulado	4	5	5	1	4	3	4	1	5	1	78
Promedio:		4,71	1,35	4,76	1,94	4,47	1,35	4,29	1,12	4,41	1,41	88,88

Tabla C.1: Tabla SUS completa, 5 representa muy de acuerdo y 1 representa completamente en desacuerdo