

MODELO EN MACHINE LEARNING PARA EL DIAGNOSTICO DEL CÁNCER DE MAMA

Jorge Armando Millán Gomez, Jaime Brandon Robles Fajardo

3 de junio de 2020



MODELO EN MACHINE LEARNING PARA EL DIAGNOSTICO DEL CÁNCER DE MAMA

Jorge Millán, Brandon Robles

Proyecto de Investigación presentado para el grado
de Especialista en Ingeniería de Software

3 de junio de 2020

Índice general

I	PROYECTO DE INVESTIGACIÓN	11
1.	Introducción	12
2.	Metodología de la Investigación	13
2.1.	Identificación del Problema	14
2.1.1.	Planteamiento del Problema	14
2.1.2.	Formulación del Problema	14
2.2.	Objetivos	15
2.2.1.	Objetivo General	15
2.2.2.	Objetivos Específicos	15
2.3.	Justificación de la investigación	16
2.4.	Hipótesis	16
2.5.	Marco Referencial	17
2.5.1.	Marco Teórico	17
2.5.2.	Marco Conceptual	25
2.6.	Metodología de la investigación	28
2.6.1.	Tipo de Estudio	28
2.6.2.	Método de Investigación	28
2.6.3.	Fuentes y Técnicas para la recolección de Información	28
2.7.	Organización del Trabajo de Grado	29
3.	Desarrollo de la Investigación	30

3.1. Recolección de datos	31
3.2. Verificación de datos	32
3.3. Clasificación de datos	33
3.4. Limpieza de datos	36
3.4.1. Importación de datos:	36
3.4.2. Formateo de datos:	37
3.4.3. Comprobación de datos faltantes:	38
3.4.4. Eliminación de datos Innecesarios:	39
3.5. Procesamiento de los Datos	40
3.6. Modelos de Machine Learning	45
3.6.1. Logistic Regression	47
3.6.2. Decision Trees	49
3.6.3. Random Forest	51
3.6.4. Gaussian Naive Bayes	53
3.6.5. K-Nearest Neighbors(KNN)	55
3.6.6. Support Vector Machines(SVM)	58
3.7. BreastApp	60
3.7.1. Back-End BreastApp	60
3.7.2. Front-End BreastApp	79
 II ARQUITECTURA Y DISEÑO	 87
 4. Organización Empresarial	 88
4.1. Misión	88
4.2. Visión	88
4.3. Objetivo	89
4.4. Actores, Roles y Funciones	90
4.5. Servicios	91
 5. Capa de Motivación	 92

5.1. Introducción	92
5.2. Punto de Vista de Interesados	93
5.3. Punto de Vista de Realización de Objetivos	95
5.4. Punto de Vista de Contribución de Objetivos	97
5.5. Punto de Vista de Principios	99
5.6. Punto de Vista de Realización de Requerimientos	100
5.7. Punto de Vista de Motivación	102
6. Capa de Estrategia	104
6.1. Introducción	104
6.2. Punto de Vista de la Estrategia	105
6.3. Punto de Vista del Mapa de Capacidad	107
6.4. Punto de Vista de la Realización de Resultado	109
6.5. Punto de Vista de Mapa de Recurso	111
7. Capa de Negocio	113
7.1. Introducción	113
7.2. Punto de Vista de la Organización	114
7.3. Punto de Vista de Cooperación de Actor	116
7.4. Punto de Vista de la Función de Negocio	118
7.5. Punto de Vista de Proceso de Negocio	120
7.6. Punto de Vista de Cooperación de Proceso de Negocio	123
7.7. Punto de Vista de Producto	125
8. Capa de Aplicación	128
8.1. Introducción	128
8.2. Punto de Vista del Comportamiento de la Aplicación	129
8.3. Punto de Vista de Cooperación de Aplicación	132
8.4. Punto de Vista de la Estructura de la Aplicación	134
8.5. Punto de Vista del Uso de la Aplicación	136

9. Capa de Tecnología	137
9.1. Introducción	137
9.2. Punto de Vista de la Infraestructura	138
9.3. Punto de Vista del Uso de la Infraestructura	140
9.4. Punto de Vista de Implementación y Organización	142
9.5. Punto de Vista de Estructura de la Información	144
9.6. Punto de Vista de Realización del Servicio	146
9.7. Punto de Vista de Capas	148
10. Capa de Migración e Implementación	151
10.1. Introducción	151
10.2. Punto de Vista de Proyecto	152
10.3. Punto de Vista de Migración	155
10.4. Punto de Vista de Migración e implementación	156
III REFLEXIONES	159
11. Resultados y Conclusiones	160
11.1. Resultados	160
11.2. Conclusiones	162
11.3. Aportes Originales	162
11.4. Trabajos Futuros	163

Índice de figuras

3.1. Muestra FNA capturada con Microscopio y procesada con el Software Xcyt	31
3.2. Medición de la lisura de un contorno nuclear	33
3.3. Medición de concavidad del núcleo celular	34
3.4. Medición de la simetría de un núcleo celular	34
3.5. Medición de la dimensión Fractal de un núcleo celular	35
3.6. Comparación Estadística casos Benignos vs Malignos	41
3.7. Gráfico de correlación cruzada de variables Oncológicas	42
3.8. Diagrama de Calor de la correlación de variables Oncológicas	43
3.9. Clasificación realizada por el método Logistic Regression	48
3.10. Clasificación realizada por el método Decision Trees	50
3.11. Clasificación realizada por el método Random Forest	52
3.12. Clasificación realizada por el método Gaussian Naive Bayes	54
3.13. Clasificación realizada por el método K-Nearest Neighbors(KNN)	57
3.14. Clasificación realizada por el método Support Vector Machines(SVM)	59
3.15. Vista para cargar Data-Sets para entrenamiento	79
3.16. Vista de Data-Sets cargados	80
3.17. Vista entrenamiento realizado correctamente	80
3.18. Vista historial de entrenamientos	81
3.19. Vista carga de datos de pacientes a diagnosticar	82
3.20. Vista de datos cargados de pacientes a diagnosticar	82
3.21. Vista de diagnostico realizado correctamente	83

3.22. Vista modelos de Machine Learning utilizados en el diagnostico	84
3.23. Vista diagnostico detallado del paciente	85
3.24. Reporte detallado en pdf del diagnostico del paciente	86
4.1. Jerarquía del Área de Investigación de análisis de datos	90
5.1. Punto de Vista de Interesados	93
5.2. Punto de Vista de Realización de Objetivos	95
5.3. Punto de Vista de Contribución de Objetivos	97
5.4. Punto de Vista de Principios	99
5.5. Punto de Vista de Realización de Requerimientos	100
5.6. Punto de Vista de Motivación	102
6.1. Punto de Vista de la Estrategia	105
6.2. Punto de Vista del Mapa de Capacidad	107
6.3. Punto de Vista de la Realización de Resultado	109
6.4. Punto de Vista de Mapa de Recurso	111
7.1. Punto de Vista de la Organización	114
7.2. Punto de Vista de Cooperación de Actor	116
7.3. Punto de Vista de la Función del negocio	118
7.4. Punto de Vista del proceso del negocio	120
7.5. Punto de Vista de Cooperación de Proceso de Negocio	123
7.6. Punto de Vista de Producto	125
8.1. Punto de Vista del Comportamiento de la Aplicación	129
8.2. Punto de Vista de Cooperación de la Aplicación	132
8.3. Punto de Vista de la Estructura de la Aplicación	134
8.4. Punto de Vista del uso de la Aplicación	136
9.1. Punto de Vista de la Infraestructura	138
9.2. Punto de Vista del Uso de la Infraestructura	140

9.3. Punto de Vista de Implementación y Organización	142
9.4. Punto de Vista de Estructura de Información	144
9.5. Punto de Vista de Realización de Servicio	146
9.6. Punto de Vista de Capas	148
10.1. Punto de Vista de Proyecto	152
10.2. Punto de Vista de Migración	155
10.3. Punto de Vista de Migración e Implementación	156
11.1. Uso de modelos enfocados en el diagnostico de cáncer de mama	160

Índice de tablas

3.1. Diseño servicio REST Data-Set	60
3.2. Diseño servicio REST ModelMl	63
3.3. Diseño servicio REST Prediction	66
3.4. Diseño servicio REST Training	70
3.5. Diseño servicio REST Patient	73
3.6. Diseño servicio REST Patientprediction	76
11.1. Precisión de los Modelos de Machine Learning seleccionados	161

Parte I

PROYECTO DE INVESTIGACIÓN

Capítulo 1

Introducción

El aprendizaje automático (Machine Learning) hace parte de una rama de la inteligencia artificial que se vale de un número importante de algoritmos desarrollados para darle la posibilidad a la máquina de aprender y responder a situaciones problemáticas, de las que la máquina ha sido entrenada, mediante un set de datos. Aplicaciones de machine learning abundan en la identificación de patrones y la clasificación de datos, las cuales han sido bien recibidos en la comunidad científica e industrial para la implementación de soluciones que van desde la detección de anomalías en procesos industriales, detección y predicción de fraude hasta el análisis de textos y la clasificación de los mismos. Este documento presenta el desarrollo de la implementación de un modelo Machine Learning para realizar la diagnosis de cáncer de mama, donde, de manera detallada, se mostrarán los antecedentes en cuanto al desarrollo de modelos existentes de Machine Learning en el área de la salud y especialmente para el diagnóstico de cáncer de mama.

Capítulo 2

Metodología de la Investigación

El número de casos de Cáncer en Colombia es muy alto, según el informe mundial de la salud del año 2018, 101.893 casos de diferentes tipos de cáncer fueron detectados de los cuales 46.057 casos resultaron en muerte[1]. El Cáncer de mama ocupa el primer lugar con mayor número de muertes para el género femenino. El pronóstico anticipado de un tipo de cáncer se ha convertido en una necesidad de investigación ya que puede facilitar el tratamiento preventivo para evitar su letalidad en un estado avanzado. Por lo tanto, surgió la idea de esta investigación con el objetivo de diagnosticar el posible padecimiento de cáncer de mama al realizar la comparación de los historiales clínicos de los pacientes con síntomas leves con el de pacientes con un estado de cáncer avanzado.

Para el año 2019 en el ámbito tecnológico y de investigación el aprendizaje automático o Machine Learning (ML) que se encuentra dentro de la rama de inteligencia artificial proporciona herramientas y métodos que permiten analizar una gran cantidad de datos y por medio de una regresión lineal llegar a un resultado diciente. Estas técnicas han sido utilizadas por diferentes investigadores para modelar la progresión y el tratamiento de afecciones cancerosas debido a su capacidad para detectar características significativas en conjuntos de datos complejos.

En este trabajo se busca realizar una revisión de los diferentes Modelos utilizados en Machine-Learning para la detección de cáncer sugeridos por diferentes investigadores para realizar la implementación un modelo en ML para diagnosticar el posible padecimiento de cáncer de mama teniendo como base los historiales clínicos de pacientes con afecciones Malignas o Benignas.

2.1. Identificación del Problema

2.1.1. Planteamiento del Problema

Según el informe de la organización mundial de la salud del año 2018 los casos detectados de cáncer de mama en Colombia fueron 13.380 de los cuales 3.702 casos terminó en muerte ocupando el primer puesto de la tasa de letalidad sobre los demás tipos de cáncer[1]. Si no se tiene un diagnostico a tiempo que detecte los primeros síntomas que caracterizan el cáncer de mama es posible que la cifra de muertes en Colombia sea mucho mayor en los años posteriores. Una alternativa para disminuir esta tasa de mortandad es poder diagnosticar a tiempo, con base en el historial de los datos obtenidos de exámenes realizados por el individuo, que probabilidad de padecer cáncer de mama tiene, y según estas estadísticas realizar un tratamiento preventivo que permita combatir el cáncer antes de que el mismo haga metástasis o que llegue a un estado avanzado en donde sea más difícil de tratarlo. Una de las disciplinas científicas del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente es el Machine Learning con el cual es posible por medio del análisis de datos comparar y diagnosticar por medio de una regresión lineal la tendencia del comportamiento de dichos datos para llegar a un resultado probabilístico.

2.1.2. Formulación del Problema

¿Tiene el cáncer de mama un patrón característico en cada individuo que al ser convertido en un Dataset permite detectarlo y según el cambio y similitud de los datos diagnosticarlo?

2.2. Objetivos

2.2.1. Objetivo General

Desarrollar un sistema con modelos de Machine Learning para diagnosticar el padecimiento de cáncer de mama.

2.2.2. Objetivos Específicos

- Efectuar el entrenamiento de los modelos en Machine Learning que apliquen para el diagnóstico de cáncer de mama.
- Realizar un análisis comparativo de la precisión de los modelos existentes utilizados en el diagnóstico de cáncer de mama.
- Elaborar un aplicativo que sea capaz de diagnosticar el padecimiento de cáncer de mama de un paciente en particular.

2.3. Justificación de la investigación

La razón de ser de esta investigación es la de utilizar los distintos modelos algorítmicos basados en Machine Learning existentes para el diagnóstico temprano del cáncer de mama que permitan ampliar este campo de investigación y que sean un complemento en los diversos estudios que se tienen, esto con el objetivo de ayudar a reducir las muertes por cáncer de mama.

La implementación está enfocada principalmente en la diagnosis de cáncer de mama con base en la información recolectada mediante un conjunto de datos, que luego de ser analizados y clasificados por los diferentes modelos de Machine Learning, tendrán como resultado un prototipo con el que se podrán realizar análisis y posteriores tomas de decisiones sobre el caso en particular que se esté estudiando. El resultado más importante de esta investigación es la implementación de varios modelos existentes que ayudaran significativamente en el diagnóstico de cáncer de mama y de esta forma aportar en materia tecnológica al área de la oncología.

El resultado de la investigación además de implementar un modelo para diagnosticar el padecimiento de cáncer de mama va a permitir contrastar con los otros modelos sugeridos por diversos investigadores y podrá emplearse en otras investigaciones posteriores para obtener una aproximación cada vez más cercana para el tratamiento de los diversos tipos de cáncer que tienen como referencia el uso de las herramientas con base en Machine Learning.

2.4. Hipótesis

Al comparar grandes cantidades de datos que contienen información de resultados diagnósticos de pacientes particulares con los datos característicos de pacientes que padecen de cáncer de mama por medio de los modelos algorítmicos de Machine Learning, la similitud del comportamiento de los datos de los pacientes particulares con el patrón característico del paciente con cáncer de mama diagnostica de manera correcta el padecimiento de cáncer de mama de los pacientes particulares.

2.5. Marco Referencial

2.5.1. Marco Teórico

Diversos métodos basados en Machine learning para la predicción y diagnóstico de enfermedades han sido propuestos por una gran cantidad de investigadores. Las investigaciones analizadas brindan una gran cantidad de información y son un punto de partida importante para implementar un modelo en Machine Learning para el diagnóstico de cáncer de mama, los artículos seleccionados como referentes de la investigación se muestran a continuación:

El artículo *A Method to Select a Good Setting for the kNN Algorithm when Using it for Breast Cancer Prognosis*, propone aplicación del método k-Nearest Neighbours (kNN) para el pronóstico del cáncer de mama basado en las herramientas de Machine learning (ML) que selecciona la mejor configuración según los parámetros que se pueden cambiar al momento de utilizar la técnica de clasificación. Para comprobar este método se usaron los datos de pronóstico de cáncer de seno de Wisconsin, y el resultado generado arrojó una precisión promedio de 76 %.[2]

En el artículo *An Ensemble Model for Breast Cancer Prediction Using Sequential Least Squares Programming Method (SLSQP)* se propone un método con mayor precisión que los utilizados normalmente con el sistema tradicional de clasificación individual. Para comprobarlo el modelo utiliza cuatro técnicas de Machine learning (ML): Support Vector Machine, Logistic Regression, Decision Tree y k-Nearest Neighbours (kNN) en donde por medio del SLSQP se asigna un peso a cada modelo de clasificación y la predicción de cada clasificador se combina mediante la técnica de votación suave[3].

En el artículo *Application of Machine Learning in Disease Prediction* se utilizan diferentes algoritmos de clasificación en tres data set de enfermedades (corazón, cáncer de seno, diabetes) disponibles en el repositorio de UCI para la predicción de enfermedades. La selección de características para cada conjunto de datos se realizó mediante el modelado backward utilizando la prueba del valor-p. Los resultados del estudio fortalecen la idea de la aplicación del Machine Learning en la detección temprana de enfermedades[4].

En el artículo de opinión *Artificial intelligence in radiology*, se establece una comprensión general de los métodos de Inteligencia Artificial (IA), particularmente los relacionados con las tareas basadas en imágenes. Se explora cómo estos métodos podrían afectar múltiples facetas de la radiología, con un enfoque general en aplicaciones en oncología, y se demuestran formas en que estos métodos están avanzando en el campo. Finalmente, se discuten los desafíos que enfrenta la implementación clínica y se brinda una perspectiva sobre cómo se podría avanzar en esta rama[5].

En el artículo BAT-ELM: A Bio Inspired Model for Prediction of Breast Cancer Data se propone un modelo de inspiración biológica llamado BATELM, que es una combinación de algoritmo Bat (BAT) y Extreme Learning Machines (ELM), es una investigación poco tradicional debido a que el estudio del análisis de datos del cáncer de mama no se basa en imágenes diagnósticas. En él se hace uso de BAT para optimizar los parámetros de ELM para que la tarea de predicción se lleve a cabo de manera eficiente. Para lograr un error mínimo, los investigadores prueban el conjunto de datos de Pronóstico del cáncer de seno de Wisconsin (WBCP) en tres funciones de aprendizaje diferentes (sigmoide, sin y tanh) mostrando como resultado de la investigación la función que presenta una mayor precisión[6].

En el artículo Breast Cancer Detection using K-nearest Neighbor Machine Learning Algorithm, se propone un nuevo método para detectar el cáncer de seno con alta precisión. Este método consta de dos partes principales, en la primera parte, las técnicas de procesamiento de imágenes se utilizan para preparar las imágenes de mamografía para el proceso de extracción de características y patrones. Las características extraídas se utilizan como entrada para dos tipos de modelos de aprendizaje supervisado, que son el modelo de Red Neural de Propagación Posterior (BPNN) y el modelo de Regresión Logística (LR) con la comparación del resultado y la precisión de ambos modelos[7].

En el artículo Breast Cancer Diagnosis Using an Unsupervised Feature Extraction Algorithm Based on Deep Learning, se demuestra un nuevo esquema, que integra un algoritmo de extracción de características no supervisado basado en Deep Learning con los stacked auto-encoders, y un modelo de support vector machine (SAE-SVM), para el diagnóstico del cáncer de mama. Los stacked auto-encoders junto con el pre-entrenamiento greedy layerwise y un algoritmo con una momentum de actualización mejorado son aplicados para capturar la información esencial y extraer las características necesarias de los datos originales. Luego, se emplea un modelo de support vector machine para clasificar las muestras con nuevas características en tumores malignos o benignos. El método propuesto se probó en el conjunto de datos de cáncer de mama de diagnósticos de Wisconsin. El rendimiento se evalúa utilizando varias medidas y se compara con los resultados publicados anteriormente. Los resultados de la comparación muestran que el método SAE-SVM propuesto mejora la precisión al 98.25 % y supera a los otros métodos[8].

En el artículo Breast Cancer Diagnostic System using Symbiotic Adaptive Neuroevolution (SANE) se desarrolla un sistema inteligente híbrido para el diagnóstico, el pronóstico y la predicción del cáncer de mama utilizando SANE (Simbiótico, Neuroevolución Adaptativa) y se compara con el conjunto ANN, la red neuronal modular, la red neuronal evolutiva de arquitectura fija (F-ENN) y la red neuronal evolutiva de arquitectura variable (V-ENN). El sistema SANE co-evolucionará a una población de neuronas que cooperan para formar una red neuronal funcional. La base de datos de cáncer de mama de la Universidad de Wisconsin disponible en el repositorio de Aprendizaje Automático de UCI se utilizó

para realizar trabajos experimentales[9].

En el artículo Breast Cancer Prediction Based On Backpropagation Algorithm, se desarrolla un sistema que puede clasificar el tumor de un síntoma que causa la enfermedad de cáncer de mama utilizando el algoritmo de red neuronal Feedforward Backpropagation. El objetivo principal de la investigación es desarrollar sistemas más rentables y fáciles de usar para apoyar a los médicos. Para el problema de diagnóstico de tumor de cáncer de mama, los resultados experimentales muestran que los modelos concisos extraídos de la red logran una alta tasa de precisión en el conjunto de datos de entrenamiento y en el conjunto de datos de prueba. La base de datos de tumores de cáncer de seno utilizada para este propósito es del repositorio de Machine Learning de la Universidad de Wisconsin (UCI)[10].

En el artículo Breast Cancer Prediction using Feature Selection and Ensemble Voting se analiza el desempeño de modelos supervisados y no supervisados para la clasificación del cáncer de seno. Los Dataset de cáncer de seno de Wisconsin se utilizan en esta investigación. La selección de características se procesa a través del método Feature scaling y el análisis de componentes principales. Los resultados finales indican que el método de la votación de agrupamiento por conjuntos (Ensemble Voting Approach) es ideal como modelo predictivo para el cáncer de mama. El modelo de referencia se crea utilizando el método de Random Forest. Entre todos los modelos evaluados, solo cuatro modelos, es decir, Ensemble Voting Classifier, Logistics Regression, SVM Tuning y AdaBoost tuvieron una precisión de al menos 98 %. Basado en los resultados de la precisión, ROC-AUC, medida F1 y tiempo computacional de los modelos, el conjunto mostró el mayor potencial en la clasificación del cáncer de mama del conjunto de datos dado[11].

En el artículo Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach, se introduce una nueva técnica en el método de agrupamiento de datos de clasificación de promedio ponderado basado en GA que superó las limitaciones del método de promedio ponderado clásico. El método de promedio ponderado basado en algoritmos genéticos se utiliza para la predicción de múltiples modelos. La comparación entre la optimización de enjambre de partículas (PSO), la evolución diferencial (DE) y el algoritmo genético (GA) concluye que el algoritmo genético supera a los métodos promedio ponderados. Una comparación más entre el método de conjunto clásico y el método de promedio ponderado basado en GA concluye que el método de promedio ponderado basado en GA supera el rendimiento[12].

En el artículo Comparison of Machine Learning Classifiers for Breast Cancer Diagnosis Based on Feature Selection, se recopilaron datos de imágenes digitalizadas de un aspirado con aguja fina (FNA) de una masa mamaria. Además, Describen las características de los núcleos celulares presentados en la imagen. Este trabajo adopta varios métodos de selección de características para seleccionar las características más relacionadas para el diagnóstico de cáncer de seno. En función de las características seleccionadas, se construyen cuatro modelos de Machine Learning, Support Vector Machine (SVM), Decision Tree (DT),

AdaBoost y Random Forest (RF) y se evalúa su rendimiento. Los resultados experimentales muestran que la precisión de RF es mayor que los otros tres métodos[13].

En el artículo Deep Convolution Neural Network Model to Predict Relapse in Breast Cancer, se desarrolló un modelo de red neuronal multicapa para clasificar los datos genómicos multidimensionales en su grupo de anotación similar. Para lograrlo se utilizó un perceptrón genómico del cáncer multicapa para anotar genes expresados diferencialmente (DEG) para diagnosticar la recaída en función del estado del Receptor de Estrógenos (ER) en el cáncer de mama. Este enfoque proporciona una identificación multivariada de genes, no solo por expresión diferencial, sino también por la causa-efecto del estado de la enfermedad. El resultado de la investigación es la demostración del se centra en que el conocimiento multicapa con datos de expresión génica para entrenar la red neuronal de convolución profunda estratifica la recaída del paciente y la dosificación del fármaco junto con las propiedades moleculares subyacentes[14].

El artículo Ensemble learning method for the prediction of breast cancer recurrence propone una mejora en la predicción de la recurrencia del cáncer de seno utilizando una técnica de Ensemble Learning y además proporciona un sitio web que permite a los médicos ingresar características relacionadas con un paciente con cáncer de seno y obtener la probabilidad de recurrencia de dicho cáncer. Los resultados demuestran que los modelos con mejor desempeño son el modelo Random Forest con precisión 0.6522, sensibilidad 0.6250 y especificidad 0.6593, luego el modelo Decisión Tree con precisión 0.6261, sensibilidad 0.63636 y especificidad 0.62500, y el modelo de Naïve Bayes con precisión 0.5913, sensibilidad 0.4889 y especificidad 0.6571[15].

En el artículo Estrogen receptor status prediction for breast cancer using artificial neural network, se estudia el patrón de los genes de un grupo de 278 muestras de cáncer de mama, etiquetados en clases de Receptor de Estrógenos positivo y Receptor de Estrógenos negativo, utilizando la red neuronal artificial (ANN). El modelo demuestra su eficacia para seleccionar genes significativos en comparación con otros estudios realizados y también muestra que los genes altamente clasificados están asociados con el desarrollo del cáncer de mama[16].

El artículo Interpretability of Artificial Hydrocarbon Networks for Breast Cancer Classification trata de demostrar la capacidad de las redes de hidrocarburos artificiales (AHN) para entregar modelos interpretables que puedan ser procesados con diversos métodos de clasificación de Machine Learning para detectar el cáncer de mama. Con el fin de evaluar la interpretabilidad de AHN, los investigadores abordan el problema del cáncer de mama utilizando un conjunto de datos públicos. Los resultados mostraron que AHN puede transformarse en modelos basados en árboles y basados en reglas, conservando una alta precisión en la clasificación de salida[17].

El artículo Learning Approaches to Improve Prediction of Drug Sensitivity in Breast Cancer Patients, presenta tres enfoques de aprendizaje para mejorar la predicción de la res-

puesta de los pacientes con cáncer de mama al medicamento de quimioterapia: el enfoque de selección de instancia, el enfoque de sobremuestreo y el enfoque híbrido. Los investigadores evalúan el rendimiento de sus enfoques y los comparan con un enfoque de referencia utilizando el Área bajo la curva ROC (AUC) en los datos de ensayos clínicos, además de probar la estabilidad de los enfoques. Los resultados experimentales muestran la estabilidad de los enfoques propuesto dan el AUC más alto con significación estadística[18].

El artículo Machine learning applications in cancer prognosis and prediction, presenta una revisión de los enfoques recientes de Machine Learning (ML) empleados en el modelado de la progresión del cáncer. Los modelos predictivos discutidos en la investigación se basan en varias técnicas de ML supervisadas, así como en diferentes características de entrada y muestras de datos. Dada la tendencia creciente en la aplicación de métodos de ML en la investigación del cáncer, el artículo presenta las publicaciones más recientes que emplean estas técnicas como un objetivo para modelar el riesgo de cáncer o los resultados de los pacientes[19].

El artículo Mobile Personal Health Record (mPHR) for Breast Cancer using Prediction Modeling proporciona herramientas predictivas de un soporte de decisión para evaluar el nivel de malignidad de un tumor, los datos fueron obtenidos del Hospital de Oncología de Surabaya (Indonesia). En esta investigación, también se desarrolla un registro de salud personal móvil (mPHR) para la enfermedad del cáncer de mama utilizando métodos predictivos de Machine Learning. Para lograrlo los investigadores realizan una regresión logística y la comparan con el método Naive Bayes utilizado para diagnosticar el riesgo de malignidad tumoral relacionada con el cáncer de mama[20].

El artículo Breast Cancer Intelligent Diagnosis based on Subtractive Clustering Adaptive Neural Fuzzy Inference System and Information Gain utiliza un nuevo método inteligente para el diagnóstico de cáncer de mama. El método combina el método de ganancia de información y el sistema de inferencia difusa neural adaptativa de agrupamiento sustractivo (IG-SCANFIS). El método de ganancia de información se aplica para reducir la dimensión de los atributos y luego aplica los atributos seleccionados como entrada a SCANFIS. El modelo SCANFIS utiliza el algoritmo de agrupamiento sustractivo para agrupar los datos de entrada para obtener las reglas difusas y establecer el sistema de razonamiento difuso neural. Los conjuntos de datos utilizados para el entrenamiento y las pruebas se obtuvieron de la librería de Machine Learning Irvine (UCI) de la Universidad de California. El resultado de la simulación muestra que el método propuesto tiene una precisión de 99.44%[21].

El artículo On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context aborda el problema de la predicción del cáncer de mama en el contexto del Big data. En él se consideran dos variedades de datos: La expresión génica (GE) y la metilación del ADN (DM). El objetivo de la investigación fue ampliar los algoritmos de Machine Learning que se utilizan para la clasificación mediante la aplicación de cada conjunto de datos por separado y de forma conjunta. Para este propósito,

eligieron Apache Spark como plataforma. En el artículo se seleccionan tres algoritmos de clasificación diferentes: Support Vector Machine (SVM), Decision Tree y Random Forest, para crear nueve modelos que ayuden a diagnosticar el cáncer de mama. Los resultados experimentales mostraron que el clasificador Scaling SVM en el entorno Spark supera a los otros clasificadores, ya que logró la mayor precisión[22].

En el artículo Post-Surgical Survival forecasting of breast cancer patient: a novel approach se propone disminuir el porcentaje de muertes a causa de cáncer de mama. Además de poder realizar el diagnóstico es necesario llevar un control de las cirugías que se realicen a los pacientes. Tumorectomía y mastectomía son los procedimientos quirúrgicos mayormente utilizados como tratamiento. Este artículo Propone la utilización de predicción de supervivencia utilizando la máquina de vectores de soporte comunicación eficiente distribuida doble coordenada ascenso (SVM)[23].

En el artículo Predicting cancer outcomes from histology and genomics using convolutional networks implementan redes neuronales convolucionales (SCNN) basados en imágenes histológicas y biomarcadores genómicos para el diagnóstico de cáncer[24].

La investigación Predicting Invasive Disease-Free Survival for Early Stage Breast Cancer Patients Using Follow-Up Clinical Data busca diagnosticar la supervivencia de cáncer de mama en pacientes en estado temprano con el análisis de datos clínicos existentes, utilizando el framework MP4Ei que es una propuesta de los autores[25].

En el artículo Prediction Models for Estimation of Survival Rate and Relapse for Breast Cancer Patients se presenta al clasificador Naive Bayes como modelo para el pronóstico de la supervivencia del cáncer en función de la tasa de supervivencia de 5 años, mientras que la Red Neuronal, según los autores, mostró mejor rendimiento en el pronóstico de la recurrencia del cáncer. Presentan una de las ventajas de la utilización de clasificadores para realizar diagnosis : “Los sistemas de clasificación pueden ayudar a minimizar los posibles errores que se pueden presentar debido a expertos sin experiencia, y también proporcionan datos médicos para ser examinados en un tiempo más corto y más detallado” [26].

En el artículo Probabilistic Graphical Models and Deep Belief Networks for Prognosis of Breast Cancer, implementan un (PGM) ,el cual es un clasificador bayesiano y Manifold Learning para la reducción dimensional para el pronóstico y diagnóstico de cáncer de seno que puede ayudar a los médicos a tomar mejores decisiones sobre el mejor tratamiento para un paciente[27].

La investigación Prognosis Prediction of Human Breast Cancer by integrating Deep Neural Network and Support Vector Machine integra una red neuronal profunda con una máquina de soporte de vectores para la prognosis de cáncer de mama utilizando información que se encuentra en <https://gdac.broadinstitute.org> donde existen Datasets de transcripción genética de diferentes tipos de cáncer entre ellos el mamario[28].

En el artículo Using Random Forest Algorithm for Breast Cancer Diagnosis, se imple-

menta el algoritmo de Random Forest para el diagnóstico de cáncer de mama. Se muestra como el Random Forest puede llegar a tener mejor rendimiento entre más de 179 algoritmos testeados. En la investigación se hace uso de la librería de Scikit-Learn para el lenguaje de programación Python en su versión 3.6. El Dataset utilizado está publicado en la página web de UC Irvine Machine Learning[29].

En el artículo A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data se hace uso de una Red neuronal profunda para la predicción de cáncer de mama implementado datos multidimensionales, haciendo uso de la librería TensorFlow 1.0 y un set de datos que muestra información genómica. Para la comparación de modelos de Machine Learning se utilizan cuatro parámetros: Acc (exactitud), Pre (Precisión), Sn, and Mcc[30].

En el artículo Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique se investigaron tres métodos: Support Vector Machine, Decision Tree and Naïve Bayes para diagnosticar la recurrencia del cáncer de mama utilizando WEKA, una herramienta que contiene varios algoritmos de Machine Learning. El Dataset utilizado fue extraído de UC Irvine Machine Learning, el cual cuenta con 34 atributos. Como resultado se puede evidenciar que el algoritmo Support Vector Machine tiene mejores resultados que sus dos competidores[31].

En el artículo Automated Detection of Breast Cancer Metastases in Whole Slide Images se implementa el método Random Forest Classifier para la clasificación de imágenes y para la localización de tumores. En donde se utiliza Python como lenguaje de programación y librerías como Scikit-learn[32].

La investigación Stacked Regression Ensemble for Cancer Class Prediction, presenta un modelo llamado Stacked Regression Ensemble (SRE) para la predicción de diferentes clases de cáncer, comparando su rendimiento con otros modelos como: SVM (Máquina de soporte de vectores) y GRNN (General regression neural network) enfocados en la clasificación y ordenamiento de datos referentes al cáncer[33].

En la publicación Enhanced Deep Learning Approach for Predicting Invasive Ductal Carcinoma from Histopathology Images se pretende diagnosticar el carcinoma ductal invasivo a partir de Imágenes de histopatología utilizando aprendizaje profundo. Este estudio entrenó a una red neuronal convolucional mejorada y se investigó el rendimiento del modelo en la tarea de clasificación basada en parches IDC (Carcinoma ductal invasivo, es el tipo más común de cáncer de mama). El conjunto de datos consta de 277,524 imágenes 50x50 píxeles que muestran la histopatología mamaria. Se utilizó Python como lenguaje de programación. Según los autores la implementación de esta solución tiene una precisión del 0,86 para casos positivos de cáncer y 0.85 para casos negativos de cáncer de mama[34].

Para la disminución de dimensionalidad de los datos, en el artículo A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA

Methylation proponen la implementación del filtro F-score y al mostrar los resultados que arrojan los métodos: Naïve Bayes, Random Forest y Support Vector Machine, para clasificar los diferentes tipos de cáncer entre ellos: mama, colon, cabeza, riñón, pulmón, tiroides, y uterino con y sin abordaje hibridizado[35].

El artículo Breast Cancer Prognosis via Gaussian Mixture Regression compara el desempeño de clasificación y regresión por árboles (CART), splines de regresión adaptativa multivariante (MARS) y un método de regresión de mezcla gaussiana (GMR) para diagnosticar el tiempo de recurrencia del cáncer de mama. Se muestra que el algoritmo basado en GMR demuestra una mejora del rendimiento en comparación con CART y MARS. Además, el rendimiento de GMR es comparable al de un predictor de referencia con la ventaja de realizar la selección automática de características y la optimización del modelo. Se utilizó la Base de datos de Wisconsin (Prognosis Breast Cancer), la cual está compuesta por 253 registros de pacientes clasificados como malignos y están disponibles al público[36].

En el artículo Prediction of Breast Cancer using Voting Classifier Technique comparan los resultados de algoritmos de clasificación de aprendizaje y combinación de estos algoritmos que utilizan la técnica de clasificador de votación. Votar es un enfoque de conjunto donde se puede combinar múltiples modelos para la mejor clasificación. El conjunto de datos se toma de la Universidad de Wisconsin que contiene 699 instancias con 12 atributos. En la investigación implementan los siguientes métodos clasificadores: Árboles de decisión, Naïve Bayes y Support Vector Machine[37].

La implementación de modelos de Machine Learning se ve aplicada en la publicación Predicting Malignancy from Mammography Findings and Surgical Biopsies donde el objetivo principal es diagnosticar el resultado de una mamografía con un pequeño set de datos de anotaciones encontrados en otras mamografías. Los datos están conformados por 348 masas mamarias realizadas entre 2005 a 2008 a 328 mujeres. Se hace reiterativo la utilización de los algoritmos Naive Bayes, DTNB y SMO con la plataforma WEKA[38].

Para la clasificación de tumores benignos y malignos basados en una máquina de vectores de soporte (SMV) en el artículo SVM Approach to Breast Cancer Classification usan una base de datos ofrecida por la universidad de Wisconsin con imágenes de biopsias. El conjunto SVM con éxito clasificó más del 99 % de los datos de prueba y en el proceso arrojó una predicción del tumor benigno con 100 % de exactitud[39].

En el artículo Predicting Cancer Relapse with Clinical Data: A Survey of Current Techniques se exploran los métodos actuales para construir modelos de riesgo de recurrencia del cáncer con estructuras de datos de pacientes clínicos[40].

La propuesta presentada en el artículo CWV-BANN-SVM: ensemble learning classifier for an accurate diagnosis of breast cancer aplica el modelo Support Vector Machine (SVM) y redes neuronales artificiales (ANN) para analizar los datos de cáncer de mama con el ya conocido conjunto de datos de cáncer de mama de Wisconsin (WBCD) [41].

2.5.2. Marco Conceptual

El eje central de la investigación está basado principalmente en las técnicas de predicción utilizadas en Machine Learning (ML). ML es el diseño y estudio de las herramientas informáticas que utilizan la experiencia pasada para tomar decisiones futuras; es el estudio de programas que pueden aprender de los datos. El objetivo fundamental del Machine Learning es generalizar, o inducir una regla desconocida a partir de ejemplos donde esa regla es aplicada[42]. Para poder comprender el funcionamiento de Machine Learning es necesario conocer las siguientes variantes que lo componen:

Aprendizaje supervisado

En los problemas de aprendizaje supervisado se enseña o se entrena al algoritmo a partir de datos que ya vienen etiquetados con la respuesta correcta. Cuanto mayor es el conjunto de datos más el algoritmo puede aprender sobre el tema. Una vez concluido el entrenamiento, se le brindan nuevos datos. Ya sin las etiquetas de las respuestas correctas, el algoritmo de aprendizaje utiliza la experiencia pasada que adquirió durante la etapa de entrenamiento para diagnosticar un resultado[42].

Aprendizaje no supervisado

En los problemas de aprendizaje no supervisado el algoritmo es entrenado usando un conjunto de datos que no tiene ninguna etiqueta; en este caso, nunca se le dice al algoritmo lo que representan los datos. La idea es que el algoritmo pueda encontrar por sí solo patrones que ayuden a entender el conjunto de datos. El aprendizaje no supervisado es similar al método que utilizamos para aprender a hablar cuando somos bebés, en un principio escuchamos hablar a nuestros padres y no entendemos nada; pero a medida que vamos escuchando miles de conversaciones, nuestro cerebro comenzará a formar un modelo sobre cómo funciona el lenguaje y comenzaremos a reconocer patrones y a esperar ciertos sonidos[42].

Aprendizaje por refuerzo

En los problemas de aprendizaje por refuerzo, el algoritmo aprende observando el mundo que le rodea. Su información de entrada es el feedback o retroalimentación que obtiene del mundo exterior como respuesta a sus acciones. Por lo tanto, el sistema aprende a base de ensayo-error[42].

Los algoritmos que más se suelen utilizar en los problemas de Machine Learning son los siguientes:

Linear Regression

Se utiliza para estimar los valores reales (costo de las viviendas, el número de llamadas, ventas totales, etc.) basados en variables continuas. La idea es tratar de establecer la relación entre las variables independientes y dependientes por medio de ajustar una mejor línea recta con respecto a los puntos[42].

Logistic Regression

Los modelos lineales, también pueden ser utilizados para clasificaciones; es decir, que primero ajustamos el modelo lineal a la probabilidad de que una cierta clase o categoría ocurra y, a luego, utilizamos una función para crear un umbral en el cual especificamos el resultado de una de estas clases o categorías. La función que utiliza este modelo es denominada regresión logística[42].

Decision Trees

Los Árboles de Decisión son diagramas con construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, para la resolución de un problema. Los Árboles de Decisión están compuestos por nodos interiores, nodos terminales y ramas que emanan de los nodos interiores[42].

Random Forest

La idea central detrás del algoritmo de Random Forest es construir una gran cantidad de árboles muy poco profundos, y luego se toma la clase que cada árbol eligió[42].

Support Vector Machines(SVM)

La idea detrás de SVM es encontrar un plano que separe los grupos dentro de los datos de la mejor forma posible. Aquí, la separación significa que la elección del plano maximiza el margen entre los puntos más cercanos en el plano; estos puntos se denominan vectores de soporte[42].

K-Nearest Neighbors(KNN)

Este es un método de clasificación no paramétrico, que estima el valor de la probabilidad a posteriori de que un elemento xx pertenezca a una clase en particular a partir de la información proporcionada por el conjunto de prototipos. La regresión KNN se calcula simplemente tomando el promedio del punto k más cercano al punto que se está probando[42].

K-Means Clustering

Este es probablemente uno de los algoritmos de agrupamiento más conocidos y, en un sentido más amplio, una de las técnicas de aprendizaje no supervisado más conocidas. K-means es en realidad un algoritmo muy simple que funciona para reducir al mínimo la suma de las distancias cuadradas desde la media dentro del agrupamiento[42].

2.6. Metodología de la investigación

2.6.1. Tipo de Estudio

El tipo de estudio de la investigación es descriptivo porque se van a presentar las características del problema de investigación planteado, que en el caso de la presente investigación tiene que ver con el análisis y caracterización de datos de pacientes que padecen de cáncer de mama.

2.6.2. Método de Investigación

Es importante definir el método que se va a seguir en la investigación, por esto se define que: al partir de un marco teórico ya establecido, que tiene como base todo el músculo cognitivo de la inteligencia artificial y en especial de Machine Learning se puede implementar o aplicar los conocimientos que este marco teórico brinda en una realidad concreta como lo es el diagnóstico del cáncer de mama según esto se determina que el método que se va a seguir en la investigación es el *Deductivo*.

2.6.3. Fuentes y Técnicas para la recolección de Información

La información sobre Machine Learning abunda en este tiempo. Basta con buscar Machine Learning en la IEEE para darse cuenta del gran número de artículos que tienen o mencionan el aprendizaje automático en sus manuscritos. Según las consultas que se han realizado para la investigación, se observa que la gran mayoría de información se encuentra en formatos de tipo artículo y publicaciones indexadas. Por lo anterior, las fuentes de información serán secundarias, basadas principalmente en producción intelectual de otras personas, especialmente las que se encuentran consignadas en bases de datos como la IEEE y Web of science, de las que se tiene acceso gracias a la Universidad Distrital Francisco José de Caldas.

Lo anterior hace parte exclusivamente de la información referente a la teoría que soporta la investigación, pero también es importante la información de los Dataset con los que se va a entrenar el modelo de Machine Learning, para esto se hará usos de algunas bases de datos públicas que contienen información en diferentes formatos para el análisis y clasificación de las diferentes fases de cáncer.

2.7. Organización del Trabajo de Grado

El trabajo en cuestión se encuentra organizado en tres secciones:

- La primera sección define la metodología, el contexto y el propósito de la investigación además de los elementos necesarios para entenderla. Posteriormente se realiza el desarrollo de la investigación
- En la segunda sección se puede encontrar la arquitectura del proyecto y cada una de las etapas realizadas para implementar los modelos enfocados en la diagnosis del cáncer de mama.
- En la tercera sección se encuentra el resultado de la investigación basada en la implementación los modelos predictivos existentes. Además se encuentra los posibles trabajos en los cuales puede incurrir la investigación en un futuro.

Capítulo 3

Desarrollo de la Investigación

Para realizar la implementación de los modelos de Machine Learning se utilizo *Python* en su versión 3 además de la librería *Scikit-Learn* que cuenta con un gran número de funciones que fortalecen el desarrollo y la implementación de modelos de Machine Learning. Con la ayuda de la herramienta *Jupyter Notebooks* se realizaron las primeras pruebas de los algoritmos, esta herramienta permite ejecutar bloques de código específico que dan una ventaja clara a la hora de probar estos tipo de modelos. La información está contenida en el Data-Set de cáncer de mama de la Universidad de Wisconsin que tiene 359 registros con 32 atributos.

A continuación se describen los procesos desde la recolección de datos hasta la implementación y desarrollo de una aplicación web denominada (BreastApp) la cual utiliza métodos de Machine Learning para diagnosticar el Cáncer de mama teniendo como base el Data-Set de la Universidad de Wisconsin.

3.1. Recolección de datos

El enfoque de Machine Learning aplicado al diagnóstico y pronóstico del cáncer de mama es el resultado de una colaboración en la Universidad de Wisconsin-Madison entre el Profesor Olvi L. Mangasarian del Departamento de Ciencias de la Computación y el Dr. William H. Wolberg de los departamentos de Cirugía y Oncología Humana[43].

La investigación surgió del deseo del Dr. Wolberg de diagnosticar con precisión las masas mamarias basadas únicamente en una aspiración con aguja fina (FNA) por sus siglas en inglés *Fine Needle Aspiration*. Identificó nueve características evaluadas visualmente de una muestra de FNA que consideró relevantes para el diagnóstico. En colaboración con el profesor Mangasarian y dos de sus estudiantes de posgrado, Rudy Setiono y Kristin Bennett, se construyó un clasificador utilizando el método multisuperficie (MSM), por sus siglas en inglés *Multisurface Method*, de separación de patrones. En estas nueve características se diagnosticaron con éxito el 97% de casos nuevos de cáncer. El conjunto de datos resultante es conocido como el **Data-Set de cáncer de mama de Wisconsin**[43].

El trabajo de análisis de imágenes comenzó en 1990. El objetivo era diagnosticar la muestra con base en una imagen digital de una pequeña sección de una muestra FNA, el ejemplo de dicha muestra se puede observar en la Figura 3.1[43]. En ella se puede observar las *Snakes* las cuales son definidas como contornos basados en la representación final de los límites del núcleo celular [44].

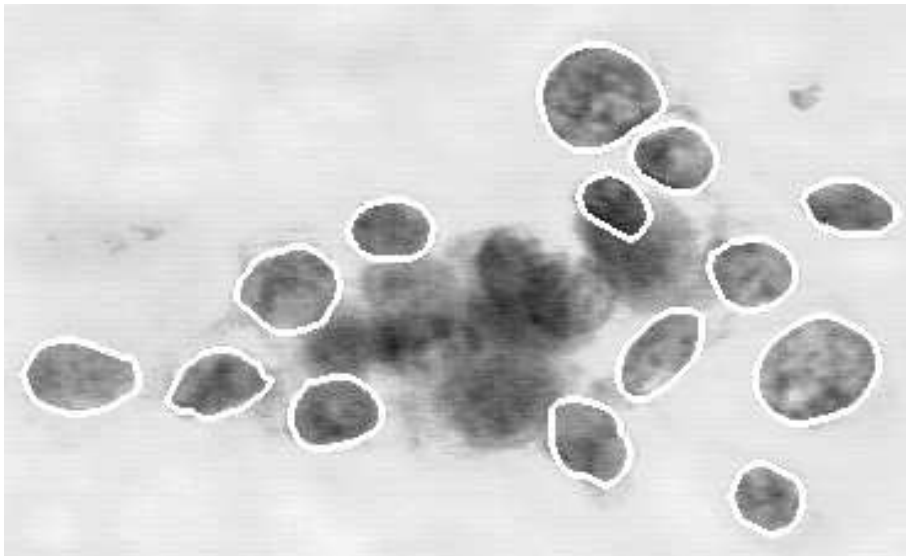


Figura 3.1: Muestra FNA capturada con Microscopio y procesada con el Software Xcyt

3.2. Verificación de datos

La verificación de los datos extraídos por medio de la muestra FNA y la veracidad del Data-Set de cáncer de mama de Wisconsin se realiza con base a los resultados obtenidos en la investigación del profesor Olvi L. Mangasarian basada en un sistema de software conocido como Xcyt. El proceso de diagnóstico fue realizado de la siguiente manera:

- Se toma una FNA de la masa mamaria. Este material se monta en la platina de muestras del microscopio y se retiene para resaltar los núcleos celulares. Una parte de la platina en la que las células están bien diferenciadas se escanea con una cámara digital y una placa de captura de fotogramas[43].
- El usuario luego aísla los núcleos individuales usando Xcyt . Usando un puntero del mouse, el usuario dibuja el límite aproximado de cada núcleo. Usando un enfoque de visión por computadora conocido como *snakes*, estas aproximaciones luego convergen a los límites nucleares exactos. Este proceso interactivo lleva entre dos y cinco minutos por muestra[43].
- Una vez que todos (o la mayoría) de los núcleos han sido aislados, el programa calcula los valores para cada una de las diez características de cada núcleo, midiendo el tamaño, la forma y la textura. Se calculan la media, el error estándar y los valores extremos de estas características, lo que da como resultado un total de 30 características nucleares para cada muestra[43].
- Basado en un conjunto de entrenamiento de 569 casos, se construyó un clasificador lineal para diferenciar las muestras benignas de las malignas. Este clasificador consta de un solo plano de separación en el espacio de tres de las características: valor extremo de área, valor extremo de suavidad y valor medio de textura. Al proyectar todos los casos en este plano de separación normal, se construyeron densidades de probabilidad aproximadas de los puntos benignos y malignos. Estos permiten un cálculo bayesiano simple de la probabilidad de malignidad para nuevos pacientes. Estas densidades se muestran al paciente, lo que le permite juzgar la *confianza* de su diagnóstico en comparación con cientos de muestras anteriores[43].

Hasta la fecha, este sistema ha diagnosticado correctamente 176 nuevos pacientes consecutivos (119 benignos, 57 malignos). En solo ocho de esos casos, Xcyt devolvió un diagnóstico *sospechoso* (es decir, una probabilidad estimada de malignidad entre 0.3 y 0.7)[43].

3.3. Clasificación de datos

El Sistema de enfoque de visión por computadora extrae diez características diferentes de los límites del núcleo celular generado en los *snakes*. Todas las características se modelan numéricamente de modo que los valores más grandes generalmente indican una mayor probabilidad de malignidad. Las características extraídas son las siguientes [44]:

- **Radio (*Radius*):** El radio de un núcleo individual se mide promediando la longitud de los segmentos lineales radiales definidos por el centroide y los puntos de individuales de la *snake*.
- **Perímetro (*Perimeter*):** El perímetro nuclear es la distancia total entre los puntos de la *snake*.
- **Área (*Area*):** El área nuclear se mide contando el número de píxeles en el interior de la *snake* y agregando la mitad de los píxeles en el perímetro.
- **Textura (*Texture*):** La textura del núcleo celular se mide al encontrar la varianza de las intensidades de escala de grises en los píxeles encontrados en la digitalización de la muestra.
- **Lisura (*Smoothness*):** La lisura de un contorno nuclear se cuantifica midiendo la diferencia entre la longitud de una línea radial y la longitud media de las líneas que lo rodean. En la Figura 3.2 se puede observar de manera gráfica la medición de la lisura [44].

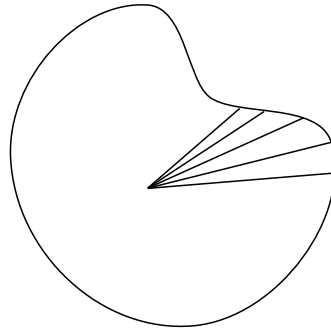


Figura 3.2: Medición de la lisura de un contorno nuclear

- **Concavidad (Concavity):** Se obtiene al medir el número y la gravedad de las concavidades o hendiduras en un núcleo celular. Para ellos se dibujan puntos no adyacentes en la *snake* y se mide el grado en que el límite real del núcleo se encuentra en el interior de cada línea. En la Figura 3.3 se puede observar de manera gráfica la medición de la concavidad [44].

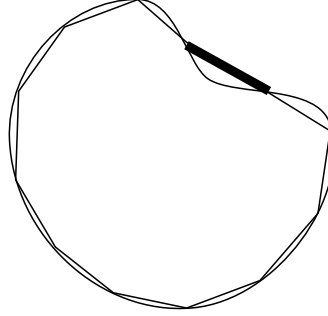


Figura 3.3: Medición de concavidad del núcleo celular

- **Puntos Cóncavos (Concave Points):** Esta característica es similar a la Concavidad, pero mide solo el número, en lugar de la magnitud, de las concavidades del contorno.
- **Simetría (Symmetry):** Para medir la simetría, se encuentra el eje principal, o la línea más larga a través del centro. Luego se mide la diferencia de longitud entre líneas perpendiculares del eje principal al límite de la celda en ambas direcciones. En la Figura 3.4 se puede observar de manera gráfica la medición de la simetría de un núcleo celular [44].

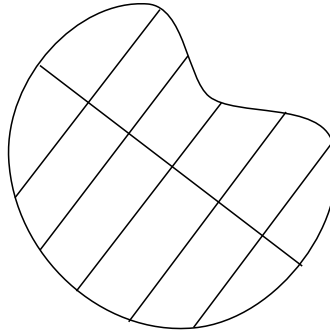


Figura 3.4: Medición de la simetría de un núcleo celular

- **Compacidad (Compactness):** El perímetro y el área se combinan para obtener una medida del espacio compacto de los núcleos celulares. Este número adimensional se minimiza mediante un disco circular y aumenta con la irregularidad del límite. Sin embargo, esta medida de forma también aumenta para los núcleos celulares alargados, que no necesariamente indican una mayor probabilidad de malignidad. La característica también está sesgada hacia arriba para las celdas pequeñas debido a la menor precisión impuesta por la digitalización de la muestra.
- **Dimensión Fractal (Fractal Dimension):** La dimensión fractal de una célula se calcula utilizando el método de la *Paradoja de la línea de costa* de Benoît Mandelbrot. En ella el perímetro del núcleo se mide utilizando intervalos de escala cada vez más grandes. A medida que aumenta el tamaño de la escala, disminuyendo la precisión de la medición, disminuye el perímetro observado. Al trazar estos valores en una escala logarítmica y medir la pendiente descendente se obtiene una aproximación a la dimensión fractal. Al igual que con todas las características de forma, un valor más alto corresponde a un contorno menos regular y, por lo tanto, a una mayor probabilidad de malignidad. En la Figura 3.5 se puede observar de manera gráfica las características de la dimensión fractal[44].

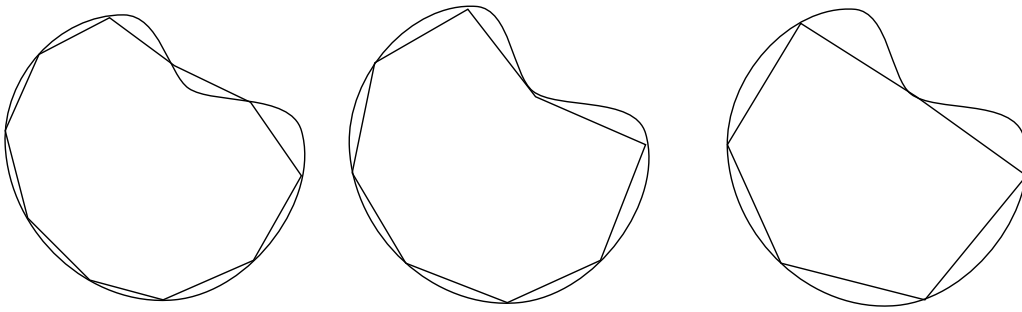


Figura 3.5: Medición de la dimensión Fractal de un núcleo celular

Todas las características de la forma de las masas mamarias se verificaron utilizando células fantasma idealizadas. Se demostró que aumentaban a medida que los límites se volvían menos regulares y que en gran medida no estaban correlacionados con el tamaño del contorno. El valor medio, el valor extremo (mayor) y el error estándar de cada característica fueron calculados para cada imagen en la cual esta basado el Data-Set de la Universidad de Wisconsin[44].

3.4. Limpieza de datos

Para el correcto entrenamiento de los algoritmos con el del Data-Set de la Universidad de Wisconsin y obtener un resultado acertado con respecto al diagnóstico del cáncer de mama es necesario tener alta calidad en los datos, para ello es necesario la corrección de los errores de los mismos para el posterior procesamiento con los algoritmos de Machine Learning. Para la correcta limpieza de los datos se realizaron los siguientes pasos:

3.4.1. Importación de datos:

El Data-Set de cáncer de mama de la Universidad de Wisconsin por defecto esta en un Formato de texto .csv, el peso del archivo es de 125 Kb , y la cantidad de registros que contiene son 569. Para la comprobación del Data-Set se utilizó la librería *Pandas* la cual cuenta con la función *read_csv()* para leer las variables del Data-Set, la función *head()* para visualizar un numero de registros determinados y la propiedad *shape* la cual realiza el conteo de las cantidad de filas y columnas. A continuación se puede observar el uso de la funciones *read_csv* , *head()* y la propiedad *shape*:

```
[In]: # Leer el Data-Set
df = pd.read_csv('data.csv')
#Visualizar n registros del Data-Set
df.head(7)
```

```
[Out]: id diagnosis  radius_mean  texture_mean  perimeter_mean  area_mean  ...
0      842302        M          17.99         10.38          122.80  ...
1      842517        M          20.57         17.77          132.90  ...
2      84300903       M          19.69         21.25          130.00  ...
3      84348301       M          11.42         20.38           77.58  ...
4      84358402       M          20.29         14.34          135.10  ...
5      843786        M          12.45         15.70           82.57  ...
6      844359        M          18.25         19.98          119.60  ...
```

```
[In]: #Generar número de filas y columnas
df.shape
```

```
[Out]: (569, 33)
```

3.4.2. Formateo de datos:

Para la comprobación de los tipos de datos del Data-Set de cáncer de mama de la Universidad de Wisconsin se utilizó la librería *Pandas* la cual cuenta con la propiedad *dtypes* la cual genera los tipos de datos de las variables de dicho Data-Set. A continuación se puede observar el uso de la propiedad *dtypes*:

```
[In]: #Tipo de Datos
df.dtypes()
```

```
[Out]: id                int64  radius_worst          float64
diagnosis              object  texture_worst          float64
radius_mean           float64  perimeter_worst       float64
texture_mean          float64  area_worst          float64
perimeter_mean        float64  smoothness_worst    float64
area_mean             float64  compactness_worst   float64
smoothness_mean       float64  concavity_worst     float64
compactness_mean      float64  concave points_worst float64
concavity_mean        float64  symmetry_worst      float64
concave points_mean   float64  fractal_dimension_worst float64
symmetry_mean         float64  dtype: object
fractal_dimension_mean float64
radius_se             float64
texture_se            float64
perimeter_se          float64
area_se              float64
smoothness_se         float64
compactness_se        float64
concavity_se          float64
concave points_se     float64
symmetry_se           float64
fractal_dimension_se  float64
```

3.4.3. Comprobación de datos faltantes:

Para la comprobación de datos faltantes se utilizó a la librería *Pandas* la cual cuenta con la función *isna()* que valida si las variables contienen valores *Nulos(Nan)*. En este caso en específico el Data-Set de cáncer de mama de la Universidad de Wisconsin no contiene ningún valor nulo. A continuación se puede observar el uso de la función *isna()*:

```
[In]: #Contando los datos vacíos (NaN, NAN, na)
df.isna().sum()
```

```
[Out]: id                                0  radius_worst                                0
      diagnosis                          0  texture_worst                                0
      radius_mean                        0  perimeter_worst                             0
      texture_mean                       0  area_worst                                 0
      perimeter_mean                     0  smoothness_worst                         0
      area_mean                          0  compactness_worst                       0
      smoothness_mean                    0  concavity_worst                         0
      compactness_mean                    0  concave points_worst                    0
      concavity_mean                     0  symmetry_worst                         0
      concave points_mean                  0  fractal_dimension_worst                 0
      symmetry_mean                       0
      fractal_dimension_mean               0
      radius_se                           0
      texture_se                           0
      perimeter_se                         0
      area_se                             0
      smoothness_se                       0
      compactness_se                      0
      concavity_se                        0
      concave points_se                   0
      symmetry_se                         0
      fractal_dimension_se                 0
```

3.4.4. Eliminación de datos Innecesarios:

Para la eliminación de datos innecesarios se utilizo a la librería *Pandas* la cual cuenta con la función *dropna()* para eliminar la cabecera. A continuación se puede observar el uso de la función *head()* y *dropna()* :

```
[In]: # Eliminar la primera fila  
df = df.dropna(axis=1)
```

```
[Out]: 0      842302      M      17.99      10.38      122.80      1001.0 ...  
1      842517      M      20.57      17.77      132.90      1326.0 ...  
2      84300903     M      19.69      21.25      130.00      1203.0 ...  
3      84348301     M      11.42      20.38      77.58      386.1 ...  
4      84358402     M      20.29      14.34      135.10      1297.0 ...  
5      843786      M      12.45      15.70      82.57      477.1 ...  
6      844359      M      18.25      19.98      119.60      1040.0 ...
```

3.5. Procesamiento de los Datos

En primera instancia se importan las librerías necesarias para el procesamiento de datos y la generación de imágenes estadísticas, en el primer grupo se encuentran las librerías *Pandas*, *Numpy* y en el segundo *Matplotlib* y *Seaborn*.

```
[In]: #Importar librería para la manipulación de datos
import numpy as np
#Importar librería para la manipulación de los datos
import pandas as pd
#Importar librería para la visualización de datos
import matplotlib.pyplot as plt
#Importar librería para la visualización de datos estadísticos
import seaborn as sns
#Importar librería para pre-procesar datos
import sklearn.preprocessing
```

Los datos se encuentran en un archivo .csv que es leído y cargado en memoria con la ayuda de la librería *Pandas*, para luego ser asignado a una variable de Data-Frame que hace que la información sea manejable y controlable de forma sencilla. Dentro del Data-Set se encuentran 357 registros de personas con tumores diagnosticados como Benignos y 212 como Malignos, esta información será utilizada para entrenar y testear el modelo.

```
[In]: # Leer el Data-Set
df = pd.read_csv('data.csv')
# Agrupar los valores de la columna indicada y sumarlos
df['diagnosis'].value_counts()
```

```
[Out]: B    357
       M    212
       Name: diagnosis, dtype: int64
```


En la Figura 3.6 se observa según el procesamiento del Data-Set la comparación de casos Malignos y Benignos.

```
[In]: #Visualización de la columna "diagnosis"  
sns_plot = sns.countplot(df['diagnosis'], label="Count")
```

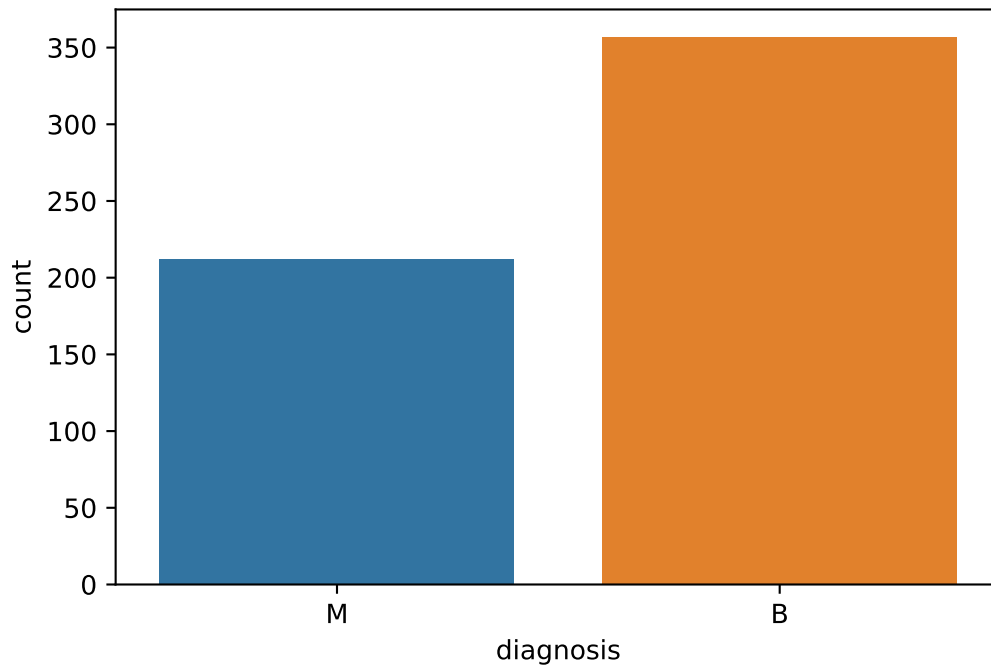


Figura 3.6: Comparación Estadística casos Benignos vs Malignos

Es importante entender las relaciones que existen entre el conjunto de variables y su respectiva diagnosis, de esta manera se establece cuáles de las variables tienen mayor influencia en el entrenamiento del sistema. Para esto se utiliza el gráfico de correlación cruzada entre todas las variables que es denominado *de parejas* o *pairplot()* para el caso de la librería Seaborn,este gráfico puede ser observado en la Figura 3.7.

```
[In]: # Visualización gráfico de correlación cruzada
sns_plot = sns.pairplot(df, hue="diagnosis", height=2.5)
```



Figura 3.7: Gráfico de correlación cruzada de variables Oncológicas

Para el caso del gráfico de calor mostrado en la Figura 3.8 ,se puede evidenciar la relación porcentual que existe entre cada variable y muestra de una forma muy concisa y útil las relaciones entre todas las magnitudes de interés en este conjunto de datos. Para este caso por ejemplo se puede evidenciar que el atributo que más influye en que la diagnosis sea Maligna es el de *concave points_worst* que cuenta con un 79% de ocurrencias respecto al atributo *diagnosis*. Con esta gráfica se pueden separar los atributos que mayor importancia tienen en el Data-Set, entre estos atributos se encuentran : *radius_mean* con 73%, *perimeter_mean* con 74%, *area_mean* con 71%, *concavity_mean* con 70% y *radius_worst* con 78%.

```
[In]: # Visualización gráfico de calor
plt.figure(figsize=(20,20))
sns_plot = sns.heatmap(df.corr(), annot=True, fmt='.0%')
```

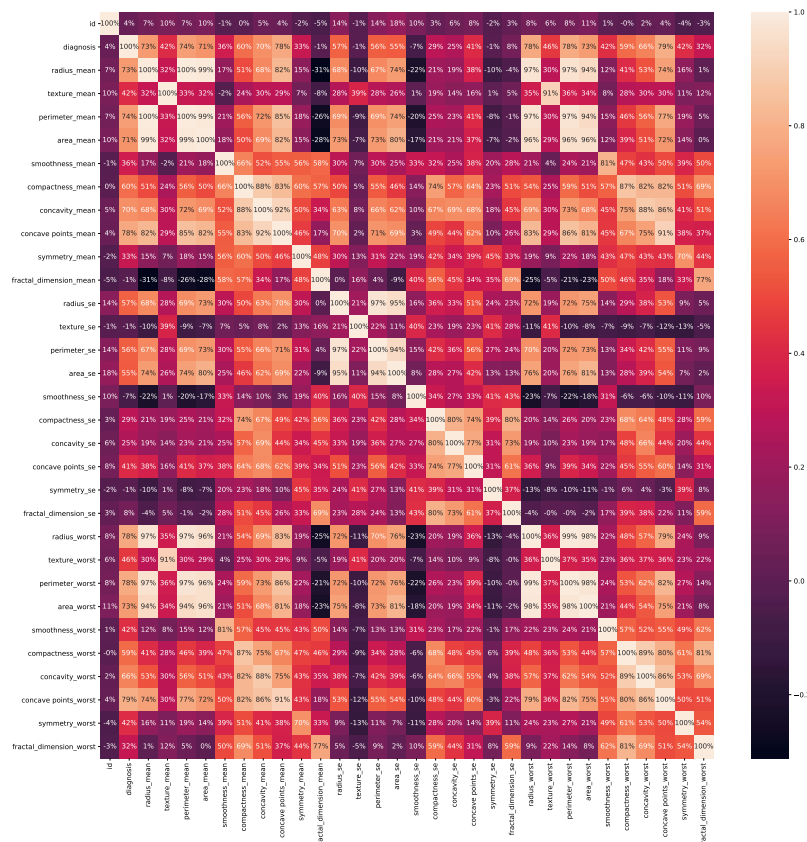


Figura 3.8: Diagrama de Calor de la correlación de variables Oncológicas

Los atributos que contiene el Data-Set en sus totalidad son de tipo *float* y corresponden a las características de las imágenes digitalizadas de los núcleos celulares, para el caso del id del registro que es de tipo *int* y la diagnosis que es de tipo *string* son los que tienen tipo de dato diferente. Para entrenar los modelos es necesario que los datos pasen por un pre-proceso que convierte los datos en valores enteros y de esta manera hace que el procesamiento de los diferentes modelos se realiza de forma más eficiente, para ello se utiliza la función *LabelEncoder* de la librería *preprocessing* de *sklearn*. Para este caso la columna de diagnosis pasará de tener las letras *M* y *B* a los números *1* y *0* respectivamente.

```
[In]: # Convertir los valores "B" y "M" en valores enteros 0 y 1
labelencoder_Y = LabelEncoder()
df.iloc[:,1]= labelencoder_Y.fit_transform(df.iloc[:,1].values)
print(labelencoder_Y.fit_transform(df.iloc[:,1].values))
```

3.6. Modelos de Machine Learning

Los modelos de Machine Learning utilizados para el diagnostico de cáncer de mama son del tipo *Aprendizaje Supervisado* (*Supervised Learning*). Este tipo de aprendizaje se realiza teniendo conocimiento previo de cuáles deberían ser los valores de salida para nuestras muestras. Por lo tanto, su objetivo es aprender una función que, dada una muestra de datos y resultados deseados, se aproxime mejor a la relación entre entrada y salida observable en los datos[45]. Para la implementación técnica se utiliza la librería creada por Google para el lenguaje de programación *Python* llamada *Scikit learn* la cual contiene un conjunto de funciones para aplicar las técnicas de Machine Learning. Para efectos de demostrar el funcionamiento de los Modelos de Machine Learning aplicados al diagnostico de cáncer de mama se va a realizar una prueba con variables de dos pacientes. Los modelos utilizados para el diagnostico de cáncer de mama son: *Logistic Regression*, *Decision Trees*, *Gaussian Naive Bayes*, *K-Nearest Neighbors* (KNN), *Random Forest*, *Support Vector Machines* (SVM).

El *paciente 1* identificado con el id *8510824* esta ubicado en el Data-Set de cáncer de mama de la Universidad de Wisconsin en la Fila 21. De antemano conocemos que el *paciente 1* tiene un diagnostico benigno. Para efectos de las pruebas se toman los valores de la columna 2 a la columna 31. No se toma la primera columna debido a que en ella se encuentra el diagnostico generado por defecto, y para demostrar que los algoritmos estiman correctamente el valor maligno o benigno esta columna no se va a tener en cuenta.

```
[In]: # paciente 1 id= 8510824  diagnosis= B
      paciente_1 = df.iloc[21, 2:31].values
      # Visualizar las variables del paciente 1
      print("datos paciente 1 : \n", paciente_1)
```

```
[Out]: datos paciente 1 :
      [9.504 12.44 60.34 273.9 0.1024 0.06492 0.029560000000000003 0.02076
      0.1815 0.06905 0.2773 0.9768 1.909 15.7 0.009606 0.01432 0.01985
      0.014209999999999999 0.02027 0.002968 10.23 15.66 65.13 314.9 0.1324
      0.1148 0.08867 0.062270000000000006 0.245]
```

El *paciente 2* identificado con el id *84358402* esta ubicado en el Data-Set de cáncer de mama de la Universidad de Wisconsin en la Fila 4. De antemano conocemos que el *paciente 1* tiene un diagnostico maligno. Para efectos de las pruebas se toman los valores de la columna 2 a la columna 31. No se toma la primera columna debido a que en ella se encuentra el diagnostico generado por defecto, y para demostrar que los algoritmos estiman correctamente el valor maligno o benigno esta columna no se va a tener en cuenta.

```
[In]: # paciente 2 id= 84358402 diagnosis= M
paciente_2 = df.iloc[4, 2:31].values
# Visualizar las variables del paciente 2
print("datos paciente 1 : \n", paciente_1)
```

```
[Out]: datos paciente 2 :
[20.29 14.34 135.1 1297.0 0.1003 0.1328 0.198 0.1043 0.1809
0.05882999999999999 0.7572 0.7813 5.438 94.44 0.01149 0.02461
0.05687999999999999 0.01885 0.01756 0.005115 22.54 16.67 152.2 1575.0
0.1374 0.205 0.4 0.1625 0.2364]
```

Una vez se conoce el valor de la columna *diagnosis* del conjunto de pacientes en el Data-Set, esta es transformada en valores enteros para que los modelos procesen la información de forma eficiente. Para este caso la columna de *diagnosis* pasará de tener las letras *M* y *B* a los números *1* y *0* respectivamente. Se asignan los datos de la columna 2 a la columna 31 en la variable *X* y los datos de la columna 1(diagnosis) en la variable *Y*. Luego se utiliza la función *train_test_split* de la librería *model_selection* de *sklearn* para asignar el 75 % de los datos para que los modelos aprendan y el 25 % para comprobar la efectividad de los mismos.

```
[In]: from sklearn.preprocessing import LabelEncoder
labelencoder_Y = LabelEncoder()
df.iloc[:,1]= labelencoder_Y.fit_transform(df.iloc[:,1].values)
```

```
[In]: # datos del Data-Set
X = df.iloc[:, 2:31].values
# label "diagnosis"
Y = df.iloc[:, 1].values
```

```
[In]: # Importar libreria para la separación de los datos de entrenamiento
from sklearn.model_selection import train_test_split
# Utilizar 25% de los Datos para Entrenar los Modelos
X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size =0.25,
random_state = 0)
```

3.6.1. Logistic Regression

La regresión logística, a pesar de su nombre, es un modelo lineal para la clasificación en lugar de ser una regresión. Esta técnica se ajusta a un modelo lineal con coeficientes definidos para minimizar la suma residual de cuadrados entre los objetivos observados en un conjunto de datos y los objetivos predichos por la aproximación lineal. En este modelo, las probabilidades que describen los posibles resultados se modelan utilizando la función *LogisticRegression* de la librería *model_selection* de *sklearn*.

Para poder realizar el diagnostico de cáncer de mama del *paciente 1* y el *paciente 2* se debe entrenar el modelo con las variables *X_train* y *Y_train* como resultado del proceso de asignar el 75 % de los datos para entrenamiento y 25 % de los datos asignados para la verificación del aprendizaje del modelo. Una vez el modelo fue entrenado se obtuvo un valor de *0.960093896713615* de precisión del aprendizaje del mismo.

```
[In]: #Importar librería para usar el Modelo Logistic Regression
from sklearn.linear_model import LogisticRegression
#Usar el modelo Logistic Regression
logisticRegression = LogisticRegression(random_state = 0)
#Entrenar el modelo Logistic Regression
logisticRegression.fit(X_train, Y_train)
#Imprimir la respuesta del Modelo Logistic Regression
print('Logistic Regression TrainingAccuracy:',
logisticRegression.score(X_train,Y_train))
```

```
[Out]: Logistic Regression Training Accuracy: 0.960093896713615
```

Una vez el modelo fue entrenado, se realiza el diagnostico con los datos del *paciente 1* el cual arroja un resultado **benigno** con una probabilidad del *0.999647190999448* y una probabilidad del *0.00035280900055200434* de **malignidad**.

```
[in]: pred_1 = logisticRegression.predict([paciente_1])
pred_proba_1 = logisticRegression.predict_proba([paciente_1])
print("Diagnosis: ", pred_1)
print("Probabilidad Benigno: ", pred_proba_1[0][0])
print("Probabilidad Maligno: ", pred_proba_1[0][1])
```

```
[Out]: Diagnosis:  [0]
Probabilidad Benigno:  0.999647190999448
Probabilidad Maligno:  0.00035280900055200434
```


Una vez el modelo fue entrenado, se realiza el diagnostico con los datos del *paciente 2* el cual arroja un resultado **maligno** con una probabilidad del 0.9999958989224104 y una probabilidad del $4.101077589635516e-06$ de **benignidad**.

```
[in]: pred_2 = logisticRegression.predict([paciente_2])
      pred_proba_2 = logisticRegression.predict_proba([paciente_2])
      print("Diagnosis: ", pred_2)
      print("Probabilidad Benigno: ", pred_proba_2[0][0])
      print("Probabilidad Maligno: ", pred_proba_2[0][1])
```

```
[Out]: Diagnosis: [1]
      Probabilidad Benigno: 4.101077589635516e-06
      Probabilidad Maligno: 0.9999958989224104
```

En la Figura 3.9 se puede observar la clasificación de las variables *radius_mean* y *texture_mean* con base en la relación con la diagnosis generada. En esta Figura el color azul representa el valor de variables que están ubicadas en una diagnosis Maligna , y en color anaranjado el valor de variables que están ubicadas en una diagnosis Benigna. Estas características fueron seleccionadas entre 32 características de los datos extraídos por medio de la muestra FNA para evidenciar el funcionamiento del modelo de Machine Learning.

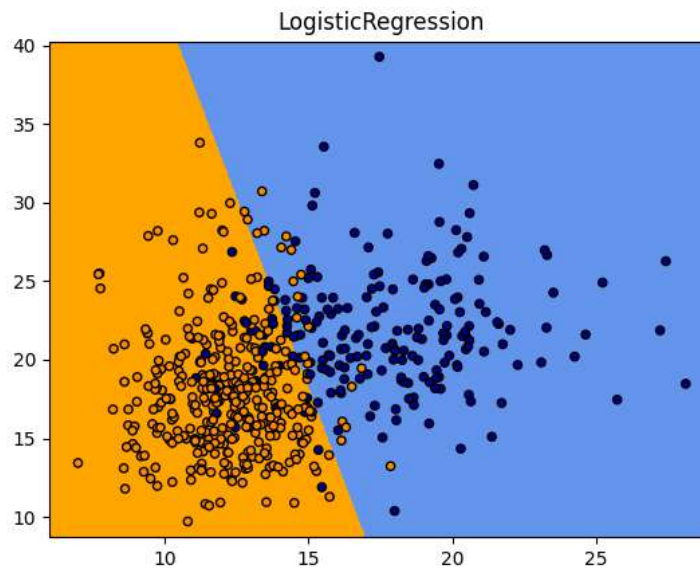


Figura 3.9: Clasificación realizada por el método Logistic Regression

3.6.2. Decision Trees

Los árboles de decisión tienen como objetivo crear un modelo que prediga el valor de una variable mediante el aprendizaje de reglas de decisión simples inferidas de las características de los datos. En este modelo, las probabilidades que describen los posibles resultados se modelan utilizando la función *DecisionTreeClassifier* de la librería *tree* de *sklearn* la cual es una clase capaz de realiza una clasificación de varias clases un conjunto de datos [46].

Para poder realizar el diagnostico de cáncer de mama del *paciente 1* y el *paciente 2* se debe entrenar el modelo con las variables *X_train* y *Y_train* como resultado del proceso de asignar el 75 % de los datos para entrenamiento y 25 % de los datos asignados para la verificación del aprendizaje del modelo. Una vez el modelo fue entrenado se obtuvo un valor de *1.0* de precisión del aprendizaje del mismo.

```
[In]: #Importar librería para usar el Modelo Decision Tree
from sklearn.tree import DecisionTreeClassifier
#Usar el modelo Decision Tree
decisionTreeClassifier = DecisionTreeClassifier(criterion = 'entropy',
                                              random_state = 0)
#Entrenar el modelo Decision Tree
decisionTreeClassifier.fit(X_train, Y_train)
#Imprimir la respuesta del Modelo Decision Trees
print('Decision Tree Classifier Training Accuracy:',
      decisionTreeClassifier.score(X_train,Y_train))
```

```
[Out]: Decision Tree Classifier Training Accuracy: 1.0
```

Una vez el modelo fue entrenado, se realiza el diagnostico con los datos del *paciente 1* el cual arroja un resultado **benigno** con una probabilidad del *1.0* y una probabilidad del *0.0* de **malignidad**.

```
[in]: pred_1 = decisionTreeClassifier.predict([paciente_1])
pred_proba_1 = decisionTreeClassifier.predict_proba([paciente_1])

print("Diagnosis: ", pred_1)
print("Probabilidad Benigno: ", pred_proba_1[0][0])
print("Probabilidad Maligno: ", pred_proba_1[0][1])
```

```
[Out]: Diagnosis:  [0]
Probabilidad Benigno:  1.0
Probabilidad Maligno:  0.0
```

Una vez el modelo fue entrenado, se realiza el diagnostico con los datos del *paciente 2* el cual arroja un resultado **maligno** con una probabilidad del *1.0* y una probabilidad del *0.0* de **benignidad**.

```
[in]: pred_2 = decisionTreeClassifier.predict([paciente_2])
      pred_proba_2 = decisionTreeClassifier.predict_proba([paciente_2])
      print("Diagnosis: ", pred_2)
      print("Probabilidad Benigno: ", pred_proba_2[0][0])
      print("Probabilidad Maligno: ", pred_proba_2[0][1])
```

```
[Out]: Diagnosis: [1]
      Probabilidad Benigno: 0.0
      Probabilidad Maligno: 1.0
```

En la Figura 3.10 se puede observar la clasificación de las variables *radius_mean* y *texture_mean* con base en la relación con la diagnosis generada. En esta Figura el color azul representa el valor de variables que están ubicadas en una diagnosis Maligna , y en color anaranjado el valor de variables que están ubicadas en una diagnosis Benigna. Estas características fueron seleccionadas entre 32 características de los datos extraídos por medio de la muestra FNA para evidenciar el funcionamiento del modelo de Machine Learning.

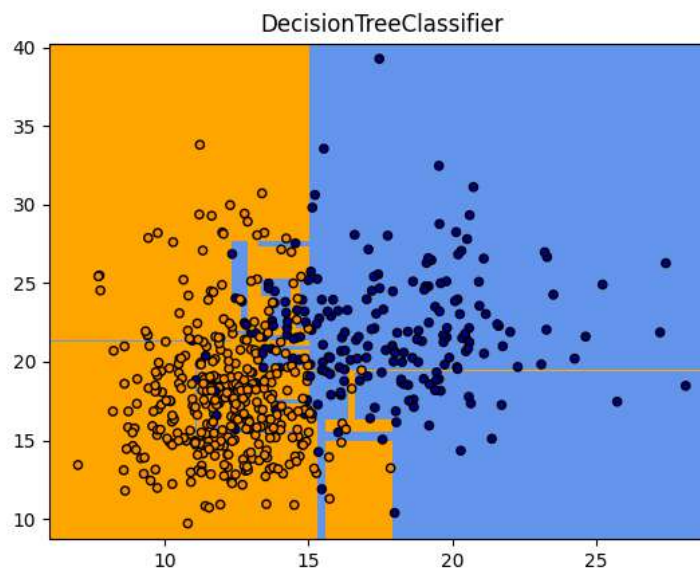


Figura 3.10: Clasificación realizada por el método Decision Trees

3.6.3. Random Forest

Un clasificador de bosque aleatorio es un meta estimador que se ajusta a varios clasificadores de árbol de decisión en varias submuestras de un conjunto de datos que utiliza el promedio para mejorar la precisión predictiva y controlar el sobreajuste.[46]. En este modelo, las probabilidades que describen los posibles resultados se modelan utilizando la función *RandomForestClassifier* de la librería *ensemble* de *sklearn*.

Para poder realizar el diagnostico de cáncer de mama del *paciente 1* y el *paciente 2* se debe entrenar el modelo con las variables *X_train* y *Y_train* como resultado del proceso de asignar el 75% de los datos para entrenamiento y 25% de los datos asignados para la verificación del aprendizaje del modelo. Una vez el modelo fue entrenado se obtuvo un valor de *0.9953051643192489* de precisión del aprendizaje del mismo.

```
[In]: #Importar librería para usar el Modelo RandomForestClassifier
from sklearn.ensemble import RandomForestClassifier
#Usar el modelo RandomForestClassifier
randomForestClassifier = RandomForestClassifier(n_estimators = 10,
criterion = 'entropy', random_state = 0)
#Entrenar el modelo RandomForestClassifier
randomForestClassifier.fit(X_train, Y_train)
#Imprimir la respuesta del Modelo RandomForestClassifier
print('Random Forest Classifier Training Accuracy:',
randomForestClassifier.score(X_train,Y_train))
```

```
[Out]: Random Forest Classifier Training Accuracy: 0.9953051643192489
```

Una vez el modelo fue entrenado, se realiza el diagnostico con los datos del *paciente 1* el cual arroja un resultado **benigno** con una probabilidad del *1.0* y una probabilidad del *0.0* de **malignidad**.

```
[in]: pred_1 = randomForestClassifier.predict([paciente_1])
pred_proba_1 = randomForestClassifier.predict_proba([paciente_1])
print("Diagnosis: ", pred_1)
print("Probabilidad Benigno: ", pred_proba_1[0][0])
print("Probabilidad Maligno: ", pred_proba_1[0][1])
```

```
[Out]: Diagnosis:  [0]
Probabilidad Benigno:  1.0
Probabilidad Maligno:  0.0
```

Una vez el modelo fue entrenado, se realiza el diagnostico con los datos del *paciente 2* el cual arroja un resultado **maligno** con una probabilidad del *1.0* y una probabilidad del *0.0* de **benignidad**.

```
[in]: pred_2 = randomForestClassifier.predict([paciente_2])
      pred_proba_2 = randomForestClassifier.predict_proba([paciente_2])
      print("Diagnosis: ", pred_2)
      print("Probabilidad Benigno: ", pred_proba_2[0][0])
      print("Probabilidad Maligno: ", pred_proba_2[0][1])
```

```
[Out]: Diagnosis: [1]
      Probabilidad Benigno: 0.0
      Probabilidad Maligno: 1.0
```

En la Figura 3.11 se puede observar la clasificación de las variables *radius_mean* y *texture_mean* con base en la relación con la diagnosis generada. En esta Figura el color azul representa el valor de variables que están ubicadas en una diagnosis Maligna , y en color anaranjado el valor de variables que están ubicadas en una diagnosis Benigna. Estas características fueron seleccionadas entre 32 características de los datos extraídos por medio de la muestra FNA para evidenciar el funcionamiento del modelo de Machine Learning.

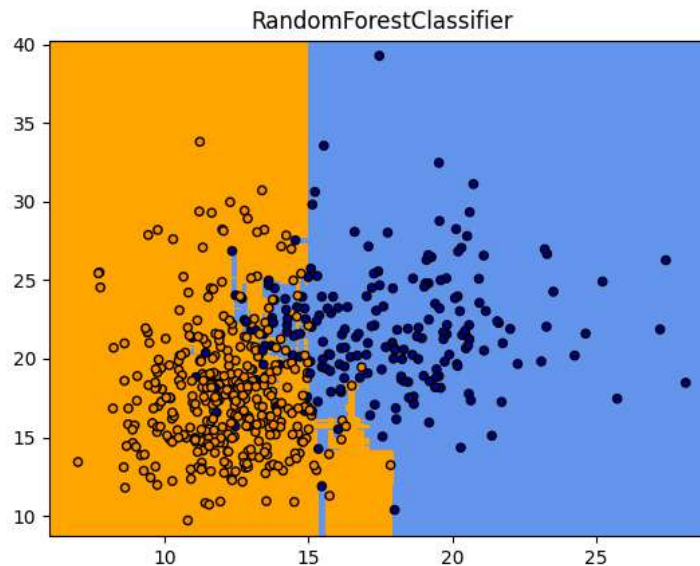


Figura 3.11: Clasificación realizada por el método Random Forest

3.6.4. Gaussian Naive Bayes

El Clasificador Bayesiano ingenuo Gaussiano es un método genérico de aprendizaje supervisado diseñado para resolver problemas de regresión y clasificación probabilística. La predicción realizada por el método es probabilística, de modo que se pueden calcular intervalos de confianza conocidos y decidir en función de si se debe reajustar la predicción en alguna región de interés [46]. En este modelo, las probabilidades que describen los posibles resultados se modelan utilizando la función *GaussianNB* de la librería *naive_bayes* de *sklearn*.

Para poder realizar el diagnóstico de cáncer de mama del *paciente 1* y el *paciente 2* se debe entrenar el modelo con las variables *X_train* y *Y_train* como resultado del proceso de asignar el 75 % de los datos para entrenamiento y 25 % de los datos asignados para la verificación del aprendizaje del modelo. Una vez el modelo fue entrenado se obtuvo un valor de *0.9507042253521126* de precisión del aprendizaje del mismo.

```
[In]: #Importar librería para usar el Modelo Gaussian Naive Bayes
from sklearn.naive_bayes import GaussianNB
#Usar el modelo Gaussian Naive Bayes
gaussianNB = GaussianNB()
#Entrenar el modelo Gaussian Naive Bayes
gaussianNB.fit(X_train, Y_train)
#Imprimir la respuesta del Modelo Gaussian Naive Bayes
print('Gaussian Naive Bayes Training Accuracy:',
gaussianNB.score(X_train, Y_train))
```

```
[Out]: Gaussian Naive Bayes Training Accuracy: 0.9507042253521126
```

Una vez el modelo fue entrenado, se realiza el diagnóstico con los datos del *paciente 1* el cual arroja un resultado **benigno** con una probabilidad del *0.9999999999999998* y una probabilidad del *2.1109106561728282e-16* de **malignidad**.

```
[in]: pred_1 = gaussianNB.predict([paciente_1])
pred_proba_1 = gaussianNB.predict_proba([paciente_1])
print("Diagnosis: ", pred_1)
print("Probabilidad Benigno: ", pred_proba_1[0][0])
print("Probabilidad Maligno: ", pred_proba_1[0][1])
```

```
[Out]: Diagnosis: [0]
Probabilidad Benigno: 0.9999999999999998
Probabilidad Maligno: 2.1109106561728282e-16
```

Una vez el modelo fue entrenado, se realiza el diagnostico con los datos del *paciente 2* el cual arroja un resultado **maligno** con una probabilidad del *1.0* y una probabilidad del *2.006997403407899e-61* de **benignidad**.

```
[in]: pred_2 = gaussianNB.predict([paciente_2])
      pred_proba_2 = gaussianNB.predict_proba([paciente_2])
      print("Diagnosis: ", pred_2)
      print("Probabilidad Benigno: ", pred_proba_2[0][0])
      print("Probabilidad Maligno: ", pred_proba_2[0][1])
```

```
[Out]: Diagnosis: [1]
      Probabilidad Benigno: 2.006997403407899e-61
      Probabilidad Maligno: 1.0
```

En la Figura 3.12 se puede observar la clasificación de las variables *radius_mean* y *texture_mean* con base en la relación con la diagnosis generada. En esta Figura el color azul representa el valor de variables que están ubicadas en una diagnosis Maligna , y en color anaranjado el valor de variables que están ubicadas en una diagnosis Benigna. Estas características fueron seleccionadas entre 32 características de los datos extraídos por medio de la muestra FNA para evidenciar el funcionamiento del modelo de Machine Learning.

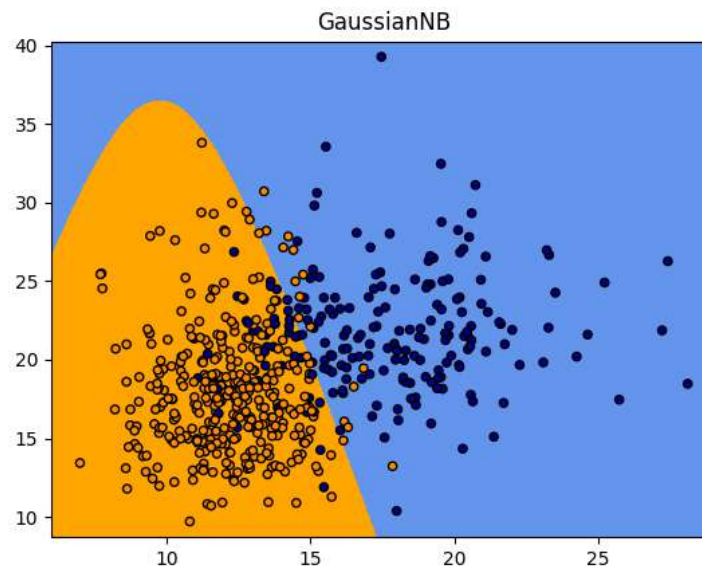


Figura 3.12: Clasificación realizada por el método Gaussian Naive Bayes

3.6.5. K-Nearest Neighbors(KNN)

El clasificador basado en vecinos cercanos es un modelo supervisado cuya función es encontrar un número predefinido de muestras de entrenamiento en una distancia cercana a un nuevo punto y predecir un valor a partir de ellas. La distancia puede ser, en general, cualquier medida métrica, por lo general la distancia euclidiana estándar es la opción más común. Este método es conocido como no generalizado, ya que simplemente recuerda todos sus datos de entrenamiento [46]. En este modelo, las probabilidades que describen los posibles resultados se modelan utilizando la función *KNeighborsClassifier* de la librería *neighbors* de *sklearn*.

Para poder realizar el diagnostico de cáncer de mama del *paciente 1* y el *paciente 2* se debe entrenar el modelo con las variables *X_train* y *Y_train* como resultado del proceso de asignar el 75% de los datos para entrenamiento y 25% de los datos asignados para la verificación del aprendizaje del modelo. Una vez el modelo fue entrenado se obtuvo un valor de *0.9413145539906104* de precisión del aprendizaje del mismo.

```
[In]: #Importar librería para usar el Modelo K-Nearest Neighbors
from sklearn.neighbors import KNeighborsClassifier
#Usar el modelo K-Nearest Neighbors
kNeighborsClassifier = KNeighborsClassifier(n_neighbors = 5,
                                          metric = 'minkowski', p = 2)
#Entrenar el modelo K-Nearest Neighbors
kNeighborsClassifier.fit(X_train, Y_train)
#Imprimir la respuesta del Modelo K-Nearest Neighbors
print('K-Nearest Neighbors Training Accuracy:',
      kNeighborsClassifier.score(X_train,Y_train))
```

```
[Out]: K-Nearest Neighbors Training Accuracy: 0.9413145539906104
```

Una vez el modelo fue entrenado, se realiza el diagnostico con los datos del *paciente 1* el cual arroja un resultado **benigno** con una probabilidad del *1.0* y una probabilidad del *0.0* de **malignidad**.

```
[in]: pred_1 = kNeighborsClassifier.predict([paciente_1])
      pred_proba_1 = kNeighborsClassifier.predict_proba([paciente_1])
      print("Diagnosis: ", pred_1)
      print("Probabilidad Benigno: ", pred_proba_1[0][0])
      print("Probabilidad Maligno: ", pred_proba_1[0][1])
```

```
[Out]: Diagnosis:  [0]
      Probabilidad Benigno:  1.0
      Probabilidad Maligno:  0.0
```

Una vez el modelo fue entrenado, se realiza el diagnostico con los datos del *paciente 2* el cual arroja un resultado **maligno** con una probabilidad del *1.0* y una probabilidad del *0.0* de **benignidad**.

```
[in]: pred_2 = kNeighborsClassifier.predict([paciente_2])
      pred_proba_2 = kNeighborsClassifier.predict_proba([paciente_2])
      print("Diagnosis: ", pred_2)
      print("Probabilidad Benigno: ", pred_proba_2[0][0])
      print("Probabilidad Maligno: ", pred_proba_2[0][1])
```

```
[Out]: Diagnosis:  [1]
      Probabilidad Benigno:  0.0
      Probabilidad Maligno:  1.0
```


En la Figura 3.13 se puede observar la clasificación de las variables *radius_mean* y *texture_mean* con base en la relación con la diagnosis generada. En esta Figura el color azul representa el valor de variables que están ubicadas en una diagnosis Maligna , y en color anaranjado el valor de variables que están ubicadas en una diagnosis Benigna. Estas características fueron seleccionadas entre 32 características de los datos extraídos por medio de la muestra FNA para evidenciar el funcionamiento del modelo de Machine Learning.

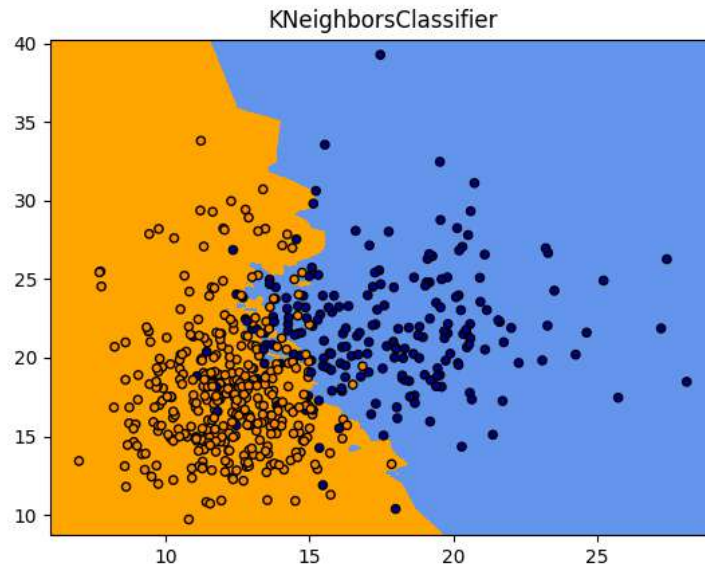


Figura 3.13: Clasificación realizada por el método K-Nearest Neighbors(KNN)

3.6.6. Support Vector Machines(SVM)

Las máquinas de vectores de soporte (SVM) son un conjunto de métodos de aprendizaje supervisado utilizados para la clasificación, regresión y detección de valores atípicos. Esta técnica utiliza un subconjunto de puntos de entrenamiento en la función de decisión (llamados vectores de soporte), por lo que es eficiente en memoria. Es efectivo en casos donde el número de dimensiones es mayor que el número de muestras [46]. En este modelo, las probabilidades que describen los posibles resultados se modelan utilizando la función *SVC* de la librería *svm* de *sklearn*.

Para poder realizar el diagnostico de cáncer de mama del *paciente 1* y el *paciente 2* se debe entrenar el modelo con las variables *X_train* y *Y_train* como resultado del proceso de asignar el 75% de los datos para entrenamiento y 25% de los datos asignados para la verificación del aprendizaje del modelo. Una vez el modelo fue entrenado se obtuvo un valor de *0.9647887323943662* de precisión del aprendizaje del mismo.

```
[In]: #Importar librería para usar el Modelo Support Vector Machines
      from sklearn.svm import SVC
      #Usar el modelo Support Vector Machines
      svc = SVC(kernel = 'linear', random_state = 0, probability=True)
      #Entrenar el modelo Support Vector Machines
      svc.fit(X_train, Y_train)
      #Imprimir la respuesta del Modelo Support Vector Machines
      print('Support Vector Machine Training Accuracy:',
            svc.score(X_train,Y_train))
```

```
[Out]: Support Vector Machine Training Accuracy:0.9647887323943662
```

Una vez el modelo fue entrenado, se realiza el diagnostico con los datos del *paciente 1* el cual arroja un resultado **benigno** con una probabilidad del *0.9990962732841208* y una probabilidad del *0.0009037267158791459* de **malignidad**.

```
[in]: pred_1 = svc.predict([paciente_1])
      pred_proba_1 = svc.predict_proba([paciente_1])
      print("Diagnosis: ", pred_1)
      print("Probabilidad Benigno: ", pred_proba_1[0][0])
      print("Probabilidad Maligno: ", pred_proba_1[0][1])
```

```
[Out]: Diagnosis:  [0]
      Probabilidad Benigno:  0.9990962732841208
      Probabilidad Maligno:  0.0009037267158791459
```

Una vez el modelo fue entrenado, se realiza el diagnostico con los datos del *paciente 2* el cual arroja un resultado **maligno** con una probabilidad del *0.9933163476063908* y una probabilidad del *0.0066836523936092546* de **benignidad**.

```
[in]: pred_2 = svc.predict([paciente_2])
      pred_proba_2 = svc.predict_proba([paciente_2])
      print("Diagnosis: ", pred_2)
      print("Probabilidad Benigno: ", pred_proba_2[0][0])
      print("Probabilidad Maligno: ", pred_proba_2[0][1])
```

```
[Out]: Diagnosis: [1]
      Probabilidad Benigno: 0.0066836523936092546
      Probabilidad Maligno: 0.9933163476063908
```

En la Figura 3.14 se puede observar la clasificación de las variables *radius_mean* y *texture_mean* con base en la relación con la diagnosis generada. En esta Figura el color azul representa el valor de variables que están ubicadas en una diagnosis Maligna , y en color anaranjado el valor de variables que están ubicadas en una diagnosis Benigna. Estas características fueron seleccionadas entre 32 características de los datos extraídos por medio de la muestra FNA para evidenciar el funcionamiento del modelo de Machine Learning.

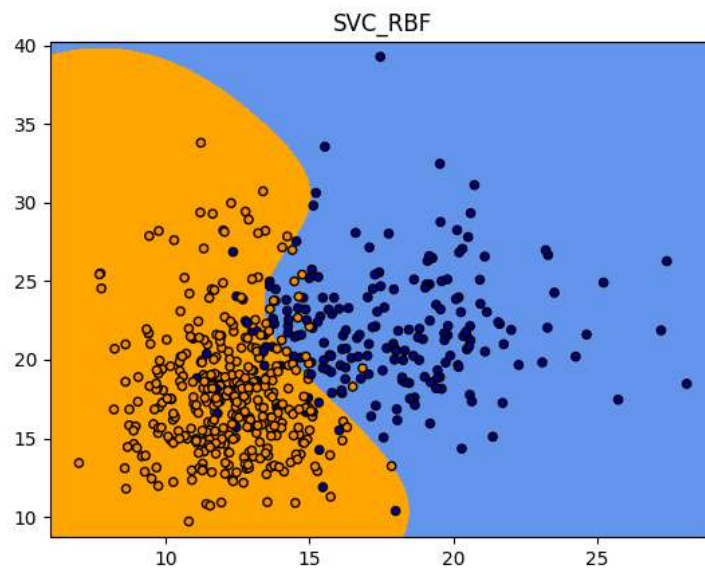


Figura 3.14: Clasificación realizada por el método Support Vector Machines(SVM)

3.7. BreastApp

La aplicación web BreastApp se diseñó desde la perspectiva de la programación orientada a servicios, donde se destaca la creación de un API REST (Back-End) que expone los servicios de los que va a consumir la aplicación web (Front-End).

3.7.1. Back-End BreastApp

El Back-End de la aplicación fue construido con el lenguaje de programación Python, apoyado en el framework Django REST Framework, que ofrece un marco de trabajo completo para la implementación de aplicaciones de este tipo y facilita la exposición de servicios web sobre Python. A continuación se describen los servicios implementados.

Servicio Data-Set

Este servicio permite cargar el Data-Set y almacenarlo en el servidor. La petición debe contener el archivo .csv que corresponde al parámetro *dtsUrlDataSet* y el nombre que se asignará en el registro que corresponde al parámetro *dtsName*. La respuesta de la petición retorna el id asignado al Data-Set, el nombre del Data-Set, la url donde está almacenado el Data-Set en el servidor, la fecha de creación y la última fecha de edición del Data-Set. La información del diseño de este servicio puede ser observado en la Tabla 3.1.

Url:	/Data-Set/
Request	
Parametro	Descripción
dtsName	Nombre del Data-Set
dtsUrlDataSet	Archivo .csv con información del Data-Set
Response	
Parametro	Descripción
id	Id asociado al Data-Set
dtsName	Nombre del Data-Set
dtsUrlDataSet	Url donde se almacena el Data-Set
dtsCreatedAt	Fecha de creación del registro
dtsUpdateAt	Fecha de actualización del registro

Tabla 3.1: Diseño servicio REST Data-Set

A continuación se puede observar el diseño de los métodos que tiene el servicio para realizar las operaciones según su funcionamiento:

```
[Info]: #Información de la entidad Data-Set
{
  "id": {
    "type": "integer",
    "required": false,
    "read_only": true,
    "label": "ID"
  },
  "dtsName": {
    "type": "string",
    "required": true,
    "read_only": false,
    "label": "DtsName",
    "max_length": 255
  },
  "dtsUrlDataSet": {
    "type": "file upload",
    "required": true,
    "read_only": false,
    "label": "DtsUrlDataSet",
    "max_length": 100
  },
  "dtsCreatedAt": {
    "type": "datetime",
    "required": false,
    "read_only": true,
    "label": "DtsCreatedAt"
  },
  "dtsUpdatedAt": {
    "type": "datetime",
    "required": false,
    "read_only": true,
    "label": "DtsUpdatedAt"
  }
}
```

[Request]: *#Data-Set Request*

```
{  
  "dtsName": "",  
  "dtsUrlDataSet":  
}
```

[Response]: *#Data-Set Response*

```
{  
  "id": 36,  
  "dtsName": "Forest Data-Set",  
  "dtsUrlDataSet": "http://127.0.0.1:8000/Data-Set/datasets/data_g7dls.csv",  
  "dtsCreatedAt": "2020-05-14T23:10:13.641167-05:00",  
  "dtsUpdatedAt": "2020-05-14T23:10:13.641210-05:00"  
}
```

Servicio ModelMI

Este servicio permite almacenar modelos de Machine Learning en el servidor. La petición debe contener el nombre del modelo que corresponde al parámetro *modName* y el nombre corto del modelo en el parámetro *modShortName*. La respuesta de la petición retorna el id asignado al modelo, el nombre del modelo, la fecha de creación y la última fecha de edición del modelo. La información del diseño de este servicio puede ser observado en la Tabla 3.2.

Url:	/modelml/
Request	
Parametro	Descripción
modName	Nombre del modelo
modShortName	Nombre corto del modelo
Response	
Parametro	Descripción
id	Id asociado al modelo
modName	Nombre del modelo
modShortName	Nombre corto del modelo
modCreatedAt	Fecha de creación del registro
modUpdatedAt	Fecha de actualización del registro

Tabla 3.2: Diseño servicio REST ModelMI

A continuación se puede observar el diseño de los métodos que tiene el servicio para realizar las operaciones según su funcionamiento:

```
[Info]: #Información de la entidad ModelMl
{
  "id": {
    "type": "integer",
    "required": false,
    "read_only": true,
    "label": "ID"
  },
  "modName": {
    "type": "string",
    "required": true,
    "read_only": false,
    "label": "ModName",
    "max_length": 255
  },
  "modShortName": {
    "type": "string",
    "required": false,
    "read_only": false,
    "label": "ModShortName",
    "max_length": 255
  },
  "modCreatedAt": {
    "type": "datetime",
    "required": false,
    "read_only": true,
    "label": "ModCreatedAt"
  },
  "modUpdatedAt": {
    "type": "datetime",
    "required": false,
    "read_only": true,
    "label": "ModUpdatedAt"
  }
}
```


[Request]: *#ModelMl Request*

```
{  
  "modelName": "",  
  "modShortName": ""  
}
```

[Response]: *#ModelMl Response*

```
{  
  "id": 21,  
  "modelName": "LogisticRegression",  
  "modShortName": "log",  
  "modCreatedAt": "2020-05-14T23:22:46.853379-05:00",  
  "modUpdatedAt": "2020-05-14T23:22:46.853416-05:00"  
}
```

Servicio Prediction

Este servicio permite realizar el diagnóstico de malignidad o benignidad, según los datos ingresados de cada paciente. La petición debe contener el archivo .csv que corresponde al parámetro *preDataToPredict* este contiene los datos de cada uno de los pacientes a los que se quiere generar el diagnóstico. El parámetro *training* debe contener el id del entrenamiento con el que se va a realizar el diagnóstico. La respuesta de la petición retorna el id asignado a la predicción, la fecha de creación y la última fecha de edición de la predicción, la url donde se almacenaron los datos de los pacientes a quienes se le realizó la predicción, el id del entrenamiento, el id del paciente diagnosticado y un arreglo con las predicciones realizadas, que se componen por: el id de la predicción, el dni del paciente, el nombre corto del modelo, el diagnóstico generado y nombre del modelo con el que se realizó el diagnóstico. La información del diseño de este servicio puede ser observado en la Tabla 3.3.

Url: /prediction/	
Request	
Parametro	Descripción
preDataToPredict	Archivo .csv con los datos de los pacientes
training	Id del entrenamiento
patient	Id del paciente
Response	
Parametro	Descripción
id	Id asociado a la predicción
preCreatedAt	Fecha de creación del registro
preUpdateAt	Fecha de actualización del registro
preDataToPredict	Archivo .csv con los datos de los pacientes
training	Id del entrenamiento
patient	Id del paciente

Tabla 3.3: Diseño servicio REST Prediction

A continuación se puede observar el diseño de los métodos que tiene el servicio para realizar las operaciones según su funcionamiento:

[Info]: *#Información de la entidad Prediction*

```
{
  "id": {
    "type": "integer",
    "required": false,
    "read_only": true,
    "label": "ID"
  },
  "preCreatedAt": {
    "type": "datetime",
    "required": false,
    "read_only": true,
    "label": "PreCreatedAt"
  },
  "preUpdatedAt": {
    "type": "datetime",
    "required": false,
    "read_only": true,
    "label": "PreUpdatedAt"
  },
  "preDataToPredict": {
    "type": "file upload",
    "required": true,
    "read_only": false,
    "label": "PreDataToPredict",
    "max_length": 100
  },
  "training": {
    "type": "field",
    "required": true,
    "read_only": false,
    "label": "Training"
  },
  "patient": {
    "type": "field",
    "required": false,
    "read_only": false,
    "label": "Patient"}}}
```

[Request]: *#Prediction Request*

```
{
  "preDataToPredict": null,
  "training": null
}
```

[Response]: *#Prediction Response*

```
{
  "data": {
    "id": 106,
    "preCreatedAt": "2020-05-14T23:36:48.158151-05:00",
    "preUpdatedAt": "2020-05-14T23:36:48.158227-05:00",
    "preDataToPredict": "prediction/data_to_predict/data_test.csv",
    "training": 1,
    "patient": 2
  },
  "prediction": [
    {
      "id": 3,
      "dni": 842302,
      "modShortName": "log",
      "label": 0,
      "modelMl": "LogisticRegression"
    },
    {
      "id": 3,
      "dni": 842302,
      "modShortName": "knn",
      "label": 1,
      "modelMl": "KNeighborsClassifier"
    },
    {
      "id": 3,
      "dni": 842302,
      "modShortName": "svc_lin",
      "label": 0,
      "modelMl": "SVC Linear"
    },
    {
      "id": 3,
      "dni": 842302,
```

```
        "modShortName": "svc_rbf",
        "label": 1,
        "modelMl": "SVC RBF"
    },
    {
        "id": 3,
        "dni": 842302,
        "modShortName": "gauss",
        "label": 1,
        "modelMl": "GaussianNB"
    },
    {
        "id": 3,
        "dni": 842302,
        "modShortName": "tree",
        "label": 1,
        "modelMl": "DecisionTreeClassifier"
    },
    {
        "id": 3,
        "dni": 842302,
        "modShortName": "forest",
        "label": 1,
        "modelMl": "RandomForestClassifier"
    },
    {
        "id": 3,
        "dni": 842302,
        "modShortName": "log",
        "label": 0,
        "modelMl": "LogisticRegression"}]
}
```

Servicio Training

Este servicio permite entrenar los modelos con el Data-set que se indique en la petición. La petición debe contener el id de Data-Set *Data-Set*. La respuesta a la petición retorna el id del entrenamiento, el id del Data-Set, el nombre del Data-Set, la fecha de creación de entrenamiento y un arreglo con el porcentaje de precisión de cada uno de los modelos entrenados. La información del diseño de este servicio puede ser observado en la Tabla 3.4.

Url: /training/	
Request	
Parametro	Descripción
Data-Set	Id del Data-Set
Response	
Parametro	Descripción
id	Id asociado al entrenamiento
Data-Set	Id del Data-Set
dataSetName	Nombre del Data-Set
traCreatedAt	Fecha de creación del registro
score	JSON con los valores asociados a la precisión

Tabla 3.4: Diseño servicio REST Training

A continuacion se puede observar el diseño de los métodos que tiene el servicio para realizar las operaciones según su funcionamiento:

[Info]: *#Información de la entidad Training*

```
{
  "id": {
    "type": "integer",
    "required": false,
    "read_only": true,
    "label": "ID"
  },
  "Data-Set": {
    "type": "field",
    "required": true,
    "read_only": false,
    "label": "Data-Set"
  },
  "dataSetName": {
    "type": "field",
    "required": false,
    "read_only": true,
    "label": "Datasetname"
  },
  "traScore": {
    "type": "string",
    "required": false,
    "read_only": false,
    "label": "TraScore",
    "max_length": 255
  },
  "traCreatedAt": {
    "type": "datetime",
    "required": false,
    "read_only": true,
    "label": "TraCreatedAt"
  }
}
```

```
[Request]: #Training Request
{
  "Data-Set": null
}
```

```
[Response]: #Training Response
{
  "data": {
    "id": 19,
    "Data-Set": 24,
    "dataSetName": "Prueba",
    "traScore": null,
    "traCreatedAt": "2020-05-14T23:39:59.011770-05:00"
  },
  "score": {
    "log": 0.9906103286384976,
    "knn": 0.9765258215962441,
    "svc_lin": 0.9882629107981221,
    "svc_rbf": 0.9835680751173709,
    "gauss": 0.9507042253521126,
    "tree": 1.0,
    "forest": 0.9953051643192489
  }
}
```


Servicio Patient

Este servicio permite crear pacientes en el sistema. La información del diseño de este servicio puede ser observado en la Tabla 3.5.

Url:	/patient/
Request	
Parametro	Descripción
patDNI	Identificación del paciente
patName	Nombre del paciente
patLastName	Apellido del paciente
Response	
Parametro	Descripción
id	Id asociado al paciente
patDNI	Identificación del paciente
patName	Nombre del paciente
patLastName	Apellido del paciente
dtsCreatedAt	Fecha de creación del registro
dtsUpdateAt	Fecha de actualización del registro

Tabla 3.5: Diseño servicio REST Patient

A continuación se puede observar el diseño de los métodos que tiene el servicio para realizar las operaciones según su funcionamiento:

```
[Info]: #Información de la entidad Patient
{
  "id": {
    "type": "integer",
    "required": false,
    "read_only": true,
    "label": "ID"
  },
  "patDNI": {
    "type": "string",
    "required": true,
    "read_only": false,
    "label": "PatDNI",
    "max_length": 255
  },
  "patName": {
    "type": "string",
    "required": true,
    "read_only": false,
    "label": "PatName",
    "max_length": 255
  },
  "patLastName": {
    "type": "string",
    "required": true,
    "read_only": false,
    "label": "PatLastName",
    "max_length": 255
  },
  "dtsCreatedAt": {
    "type": "datetime",
    "required": false,
    "read_only": true,
    "label": "DtsCreatedAt"
  },
  "dtsUpdatedAt": {
    "type": "datetime",
    "required": false,
```

```
        "read_only": true,  
        "label": "DtsUpdateAt"  
    }  
}
```

```
[Request]: {  
  "patDNI": "",  
  "patName": "",  
  "patLastName": ""  
}
```

```
[Response]: {  
  "id": 27,  
  "patDNI": "1012397483",  
  "patName": "Jaime",  
  "patLastName": "Robles Fajardo",  
  "dtsCreatedAt": "2020-05-14T23:48:16.998190-05:00",  
  "dtsUpdateAt": "2020-05-14T23:48:16.998240-05:00"  
}
```

Servicio Patient Prediction

Este servicio permite asociar a los pacientes con el diagnóstico realizado por cada uno de los modelos de Machine Learning. La información del diseño de este servicio puede ser observado en la Tabla 3.6.

Url: /patientprediction/	
Request	
Parametro	Descripción
patient	Id del paciente
prediction	Id de la predicción
ModelMl	Id del modelo de Machine Learning
patPreLabel	Etiqueta resultado del diagnóstico
Response	
Parametro	Descripción
patient	Id del paciente
prediction	Id de la predicción
ModelMl	Id del modelo de Machine Learning
patPreLabel	Etiqueta resultado del diagnóstico
traId	Id de entrenamiento
traScore	JSON con los valores asociados a la precisión
modelMl	Id del modelo de Machine Learning
modShortName	Nombre corto del modelo con el q
modName	Nombre Largo del modelo
patCreatedAt	Fecha de creación del registro
patUpdateAt	Fecha de actualización del registro
patPreLabel	Etiqueta resultado del diagnóstico

Tabla 3.6: Diseño servicio REST Patientprediction

A continuación se puede observar el diseño de los métodos que tiene el servicio para realizar las operaciones según su funcionamiento:

[Info]: *#Información de la entidad Patientprediction*

```
{
  "id": {
    "type": "integer",
    "required": false,
    "read_only": true,
    "label": "ID"
  },
  "patient": {
    "type": "field",
    "required": false,
    "read_only": false,
    "label": "Patient"
  },
  "prediction": {
    "type": "field",
    "required": false,
    "read_only": false,
    "label": "Prediction"
  },
  "traScore": {
    "type": "field",
    "required": false,
    "read_only": true,
    "label": "Trascore"
  },
  "modelMl": {
    "type": "field",
    "required": false,
    "read_only": false,
    "label": "ModelMl"
  },
  "modShortName": {
    "type": "field",
    "required": false,
    "read_only": true,
    "label": "Modshortname"
  },
}
```

```

        "modName": {
            "type": "field",
            "required": false,
            "read_only": true,
            "label": "Modname"
        },
        "patCreatedAt": {
            "type": "datetime",
            "required": false,
            "read_only": true,
            "label": "PatCreatedAt"
        },
        "patUpdatedAt": {
            "type": "datetime",
            "required": false,
            "read_only": true,
            "label": "PatUpdatedAt"
        },
        "patPreLabel": {
            "type": "string",
            "required": true,
            "read_only": false,
            "label": "PatPreLabel",
            "max_length": 255}}

```

```

[Request]: {"patient": null,
            "prediction": null,
            "modelMl": null}

```

```

[Response]: {
    "id": 5745,
    "patient": 3,
    "prediction": 106,
    "traScore": "{\"log\": 0.9420, \"knn\": 0.9472, \"svc_lin\": 0.9630}",
    "modelMl": 14,
    "modShortName": "log",
    "modName": "LogisticRegression",
    "patCreatedAt": "2020-05-14T23:49:47.410144-05:00",
    "patUpdatedAt": "2020-05-14T23:49:47.410177-05:00",
    "patPreLabel": "sdasd"
}

```

3.7.2. Front-End BreastApp

El Front-End de la aplicación, está implementado en Angular el cual es un framework que brinda una gran variedad de prestaciones para la construcción de interfaces de usuario con características como un diseño web adaptable y una experiencia de usuario satisfactoria gracias a una visualización agradable, sin contar con el buen rendimiento que ofrece para el consumo de servicios Web. Para la implementación y desarrollo de la aplicación Web BreastApp, se implementaron 5 vistas que conforman el flujo completo de la versión actual, que van desde Cargar y Entrenar el Data-Set, realizar el entrenamiento de los Moldeos en Machine Learning, Realizar el diagnóstico, generar el Reporte final y permitir descargar el Reporte final en PDF. Estas vistas se describen a continuación:

Gestor de Data-Sets

Esta vista tiene como objetivo permitir la subida de un Data-Set para luego ser entrenado por los diferentes modelos de Machine Learning. esta vista consume el *Servicio Data-Set*. En la Figura 3.15 se puede observar esta funcionalidad.

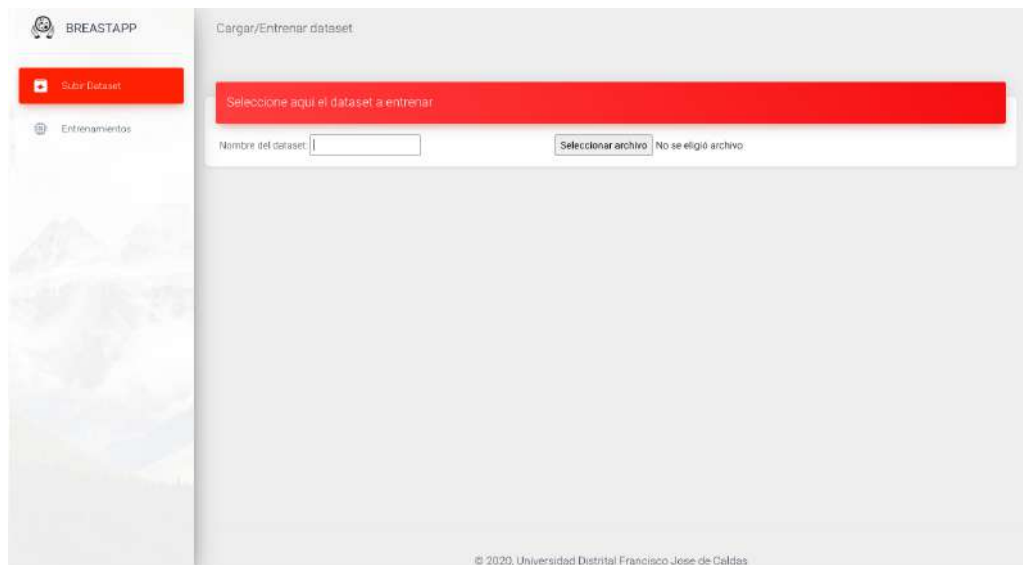
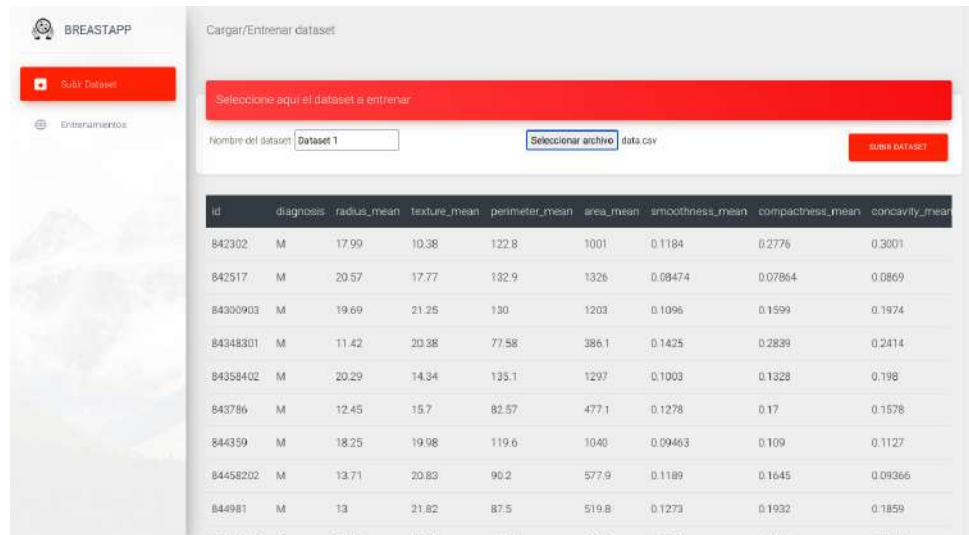


Figura 3.15: Vista para cargar Data-Sets para entrenamiento

Al ingresar el Data-set se puede observar la vista previa con la información que contiene el mismo. En la Figura 3.16 se puede observar esta funcionalidad.



Cargar/Entrenar dataset

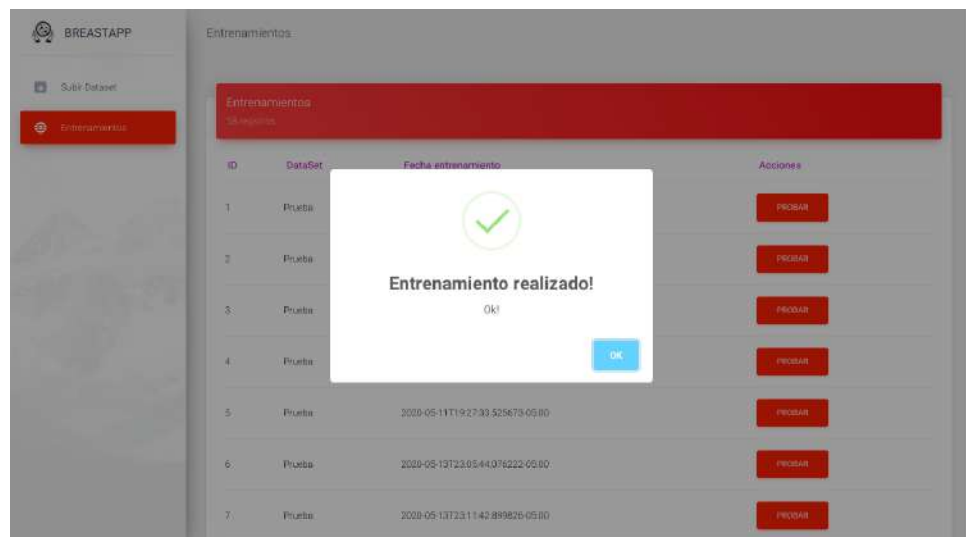
Seleccione aquí el dataset a entrenar

Nombre del dataset:

id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001
842517	M	20.57	17.77	182.9	1326	0.08474	0.07864	0.0869
8430903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859

Figura 3.16: Vista de Data-Sets cargados

Una vez se presiona el botón de subir Data-Set, un Pop-Up realiza la evidencia de que el entrenamiento realizado correctamente. En la Figura 3.17 se puede observar esta funcionalidad.



Entrenamientos

Entrenamientos 13 realizados

ID	DataSet	Fecha entrenamiento	Acciones
1	Prueba		<input type="button" value="PROBAR"/>
2	Prueba		<input type="button" value="PROBAR"/>
3	Prueba		<input type="button" value="PROBAR"/>
4	Prueba		<input type="button" value="PROBAR"/>
5	Prueba	2020-05-11T19:27:33.525673-05:00	<input type="button" value="PROBAR"/>
6	Prueba	2020-05-13T23:05:44.076222-05:00	<input type="button" value="PROBAR"/>
7	Prueba	2020-05-13T23:11:42.899826-05:00	<input type="button" value="PROBAR"/>

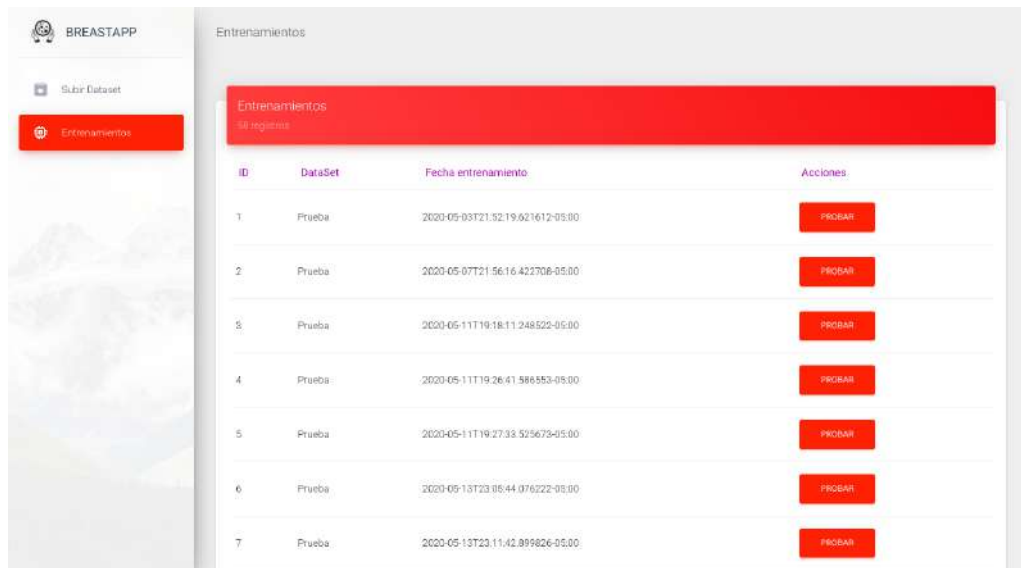
Entrenamiento realizado!

OK

Figura 3.17: Vista entrenamiento realizado correctamente

Gestor de Entrenamiento

Esta vista tiene como objetivo visualizar en una tabla los entrenamientos que ha llevado a cabo el sistema. Cada fila de la tabla tiene un botón que permite probar el entrenamiento. Para listar los entrenamientos esta vista consume el *Servicio Training*. En la Figura 3.18 se puede observar esta funcionalidad.



ID	DataSet	Fecha entrenamiento	Acciones
1	Prueba	2020-05-03T21:52:19.021612+05:00	PROBAR
2	Prueba	2020-05-07T21:56:16.423706+05:00	PROBAR
3	Prueba	2020-05-11T19:18:11.248522+05:00	PROBAR
4	Prueba	2020-05-11T19:26:01.586553+05:00	PROBAR
5	Prueba	2020-05-11T19:27:33.525673+05:00	PROBAR
6	Prueba	2020-05-13T23:05:44.076222+05:00	PROBAR
7	Prueba	2020-05-13T23:11:42.899826+05:00	PROBAR

Figura 3.18: Vista historial de entrenamientos

Gestor de Predicción

Esta vista tiene como objetivo subir y diagnosticar el archivo .csv que contiene los datos de los pacientes. Para subir el Data-Set y realizar el diagnóstico esta vista consume el *Servicio Prediction*. Al ingresar el archivo .csv con la información de los pacientes se muestra una vista previa de los datos de los pacientes a los que se le va a realizar el diagnóstico. En la Figura 3.19 y en la Figura 3.20 se puede observar estas funcionalidades.

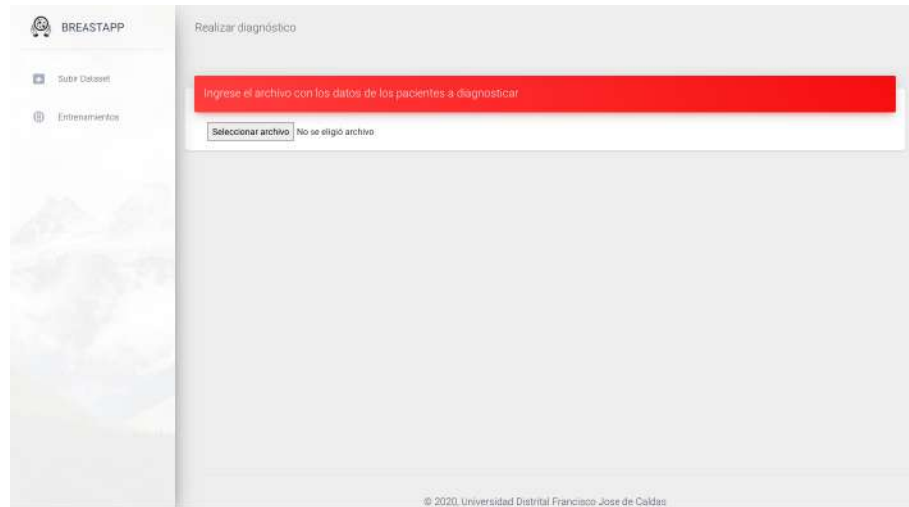


Figura 3.19: Vista carga de datos de pacientes a diagnosticar

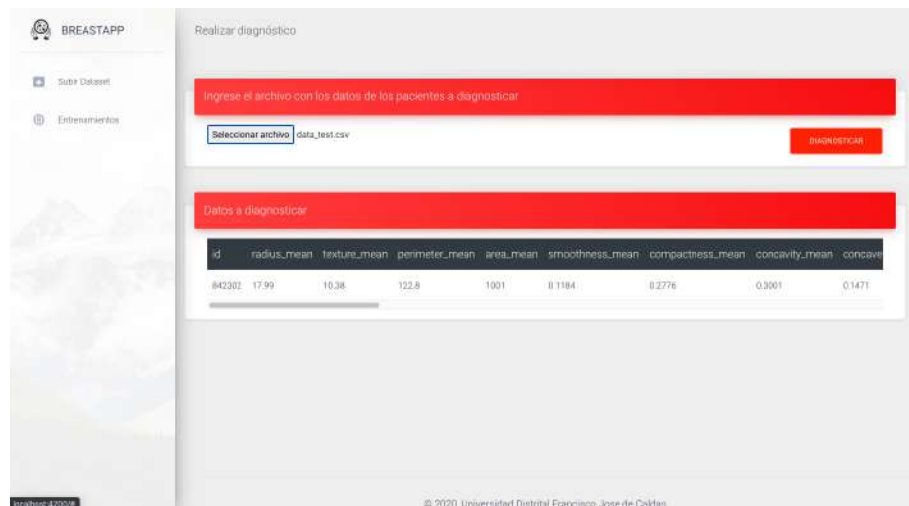


Figura 3.20: Vista de datos cargados de pacientes a diagnosticar

Cuando el diagnostico se ha realizado, se visualiza un Pop-Up avisando que el diagnóstico fue realizado correctamente. Para subir el Data-Set y realizar el diagnóstico esta vista consume el *Servicio ModelML*. En la Figura 3.21 se puede observar esta funcionalidad.

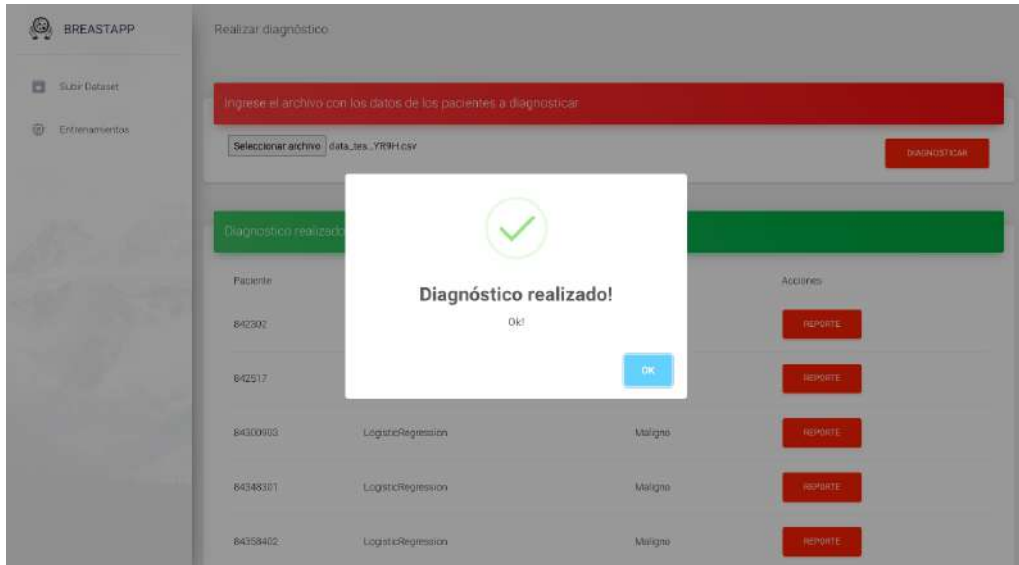
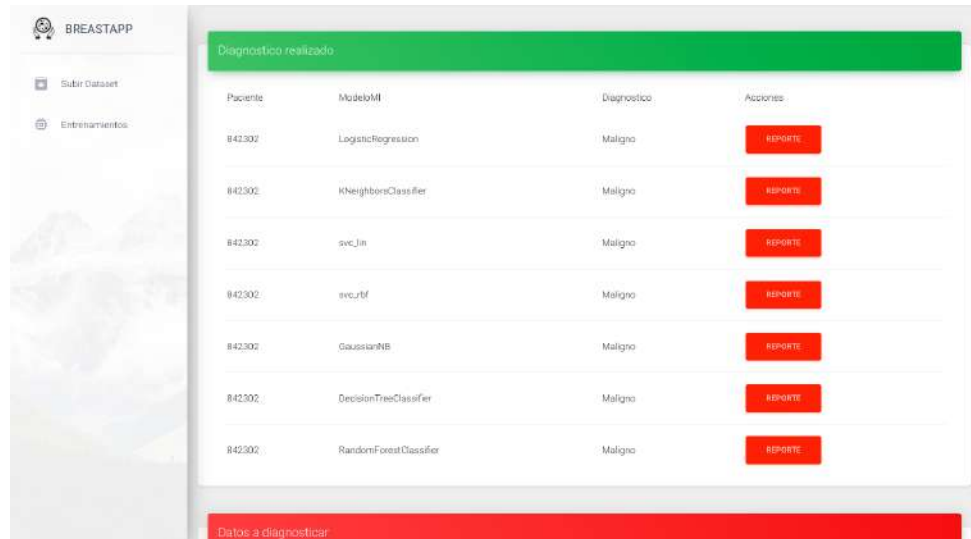


Figura 3.21: Vista de diagnostico realizado correctamente

Gestor de Modelos

En esta vista se pueden observar los diferentes diagnósticos realizados por los modelos de Machine Learning con los que cuenta el sistema. En la Figura 3.22 se puede observar esta funcionalidad.



The screenshot shows the BREASTAPP interface. On the left is a sidebar with a logo and two menu items: 'Subir Dataset' and 'Entrenamientos'. The main area has a green header 'Diagnostico realizado' and a red footer 'Datos a diagnosticar'. Between them is a table with four columns: 'Paciente', 'Modelo ML', 'Diagnostico', and 'Acciones'. The table contains seven rows, all with the patient ID '842302' and a 'Maligno' diagnosis. Each row has a red 'REPORTE' button in the 'Acciones' column.

Paciente	Modelo ML	Diagnostico	Acciones
842302	LogisticRegression	Maligno	REPORTE
842302	KNeighborsClassifier	Maligno	REPORTE
842302	svc_lin	Maligno	REPORTE
842302	svc_rbf	Maligno	REPORTE
842302	GaussianNB	Maligno	REPORTE
842302	DecisionTreeClassifier	Maligno	REPORTE
842302	RandomForestClassifier	Maligno	REPORTE

Figura 3.22: Vista modelos de Machine Learning utilizados en el diagnostico

Gestor de Informes

Esta vista tiene como objetivo mostrar una vista detallada del diagnóstico de un paciente en particular. En esta se puede ver cuales son los diagnósticos realizados a dicho paciente. Esta vista consume el *Servicio PatientPrediction*. En la Figura 3.23 se puede observar esta funcionalidad.

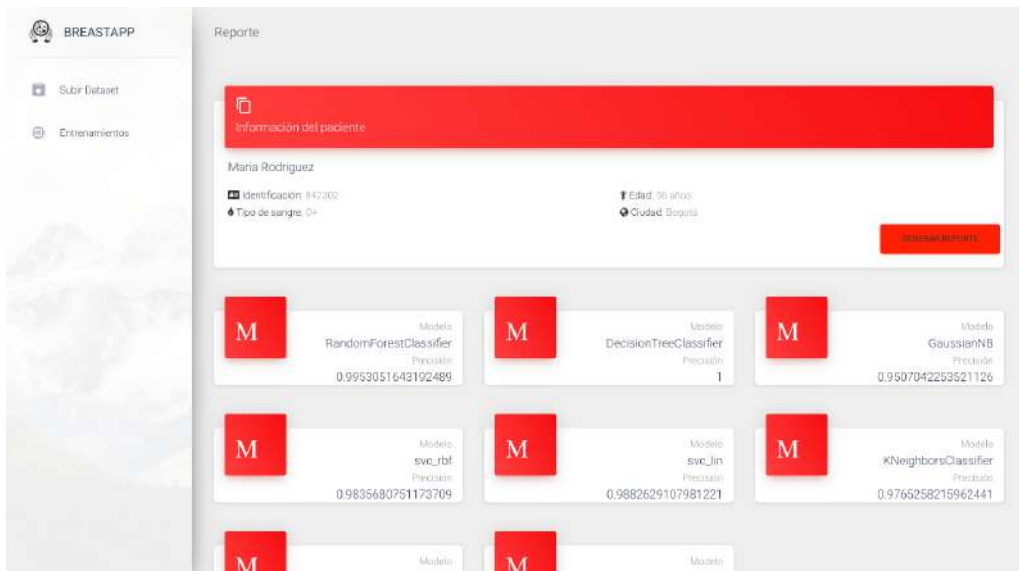


Figura 3.23: Vista diagnostico detallado del paciente

Esta vista cuenta con la opción de generar un reporte en formato.pdf en el que muestra información detallada del diagnóstico del usuario y un resumen gráfico con información de los modelos de Machine Learning. En la Figura 3.24 se puede observar esta funcionalidad.



Figura 3.24: Reporte detallado en pdf del diagnostico del paciente

Parte II

ARQUITECTURA Y DISEÑO

Capítulo 4

Organización Empresarial

La estructura organizativa planteada para el proyecto de investigación está basada de forma ideal en el *Área de Investigación de análisis de datos* de un *Instituto de Cancerología*.

4.1. Misión

Somos una institución que trabaja por el control integral del cáncer a través de la atención y el cuidado de pacientes, la investigación, la formación de talento humano y el desarrollo de acciones en salud pública.

4.2. Visión

En 2026 el Instituto de Cancerología será referente por sus logros en la reducción de la mortalidad por cáncer, sobre la base de la innovación y la tecnología, con un actuar ético y sostenible y con un talento humano motivado y comprometido.

4.3. Objetivo

La investigación en su formulación y desarrollo genera dudas o problemas de carácter estadístico que necesitan un enfoque experto. El Área de Investigación de Análisis de datos tiene por objetivo dar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología, en métodos para el manejo y análisis de datos, a través de las siguientes actividades:

- Proponer e implementar estrategias que permitan mantener una calidad elevada en el análisis de datos de los proyectos de investigación de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología.
- Apoyar el diseño de base de datos y el desarrollo de soluciones informáticas para la captura de información de los proyectos de investigación y otros proyectos del Instituto de Cancerología.
- Administrar las bases de datos de los proyectos de investigación desarrollados por el Instituto de Cancerología.
- Desarrollar soluciones informáticas que permitan la sistematización de los diferentes procesos relacionados con la gestión y análisis de datos resultado de proyectos de investigación u otros proyectos del Instituto de Cancerología.
- Apoyar, desde el punto de vista estadístico, la formulación de proyectos de investigación.
- Elaborar e implementar el plan de análisis estadístico de los proyectos de investigación.
- Apoyar los procesos de publicación científica en el componente de análisis de datos.
- Proponer e implementar mecanismos que permitan hacer seguimiento a la ejecución de proyectos de investigación.
- Desarrollar modelos de simulación para estudios de evaluación económica.
- Desarrollar modelos de simulación para la toma de decisiones en salud.

4.4. Actores, Roles y Funciones

Los actores, Roles y funciones que conforman La jerarquía del *Área de Investigación de análisis de datos* puede ser observada en la Figura4.1

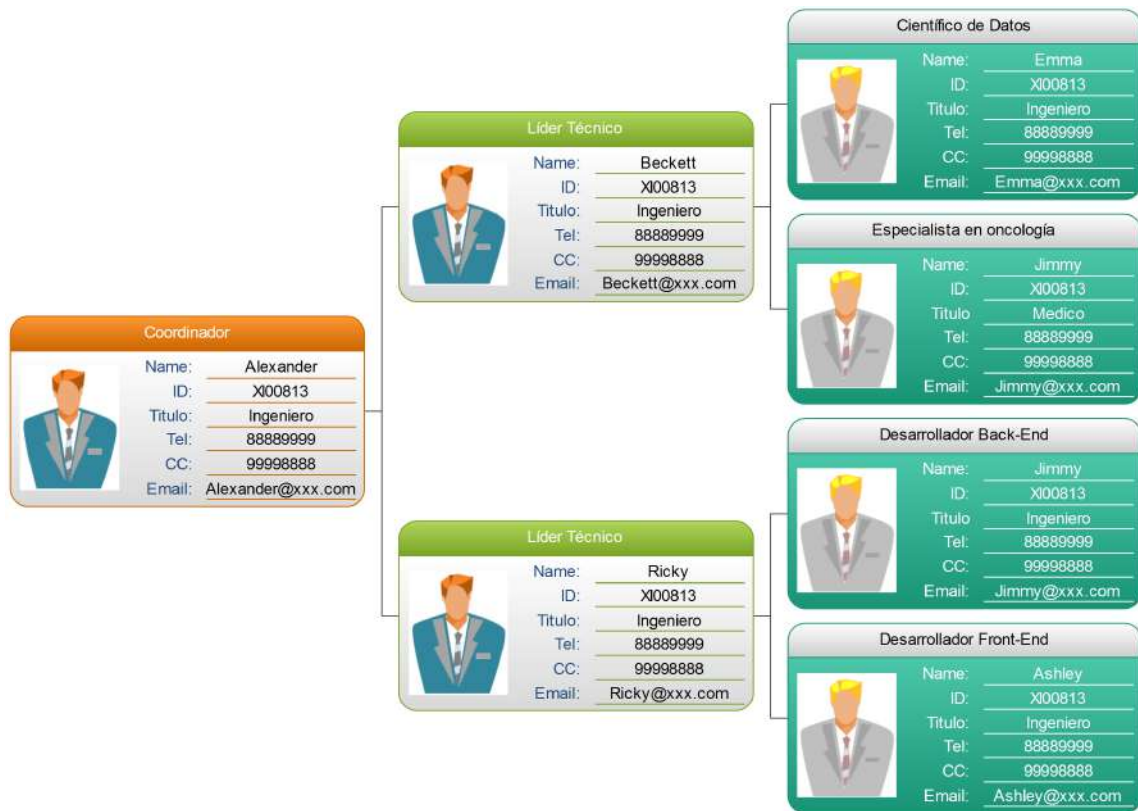


Figura 4.1: Jerarquía del Área de Investigación de análisis de datos

4.5. Servicios

El grupo de análisis de datos ofrece los siguientes servicios:

- Desarrollo de modelos de decisión (árboles de decisión, modelos de Markov, simulación de eventos discretos y simulación dinámica).
- Desarrollo de modelos estadísticos predictivos y explicativos.
- Análisis de datos avanzados (modelos lineales generalizados, análisis multinivel, etc.).
- Análisis bioinformáticas.
- Estadística espacial.
- Minería de datos.
- Diseños de formularios para la captura electrónica de información (web, escritorio, apps, etc.).
- Administración de bases de datos.
- Elaboración de proyectos de investigación (métodos).
- Programación en R, SAS, SPSS, Stata.
- Evaluación económica

Capítulo 5

Capa de Motivación

5.1. Introducción

Esta capa retoma, en primer lugar, algunos de los conceptos utilizados en la capa de negocio como por ejemplo el concepto de rol representado en esta capa como un Interesado (StakeHolder) y el concepto de servicio el cual tiene como equivalente un Objetivo(Goal)[47].

Es importante entender el uso e interacción de los elementos que definen la estructura de la organización, durante el proceso de apropiación de conceptos que se presenta en esta capa, se destacan los diferentes aspectos motivacionales que denotan principalmente aspectos organizacionales, en términos de cada una de las entidades encontradas y sus interacciones. Entidades como objetivos y principios que determinan de forma detallada aspectos fundamentales de la estructura organizacional (objetivos, plan estratégico, misión y visión)[47].

Por otra parte, es importante resaltar que, dentro de la estructura organizacional un aspecto clave corresponde a las relaciones que se presentan entre los diferentes elementos. A nivel externo (principios, interesados) e interno (objetivos, requerimientos) de la organización se pueden ver estas interacciones que representan los diferentes servicios ofrecidos, mencionando sus detalles a través de requerimientos y restricciones, y también los diferentes roles que interactúan en este escenario (interesados, manejador)determinando así la comunicación entre los conceptos organizacionales [47].

A continuación se presentan cada uno de los puntos de vista de la capa de Motivación a partir del soporte realizado por el Área de Investigación de Análisis de datos a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología.

5.2. Punto de Vista de Interesados

El Punto de Vista de Interesados permite al analista modelar las partes interesadas, los impulsores internos y externos del cambio y las evaluaciones en términos de fortalezas, debilidades, oportunidades y amenazas de dichos controladores. Además, se pueden describir los vínculos con los objetivos iniciales de alto nivel que abordan estas preocupaciones y evaluaciones. Estos objetivos forman la base para el proceso de ingeniería de requisitos, incluyendo refinamiento de objetivos, contribución y análisis de conflictos, y la derivación de requisitos que realicen las metas[47].

En la Figura 5.1, se plantea el Caso para el Punto de Vista de Interesados con cada uno de los elementos que interactúan entre sí.

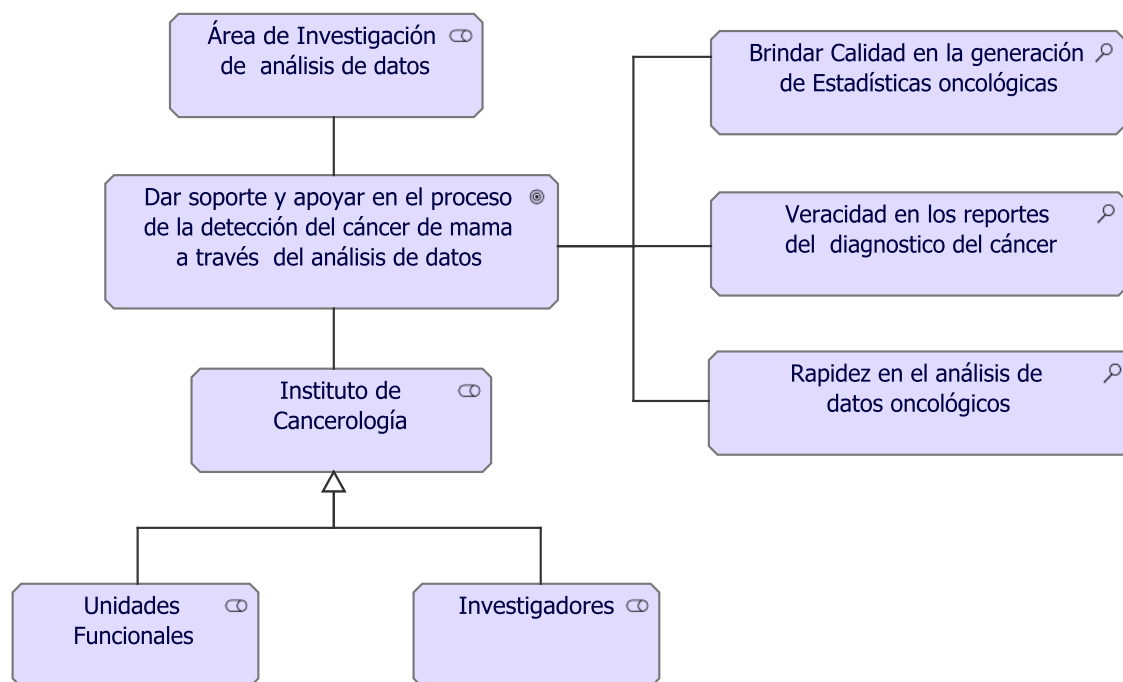


Figura 5.1: Punto de Vista de Interesados

- 1) **Área de Investigación de análisis de datos:** Corresponde a una de las dependencias del área de investigación la cual tiene por objetivo dar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología, haciendo uso métodos computacionales para el manejo y análisis de datos Oncológicos.

- 2) **Dar soporte y apoyar en el proceso de la detección del cáncer de mama a través del análisis de datos:** Este objetivo organizacional hace referencia a el servicio de generar diagnósticos relacionados con el cáncer de mama solicitados por los diferentes funcionarios del Instituto de Cancerología. Se asocia con dos Stakeholder: *Área de Investigación de análisis de datos* e *Instituto de Cancerología*; el primero es el que vela por el cumplimiento del objetivo; el segundo es a quien se brinda el servicio del Diagnóstico del Cáncer de mama.
- 3) **Instituto de Cancerología:** Este rol de aplicación corresponde a los diferentes usuarios del instituto de nacional de cancerología. Tienes asociado el actor *Investigadores* y el actor *Unidades funcionales*. Estos roles se describen a continuación:
 - **Investigadores:** Este rol está conformado por todos los grupos de investigación en cáncer del país registrados ante Colciencias y adicionalmente, con representantes de diferentes tipos de usuarios del conocimiento generado por la investigación como son las sociedades médicas, los prestadores de servicios oncológicos, los aseguradores, las autoridades sanitarias y los pacientes entre otros.
 - **Unidades Funcionales :** Este rol está conformado por las unidades clínicas ubicadas al interior del Instituto de Cancerología cuya función es evaluar la situación de salud del paciente con diagnóstico presuntivo de cáncer.
- 4) **Brindar Calidad en la generación de Estadísticas Oncológicas:** Teniendo en cuenta el objetivo *Dar soporte y apoyar en el proceso de la detección del cáncer de mama a través del análisis de datos*, se tiene esta valoración. En este caso, el resultado es la oportunidad de generar calidad en la información estadística acerca de la identificación del cáncer de mama diversos pacientes.
- 5) **Veracidad en los reportes del diagnostico:** Teniendo en cuenta el objetivo *Dar soporte y apoyar en el proceso de la detección del cáncer de mama a través del análisis de datos*, se tiene esta valoración. En este caso, el resultado es la oportunidad de generar diagnósticos exactos acerca de las posibilidad de padecer cáncer de mama.
- 6) **Rapidez en el análisis de datos oncológicos:** Teniendo en cuenta el objetivo *Dar soporte y apoyar en el proceso de la detección del cáncer de mama a través del análisis de datos*, se tiene esta valoración. En este caso, el resultado es la oportunidad de generar de forma eficiente diversos reportes enfocados en el diagnostico de cáncer de mama en el menor tiempo posible para poder dar un tratamiento oportuno a cada paciente.

5.3. Punto de Vista de Realización de Objetivos

El Punto de Vista de Realización de Objetivos permite modelar el refinamiento de metas (de alto nivel) en metas más concretas y el refinamiento de objetivos concretos en requisitos o restricciones que describen las propiedades que se necesitan para realizar las metas[47].

En la Figura 5.2, se plantea el Caso para el Punto de Vista de Realización de Objetivos con cada uno de los elementos que interactúan entre sí.

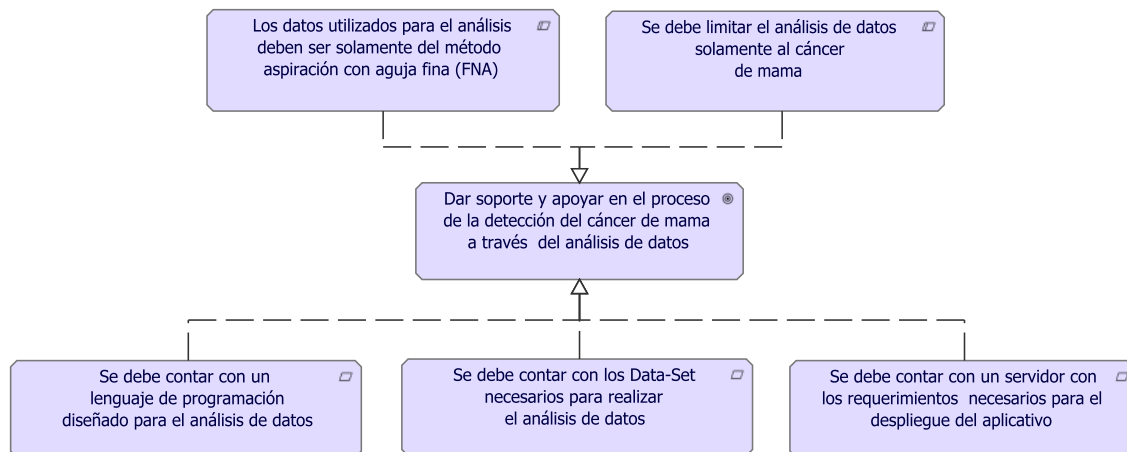


Figura 5.2: Punto de Vista de Realización de Objetivos

- 1) **Dar soporte y apoyar en el proceso de la detección del cáncer de mama a través del análisis de datos:** el objetivo organizacional cuenta con tres requerimientos y dos restricciones. En este caso, este objetivo es el encargado de realizarlos.
- 2) **Se debe contar con un lenguaje de programación diseñado para el análisis de datos:** Este requerimiento plantea que para poder generar el diagnostico de cáncer es necesario utilizar un lenguaje de programación que contenga herramientas y funcionalidades enfocados en el procesamiento y análisis de datos.
- 3) **Se debe contar con los Data-Set necesarios para realizar el análisis de datos:** Este requerimiento se refiere a que los datos sobre los que se va a realizar el análisis deben contener variables de pacientes ya diagnosticados con cáncer de mama que permitan el entrenamiento de los algoritmos que clasificaran nuevos pacientes candidatos de padecer dicho cáncer.

- 4) **Se debe contar con un servidor con los requerimientos necesarios para el despliegue del aplicativo:** Este requerimiento se refiere a que el servidor en el cual se va a desplegar el Back-End y el Front-End de la aplicación BreastApp debe contener las prestaciones de almacenamiento y procesamiento suficiente para la generación optima de diagnósticos asociados al cáncer de mama.
- 5) **Los datos utilizados para el análisis deben ser solamente del método aspiración con aguja fina (FNA):** Esta restricción hace referencia a que las variables necesarias para el diagnostico de cáncer de mama deben ser solamente las obtenidas por el método FNA, esto debido a que el aplicativo esta diseñado para generar los diagnósticos con base solamente a este método.
- 6) **Se debe limitar el análisis de datos solamente al cáncer de mama:** Esta restricción hace referencia que el aplicativo esta enfocado solamente al cáncer de mama y no a otro tipo de cáncer. Se realiza esta limitación para que la generación de reportes sea rápida y el cáncer pueda tratarse a tiempo.

5.4. Punto de Vista de Contribución de Objetivos

El Punto de Vista de Contribución de Objetivos permite a un analista modelar las relaciones de influencia entre objetivos y requisitos. Las vistas resultantes en este punto de vista pueden usarse para analizar el impacto que las metas tienen entre sí o para detectar conflictos entre los objetivos de las partes interesadas[47].

En la Figura 5.3, se plantea el Caso para el Punto de Vista de Contribución de Objetivos con cada uno de los elementos que interactúan entre sí.

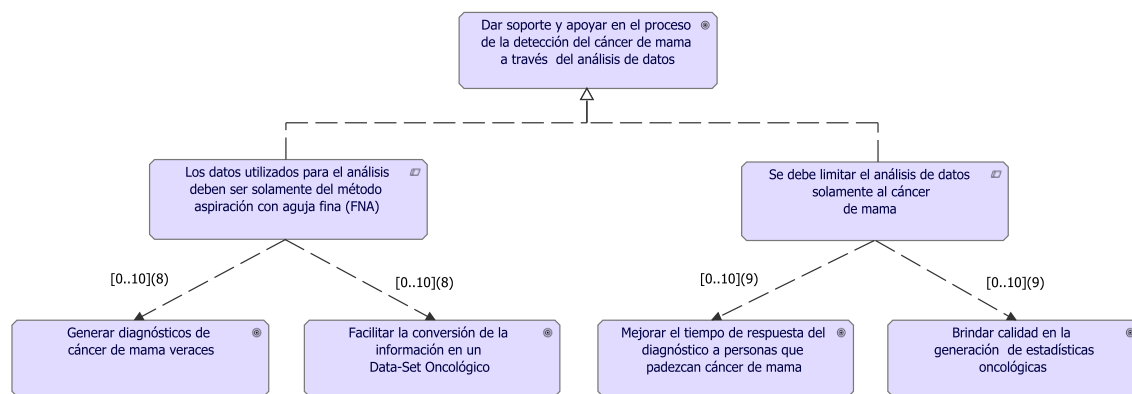


Figura 5.3: Punto de Vista de Contribución de Objetivos

- 1) **Dar soporte y apoyar en el proceso de la detección del cáncer de mama a través del análisis de datos:** para este punto de vista, se toman en cuenta las restricciones que realiza este objetivo organizacional, planteada en el modelo del punto de vista anterior. Esta se describe en el apartado siguiente.
- 2) **Los datos utilizados para el análisis deben ser solamente del método aspiración con aguja fina (FNA):** Esta restricción hace referencia a que las variables necesarias para el diagnóstico de cáncer de mama deben ser solamente las obtenidas por el método FNA, esto debido a que el aplicativo está diseñado para generar los diagnósticos con base solamente a este método. Esta restricción tiene dos objetivos con impacto positivo. El impacto se mide en una escala de 1 a 10, siendo 10 el más alto y el impacto más positivo.
 - **Generar diagnósticos de cáncer de mama veraces:** Teniendo en cuenta la restricción anterior se tiene un impacto positivo en este objetivo debido a que al procesar solamente datos obtenidos por el método de aguja fina (FNA) el entrenamiento de los algoritmos van a generar resultados cada vez más precisos en la detección del cáncer de mama. El nivel de impacto asociado es de 8.

- ***Facilitar la conversión de la información en un Data-Set Oncológico:***

Teniendo en cuenta la restricción anterior se tiene un impacto positivo en este objetivo debido a que el método de aguja fina(FNA) es bastante utilizado en el ámbito medico,por lo que las estandarización e identificación de variables facilita la generación diversos Data-Set para el diagnostico de cáncer de mama.El nivel de impacto asociado es de 8.

3) Se debe limitar el análisis de datos solamente al cáncer de mama: Este restricción hace referencia que el aplicativo esta enfocado solamente al cáncer de mama y no a otro tipo de cáncer. Se realiza esta limitación para que la generación de reportes sea rápida y el cáncer pueda tratarse a tiempo.Esta restricción tiene dos objetivos con impacto positivo. El impacto se mide en una escala de 1 a 10, siendo 10 el más alto y el impacto más positivo.

- ***Mejorar el tiempo de respuesta del diagnostico a personas que padecen cáncer de mama:***

Teniendo en cuenta la restricción anterior se tiene un impacto positivo debido a que al enfocarse solamente en el cáncer de mama la generación de diagnósticos es mas rápida ya que el sistema esta diseñado solamente para este tipo de cáncer.El nivel de impacto asociado es de 9.

- ***Brindar calidad en la generación de estadísticas Oncológicas:***

Teniendo en cuenta la restricción anterior se tiene un impacto positivo debido a que al enfocarse solamente en el cáncer de mama los algoritmos del sistema cada vez van a irse entrenando con una gran cantidad de datos mayor que va a garantizar cada vez un resultado mas exacto del diagnostico de cáncer de mama.

5.5. Punto de Vista de Principios

El Punto de Vista de Principios permite al analista modelar los principios que son relevantes para el problema de diseño en cuestión, incluyendo los objetivos que motivan dichos principios [47].

En la Figura 5.4, se plantea el Caso para el Punto de Vista de principios con cada uno de los elementos que interactúan entre sí.

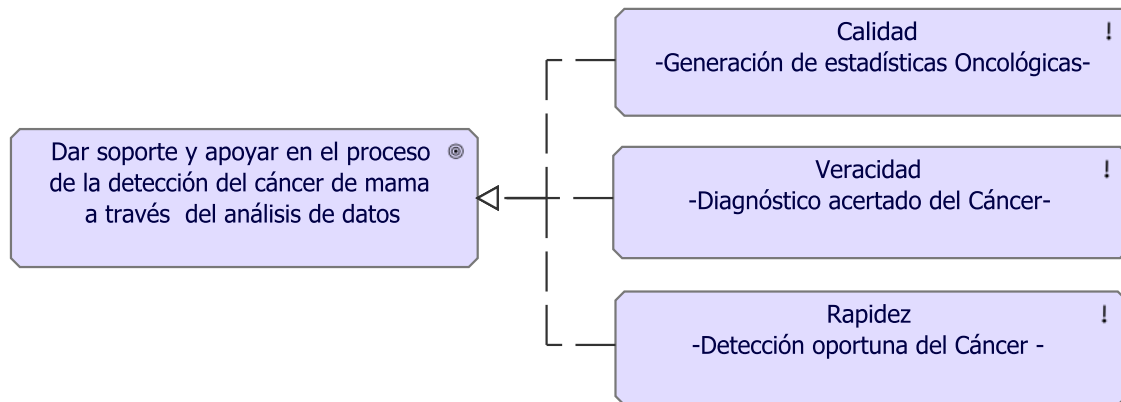


Figura 5.4: Punto de Vista de Principios

1) **Dar soporte y apoyar en el proceso de la detección del cáncer de mama a través del análisis de datos:** Este objetivo organizacional hace referencia a el servicio de generar diagnósticos relacionados con el cáncer de mama solicitados por los diferentes funcionarios del Instituto de Cancerología. Realiza tres principios representativos para la organización, descritos a continuación.

- **Calidad:** Generación de estadísticas oncológicas con una adecuada organización para una mayor comprensión y análisis.
- **Veracidad:** Precisión alta en el diagnóstico y detección del Cáncer de mama.
- **Rapidez:** Generación de diagnósticos del cáncer de mama de forma eficiente para un tratamiento oportuno.

5.6. Punto de Vista de Realización de Requerimientos

El Punto de Vista de Realización de Requerimientos permite a los implicados modelar la realización de los requisitos por parte de los elementos básicos, como los actores empresariales y los servicios empresariales. Este punto de vista tiene en cuenta elementos como los procesos empresariales, los servicios y componentes de la aplicación. Además, puede usarse para refinar requisitos en requisitos mas detallados[47].

En la Figura 5.5, se plantea el Caso para el Punto de Vista de Realización de Requerimientos con cada uno de los elementos que interactúan entre sí.

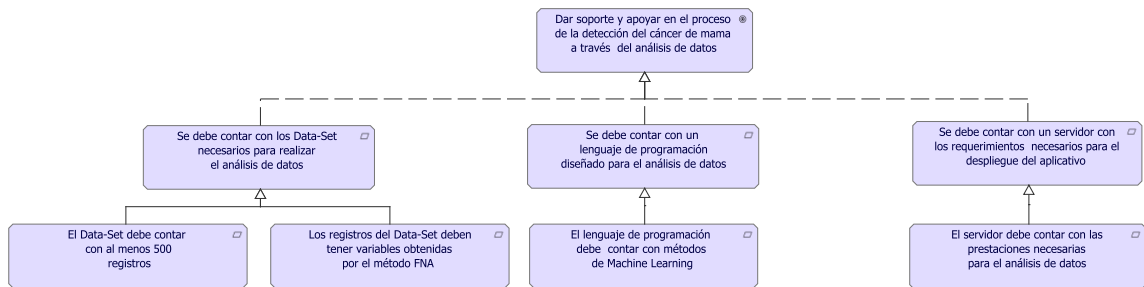


Figura 5.5: Punto de Vista de Realización de Requerimientos

- 1) **Dar soporte y apoyar en el proceso de la detección del cáncer de mama a través del análisis de datos:** El objetivo organizacional cuenta con tres requerimientos. En este caso, este objetivo es el encargado de realizarlos.
- 2) **Se debe contar con un lenguaje de programación diseñado para el análisis de datos:** Este requerimiento plantea que para poder generar el diagnostico de cáncer es necesario utilizar un lenguaje de programación que contenga herramientas y funcionalidades enfocados en el procesamiento y análisis de datos. Este requerimiento tiene asociado los dos requerimientos que se detallan a continuación:
 - **El Data-Set debe contar con al menos 500 registros:** Es necesario para el entrenamiento de los algoritmos de Machine Learning contar con información mayor o igual a 500 registros de pacientes los cuales incluyan las variables necesarias para el entrenamiento de dichos algoritmos para realizar posteriormente el diagnostico de nuevos pacientes.
 - **Los registros del Data-Set deben tener variables obtenidas por el método FNA:** Es necesario que las variables obtenidas sean del método medico de aspiración con aguja fina (FNA) esto debido a que el aplicativo esta diseñado para generar los diagnósticos con base solamente a este método.

- 3) **Se debe contar con un lenguaje de programación diseñado para el análisis de datos:** Este requerimiento plantea que para poder generar que el diagnostico de cáncer es necesario utilizar un lenguaje de programación que contenga herramientas y funcionalidades enfocados en el procesamiento y análisis de datos. Este requerimiento tiene asociado el requerimiento que se detalla a continuación:
 - ***El lenguaje de programación debe contar con métodos de Machine Learning:*** Es necesario que el lenguaje de programación cuente con métodos de Machine Learning debido a que por medio de ellos es que se realiza el diagnostico de cáncer de mama.
- 4) **Se debe contar con un servidor con los requerimientos necesarios para el despliegue del aplicativo:** Este requerimiento se refiere a que el servidor en el cual se va a desplegar el Back-End y el Front-End de la aplicación BreastApp debe contener las prestaciones de almacenamiento y procesamiento suficiente para la generación optima de diagnósticos asociados al cáncer de mama. Este requerimiento tiene asociado el requerimiento que se detalla a continuación:
 - ***El servidor debe contar con las prestaciones necesarias para el análisis de datos:*** Es necesario que el servidor cuente con la memoria suficiente de almacenamiento y procesamiento debido a que la cantidad de datos y cálculos realizados por los métodos de Machine Learning es bastante alta.

5.7. Punto de Vista de Motivación

El Punto de Vista de Motivación permite al diseñador o analista modelar el aspecto de las razones (motivaciones), que guían el diseño o el cambio de una Arquitectura Empresarial. Este punto de vista puede utilizarse para presentar un panorama completo o parcial del aspecto de la motivación relacionando a las partes interesadas, sus objetivos principales, los principios que se aplican y los principales requisitos de servicios, procesos, aplicaciones y objetos[47].

En la Figura 5.6, se plantea el Caso para el Punto de Vista de Motivación con cada uno de los elementos que interactúan entre sí.

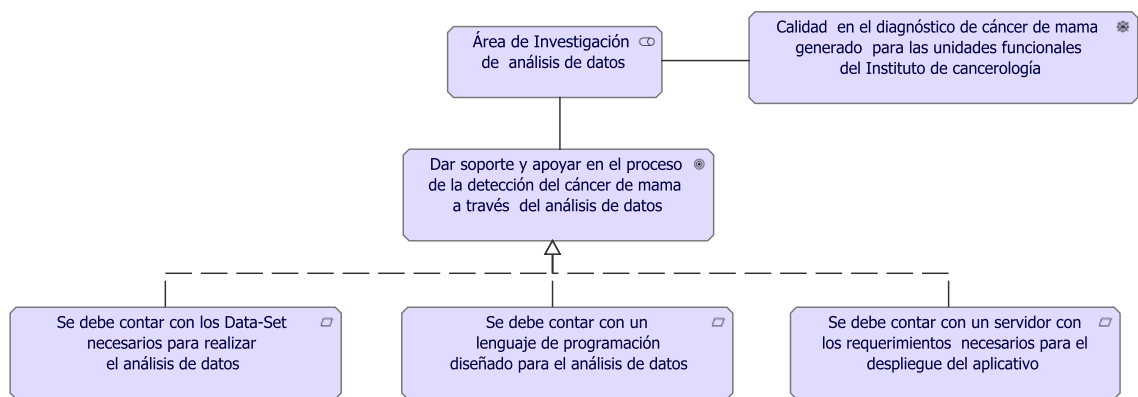


Figura 5.6: Punto de Vista de Motivación

- 1) **Dar soporte y apoyar en el proceso de la detección del cáncer de mama a través del análisis de datos:** El objetivo organizacional cuenta con tres requerimientos. En este caso, este objetivo es el encargado de realizarlos. Además, se asocia con el StakeHolder *Área de Investigación de análisis de datos*.
- 2) **Se debe contar con un lenguaje de programación diseñado para el análisis de datos:** Este requerimiento plantea que para poder generar el diagnóstico de cáncer es necesario utilizar un lenguaje de programación que contenga herramientas y funcionalidades enfocados en el procesamiento y análisis de datos.
- 3) **Se debe contar con los Data-Set necesarios para realizar el análisis de datos:** Este requerimiento se refiere a que los datos sobre los que se va a realizar el análisis deben contener variables de pacientes ya diagnosticados con cáncer de mama que permitan el entrenamiento de los algoritmos que clasificarán nuevos pacientes candidatos de padecer dicho cáncer.

- 4) **Se debe contar con un servidor con los requerimientos necesarios para el despliegue del aplicativo:** Este requerimiento se refiere a que el servidor en el cual se va a desplegar el Back-End y el Front-End de la aplicación BreastApp debe contener las prestaciones de almacenamiento y procesamiento suficiente para la generación optima de diagnósticos asociados al cáncer de mama.
- 5) **Calidad en el diagnóstico de cáncer de mama generado para las unidades funcionales del Instituto de cancerología:** Este conductor hace referencia a la búsqueda de mejoramiento en la atención a pacientes de los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología con base en la generación de diagnosticos de cáncer de mama con un nivel de calidad alto .Se asocia con el StakeHolder *Área de Investigación de análisis de datos*.

Capítulo 6

Capa de Estrategia

6.1. Introducción

Esta capa plantea los aspectos estratégicos de la empresa. Cada uno de los puntos de vista que contiene esta capa presenta una perspectiva diferente sobre el modelado de la dirección estratégica de alto nivel y la composición de la organización[48].

Esta capa busca establecer y conocer los diferentes planes estratégicos que van a tomar lugar en los procesos de negocio actuales de la organización, por lo tanto, es importante tener en cuenta los conceptos utilizados en la capa de Motivación para así facilitar la identificación de los diferentes usos e interacciones que se presentan con esta capa. Para una adecuada apropiación de la conceptualización de esta capa, se implementa el uso de cuatro diferentes puntos de vista: Estrategia, Mapa de capacidad, Realización de Resultado y Mapa de recurso.

A continuación se presentan cada uno de los puntos de vista de la capa de Estrategia a partir del soporte realizado por el Área de Investigación de Análisis de datos a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología.

6.2. Punto de Vista de la Estrategia

El punto de vista de la estrategia permite a la organización modelar una visión general estratégica de alto nivel de los cursos de acción de la empresa, las capacidades y recursos que los respaldan, y los resultados previstos[48].

En la Figura 6.1 se plantea el Caso para el Punto de Vista de la Estrategia con cada uno de los elementos que interactúan entre sí.

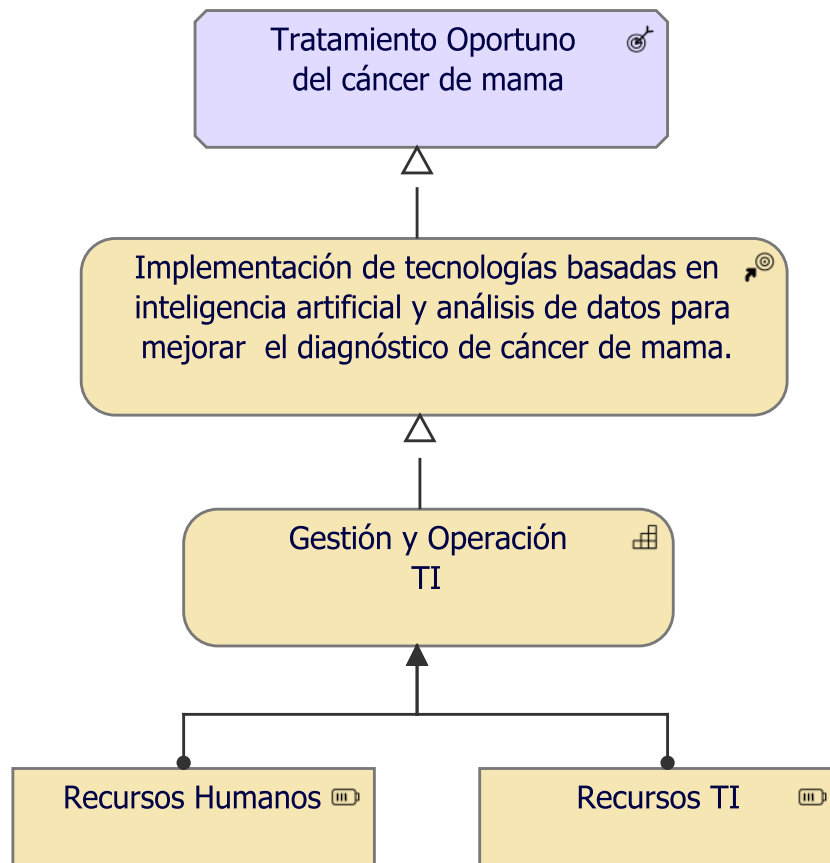


Figura 6.1: Punto de Vista de la Estrategia

- 1) **Tratamiento Oportuno del cáncer de mama:** Hace referencia al *resultado* que espera obtener el Instituto de Cancerología a través del Área de Investigación de Análisis de datos. Este resultado está basado en la rapidez y veracidad de los diagnósticos de cáncer de mama generados para cada paciente que tiene como consecuencia un tratamiento a tiempo que permita sanar dicho cáncer.

- 2) **Implementación de tecnologías basadas en inteligencia artificial y análisis de datos para mejorar el diagnóstico de cáncer de mama :** Hace referencia al *curso de acción* planteado por el Instituto de Cancerología a través del Área de Investigación de Análisis de datos para obtener un *Tratamiento Oportuno del cáncer de mama*.
- 3) **Gestión y Operación TI:** Hace referencia a la *capacidad* con respecto a la gestión y operación de Tecnologías de la Información del Área de Investigación de Análisis de datos para realizar el curso de acción de *Implementación de tecnologías basadas en inteligencia artificial y análisis de datos para mejorar el diagnóstico de cáncer de mama*. Esta asignado a los siguientes recursos:
 - **Recursos Humanos:** Este recurso hace referencia a los coordinadores, líderes técnicos, Especialistas en Oncología, científicos de datos y desarrolladores del Área de Investigación de Análisis de datos los cuales poseen habilidades en la gestión y operación de las Tecnologías de la Información.
 - **Recursos TI:** Este recurso hace referencia a las herramientas basadas en las Tecnologías de la Información que influyen en la realización de diagnósticos oncológicos. En este caso la principal herramienta para la generación de diagnósticos de cáncer de mama es la aplicación web *BreastApp*.

6.3. Punto de Vista del Mapa de Capacidad

El punto de vista del mapa de capacidad permite crear una visión general estructurada de las capacidades de la empresa. Un mapa de capacidades generalmente muestra dos o tres niveles de las capacidades en toda la organización. Puede, por ejemplo, usarse como un mapa de calor para identificar áreas de inversión. En algunos casos, un mapa de capacidades también puede mostrar resultados específicos entregados por estas capacidades[48].

En la Figura 6.2 se plantea el Caso para el Punto de Vista del Mapa de Capacidad con cada uno de los elementos que interactúan entre sí.

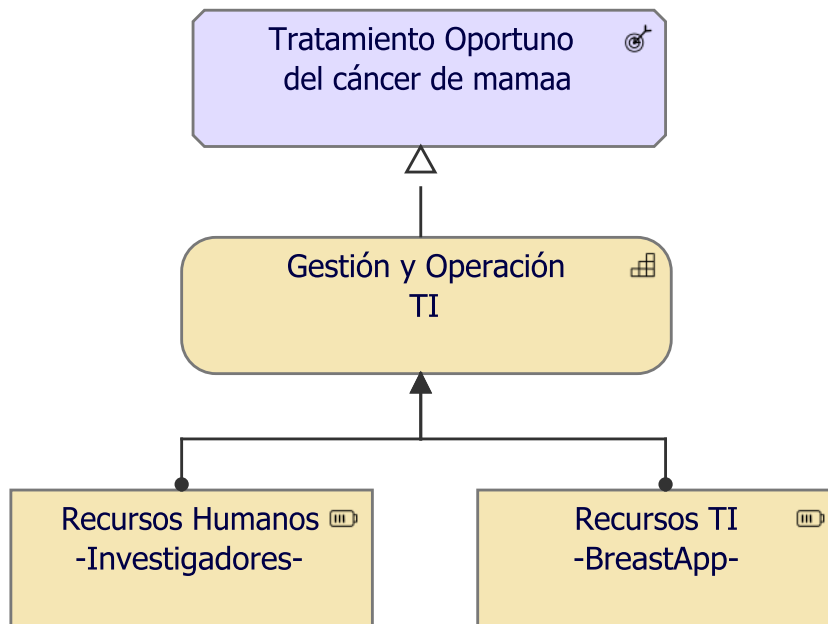


Figura 6.2: Punto de Vista del Mapa de Capacidad

- 1) **Tratamiento Oportuno del cáncer de mama:** Hace referencia al *resultado* que espera obtener el Instituto de Cancerología a través del Área de Investigación de Análisis de datos. Este resultado está basado en la rapidez y veracidad de los diagnósticos de cáncer de mama generados para cada paciente que tiene como consecuencia un tratamiento a tiempo que permita sanar dicho cáncer.

- 2) **Gestión y Operación TI:** Hace referencia a la *capacidad* con respecto a la gestión y operación de Tecnologías de la Información del Área de Investigación de Análisis de datos para realizar el curso de acción de *Implementación de tecnologías basadas en inteligencia artificial y análisis de datos para mejorar el diagnóstico de cáncer de mama*. Esta asignado a los siguientes recursos:
- **Recursos Humanos:** Este recurso hace referencia a los coordinadores, líderes técnicos, Especialistas en Oncología, científicos de datos y desarrolladores del Área de Investigación de Análisis de datos los cuales poseen habilidades en la gestión y operación de las Tecnologías de la Información.
 - **Recursos TI:** Este recurso hace referencia a las herramientas basadas en las Tecnologías de la Información que influyen en la realización de diagnósticos oncológicos. En este caso la principal herramienta para la generación de diagnósticos de cáncer de mama es la aplicación web *BreastApp*.

6.4. Punto de Vista de la Realización de Resultado

El punto de vista de realización del resultado se utiliza para mostrar los resultados orientados al negocio ,considerados de alto nivel, son realizados por las capacidades y los elementos centrales de la organización[48].

En la Figura 6.3 se plantea el Caso para el Punto de Vista de la realización de resultado con cada uno de los elementos que interactúan entre sí.

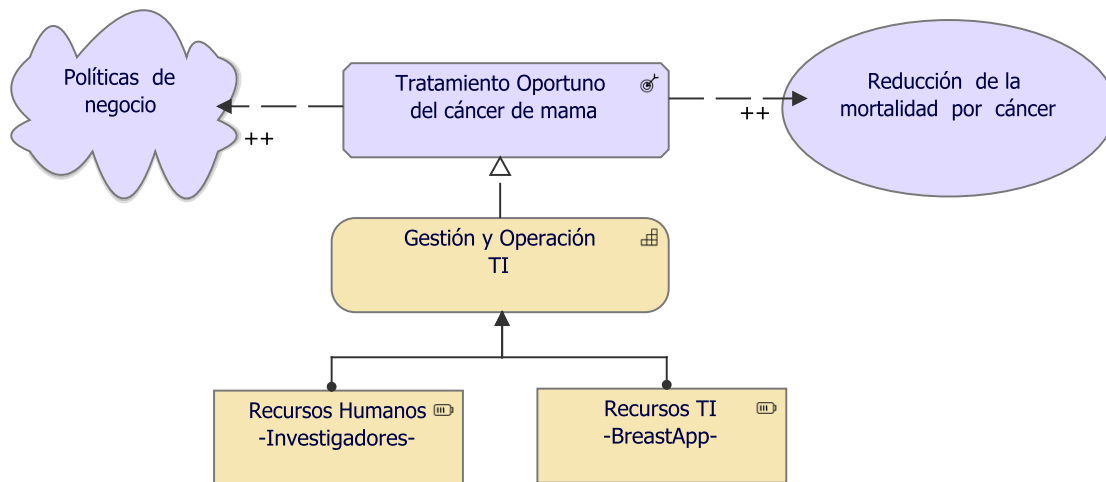


Figura 6.3: Punto de Vista de la Realización de Resultado

- 1) **Tratamiento Oportuno del cáncer de mama:** Hace referencia al *resultado* que espera obtener el Instituto de Cancerología a través del Área de Investigación de Análisis de datos. Este resultado esta basado en la rapidez y veracidad de los diagnósticos de cáncer de mama generados para cada paciente que tiene como consecuencia un tratamiento a tiempo que permita sanar dicho cáncer.
- 2) **Políticas de negocio:** Hace referencia al elemento asociado al core del instituto de Cancerología el cual esta basado en el control integral del cáncer a través de la atención y el cuidado de pacientes, la investigación, la formación de talento humano y el desarrollo de acciones en salud pública. Se puede observar que el *Tratamiento Oportuno del cáncer de mama* obtenido como resultado de la estrategia planteada por el Área de Investigación de Análisis de datos tiene un efecto positivo en las políticas de negocio.

- 3) **Reducción de la mortalidad por cáncer:** Hace referencia a el valor del core de la visión del instituto de Cancerología fundamentado en la reducción de la mortalidad por cáncer, sobre la base de la innovación y la tecnología. Se puede observar que el *Tratamiento Oportuno del cáncer de mama* obtenido como resultado de la estrategia planteada por el Área de Investigación de Análisis de datos de datos tiene un efecto positivo en este valor de la organización.
- 4) **Gestión y Operación TI:** Hace referencia a la *capacidad* con respecto a la gestión y operación de Tecnologías de la Información del Área de Investigación de Análisis de datos para realizar el curso de acción de *Implementación de tecnologías basadas en inteligencia artificial y análisis de datos para mejorar el diagnóstico de cáncer de mama*. Esta asignado a los siguientes recursos:
 - **Recursos Humanos:** Este recurso hace referencia a los coordinadores, líderes técnicos, Especialistas en Oncología, científicos de datos y desarrolladores del Área de Investigación de Análisis de datos los cuales poseen habilidades en la gestión y operación de las Tecnologías de la Información.
 - **Recursos TI:** Este recurso hace referencia a las herramientas basadas en las Tecnologías de la Información que influyen en la realización de diagnósticos oncológicos. En este caso la principal herramienta para la generación de diagnósticos de cáncer de mama es la aplicación web *BreastApp*.

6.5. Punto de Vista de Mapa de Recurso

El punto de vista de mapa de recurso permite tener una visión general estructurada de los recursos de la organización describiendo las relaciones entre los recursos y las capacidades a las que están asignados [48].

En la Figura 6.4 se plantea el Caso para el Punto de Vista de Mapa de Recurso con cada uno de los elementos que interactúan entre sí.

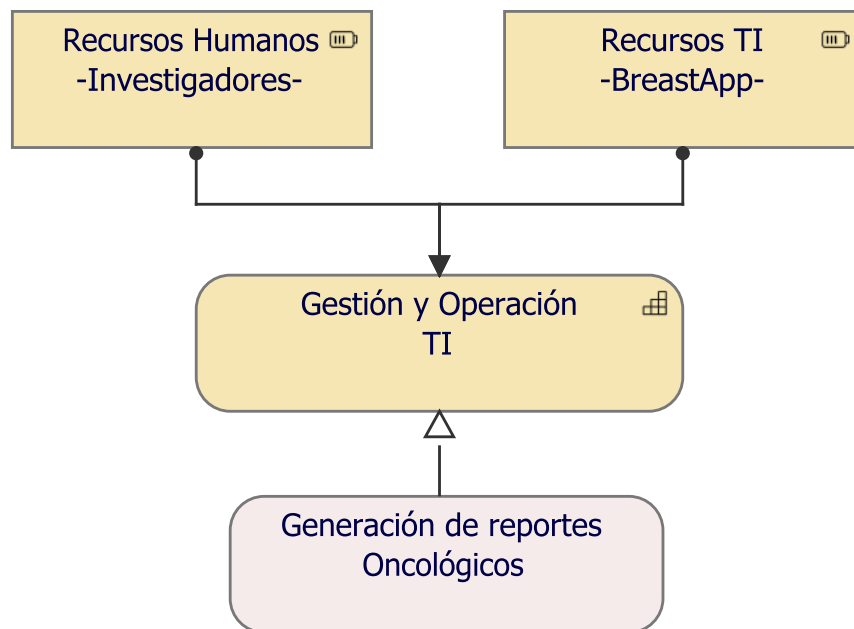


Figura 6.4: Punto de Vista de Mapa de Recurso

- 1) **Gestión y Operación TI:** Hace referencia a la *capacidad* con respecto a la gestión y operación de Tecnologías de la Información del Área de Investigación de Análisis de datos para realizar el curso de acción de *Implementación de tecnologías basadas en inteligencia artificial y análisis de datos para mejorar el diagnóstico de cáncer de mama*. Esta asignado a los siguientes recursos:

- **Recursos Humanos:** Este recurso hace referencia a los coordinadores, líderes técnicos, Especialistas en Oncología, científicos de datos y desarrolladores del Área de Investigación de Análisis de datos los cuales poseen habilidades en la gestión y operación de las Tecnologías de la Información.

- **Recursos TI:** Este recurso hace referencia a las herramientas basadas en las Tecnologías de la Información que influyen en la realización de diagnósticos oncológicos. En este caso la principal herramienta para la generación de diagnósticos de cáncer de mama es la aplicación web *BreastApp*.
- 2) **Generación de reportes oncológicos:** Es el paquete de trabajo principal del instituto de Cancerología basado en la estrategia planteada y los recursos disponibles identificados. La meta es construir una aplicación web que genere informes tipo reporte con base a los resultados arrojados por diferentes modelos de Machine Learning, en donde se de un resultado definitivo acerca del padecimiento de Cáncer de mama. Este diagnóstico se realiza con el propósito de dar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología.

Capítulo 7

Capa de Negocio

7.1. Introducción

Esta capa consiste en entender los conceptos que definen la estructura estática de la organización en términos de las entidades que componen la organización y las interacciones presentadas entre estas. Entidades como actores de negocio, roles de negocio, funciones de negocio, etc, que se comportan por medio de un diseño orientado al servicio, es decir, las funciones o procesos de negocio interno y externo de la organización como las responsabilidades [47].

Por otra parte, es importante resaltar que, dentro de la arquitectura de negocio se tiene que la relación entre entidades como se menciona anteriormente, de forma externa e interna, determina un proceso de comunicación entre estas, contando con interfaces que permiten el acceso a diferentes servicios mediante colaboraciones dadas por las entidades organizacionales [47].

A continuación se presentan cada uno de los puntos de vista de la capa de negocio a partir del soporte realizado por el Área de Investigación de Análisis de datos a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología.

7.2. Punto de Vista de la Organización

El Punto de Vista de Organización se centra en la organización interna de una empresa, un departamento, una red de empresas o de otra entidad organizativa. Se utiliza para identificar las competencias, la autoridad y las responsabilidades en una organización[47].

En la Figura 7.1, se plantea el Caso para el Punto de Vista de Organización con cada uno de los elementos que interactúan entre sí.

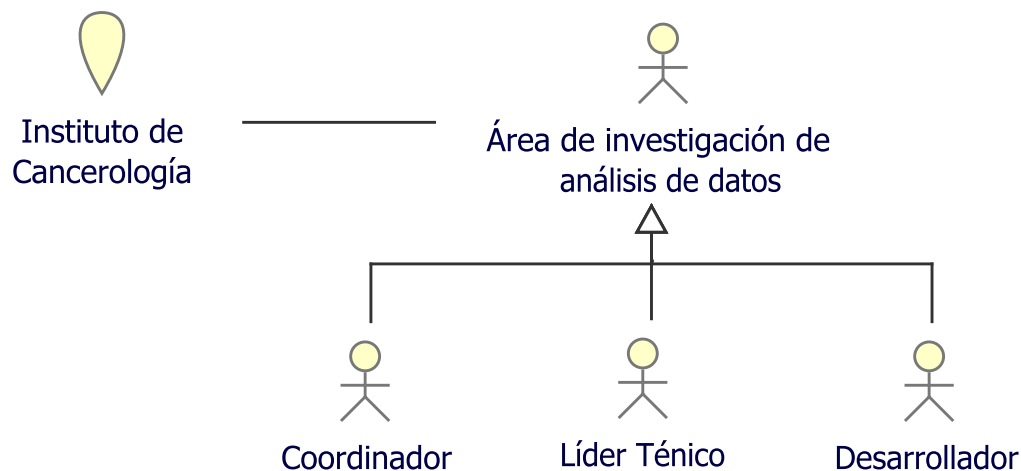


Figura 7.1: Punto de Vista de la Organización

- 1) **Instituto de Cancerología:** Este elemento hace referencia a la localización geográfica donde se encuentra ubicada la organización. El instituto de Cancerología cuenta con siete grupos de investigación que desarrollan diversas actividades según su campo de acción: investigación clínica, investigación epidemiológica, biología del cáncer e investigación en el área de la salud pública. Esta localización está asociada al actor *Área de investigación de análisis de datos*.
- 2) **Área de Investigación de análisis de datos:** Corresponde a una de las dependencias del área de investigación la cual tiene por objetivo dar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología, haciendo uso de métodos computacionales para el manejo y análisis de datos Oncológicos. Este actor tiene como *especialización* tres actores los cuales se muestran a continuación:

- **Coordinador:** Es la persona encargada de regular, gestionar, dirigir y supervisar el Área de Investigación de análisis de datos de Oncología para que el soporte realizado a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales se realice correctamente.
- **Líder Técnico:** Es la persona con conocimiento técnico avanzado en los temas de análisis de datos de Oncología y es el responsable de asignar y definir las tareas y el tiempo necesario en el desarrollo e implementación de los recursos según las necesidades presentadas por los investigadores y las unidades funcionales del Instituto de Cancerología.
- **Desarrollador:** Es la persona encargada de cumplir con las implementaciones presentadas en el ámbito de análisis de datos de Oncología y que da solución a cada uno de los requerimientos y necesidades presentadas por los investigadores y las unidades funcionales del Instituto de Cancerología.

7.3. Punto de Vista de Cooperación de Actor

El Punto de Vista de Cooperación de Actor se centra en las relaciones de los actores entre sí y su entorno. Se utiliza para determinar las dependencias externas y colaboraciones, y muestra la cadena de valor o la red en la que actúa el actor[47].

En la Figura 7.2, se plantea el Caso para el Punto de Vista de cooperación de Actor donde se describen los elementos que interactúan este punto de vista.

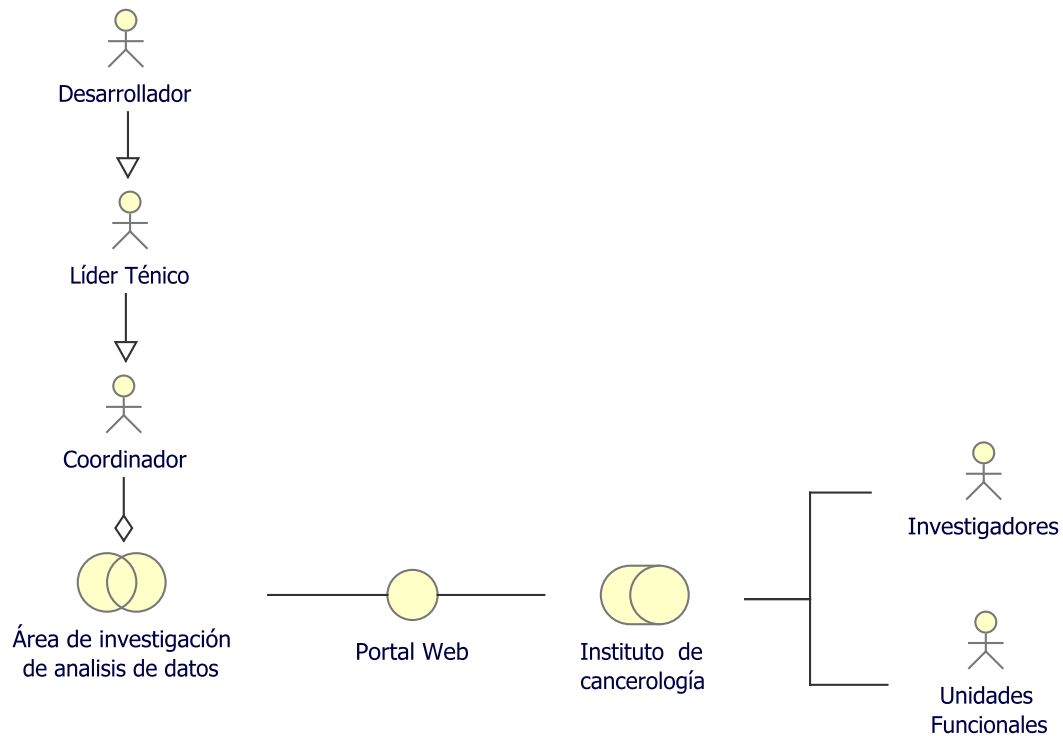


Figura 7.2: Punto de Vista de Cooperación de Actor

- 1) **Portal Web** : Es el medio de comunicación entre el Área de Investigación de análisis de datos y los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología. Esta interfaz tiene una asociación con el rol del *Área de Investigación de análisis de datos* y es usada por el rol *Instituto de Cancerología*.

2) **Instituto de Cancerología:** A el rol instituto de de cancerología están asociados el actor *Investigador* y el actor *Unidades funcionales* . Estos roles se describen a continuación:

- ***Investigadores*** : Este rol está conformado por todos los grupos de investigación en cáncer del país registrados ante Colciencias y adicionalmente, con representantes de diferentes tipos de usuarios del conocimiento generado por la investigación como son las sociedades médicas, los prestadores de servicios oncológicos, los aseguradores, las autoridades sanitarias y los pacientes entre otros.
- ***Unidades Funcionales*** : Este rol está conformado por las unidades clínicas ubicadas al interior del Instituto de Cancerología cuya función es evaluar la situación de salud del paciente con diagnóstico presuntivo de cáncer.

7.4. Punto de Vista de la Función de Negocio

El Punto de Vista de la Función de Negocio muestra las principales funciones de negocio de una organización y sus relaciones en términos de los flujos de información, valor o bienes entre ellos.

Adicionalmente, proporciona una visión de alto nivel de las operaciones generales de la empresa y puede utilizarse para identificar las competencias necesarias o estructurar una organización de acuerdo con sus principales actividades [47].

En la Figura 7.3, se plantea el Caso para el Punto de Vista de Función de Negocio donde se describen los elementos que interactúan en este punto de vista.

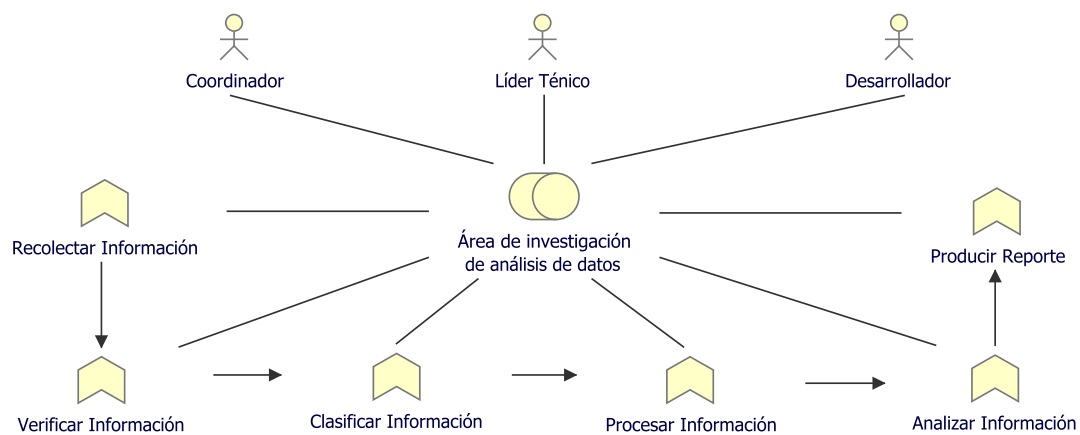


Figura 7.3: Punto de Vista de la Función del negocio

- 1) **Área de Investigación de análisis de datos:** Este rol corresponde a una de las dependencias del área de investigación la cual tiene por objetivo dar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología. Este rol está asociado a los actores coordinador, líder técnico y desarrollador y a seis diferentes funciones las cuales se muestran a continuación:

- **Recolectar Datos:** Esta función se encarga de recopilar y medir la información de los Data Set que contienen variables oncológicas específicas que requieren ser analizadas.
- **Verificar Datos:** Esta función se basa en la identificación y corrección o eliminación de registros de datos erróneos encontrados en los Data Set oncológicos. Este proceso permite encontrar datos incompletos, incorrectos, inexactos, no pertinentes, etc. y luego substituir, modificar o eliminar para realizar posteriormente la clasificación de los mismos.
- **Clasificar Datos:** Esta función se basa en la distribución por medio de algoritmos de Machine Learning de los Data Set que contienen variables oncológicas específicas que al ser comparados con trazas características de casos de cáncer avanzados permiten agruparlos en diferentes categorías según el porcentaje de padecimiento del mismo.
- **Procesar Datos:** Esta función se basa en la manipulación de los Data Set oncológicos para generar información por medio de gráficos y estadísticas haciendo uso de los modelos de Machine Learning.
- **Analizar Datos:** Esta función se encarga de examinar los Data Set que dieron como resultado de las funciones anteriores y los gráficos obtenidos con el propósito de sacar conclusiones diagnosticas sobre el posible padecimiento de cáncer.
- **Generar Reportes:** Esta función se encarga de crear en un módulo informativo los datos relevantes obtenidos de los Data Set oncológicos solicitados para ser analizados por medio de tablas, estadísticas y gráficos que dan el resumen de los resultados, conclusiones y diagnósticos sobre el posible padecimiento de cáncer en el individuo.

7.5. Punto de Vista de Proceso de Negocio

El Punto de Vista de Proceso de Negocio se utiliza para mostrar la estructura y composición de alto nivel de uno o más procesos empresariales[47].

Junto a los procesos mismos, este punto de vista contiene otros conceptos directamente relacionados, tales como: los servicios que un proceso de negocio ofrece de manera global, mostrando como un proceso contribuye a la realización de los productos de la empresa; la asignación de los procesos de negocio a las funciones, lo que da una idea de las responsabilidades de los actores asociados; la información utilizada por el proceso de negocio. Cada uno de estos puede ser considerado como una sub-vista de la vista del proceso empresarial[47].

En la Figura 7.4, se plantea el Caso para el Punto de Vista de Proceso de Negocio aplicado al Área de Investigación de análisis de datos. A continuación, se describen los elementos que interactúan en este punto de vista.

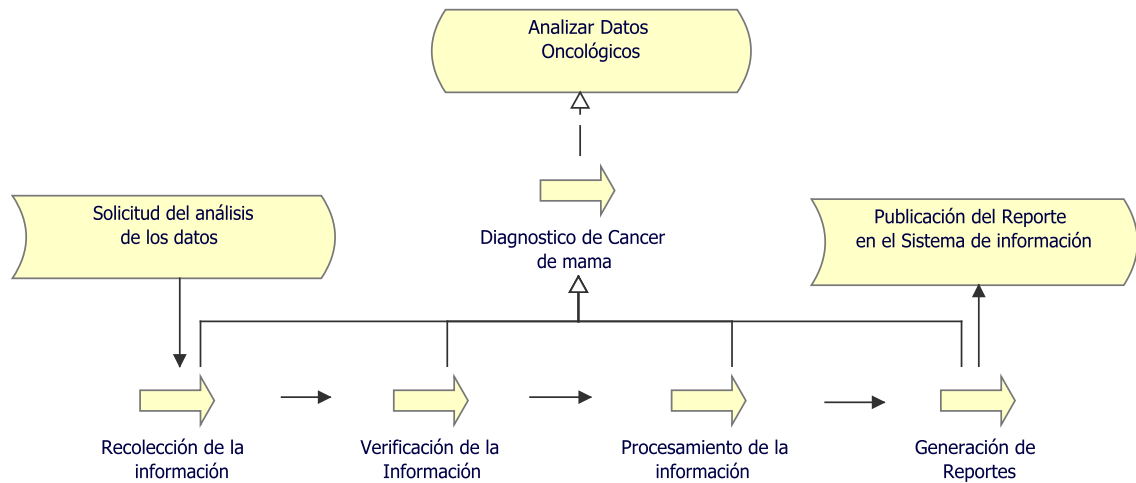


Figura 7.4: Punto de Vista del proceso del negocio

- 1) **Analizar Datos Oncológicos:** Este servicio realiza el análisis de los datos entregados por los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología para generar diagnósticos oncológicos. Es brindando por el proceso de *Análisis de Datos*.

- 2) **Diagnostico de Cáncer de mama:** El proceso de Diagnóstico de Cáncer de mama es fundamental debido que cumple con el objetivo principal de facilitar el análisis de datos oncológicos a las diferentes áreas y unidades funcionales de la organización. Además, este proceso está conformado de otros procesos en secuencia que dan solución al servicio final.
- 3) **Solicitud de análisis de datos:** Este evento es el que origina los sub-procesos para lograr el objetivo del servicio principal. Este corresponde a la acción en donde los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología desean que se analicen diferentes Data Set oncológicos para obtener un reporte. Está relacionado directamente con el sub-proceso de *Recolección de información* que es el primer proceso para analizar los datos solicitados.
- 4) **Recolección de información:** Este sub-proceso se encarga de recopilar y organizar toda la información que contiene datos oncológicos de cada paciente que son necesarios para su posterior verificación. Es el primer subproceso llevado a cabo para el objetivo de análisis de datos oncológicos que a su vez, desencadena un nuevo subproceso mediante una relación causal.
- 5) **Verificación de la información:** Este sub-proceso se encarga de identificar si las variables oncológicas solicitadas están completas. Además, si la información entregada esta completa pero algunos de los datos se encuentran erróneos o defectuosos se sustituyen para realizar posteriormente la clasificación de los mismos. Este sub-proceso desencadena un nuevo subproceso mediante una relación causal.
- 6) **Procesamiento de la información:** Este sub-proceso se encarga de usar los datos oncológicos verificados para generar información relevante por medio de gráficos y estadísticas. Este sub-proceso desencadena un nuevo subproceso mediante una relación causal.
- 7) **Generación de Reportes:** Este sub-proceso se encarga de organizar en un módulo la información de los resultados obtenidos de los Data Set oncológicos solicitados para ser analizados por medio de tablas, estadísticas y gráficos. Este sub-proceso desencadena un nuevo subproceso mediante una relación causal.

- 8) **Publicación del Reporte en el portal web :** Este evento es el que finaliza el flujo conformado por los sub-procesos anteriores y representa la acción de difundir la información oncológica analizada y los resultados diagnósticos obtenidos en el Portal web de cada uno de los datos solicitados para análisis. Este evento es accionado por el sub-proceso *Generación de Reportes*.

7.6. Punto de Vista de Cooperación de Proceso de Negocio

El Punto de Vista de Cooperación de Proceso de Negocio se utiliza para mostrar las relaciones de uno o más procesos de negocio entre si y/o con su entorno[47].

Puede utilizarse tanto para crear un diseño de alto nivel de procesos empresariales dentro de su contexto, como para proporcionar un gestor operativo responsable de uno o más de dichos procesos con información sobre sus dependencias [47].

En la Figura 7.5, se plantea el Caso para el Punto de Vista de Cooperación de Proceso de Negocio aplicado al Área de Investigación de análisis de datos. A continuación, se describen los elementos que interactúan en el punto de vista.

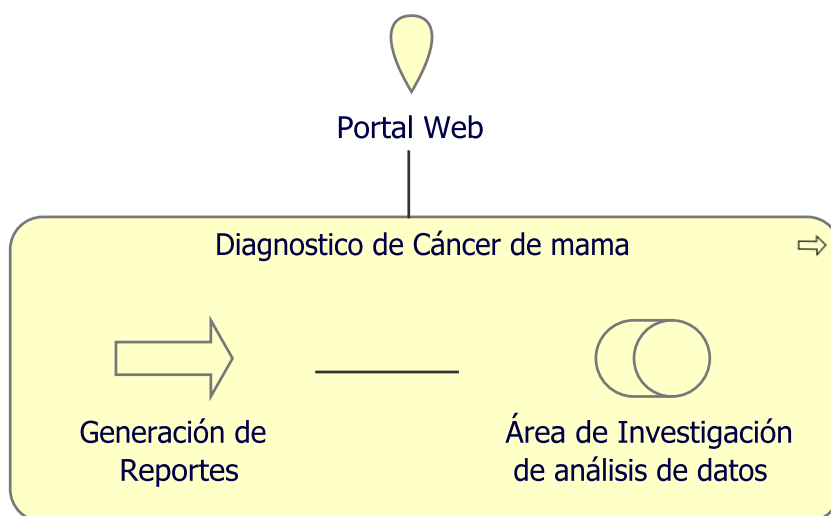


Figura 7.5: Punto de Vista de Cooperación de Proceso de Negocio

- 1) **Portal Web :** Es el medio de comunicación entre el Área de Investigación de análisis de datos y los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de cancerología. Está relacionado directamente con el servicio de *Diagnostico de Cáncer de mama*.
- 2) **Diagnóstico de Cáncer de mama:** Este servicio cumple el objetivo dar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología con respecto al análisis de datos oncológicos. Es brindado por el proceso de *Análisis de Datos*.

- 3) **Generación de Reportes:** Este sub-proceso hace referencia a la organización de la información de los resultados obtenidos de los Data Set oncológicos solicitados para ser analizados por medio de tablas, estadísticas y gráficos.
- 4) **Área de Investigación de análisis de datos:** Corresponde área de investigación la cual da soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología por medio reportes generados con base a los datos solicitados.

7.7. Punto de Vista de Producto

El Punto de Vista de Producto representa el valor que estos productos ofrecen a los clientes u otras partes externas involucradas[47].

Adicionalmente, muestra la composición de uno o más productos en términos de los servicios constitutivos-comerciales o de aplicación, y los contratos asociados u otros acuerdos. También se puede utilizar para mostrar las interfaces a través de las cuales se ofrece este producto, y los eventos asociados al mismo[47].

En la Figura 7.6, se plantea el Caso para el Punto de Vista de Producto. A continuación, se describen los elementos que interactúan en el punto de vista.

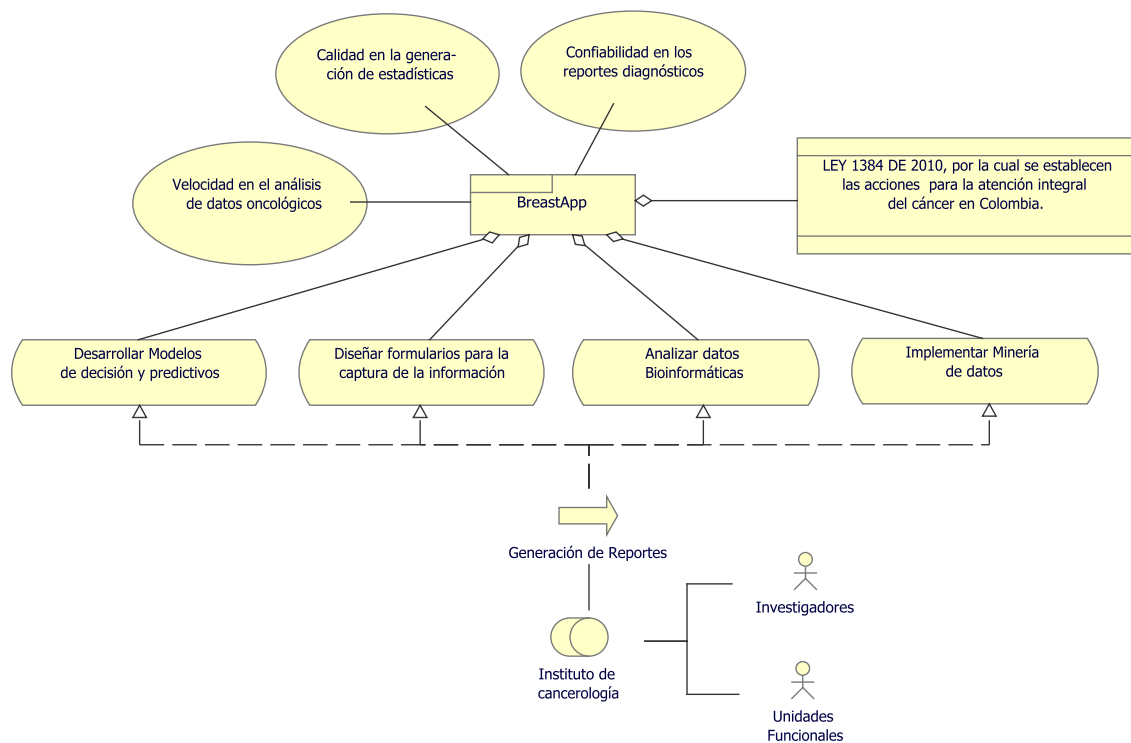


Figura 7.6: Punto de Vista de Producto

- 1) **BreastApp**: Este producto corresponde a la generación de reportes diagnósticos según el análisis de datos oncológicos elaborados por los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología. Está compuesto de los siguientes servicios:

- **Desarrollar Modelos de Decisión y predictivos:** Este servicio corresponde al uso de modelos de Machine Learning para diagnosticar el padecimiento de cáncer. Este servicio sirve como herramienta para realizar diversas investigaciones y poder disminuir su tasa de mortalidad en Colombia.
- **Diseñar formularios para la captura de la información:** Este servicio corresponde a la creación de formularios web de ámbito científico con énfasis en el diagnóstico del cáncer para que los diferentes investigadores y áreas funcionales de la organización proporcionen la información relacionada con los datos de cada individuo.
- **Analizar Datos Bioinformáticos:** Este servicio corresponde a los métodos y modelos basados en diferentes sistemas de cómputo numéricos con base a variables del ámbito de la biología para realizar simulaciones que permitan observar el comportamiento de los mismos en diferentes ambientes y determinar una posible solución al problema del padecimiento de cáncer.
- **Implementar Minería de Datos:** Este servicio corresponde a la búsqueda de correlaciones o patrones entre los millones de datos obtenidos de todos los pacientes con cáncer para proporcionar información diagnóstica cada vez más específica y exacta de los datos solicitados por los investigadores y las unidades funcionales de la organización.

Además, cuenta con los siguientes valores:

- **Velocidad en el análisis de datos oncológicos:** El objetivo principal es realizar de manera eficaz el análisis de los datos oncológicos en cuestión para dar un diagnóstico en el menor tiempo posible.
 - **Calidad en la generación de estadísticas:** El objetivo principal es analizar los datos de una manera profesional para satisfacer necesidades explícitas de los investigadores y las unidades funcionales de la organización.
 - **Confiabilidad en los reportes diagnósticos:** El objetivo principal es generar los reportes oncológicos analizados con una precisión superior para que la información generada por tablas y gráficos genere un diagnóstico exacto que ayude en las diferentes actividades de los investigadores y las unidades funcionales de la organización.
- 2) **Ley 1384 de 2010:** Esta resolución establece las acciones para el control integral del cáncer en la población colombiana, de manera que se reduzca la mortalidad y la morbilidad por cáncer adulto, así como mejorar la calidad de vida de los pacientes oncológicos, a través de la garantía por parte del Estado y de los actores que intervienen en el Sistema General de Seguridad Social en Salud vigente, de la prestación de todos los servicios que se requieran para su prevención, detección temprana, tratamiento integral, rehabilitación y cuidado paliativo.

- 3) **Generación de Reportes:** Implica la organización de la información de los resultados obtenidos en los diagnósticos oncológicos solicitados para ser visualizados por medio de tablas, estadísticas y gráficos. Se asocia con el rol de *Instituto de cancerología*. A su vez, realiza los servicios descritos anteriormente.
- 4) **Instituto de Cancerología:** Este elemento hace referencia el rol Instituto de de cancerología el cual cuenta con siete grupos de investigación que desarrollan diversas actividades según su campo de acción: investigación clínica, investigación epidemiológica, biología del cáncer e investigación en el área de la salud pública. Este rol usa el proceso *Generación de Reporte* y se asigna a los dos actores a continuación:
 - **Investigadores :** Este rol está conformado por todos los grupos de investigación en cáncer del país registrados ante Colciencias y adicionalmente, con representantes de diferentes tipos de usuarios del conocimiento generado por la investigación como son las sociedades médicas, los prestadores de servicios oncológicos, los aseguradores, las autoridades sanitarias y los pacientes entre otros.
 - **Unidades Funcionales :** Este rol está conformado por las unidades clínicas ubicadas al interior del Instituto de Cancerología cuya función es evaluar la situación de salud del paciente con diagnóstico presuntivo de cáncer.

Capítulo 8

Capa de Aplicación

8.1. Introducción

Esta capa retoma la mayoría de conceptos utilizados en la capa de negocio con la distinción de los diferentes usos e interacciones que se presentan. Para una adecuada apropiación de los conceptos de esta capa, se implementa el uso de cuatro diferentes puntos de vista: comportamiento, cooperación, estructura y uso de la aplicación [47].

Entidades como componentes y objeto de datos que se comportan por medio de una relación explícita en el modelo, es decir, el comportamiento a nivel interno y externo de la organización; el comportamiento externo dado en términos de los servicios de aplicación mientras que el comportamiento interno en funciones de aplicación que realizan estos servicios [47].

Dentro de la arquitectura de aplicación un aspecto fundamental corresponde a las interrelaciones de componentes, que determinan la comunicación entre estos, contando con el concepto de colaboración. Además, el concepto de interfaz que se define como el canal lógico mediante el cual se accede a un servicio de componente, detalla características de comportamiento tales como el conjunto de operaciones y eventos que expone el componente [47].

También, se puede distinguir los servicios de aplicaciones internas (comunicación de aplicación a aplicación) y servicios de aplicaciones externas (comunicación de aplicación a negocio), ambos servicios presentados por la interfaz de aplicación [47].

A continuación se presentan cada uno de los puntos de vista de la capa de aplicación a partir del soporte realizado por el Área de Investigación de Análisis de datos a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología.

8.2. Punto de Vista del Comportamiento de la Aplicación

El Punto de Vista de comportamiento de la aplicación describe el comportamiento interno de una aplicación, es útil para diseñar el comportamiento principal de las aplicaciones o para identificar la superposición funcional entre las mismas.[47].

En la Figura 8.1 se plantea el Caso para el Punto de Vista del comportamiento de la aplicación con cada uno de los elementos que interactúan entre sí.

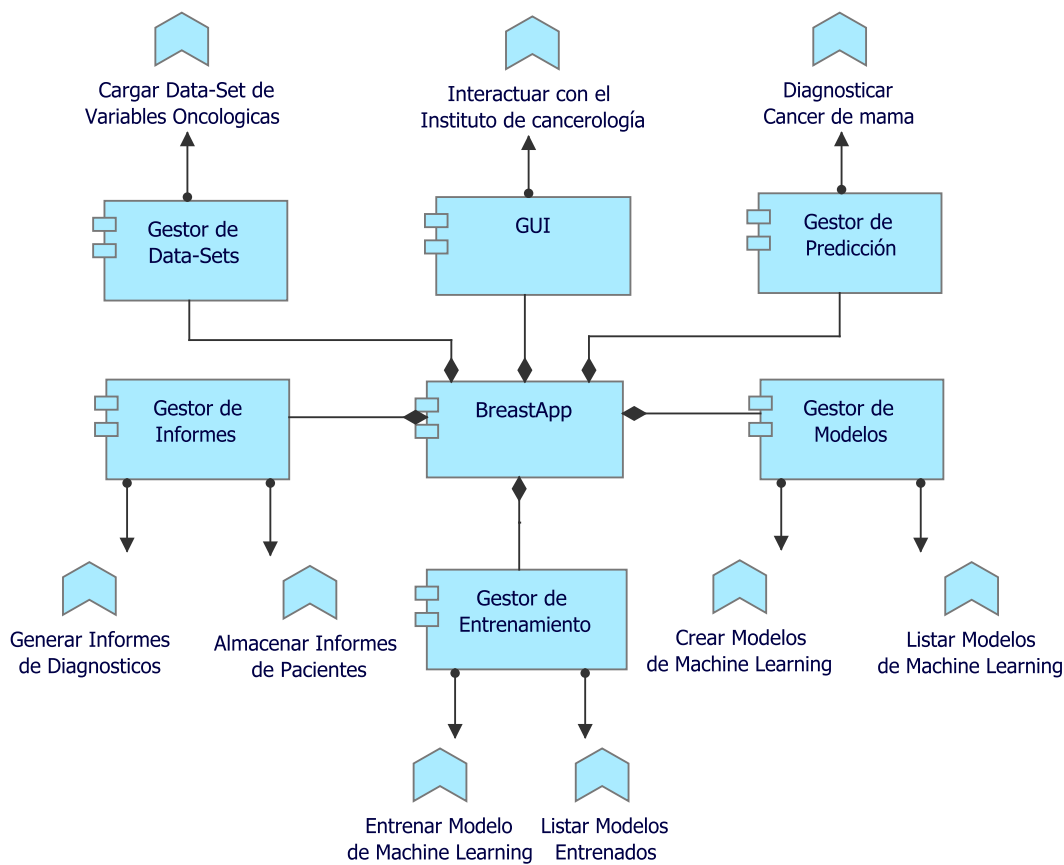


Figura 8.1: Punto de Vista del Comportamiento de la Aplicación

- 1) **BreastApp:** El componente de BreastApp es el framework de la aplicación. En este componente se agregan los componentes correspondientes para el diagnóstico de cáncer de mama y se realiza la carga de modelos, carga de Data-Sets, entrenamiento de modelos y generación de informes sobre el cáncer de mama.

- 2) **GUI:** Este componente esta agregado al componente principal de BreastApp. Este componente hace referencia a la interfaz gráfica de usuario por la cual los investigadores y unidades funcionales del Instituto de Cancerología acceden a la aplicación, para llevar el análisis según diagnostico de cáncer de mama generado por la aplicación. Este componente cumple con la siguientes funciones:
 - ***Interactuar con el Instituto de Cancerologia:*** Hace referencia a las acciones y herramientas que contiene el Front-End de la aplicacion BreastApp las cuales pueden ser usadas po los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología
- 3) **Gestor de Data-Sets:** Este componente esta agregado al componente principal de BreastApp. En este componente se maneja el proceso de carga y gestión de Data-Sets para ser almacenados en el sistema. Este componente cumple con la siguientes funciones:
 - ***Cargar Data-Set de Variables Oncológicas :*** Hace referencia a la carga del Data-Set para entrenamiento de los Modelos en Machine Learning y el Data-Set que contiene las variables Oncológicas de los pacientes a los cuales se les quiere detectar el padecimiento de cáncer de mama.
- 4) **Gestor de Entrenamiento:** Este componente esta agregado al componente principal de BreastApp. En este componente se maneja el proceso de entrenamientos de los modelos de Machine Learning haciendo uso de los Data-Sets que estén almacenados en el sistema. Este componente cumple con la siguientes funciones:
 - ***Entrenar Modelos de Machine Learning :*** Esta función realiza la mejora incremental para la predicción con Base en el Data-Set que contiene variables con resultados ya definidos de pacientes a los que ya se les detecto si el cáncer de mama era maligno o Benigno.
 - ***Listar Modelo Entrenados:*** Esta función realiza una lista de los modelos a los cuales ya se les realizo una mejora incremental de predicción para no repetir el proceso de entrenamiento cada vez que se requiere un diagnostico de cáncer de mama de diversos pacientes.

- 5) **Gestor de Predicción:** Este componente esta agregado al componente principal de BreastApp. En este componente se maneja el proceso de diagnosticar el padecimiento de cáncer de mama. Este componente cumple con la siguiente función:
- ***Diagnosticar Cáncer de mama:*** Esta función realiza el diagnostico de cáncer de mama con base al entrenamiento anteriormente realizado , las variables de nuevos pacientes a los que se requiere realizar el análisis ontológico y las diversas técnicas de predicción y decisión de Machine Learning.
- 6) **Gestor de Informes:** Este componente esta agregado al componente principal de BreastApp. En este componente se maneja el proceso generar informes con detalles del diagnóstico. Este componente cumple con la siguientes funciones:
- ***Generar Informes de Diagnósticos:*** Esta función genera los informes tipo reporte con base a los resultados arrojados por los diferentes Modelos en Machine Learning, en donde se da un resultado definitivo acerca del padecimiento de Cáncer de mama.
 - ***Almacenar Informes de Pacientes:*** Esta función almacena los diferentes informes determinados a los pacientes a los que se quiere determinar el padecimiento de cáncer de mama.
- 7) **Gestor de Modelos:** Este componente esta agregado al componente principal de BreastApp. En este componente se maneja el proceso implementación de modelos de Machine Learning. Este componente cumple con la siguientes funciones:
- ***Crear Modelos de Machine Learning:*** Esta función asocia los Data-Sets a los modelos de Machine Learning con los que realiza el respectivo diagnostico.
 - ***Listar Modelos de Machine Learning:*** Esta función permite obtener los modelos de Machine Learning que utiliza el aplicativo de BreastApp.

8.3. Punto de Vista de Cooperación de Aplicación

El Punto de Vista de cooperación de la aplicación describe las relaciones entre los componentes de las aplicaciones en términos de los flujos de información entre ellos o en términos de los servicios que ofrecen y utilizan.[47].

En la Figura 8.2 se plantea el Caso para el Punto de Vista de cooperación de la aplicación con cada uno de los elementos que interactúan entre sí.

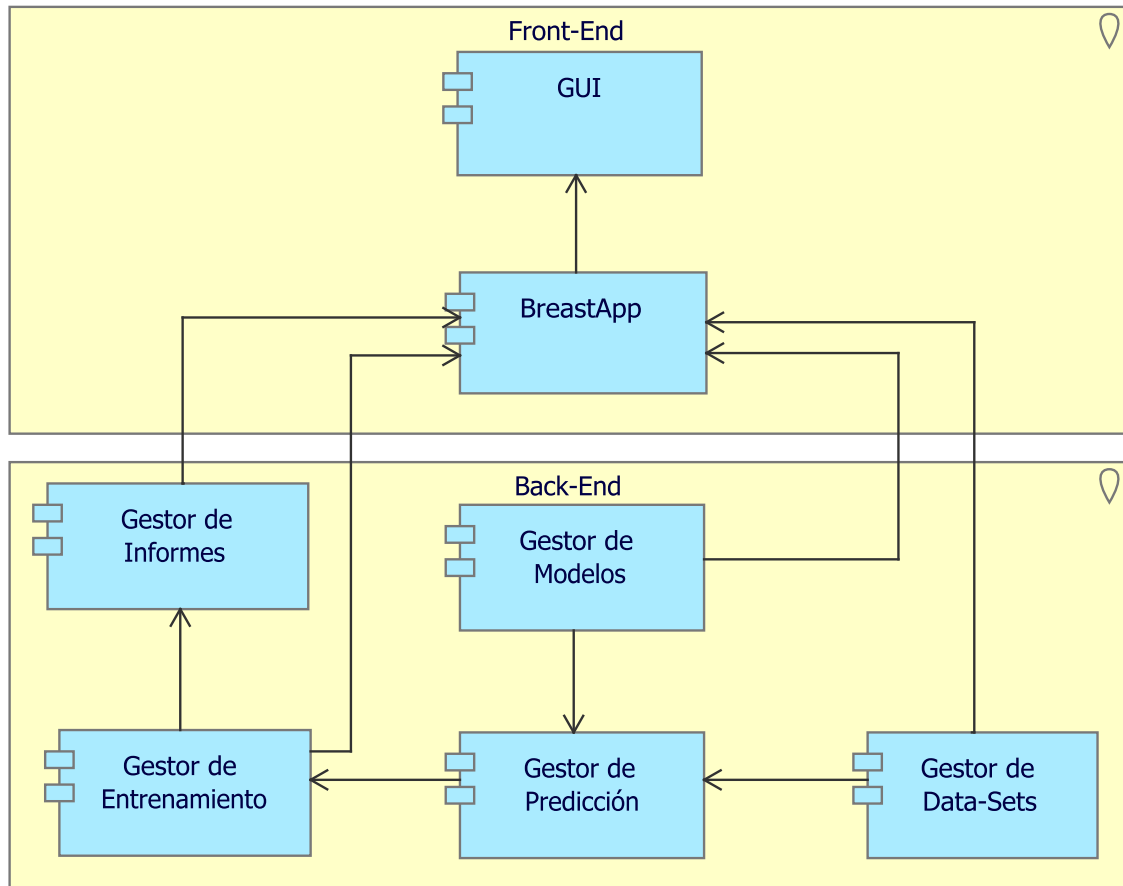


Figura 8.2: Punto de Vista de Cooperación de la Aplicación

- 1) **BreastApp:** El componente de BreastApp es el framework de la aplicación. En este se componente agregan los componentes correspondientes para el diagnóstico de cáncer de mama y se realiza la carga de modelos, carga de Data-Sets, entrenamiento de modelos y generación de informes sobre el cáncer de mama.

- 2) **GUI:** Este componente esta agregado al componente principal de BreastApp. Este componente hace referencia a la interfaz gráfica de usuario por la cual los investigadores y unidades funcionales del Instituto de Cancerología acceden a la aplicación, para llevar el análisis según el diagnostico de cáncer de mama generado por la aplicación.
- 3) **Gestor de Modelos:** Este componente es usado por el componente principal de BreastApp y el componente Gestor de entrenamiento. En este componente se maneja el proceso de crear modelos.
- 4) **Gestor de Predicción:** Este componente es usado por el componente de Gestor de Data-Sets y el componente Gestor de Modelos. En este componente se maneja el proceso de diagnosticar padecimiento de cáncer de mama.
- 5) **Gestor de Data-Sets:** Este componente es usado por el componente principal de BreastApp y el componente Gestor de Predicción . En este componente se maneja el proceso de carga y gestión de Data-Sets para ser almacenados en el sistema.
- 6) **Gestor de Entrenamiento:** Este componente es usado por el componente principal de BreastApp y el componente de Gestor de Predicción. En este componente se maneja el proceso de entrenamientos de los modelos de Machine Learning haciendo uso de los Data-Sets que estén almacenados en el sistema.
- 7) **Gestor de Informes:** Este componente es usado por el componente principal de BreastApp. En este componente se maneja el proceso Generar informes con detalles del diagnóstico.

8.4. Punto de Vista de la Estructura de la Aplicación

El Punto de Vista de estructura de la aplicación describe la estructura principal de aplicaciones o componentes y los datos asociados.[47].

En la Figura 8.3 se plantea el Caso para el Punto de Vista de la estructura de la aplicación con cada uno de los elementos que interactúan entre sí.

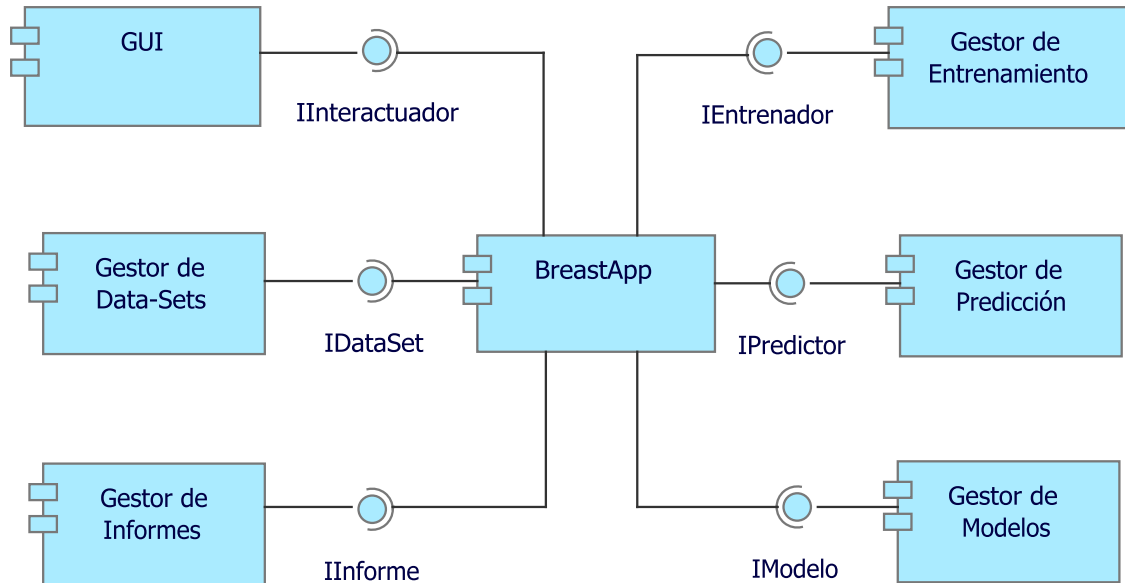


Figura 8.3: Punto de Vista de la Estructura de la Aplicación

- 1) **BreastApp:** El componente de BreastApp es el framework de la aplicación. En este se componente agregan los componentes correspondientes para el diagnóstico de cáncer de mama y se realiza la carga de modelos, carga de Data-Sets, entrenamiento de modelos y generación de informes sobre el cáncer de mama.
- 2) **GUI:** Este componente esta agregado al componente principal de BreastApp. Este componente hace referencia a la interfaz gráfica de usuario por la cual los investigadores y unidades funcionales del Instituto de Cancerología acceden a la aplicación, para llevar el análisis según diagnostico de cáncer de mama generado por la aplicación, a través de la interfaz *IInteractuador*.
- 3) **Gestor de Data-Sets:** Este componente ofrece un servicio a través de la interfaz *IDataSet* al componente principal de BreastApp. En este componente se maneja el proceso de carga y gestión de Data-Sets para ser almacenados en el sistema.

- 4) **Gestor de Informes:** Este componente ofrece un servicio a través de la interfaz *IInforme* al componente principal de BreastApp. Este componente es usado por el componente principal de BreastApp. En este componente se maneja el proceso Generar informes con detalles del diagnóstico de cada paciente.
- 5) **Gestor de Entrenamiento:** Este componente ofrece un servicio a través de la interfaz *IEntrenador* al componente principal de BreastApp. En este componente se maneja el proceso de entrenamientos de los modelos de Machine Learning haciendo uso de los Data-Sets que estén almacenados en el sistema.
- 6) **Gestor de Predicción:** Este componente ofrece un servicio a través de la interfaz *IPredictor* al componente principal de BreastApp. En este componente se maneja el proceso de diagnosticar padecimiento de cáncer de mama.
- 7) **Gestor de Modelos:** Este componente ofrece un servicio a través de la interfaz *IModelo* al componente principal de BreastApp. En este componente se maneja el proceso de crear modelos.

8.5. Punto de Vista del Uso de la Aplicación

El Punto de Vista del uso de la aplicación describe el diseño de una aplicación mediante la identificación de los servicios necesarios por los procesos de negocio y otras aplicaciones, o en el diseño de procesos de negocio mediante la descripción de los servicios que están disponibles [47].

En la Figura 8.4 se plantea el Caso para el Punto de Vista del uso de la aplicación con cada uno de los elementos que interactúan entre sí.

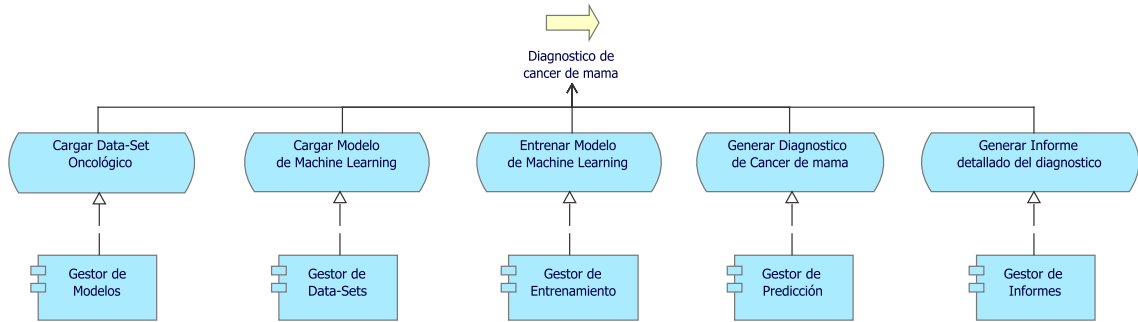


Figura 8.4: Punto de Vista del uso de la Aplicación

- 1) **Diagnóstico de cáncer de mama:** Este proceso consiste en el diagnóstico de padecimiento de cáncer de mama. Es logrado mediante cinco servicios: *Cargar modelo*, *Cargar Data-Set*, *Entrenar Modelo*, *Generar diagnóstico* y *Generar informe detallado del diagnóstico*.
- 2) **Cargar Data-Set Oncológico:** Este servicio de la aplicación recibe los Data-Set con las variables oncológicas en formato .csv, los valida y los almacena en el sistema. Este servicio es realizado por el componente *Gestor de Data-Sets*.
- 3) **Cargar Modelo de Machine Learning:** Este servicio de la aplicación es el encargado de crear modelos de Machine Learning. Este servicio es realizado por el componente *Gestor de Modelos*.
- 4) **Generar Diagnostico de cáncer de mama:** Este servicio de aplicación recibe los datos de un paciente en particular y realiza el diagnóstico. El servicio responde si según los datos tiene un diagnóstico Benigno y Maligno. Este servicio es realizado por el componente *Gestor de Predicción*.
- 5) **Generar Informe Detallado del Diagnostico:** Este servicio genera un informe detallado en formato .pdf con la información respectiva de las predicciones y gráficos de cada cada modelo. Este servicio es realizado por el componente *Gestor de Informes*.

Capítulo 9

Capa de Tecnología

9.1. Introducción

Esta capa toma, en primer lugar, la utilización de conceptos como Nodo y Dispositivo para identificar las diferentes entidades físicas o recursos computacionales de la infraestructura del sistema [47].

Esta capa busca establecer y conocer los diferentes elementos a nivel de software y hardware que interactúan en los procesos de negocio establecidos anteriormente, por lo tanto, es importante tener en cuenta los conceptos utilizados en la capa de negocio y en la capa de aplicación para así facilitar la identificación de los diferentes usos e interacciones que se presentan con esta capa. Para una adecuada apropiación de la conceptualización de esta capa, se implementa el uso de seis diferentes puntos de vista: infraestructura, uso de infraestructura, implementación y organización, estructura de información, realización de servicio y capas [47].

Por otra parte, es importante resaltar que, dentro de la infraestructura, un aspecto fundamental corresponde a las interacciones entre los diferentes elementos a nivel de software (componentes) y a nivel de hardware (dispositivos), que determinan la comunicación entre los conceptos de aplicación y conceptos de negocio mencionados en las dos capas anteriores[47].

A continuación se presentan cada uno de los puntos de vista de la capa de tecnología a partir del soporte realizado por el Área de Investigación de Análisis de datos a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología.

9.2. Punto de Vista de la Infraestructura

El Punto de Vista de la Infraestructura contiene los elementos de infraestructura de software y hardware que soportan la capa de aplicación, como dispositivos físicos, redes o software del sistema[47].

En la Figura 9.1, se plantea el Caso para el Punto de Vista de la infraestructura con cada uno de los elementos que interactúan entre sí.

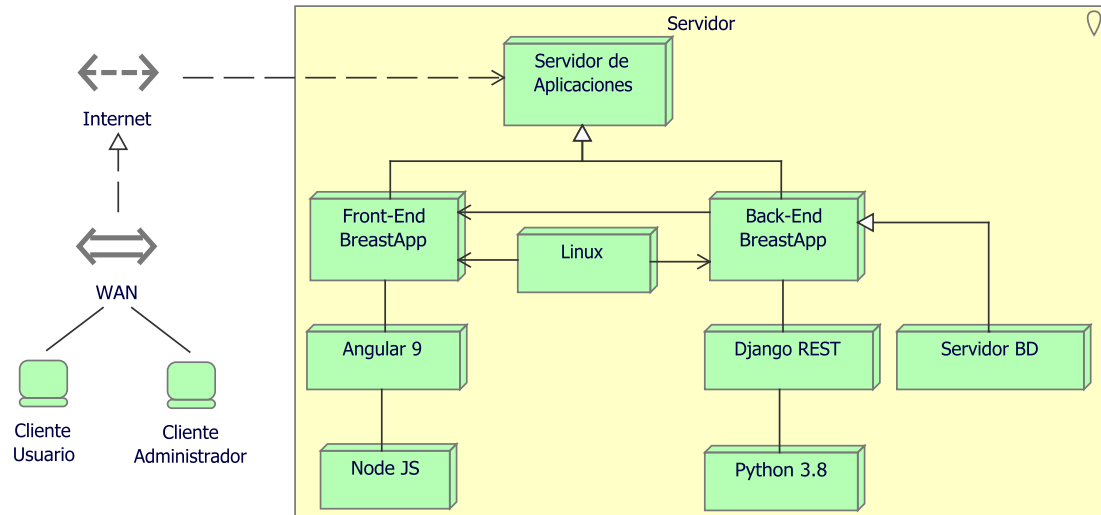


Figura 9.1: Punto de Vista de la Infraestructura

- 1) **Servidor de Aplicaciones:** Hace referencia a el nodo que se conecta directamente a la red para exponer la disponibilidad de la aplicación. Es el nodo general donde interactúan tanto componentes, como sistemas de software y demás elementos necesarios para la aplicación.
- 2) **Front-End BreastApp:** Corresponde a la interfaz gráfica de usuario por la cual los investigadores y unidades funcionales del Instituto de Cancerología acceden a la aplicación, para llevar el análisis según diagnostico de cáncer de mama generado por la aplicación.
- 3) **Angular 9:** Hace referencia a el framework seleccionado para desarrollar la interfaz gráfica de usuario de la aplicación BreastApp.
- 4) **Node JS:** Es el entorno basado en JavaScript imprescindible para poder usar el framework de Angular 9.

- 5) **Back-End BreastApp:** Hace referencia a el desarrollo de la lógica para la generación de la capa de servicios REST para las diferentes funcionalidades necesarias para generar el diagnostico relacionado con el cáncer de mama.
- 6) **Django:** Hace referencia a el framework seleccionado para desarrollar la capa de servicios REST para la aplicación BreastApp.
- 7) **Python 3.8:** Corresponde a el lenguaje de programación seleccionado para el desarrollo del Back-End de la aplicación BreastApp.
- 8) **Linux:** Este sistema de software es el sistema operativo donde se desarrolla el Back-End y Front-End de la aplicación BreastApp.
- 9) **Servidor BD:** Este sistema de software hace referencia a la persistencia de los datos de la aplicación BreastApp. Se realiza automáticamente en el Back-End de la aplicación por medio del framework Django.
- 10) **Internet:** Hace referencia a el acceso a la aplicación través de la red, debido a que los investigadores y unidades funcionales del Instituto de Cancerología podrán acceder desde sus casas o lugares externos.
- 11) **WAN:** Debido a que los usuarios pueden acceder desde cualquier punto, es necesario una red de área amplia que reúna varias redes locales y brinde la conexión a los investigadores y unidades funcionales del Instituto de Cancerología.
- 12) **Cliente:** En la aplicación BreastApp, existen dos tipos de clientes:
 - **Cliente Usuario:** Hace referencia a los investigadores y unidades funcionales del Instituto de Cancerología.
 - **Cliente Administrador:** Hace referencia a las personas autorizadas de administrar el uso y funcionamiento de la aplicación BreastApp, así como utilizar la información del Instituto de Cancerología de manera pertinente.

9.3. Punto de Vista del Uso de la Infraestructura

El Punto de Vista del uso de la Infraestructura describe como las aplicaciones son compatibles con la infraestructura de software y hardware. Es muy útil para determinar los requisitos de rendimiento y calidad de la infraestructura basados en las demandas de las diversas aplicaciones que la utilizan [47].

En la Figura 9.2, se plantea el Caso para el Punto de Vista del Uso de la infraestructura con cada uno de los elementos que interactúan entre sí.

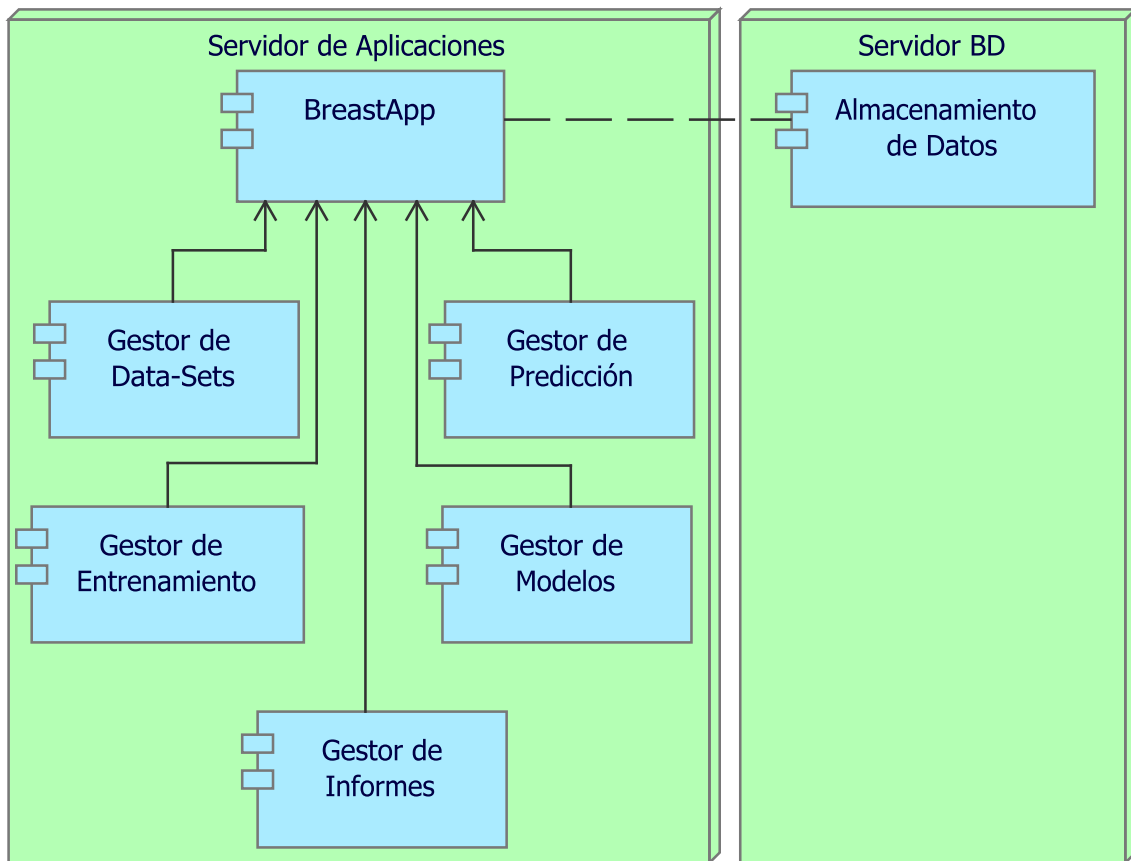


Figura 9.2: Punto de Vista del Uso de la Infraestructura

- 1) **Servidor de Aplicaciones:** En este servidor, interactúan diferentes componentes de la aplicación de BreastApp y se comunica con mas servidores a través de dichos componentes. Los componentes se describen a continuación:

- **Gestor de Modelos:** Este componente es usado por el componente principal de BreastApp. En este componente se maneja el proceso Crear modelos.
 - **Gestor de Predicción:** Este componente es usado por el componente principal de BreastApp. En este componente se maneja el proceso de diagnosticar el padecimiento de cáncer de mama.
 - **Gestor de Data-Sets:** Este componente es usado por el componente principal de BreastApp. En este componente se maneja el proceso de carga y gestión de Data-Sets para ser almacenados en el sistema.
 - **Gestor de Entrenamiento:** Este componente es usado por el componente principal de BreastApp. En este componente se maneja el proceso de entrenamientos de los modelos de Machine Learning haciendo uso de los Data-Sets que estén almacenados en el sistema.
 - **Gestor de Informes:** Este componente es usado por el componente principal BreastApp. En este componente se maneja el proceso de generar informes con detalles del diagnóstico.
- 2) **Servidor BD:** Este sistema de software hace referencia a la persistencia de los datos de la aplicación BreastApp. Se realiza automáticamente en el Back-End de la aplicación por medio del framework Django. Este servidor hace uso del componente que se describe a continuación:
- **Almacenamiento de Datos:** Este componente hace referencia a la persistencia de la aplicación BreastApp. En este se desarrollan las funciones CRUD de la Capa de servicios REST implementada en el Back-End de la aplicación.

9.4. Punto de Vista de Implementación y Organización

El Punto de Vista de implementación y Organización describe como se despliegan una o más aplicaciones en la infraestructura. Este punto de vista comprende la asignación de aplicaciones y componentes lógicos a artefactos físicos [47].

En la Figura 9.3, se plantea el Caso para el Punto de Vista de Implementación y Organización con cada uno de los elementos que interactúan entre sí.

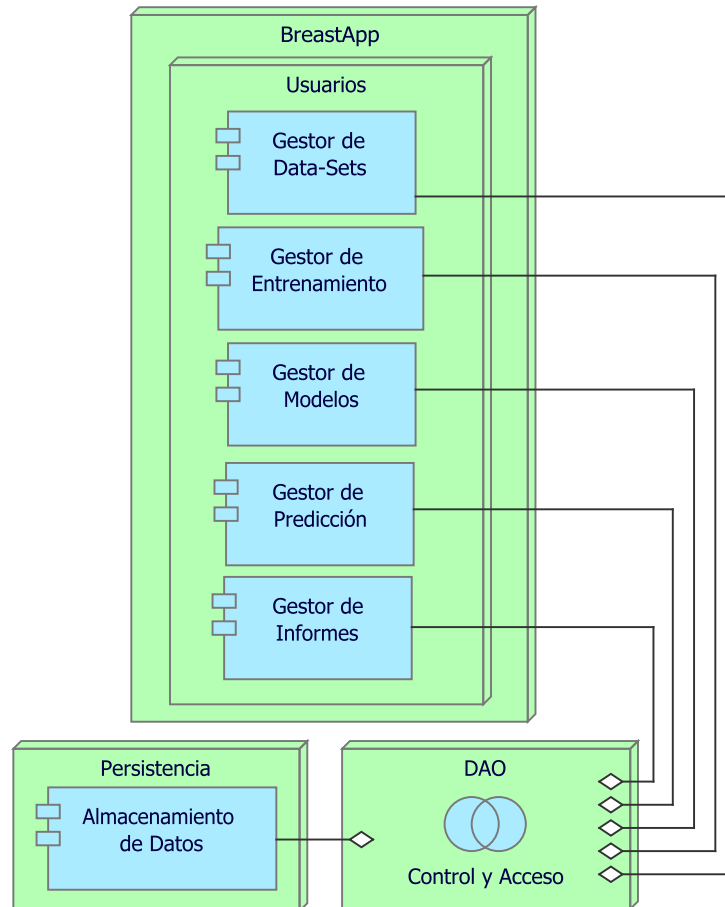


Figura 9.3: Punto de Vista de Implementación y Organización

- 1) **BreastApp**: Este sistema de Software representa a la aplicación en donde se desarrollan los componentes correspondientes para el diagnóstico de cáncer de mama. En este caso se tiene un nodo el cual interactúa con los componentes de este sistema.

- 2) **Usuarios:** Este nodo contiene los componentes que tienen directa relación con los investigadores y unidades funcionales del Instituto de Cancerología y sus diferentes funciones. En este caso, se especifica un componente agregado a la colaboración de aplicación *Control y Acceso*.
- **Gestor de Modelos:** En este componente se maneja el proceso de crear modelos de Machine Learning.
 - **Gestor de Predicción:** En este componente se maneja el proceso de diagnosticar el padecimiento de cáncer de mama.
 - **Gestor de Data-Sets:** En este componente se maneja el proceso de carga y gestión de Data-Sets para ser almacenados en el sistema.
 - **Gestor de Entrenamiento:** En este componente se maneja el proceso de entrenamientos de los modelos de Machine Learning haciendo uso de los Data-Sets que estén almacenados en el sistema.
 - **Gestor de Informes:** En este componente se maneja el proceso de generar informes con detalles del diagnóstico.
- 3) **Persistencia:** Este nodo hace referencia a la persistencia de la aplicación. Tiene el componente encargado de el almacenamiento y uso de información. Además, esta agregado a la colaboración de aplicación *Control y Acceso*.
- **Almacenamiento de Datos:** Este componente hace referencia a la persistencia de la aplicación BreastAPP. En este se desarrollan las funciones CRUD de la Capa de servicios REST implementada en el Back-End de la aplicación.
- 4) **DAO:** Este sistema de software se refiere al componente que suministra la interfaz de comunicación entre la aplicación BreastApp y la base de datos. En este caso, se hace a través de una colaboración de aplicación, que se describe a continuación.
- **Control y Acceso:** Esta colaboración trata del control de la información de las diferentes funciones de la aplicación BreastApp y su almacenamiento, y el acceso a dicha información. Tiene agregado los componentes de *Gestor de Modelos*, *Gestor de Predicción*, *Gestor de Data-Sets*, *Gestor de Entrenamiento*, *Gestor de Informes* y *Almacenamiento de Datos*. Esta colaboración es el puente para el CRUD de la Capa de servicios REST implementada en el Back-End de la aplicación BreastApp y el almacenamiento de los datos.

9.5. Punto de Vista de Estructura de la Información

El Punto de Vista de Estructura de la Información describe la estructura de la información usada en la empresa o en un proceso específico de negocio o aplicación, en términos de tipos de datos o estructuras de clases [47].

En la Figura 9.4, se plantea el Caso para el Punto de Vista de Estructura de la Información con cada uno de los elementos que interactúan entre sí.

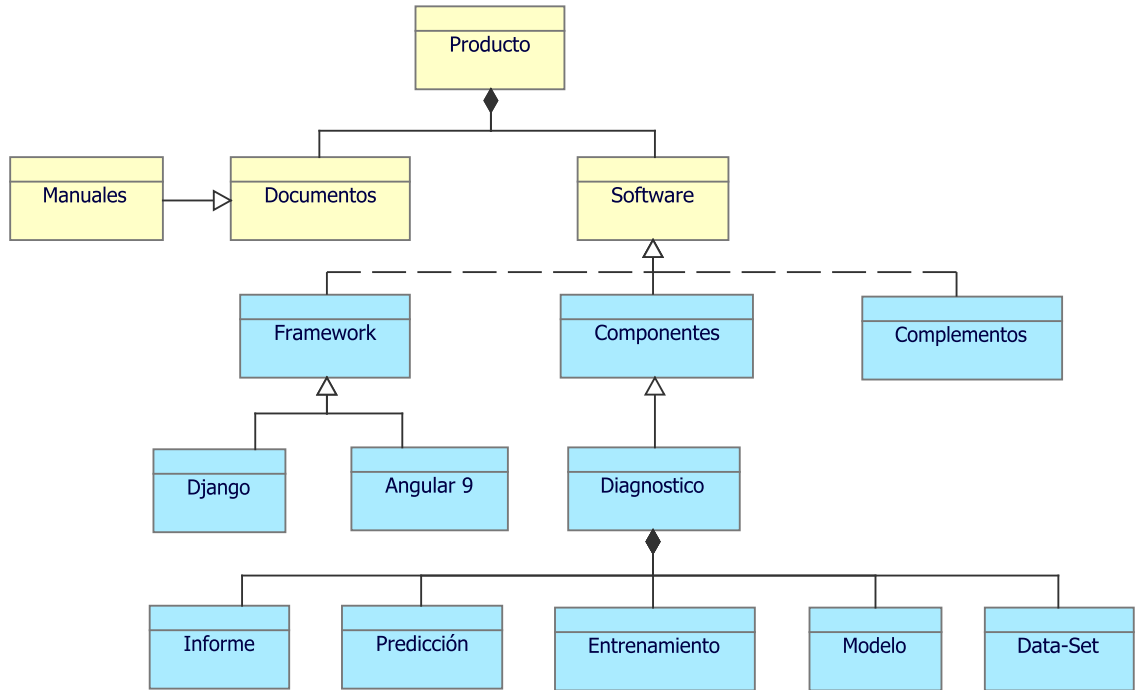


Figura 9.4: Punto de Vista de Estructura de Información

- 1) **Producto:** Este objeto de negocio se refiere al producto de software que se tiene en su totalidad. Este a su vez está compuesto por *Documentos* y *Software*.
- 2) **Documentos:** Este objeto de negocio que hace referencia a toda la documentación correspondiente al producto como los *Manuales*. Todo esto se va desarrollando gradualmente y se obtiene con el producto final.
- 3) **Software:** Este objeto de negocio hace referencia a la implementación en software de la aplicación BreastApp. Este a su vez es realizado por diferentes objetos de negocio: *Framework*, *Componentes* y *Complementos*.

- 4) **Framework:** En este objeto se desarrolla el software BreastApp. Esta conformado por el Framework *Django* en el cual se realiza toda la lógica de la capa de servicios REST implementada en el Back-End y el Framework *Angular 9* en el cual se pueden visualizar todas las funciones de la aplicación implementadas en el Front-End.
- 5) **Componentes:** Este objeto contiene los diferentes componentes que interactúan en la aplicación BreastApp. Es una especialización del objeto *Diagnostico* el cual es el resultado final de la aplicación y que se encuentra compuesto por los objetos *Informe, Predicción, Entrenamiento, Modelo* y *Data-Set*.
- 6) **Complementos:** Este objeto que se refiere a los diferentes complementos utilizados en la aplicación BreastApp. En este caso se tienen complementos para la generación de la persistencia de los datos, el diseño del Front y la publicación y consumo de la capa de servicios de la aplicación.

9.6. Punto de Vista de Realización del Servicio

El Punto de Vista de Realización de Servicio describe cómo uno o más servicios de negocio son realizados por un proceso fundamental o, algunas veces, por componentes de aplicación. Esto forma el puente entre productos y las vistas de los procesos de negocio [47].

En la Figura 9.5, se plantea el Caso para el Punto de Vista de Realización del Servicio con cada uno de los elementos que interactúan entre sí.

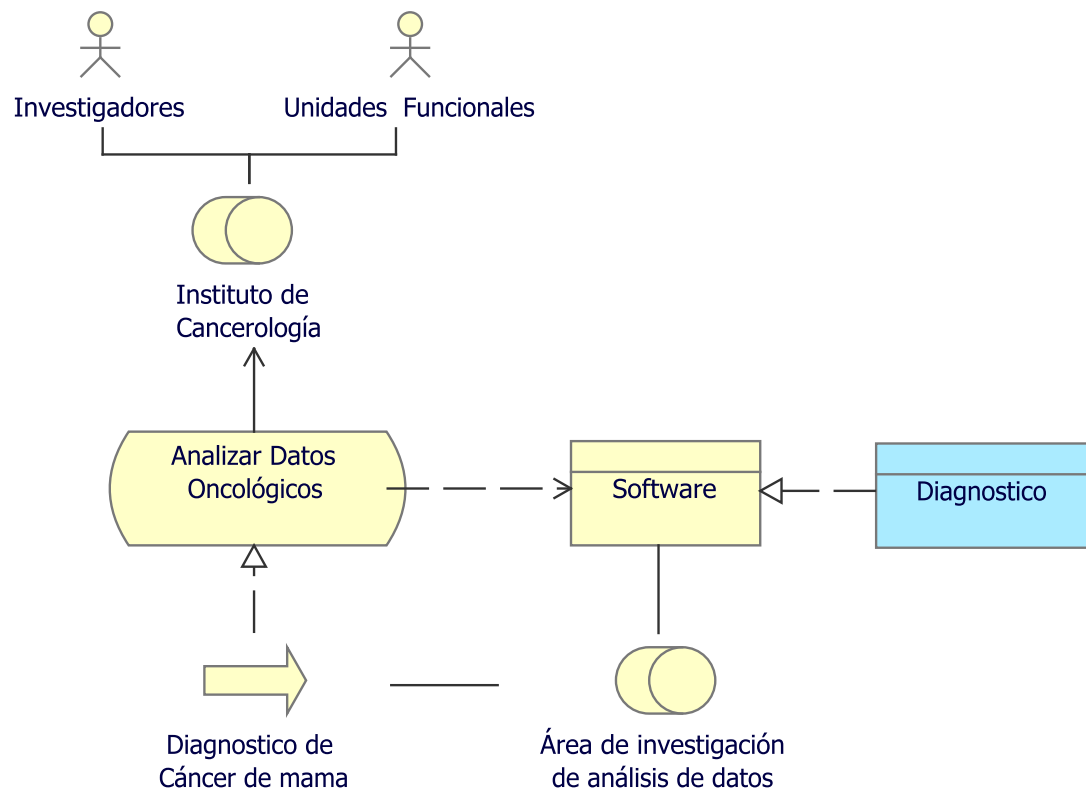


Figura 9.5: Punto de Vista de Realización de Servicio

- 1) **Analizar Datos Oncológicos:** Este servicio realiza el análisis de los datos entregados por los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología para generar diagnósticos oncológicos. Es realizado por el proceso de *Diagnostico de Cáncer de mama* y usado por el *Instituto de Cancerología*.

- 2) **Diagnostico de Cáncer de mama:** Este proceso cumple el objetivo dar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología con respecto al análisis de datos oncológicos. Tiene asignado el rol *Área de investigación de análisis de datos*.
- 3) **Área de investigación de análisis de datos:** Corresponde al área de investigación la cual da soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología por medio reportes generados con base a los datos solicitados. Esta asignado al proceso *Diagnostico de Cáncer de mama*.
- 4) **Instituto de Cancerología :** Este rol de aplicación corresponde a los diferentes usuarios del Instituto de Cancerología. Tienes asociado el actor *Investigadores* y el actor *Unidades funcionales* . Estos roles se describen a continuación:
 - **Investigadores :** Este rol está conformado por todos los grupos de investigación en cáncer del país registrados ante Colciencias y adicionalmente, con representantes de diferentes tipos de usuarios del conocimiento generado por la investigación como son las sociedades médicas, los prestadores de servicios oncológicos, los aseguradores, las autoridades sanitarias y los pacientes entre otros.
 - **Unidades Funcionales :** Este rol está conformado por las unidades clínicas ubicadas al interior del Instituto de Cancerología cuya función es evaluar la situación de salud del paciente con diagnóstico presuntivo de cáncer.
- 5) **Software:** Este objeto de negocio hace referencia a la implementación en software de la aplicación BreastApp. Es realizado por el objeto *Diagnostico*.

9.7. Punto de Vista de Capas

El Punto de Vista de capas describe varios los aspectos de una arquitectura empresarial en un diagrama como resultado del uso de la relación de agrupación para una partición natural de todo el conjunto de objetos y relaciones que pertenecen a el modelo [47].

En la Figura 9.6, se plantea el Caso para el Punto de Vista de Capas con cada uno de los elementos que interactúan entre sí.

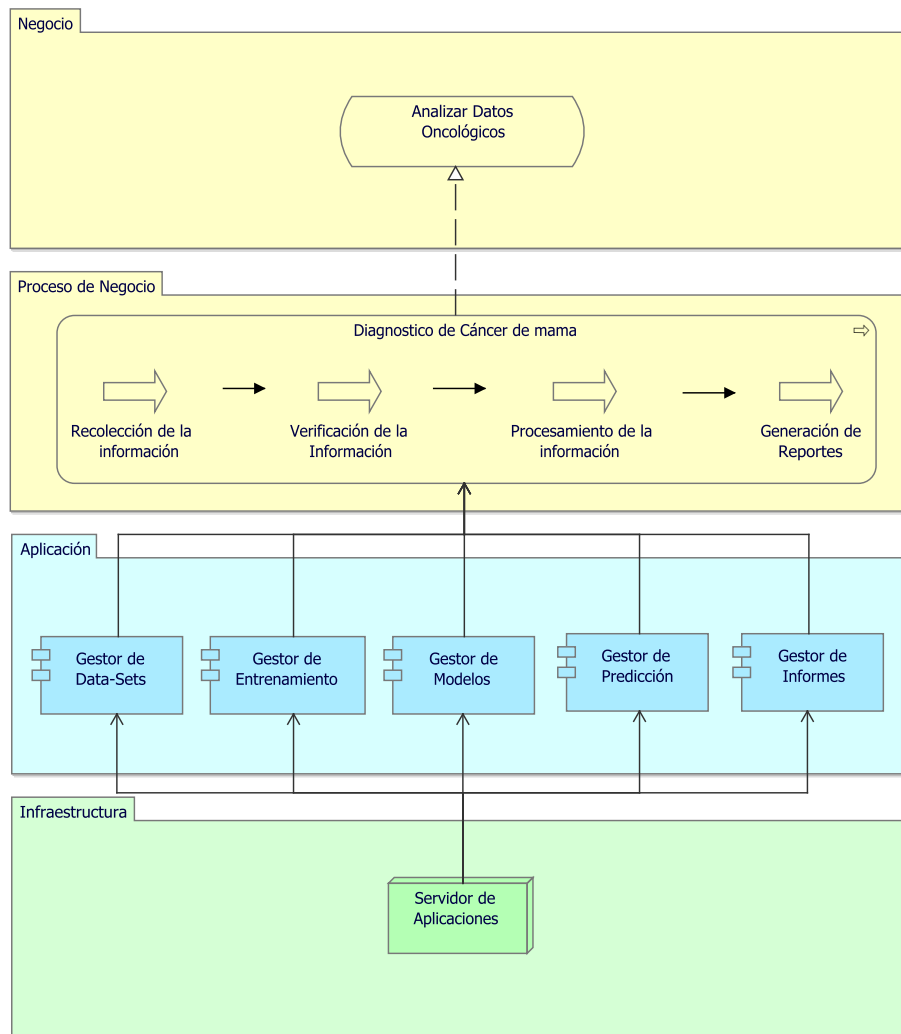


Figura 9.6: Punto de Vista de Capas

- 1) **Negocio:** Esta capa expone un servicio importante el cual fue planteado con anterioridad en la Capa de Negocio. Estos servicio es realizado por un proceso específico, expuesto en la siguiente capa del modelo de Proceso de Negocio. La descripción de los servicios se expone a continuación:
 - **Analizar Datos Oncológicos:** Este servicio realiza el análisis de los datos entregados por los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología para generar diagnósticos oncológicos. Es realizado por el proceso de *Diagnostico de Cáncer de mama*.
- 2) **Proceso de Negocio:** Esta capa, expone los procesos encargados de realizar los servicios expuestos en el entorno. Son usados por distintos componentes planteados en la capa de Aplicación. La descripción de los procesos se expone a continuación.
 - **Diagnostico de Cáncer de mama:** Este proceso cumple el objetivo dar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología con respecto al análisis de datos oncológicos. Esta compuesto de los siguientes procesos de negocio secuenciales: *Recolección de información, Verificación de la información, Procesamiento de la información y Generación de Reportes*.
- 3) **Aplicación:** Esta capa, plantea los diferentes componentes que usan los procesos de negocio expuestos en la capa de Aplicación. A su vez, son usados en la capa siguiente, a través de un nodo. Esta capa esta conformada de por los siguientes componentes:
 - **Gestor de Modelos:** En este componente se maneja el proceso de crear modelos de Machine Learning.
 - **Gestor de Predicción:** En este componente se maneja el proceso de diagnosticar el padecimiento de cáncer de mama.
 - **Gestor de Data-Sets:** En este componente se maneja el proceso de carga y gestión de Data-Sets para ser almacenados en el sistema.
 - **Gestor de Entrenamiento:** En este componente se maneja el proceso de entrenamientos de los modelos de Machine Learning haciendo uso de los Data-Sets que estén almacenados en el sistema.
 - **Gestor de Informes:** En este componente se maneja el proceso generar informes con detalles del diagnóstico.
- 4) **Infraestructura:** En esta Capa, se plantea el nodo Servidor de Aplicaciones, que usa los diferentes componentes planteados en la Capa de Aplicación. El nodo, se describe a continuación:

- **Servidor de Aplicaciones:** Hace referencia a el nodo que se conecta directamente a la red para exponer la disponibilidad de la aplicación. Es el nodo general donde interactúan tanto componentes, como sistemas de software y demás elementos necesarios para la aplicación. Los componentes usados de la capa de aplicación por este nodo son: *Gestor de Modelos*, *Gestor de Predicción*, *Gestor de Data-Sets*, *Gestor de Entrenamiento* y *Gestor de Informes*.

Capítulo 10

Capa de Migración e Implementación

10.1. Introducción

Esta capa describe la implementación de modelos y características de migración para ejecutar un cambio en la arquitectura del negocio. Cada uno de los puntos de vista que contiene esta capa presenta una perspectiva para modelar la gestión del cambio de arquitectura, modelar la transición de una arquitectura existente a una arquitectura de destino y definir las relaciones entre los programas y proyectos y las partes de la arquitectura que implementan [48].

Esta capa da conocer el contexto sobre el cual se busca modelar el cambio estructural de la arquitectura empresarial. Para una adecuada apropiación de la conceptualización de esta capa, se implementa el uso de tres diferentes puntos de vista (proyecto, migración y migración e implementación). En esta capa se retoman algunos de los conceptos utilizados en la capa de negocio como por ejemplo el concepto de objeto de negocio representado en esta capa como un Paquete de Trabajo (Work Package) y el concepto de representación el cual tiene como equivalente un Liberable (Derivable)[47].

A continuación se presentan cada uno de los puntos de vista de la capa de Implementación a partir del soporte realizado por el Área de Investigación de Análisis de datos a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología.

10.2. Punto de Vista de Proyecto

El punto de vista de proyecto se utiliza principalmente para modelar la gestión del cambio de la arquitectura permitiendo al diseñador o analista plasmar mediante un diagrama el proceso de interacción de los diferentes elementos[47].

En la Figura 10.1 se plantea el Caso para el Punto de Vista de Proyecto con cada uno de los elementos que interactúan entre sí.

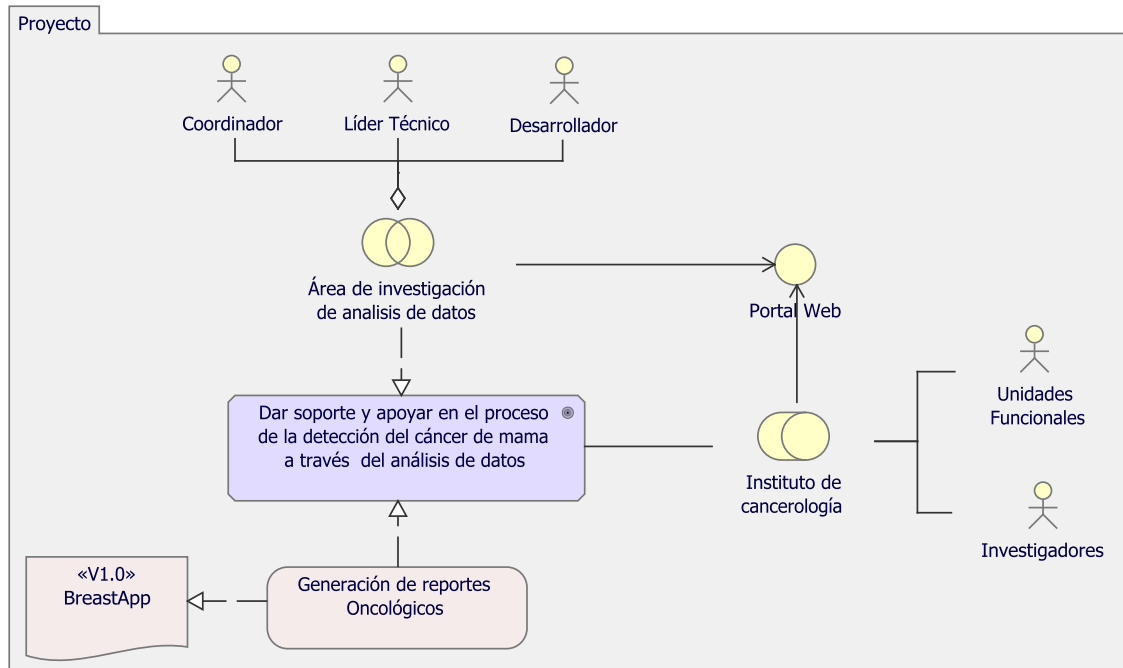


Figura 10.1: Punto de Vista de Proyecto

- 1) **Instituto de Cancerología:** Este elemento hace referencia a el rol de la organización el cual cuenta con siete grupos de investigación que desarrollan diversas actividades según su campo de acción: investigación clínica, investigación epidemiológica, biología del cáncer e investigación en el área de la salud pública. A este rol están asociados el actor *Investigador* y el actor *Unidades funcionales*. Estos roles se describen a continuación:

- **Investigadores** : Este rol está conformado por todos los grupos de investigación en cáncer del país registrados ante Colciencias y adicionalmente, con representantes de diferentes tipos de usuarios del conocimiento generado por la investigación como son las sociedades médicas, los prestadores de servicios oncológicos, los aseguradores, las autoridades sanitarias y los pacientes entre otros.
- **Unidades Funcionales** : Este rol está conformado por las unidades clínicas ubicadas al interior del Instituto de Cancerología cuya función es evaluar la situación de salud del paciente con diagnóstico presuntivo de cáncer.

2) **Área de Investigación de análisis de datos:** Corresponde a una de las dependencias del área de investigación la cual tiene por objetivo dar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología, haciendo uso de métodos computacionales para el manejo y análisis de datos Oncológicos. Este actor tiene *agregado* tres actores los cuales se muestran a continuación:

- **Coordinador:** Es la persona encargada de regular, gestionar, dirigir y supervisar el Área de Investigación de análisis de datos de Oncología para que el soporte realizado a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales se realice correctamente.
- **Líder Técnico:** Es la persona con conocimiento técnico avanzado en los temas de análisis de datos de Oncología y es el responsable de asignar y definir las tareas y el tiempo necesario en el desarrollo e implementación de los recursos según las necesidades presentadas por los investigadores y las unidades funcionales del Instituto de Cancerología.
- **Desarrollador:** Es la persona encargada de cumplir con las implementaciones presentadas en el ámbito de análisis de datos de Oncología y que da solución a cada uno de los requerimientos y necesidades presentadas por los investigadores y las unidades funcionales del Instituto de Cancerología.

3) **Portal Web** : Es el medio de comunicación entre el Área de Investigación de análisis de datos y los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología. Esta interfaz tiene una asociación con el rol del *Área de Investigación de análisis de datos* y es usada por el rol *Instituto de Cancerología*.

4) **Dar soporte y apoyar en el proceso de la detección del cáncer de mama a través del análisis de datos:** Este objetivo organizacional hace referencia a el servicio de generar diagnósticos relacionados con el cáncer de mama solicitados por los diferentes funcionarios del Instituto de Cancerología.

- 5) **Generación de reportes oncológicos:** Es el paquete de trabajo principal del instituto de Cancerología basado en la estrategia planteada y los recursos disponibles identificados. La meta es construir una aplicación web que genere informes tipo reporte con base a los resultados arrojados por diferentes modelos de Machine Learning, en donde se de un resultado definitivo acerca del padecimiento de Cáncer de mama. Este diagnostico se realiza con el propósito de dar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología.
- 6) **BreastApp:** Representa el liberable, consecuencia del paquete de trabajo definido anteriormente. Este producto corresponde a la generación de reportes diagnósticos según el análisis de datos oncológicos elaborados por los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología. Es el resultado obtenido que cumple con los objetivos planteados y satisface las necesidades de la organización.

10.3. Punto de Vista de Migración

El Punto de Vista de Migración implica modelos y conceptos que pueden usarse para especificar la transición de una arquitectura existente a una arquitectura deseada.[47].

En la Figura 10.2 se plantea el Caso para el Punto de Vista de Migración con cada uno de los elementos que interactúan entre sí.

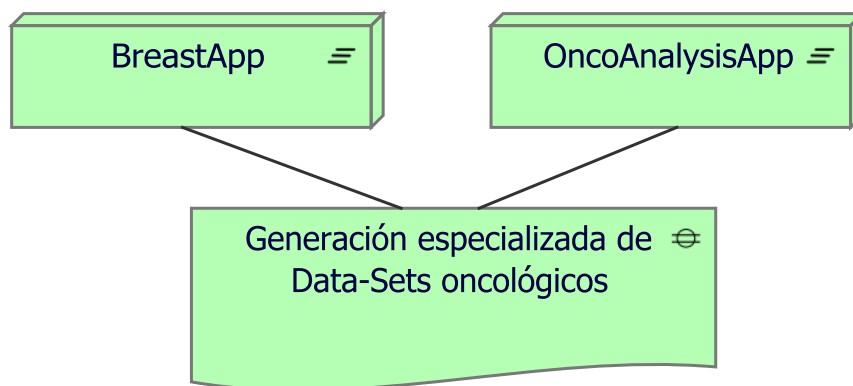


Figura 10.2: Punto de Vista de Migración

- 1) **BreastApp:** Es la Meseta inicial que corresponde a una aplicación web la cual genera reportes diagnósticos según el análisis de datos relacionados con el Cáncer de mama elaborados para brindar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología.
- 2) **Generación especializada de Data-Sets oncológicos:** Hace referencia a la brecha que debe resolverse para poder generar la Meseta futura *OncoAnalysisApp*. Esta brecha está relacionada con la generación dinámica de Data-Sets para cualquier tipo de cáncer.
- 3) **OncoAnalysisApp:** Es la Meseta final de la aplicación. Representa un producto a futuro con nuevas implementaciones y funcionalidades adicionales. En este caso se pretende generar diagnósticos para cualquier tipo de cáncer para brindar un soporte mas amplio a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología.

10.4. Punto de Vista de Migración e implementación

El Punto de Vista de Migración e Implementación se utiliza para relacionar programas y proyectos con las partes de la arquitectura que implementan[47].

En la Figura 10.3 se plantea el Caso para el Punto de Vista de Migración con cada uno de los elementos que interactúan entre sí.

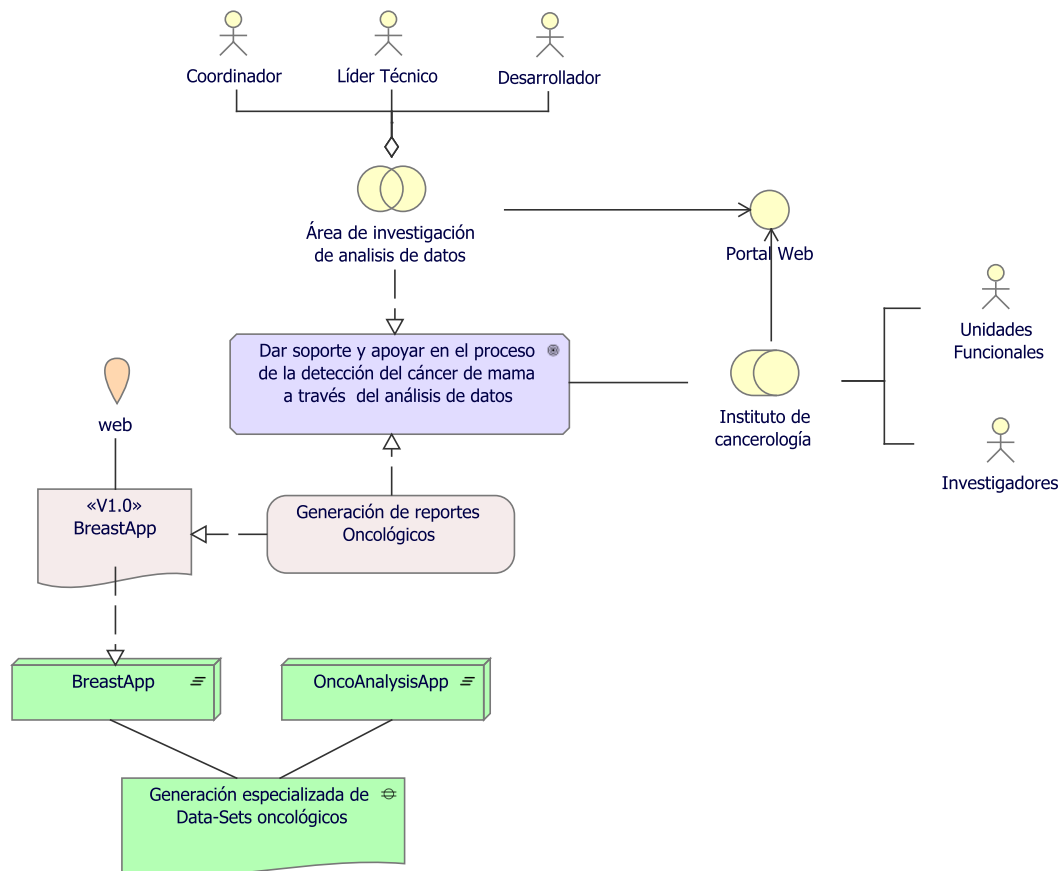


Figura 10.3: Punto de Vista de Migración e Implementación

- 1) **Instituto de Cancerología:** Este elemento hace referencia a el rol de la organización el cual cuenta con siete grupos de investigación que desarrollan diversas actividades según su campo de acción: investigación clínica, investigación epidemiológica, biología del cáncer e investigación en el área de la salud pública. A este rol están asociados el actor *Investigador* y el actor *Unidades funcionales*. Estos roles se describen a continuación:

- **Investigadores** : Este rol está conformado por todos los grupos de investigación en cáncer del país registrados ante Colciencias y adicionalmente, con representantes de diferentes tipos de usuarios del conocimiento generado por la investigación como son las sociedades médicas, los prestadores de servicios oncológicos, los aseguradores, las autoridades sanitarias y los pacientes entre otros.
 - **Unidades Funcionales** : Este rol está conformado por las unidades clínicas ubicadas al interior del Instituto de Cancerología cuya función es evaluar la situación de salud del paciente con diagnóstico presuntivo de cáncer.
- 2) **Área de Investigación de análisis de datos:** Corresponde a una de las dependencias del área de investigación la cual tiene por objetivo dar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología, haciendo uso de métodos computacionales para el manejo y análisis de datos Oncológicos. Este actor tiene *agregado* tres actores los cuales se muestran a continuación:
- **Coordinador:** Es la persona encargada de regular, gestionar, dirigir y supervisar el Área de Investigación de análisis de datos de Oncología para que el soporte realizado a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales se realice correctamente.
 - **Líder Técnico:** Es la persona con conocimiento técnico avanzado en los temas de análisis de datos de Oncología y es el responsable de asignar y definir las tareas y el tiempo necesario en el desarrollo e implementación de los recursos según las necesidades presentadas por los investigadores y las unidades funcionales del Instituto de Cancerología.
 - **Desarrollador:** Es la persona encargada de cumplir con las implementaciones presentadas en el ámbito de análisis de datos de Oncología y que da solución a cada uno de los requerimientos y necesidades presentadas por los investigadores y las unidades funcionales del Instituto de Cancerología.
- 3) **Portal Web** : Es el medio de comunicación entre el Área de Investigación de análisis de datos y los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología. Esta interfaz tiene una asociación con el rol del *Área de Investigación de análisis de datos* y es usada por el rol *Instituto de Cancerología*.
- 4) **Dar soporte y apoyar en el proceso de la detección del cáncer de mama a través del análisis de datos:** Este objetivo organizacional hace referencia a el servicio de generar diagnósticos relacionados con el cáncer de mama solicitados por los diferentes funcionarios del Instituto de Cancerología.

- 5) **Generación de reportes oncológicos:** Es el paquete de trabajo principal del instituto de Cancerología basado en la estrategia planteada y los recursos disponibles identificados. La meta es construir una aplicación web que genere informes tipo reporte con base a los resultados arrojados por diferentes modelos de Machine Learning, en donde se de un resultado definitivo acerca del padecimiento de Cáncer de mama. Este diagnostico se realiza con el propósito de dar soporte a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología.
- 6) **BreastApp:** Representa el liberable, consecuencia del paquete de trabajo definido anteriormente. Este producto corresponde a la generación de reportes diagnósticos según el análisis de datos oncológicos elaborados por los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología. Es el resultado obtenido que cumple con los objetivos planteados y satisface las necesidades de la organización. Realiza a la Meseta inicial *BreastApp*
- 7) **Generación especializada de Data-Sets oncológicos:** Hace referencia a la brecha que debe resolverse para poder generar la Meseta futura *OncoAnalysisApp*. Esta brecha esta relacionada con la generación dinámica de Data-Sets para cualquier tipo de cáncer.
- 8) **OncoAnalysisApp:** Es la Meseta final de la aplicación. Representa un producto a futuro con nuevas implementaciones y funcionalidades adicionales. En este caso se pretende generar diagnósticos para cualquier tipo de cáncer para brindar un soporte mas amplio a los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología
- 9) **Web:** Hace referencia a la ubicación en donde los investigadores de la Subdirección de Investigaciones, otras Subdirecciones y demás unidades funcionales del Instituto de Cancerología pueden acceder a la aplicación *BreastApp*.

Parte III

REFLEXIONES

Capítulo 11

Resultados y Conclusiones

11.1. Resultados

- 1) Según la investigación realizada donde se consultaron 40 artículos que hacen énfasis en modelos algorítmicos para el diagnóstico de cáncer de mama, se plantearon diferentes categorías con base a modelos de Machine Learning, Deep Learning y Algoritmos Genéticos. Una vez teniendo los modelos por categorías se asignó un peso a cada uno dependiendo del uso de los mismos por parte de los investigadores, estos pesos pueden ser observados en la Gráfica 11.1.

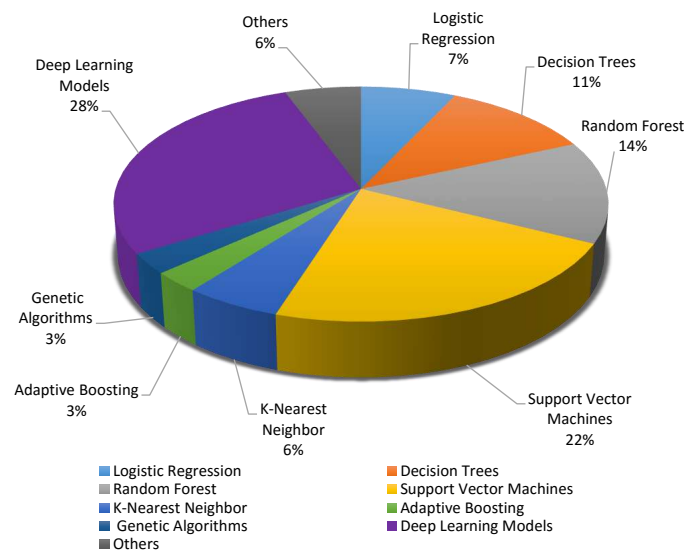


Figura 11.1: Uso de modelos enfocados en el diagnostico de cáncer de mama

Con base en el análisis realizado se determinó que los modelos de Deep Learning son los más usados con un porcentaje del 28 %. Pero como la investigación realizada se basa en modelos de Machine Learning se seleccionaron los que tiene el porcentaje de uso más alto después de los modelos Deep Learning, estos modelos son los siguientes: Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM) y K-Neares Neighbor(KNN), siendo el SVM el modelo de Machine Learning más utilizado con un porcentaje del 22 %. Adicionalmente, se seleccionó el modelo Gaussian Naive Bayes para complementar la investigación. Para realizar el entrenamiento de los modelos de Machine Learning seleccionados, se utilizó el data set de cáncer de mama, elaborado por la Universidad de Wisconsin, el cual cuenta con 569 registros cada uno de ellos con 32 características y se utilizó la librería Scikit-Learn para el procesamiento de datos y el posterior entrenamiento de dichos modelos.

- 2) La determinación de la precisión de los modelos es expresada como la relación entre los diagnósticos realizados correctamente y el número total de diagnósticos, expresado de otra manera, con qué frecuencia la clasificación de las personas con un diagnostico maligno o benigno se realiza de forma correcta. Para la verificación del aprendizaje se utilizó el 25 % de los datos, esto corresponde a 144 de los 569 registros del Data-Set de la Universidad de Wisconsin. En la tabla 11.1 se puede observar la comparación de la precisión de los Modelos en orden Descendente.

Modelo	Presición
Decision Trees	1.0
Random Forest	0.9953
Support Vector Machines(SVM)	0.9647
Logistic Regression	0.9600
Gaussian Naive Bayes	0.9507
K-Nearest Neighbor (KNN)	0.9413

Tabla 11.1: Precisión de los Modelos de Machine Learning seleccionados

Se puede evidenciar que el modelo Decision Trees realizo la clasificación de manera correcta un 100 % del total de los datos destinados para la verificación teniendo una precisión superior a los demás modelos seleccionados. Por otra parte se puede evidenciar que el modelo K-Nearest Neighbor (KNN) realizo la clasificación de manera correcta un 94,13 % del total de los datos destinados para la verificación teniendo las precisión más baja con relación a los demás modelos seleccionados. Esto significa que el modelo Decision Trees es el más indicado para realizar diagnósticos de cáncer de mama con datos obtenidos por el método de Aspiración con aguja Fina (FNA), por consiguiente el modelo KNN es el menos indicado para realizar diagnósticos de cáncer de mama con dichos datos.

- 3) Desarrollo de una aplicación web para el diagnóstico del cáncer de mama donde se utilizaron modelos de Machine Learning seleccionados a partir de la exploración y comparación de los modelos más utilizados por diferentes investigadores y la precisión de dichos modelos en el diagnóstico de este tipo de Cáncer.
- 4) Implementación de la aplicación web llamada BreastApp V1.0 la cual está conformada por dos componentes: El back-End de la aplicación el cual fue realizado en Python y el Front-End de la aplicación el cual fue realizado en Angular. La función principal de la aplicación es generar informes diagnósticos con base en la información de los pacientes. Este diagnóstico brinda un dictamen final del padecimiento cáncer de mama soportado en datos y graficas proporcionados por los modelos de Machine Learning utilizados.

11.2. Conclusiones

- 1) Con respecto a los modelos de Machine Learning utilizados por diversos investigadores para predecir el Cáncer de mama se puede evidenciar que el algoritmo más utilizado es el Support Vector Machines (SVM), pero al realizar la comparación con los resultados obtenidos de la precisión de los métodos de Machine Learning observados anteriormente, el algoritmo Decision Trees es el que mejor resultado tiene, con una exactitud del 100 %, esto quiere decir que clasifico correctamente el total de las muestras. Por lo tanto según la investigación realizada se concluye que si se va a diagnosticar el Cáncer de mama los modelos más indicados para hacerlo son el Support Vector Machines(SVM) y Decision Trees.
- 2) Según el análisis realizado con base en el diagrama de calor conformado por la correlación de las variables del Data-Set de la Universidad de Wisconsin se puede evidenciar que las variables *concave_points_worst* y *area_worst* generan información relevante en la realización del diagnóstico de Cáncer de mama debido a que expresan una deformidad mayor de los núcleos celulares encontrados en las masas mamarias extraídas por el método de Aspiración con Aguja Fina(FNA).

11.3. Aportes Originales

- 1) Implementación de una capa de servicios REST basada modelos de Machine Learning para el diagnóstico de Cáncer de mama que podría ser utilizada en diferentes ámbitos en la detección y el diagnóstico de dicho Cáncer.
- 2) Diseño y Arquitectura de un aplicativo web enfocado en el uso de modelos de Machine Learning aplicados en la rama de la Medicina especializada en oncología.

11.4. Trabajos Futuros

- 1) Creación de una aplicación web llamada OncoAnalysisApp la cual permita el diagnóstico de cualquier tipo de Cáncer teniendo como entrada Data-Sets obtenidos por diversos métodos médicos.
- 2) Creación de una aplicación que permita el análisis de imágenes y que diagnostique el padecimiento de Cáncer de mama con base a los modelos de Deep-Learning existentes.
- 3) Creación de una aplicación que permita crear nuevos Data-Set dinámicamente según parámetros proporcionados por el usuario.

Bibliografía

- [1] International Agency for Research on Cancer. Colombia: Globocan 2018. 380:1–2, 2018.
- [2] Alberto Palacios Pawlovsky and Mai Nagahashi. A method to select a good setting for the kNN algorithm when using it for breast cancer prognosis. *2014 IEEE-EMBS International Conference on Biomedical and Health Informatics, BHI 2014*, pages 189–192, 2014.
- [3] Madhuri Gupta and Bharat Gupta. An Ensemble Model for Breast Cancer Prediction Using Sequential Least Squares Programming Method (SLSQP). *2018 11th International Conference on Contemporary Computing, IC3 2018*, pages 1–3, 2018.
- [4] Pahulpreet Singh Kohli and Shriya Arora. Application of Machine Learning in Disease Prediction. *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, pages 1–4, 2019.
- [5] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, Hugo J W L Aerts, and Hugo_Aerts@dfci Harvard Edu. Artificial intelligence in radiology. *Nat Rev Cancer*, 18(8):500–510, 2018.
- [6] Doreswamy and M. Umme Salma. BAT-ELM: A bio inspired model for prediction of breast cancer data. *Proceedings of the 2015 International Conference on Applied and Theoretical Computing and Communication Technology, iCATccT 2015*, pages 501–506, 2016.
- [7] Moh'D Rasoul Al-Hadidi, Abdulsalam Alarabeyyat, and Mohannad Alhanahnah. Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm. *Proceedings - 2016 9th International Conference on Developments in eSystems Engineering, DeSE 2016*, pages 35–39, 2017.
- [8] Yawen Xiao, Jun Wu, Zongli Lin, and Xiaodong Zhao. Breast Cancer Diagnosis Using an Unsupervised Feature Extraction Algorithm Based on Deep Learning. *Chinese Control Conference, CCC*, 2018-July:9428–9433, 2018.

- [9] R. R. Janghel, Anupam Shukla, Ritu Tiwari, and Rahul Kala. Breast cancer diagnostic system using Symbiotic Adaptive Neuro-Evolution (SANE). *Proceedings of the 2010 International Conference of Soft Computing and Pattern Recognition, SoCPaR 2010*, pages 326–329, 2010.
- [10] M. S B M Azmi and Zaihisma Che Cob. Breast cancer prediction based on back-propagation algorithm. *Proceeding, 2010 IEEE Student Conference on Research and Development - Engineering: Innovation and Beyond, SCORED 2010*, (SCORED):164–168, 2010.
- [11] Quang H. Nguyen, Trang T. T. Do, Yijing Wang, Sin Swee Heng, Kelly Chen, Wei Hao Max Ang, Conceicao Edwin Philip, Misha Singh, Hung N. Pham, Binh P. Nguyen, and Matthew C. H. Chua. Breast Cancer Prediction using Feature Selection and Ensemble Voting. *2019 International Conference on System Science and Engineering (ICSSE)*, pages 250–254, 2019.
- [12] Pragya Chauhan and Amit Swami. Breast Cancer Prediction Using Genetic Algorithm Based Ensemble Approach. *2018 9th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2018*, pages 1–8, 2018.
- [13] Bo Liu, Xingrui Li, Jianqiang Li, Yong Li, Jianlei Lang, Rentao Gu, and Fei Wang. Comparison of Machine Learning Classifiers for Breast Cancer Diagnosis Based on Feature Selection. *Proceedings - 2018 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2018*, pages 4399–4404, 2019.
- [14] Alokumar Jha, Ghanshyam Verma, Yasar Khan, Qaiser Mehmood, Dietrich Rebholz-Schuhmann, and Ratnesh Sahay. Deep Convolution Neural Network Model to Predict Relapse in Breast Cancer. *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*, pages 351–358, 2019.
- [15] Daad Abdullah Almuheidib, Hadil Ahmed Shaiba, Najla Ghazi Alharbi, Sara Muhammad Alotaibi, Fatima Moteb Albusayyis, Mashael Abdulalim Alzaid, and Reem Mohammed Almadhi. Ensemble Learning Method for the Prediction of Breast Cancer Recurrence. *1st International Conference on Computer Applications and Information Security, ICCAIS 2018*, pages 1–6, 2018.
- [16] Gopal K. Dhondalay, Dong L. Tong, and Graham R. Ball. Estrogen receptor status prediction for breast cancer using artificial neural network. *Proceedings - International Conference on Machine Learning and Cybernetics*, 2:727–731, 2011.
- [17] Hiram Ponce and Ma De Lourdes Martinez-Villasenor. Interpretability of artificial hydrocarbon networks for breast cancer classification. *Proceedings of the International Joint Conference on Neural Networks*, 2017-May:3535–3542, 2017.

- [18] Turki Turki and Zhi Wei. Learning approaches to improve prediction of drug sensitivity in breast cancer patients. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2016-Octob(i):3314–3320, 2016.
- [19] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis, and Dimitrios I. Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13:8–17, 2015.
- [20] Tessy Badriyah, Rimawanti Fauzyah, Iwan Syarif, and Prima Kristalina. Mobile personal health record (mPHR) for Breast Cancer using prediction modeling. *Proceedings of the 2nd International Conference on Informatics and Computing, ICIC 2017*, 2018-Janua:1–4, 2018.
- [21] Miaomiao Liang, Lican Huang, and Waheed Ahmad. Breast cancer intelligent diagnosis based on subtractive clustering adaptive neural fuzzy inference system and information gain. *2017 International Conference on Computer Systems, Electronics and Control, ICCSEC 2017*, (x):152–156, 2018.
- [22] Sara Alghunaim and Heyam H. Al-Baity. On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context. *IEEE Access*, 7:91535–91546, 2019.
- [23] Devender Kaushik, Bakshi Rohit Prasad, Sanjay Kumar Sonbhadra, and Sonali Agarwal. Post-Surgical Survival Forecasting of Breast Cancer Patient: A Novel Approach. *2018 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2018*, pages 37–41, 2018.
- [24] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A. Gutman, Jill S. Barnholtz-Sloan, José E. Velázquez Vega, Daniel J. Brat, and Lee A.D. Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences of the United States of America*, 115(13):E2970–E2979, 2018.
- [25] Bo Fu, Pei Liu, Jie Lin, Ling Deng, Kejia Hu, and Hong Zheng. Predicting Invasive Disease-Free Survival for Early Stage Breast Cancer Patients Using Follow-Up Clinical Data. *IEEE Transactions on Biomedical Engineering*, 66(7):2053–2064, 2019.
- [26] Bojana R.Andjelkovic Cirkovic, Aleksandar M. Cvetkovic, Srdjan M. Ninkovic, and Nenad D. Filipovic. Prediction models for estimation of survival rate and relapse for breast cancer patients. *2015 IEEE 15th International Conference on Bioinformatics and Bioengineering, BIBE 2015*, pages 1–6, 2015.

- [27] Mahmoud Khademi and Nedialko S. Nedialkov. Probabilistic graphical models and deep belief networks for prognosis of breast cancer. *Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015*, pages 727–732, 2016.
- [28] Dongdong Sun, Minghui Wang, Huanqing Feng, and Ao Li. Prognosis prediction of human breast cancer by integrating deep neural network and support vector machine: Supervised feature extraction and classification for breast cancer prognosis prediction. *Proceedings - 2017 10th International Congress on Image and Signal Processing, Bio-Medical Engineering and Informatics, CISP-BMEI 2017*, 2018-Janua(61471331):1–5, 2018.
- [29] Bin Dai, Rung Ching Chen, Shun Zhi Zhu, and Wei Wei Zhang. Using random forest algorithm for breast cancer diagnosis. *Proceedings - 2018 International Symposium on Computer, Consumer and Control, IS3C 2018*, pages 449–452, 2019.
- [30] Dongdong Sun, Minghui Wang, and Ao Li. A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 16(3):841–850, 2019.
- [31] Ahmed Iqbal Pritom, Md Ahadur Rahman Munshi, Shahed Anzarus Sabab, and Shihabuzzaman Shihab. Predicting breast cancer recurrence using effective classification and feature selection technique. *19th International Conference on Computer and Information Technology, ICCIT 2016*, pages 310–314, 2017.
- [32] Pallvi Grover and Raj Mohan Singh. Automated Detection of Breast Cancer Metastases in Whole Slide Images. *ICSCCC 2018 - 1st International Conference on Secure Cyber Computing and Communications*, pages 111–116, 2019.
- [33] Muhammad Shoaib B. Sehgal, Iqbal Gondal, and Laurence Dooley. Stacked regression ensemble for cancer class prediction. *2005 3rd IEEE International Conference on Industrial Informatics, INDIN*, 2005:831–835, 2005.
- [34] Aiza M. Romano and Alexander A. Hernandez. Enhanced Deep Learning Approach for Predicting Invasive Ductal Carcinoma from Histopathology Images. *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 142–148, 2019.
- [35] Abeer A. Raweh, Mohammed Nassef, and Amr Badr. A Hybridized Feature Selection and Extraction Approach for Enhancing Cancer Prediction Based on DNA Methylation. *IEEE Access*, 6:15212–15223, 2018.

- [36] Tiago H. Falk, Hagit Shatkay, and Wai Yip Chan. Breast cancer prognosis via Gaussian mixture regression. *Canadian Conference on Electrical and Computer Engineering*, (May):987–990, 2007.
- [37] U. Karthik Kumar, M. B.Sai Nikhil, and K. Sumangali. Prediction of breast cancer using voting classifier technique. *2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2017 - Proceedings*, (August):108–114, 2017.
- [38] Pedro Ferreira, Nuno A. Fonseca, Inês Dutra, Ryan Woods, and Elizabeth Burnside. Predicting malignancy from mammography findings and surgical biopsies. *Proceedings - 2011 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2011*, pages 339–344, 2011.
- [39] Mihir Sewak, Priyanka Vaidya, Chien-chung Chan, and Zhong-hui Duan. SVM Approach to Breast Cancer Classification. *IEEE Computer society*, pages 32–37, 2007.
- [40] Aaron N. Richter and Taghi M. Khoshgoftaar. Predicting cancer relapse with clinical data: A survey of current techniques. *Proceedings - 2016 IEEE 17th International Conference on Information Reuse and Integration, IRI 2016*, pages 369–376, 2016.
- [41] Moloud Abdar and Vladimir Makarenkov. CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. *Measurement: Journal of the International Measurement Confederation*, 146:557–570, 2019.
- [42] Raul E. Briega Lopez. Machine Learning con Python, 2015.
- [43] Olvi L. Mangasarian and William H. Wolberg. Machine Learning for Cancer Diagnosis and Prognosis. *University of Wisconsin-Madison*, pages 1–4, 1990.
- [44] Street W.N., Wolberg W.H., and Mangasarian O.L. Nuclear feature extraction for breast tumor diagnosis. *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, 1905(870):861–870, 1993.
- [45] Jorge Leonel. Supervised Learning, 2018.
- [46] Scikit-learn Developers. Scikit-Learn, 2019.
- [47] Sandro Javier Bolaños Castro, Maddyzeth Ariza Riaño, and Heiner Santiago Alfonso Casallas. *Enfoque de Ingeniería de Software desde el proceso , la arquitectura y la implementación*. Universidad Distrital Francisco José de Caldas, Bogota, Colombia, 2019.
- [48] Open Group Standard and The Open Group. *Open Group Standard (ArchiMate 3.0.1)*. 2017.