

Análisis correlacional y predictivo entre datos los de inscripción y el historial académico de los estudiantes de Ingeniería de Sistemas de la Universidad de Pamplona.

autor
Zulay Viviana Medina

**Programa de Ingeniería de Sistemas
DEPARTAMENTO ELECTRICA, ELECTRONICA, SISTEMAS Y
TELECOMUNICACIONES
FACULTAT DE INGENIERIAS Y ARQUITECTURA**



**UNIVERSIDAD DE PAMPLONA
PAMPLONA, JUNIO 20 DE 2022**

Análisis correlacional y predictivo entre datos los de inscripción y el historial académico de los estudiantes de Ingeniería de Sistemas de la Universidad de Pamplona.

autor

Zulay Viviana Medina

Director

José Orlando Maldonado Bautista

Doctor en Ingeniería Informática

**Programa de Ingeniería de Sistemas
DEPARTAMENTO ELECTRICA, ELECTRONICA, SISTEMAS Y
TELECOMUNICACIONES
FACULTAT DE INGENIERIAS Y ARQUITECTURA**



**UNIVERSIDAD DE PAMPLONA
PAMPLONA, JUNIO 20 DE 2022**

Agradecimientos

En primer lugar, Doy Gracias a Dios, por permitir culminar esta etapa de estudio.

A mi director de tesis José Orlando Maldonado Bautista por su paciencia, colaboración y motivación para guiarme durante todo el desarrollo de este trabajo.

A la Universidad de Pamplona por su aporte a mi formación como persona y como profesional.

Finalmente, gracias a mi familia y amigos por estar a mi lado apoyándome en los momentos más difíciles.

Dedico esta tesis a mis padres

Jose R Medina y Bertha Villamizar,

A mis hermanos que guía

mi camino día a día.

Contenido

1.	Introducción	11
1.1.	Descripción del problema y justificación	12
1.2.	Objetivos	13
1.2.1.	Objetivo General	13
1.2.2.	Objetivos Específicos	13
2.	Marco Teórico	14
2.1.	Minería de datos	14
2.2.	Modelos predictivos	14
2.2.1.	Clasificación	15
2.2.1.1.	Algoritmo del vecino más cercano	15
2.2.1.2.	Algoritmo naïves bayes	15
2.2.1.3.	Arboles de Regresión	16
2.2.1.4.	Máquina de vectores de soportes	16
2.2.1.5.	Análisis discriminante lineal	16
2.2.1.6.	Redes neuronales	17
2.3.	Regresión	17
2.3.1.	Regresión lineal	18
2.3.2.	Regresión polinómica o Regresión Múltiple	18
2.3.3.	Regresión no lineal	18
2.4.	Análisis correlacional	18
2.5.	Modelo KDD	21
3.	Análisis de datos	22
3.1.	Descripción de datos y fuentes	22
3.2.	Preparación de datos	22
3.3.	Análisis descriptivo	23
3.4.	Análisis predictivo	71
3.5.	Síntesis de resultados	91
4.	Conclusiones y trabajos futuros	95
5.	Referencia	97
6.	Anexo	98

GRAFICAS

Ilustración 1 Estructura de red neuronal simple y profunda	17
Ilustración 2 Pearson $r = 0$	19
Ilustración 3 Pearson $r = 0.476$	19
Ilustración 4 Pearson $r = 0.93$	20
Ilustración 5 Pearson $r = -0.968$	20
Ilustración 6 Fases del método KDD	21
Ilustración 7 Resultado de Ficheros en estudio	24
Ilustración 8 grafica de notas obtenidas en Matemáticas en las pruebas de estado icfes	25
Ilustración 9 Grafica de notas obtenidas en la materia lenguaje en las pruebas de estado icfes.....	25
Ilustración 10 grafica de notas obtenidas en Filosofía en las pruebas de estado icfes	26
Ilustración 11 Grafica para notas en Biología en las Pruebas icfes	26
Ilustración 12 Notas de Química en las Icfes	27
Ilustración 13 Grafica puntajes de Física en las icfes	27
Ilustración 14 Grafica Resultados ICFES	28
Ilustración 15 Resultados ICFES Matemáticas, Lenguaje y Física.....	29
Ilustración 16 Distribución de calificaciones de Calculo diferencia	32
Ilustración 17 Distribución de calificaciones de Biología general	32
Ilustración 18 Distribución de calificaciones Catedra Faria	33
Ilustración 19 Distribución de calificaciones Educacion Ambiental	33
Ilustración 20 Distribución de calificaciones Habilidades comunicativas	34
Ilustración 21 Distribución de calificaciones Laboratorio biología general	34
Ilustración 22 Distribución de calificaciones Programación estructurada	35
Ilustración 23 Distribución de calificaciones Calculo integral	35
Ilustración 24 Distribución de calificaciones laboratorio de mecánica	36
Ilustración 25 Distribución de calificaciones laboratorio de química	36
Ilustración 26 Distribución de calificaciones Mecánica	37
Ilustración 27 Distribución de calificaciones POO	37
Ilustración 28 Distribución de calificaciones Química General	38
Ilustración 29 Distribución de calificaciones Algebra Lineal	38
Ilustración 30 Distribución de calificaciones Calculo Multivariable.....	39
Ilustración 31 Distribución de calificaciones Electromagnetismo	39
Ilustración 32 Distribución de calificaciones Estructuras de datos dinámicos y algoritmos	40
Ilustración 33 Distribución de calificaciones Laboratorio de Electromagnetismo	40
Ilustración 34 Distribución de calificaciones Análisis de Algoritmos	41
Ilustración 35 Distribución de calificaciones Ecuaciones diferenciales	41
Ilustración 36 Distribución de calificaciones Expresión grafica	42
Ilustración 37 Distribución de calificaciones introducción a las ciencias computacionales	42
Ilustración 38 Distribución de calificaciones Lab de oscilaciones y ondas	43
Ilustración 39 Distribución de calificaciones oscilaciones y ondas	43
Ilustración 40 Distribución de calificaciones fundamentos y diseños de BD	44
Ilustración 41 Distribución de calificaciones Fundamentos y diseños SO.....	44
Ilustración 42 Distribución de calificaciones Lógica computacional	45
Ilustración 43 Distribución de calificaciones Matemáticas especiales	45

Ilustración 44 Distribución de calificaciones Métodos numéricos.....	46
Ilustración 45 Distribución de calificaciones paradigmas	46
Ilustración 46 Distribución de calificaciones Bases de Datos II	47
Ilustración 47 Distribución de calificaciones electiva sociohumanística I.....	47
Ilustración 48 Distribución de calificaciones Estadística II	48
Ilustración 49 Distribución de calificaciones Modelamiento de sistemas de información	48
Ilustración 50 Distribución de calificaciones Sistemas y modelado I	49
Ilustración 51 Distribución de calificaciones Teoría de la computación	49
Ilustración 52 Distribución de calificaciones electiva de ingeniería.....	50
Ilustración 53 Distribución de calificaciones Electiva Profesional	50
Ilustración 54 Distribución de calificaciones redes de computadores y sistemas distribuidos.....	51
Ilustración 55 Distribución de calificaciones Sistemas y modelados	52
Ilustración 56 Distribución de calificaciones tópicos avanzados de ingeniería del software	52
Ilustración 57 Distribución de calificaciones Electiva de ingeniería.....	53
Ilustración 58 Distribución de calificaciones electiva profesional II	53
Ilustración 59 Distribución de calificaciones ingeniería de proyectos I	54
Ilustración 60 Distribución de calificaciones Investigación de operaciones	54
Ilustración 61 Distribución de calificaciones Legislación informática	55
Ilustración 62 Distribución de calificaciones redes y sistemas distribuidos.....	55
Ilustración 63 Distribución de calificaciones Electiva profesional III	56
Ilustración 64 Distribución de calificaciones Electiva Socio humanística II.....	56
Ilustración 65 Distribución de calificaciones Ética	57
Ilustración 66 Distribución de calificaciones Metodología de la investigación.....	57
Ilustración 67 Protección de sistemas informativos	58
Ilustración 68 Distribución de calificaciones recreación y deportes	58
Ilustración 69 Distribución de calificaciones Cívica	59
Ilustración 70 Distribución de calificaciones Informática Básica	59
Ilustración 71 Distribución de calificaciones Trabajo Grado	60
Ilustración 72 Formación básica	61
Ilustración 73 Materias Sociohumanísticas	62
Ilustración 74 Formación Profesional.....	64
Ilustración 75 Formación de Profundización	65
Ilustración 76 Materias primer semestre y Materias Segundo Semestre	66
Ilustración 78 Materias Tercer Semestre y Materias Cuarto Semestre	67
Ilustración 80 Materias Quinto Semestre y Materias Sexto Semestre	68
Ilustración 82 Materias Séptimo Semestre y Materias Octavo Semestre.....	69
Ilustración 84 Materias Noveno Semestre y Materias Decimo Semestre	70
Ilustración 86 Notas obtenidas en al usar las maquinas de regresión. Para las siguientes materias fundamentos y diseño de base de datos, metodología de la investigación, electiva profesional II, redes de computadores y sistemas distribuidos II, modelamiento de sistemas de información, estadística II, electiva de ingeniería i, electiva profesional III, tópicos avanzados de ingeniería del software, matemáticas especiales	79
Ilustración 77. Comportamiento de los resultados en pruebas ICFES en matemáticas, física y lenguaje.....	91

Ilustración 78. Comportamiento de los resultados ICFES en todas las áreas	92
---	----

TABLAS

Tabla 1 Coeficientes de correlación	19
Tabla 2 Resultados de ICFES	23
Tabla 3 Análisis estadístico notas ICFES	23
Tabla 4 Materias de pregrado	30
Tabla 5 Numero de vistas por materia	31
Tabla 6 Listado de pensum2006.....	31
Tabla 7 Resultado ICFES vs Programación estructurada.....	72
Tabla 8 Resultados de Programación estructurada Vs Métodos de Predicción	73
Tabla 9 Resultados Calculo diferencial Vs Métodos de Predicción	74
Tabla 10 Resultado Biología general Vs Métodos de Predicción	74
Tabla 11 Resultado Catedra Faria Vs Métodos de Predicción	75
Tabla 12 Resultado Educación ambiental Vs Métodos de Predicción	75
Tabla 13 Resultado Habilidades comunicativas Vs Métodos de Predicción	75
Tabla 14 Resultados maquinas aprendizaje Error ± 0.5 puntos.....	76
Tabla 15 Notas obtenidas según resultado de error ± 0.5	77
Tabla 16 Discretización de notas en 5 rangos	80
Tabla 17 Clasificadores aplicados a Programación estructurada.....	81
Tabla 18 Clasificadores aplicados a Calculo diferencial	82
Tabla 19 Clasificadores aplicados a Biología general.....	82
Tabla 20 Clasificadores aplicados a Catedra Faria	83
Tabla 21 Clasificadores aplicados a Educación ambiental	83
Tabla 22 Clasificadores aplicados a Habilidades Comunicativas.....	84
Tabla 23 Resultados con exactitud sobre 70%	84
Tabla 24 Matriz de confusión aplicada a Matematicas especiales	87
Tabla 25 Matriz de confusión aplicada a Modelamiento de Sistemas de Información	88
Tabla 26 Matriz de confusión aplicada a Tópicos avanzados	88
Tabla 27 Matriz de confusión aplicada a Investigación de Operaciones	89
Tabla 28 Matriz de confusión aplicada a Electiva Profesional III	90
Tabla 29 Matriz de confusión aplicada a Actividad Recreativa e Informática Básica.....	90

1. Introducción

Mediante este trabajo se espera determinar si existe algún grado o tipo de correlación entre los datos académicos y socioeconómicos proporcionados por los aspirantes al programa de Ingeniería de Sistemas a la hora de la inscripción al programa, y los resultados académicos obtenidos durante su trayectoria en el pregrado.

En especial se espera realizar énfasis para encontrar una posible correlación entre los datos reportados por los estudiantes sobre sus resultados en las pruebas de estado SABER 11, y las calificaciones obtenidas durante los primeros cinco semestres del programa, considerados como una etapa sensible respecto a condiciones de mortalidad y deserción.

En función del grado de correlación encontrada, se espera determinar la viabilidad de predicción en cuanto a la mortalidad y/o deserción escolar universitaria, y en consecuencia determinar el error de predicción si es el caso.

1.1. Descripción del problema y justificación

La deserción y la mortalidad estudiantil en educación superior es un asunto que preocupa tanto a los organismos gubernamentales, como el mismo Ministerio de Educación, hasta los programas académicos, pasando por la dirección de las instituciones de educación superior. Abandonar la universidad y repetir cursos escolares tiene gran repercusión económica, social y académica en la comunidad universitaria. Según un informe del banco mundial citado en (Espectador, 2018), Colombia es el segundo país en América Latina con mayor tasa de deserción en educación superior y se estima que el 42 % de los que ingresan a planteles educativos termina desertando en los primeros años.

El Ministerio de Educación Nacional en su intento por entender dicho fenómeno y tomar medidas que puedan reducir las tasas de deserción y mortalidad implementó el sistema SPADIES. Este sistema es la herramienta para hacer seguimiento sobre las cifras de deserción de estudiantes de la educación superior.

Con los datos suministrados por las instituciones de educación superior a SPADIES, se identifican y se ponderan los comportamientos, las causas, variables y riesgos determinantes para desertar. Sin embargo, los reportes obtenidos por el sistema parecen bastante genéricos. Para entender el comportamiento académico de un programa en particular, se requiere un análisis con los datos propios del mismo. Es por ello que, utilizando la información recopilada desde diferentes fuentes del programa, como la proporcionada por la oficina de Registro y Control Académico, la oficina de Planeación, y el Programa de Ingeniería de Sistemas, se espera obtener una visión clara del comportamiento académico de la población, sobre todo en la primera mitad de la carrera, en donde las tasas de mortalidad y deserción son más altas, que permitan reducir los altos índices encontrados.

1.2. Objetivos

1.2.1. Objetivo General

Analizar los datos de inscripción obtenidos por los aspirantes al programa de Ingeniería de Sistemas, y el historial académico obtenido durante la primera mitad del programa para determinar si existe correlación entre los mismos con posibilidades de predicción.

1.2.2. Objetivos Específicos

- Realizar la recopilación e integración de los datos disponibles desde las diversas fuentes requeridas.
- Aplicar herramientas para el análisis de datos que permitan seleccionar, limpiar y transformar los mismos como paso previo al análisis correlacional y predictivo.
- Aplicar modelos correlacionales y predictivo utilizando los enfoques obtenidos mediante exploración bibliográfica.
- Difundir los resultados a la comunidad académica mediante la presentación del informe técnico.

2. Marco Teórico

2.1. Minería de datos

La minería de datos puede definirse como el conjunto de pasos usados para el hallazgo de nuevas relaciones, patrones o tendencias relacionadas en grandes cantidades de datos.

Data Mining o minería de datos es la técnica especializada la cual surgió como resultado del estudio de grandes volúmenes de información y el uso de herramientas informáticas para dichos análisis.

El descubrimiento del conocimiento logra evitar el análisis de datos dirigido a la verificación, pero abre caminos a la minería de datos, la minería de datos ayuda a facilitar la detección de patrones en los datos. Los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento (Weiss, 1998)

Para el aprendizaje supervisado, el programador tiene las respuestas esperadas al problema, ejecuta el algoritmo el cual identifica, aprende y hace predicciones a medida que analiza los patrones, al finalizar el programador genera las correcciones si es necesario, tantas veces hasta que se alcanza un alto nivel de precisión en el algoritmo.

Cuando el algoritmo de aprendizaje realiza automáticamente la identificación de patrones sin tener un programador que cuente con respuestas esperadas o deba proporcionar instrucciones se llama aprendizaje no supervisado.

2.2. Modelos predictivos

Los datos discretos, que a veces se llaman datos temáticos, categóricos o discontinuos, la mayoría de las veces representan objetos en los sistemas de almacenamiento de datos ráster y de entidades tipo vector. Un objeto discreto tiene límites conocidos y definibles: es fácil definir con precisión dónde comienza el objeto y dónde termina. Los objetos discretos por lo general son sustantivos.

Los datos continuos se les suele conocer también como datos de campo, no discretos o de superficie. Un tipo de superficie continua deriva de aquellas características que definen una superficie, en la que cada ubicación se mide desde un punto de registro fijo.

El Machine Learning es una disciplina científica que trata de que los sistemas aprendan automáticamente. Aprender, en este contexto, quiere decir identificar patrones complejos en

millones de datos. La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros.

El Machine Learning es, por tanto, un proceso de inducción del conocimiento, es decir, un método que permite obtener por generalización un enunciado general a partir de enunciados que describen casos particulares. Machine Learning o Aprendizaje Automático es aprender de los datos, es descubrir la estructura y los patrones que subyacen en ellos.

Por lo cual un modelo predictivo es un conjunto de procesos ejecutados utilizando técnica de análisis de datos que ayudan a obtener la probabilidad en la cual sucedan las cosas, o se obtengan ciertos resultados.

2.2.1. Clasificación

Los modelos predictivos son usados para optimizar recursos a la hora de toma de decisiones para reducir el riesgo, mejorar la respuesta a los requerimientos de un usuario teniendo en cuenta su comportamiento, reducción de costos e incremento de beneficios, para todo este se tienen en cuenta las siguientes técnicas o métodos para el análisis predictivo.

2.2.1.1. Algoritmo del vecino más cercano

Método de clasificación en base al parecido con otros casos, el cual reconoce patrones sin necesidad de una coincidencia exacta o casos almacenados, la palabra vecino es asignada a los casos próximos entre ellos según la distancia definida, se define un número k de vecinos más próximos a examinarse según el estudio a realizar, este algoritmo se lo debemos a Rosenkrantz, Stearns y Lewis (1977). (Trujillo Cabezas, 2022)

El objetivo y las características a encontrar son nominales, ordinales o de escala, nominal se da cuando una variable se encuentra sin clasificación intrínseca por ejemplo la religión o afiliación a seguridad social, una característica ordinal se da cuando una variable tiene alguna clasificación intrínseca ejemplo nivel de satisfacción bajo, medio o alta, cuando se presenta una categoría ordenada bajo alguna función o distancias aproximadas entre dichos valores se denomina escala la cual puede ser la edad o ingreso salarial.

2.2.1.2. Algoritmo naïves bayes

Es un clasificador probabilístico basado en el teorema de Bayes donde se evalúan las características de forma independiente, la ausencia de una no se asocia a la ausencia de otra característica, y requiere

de pocos datos es fácil y rápido para su entrenamiento lo cual se toma como una ventaja, se puede usar en clasificaciones binarias como multiclase. Al tomar las características en el conjunto de datos como independientes, aunque esta consideración no aplica a todos los casos, la convierte en una desventaja. (Pertuz, 2021, pág. 202)

Este algoritmo es basado en la probabilidad condicional la cual consiste en determinar la ocurrencia entre un evento A vs un evento B, expresada de la siguiente manera:

$$P(A/B) = P(A \cap B) / P(B)$$

Se lee la probabilidad de A dado B es igual a la unión de la probabilidad de que suceda el evento A y el evento B Sobre la probabilidad de que suceda el evento B. (Pertuz, 2021)

Es utilizado para clasificación de texto que implica datos de entrenamiento de alta complejidad.

2.2.1.3. Árboles de Regresión

Los árboles de decisión se usan frecuentemente por los Data Scientists para la clasificación propuesta por Leo Breiman, usados para predecir la variable Y cualitativa, en función de covariables, basado en un conjunto de reglas sucesivas que ayudan a tomar decisiones.

Las ventajas de los árboles de regresión son: fácil de entender e interpretar, los datos requieren poca preparación, las covariables pueden ser cualitativas o cuantitativas, donde los nodos representan un evento o una elección, las reglas son representados por los bordes a si mismo las condiciones de decisión.

2.2.1.4. Máquina de vectores de soportes

Las SVM máquinas de vectores de soporte en aprendizaje automático busca los puntos más cercanos y este resultado se le da el nombre de vector de soporte. Una vez conocido este vector y los valores más cercanos traza una línea desde el valor menor al mayor haciendo una recta vectorial.

2.2.1.5. Análisis discriminante lineal

Es una técnica estadística la cual cumple la función de clasificar los fenómenos basándose en una función con variables discriminativas y una probabilidad de pertenencia. La cual busca agrupar las por sus semejanzas, cuantificando la probabilidad de pertenecer a un grupo u otro.

2.2.1.6. Redes neuronales

Simulando el comportamiento del cerebro, son neuronas o equipos de cómputo interconectadas en diversas capas que transmiten mensajes de la una a otra, dichas unidades son ordenadas en capas y son usadas cuando se desconoce la naturaleza de la relación entre los valores de entrada y salida; estas redes pueden resolver tareas complejas como reconocimiento de imágenes, aunque son lentas y necesitan mucha potencia.

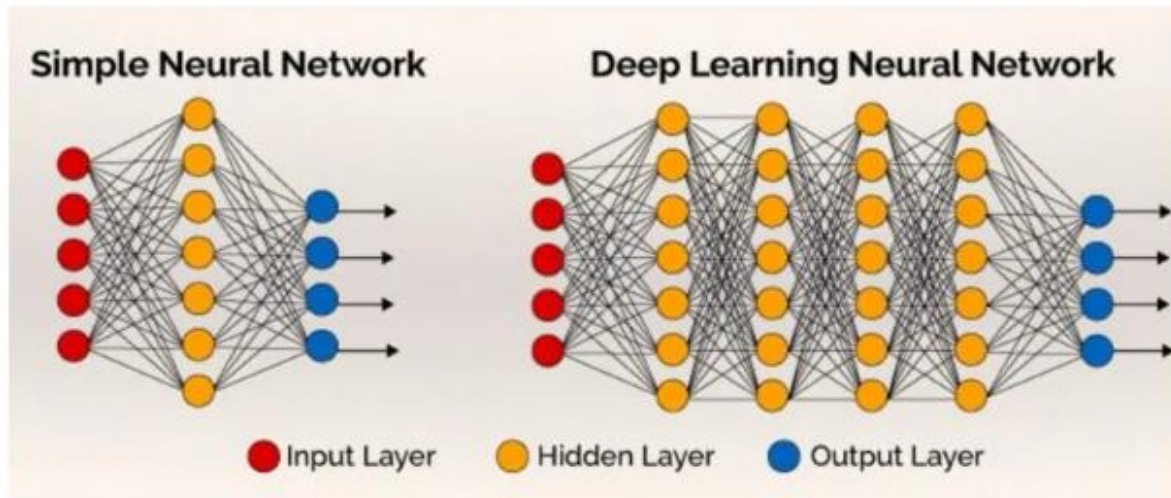


Ilustración 1 Estructura de red neuronal simple y profunda

Con el fin de entrenar redes neuronales para modelar los datos proporcionados se busca crear un modelo basado en el establecer reglas en la etapa de entrenamiento, seguido evaluar el modelo con el conjunto de datos seleccionados para la verificación, ingrese datos nuevos y tome decisiones basadas en los resultados obtenidos. Los cuales pueden ser continuos como el valor de una compra o pueden ser binarios donde la respuesta puede ser 1 o 0, donde se determina un si o un no, determinado por dos categorías. (Heras, 2022)

2.3. Regresión

En la minería de datos encontramos métodos de regresión, en el cual el programa de aprendizaje automático debe evaluar y comprender las relaciones entre las variables. El análisis de regresión se enfoca en una variable dependiente y una serie de otras variables independientes, lo cual lo hace efectivo en diversos estudios. (Hernández, 2021)

La relación es una técnica paramétrica que existe entre una variable dependiente y un conjunto de variables independientes antes de analizar los datos se debe conocer el número de parámetros o coeficientes que se van necesitar esto recibe el nombre de modelo de regresión.

2.3.1. Regresión lineal

Regresión lineal o método de los mínimos cuadrados dado que calcula la suma de los cuadrados entre los puntos que representan los datos y los puntos de la línea generada por el modelo, de esa forma la mejor estimación será la que minimice estas distancias.

Las regresiones logísticas se usan para predecir el resultado de una variable categórica en función de las variables predictivas o independientes. Es práctico para modelar y obtener la probabilidad de un evento dado por función de otros factores es decir en muchos casos se determina si la relación entre las variables estipulada por una ecuación lineal. (Alberca, 2010)

2.3.2. Regresión polinómica o Regresión Múltiple

Como su nombre lo indica se encarga de analizar situaciones que están sujetas a varias variables, este método busca definir cuáles son las variables independientes, define las causas y predice los valores de forma aproximada. (ESTAMÁTICA)

2.3.3. Regresión no lineal

La regresión no lineal es más compleja dado que se presenta que el número de parámetros a analizar no coincida con el número de variables explicativas, este modelo puede transformar las variables originales que vienen de una función no lineal en una lineal, con los coeficientes pueden definir nuevas variables. (questionpro)

2.4. Análisis correlacional

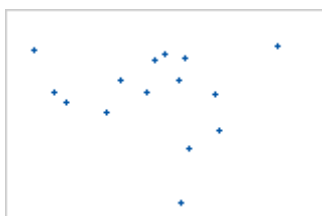
El concepto de **CORRELACION** es una técnica de análisis de información con base estadística que se refiere al grado de variación conjunta existente entre dos o más variables. Cuando se realiza un análisis de información sobre un conjunto de datos sin quitar importancia a su origen ya sea una base de datos o una hoja de cálculo las correlaciones son el medio más efectivo para obtener una conclusión dado que muestra el sentido y la fuerza de dicha relación. Para analizar la relación entre variables se utilizan los llamados «coeficientes de correlación». Se realizan sobre variables cuantitativas o cualitativas. Ello determinará si se calcula o bien el coeficiente de correlación de Pearson, el de Spearman, o el de Kendall. Esto si estamos hablando de correlaciones bivariadas. Existen otras como pueden ser las correlaciones o las medidas de distancia o disimilitud de intervalos, recuentos o binarias.

Para interpretar el coeficiente de correlación utilizamos la siguiente escala:

Valor	Significado
-1	Correlación Negativa Grande Y Perfecta
-0,9 a -0,99	Correlación Negativa Muy Alta
-0,7 a -0,89	Correlación Negativa Alta
-0,4 a -0,69	Correlación Negativa Moderada
-0,2 a -0,39	Correlación Negativa Baja
-0,01 a -0,19	Correlación Negativa Muy Baja
0	Correlación Nula
0,01 a 0,19	Correlación Positiva Muy Baja
0,2 a 0,39	Correlación Positiva Baja
0,4 a 0,69	Correlación Positiva Moderada
0,7 a 0,89	Correlación Positiva Alta
0,9 a 0,99	Correlación Positiva Muy Alta
1	Correlación Positiva Grande Y Perfecta

Tabla 1 Coeficientes de correlación

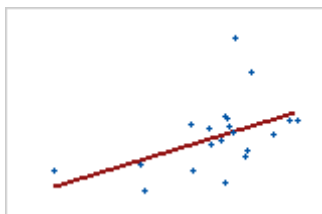
Las siguientes gráficas muestran datos con valores específicos del coeficiente de correlación para ilustrar diferentes patrones en la fuerza y la dirección de las relaciones entre dos grupos de variables.



Ninguna relación o correlación nula: Pearson $r = 0$

Los puntos se ubican sin ningún patrón en común en la gráfica, por lo tanto, no existe relación lineal entre las variables.

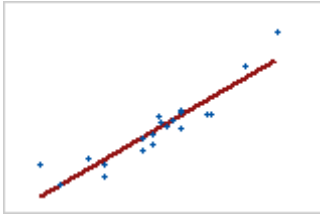
Ilustración 2 Pearson $r = 0$



Relación Positiva Moderada: Pearson $r = 0.476$

Al observar que algunos puntos están cerca de la línea, pero otros están lejos de ella, esto indica que existe una relación lineal moderada en estos grupos de variables.

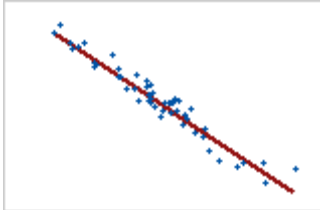
Ilustración 3 Pearson $r = 0.476$



Relación Positiva Muy Alta: Pearson $r = 0.93$

Se observa la mayor cantidad de puntos ubicado cerca de la línea, lo que indica que existe una relación lineal fuerte entre las variables. La relación es positiva dado que aumentan las dos variables de forma simultánea.

Ilustración 4 Pearson $r = 0.93$



Relación Negativa Muy Alta: Pearson $r = -0.968$

Al graficar se ven puntos cerca de la línea, por lo cual se concluye que existe una relación negativa fuerte entre las variables. La relación es negativa porque disminuye las dos funciones de forma simultánea.

Ilustración 5 Pearson $r = -0.968$

2.5. Modelo KDD

Fases del Proceso de Extracción de Conocimiento El proceso de extracción de conocimiento, en inglés KDD (Knowledge Discovery from Databases), es como se denomina al proceso que se encarga de sacar conclusiones o información “limpia” a partir de unos datos que generalmente están organizados en una base de datos. Este proceso es iterativo debido a que la salida de algunas de sus fases puede requerir que se vuelva a una fase anterior y también a que en ocasiones es necesario realizar varias iteraciones para poder extraer conocimiento de los datos.

El método KDD consta cinco grandes fases:

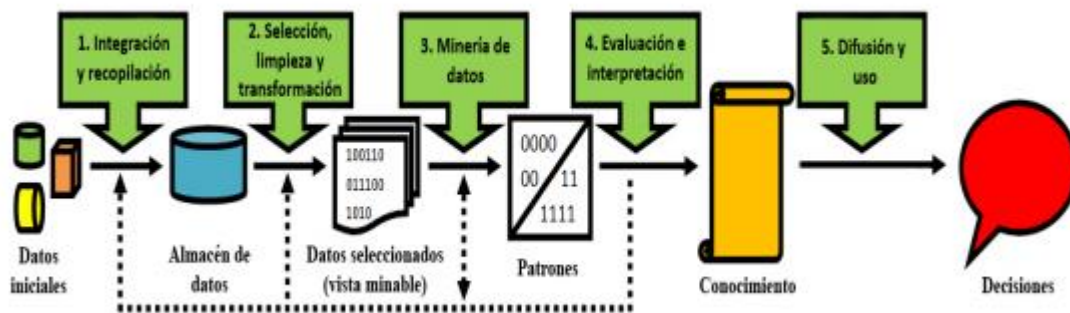


Ilustración 6 Fases del método KDD

1. Fase de integración y recopilación de datos: define la fuente de los datos que son útiles para la investigación.
2. Fase de Selección, limpieza y transformación: se verifican si los datos recolectados son correctos y/o se busca detectar para corregir errores.
3. Fase de minería de datos: en esta etapa se decide qué tarea realizar con estos datos y se elige el método que se va a emplear para ello.
4. Fase de evaluación e interpretación: etapa en la que se obtienen patrones definidos en los datos tratados, se evalúa e interpreta los resultados para lograr entender dichos patrones.
5. Fase de difusión y uso: se tiene como objetivo en esta etapa comunicar y aplicar los resultados obtenidos.

3. Análisis de datos

Este estudio de desarrollo bajo una metodología correlacional para la búsqueda de modelos correlacionales o predictivos y analizar los datos obtenidos mediante graficas.

3.1. Descripción de datos y fuentes

Para la recolección de los datos, teniendo en cuenta que es información clasificada, se solicita a la oficina de admisiones, registro y control de la universidad de Pamplona la información la cual una parte es proporcionada por los aspirantes al programa de Ingeniería de sistemas mediante el diligencia los formularios a la hora de realizar la inscripción a este programa, en dicho formulario se encuentran relacionados los puntajes obtenidos en las pruebas de estado icfes, y el faltante de la información son las notas alcanzadas por los estudiantes del pregrado, dichos registros son proporcionados en formato .xlsx (libros de Excel).

3.2. Preparación de datos

Respecto a los datos recibidos con los puntajes de las pruebas de estado se obtuvieron 24 libros cada uno con una sola hoja de cálculo en total con datos registrados en el periodo comprendido desde 2006-2 hasta 2018, en total de 1804 registro de dichos puntajes.

Cada archivo está conformado con la siguiente estructura Tipo, Metodología, Modalidad, Nivel Educativo, Estado, Admisión, Sede, Primera opción, Documento, Tipo Doc, Nombres, Apellidos, Carga de Documentos, Verificación de Documentos, Jornada Primero, Admitido Primera, Total entregarPO, Total entregadosPO, Segunda opción, Jornada Segunda, Und segunda opc, Admitido segunda, Jornada Segunda, Total entregarSO, Total entregadosSO, Lugar Residencia, Dirección, Teléfono, Email, Fecha Preinscripción, Circunscripción, Convenio, Cod SNP, Lenguaje, Matemáticas, C Sociales, Filosofía, Biología, Química, Física, Idioma, Estrato, Promedio, Programa.

El reporte de las notas por las asignaturas cursadas en el plan de estudios desde el primer semestre de 2006 con fecha de consulta hasta el segundo semestre de 2018, Se cuenta con un libro de Excel con una hora única con 1763 registros. Después de limpiar los datos se obtiene una tabla con el identificador de cada estudiante y 59 materias contempladas en el pensum 2006.

3.3. Análisis descriptivo

Se realiza un análisis de cada materia con la distribución de la nota definitiva de cada curso, en este caso solo se tiene en cuenta la nota final obtenida. Para cada curso se realiza un diagrama de barras que muestra la distribución de la nota definitiva registrada en la fuente de datos, así como una descripción estadística de los datos

Estos archivos se cargan en un DataFrame de pandas y se realiza limpieza a los datos, eliminando datos repetidos e incompletos. En la variable datos almacenamos los datos por el archivo seleccionado ubicado en el directorio /Data/icfes+matricula/icfes2006-2.xlsx para el primer ejemplo, este ejercicio se puede ejecutar con todos los archivos.

Los archivos leídos nos dan una tabla como esta:

	Documento	Tipo Doc	Estrato	Lenguaje	Matemáticas	C Sociales	Filosofía	Biología	Química	Física	Idioma
0	15879924	CC	1	43.00	41.00	Nan	46.00	49.00	43.00	51.00	37
1	41058630	CC	1	59.00	Nan	59.0	Nan	65.00	51.00	44.00	Nan
2	68297165	CC	1	51.00	45.00	Nan	37.00	54.00	34.00	48.00	39
3	89092680526	TI	1	40.14	51.11	Nan	41.62	53.42	33.62	51.53	36.20
4	89052771987	TI	1	44.61	55.73	Nan	44.09	51.63	52.03	58.70	41.36

Tabla 2 Resultados de ICFES

Los datos estadísticos de las materias evaluados en las icfes se presentan en esta tabla donde encontramos la cantidad de registros por materia, la nota max y mínima, adicional tenemos calculados los cuartiles en el cual tenemos en cuenta el cuartil 25, 50 y 75, la desviación estándar y la moda.

	Documento	Estrato	Lenguaje	Matemáticas	C Sociales	Filosofía	Biología	Química	Física
contar	4.200000e+01	42.000000	39.000000	36.000000	5.000000	36.000000	39.000000	39.000000	39.000000
significar	3.424402e+10	1.285714	47.957692	43.112500	53.298000	42.362500	48.115641	44.302308	48.394872
ETS	4.333842e+10	0.553733	7.637576	7.324345	12.692124	6.128273	8.167050	9.720410	7.744459
min	1.587992e+07	1.000000	32.820000	26.330000	33.450000	28.300000	35.000000	30.380000	29.000000
25%	3.195711e+08	1.000000	42.695000	41.000000	48.040000	38.930000	43.245000	38.670000	43.695000
50%	1.117845e+09	1.000000	46.830000	43.270000	59.000000	43.425000	47.000000	43.000000	48.900000
75%	8.812239e+10	1.000000	52.725000	48.670000	62.000000	45.205000	50.755000	46.850000	51.530000
máximo	9.009288e+10	3.000000	64.000000	55.730000	64.000000	55.130000	73.000000	76.000000	74.000000

Tabla 3 Análisis estadístico notas ICFES

A hora se leen los 24 archivos .xlsx en panda tomando los siguientes datos, usando un for recorreremos el directorio de /Data/icfes+matricula/, los cuales son 24.

```
for nombre in ficheros:
    datos = pd.read_excel('Data/icfes+matricula/'+ nombre, header=1)
    datos = datos.dropna(axis=1, how = 'all')
    datosFiltrado = datos[listaFiltro]
    datosIcfes.append(datosFiltrado)
    print(nombre)
```

```
icfes2006-2.xlsx
icfes2007-1.xlsx
icfes2007-2.xlsx
icfes2008-1.xlsx
icfes2008-2.xlsx
icfes2009-1.xlsx
icfes2009-2.xlsx
icfes2010-1.xlsx
icfes2010-2.xlsx
icfes2011-1.xlsx
icfes2011-2.xlsx
icfes2012-1.xlsx
icfes2012-2.xlsx
icfes2013-1.xlsx
icfes2013-2.xlsx
icfes2014-1.xlsx
icfes2014-2.xlsx
icfes2015-1.xlsx
icfes2015-2.xlsx
icfes2016-1.xlsx
icfes2016-2.xlsx
icfes2017-1.xlsx
icfes2017-2.xlsx
icfes2018.xlsx
```

Ilustración 7 Resultado de Ficheros en estudio

El siguiente paso es realizar este análisis estadístico a cada materia por archivo y usando todos los archivos se obtienen las siguientes graficas.

En esta grafica observamos los resultados obtenidos por los aspirantes al pregrado en las icfes en la asignatura de Matemáticas, en la cual se observa que con el pasar de los años el promedio de las notas a aumentado.

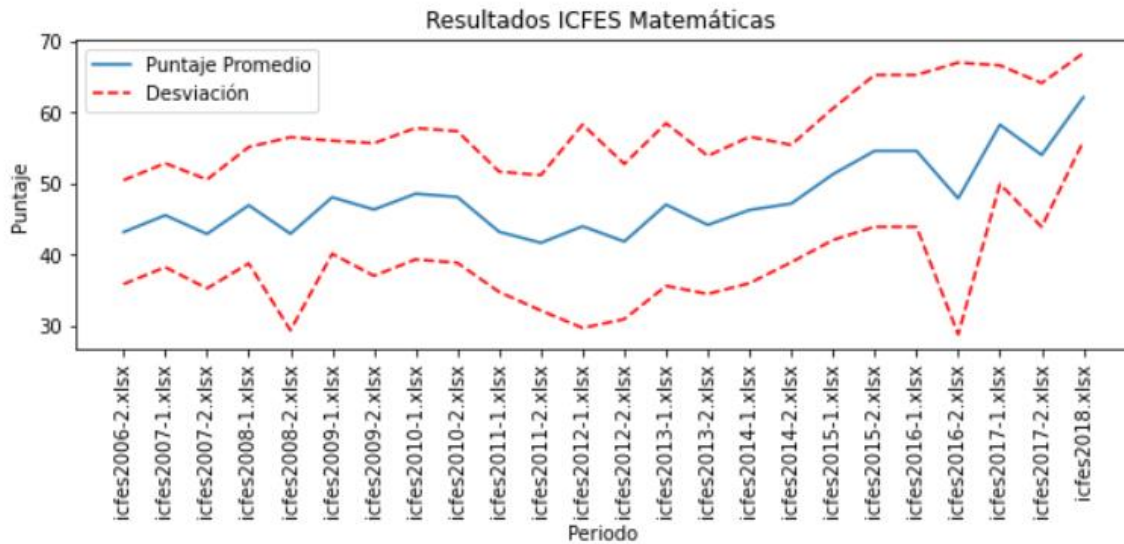


Ilustración 8 grafica de notas obtenidas en Matemáticas en las pruebas de estado icfes

Usando el mismo código utilizado para las notas de Matemáticas se usa para los puntajes obtenidos en estos 24 periodos académicos para la materia de Lenguaje, para este caso la desviación estándar es mayor tenemos desviaciones de 6 a 15 puntos.

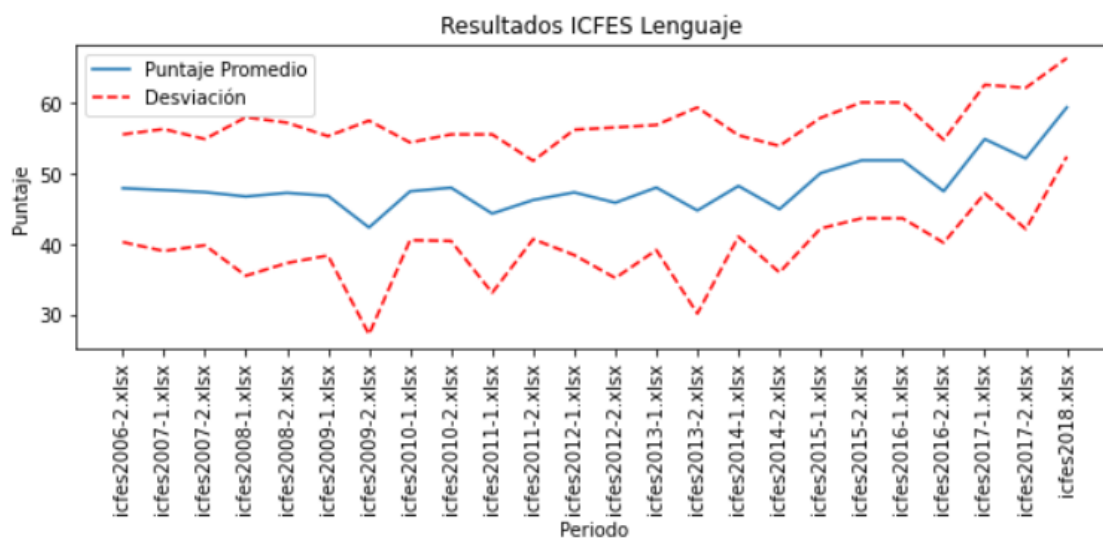


Ilustración 9 Grafica de notas obtenidas en la materia lenguaje en las pruebas de estado icfes.

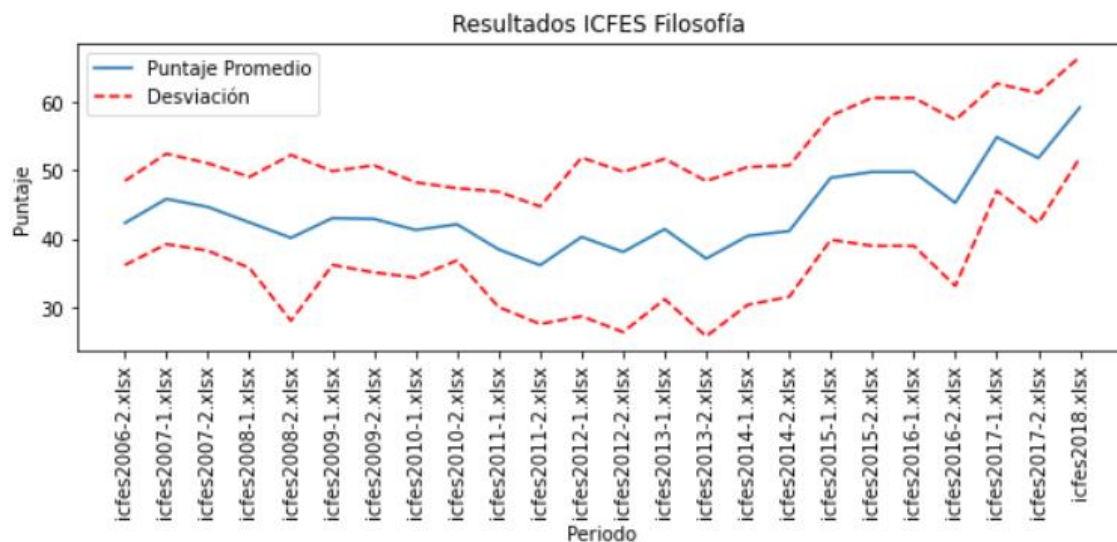


Ilustración 10 grafica de notas obtenidas en Filosofía en las pruebas de estado icfes

En filosofía los aspirantes al programa de ingeniería de sistemas obtuvieron un promedio desde 37 a 59 puntos y mantiene una desviación estándar entre 5 y 11 según los promedios obtenidos en cada semestre.

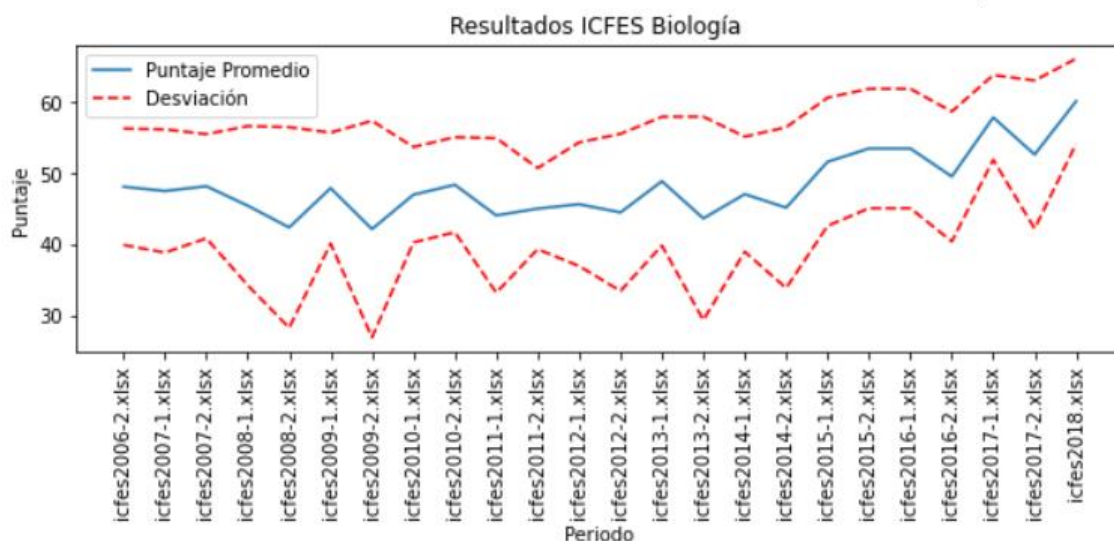


Ilustración 11 Grafica para notas en Biología en las Pruebas icfes

Para biología en la recopilación de notas observamos que tuvieron dos periodos con una desviación de alrededor de 14 puntos por esta razón se observan esos dos picos en icfes2008-2 y icfes2009-2.

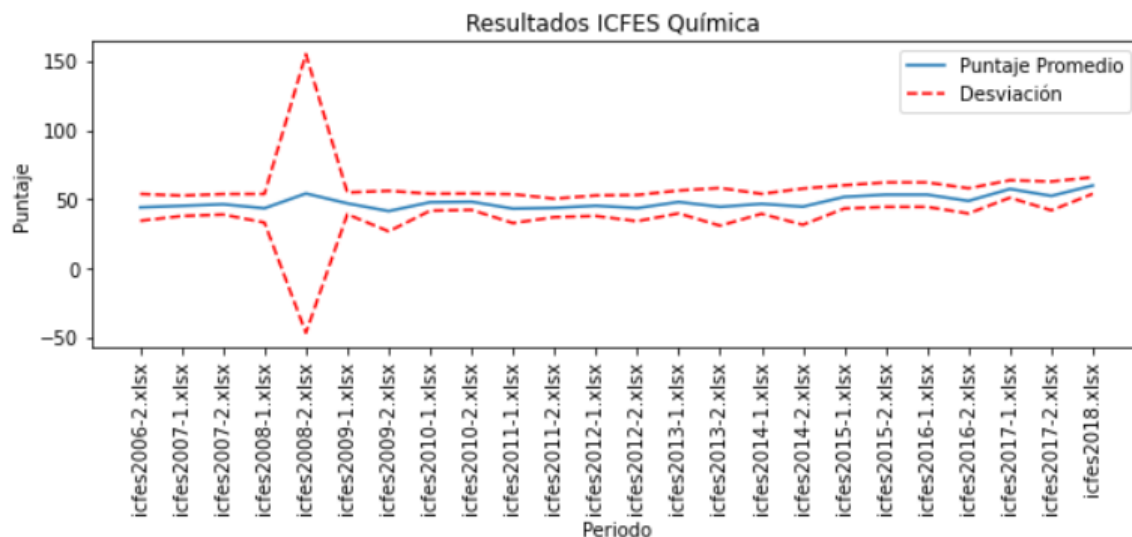


Ilustración 12 Notas de Química en las Icfes

En esta grafica observamos que el promedio está muy cerca de un puntaje de 50, adicional observamos que en algunos periodos la desviación esta muy elevada alrededor de un diferencial de 100.

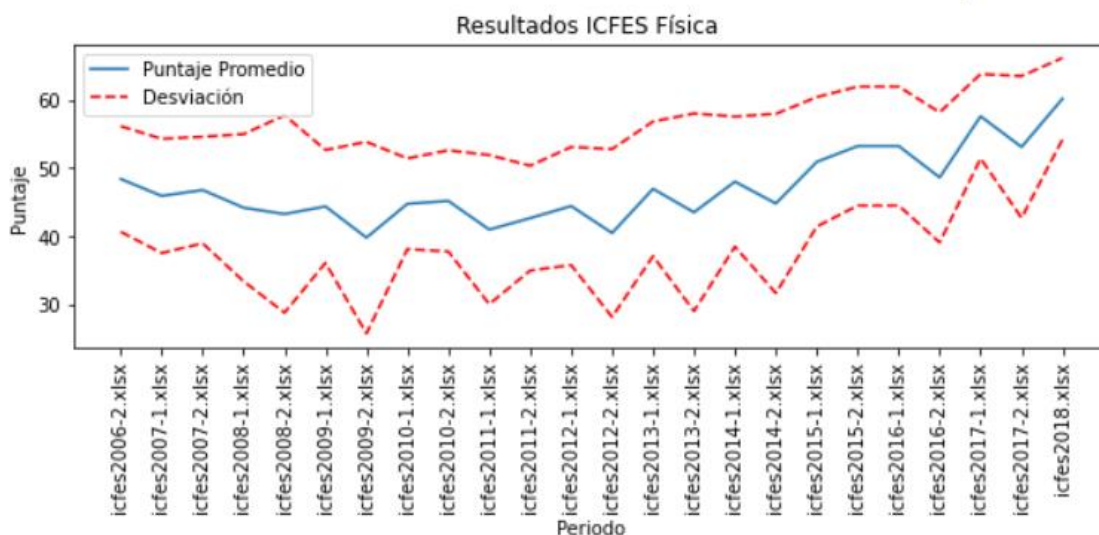


Ilustración 13 Grafica puntajes de Física en las icfes

Los resultados obtenidos en las pruebas de estado para Física todas oscilan en promedios de 39 a 60 y la desviación estándar para estos periodos esta desde 5.9 a 14.

A continuación, en la misma grafica veremos los resultados de todas las materias periodo a periodo, en el cual se puede observar que la mayoría de notas obtenidas en estos exámenes esta alrededor de los 40 y 50 puntos, podemos evidenciar que filosofía en diversos periodos es la materia con menor puntajes, para química fueron mejorando los puntajes a medida que avanzaban los semestres

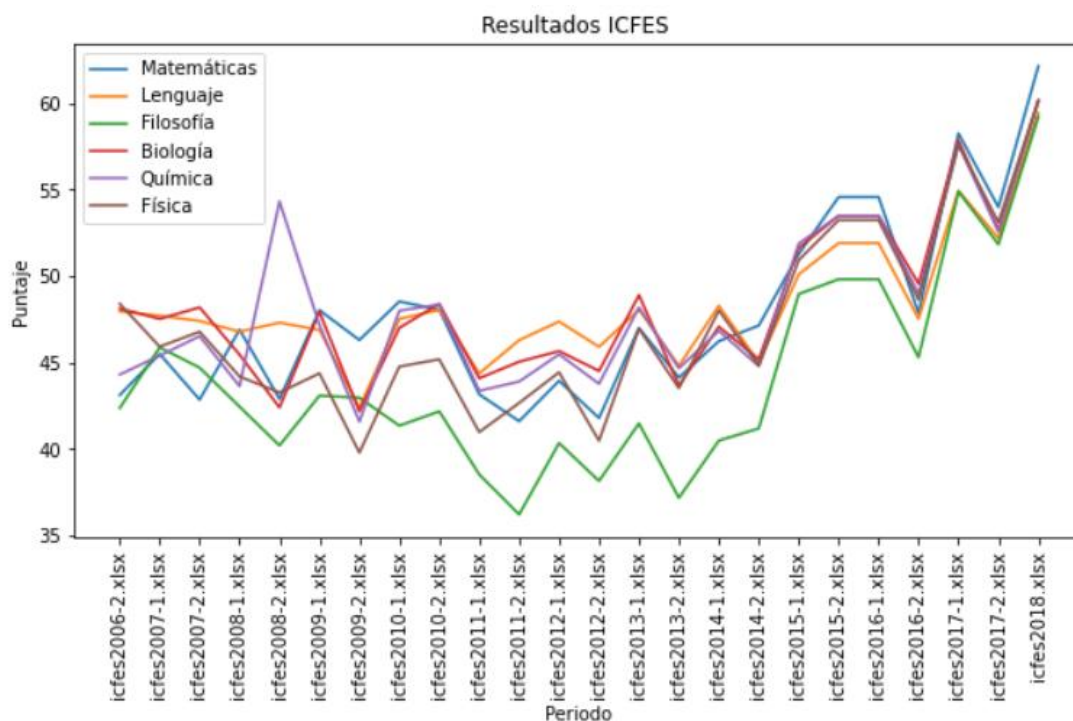


Ilustración 14 Grafica Resultados ICFES

Para siguiente grafica tenemos las notas de Matemáticas, Lenguaje y Física. En la gráfica observamos que a pesar de que en lenguaje no usamos números la tendencia de las notas de estas tres materias es muy similar.

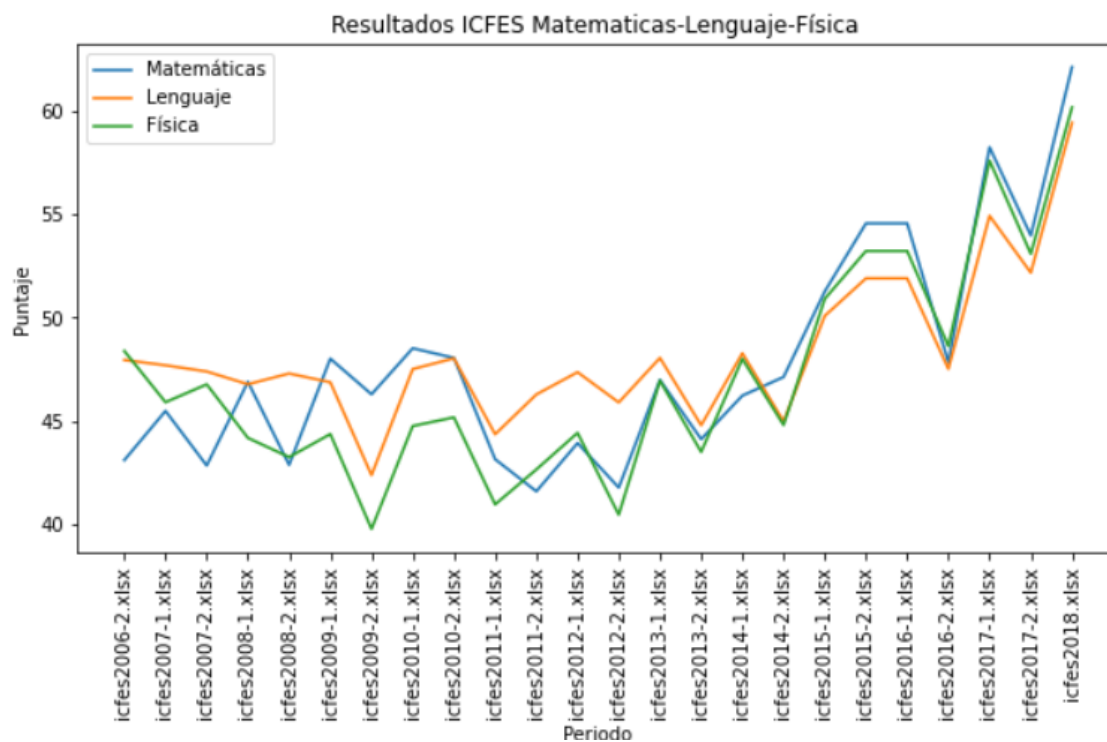


Ilustración 15 Resultados ICFES Matemáticas, Lenguaje y Física

MATERIAS DEL PREGRADO

A la hora de graficar las notas de las materias de pregrado se tienen en cuenta la muestra de 1763 registros de cursar una materia y conservando para el cálculo del promedio con la obviedad de los espacios en blanco. Al ejecutar el `datosMat.head()` nos muestra las primeras 5 filas y nos indica que trae relación con 62 columnas.

	DOCUMENTO	PROG_NOMBRE	SITUACION	PENSUM	157005	156001	157019	157006	157020	1672
0	1000516405	INGENIERÍA DE SISTEMAS	ACTIVO	INGENIERÍA DE SISTEMAS 2006	Nan	3.2	Nan	Nan	Nan	N
1	1001993499	INGENIERÍA DE SISTEMAS	ACTIVO	INGENIERÍA DE SISTEMAS 2006	3.3	Nan	Nan	3.7	Nan	N
2	1002047981	INGENIERÍA DE SISTEMAS	ACTIVO	INGENIERÍA DE SISTEMAS 2006	3.5	3.8	Nan	Nan	Nan	N
3	1002278170	INGENIERÍA DE SISTEMAS	ACTIVO	INGENIERÍA DE SISTEMAS 2006	Nan	4.1	Nan	Nan	Nan	N
4	1002375281	INGENIERÍA DE SISTEMAS	ACTIVO	INGENIERÍA DE SISTEMAS 2006	3.4	3.4	3.1	3.3	Nan	N

5 filas × 62 columnas

Tabla 4 Materias de pregrado

CREAR TABLA CON EL NUMERO DE REGISTRO DE NOTA

Crear un vector con las columnas del arreglo para consultar las materias que se la han registrado varias notas definitivas. Hacer un recorrido recurrente para verificar en cada ubicación, para verificar cuantas veces se le han registrado notas.

```
tablaVistas['DOCUMENTO'] = datosMat['DOCUMENTO']
tablaVistas.head(10)
```

	DOCUMENTO	157005	156001	157019	157006	157020	167248	167202	164018	167270	...	150001	157102	157002	167238	167213	167214	167252	168005	167225	167328
0	1000516405	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	1001993499	1	0	0	1	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0
2	1002047981	1	1	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	0
3	1002278170	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	1002375281	1	1	1	1	0	0	0	0	0	...	0	0	1	1	0	0	0	0	0	0
5	1002497374	1	1	1	1	0	0	0	0	0	...	1	0	1	0	0	0	0	0	0	0
6	1003080578	2	1	1	0	0	0	0	0	0	...	1	0	2	0	0	0	0	0	0	0
7	1004319762	1	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
8	1004320583	1	0	1	0	0	0	0	0	0	...	0	0	2	0	0	0	0	0	0	0
9	1004802359	2	2	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

10 rows × 59 columns

Tabla 5 Numero de vistas por materia

Creamos una relación en Materias2006 con su código y nombre de la materia de pregrado

```
materias2006 = {'156001': 'BIOLOGÍA GENERAL',
'157005': 'CÁLCULO DIFERENCIAL',
'153002': 'CÁTEDRA FARÍA',
'164004': 'EDUCACIÓN AMBIENTAL',
'162003': 'HABILIDADES COMUNICATIVAS',
'156005': 'LABORATORIO BIOLOGÍA GENERAL',
'167250': 'PROGRAMACIÓN ESTRUCTURADA',

'157006': 'CÁLCULO INTEGRAL',
'157015': 'LABORATORIO DE MECÁNICA',
'156006': 'LABORATORIO DE QUÍMICA GENERAL',
'157019': 'MECÁNICA',
'167251': 'PROGRAMACIÓN ORIENTADA A OBJETOS',
'156009': 'QUÍMICA GENERAL',

'157002': 'ÁLGEBRA LINEAL',
'157007': 'CÁLCULO MULTIVARIABLE',
'157009': 'ELECTROMAGNETISMO',
'167231': 'ESTRUCTURAS DE DATOS DINÁMICOS Y ALGORITMOS',
'157014': 'LABORATORIO DE ELECTROMAGNETISMO',

'167201': 'ANÁLISIS DE ALGORITMOS',
'157008': 'ECUACIONES DIFERENCIALES',
'168003': 'EXPRESIÓN GRÁFICA',
'167238': 'INTRODUCCIÓN A LAS CIENCIAS COMPUTACIONALES',
'157016': 'LABORATORIO DE OSCILACIONES Y ONDAS',
'157020': 'OSCILACIONES Y ONDAS',

'167232': 'FUNDAMENTOS Y DISEÑO DE BASE DE DATOS',
'167233': 'FUNDAMENTOS Y DISEÑO DE SISTEMAS OPERATIVOS',
'167243': 'LÓGICA COMPUTACIONAL',
'157102': 'MATEMÁTICAS ESPECIALES',
'157103': 'MÉTODOS NUMÉRICOS',
'167248': 'PARADIGMAS DE PROGRAMACIÓN',

'167202': 'BASES DE DATOS II',
'150001': 'ELECTIVA SOCIOHUMANÍSTICA I',
'157012': 'ESTADÍSTICA II',
'167246': 'MODELAMIENTO DE SISTEMAS DE INFORMACIÓN',
'167265': 'SISTEMAS Y MODELADO I',

'167270': 'TEORÍA DE LA COMPUTACIÓN',

'167213': 'ELECTIVA DE INGENIERÍA',
'167219': 'ELECTIVA PROFESIONAL',
'167256': 'REDES DE COMPUTADORES Y SISTEMAS DISTRIBUIDOS',
'167266': 'SISTEMAS Y MODELADO II',
'167271': 'TÓPICOS AVANZADOS DE INGENIERÍA DEL SOFTWARE',

'167214': 'ELECTIVA DE INGENIERÍA',
'167222': 'ELECTIVA PROFESIONAL II',
'168106': 'INGENIERÍA DE PROYECTOS I',
'168005': 'INVESTIGACIÓN DE OPERACIONES',
'167240': 'LEGISLACIÓN INFORMÁTICA',
'167257': 'REDES DE COMPUTADORES Y SISTEMAS DISTRIBUIDOS II',

'167215': 'ELECTIVA DE INGENIERÍA III',
'167225': 'ELECTIVA PROFESIONAL III',
'150002': 'ELECTIVA SOCIOHUMANÍSTICA II',
'164010': 'ÉTICA',
'164018': 'METODOLOGÍA DE LA INVESTIGACIÓN',
'167252': 'PROTECCIÓN DE SISTEMAS INFORMÁTICOS',

'173238': 'ACTIV DEPOR, RECRE Y CULT-DEP UNIVERSITARIO',
'369006': 'CÍVICA Y CONSTITUCIÓN (EXTRAPLAN)',
'167281': 'INFORMÁTICA BÁSICA (EXTRAPLAN)',
'167328': 'TRABAJO DE GRADO',

}
```

Tabla 6 Listado de pensum2006

Luego de tener la relacion de materias basado en el pensum 2006, usando estas lineas de codigo creamos las graficas de descripcion de las notas obtenidas en las materias de pregrado.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: CÁLCULO DIFERENCIAL

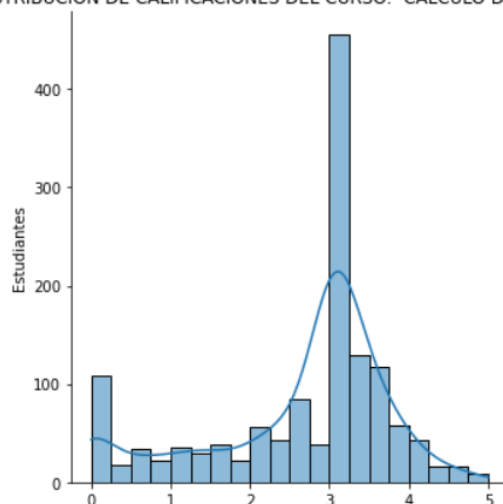


Ilustración 16 Distribución de calificaciones de Calculo diferencia

Para CÁLCULO DIFERENCIAL se obtuvo un promedio de 2,62 con un valor moda de 3.0, como nota máxima 5 y como nota minina 0.0 para un total de 1373 registros de notas en la muestra, con una desviación estándar de 1,15.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: BIOLOGÍA GENERAL

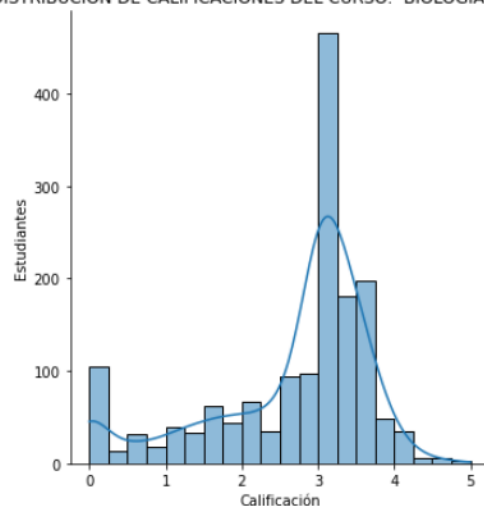


Ilustración 17 Distribución de calificaciones de Biología general

Para biología general se obtuvo un promedio de 2,65 con un valor moda de 3.0, como nota máxima 5 y como nota minina 0.0 para un total de 1580 notas en la muestra, para esta materia la desviación estándar de 1,06.

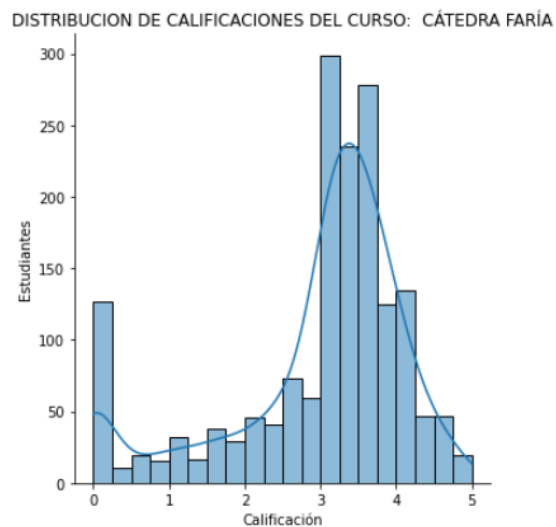


Ilustración 18 Distribución de calificaciones Catedra Faria

Para Catedra faria se obtuvo un promedio de 2,967 con un valor moda de 3.3, como nota máxima 5 y como nota minina 0.0 para un total de 1691 notas en la muestra, para catedra la desviación es 1,18.

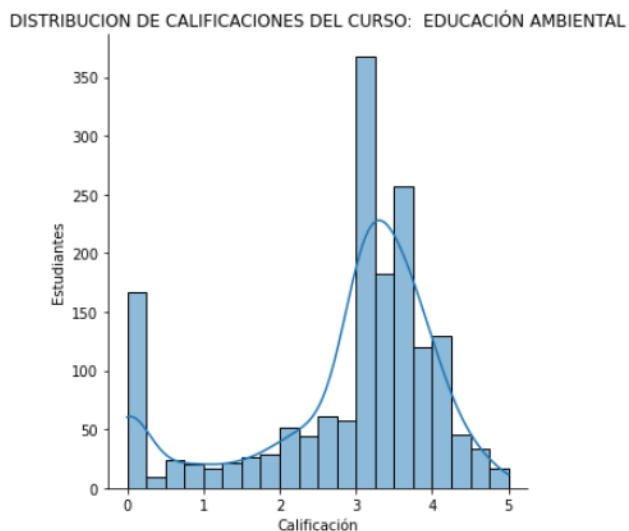


Ilustración 19 Distribución de calificaciones Educacion Ambiental

Para Educación ambiental se obtuvo un promedio de 2,867 con un valor moda de 3.2, como nota máxima 5 y como nota minina 0.0 para un total de 1681 notas en la muestra, con una desviación de 1,24.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: HABILIDADES COMUNICATIVAS

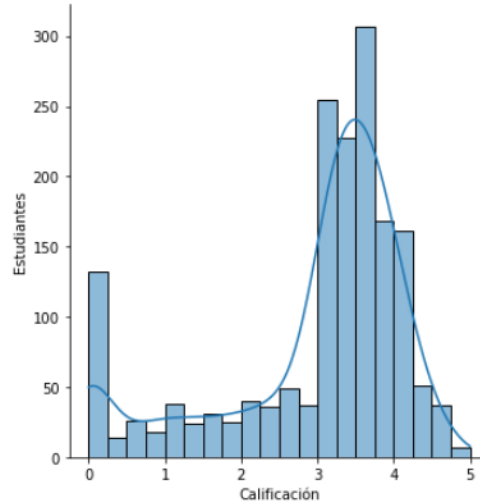


Ilustración 20 Distribución de calificaciones Habilidades comunicativas

Para Habilidades comunicativas se obtuvo un promedio de 2,96 con un valor moda de 3.36, como nota máxima 5 y como nota mínima 0.0 para un total de 1683 notas en la muestra, para habilidades la desviación es de 1,22.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: LABORATORIO BIOLOGÍA GENERAL

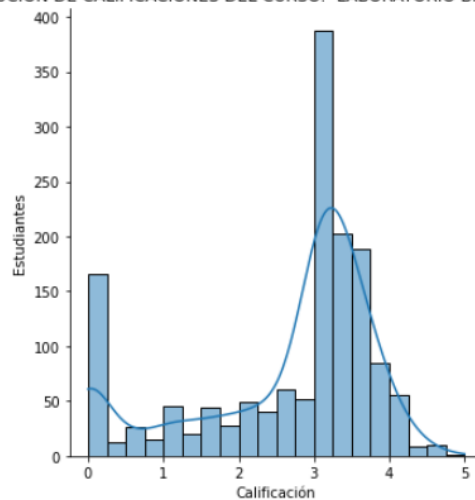


Ilustración 21 Distribución de calificaciones Laboratorio biología general

Para Biología general se obtuvo un promedio de 2,62 con un valor moda de 3.1, como nota máxima 5 y como nota mínima 0.0 para un total de 1499 notas en la muestra, para laboratorio de biología es de 1,22.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: PROGRAMACIÓN ESTRUCTURADA

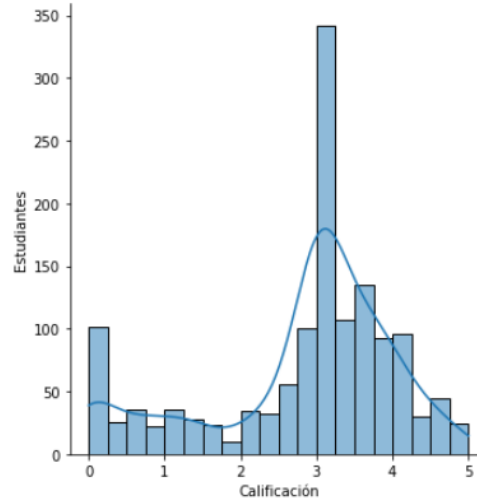


Ilustración 22 Distribución de calificaciones Programación estructurada

Para programación estructurada se obtuvo un promedio de 2,81 con un valor moda de 3.1, como nota máxima 5.0 y como nota mínima 0.0 para un total de 1374 notas en la muestra, 1,24 es la desviación calculada para programación.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: CÁLCULO INTEGRAL

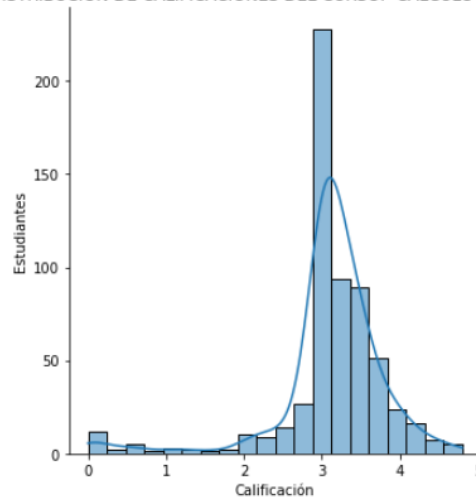


Ilustración 23 Distribución de calificaciones Calculo integral

Para programación estructurada se obtuvo un promedio de 3,113 con un valor moda de 3.1, como nota máxima 4.8 y como nota mínima 0.0 para un total de 602 notas en la muestra, con una desviación de 0,71.

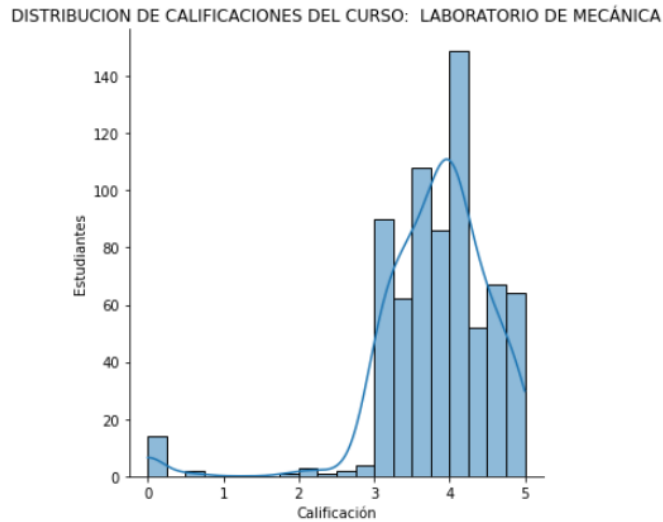


Ilustración 24 Distribución de calificaciones laboratorio de mecánica

Para Laboratorio mecánica se obtuvo un promedio de 3,81 con un valor moda de 3.9, como nota máxima 5 y como nota mínima 0.0 para un total de 705 notas en la muestra, con una desviación de 0,80.

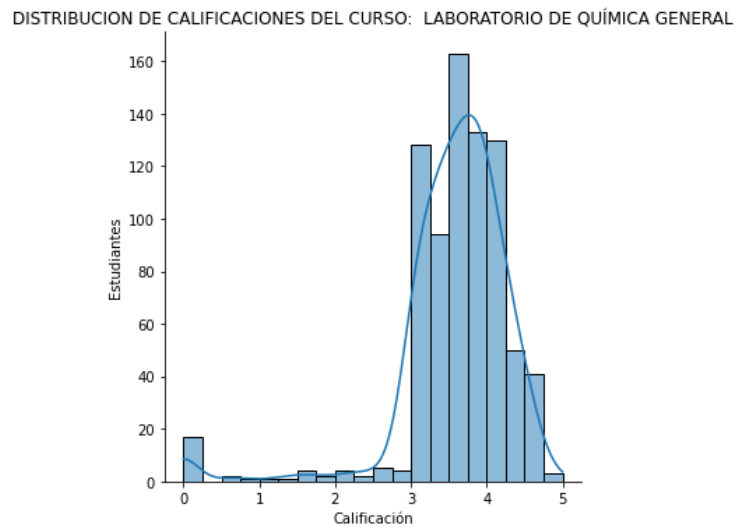


Ilustración 25 Distribución de calificaciones laboratorio de química

Para Laboratorio mecánica se obtuvo un promedio de 3,57 con un valor moda de 3.7, como nota máxima 5 y como nota mínima 0.0 para un total de 785 notas en la muestra, con una desviación de 0,758.

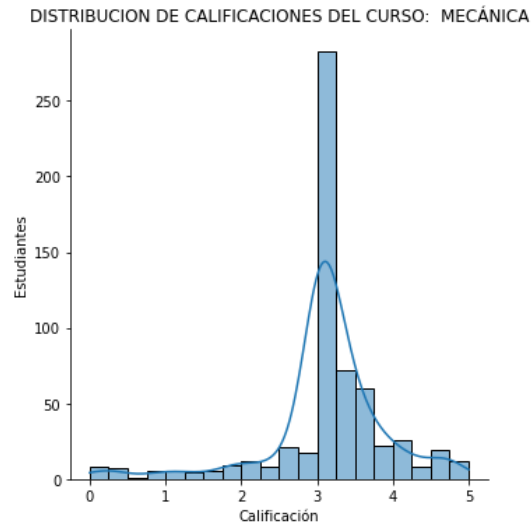


Ilustración 26 Distribución de calificaciones Mecánica

Para mecánica se obtuvo un promedio de 3,12 con un valor moda de 3.1, como nota máxima 5 y como nota mínima 0.0 para un total de 604 notas en la muestra, con una desviación de 0,801.

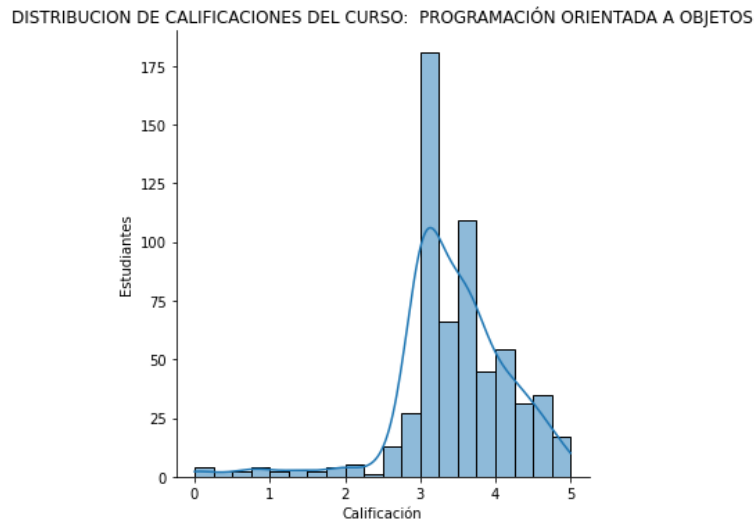


Ilustración 27 Distribución de calificaciones POO

Para programación orientada a objetos se obtuvo un promedio de 3,43 con un valor moda de 3.4, como nota máxima 5 y como nota mínima 0.0 para un total de 607 notas en la muestra, con una desviación de 0,760.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: QUÍMICA GENERAL

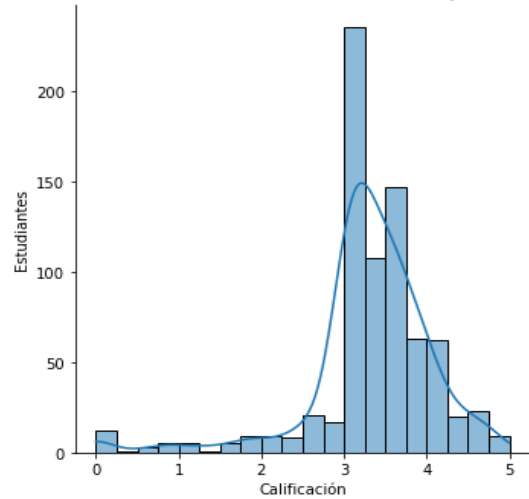


Ilustración 28 Distribución de calificaciones Química General

Para química general se obtuvo un promedio de 3,31 con un valor moda de 3.3, como nota máxima 5 y como nota minina 0.2 para un total de 764 notas en la muestra, con una desviación de 0,769.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ÁLGEBRA LINEAL

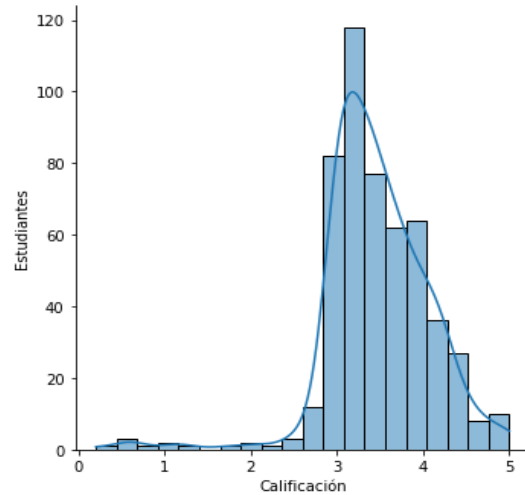


Ilustración 29 Distribución de calificaciones Algebra Lineal

Para algebra lineal se obtuvo un promedio de 3,47 con un valor moda de 3.4, como nota máxima 5 y como nota minina 0.2 para un total de 511 notas en la muestra, con una desviación de 0,612.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: CÁLCULO MULTIVARIABLE

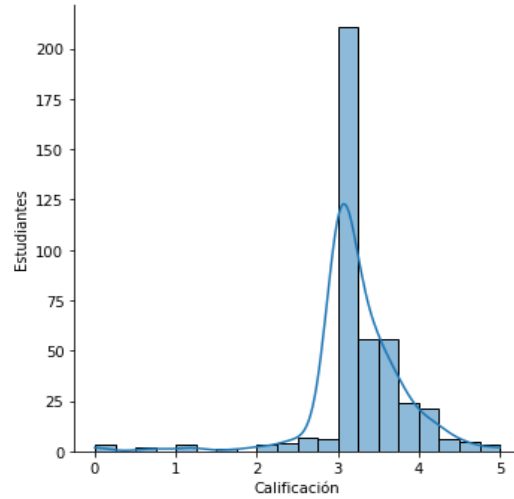


Ilustración 30 Distribución de calificaciones Calculo Multivariable

Para Calculo Multivariable se obtuvo un promedio de 3,24 con un valor moda de 3.1, como nota máxima 5 y como nota mínima 0.0 para un total de 411 notas en la muestra, con una desviación de 0,579.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ELECTROMAGNETISMO

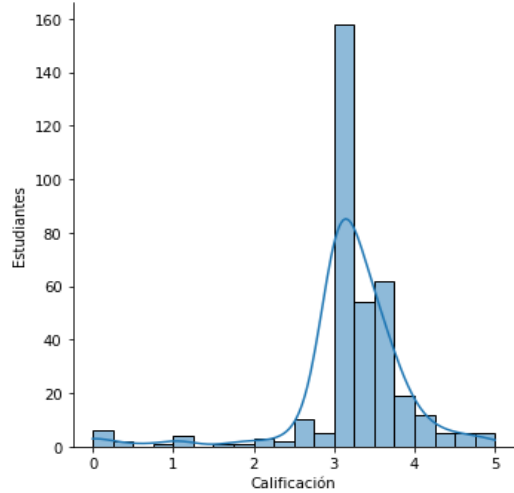


Ilustración 31 Distribución de calificaciones Electromagnetismo

Para Calculo Multivariable se obtuvo un promedio de 3,2 con un valor moda de 3.2, como nota máxima 5 y como nota mínima 0.0 para un total de 355 notas en la muestra, con una desviación de 0,709.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ESTRUCTURAS DE DATOS DINÁMICOS Y ALGORITMOS

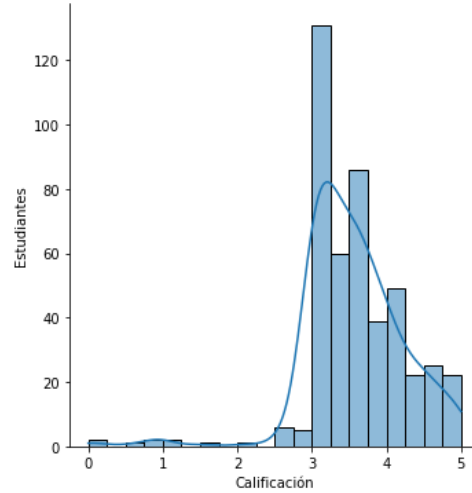


Ilustración 32 Distribución de calificaciones Estructuras de datos dinámicos y algoritmos

Para estructuras de datos dinámicos y algoritmos se obtuvo un promedio de 3,56 con un valor moda de 3.5, como nota máxima 5 y como nota mínima 0.0 para un total de 454 notas en la muestra, con una desviación de 0,672.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: LABORATORIO DE ELECTROMAGNETISMO

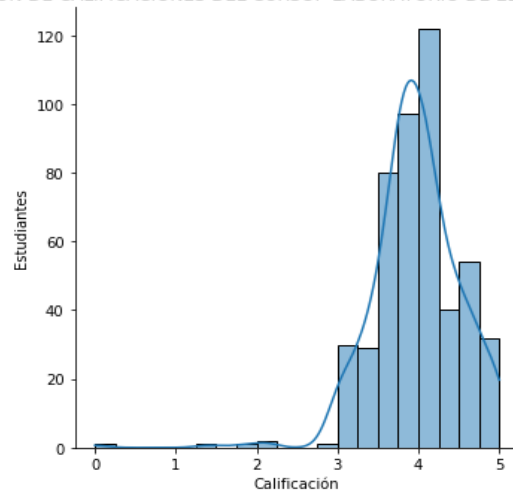


Ilustración 33 Distribución de calificaciones Laboratorio de Electromagnetismo

Para laboratorio de electromagnetismo se obtuvo un promedio de 3,96 con un valor moda de 4.0, como nota máxima 5.0 y como nota mínima 0.0 para un total de 490 notas en la muestra, con una desviación de 0,541.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ANÁLISIS DE ALGORITMOS

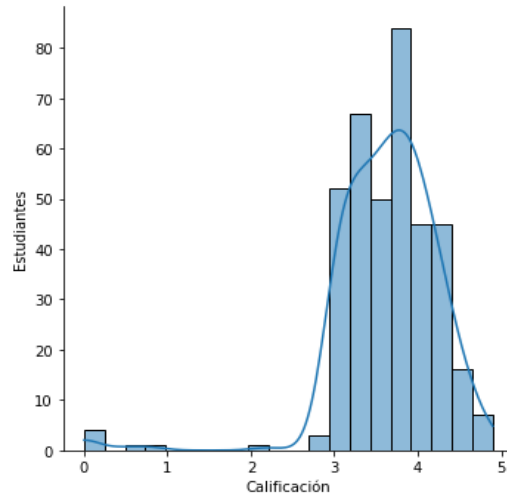


Ilustración 34 Distribución de calificaciones Análisis de Algoritmos

Para análisis de algoritmo se obtuvo un promedio de 3,64 con un valor moda de 3.7, como nota máxima 4,9 y como nota minina 0.0 para un total de 376 notas en la muestra, con una desviación de 0,641.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ECUACIONES DIFERENCIALES

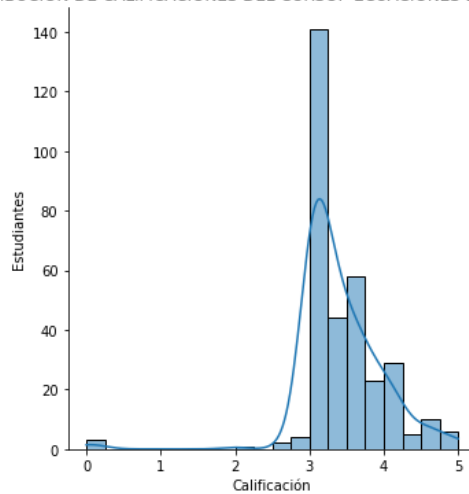


Ilustración 35 Distribución de calificaciones Ecuaciones diferenciales

Para ecuaciones diferenciales se obtuvo un promedio de 3,41 con un valor moda de 3.3, como nota máxima 5.0 y como nota minina 0.0 para un total de 326 notas en la muestra, con una desviación de 0,575.

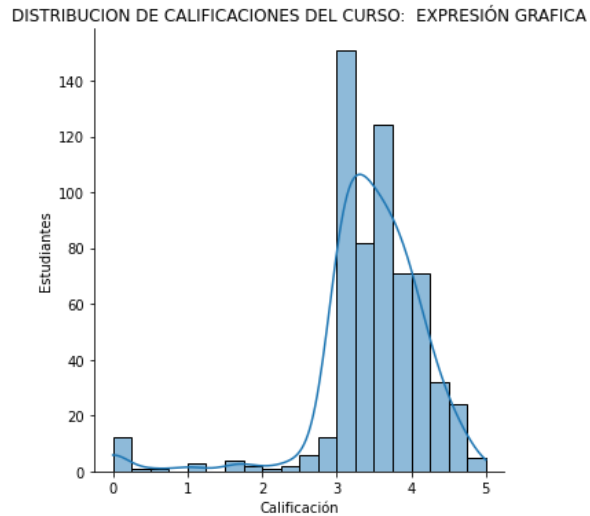
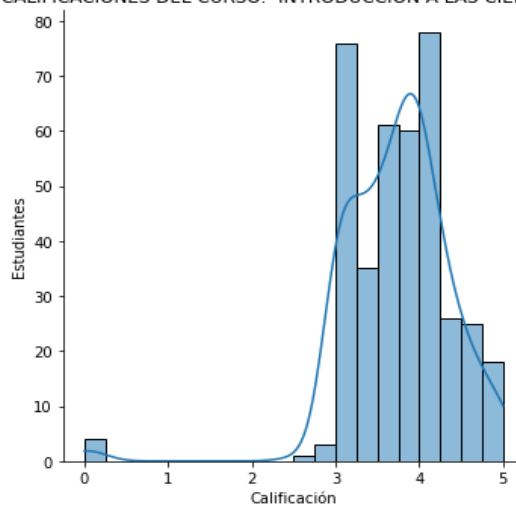


Ilustración 36 Distribución de calificaciones Expresión grafica

Para expresión grafica se obtuvo un promedio de 3,46 con un valor moda de 3.5, como nota máxima 5.0 y como nota minina 0.0 para un total de 604 notas en la muestra, con una desviación

DISTRIBUCION DE CALIFICACIONES DEL CURSO: INTRODUCCIÓN A LAS CIENCIAS COMPUTACIONALES



de 0,752.

Ilustración 37 Distribución de calificaciones introducción a las ciencias computacionales

Para Introducción a las ciencias computacionales se obtuvo un promedio de 3,74 con un valor moda de 3.8, como nota máxima 5.0 y como nota minina 0.0 para un total de 387 notas en la muestra, con una desviación de 0,651.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: LABORATORIO DE OSCILACIONES Y ONDAS

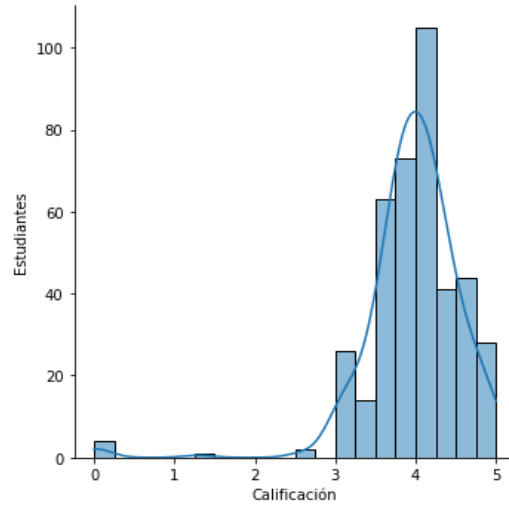


Ilustración 38 Distribución de calificaciones Lab de oscilaciones y ondas

Para laboratorio de Oscilaciones y ondas se obtuvo un promedio de 3,96 con un valor moda de 4, como nota máxima 5.0 y como nota minina 0.0 para un total de 401 notas en la muestra, con una desviación de 0,635.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: OSCILACIONES Y ONDAS

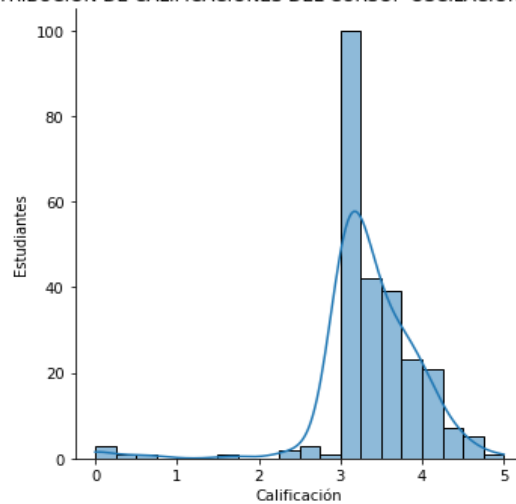


Ilustración 39 Distribución de calificaciones oscilaciones y ondas

Para Oscilaciones y ondas se obtuvo un promedio de 3,35 con un valor moda de 3.3, como nota máxima 5.0 y como nota minina 0.0 para un total de 250 notas en la muestra, con una desviación de 0,635.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: FUNDAMENTOS Y DISEÑO DE BASE DE DATOS

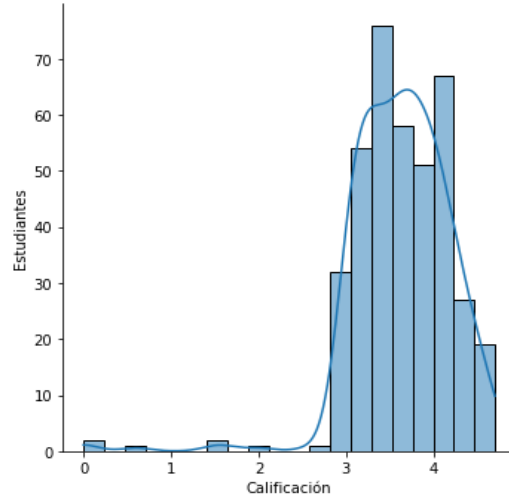


Ilustración 40 Distribución de calificaciones fundamentos y diseños de BD

Para Fundamentos y diseños de base de datos se obtuvo un promedio de 3,63 con un valor moda de 3,7, como nota máxima 4,7 y como nota minina 0,0 para un total de 391 notas en la muestra, con una desviación de 0,566.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: FUNDAMENTOS Y DISEÑO DE SISTEMAS OPERATIVOS

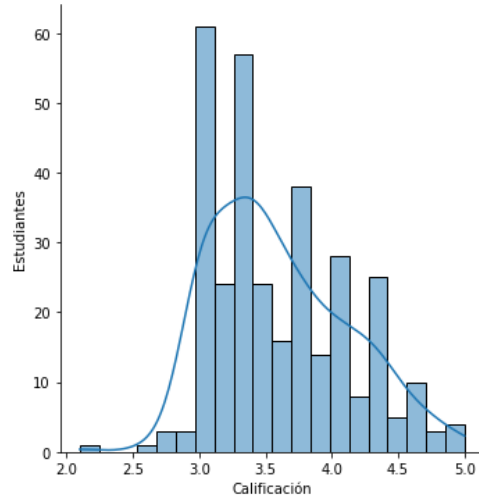


Ilustración 41 Distribución de calificaciones Fundamentos y diseños SO

Para Fundamentos y diseños de sistemas operativos se obtuvo un promedio de 3,6 con un valor moda de 3,5, como nota máxima 5,0 y como nota minina 2,1 para un total de 325 notas en la muestra, con una desviación de 0,516.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: LÓGICA COMPUTACIONAL

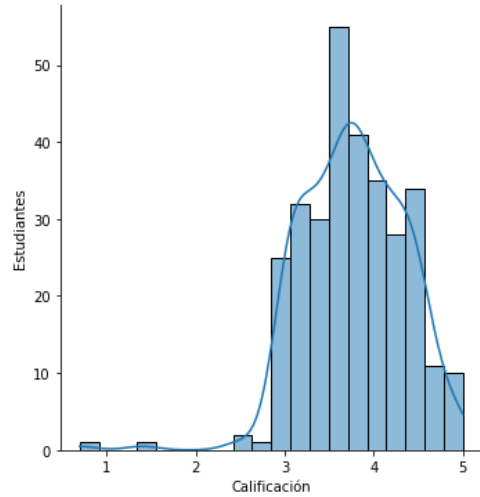


Ilustración 42 Distribución de calificaciones Lógica computacional

Para Lógica computacional se obtuvo un promedio de 3,77 con un valor moda de 3.8, como nota máxima 5.0 y como nota minina 0.7 para un total de 306 notas en la muestra, con una desviación de 0,570.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: MATEMÁTICAS ESPECIALES

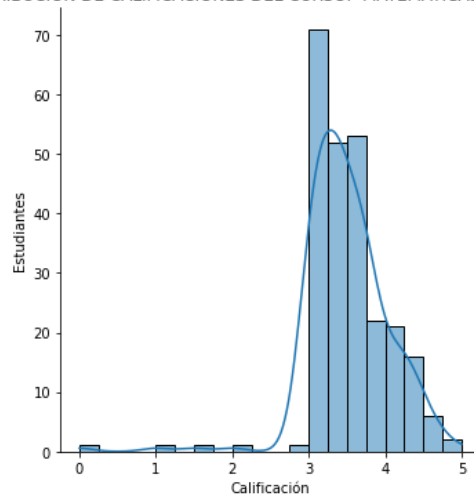


Ilustración 43 Distribución de calificaciones Matemáticas especiales

Para Matemáticas especiales se obtuvo un promedio de 3,5 con un valor moda de 3.4, como nota máxima 5.0 y como nota 0,0 para un total de 248 notas en la muestra, con una desviación de 0,545.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: MÉTODOS NUMÉRICOS

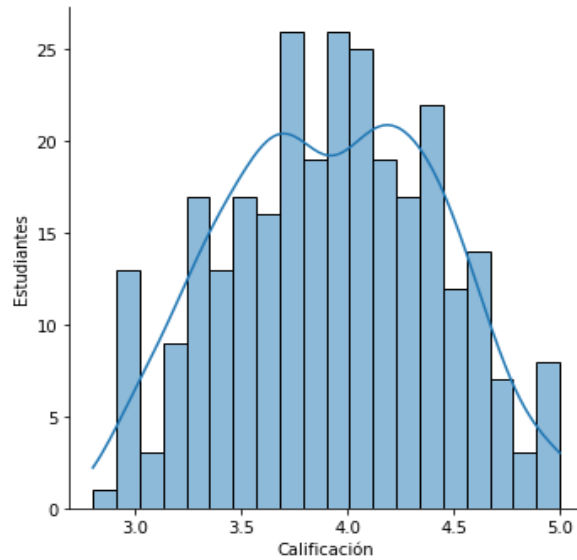


Ilustración 44 Distribución de calificaciones Métodos numéricos

Para Métodos Numéricos se obtuvo un promedio de 3,9 con un valor moda de 3.9, como nota máxima 5.0 y como nota minina 2,8 para un total de 287 notas en la muestra, con una desviación de 0,498.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: PARADIGMAS DE PROGRAMACIÓN

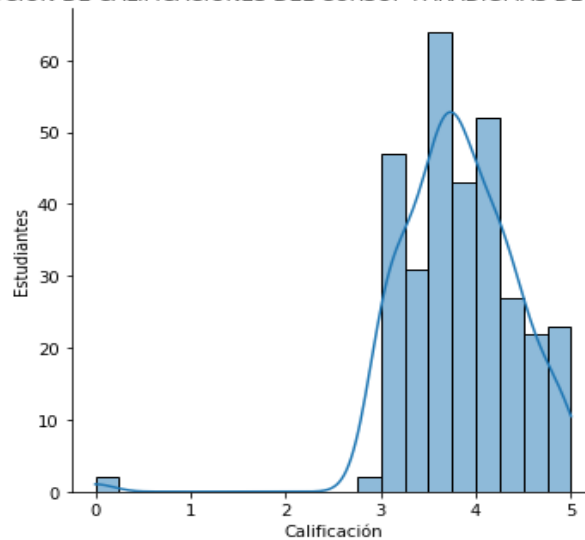


Ilustración 45 Distribución de calificaciones paradigmas

Para paradigmas de programación se obtuvo un promedio de 3,8 con un valor moda de 3.8, como nota máxima 5.0 y como nota minina 0,0 para un total de 313 notas en la muestra, con una desviación de 0,615.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: BASES DE DATOS II

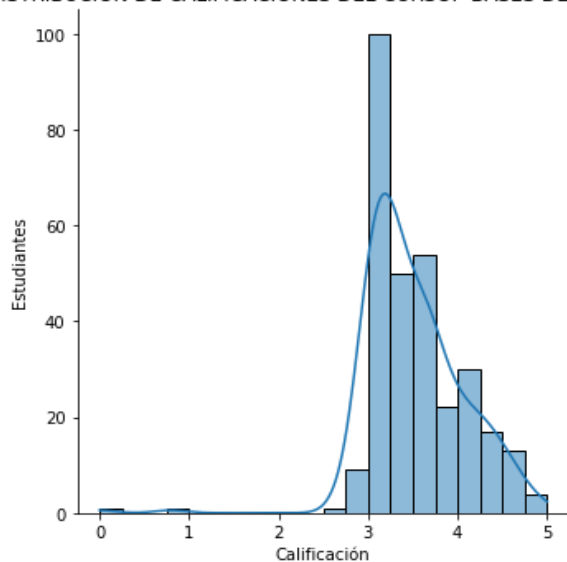


Ilustración 46 Distribución de calificaciones Bases de Datos II

Para bases de datos II se obtuvo un promedio de 3,512 con un valor moda de 3,4, como nota máxima 5.0 y como nota mínima 0,0 para un total de 302 notas en la muestra, con una desviación de 0,553.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ELECTIVA SOCIOHUMANÍSTICA I

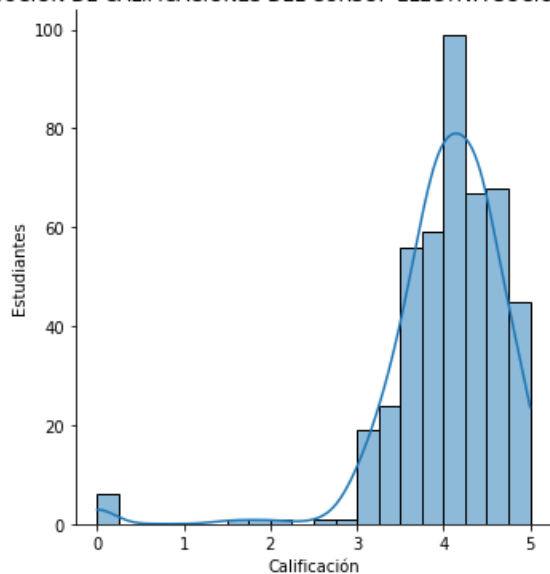


Ilustración 47 Distribución de calificaciones electiva sociohumistica I

Para socio humanística I se obtuvo un promedio de 4,03 con un valor moda de 4,1, como nota máxima 5.0 y como nota mínima 0 para un total de 448 notas en la muestra, con una desviación de 0,700.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ESTADÍSTICA II

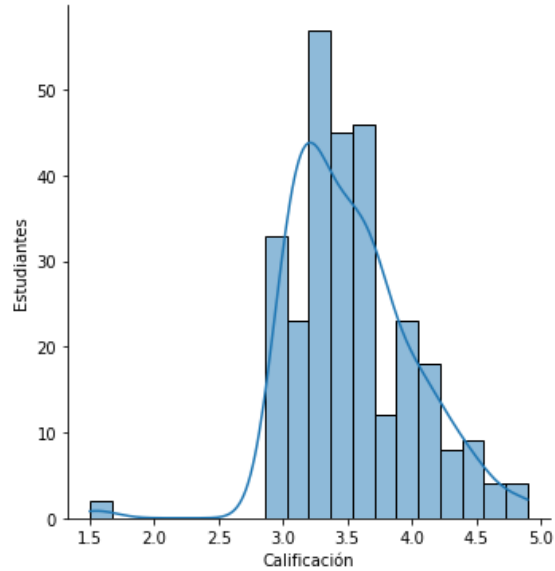


Ilustración 48 Distribución de calificaciones Estadística II

Para estadística II se obtuvo un promedio de 3,53 con un valor moda de 3,5 como nota máxima 4,9 y como nota mínima 1,5 para un total de 248 notas en la muestra, con una desviación de 0,476.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: MODELAMIENTO DE SISTEMAS DE INFORMACIÓN

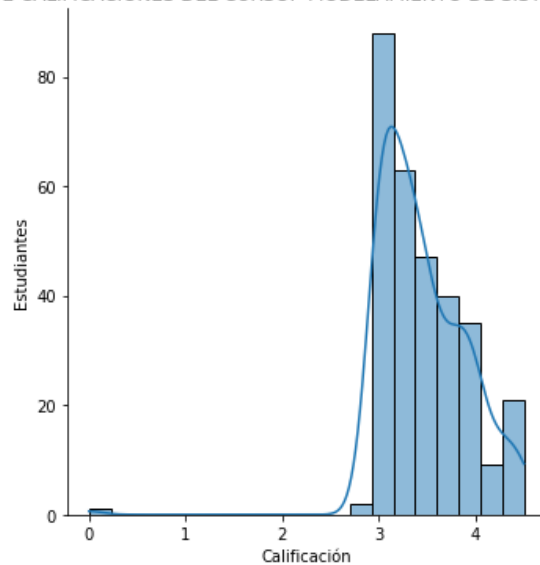


Ilustración 49 Distribución de calificaciones Modelamiento de sistemas de información

Para Modelamiento de sistemas de información se obtuvo un promedio de 3,45 con un valor moda de 3,3 como nota máxima 4,5 y como nota mínima 0,0 para un total de 306 notas en la muestra, con una desviación de 0,464.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: SISTEMAS Y MODELADO I

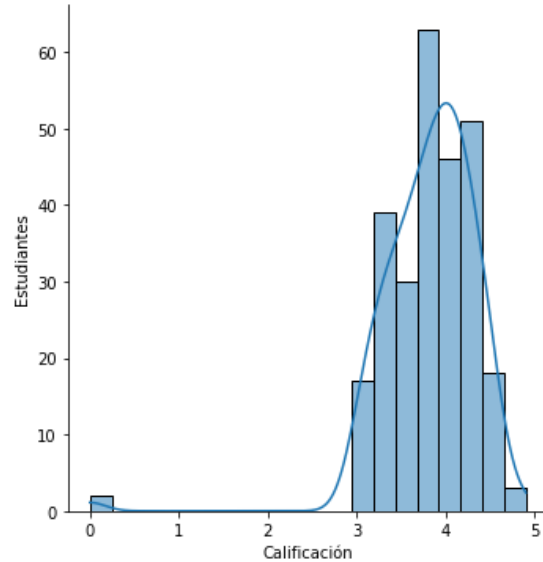


Ilustración 50 Distribución de calificaciones Sistemas y modelado I

Para sistemas y modelados I se obtuvo un promedio de 3,82 con un valor moda de 3,9 como nota máxima 4,9 y como nota mínima 0,0 para un total de 269 notas en la muestra, con una desviación de 0,538.

Para Métodos Numéricos se obtuvo un promedio de 3,9 con un valor moda de 3.9, como nota máxima 5.0 y como nota mínima 2,8 para un total de 287 notas en la muestra, con una desviación de 0,498.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: TEORÍA DE LA COMPUTACIÓN

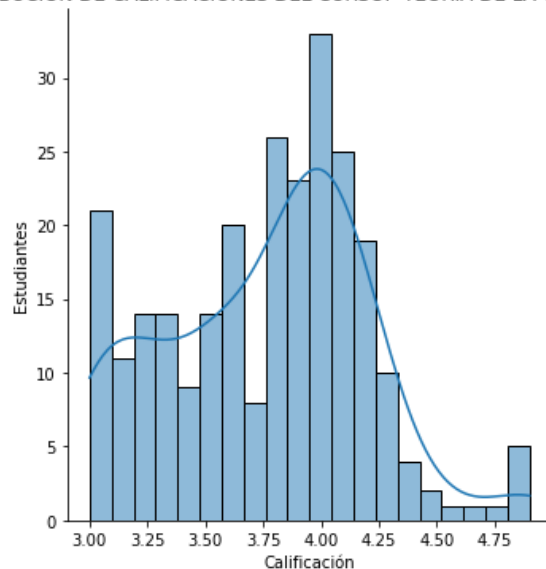


Ilustración 51 Distribución de calificaciones Teoría de la computación

Para teoría de la computación se obtuvo un promedio de 3,75 con un valor moda de 3,8 como nota máxima 4,9 y como nota mínima 0,0 para un total de 261 notas en la muestra, con una desviación de 0,538.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ELECTIVA DE INGENIERÍA

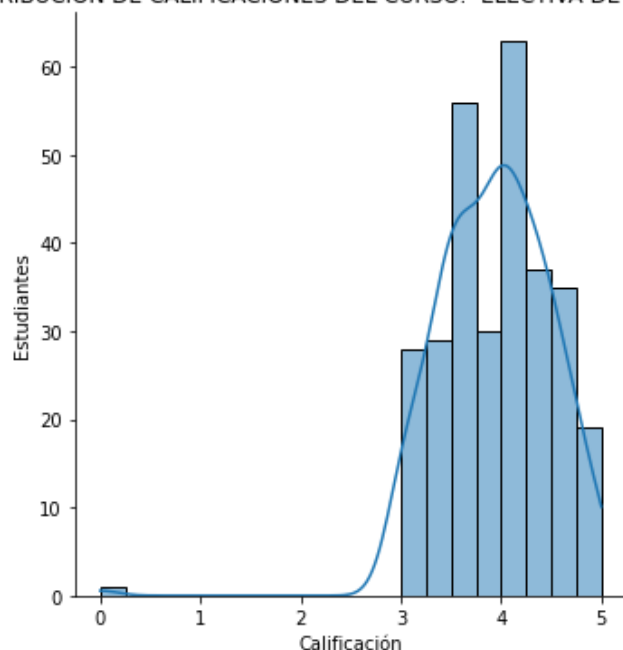


Ilustración 52 Distribución de calificaciones electiva de ingeniería

Para electiva de ingeniería un promedio de 3,93 con un valor moda de 4,0 como nota máxima 5.0 y como nota mínima 0,0 para un total de 298 notas en la muestra, con una desviación de 0,562.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ELECTIVA PROFESIONAL

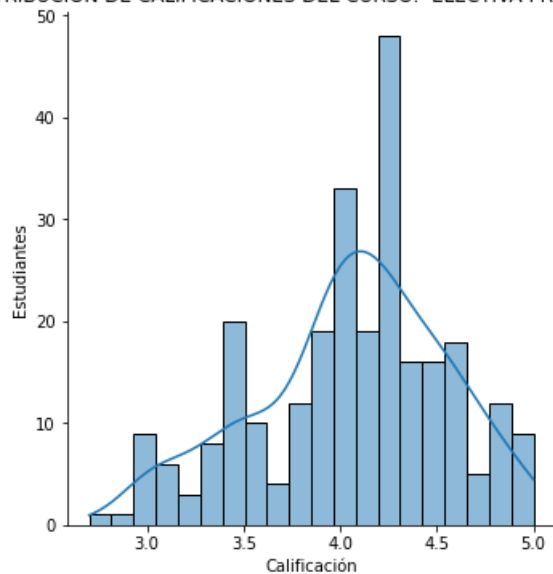


Ilustración 53 Distribución de calificaciones Electiva Profesional

Para electiva profesional se obtuvo se obtuvo un promedio de 4,05 con un valor moda de 4,1 como nota máxima 5.0 y como nota mínima 2,7 para un total de 269 notas en la muestra, con una desviación de 0,490.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: REDES DE COMPUTADORES Y SISTEMAS DISTRIBUIDOS

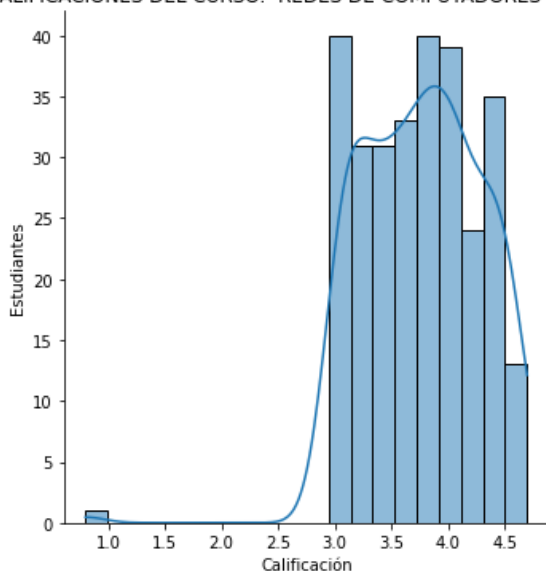


Ilustración 54 Distribución de calificaciones redes de computadores y sistemas distribuidos

Para redes de computadores y sistemas distribuidos se obtuvo un promedio de 3,8 con un valor moda de 3,77 como nota máxima 4,7 y como nota mínima 0,8 para un total de 287 notas en la muestra, con una desviación de 0,522.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: SISTEMAS Y MODELADO II

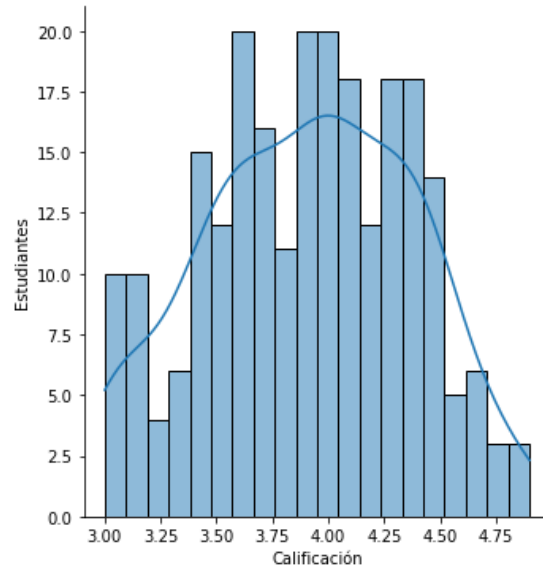


Ilustración 55 Distribución de calificaciones Sistemas y modelados

Para sistemas y modelados II se obtuvo un promedio de 3,9 con un valor moda de 3,9 como nota máxima 4,9 y como nota minina 3.0 para un total de 241 notas en la muestra, con una desviación de 0,469.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: TÓPICOS AVANZADOS DE INGENIERÍA DEL SOFTWARE

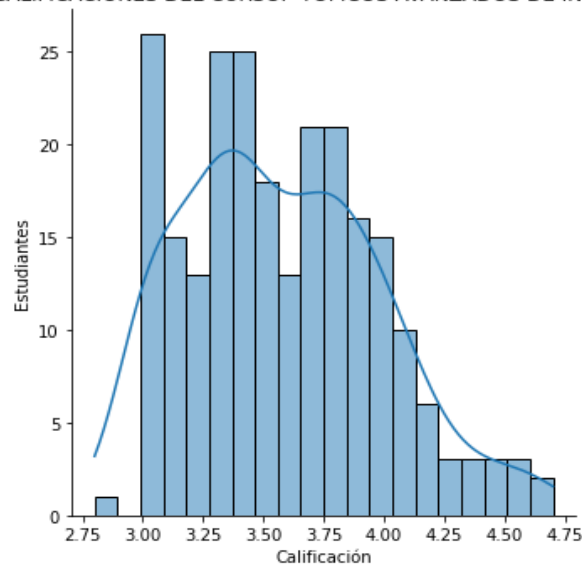


Ilustración 56 Distribución de calificaciones tópicos avanzados de ingeniería del software

Para Avanzados de ingeniería del software se obtuvo un promedio de 3,57 con un valor moda de 3,5 como nota máxima 4,7 y como nota minina 2,8 para un total de 239 notas en la muestra, con una desviación de 0,409.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ELECTIVA DE INGENIERÍA

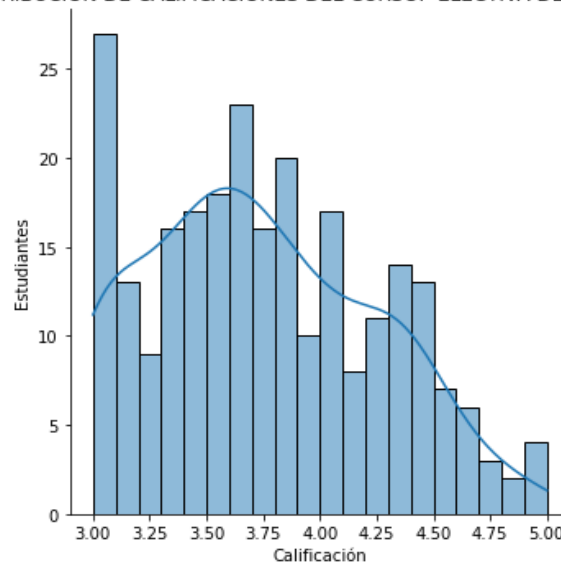


Ilustración 57 Distribución de calificaciones Electiva de ingeniería

Para electiva de ingeniería del software se obtuvo un promedio de 3,73 con un valor moda de 3,7 como nota máxima 5.0 y como nota minina 3.0 para un total de 354 notas en la muestra, con una desviación de 0,5.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ELECTIVA PROFESIONAL II

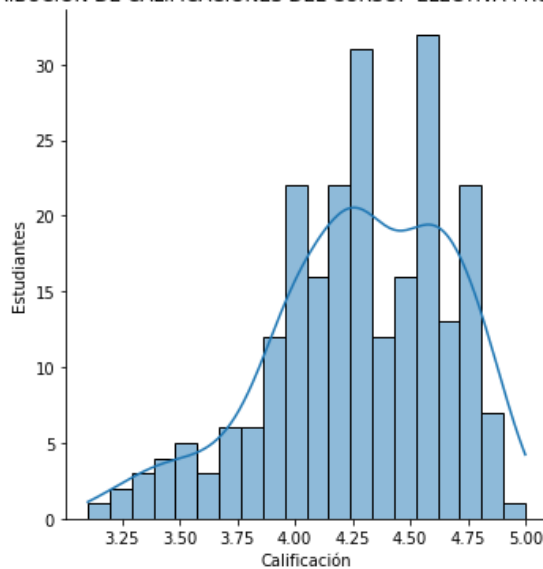


Ilustración 58 Distribución de calificaciones electiva profesional II

Para electiva de profesional II se obtuvo un promedio de 4.28 con un valor moda de 4,3 como nota máxima 5.0 y como nota minina 3.1 para un total de 236 notas en la muestra, con una desviación de 0,39.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: INGENIERÍA DE PROYECTOS I

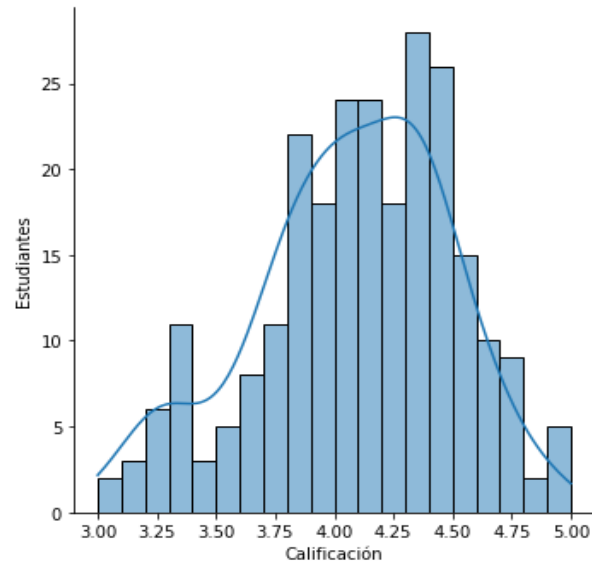


Ilustración 59 Distribución de calificaciones ingeniería de proyectos I

Para ingeniería de proyectos I se obtuvo un promedio de 4.06 con un valor moda de 4,1 como nota máxima 5.0 y como nota minina 3.0 para un total de 250 notas en la muestra, con una desviación de 0,416.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: INVESTIGACIÓN DE OPERACIONES

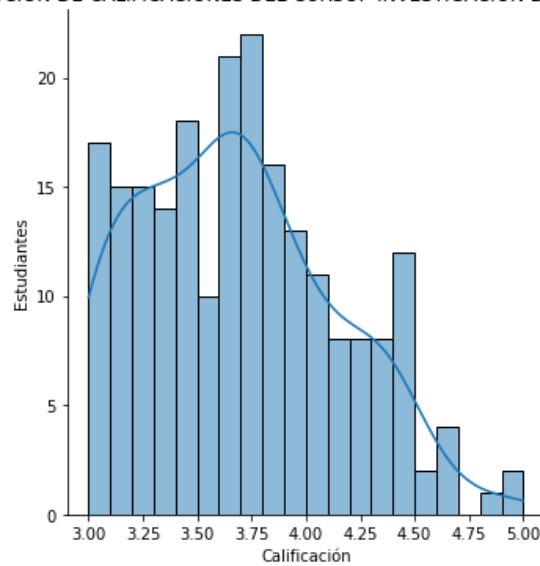


Ilustración 60 Distribución de calificaciones Investigación de operaciones

Para investigación de operaciones se obtuvo un promedio de 3,66 con un valor moda de 3,6 como nota máxima 5.0 y como nota minina 3.0 para un total de 217 notas en la muestra, con una desviación de 0,452.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: LEGISLACIÓN INFORMÁTICA

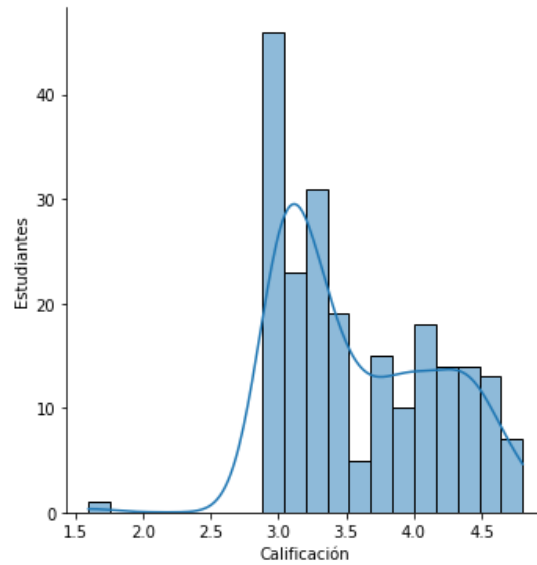


Ilustración 61 Distribución de calificaciones Legislación informática

Para Legislación Informática se obtuvo un promedio de 3,59 con un valor moda de 3,4 como nota máxima 4,8 y como nota minina 1,6 para un total de 216 notas en la muestra, con una desviación de 0,577.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: REDES DE COMPUTADORES Y SISTEMAS DISTRIBUIDOS II

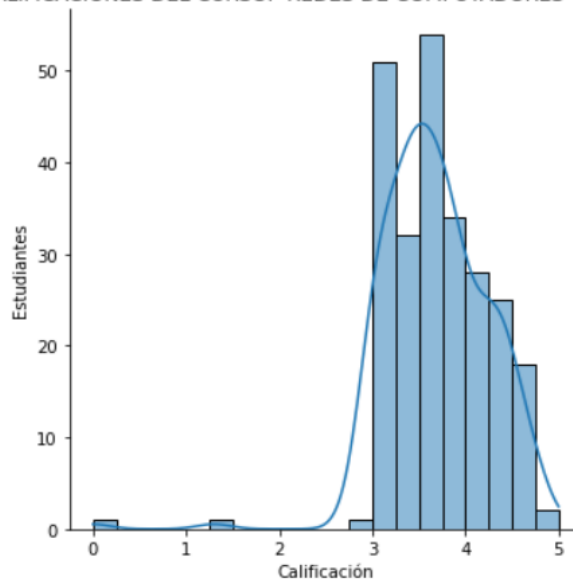


Ilustración 62 Distribución de calificaciones redes y sistemas distribuidos

Para redes de computadores y sistemas distribuidos II se obtuvo un promedio de 3,68 con un valor moda de 3,6 como nota máxima 5.0 y como nota minina 0.0 para un total de 247 notas en la muestra, con una desviación de 0,568.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ELECTIVA PROFESIONAL III

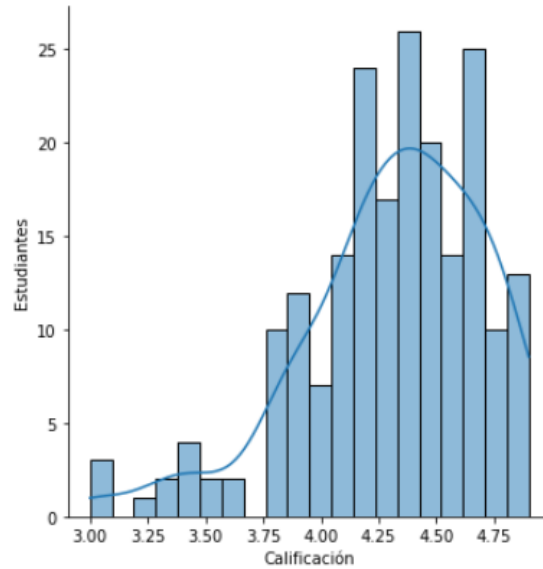


Ilustración 63 Distribución de calificaciones Electiva profesional III

Para redes de Electiva de ingeniería III se obtuvo un promedio de 3,87 con un valor moda de 3,8 como nota máxima 5.0 y como nota mínima 3.0 para un total de 218 notas en la muestra, con una desviación de 0,401.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ELECTIVA SOCIOHUMANISTICA II

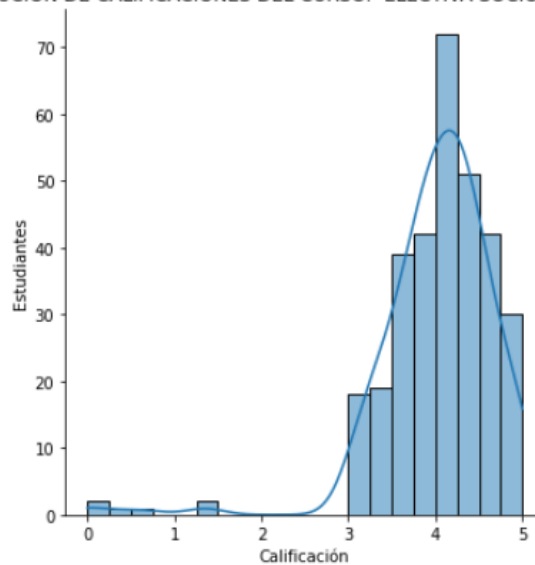


Ilustración 64 Distribución de calificaciones Electiva Socio humanística II

Para redes de Electiva de Socio humanística II se obtuvo un promedio de 4,01 con un valor moda de 4,1 como nota máxima 5.0 y como nota mínima 0.0 para un total de 319 notas en la muestra, con una desviación de 0,677.

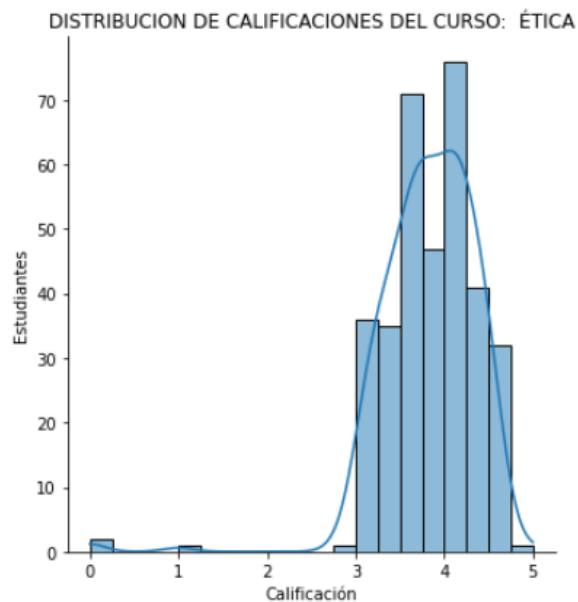


Ilustración 65 Distribución de calificaciones Ética

Para redes de Ética se obtuvo un promedio de 3,82 con un valor moda de 3.9 como nota máxima 5.0 y como nota minina 0.0 para un total de 343 notas en la muestra, con una desviación de 0,548.

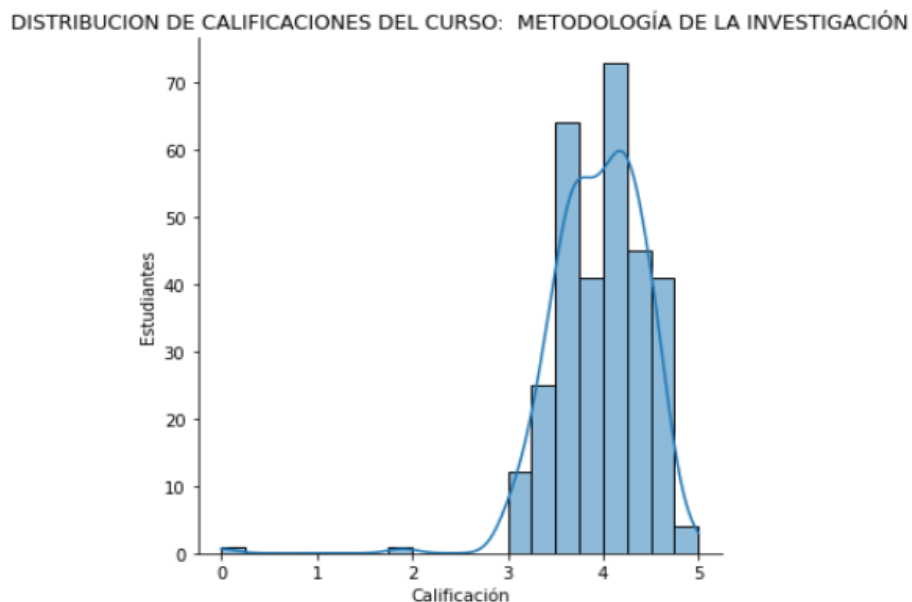


Ilustración 66 Distribución de calificaciones Metodología de la investigación

Para redes de Metodología de la Investigación se obtuvo un promedio de 3,95 con un valor moda de 4,0 como nota máxima 5.0 y como nota minina 0.0 para un total de 307 notas en la muestra, con una desviación de 0,492.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: PROTECCIÓN DE SISTEMAS INFORMATICOS

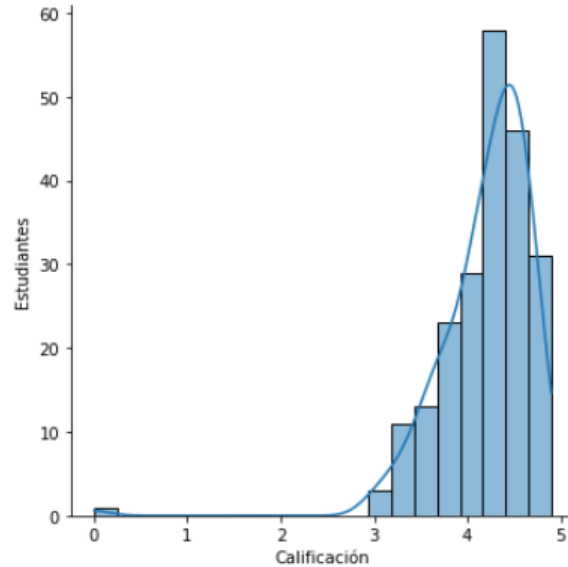


Ilustración 67 Protección de sistemas informativos

Para redes de protección de sistemas informáticos se obtuvo un promedio de 4,2 con un valor moda de 4,3 como nota máxima 4,9 y como nota minina 0.0 para un total de 215 notas en la muestra, con una desviación de 0,509.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: ACTIV DEPOR,RECRE Y CULT-DEP UNIVERSITARIO

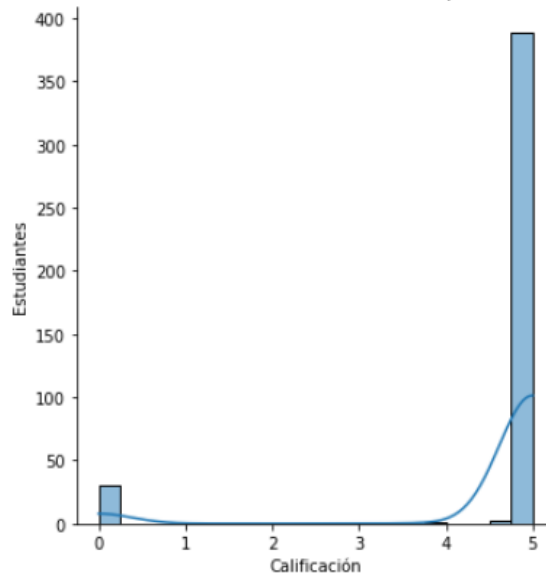


Ilustración 68 Distribución de calificaciones recreación y deportes

Para redes de protección de Actividad deportiva, recreación y cultura departamento universitario se obtuvo un promedio de 4,6 con un valor moda de 5 como nota máxima 5 y como nota minina 0.0 para un total de 422 notas en la muestra, con una desviación de 1,28.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: CÍVICA Y CONSTITUCIÓN (EXTRAPLAN)

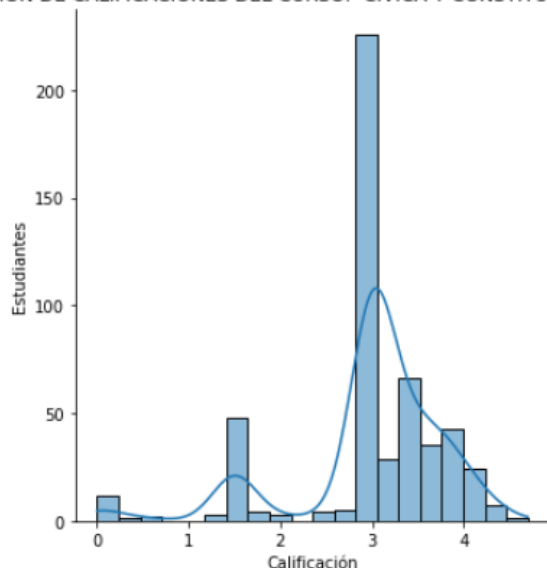


Ilustración 69 Distribución de calificaciones Cívica

Para cívica y constitución se obtuvo un promedio de 2,99 con un valor moda de 3.0 como nota máxima 4,7 y como nota minina 0.0 para un total de 513 notas en la muestra, con una desviación de 0,828.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: INFORMÁTICA BÁSICA (EXTRAPLAN)

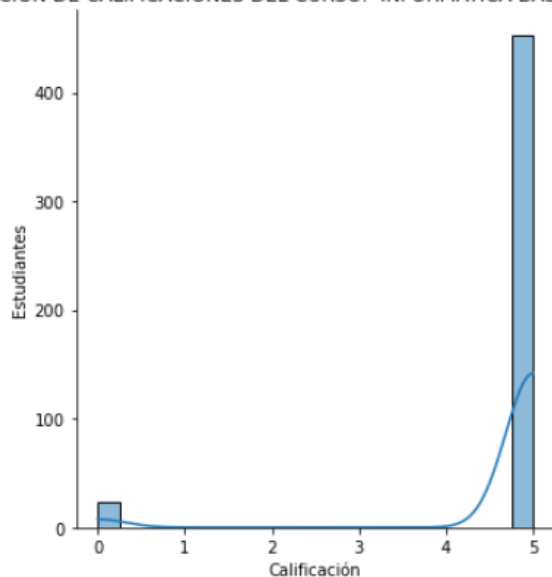


Ilustración 70 Distribución de calificaciones Informática Básica

Para redes de Metodología de la Investigación se obtuvo un promedio de 4.74 con un valor moda de 5,0 como nota máxima 5.0 y como nota minina 0.0 para un total de 477 notas en la muestra, con una desviación de 1.094.

DISTRIBUCION DE CALIFICACIONES DEL CURSO: TRABAJO DE GRADO

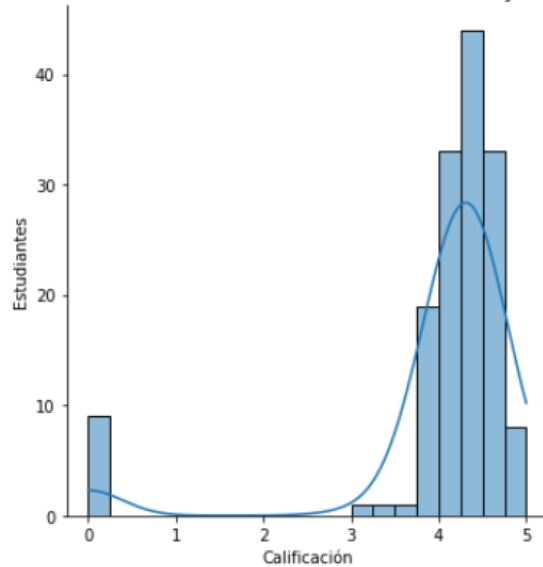


Ilustración 71 Distribución de calificaciones Trabajo Grado

Para trabajo de grado se obtuvo un promedio de 4.03 con un valor moda de 4,3 como nota máxima 5.0 y como nota minina 0.0 para un total de 149 notas en la muestra, con una desviación de 1.07

Respecto al pensum del 2006 las materias se clasifican en cuatro grandes áreas formación básica, socio humanística, profesional y profundización, para cada área se realizó la gráfica de cajas y bigotes, en la cual se creó basado en este código con el cual extraemos las notas relacionas a estas materias básicas

156001 BIOLOGÍA GENERAL
 157005 CÁLCULO DIFERENCIAL
 156006 LABORATORIO DE QUÍMICA GENERAL
 157019 MECÁNICA
 156009 QUÍMICA GENERAL
 157002 ÁLGEBRA LINEAL
 157007 CÁLCULO MULTIVARIABLE
 157009 ELECTROMAGNETISMO
 157008 ECUACIONES DIFERENCIALES
 157020 OSCILACIONES Y ONDAS

En cual pasamos como parámetro un vector con los códigos de las materias a analizar para este caso tenemos diez. Obteniendo la siguiente gráfica, en cual podemos observar que la mediana de biología general y calculo integral es igual a 3 tienen una asimetría negativa y manejan valores atípicos inferiores a 0,7 y valores atípicos superiores a 4,7. Para Laboratorio de química los datos están concentrados desde 2,2 hasta 5, tienen un valor medio de 3,7 y valores atípicos menores desde cero a 2. Para mecánica tenemos mayor cantidad de valores atípicos por fuera de los Q1 desde cero a 2,5 y Q3 por encima de cuatro hasta cinco a pesar de que el bigote superior tiene un valor máximo de 4 y su valor en el bigote inferior es 2,5.

Las materias como biología general, calculo diferencial, laboratorio de química general presentan una asimetría negativa donde se observa que la media es menor a la media y la moda.

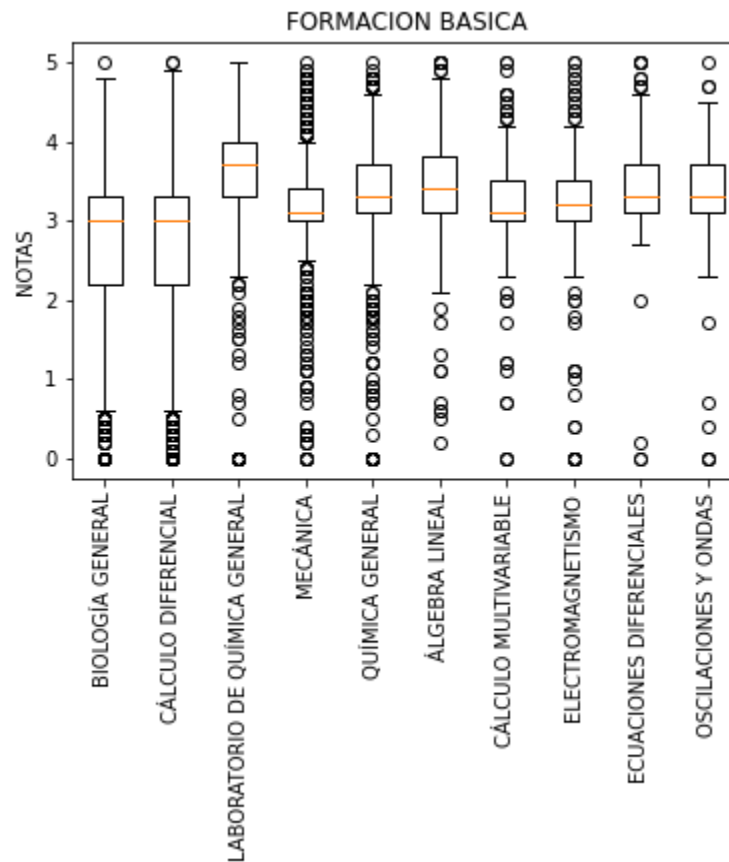


Ilustración 72 Formación básica

A comparación de mecánica, química general, oscilaciones y ondas, algebra lineal, calculo multivariable, electromagnetismo, y ecuaciones diferenciales presentan una dispersión asimétrica positiva, adicional todas presentan valores atípicos inferiores a dos.

Basada en el mismo código generamos la grafica para las asignaturas socio humanísticas para el programa de ingeniería de sistemas son las siguientes ocho relaciones.

162003 HABILIDADES COMUNICATIVAS
 153002 CÁTEDRA FARÍA
 164004 EDUCACIÓN AMBIENTAL
 369006 CÍVICA Y CONSTITUCIÓN (EXTRAPLAN)
 173238 ACTIV DEPOR, RECRE Y CULT-DEP UNIVERSITARIO
 150001 ELECTIVA SOCIOHUMANÍSTICA I
 164010 ÉTICA
 150002 ELECTIVA SOCIOHUMANISTICA II

Las notas para dichas materias se observan en la siguiente grafica.

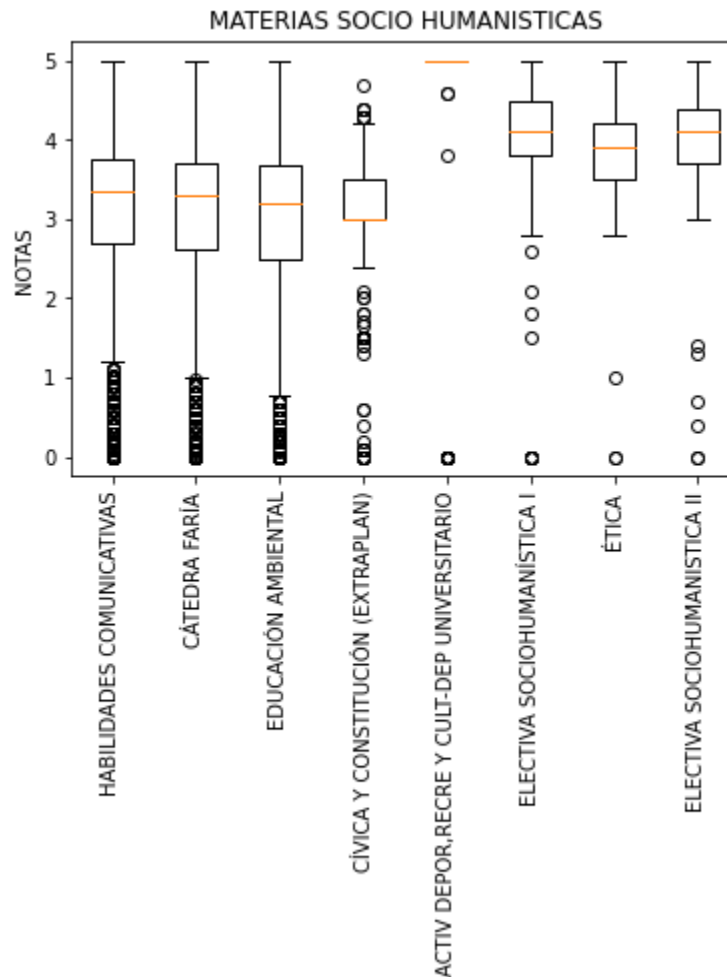


Ilustración 73 Materias Sociohumanisticas

Para habilidades comunicativas, catedra faria y educación ambiental observamos un comportamiento parecido donde la media este alrededor de 3,3 y solo tenemos valores atípicos bajos a 1,2 presenta una dispersión asimétrica negativa; para cívica encontramos que la media y el Q1 tienen el mismo valor dado en 3, presenta valores atípicos inferiores a su Q1; la electiva socio humanista I presenta una dispersión asimétrica positiva.

A diferencia de las demás materias deportes, recreación y cultura tiene un comportamiento super atípico diferente con valor igual a 5 ya que este es el valor de la moda, la mediana y la media.

167250 PROGRAMACIÓN ESTRUCTURADA
167251 PROGRAMACIÓN ORIENTADA A OBJETOS
167231 ESTRUCTURAS DE DATOS DINÁMICOS Y ALGORITMOS
167238 INTRODUCCIÓN A LAS CIENCIAS COMPUTACIONALES
167201 ANÁLISIS DE ALGORITMOS
167243 LÓGICA COMPUTACIONAL
167233 FUNDAMENTOS Y DISEÑO DE SISTEMAS OPERATIVOS
167232 FUNDAMENTOS Y DISEÑO DE BASE DE DATOS
167248 PARADIGMAS DE PROGRAMACIÓN
167246 MODELAMIENTO DE SISTEMAS DE INFORMACIÓN
167202 BASES DE DATOS II
167270 TEORÍA DE LA COMPUTACIÓN
167256 REDES DE COMPUTADORES Y SISTEMAS DISTRIBUIDOS
167271 TÓPICOS AVANZADOS DE INGENIERÍA DEL SOFTWARE
167213 ELECTIVA DE INGENIERÍA
167266 SISTEMAS Y MODELADO II
167257 REDES DE COMPUTADORES Y SISTEMAS DISTRIBUIDOS II
167240 LEGISLACIÓN INFORMÁTICA
167214 ELECTIVA DE INGENIERÍA
167252 PROTECCIÓN DE SISTEMAS INFORMATICOS
167215 ELECTIVA DE INGENIERÍA III

Para las materias del área profesional son veintiún mencionadas anteriormente y graficadas a continuación.

En la gráfica de materias de formación profesional observamos con dispersión asimétrica negativa las siguientes asignaturas: programación estructurada, introducción a las ciencias computacionales, análisis de algoritmos, protección de sistemas de informáticos, fundamentos y diseños de bases de datos, teoría de la computación, electiva de ingeniería y protección de sistemas informáticos de los cuales PE tienen gran cantidad de valores atípicos leves inferiores a 0,7 respectivamente a diferencia de los demás que presentan muy pocos.

Programación orientada a objetos POO, estructuras de datos dinámicos y algoritmos, fundamentos y diseños de SO, modelamientos de sistemas de información, bases de datos II, tópicos avanzados de la ingeniería del software, sistemas y modelados II, redes de computadores y sistemas distribuidos II, legislación informática, y electiva de ingeniería III presentan una dispersión atípica positiva de las cuales POO es la que más presenta valores atípicos inferiores a 1,7.

En grafica esta encontramos que lógica computacional, y paradigmas de programación son simétricas y presentan uno o dos valores atípicos.

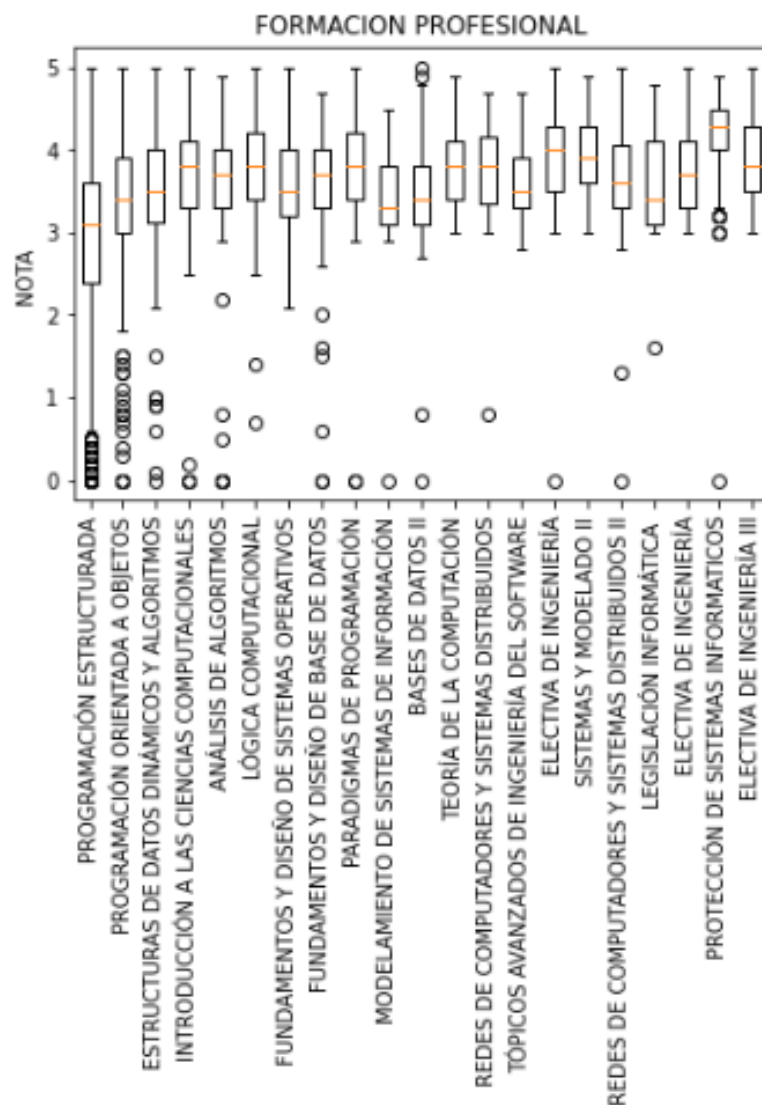


Ilustración 74 Formación Profesional

A continuación, realizaremos el mismo ejercicio para las materias de formación de profundización. En el cual son cuatro materias las 3 electivas y trabajo de grado.

167219 ELECTIVA PROFESIONAL
 167222 ELECTIVA PROFESIONAL II
 167225 ELECTIVA PROFESIONAL III
 167328 TRABAJO DE GRADO

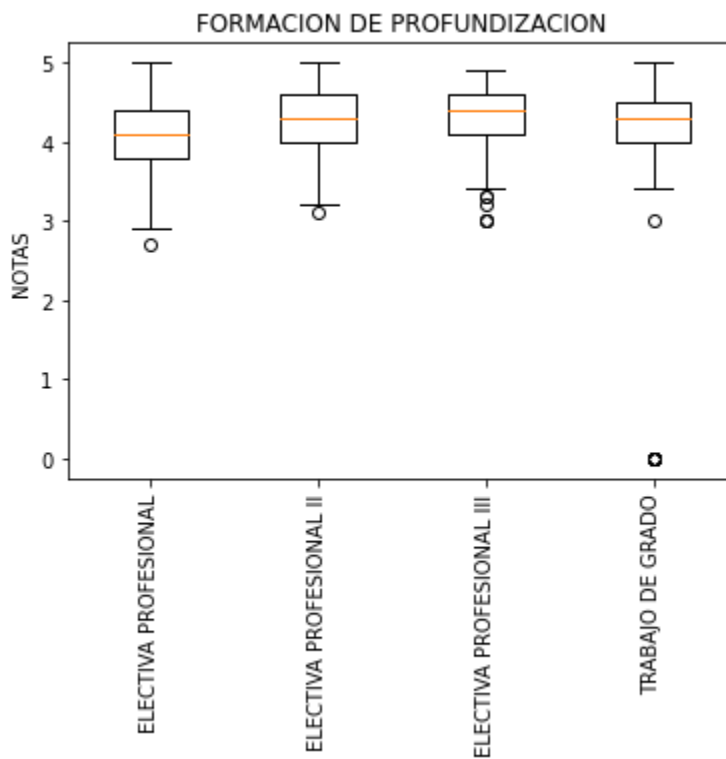


Ilustración 75 Formación de Profundización

En esta grafica observamos que electiva profesional I y II tiene una distribución simétrica con vales media superior a cuatro y un valor a típico leve al redor de 3.

Para la electiva profesional III y trabajo de grado presentan una distribución negativa dado que el valor de la moda es mayor a la mediana y la media respectivamente.

En esa parte se realizará graficas comparativas en las notas obtenidas por semestres,

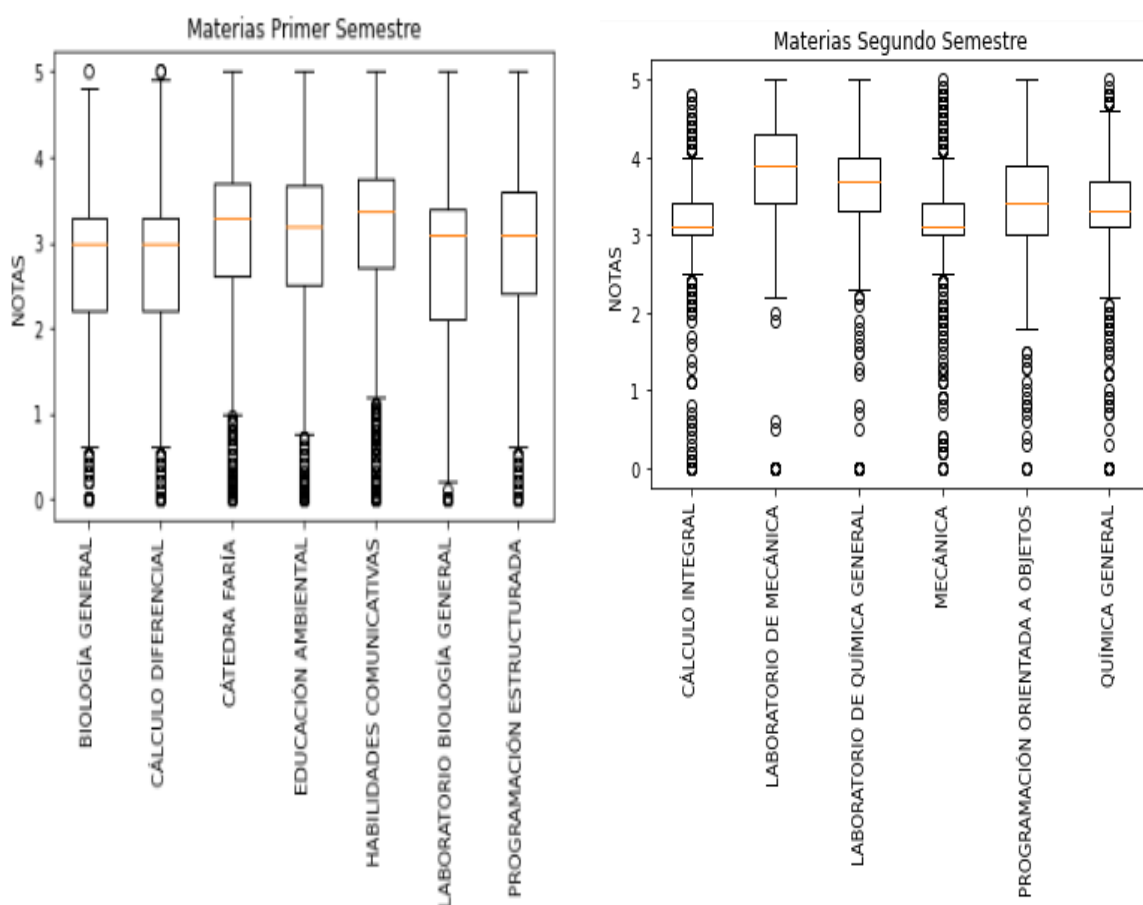


Ilustración 76 Materias primer semestre y Materias Segundo Semestre

Observamos que en primer semestre se presentan una distribución asimétrica negativa en todas las materias dado que la media está más cerca del cuartil tres las notas están concentradas en un intervalo de 2, 2 a 3,2, donde se caracteriza por que la moda es mayor a la media y la mediana, se observa que, aunque los rangos de notas con amplios te encuentran valores atípicos para las notas más bajas, para este semestre la el promedio esta sobre 3.

En esta grafica las materias del segundo semestre con notas con asimetría positiva como calculo integral, Mecánica y química general y programación Orienta a Objetos y dispersión asimetría negativa Laboratorio de mecánica y laboratorio de química.

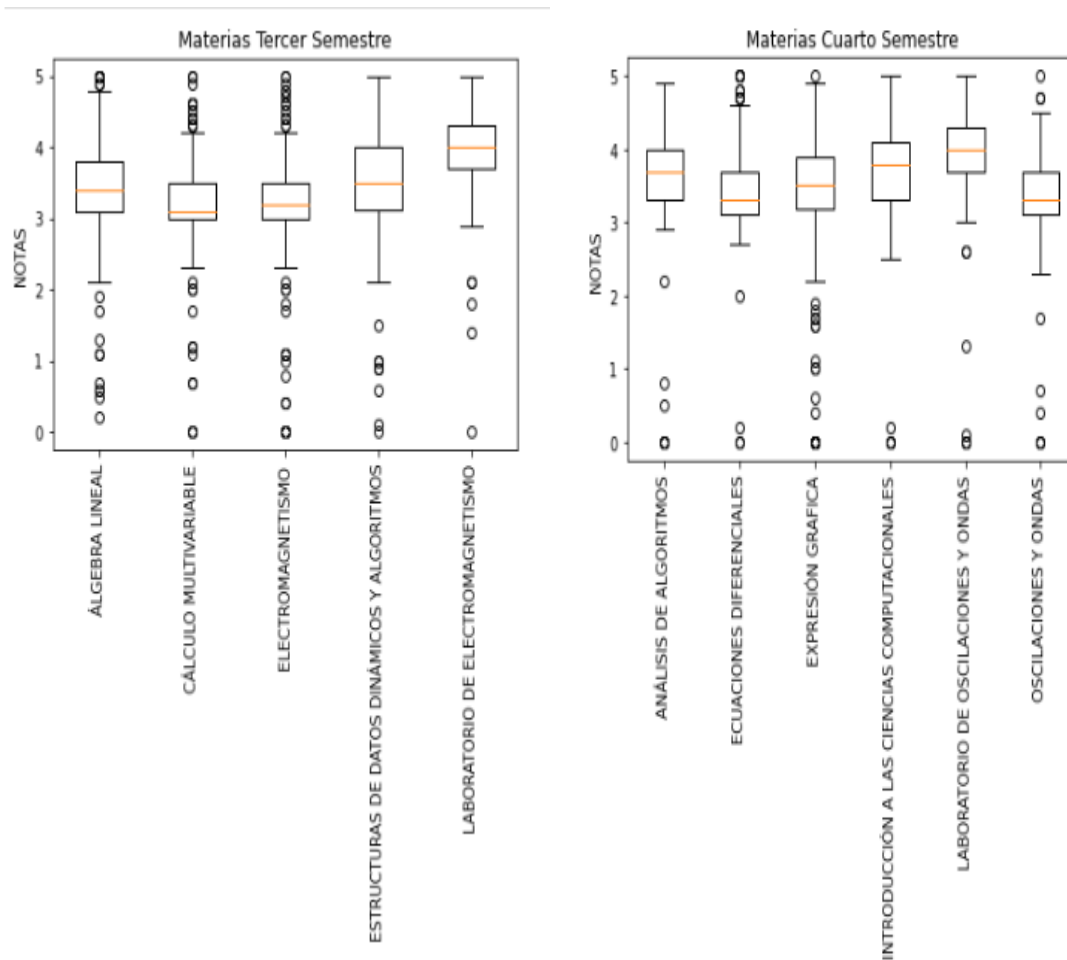


Ilustración 77 Materias Tercer Semestre y Materias Cuarto Semestre

En el tercer semestre encontramos que en las materias algebra lineal, calculo multivariable, electromagnetismo y Estructuras de datos y algoritmos presentan una dispersión asimétrica positiva, y laboratorio de electromagnetismo presenta dispersión simétrica, en la siguiente grafica.

Para el cuarto semestre observamos que todas las materias presentan valores atípicos y se observa que la media esta sobre 3, encontramos que análisis de algoritmos, Introducción a las ciencias computacionales son asimétricas negativas y asimétricas positivas para ecuaciones diferenciales y oscilaciones y ondas, en dispersión simétrica observamos laboratorio de oscilación y ondas y expresión grafica.

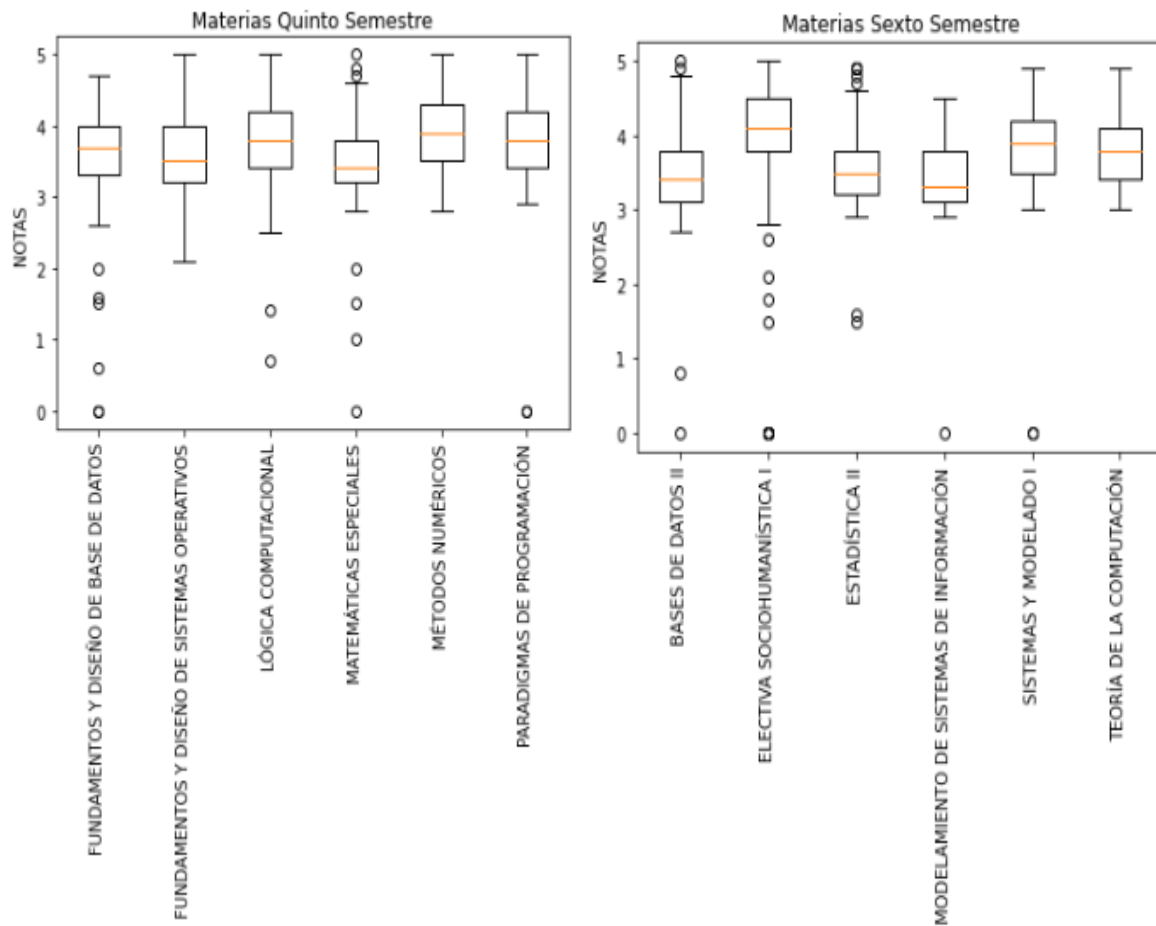


Ilustración 78 Materias Quinto Semestre y Materias Sexto Semestre

En el quinto semestre que las materias matemáticas especiales, lógica computacional y paradigmas se tiene una dispersión simétrica, en dispersión asimétrica negativa fundamentos y diseño de datos, dispersión asimétrica positiva fundamentos y se diseño de sistemas operativos

Para el sexto semestre observamos concentración de datos sobre una nota de 2,8 aunque presentan algunos datos atípicos tanto en bases de datos II y estadística II superiores alrededor de la nota máxima 5.

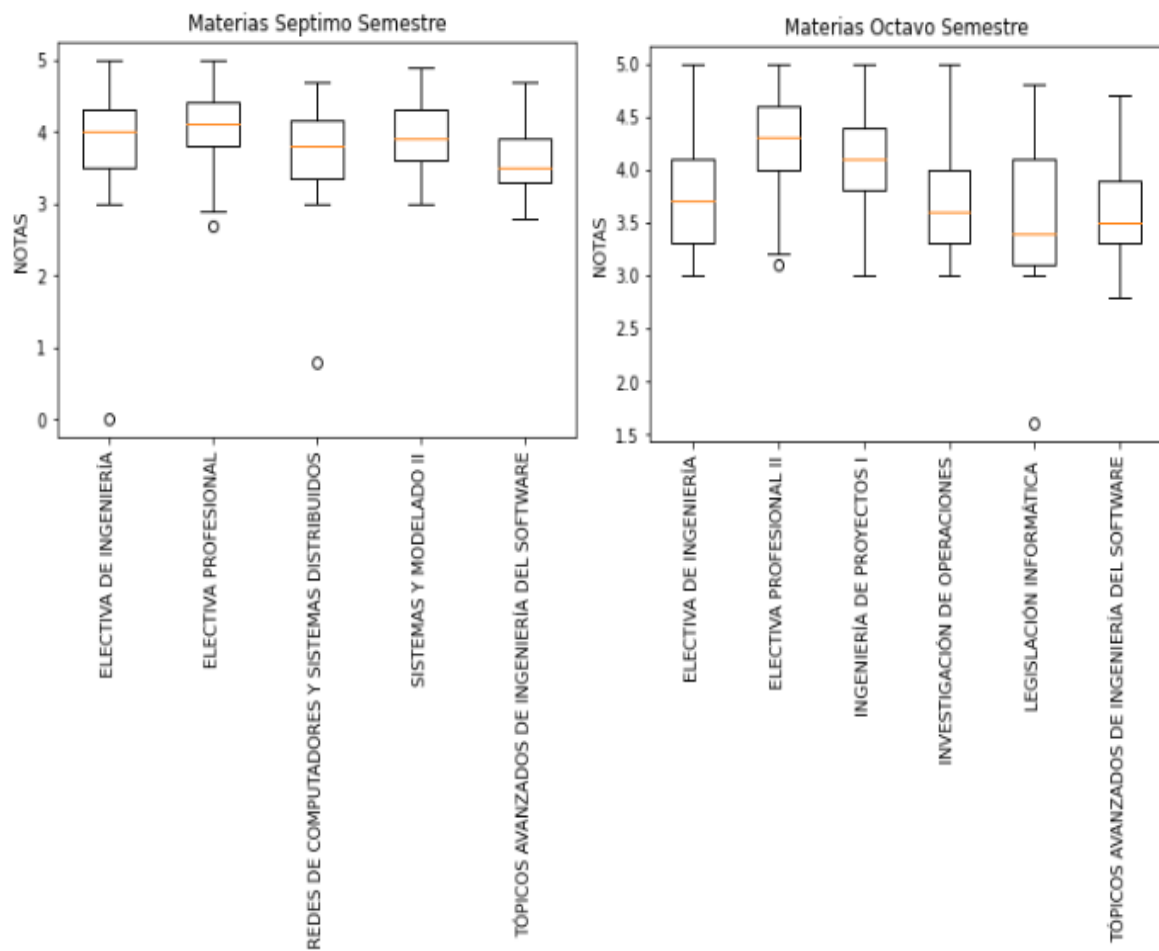


Ilustración 79 Materias Séptimo Semestre y Materias Octavo Semestre

Para la graficas de las materias para el semestre séptimo y octavo se observa con claridad que las notas se encuentran concentradas en el rango de 3 a 5 y las notas atípicas con muy pocas.

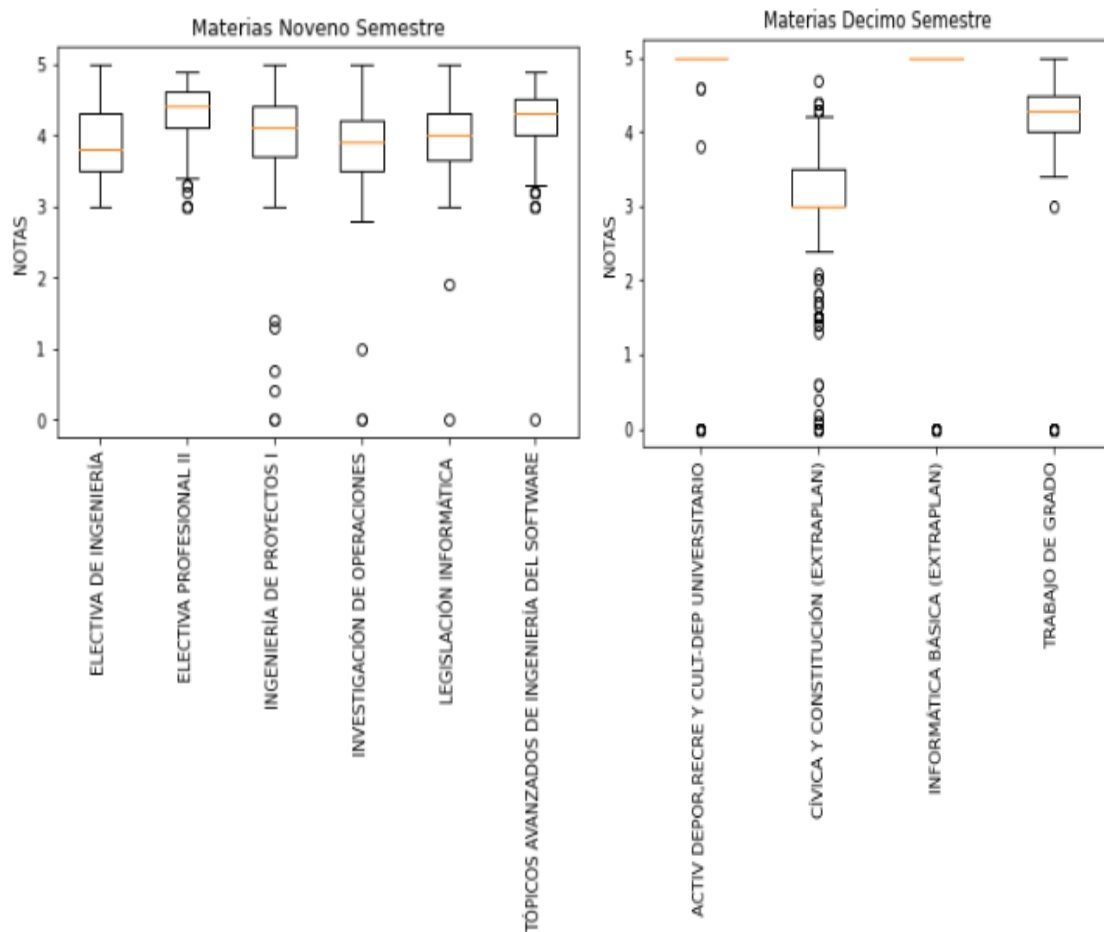


Ilustración 80 Materias Noveno Semestre y Materias Decimo Semestre

En esta grafica donde observamos que la electiva profesional y tópicos avanzados de ingeniería del software tienen concentradas las notas junto a sus valores atípicos sobre 3, a esta altura de estos semestres en las materias de pregrado están con muy pocas notas atípicas y la mayor cantidad de asignaturas aprobadas.

Para la gráfica del décimo semestre observamos que la materia de Extraplán cívica tiene gran variedad de valores atípicos por el límite inferior y superior. Para deporte recreación y Cultura e informática observamos que la nota que prevalece el 5. En trabajo de grado se observan datos atípicos, aunque su mayor cantidad de datos están concentrados sobre 3,2

3.4. Análisis predictivo

En este apartado se realiza un análisis que intenta determinar la capacidad de predicción que pueda tener un modelo de aprendizaje automático, para establecer los resultados obtenidos en los diferentes cursos del plan de estudios, teniendo como información previa, los resultados obtenidos en las pruebas ICFES.

DESCRIPCION DE LOS DATOS:

RECOPIACION E INTEGRACION DE DATOS

Inicialmente se cuenta con 24 libros de Excel, correspondientes a los resultados de las pruebas ICFES de los periodos 2006-2 hasta 2018-1, cada uno de los archivos contienen la identificación de los aspirantes, junto a los resultados obtenidos en las áreas de Lenguaje, Matemáticas, Ciencias Sociales, Filosofía, Biología, Química, Física e Idioma extranjero.

Por otro lado, se tiene un archivo con el reporte de notas de los estudiantes matriculados en el pensum 2006, hasta el periodo 2018, con un total de 1763 Registros, consignados en una sola hoja, con las calificaciones obtenidas en 58 materias. Incluido el caso en el que se han repetido asignaturas.

SELECCIÓN, LIMPIEZA Y TRANSFORMACION

Los datos relacionados con las pruebas ICFES, se recopilaron en un único DataFrame de Pandas, se eliminó a columna correspondiente a Ciencias Sociales, debido a que presentaba una gran cantidad de datos faltantes. Conformada la tabla se eliminan las filas con datos nulos o inválidos, con lo cual se obtiene una tabla de dimensiones 1948 filas por 8 columnas.

En el caso del archivo con las calificaciones del pensum 2006, se cargan los datos en un DataFrame de Pandas, seleccionando el identificador de cada estudiante y las columnas correspondientes a las calificaciones por materia, con lo cual, se obtiene una tabla de 1763 filas x 59 columnas.

Finalmente, para poder realizar el análisis predictivo, se realiza una mezcla de las tablas de ICFES y de notas en materias, seleccionando los registros con identificador común en ambas tablas y conformando una única tabla donde aparece el identificador del estudiante, los resultados de las pruebas ICFES y las calificaciones obtenidas, consiguiendo así, una tabla de 1211 filas x 66 columnas. En este caso se tratará de predecir la primera calificación obtenida, por los estudiantes en cada materia. Eso, para aclarar el caso de los estudiantes que han repetido asignaturas.

EXPERIMENTACION CON MODELOS PREDICTIVOS

MAQUINAS DE APRENDIZAJE PARA REGRESION.

En un primer experimento se selecciona un conjunto de máquinas de aprendizaje para problemas de regresión, que será aplicado a cada una de las materias del pensum 2006. Para la selección se realizan algunas pruebas con algunos de los algoritmos implementados en ScikitLearn, y se establecen las que obtienen mejores resultados.

Las máquinas seleccionadas finalmente son:

- Máquina de Vectores de Soporte para regresión: LinearSVR
- Árboles de decisión para regresión: DecisionTreeRegressor
- Vecino más cercano para regresión: KNeighborsRegressor
- Perceptrón multicapa para regresión: MLPRegressor

Para cada una de las asignaturas del plan de estudios se realiza el siguiente procedimiento:

- Se realiza un filtrado del Dataframe materias/ICFES, para obtener una nueva tabla con cada asignatura donde se tenga, identificador del estudiante, el resultado de las pruebas ICFES y la calificación obtenida en cada materia. La tabla siguiente ilustra parte del resultado para el curso Programación Estructurada, con código 167250:

	<i>Documento</i>	<i>Lenguaje</i>	<i>Matemáticas</i>	<i>Biología</i>	<i>Química</i>	<i>Física</i>	<i>167250</i>
2	68297165	51.00	45.00	54.00	34.00	48.00	3.2
4	89052771987	44.61	55.73	51.63	52.03	58.70	1
5	1014182764	52.00	42.00	53.00	54.00	51.00	4.6
8	1098615945	51.60	41.46	45.79	40.87	44.61	1.1
9	1115851931	57.73	36.11	45.79	47.16	50.36	2.8

Tabla 7 Resultado ICFES vs Programación estructurada

En este caso se tiene una tabla de 868 filas x 7 columnas.

- Con las “subtablas” generadas para cada materia, se realiza una prueba de validación cruzada, en la cual se divide la tabla en 10 conjuntos de entrada/salida, donde las entradas son las puntuaciones obtenidas en las pruebas ICFES, y las salidas, la calificación obtenida en la materia. En 10 iteraciones se toman 9 conjuntos para entrenar el clasificador y 1 para probarlo, y se calcula el error de clasificación, que en este caso es la raíz del error cuadrático medio (RMSE) que está dado por la siguiente expresión:

$$RMSE = \frac{1}{2} \sum_{i=1}^m \sqrt{(Y_{predict}^i - Y_{prueba}^i)^2}$$

donde $Y_{predict}$, es la calificación estimada por la máquina de aprendizaje, Y_{prueba} , es la calificación registrada en el conjunto de prueba de la subtabla y m , el cardinal del conjunto de prueba.

- La validación cruzada se realiza para estimar todas las asignaturas del pensum 2006, con las máquinas de aprendizaje seleccionadas.

Inicialmente se realizan pruebas individuales con algunos de los cursos de primeros semestres, que son los que hipotéticamente están más cerca de ser influenciados por los resultados obtenidos en las pruebas ICFES. En la tabla siguiente se muestran los resultados obtenidos para el curso de **Programación estructurada**:

MATERIA	Modelo	MSE	std	R_2	RMSE	RMSE_porcent
167250	MPL	1.22	0.09	0.18	1.12	22.52
167250	linearSVM	1.26	0.19	0.16	1.13	22.74
167250	RegressorTree	1.52	0.11	-0.46	1.51	30.20
167250	KNRegressor	1.26	0.09	-0.06	1.28	25.76

Tabla 8 Resultados de Programación estructurada Vs Métodos de Predicción

En la tabla, *MSE*, corresponde a la puntuación dada por el error cuadrático medio, *std*, la desviación estándar del error en los conjuntos de validación, *R_2*, el coeficiente de determinación, que explica que tan bien ajusta el modelo a los datos, *RMSE*, la raíz del error cuadrático medio, y *RMSE_porcent*, indica a que valor porcentual de la calificación, corresponde la raíz del error. Este más fácil de interpretar ya que la raíz del error cuadrático medio está dada en la misma escala de la calificación, y su valor porcentual no da un indicador del margen de error en la predicción. Por tanto, para el

caso que se muestra en la tabla anterior, los márgenes de error son muy altos, en el mejor de los casos, el MLP (perceptrón multicapa) predice la calificación obtenida por un estudiante de programación estructurada con un 22.52% de error, es decir, si alguien obtiene 3.5 como definitiva en el curso, el MLP puede predecir un valor entre 2.38 y 4.77, siendo un margen de error muy amplio.

En la tabla siguiente se ven los resultados para el curso **Calculo Diferencial**:

MATERIA	Modelo	MSE	std	R_2	RMSE	RMSE_porcent
157005	MPL	1.17	0.10	0.02	1.11	22.29
157005	linearSVM	1.17	0.16	-0.00	1.1	22.67
157005	RegressorTree	1.50	0.14	-0.57	1.418	28.36
157005	KNRegressor	1.21	0.09	-0.19	1.23	24.68

Tabla 9 Resultados Calculo diferencial Vs Métodos de Predicción

En la tabla se aprecia un margen de error similar al obtenido en Programación Estructurada.

La tabla siguiente ilustra los resultados en el caso de **Biología General**:

MATERIA	Modelo	MSE	std	R_2	RMSE	RMSE_porcent
156001	MPL	1.08	0.04	0.04	1.01	20.26
156001	linearSVM	1.11	0.05	0.04	1.01	20.31
156001	RegressorTree	1.48	0.11	-0.35	1.20	24.17
156001	KNRegressor	1.17	0.08	-0.06	1.07	21.42

Tabla 10 Resultado Biología general Vs Métodos de Predicción

El resultado de las predicciones para el curso de **Cátedra Faría** lo encontramos en la siguiente tabla:

MATERIA	Modelo	MSE	std	R_2	RMSE	RMSE_porcent
153002	MPL	1.234	0.066	-0.065	1.320	26.401
153002	linearSVM	1.291	0.089	-0.685	1.659	33.199
153002	RegressorTree	1.598	0.125	-0.809	1.720	34.407
153002	KNRegressor	1.309	0.088	-0.022	1.293	25.861

Tabla 11 Resultado Catedra Faria Vs Métodos de Predicción

Para el curso de **Educación Ambiental** se obtienen los siguientes resultados:

MATERIA	Modelo	MSE	std	R_2	RMSE	RMSE_porcent
164004	MPL	1.322	0.113	0.031	1.331	26.625
164004	linearSVM	1.493	0.304	-0.004	1.356	27.122
164004	RegressorTree	1.826	0.142	-0.750	1.789	35.792
164004	KNRegressor	1.401	0.112	-0.161	1.457	29.155

Tabla 12 Resultado Educación ambiental Vs Métodos de Predicción

Y finalmente se muestra en la siguiente tabla, los resultados obtenidos para el curso de **Habilidades Comunicativas**:

MATERIA	Modelo	MSE	std	R_2	RMSE	RMSE_porcent
162003	MPL	1.327	0.068	0.004	1.238	24.764
162003	linearSVM	1.485	0.252	-0.084	1.292	25.851
162003	RegressorTree	1.770	0.164	-0.923	1.721	34.426
162003	KNRegressor	1.450	0.067	-0.105	1.304	26.096

Tabla 13 Resultado Habilidades comunicativas Vs Métodos de Predicción

La siguiente tabla muestra los resultados obtenidos por las diferentes máquinas de aprendizaje en las cuales se obtiene un error menor o igual al 10%. Es decir, aquellas predicciones en las cuales el error de la calificación varía en un intervalo de ± 0.5 puntos.

	MATERIA	Modelo	MSE	std	R_2	RMSE	RMSE_porcent
2	167232	RegressorTree	0.572434	0.128049	0.297704	0.328449	6.568982
3	164018	KNRegressor	0.353479	0.068790	-0.362439	0.333866	6.677325
0	167222	MPL	0.599997	0.246146	-0.110357	0.355033	7.100654
1	167257	linearSVM	0.654465	0.139952	0.255826	0.382952	7.659031
3	167246	KNRegressor	0.449174	0.120469	0.044710	0.400048	8.000951
3	157012	KNRegressor	0.539273	0.295032	0.356200	0.418927	8.378543
3	167213	KNRegressor	0.518091	0.091482	-0.589812	0.432718	8.654351
1	167225	linearSVM	0.543347	0.149071	-0.212353	0.446031	8.920624
1	167271	linearSVM	0.418043	0.107487	-0.096788	0.446729	8.934577
3	157102	KNRegressor	0.523352	0.365092	-0.095816	0.480539	9.610778

Tabla 14 Resultados maquinas aprendizaje Error ± 0.5 puntos

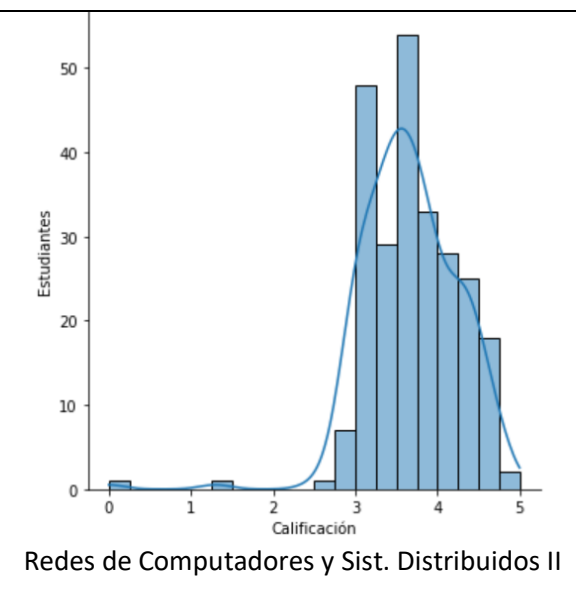
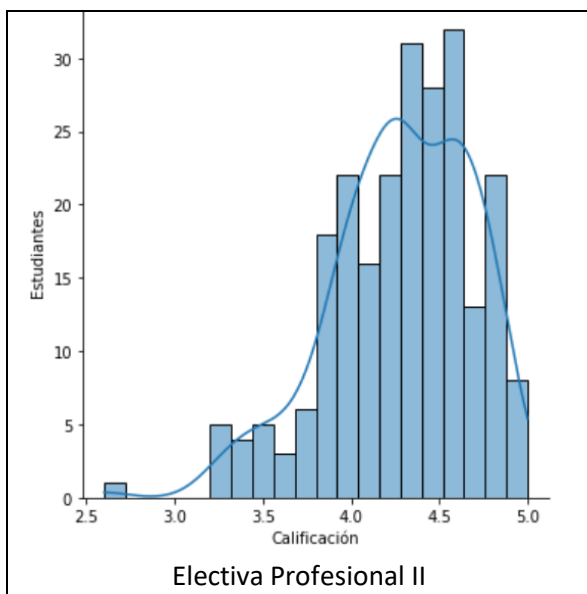
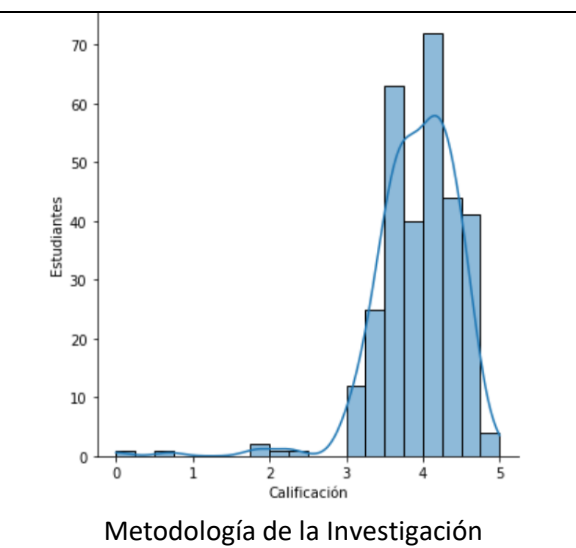
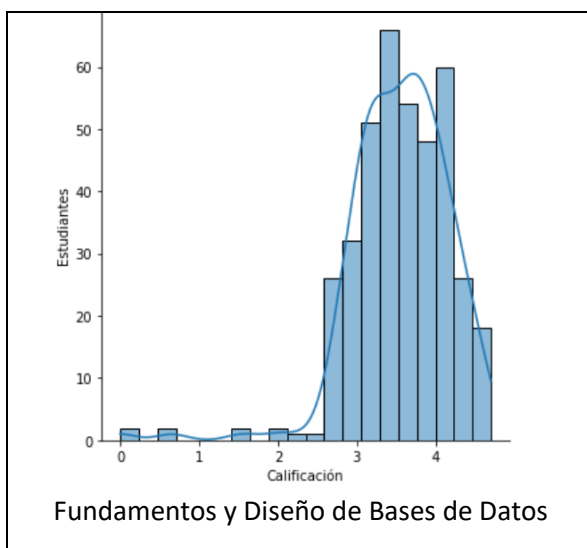
En ese sentido, las asignaturas que una vez aplicada la batería de máquinas de regresión obtiene un error porcentual inferior al 10% son las siguientes:

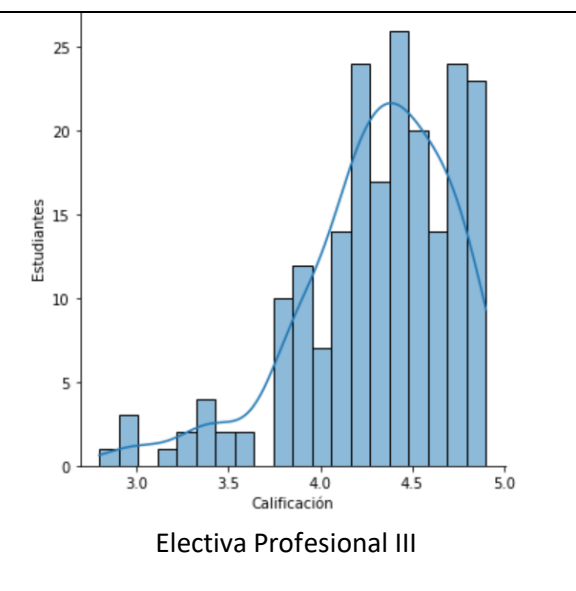
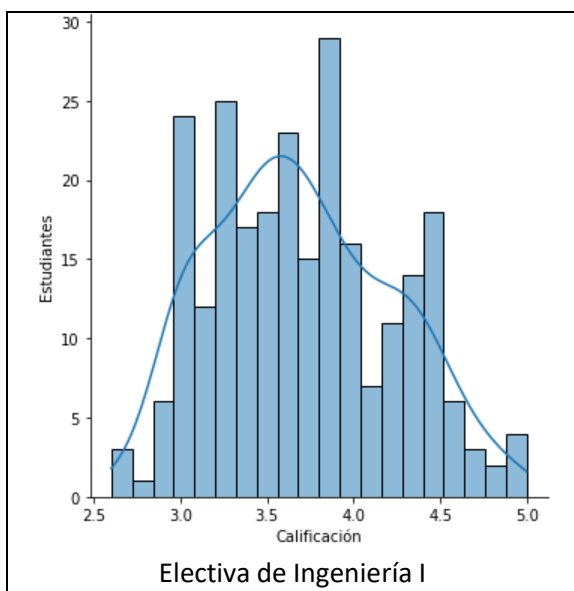
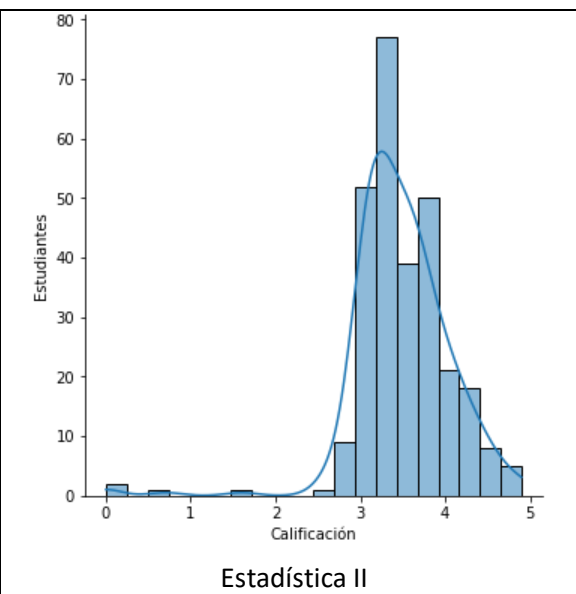
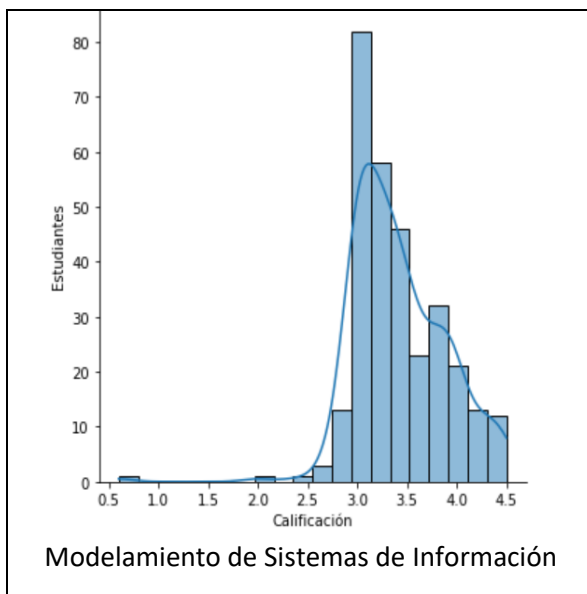
Asignatura	Error en la predicción	Porcentaje de error	Nota mínima	Nota máxima	Promedio
<i>FUNDAMENTOS Y DISEÑO DE BASE DE DATOS</i>	0.33	6.57	0	4.70	3.55
<i>METODOLOGÍA DE LA INVESTIGACIÓN</i>	0.33	6.68	0	5.0	3.92
<i>ELECTIVA PROFESIONAL II</i>	0.36	7.10	2.6	5.0	4.27
<i>REDES DE COMPUTADORES Y SISTEMAS DISTRIBUIDOS II</i>	0.38	7.66			
<i>MODELAMIENTO DE SISTEMAS DE INFORMACIÓN</i>	0.40	8.00	0.60	4.50	3.42
<i>ESTADÍSTICA II</i>	0.42	8.39	0	4.90	3.48

ELECTIVA DE INGENIERÍA I
ELECTIVA PROFESIONAL III
TÓPICOS AVANZADOS DE
INGENIERÍA DEL SOFTWARE
MATEMÁTICAS ESPECIALES

0.43	8.65	2.6	5.0	3.69
0.45	8.92			
0.45	8.93	2.4	4.7	3.51
0.49	9.89	0	5.0	3.44

Tabla 15 Notas obtenidas según resultado de error ± 0.5





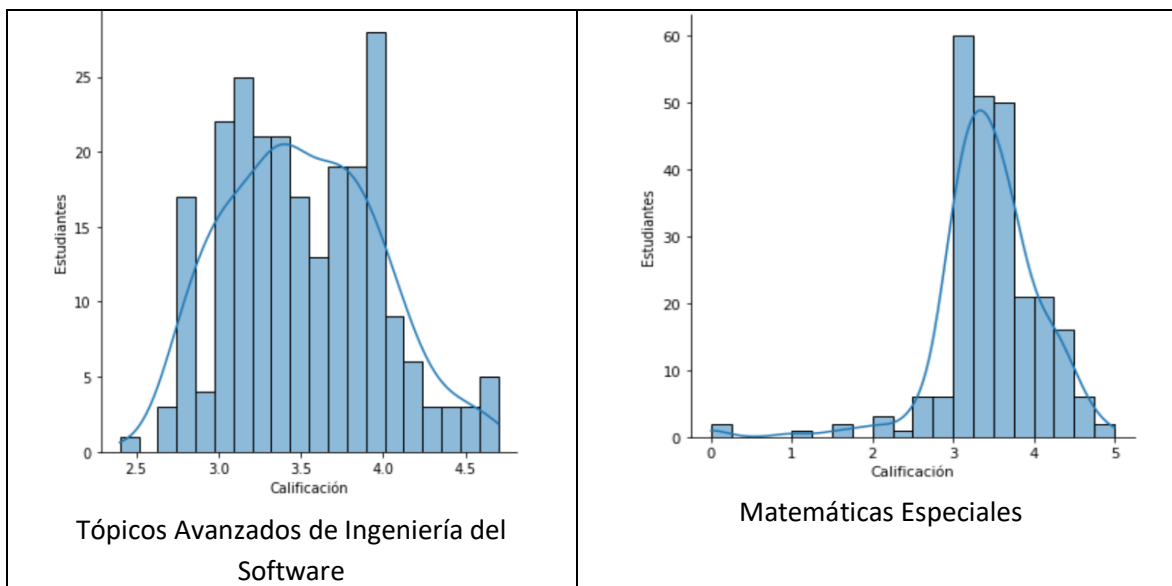


Ilustración 81 Notas obtenidas en al usar las maquinas de regresión. Para las siguientes materias fundamentos y diseño de base de datos, metodología de la investigación, electiva profesional II, redes de computadores y sistemas distribuidos II, modelamiento de sistemas de información, estadística II, electiva de ingeniería i, electiva profesional III, tópicos avanzados de ingeniería del software, matemáticas especiales

MAQUINAS DE APRENDIZAJE PARA CLASIFICACION.

En esta segunda etapa del análisis predictivo, las calificaciones obtenidas por los estudiantes matriculados en el pensum 2006, se organizan en categorías para verificar si es posible mejorar el error de predicción.

Además, se selecciona una batería de máquinas de aprendizaje para clasificación, según las clases disponibles en la librería SickitLearn.

En esta etapa se seleccionaron las siguientes máquinas de aprendizaje:

- Regresión logística: LogisticRegression (LR)
- Árboles de decisión: DecisionTreeClassifier (CART)
- K-vecinos más cercanos: KNeighborsClassifier (KNN)
- Análisis discriminante lineal: LinearDiscriminantAnalysis (LDA)
- Naive Bayes Gausiano: GaussianNB (NB)
- Máquinas de soporte vectorial: SVC
- Percetrón multicapa: MLPClassifier (MLP)

Experimento 1.

Se discretizan todas las calificaciones, tomando de cada número real que representa la nota, el número entero inmediatamente inferior, realizando la reasignación como se ilustra en la siguiente tabla:

Nota original	Discretización
Entre [0, 1)	0
Entre [1, 2)	1
Entre [2, 3)	2
Entre [3, 4)	3
Entre [4, 5)	4

Tabla 16 Discretización de notas en 5 rangos

Con esa categorización, se procede a aplicar la batería de clasificadores, mediante validación cruzada en 10 conjuntos. Para evaluar los resultados de clasificación se utiliza como métrica la exactitud (*accuracy*), que corresponde al porcentaje de aciertos obtenidos al comparar la predicción obtenida por el clasificador en el conjunto de prueba frente a los valores reales de la categoría en dicho conjunto.

En la tabla siguiente se muestran los resultados obtenidos aplicando la batería de clasificadores para tratar de predecir las categorías en el curso de **Programación Estructurada**:

CLASIFICADOR	EXACTITUD	DESVIACION
LR	0.415403	0.025216
LDA	0.431429	0.023429
KNN	0.339543	0.040869
CART	0.405120	0.065708
NB	0.393536	0.025258
SVM	0.422670	0.028753
MPL	0.411002	0.028142

Tabla 17 Clasificadores aplicados a Programación estructurada

Los mejores resultados se obtienen con el clasificador LDA – Análisis Discriminante Lineal, con un 43% de exactitud, lo cual es un resultado poco satisfactorio.

Se realiza una prueba similar para las asignaturas de los primeros semestres, que son las que pueden resultar más influenciadas por los resultados de las pruebas ICES.

En el caso del **Cálculo Diferencial**, los mejores resultados se obtienen para el clasificador SVM: Máquina de Vectores de Soporte, con un 43% de exactitud como se aprecia en la tabla:

CLASIFICADOR	EXACTITUD	DESVIACION
LR	0.389232	0.024882
LDA	0.396727	0.015103
KNN	0.346224	0.015506
CART	0.390868	0.035655
NB	0.316440	0.028294
SVM	0.435334	0.013584
MPL	0.389276	0.025090

Tabla 18 Clasificadores aplicados a Calculo diferencial

Aplicando la batería de clasificadores en el curso de **Biología General** se obtienen los resultados que se muestran en la tabla siguiente, donde nuevamente, la Máquina de Vectores de Soporte, es la que muestra mejores resultados, con una exactitud del 48%.

CLASIFICADOR	EXACTITUD	DESVIACION
LR	0.421471	0.021178
LDA	0.423948	0.021679
KNN	0.364665	0.017812
CART	0.393076	0.042306
NB	0.283107	0.022457
SVM	0.484533	0.024554
MPL	0.419017	0.023533

Tabla 19 Clasificadores aplicados a Biología general

Aplicando nuevamente los clasificadores en la asignatura **Cátedra Faría**, nuevamente el mejor resultado es obtenido por la Máquina de Soporte Vectorial, con un 53% de exactitud:

CLASIFICADOR	EXACTITUD	DESVIACION
LR	0.499415	0.002188
LDA	0.500578	0.010956
KNN	0.394397	0.031018
CART	0.396709	0.022028
NB	0.282259	0.094853
SVM	0.532082	0.010266
MPL	0.505256	0.008145

Tabla 20 Clasificadores aplicados a Catedra Faria

Similarmente con la asignatura **Educación Ambiental**, la SVM obtiene los mejores resultados con un 51% de exactitud.

CLASIFICADOR	EXACTITUD	DESVIACION
LR	0.473505	0.005234
LDA	0.472315	0.003581
KNN	0.391020	0.014900
CART	0.416979	0.010799
NB	0.266119	0.091973
SVM	0.515921	0.011760
MPL	0.469955	0.006061

Tabla 21 Clasificadores aplicados a Educación ambiental

Finalmente, la misma prueba se aplica en la asignatura Habilidades Comunicativas, donde nuevamente con el SVM se obtiene un 55% de exactitud. Esta es la asignatura en la que se obtienen mejores resultados en la predicción, sin embargo, sigue siendo un resultado bajo:

CLASIFICADOR	EXACTITUD	DESVIACION
LR	0.506433	0.004678
LDA	0.504094	0.006820
KNN	0.384795	0.015865
CART	0.378947	0.054784
NB	0.452632	0.015951
SVM	0.550877	0.013024
MPL	0.507602	0.005730

Tabla 22 Clasificadores aplicados a Habilidades Comunicativas

Finalmente, para comprobar el comportamiento de los clasificadores con el resto de las materias registradas, se ejecuta un script que permite aplicar la batería de clasificadores, con validación cruzada para todas las asignaturas registradas. Una vez ejecutado el script, se seleccionan los resultados sobre aquellas materias que tienen una exactitud promedio entre los clasificadores por encima del 70%, con lo cual se consigue la siguiente tabla:

Código	LR	LDA	KNN	CART	NB	SVM	MLP	promedio
157102	0.743333	0.823333	0.790000	0.743333	0.786667	0.793333	0.756667	0.776667
167246	0.714286	0.714286	0.757143	0.857143	0.771429	0.842857	0.757143	0.773469
167271	0.680000	0.740000	0.780000	0.660000	0.740000	0.780000	0.720000	0.728571
168005	0.770000	0.730000	0.765000	0.720000	0.655000	0.770000	0.700000	0.730000
167225	0.810000	0.815000	0.860000	0.600000	0.815000	0.860000	0.785000	0.792143
173238	0.817143	0.817143	0.775238	0.752857	0.690476	0.859048	0.810476	0.788912
167281	0.888971	0.888971	0.888971	0.852574	0.845588	0.882721	0.870588	0.874055

Tabla 23 Resultados con exactitud sobre 70%

En este caso, 7 materias obtienen la exactitud mayor o igual al 70%, estas son: Matemáticas Espaciales, Modelamiento de Sistemas de Información, Tópicos Avanzados de Ingeniería del

Software, Investigación de Operaciones, Electiva Profesional III, Deporte Universitario, e Informática Extraplán.

Sin embargo, se quiere hacer un análisis más detallado, respecto a los resultados de la clasificación, para lo cual se calcula la matriz de confusión y una tabla de métricas de exactitud para cada una de las materias obtenidas en el último filtro, aplicando el clasificador de Máquinas de Soporte Vectorial (SVM) que en la mayoría de los casos es el que ha obtenido mejores resultados de predicción o uno de los mejores resultados.

La matriz de confusión es una herramienta que permite la verificar del desempeño de un algoritmo de clasificación supervisada. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. La matriz nos permite ver los aciertos y errores que consigue el clasificador a la hora de pasar por el proceso de aprendizaje con los datos.

En la matriz de confusión se encuentran los siguientes elementos:

Positivo (P): La observación es positiva (por ejemplo, es un perro).

Negativo (N): La observación no es positiva (por ejemplo, no es un perro).

Verdadero Positivo (TP): Resultado en el que el modelo predice correctamente la clase positiva.

Verdadero Negativo (TN): Resultado donde el modelo predice correctamente la clase negativa.

Falso Positivo (FP): También llamado error de tipo 1, resultado donde el modelo predice incorrectamente la clase positiva cuando en realidad es negativa.

Falso Negativo (FN): También llamado error de tipo 2, un resultado en el que el modelo predice incorrectamente la clase negativa cuando en realidad es positiva.

Respecto a las métricas calculadas sobre la matriz de confusión se tienen:

Precisión

La precisión también se conoce como valor predictivo positivo y es la proporción de instancias relevantes entre las instancias recuperadas. En otras palabras, responde a la pregunta "¿Qué proporción de identificaciones positivas fue realmente correcta?"

$$Precision = \frac{TP}{TP + FP}$$

Sensibilidad

La sensibilidad, tasa de aciertos o tasa positiva real (TPR), es la proporción de la cantidad total de instancias pertinentes que se recuperaron realmente. Responde a la pregunta "¿Qué proporción de positivos reales se identificó correctamente?"

$$Recall = \frac{TP}{TP + FN}$$

Especificidad

La especificidad, también conocida como tasa negativa real (TNR), mide la proporción de negativos reales que se identifican correctamente como tales. Es lo opuesto a la sensibilidad.

$$Specificity = \frac{TN}{TN + FP}$$

Puntuación F1

El puntaje F1 es una medida de la precisión de una prueba, es la media armónica de precisión y recuperación. Puede tener una puntuación máxima de 1 (precisión y recuerdo perfectos) y una mínima de 0. En general, es una medida de la precisión y robustez de su modelo.

$$F1\ score = \frac{2 * (precision * recall)}{precision + recall} = \frac{2TP}{2TP + FP + FN}$$

En las ilustraciones siguientes se observan los resultados obtenidos:

La siguiente figura ilustra los resultados para el curso de **Matemáticas Especiales**, en el que se obtuvo una exactitud de 70%. Sin embargo, al observar la matriz de confusión se verifican los detalles en la predicción. El conjunto de prueba tiene un total de 41 estudiantes, 26 de los cuales obtienen una calificación en categoría 3 (en el intervalo [3,4)) y 15 en la categoría 4 (en el intervalo [4,5)). De las 26 calificaciones en la categoría 3, todas son predichas correctamente. Sin embargo, solo 3 de las calificaciones en categoría 4 se predicen correctamente, y en cambio se tienen 12 falsos positivos.

La precisión para la clase 3 se da al dividir $TP / (TP+FP)$, es decir $26/(12+3)$ da como resultado 0.68. La precisión de la clase 4 es entonces $3/(3+0)$ que da como resultado 1.0

La sensibilidad (recall) en la clase 3 se obtiene de la fórmula $TP/(TP+FN)$, es decir $26/(26+0) = 1.0$: Por otro lado en la clase 4 la sensibilidad se obtiene como $12/(12+3) = 0.20$.

En general podemos decir que, en este caso, el clasificador acierta muy bien para los datos clasificados en la clase 3, pero tiene problemas para identificar elemento en la clase 3, donde genera falsos positivos.

```

===  MATEMÁTICAS ESPECIALES  ===

0.7073170731707317
[[26  0]
 [12  3]]

              precision    recall  f1-score   support

     3         0.68         1.00         0.81         26
     4         1.00         0.20         0.33         15

 accuracy                   0.71         41
 macro avg         0.84         0.60         0.57         41
 weighted avg         0.80         0.71         0.64         41

```

Tabla 24 Matriz de confusión aplicada a Matamaticas especiales

Los resultados obtenidos para a asignatura **Modelamiento de Sistemas de Información** se pueden apreciar en la figura siguiente. En este caso se tienen calificaciones en las categorías 2, 3 y 4. En este caso el clasificador predice correctamente los elementos de la clase 3, pero tiene dificultades con las clases 2 y 4. En la primera no reconoce verdaderos positivos y en la última genera un considerable número de falsos positivos.

=== MODELAMIENTO DE SISTEMAS DE INFORMACIÓN ===

0.803921568627451

```
[[ 0  4  0]
 [ 0 37  0]
 [ 0  6  4]]
```

	precision	recall	f1-score	support
2	0.00	0.00	0.00	4
3	0.79	1.00	0.88	37
4	1.00	0.40	0.57	10
accuracy			0.80	51
macro avg	0.60	0.47	0.48	51
weighted avg	0.77	0.80	0.75	51

Tabla 25 Matriz de confusión aplicada a Modelamiento de Sistemas de Información

En relación con **Tópicos Avanzados de Ingeniería del Software** que obtuvo una exactitud del 75% se genera la siguiente matriz de confusión, junto a sus métricas.

=== TÓPICOS AVANZADOS DE INGENIERÍA DEL SOFTWARE ===

0.75

```
[[ 0  4  0]
 [ 0 27  0]
 [ 0  5  0]]
```

	precision	recall	f1-score	support
2	0.00	0.00	0.00	4
3	0.75	1.00	0.86	27
4	0.00	0.00	0.00	5
accuracy			0.75	36
macro avg	0.25	0.33	0.29	36
weighted avg	0.56	0.75	0.64	36

Tabla 26 Matriz de confusión aplicada a Tópicos avanzados

En la figura se puede ver que, en forma similar al curso anterior, los datos pertenecen a las clases 2, 3 y 4. La clase 3 predice correctamente todos los valores en su categoría, pero no se detectan los datos en las clases 2 y 4, generando falsos negativos en ambas clases.

Con respecto a **Investigación de Operaciones**, se obtiene la siguiente tabla y sus métricas asociadas:

```

===  INVESTIGACIÓN DE OPERACIONES  ===

0.6875
[[22  0]
 [10  0]]

              precision    recall  f1-score   support

     3         0.69         1.00         0.81         22
     4         0.00         0.00         0.00         10

 accuracy          0.69         0.69         0.69         32
 macro avg         0.34         0.50         0.41         32
 weighted avg         0.47         0.69         0.56         32

```

Tabla 27 Matriz de confusión aplicada a Investigación de Operaciones

Si bien la exactitud obtenida con el clasificador está en 68%, en la validación cruzada sobre pasó el umbral del 70%. Podemos apreciar de la gráfica que los datos pertenecen a la clase 3 y 4, y el clasificador tiene dificultad para identificar los datos en la clase 4, obteniendo un 100% de falsos negativos en dicha clase.

Respecto a la **Electiva Profesional III**, nuevamente se tienen elementos en la clase 3 y 4 y se aprecia que no se identifican correctamente los elementos en la clase 3, generando 100% de falsos negativos, que se convierten en falsos positivos de la clase 4.

```

=== ELECTIVA PROFESIONAL III ===

0.7666666666666667
[[ 0  7]
 [ 0 23]]

```

	precision	recall	f1-score	support
3	0.00	0.00	0.00	7
4	0.77	1.00	0.87	23
accuracy			0.77	30
macro avg	0.38	0.50	0.43	30
weighted avg	0.59	0.77	0.67	30

Tabla 28 Matriz de confusión aplicada a Electiva Profesional III

Las siguientes figuras muestran los resultados obtenidos con **Actividad Recreativa e Informática Básica**. Ambos cursos mantienen un patrón similar, en el sentido de que existen las clases 0 y 4, con mayor cantidad de datos en la segunda clase y dificultad para detectar elementos en la clase 0, generando más del 85% de falsos negativos.

=== ACTIV DEPOR,RECRE Y CULT-DEP UNIVERSITARIO ===					=== INFORMÁTICA BÁSICA (EXTRAPLAN) ===				
0.8529411764705882					0.8793103448275862				
[[1 15]					[[2 14]				
[0 86]]					[0 100]]				
	precision	recall	f1-score	support		precision	recall	f1-score	support
0	1.00	0.06	0.12	16	0	1.00	0.12	0.22	16
4	0.85	1.00	0.92	86	4	0.88	1.00	0.93	100
accuracy			0.85	102	accuracy			0.88	116
macro avg	0.93	0.53	0.52	102	macro avg	0.94	0.56	0.58	116
weighted avg	0.87	0.85	0.79	102	weighted avg	0.89	0.88	0.84	116

Tabla 29 Matriz de confusión aplicada a Actividad Recreativa e Informática Básica

3.5. Síntesis de resultados

En relación con los resultados mostrados por estudiantes aceptados para ingresar al programa, se puede apreciar una tendencia creciente en el puntaje desde el periodo 206-2 hasta el 2018-1. Específicamente en las áreas del conocimiento que más peso llegan a tener en el desarrollo del programa de Ingeniería de Sistemas, podemos ver claramente la tendencia en la siguiente ilustración:

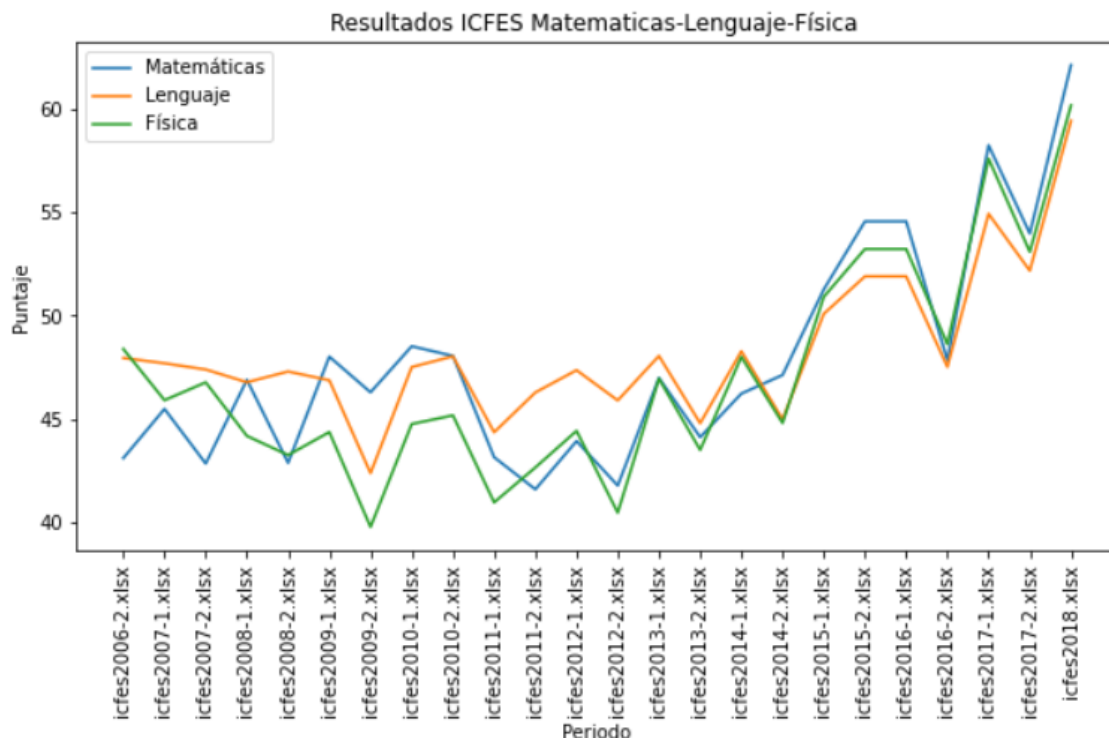


Ilustración 82. Comportamiento de los resultados en pruebas ICFES en matemáticas, física y lenguaje

Sin embargo, de los archivos obtenidos no se puede comparar esa temporalidad con las calificaciones obtenidas por los estudiantes activos del programa, ya que no es posible con los datos disponibles seguir esta línea de tiempo. Lo que podría inferirse es que con el paso del tiempo, el aumento en el número de aspirantes al programa ha permitido que se realice una selectividad de los estudiantes, consiguiendo aprobar el ingreso a jóvenes con cada vez, mejores resultados en sus pruebas de estado.

En la gráfica siguiente se puede apreciar incluso que los resultados en matemáticas en los últimos periodos se aprecian por encima de las demás áreas del conocimiento. Por otro lado, aunque no se tiene en cuenta en el proceso de selección, en el área de filosofía, que está registrada en el sistema, los resultados se muestran en su mayoría en la cota inferior de los puntajes.

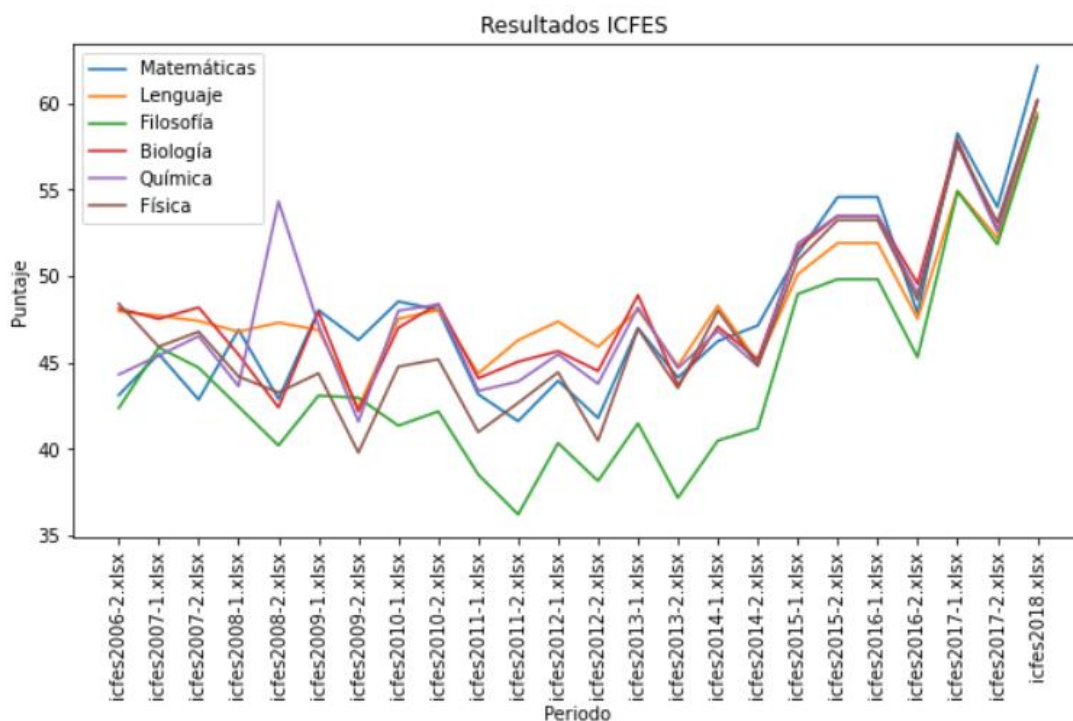


Ilustración 83. Comportamiento de los resultados ICFES en todas las áreas

En relación con los resultados obtenidos en los cursos regulares del programa de los estudiantes que cursaron el pensum 2006, entre 2006 y 2018, se pueden apreciar algunas singularidades:

En los cursos de los primeros semestres se observa que el rango de calificaciones finales se encuentra en todo el espectro posible. Es decir hay calificaciones desde cero, hasta cinco. Los histogramas muestran una tendencia mayor a calificaciones sobre tres, pero siempre aparecen notas registradas como pedidas, a medida que aumentan los semestres académicos, la tendencia es a tener menor registro en notas pedidas, presentándose la mayoría sobre 3. En los semestres avanzados, muchos cursos no tienen registro de notas perdidas, con calificaciones altas de media 4.0. Lo que permite apreciar este comportamiento, es que efectivamente los primeros semestres, y particularmente el primero tiene una alta tasa de mortalidad académica, donde probablemente ocurre la mayor cantidad de deserción, la cual se reduce al avanzar los semestres académicos.

En relación al análisis predictivo. Se realizan dos experimentos. En el primero se busca predecir mediante regresión la calificación de cada materia, teniendo como variables predictoras, los resultados obtenidos en la pruebas ECAES. De forma intuitiva se supone que las primeros cursos estarán más influenciados por la recién presentadas pruebas, por lo cual se aplica la batería de máquinas de aprendizaje especialistas en regresión a los primeros cursos: Cálculo Diferencias, Programación Estructurada, Biología General, Habilidades Comunicativas y Cátedra Faría. En todos los casos el error de predicción está por encima al 20%. Eso significa más de 1.0 de error en la nota

que podemos predecir (en la escala de 0 a 5), por lo cual no parecen buenos resultados en la predicción.

Para verificar el comportamiento en las materias de cursos superiores, se aplica el proceso de predicción a todos los cursos, y se filtran aquellos con mejores resultados. Específicamente aquellos donde el error de predicción está por debajo del 10%, es decir 0.5 puntos arriba o debajo de la nota esperada. En dicho caso se obtienen los siguientes resultados:

Asignatura	Error en la predicción	Porcentaje de error	Nota mínima	Nota máxima	Promedio
<i>FUNDAMENTOS Y DISEÑO DE BASE DE DATOS</i>	0.33	6.57	0	4.70	3.55
<i>METODOLOGÍA DE LA INVESTIGACIÓN</i>	0.33	6.68	0	5.0	3.92
<i>ELECTIVA PROFESIONAL II</i>	0.36	7.10	2.6	5.0	4.27
<i>REDES DE COMPUTADORES Y SISTEMAS DISTRIBUIDOS II</i>	0.38	7.66			
<i>MODELAMIENTO DE SISTEMAS DE INFORMACIÓN</i>	0.40	8.00	0.60	4.50	3.42
<i>ESTADÍSTICA II</i>	0.42	8.39	0	4.90	3.48
<i>ELECTIVA DE INGENIERÍA I</i>	0.43	8.65	2.6	5.0	3.69
<i>ELECTIVA PROFESIONAL III</i>	0.45	8.92			
<i>TÓPICOS AVANZADOS DE INGENIERÍA DEL SOFTWARE</i>	0.45	8.93	2.4	4.7	3.51
<i>MATEMÁTICAS ESPECIALES</i>	0.49	9.89	0	5.0	3.44

Tabla 30. Predicción de notas con margen de error del 10%

En este caso es importante aclarar que dichos cursos sobresalen porque no están ubicados en los primeros semestres y aunque algunos de ellos tienen calificaciones por debajo de 3.0 (incluso en 0.0) Estos podrían considerarse como datos atípicos, encontrándose la mayoría de las calificaciones entre 3.0 y 5.0, lo cual implicaría un menor rango de calificación.

El segundo experimento de predicción es un ejercicio de clasificación. En este caso se discretizan las calificaciones para crear 5 intervalos o categorías, como se muestra en la siguiente tabla:

Nota original	Discretización
Entre [0, 1)	0
Entre [1, 2)	1
Entre [2, 3)	2

Entre [3, 4)	3
Entre [4, 5)	4

Tabla 31. Discretización de las calificaciones en 5 categorías.

Se aplicó como en el caso anterior una batería de máquinas de aprendizaje, en este caso para resolver problemas de clasificación. Específicamente los clasificadores utilizados fueron: Regresión logística, Árboles de decisión, k-vecinos más cercanos, Análisis discriminante lineal, Naive Bayes Gausiano, Máquinas de soporte vectorial y Perceptrón multicapa.

Inicialmente se pretende verificar la exactitud en la predicción de categorías en los cursos del primer semestre, nuevamente bajo el supuesto que son las que más influenciadas están por los resultados de las pruebas ICFES, es así como se aplican pruebas mediante validación cruzada para los cursos de Calculo Diferencias, Programación Estructurada, Cátedra Faría, Educación Ambiental y Habilidades Comunicativas. En la mayoría de los casos la exactitud en la predicción no supera el 50%.

Se plantea entonces extender las pruebas a las demás asignaturas, y seleccionar aquellos cursos en los cuales la exactitud en la predicción supere el 70%

De esta forma, se obtiene los mejores resultados en las predicciones para las asignaturas de Matemáticas Especiales con Exactitud del 70%, Modelamiento de sistemas de Información con Exactitud del 80%, Tópicos Avanzados en Ingeniería del Software con exactitud del 75%, Investigación de operaciones con exactitud del 70%, Electiva Profesional III con exactitud del 76% y Deportes e Informática Básica, ambos sobre el 85%.

Salvo Deportes e Informática, las materias coinciden con las que mejor se comportaron para el caso de la regresión. Sin embargo, al realizar un análisis de la matriz de confusión y sus métricas, se aprecia que por un lado, las categorías no quedan simétricamente distribuidas entre los datos de prueba, y que hay una penalización en la clase menos frecuente, en la cual el número de verdaderos positivos es bajo. Es decir los clasificadores tiene problemas para detectar la clase menos frecuente, y el resultado alto en el rendimiento del pronóstico se beneficia por los aciertos en la clase más frecuente.

4. Conclusiones y trabajos futuros

En el presente trabajo se realizó el análisis de los datos de inscripción obtenidos por los aspirantes al programa de Ingeniería de Sistemas, y particularmente los resultados de las pruebas ICFES de los aspirante, y el historial académico obtenido por los estudiantes matriculados en el pensum 2006, entre los años 2006 a 2018.

Se utilizaron los datos suministrados por la Oficina de Registro y Control Académico, los cuales se manejaron de forma anónima. Se realiza la limpieza, transformación y organización de los datos para realizar un análisis descriptivo y predictivo

En el análisis descriptivo queda manifiesta la tendencia creciente de los puntajes obtenidos por los aspirantes en las pruebas ICFES, en todas las áreas del conocimiento durante el periodo analizado, sobresaliendo el área de Matemáticas.

Al aplicar los índices de correlación, se observa valores bajos de correlación, lo cual implica que no parece haber una fuerte correlación lineal. Sin embargo, se procede a aplicar algunos modelos para el análisis de regresiones no lineales y de clasificación

En los experimentos aplicados, no se encuentra una alta capacidad de predicción a partir de los resultados obtenidos en las pruebas ICFES, para determinar el comportamiento académico en los primeros cursos del programa, donde se presume que puede haber mayor influencia por su cercanía en el tiempo. Los resultados tanto en regresión como en clasificación son por decirlo menos, muy modestos.

Se aprecia que los mejores resultados, medidos en términos del error de predicción para correlación y por la exactitud en la clasificación se da en cursos avanzados, en ambos casos en las mismas materias. Sin embargo, los resultados deben interpretarse con precaución ya que estos cursos se caracterizan por tener un rango de notas entre 3.0 y 5.0, salvo algunos valores atípicos, lo cual puede beneficiar la bondad en los resultados.

Dentro de los trabajos futuros, existen varios experimentos que no se aplicaron en este trabajo, pero que pueden ser interesantes, con el fin de mejorar los resultados, entro otras cosas redefinir

categorías, y predecir asignaturas, incluyendo dentro de las variables predictoras cursos del pregrado ya cursados. Podría encontrarse potencial para definir un modelo para el desarrollo de sistemas de alertas tempranas.

5. Referencia

- Alberca, A. S. (1 de Julio de 2010). *Aprende con Alf*. Obtenido de Recursos Educativos Libres - Matemáticas, Estadística y Programación para Educación Superior - Universidad CEU San Pablo: <https://aprendeconalf.es/docencia/estadistica/manual/regresion/>
- Espectador. (2018). «El problema no es solo plata: 42% de los universitarios deserta.,». Bogota: El Espectador.
- ESTAMÁTICA. (s.f.). <https://estamatica.net/tipos-de-regresion-en-estadistica/>. Obtenido de estamatica@gmail.com.
- Heras, J. M. (2022). *IArtificial.net*. Obtenido de IArtificial.net
- Hernández, F. (18 de 03 de 2021). Obtenido de Modelos Predictivos: https://fhernanb.github.io/libro_mod_pred/
- Pertuz, C. P. (12 de 10 de 2021). *Aprendizaje automático y profundo en python*. Bogota: Ediciones de la U. Obtenido de <https://books.google.com.co/books?id=mgNcEAAAQBAJ&pg=PA202&dq=Algoritmo+na%C3%AFves+bayes&hl=es&sa=X&ved=2ahUKEwisw9Seraj4AhUQRTABHZUUBCwQ6AF6BAgKEAl#v=onepage&q&f=false>
- questionpro. (s.f.). *Análisis de regresión: Qué es, tipos y cómo realizarlo*. Obtenido de <https://www.questionpro.com/blog/es/analisis-de-regresion/>
- Trujillo Cabezas, R. (12 de 6 de 2022). Obtenido de <https://books.google.com.co/books?id=8qGkDwAAQBAJ&lpg=PA43&dq=algoritmo%20del%20vecino%20m%C3%A1s%20cercano&hl=es&pg=PA43#v=onepage&q=algoritmo%20del%20vecino%20m%C3%A1s%20cercano&f=true>
- Weiss, S. (1998). En S. y. Weiss, *Predictive Data Mining. A Practical Guide*. San Fransisco.

6. Anexo

Pensum 2006

