# ANÁLISIS DESCRIPTIVO DE LA DESERCIÓN EN LOS ESTUDIANTES DE LA ESCUELA COLOMBIANA DE INGENIERÍA MEDIANTE EL USO DE MÉTODOS DE CLASIFICACIÓN EN PYTHON

PRESENTADO POR: JULIÁN BENITO BAQUERO

TRABAJO DE GRADO DIRIGIDO POR:

LUIS FRANCISCO LÓPEZ CASTRO

INGENIERO INDUSTRIAL

INGENIERÍA INDUSTRIAL

BOGOTÁ DC

2020

TABLA DE CONTENIDO

# LISTADO DE TABLAS

# LISTADO DE FIGURAS

# 1. RESUMEN

El siguiente documento contiene un análisis descriptivo del fenómeno de la deserción en los estudiantes de la Escuela Colombiana de Ingeniería – Julio Garavito. Este proceso se llevó a cabo bajo los parámetros del modelo CRISP-DM utilizado principalmente en proyectos de minería de datos.

Para esto se hace en primer lugar un análisis de la situación partiendo desde Latinoamérica para finalmente llegar a La Escuela donde se evaluaron que factores se podían utilizar para tener una buena medición de este fenómeno en los estudiantes de todos los programas académicos. Con una revisión teórica y acceso a las bases de datos de La Escuela, se logró tener suficiente observación para poder hacer un análisis de clasificación de los estudiantes, proceso realizado con ayuda de la herramienta informática Python.

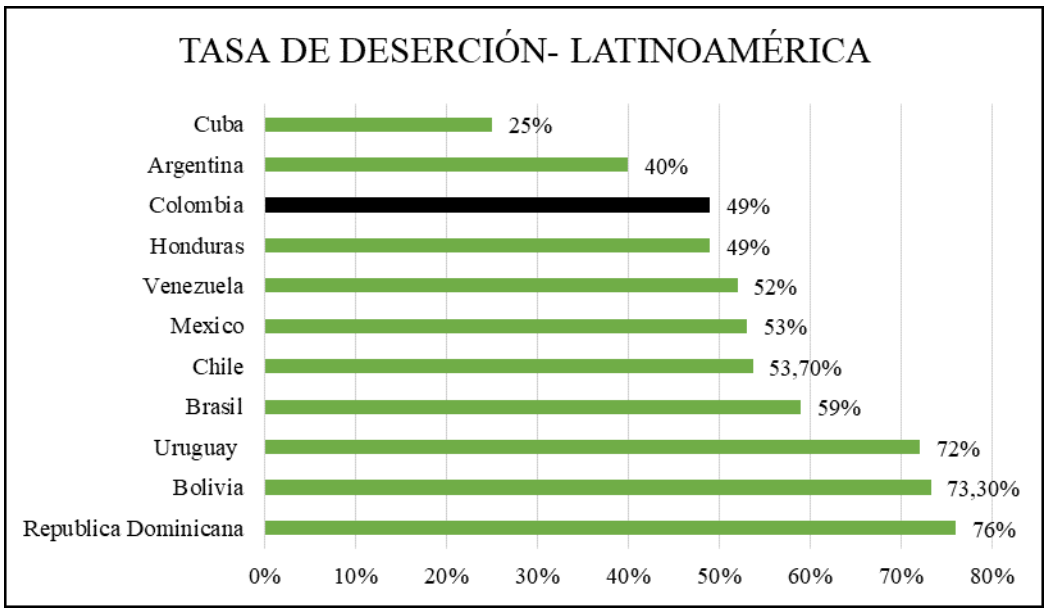
Con la elaboración de los modelos en Python, finalmente se logra llegar a conclusión respecto al proceso y se hace un análisis de lo que fue el proceso en si de la implementación de estos modelos a esta situación.

## 3. DESCRIPCION DEL PROBLEMA DE INVESTIGACION

El tema de la deserción estudiantil ha sido objeto de investigación durante los últimos años para las universidades públicas y privadas. Según la revista Dinero un estudio del Banco Mundial afirmo que el 42% de los jóvenes que logran ingresar a una universidad terminan por desertar, opacando las grandes oportunidades que actualmente se están dando para tener acceso a la educación superior. Entre las razones para tomar esta decisión se encuentran las razones individuales como el entorno familiar , la integración social, las expectativas no satisfechas ; Las razones académicas, como el método utilizado por los docentes para educar, la cantidad de materias cursadas por semestre, la calidad del programa; Las razones institucionales como los recursos universitarios disponibles para los estudiantes, el apoyo académico y psicológico, el nivel de interacción de los profesores y los estudiantes y las razones socioeconómicas como el estrato, la situación laboral, la dependencia económica, entre otras.

El tipo de deserción también se puede clasificar teniendo en cuenta el tiempo. Se clasifica en: Deserción precoz, un individuo que habiendo sido admitido en la universidad no se matricula; Deserción temprana, un individuo que abandona sus estudios en los primeros semestres de su programa de pregrado y deserción tardía, un individuo que abandona sus estudios en los últimos semestres.

A nivel Latinoamérica, el ministerio de Educación mostro una estadística de deserción para el año 2005.




*Grafica 1 : Tasa de deserción en Latinoamérica*

Colombia tiene una tasa de deserción de educación superior de Colombia 49% siendo una de las mejores en América Latina para este año. La dificultad para realizar este estudio anualmente cada vez aumenta debido a la cantidad de estudiantes que se matriculan por año y la complejidad para recolectar esa información para todos los países.

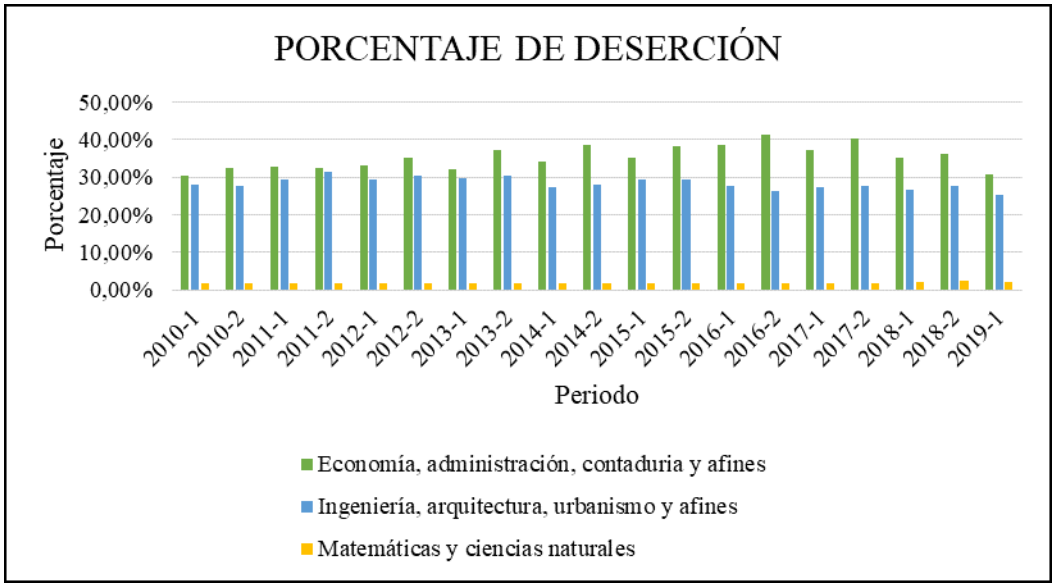
Con el fin de tener una base de datos que pueda ayudar a analizar la deserción en Colombia, se creó SPADIES (Sistema de prevención y análisis de la deserción en las instituciones de educación superior). El SPADIES es un sistema de información especializado para analizar la permanencia en la educación superior colombiana.

A nivel nacional, para las áreas de conocimiento de:

- Economía, administración, contaduría y afines.
- Ingeniería, arquitectura, urbanismo y afines.
- Matemáticas y ciencias naturales.

| Area de conocimiento | Economía, administración, contaduria | Ingeniería, arquitectura, urbanismo | Matemáticas y ciencias naturales |
|---|---|---|---|
| **2010-1** | 30,30% | 28,14% | 1,75% |
| **2010-2** | 32,57% | 27,64% | 1,89% |
| **2011-1** | 32,65% | 29,25% | 1,74% |
| **2011-2** | 32,39% | 31,24% | 1,75% |
| **2012-1** | 33,20% | 29,49% | 1,84% |
| **2012-2** | 35% | 30,39% | 1,73% |
| **2013-1** | 32,20% | 29,83% | 1,75% |
| **2013-2** | 37,16% | 30,24% | 1,63% |
| **2014-1** | 33,98% | 27,23% | 1,71% |
| **2014-2** | 38,64% | 28,15% | 1,87% |
| **2015-1** | 35,04% | 29,22% | 1,66% |
| **2015-2** | 38,10% | 29,37% | 1,69% |
| **2016-1** | 38,69% | 27,49% | 1,67% |
| **2016-2** | 41,22% | 26,23% | 1,70% |
| **2017-1** | 37,26% | 27,41% | 1,74% |
| **2017-2** | 40,28% | 27,68% | 1,73% |
| **2018-1** | 35,30% | 26,60% | 1,93% |
| **2018-2** | 36,16% | 27,55% | 2,27% |
| **2019-1** | 30,58% | 25,34% | 2,01% |

*Tabla 1 : Tasa de deserción por área de conocimiento. Datos tomados del "Sistema de prevención de la Deserción en la Educación Superior" SPADIES*

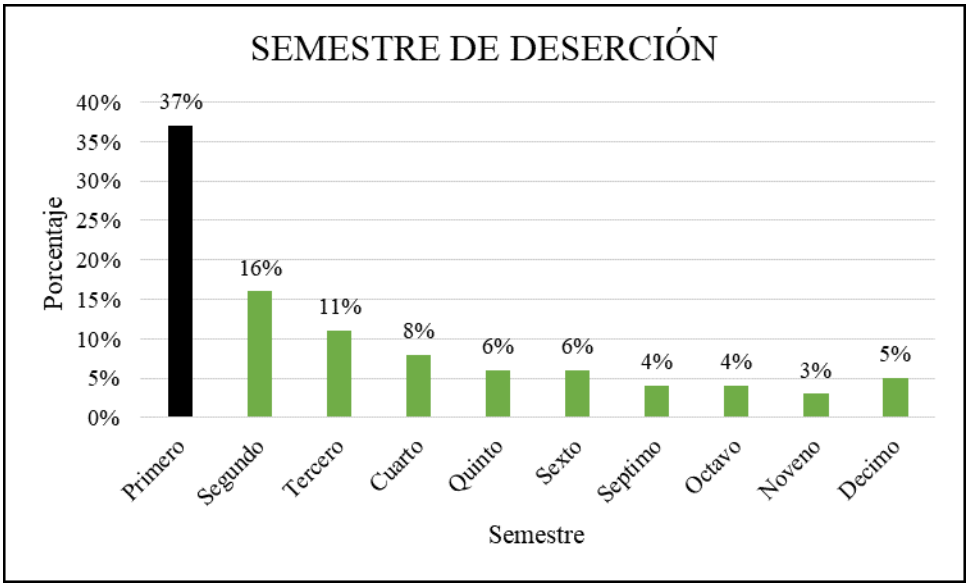



*Grafica 2 Comparación de la tasa de deserción por área de conocimiento. Datos tomados del "Sistema de prevención de la Deserción en la Educación Superior" SPADIES*

Teniendo en cuenta los datos con los que cuenta el Ministerio de Educación respecto a la cantidad de matriculados y desertores en las universidades de Colombia, en promedio la tasa de

deserción para las carreras de Economía, administración contaduría y afines es de 35,30%, para Ingeniería, Arquitectura, urbanismo y afines es de 28,34% y Matemáticas y ciencias naturales es del 1,79%.

Identificando que el porcentaje de deserción es bastante alto, es importante no solo identificar esta cifra sino hacer un análisis más específico de los elementos que pueden estar generando este indicador. Uno de los más importantes y que es necesario identificar es que tipo de deserción es teniendo en cuenta el tiempo, como se mencionó anteriormente. El ministerio de Educación muestra el siguiente resumen:

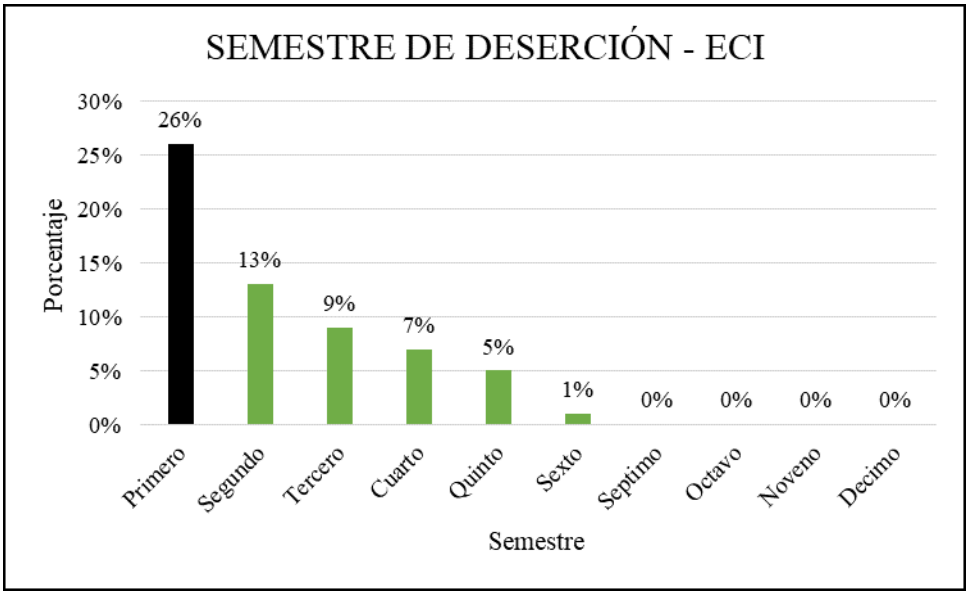



*Grafica 3 Tasa de deserción por semestre académico .Datos tomados del "Deserción estudiantil en la Educación superior Colombiana", 2016, p.75, Bogotá, Colombia.*

La deserción es temprana, siendo el primer año de estudio el periodo de tiempo donde los estudiantes más desertan de sus carreras universitarias a nivel general.

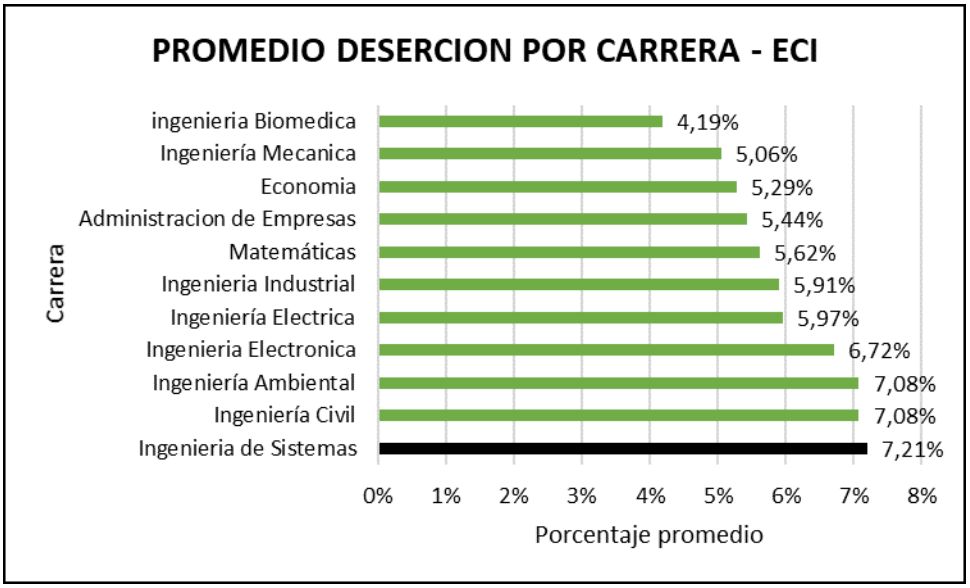
Se tienen en cuenta como base estas tres áreas de conocimiento (Áreas de Economía, Ingeniería y Matemáticas) debido que son las carreras de pregrado que tiene actualmente la Escuela

Colombiana de Ingeniería – Julio Garavito. Con los datos de los estudiantes matriculados en los últimos cinco años de los programas de pregrado se pueden sacar varias conclusiones, teniendo en cuenta que se tienen en cuenta todos los factores que pueden influir en que un estudiante decida desertar de su carrera:




*Grafica 4 Tasa de deserción por semestre - Escuela Colombiana de Ingeniería Julio Garavito. Datos tomados base de datos – Escuela Colombiana de Ingeniería Julio Garavito*

De la misma manera que el estudio nacional, en su mayoría la Escuela Colombiana de Ingeniera-Julio Garavito presenta deserción temprana, siendo el primer año académico (Primer y segundo semestre) el momento en que más desertan los estudiantes. Después de haber cursado mitad del programa, los estudiantes no desertan de sus carreras, por lo que el estudio se debe desarrollar en los primeros semestres académicos. Haciendo el mismo análisis por carrera tenemos que:




*Grafica 5 Promedio de deserción por programa académico- Escuela Colombiana de Ingeniería Julio Garavito. Datos tomados Base de datos- Escuela Colombiana de Ingeniería Julio Garavito*

En promedio por semestre, las carreras de Ingeniería Ambiental, Civil y de sistemas son las que más presentan desertores. El valor es muy bajo debido a que luego del quinto semestre ya no hay desertores mejorando el indicativo visto durante los 5 semestres que idealmente tardaría un estudiante en graduarse.

# 4. OBJETIVOS

## 4.1 OBJETIVO GENERAL

Realizar un análisis sobre la deserción en los estudiantes de los primeros semestres de los programas de pregrado de la Escuela Colombiana de Ingeniería - Julio Garavito

## 4.2 OBJETIVOS ESPECÍFICOS

- Identificar el comportamiento de la deserción en los últimos dos años de los estudiantes de los primeros semestres en los programas de pregrado de la Escuela Colombiana de Ingeniería - Julio Garavito.
- Identificar en qué periodo y en qué condiciones se presentan más casos de deserción.
- Identificar si existe una relación entre las pruebas de conocimiento y la deserción como un fenómeno causa - efecto y proponer posibles cambios para implementar a futuro.
- Realizar un estudio de clasificación de los estudiantes mediante el uso de la herramienta Python.

# 5. MARCO TEORICO

## 5.1 LA EDUCACION SUPERIOR EN COLOMBIA

La Ley 30 de 1992 establece que la Educación Superior a nivel pregrado está dividida en tres niveles de formación a saber: el nivel profesional, el nivel tecnológico y el nivel técnico. Estos programas son ofrecidos por las universidades, instituciones universitarias o escuelas tecnológicas y las instituciones técnicas profesionales. Para analizar de manera efectiva cómo está la educación en el país, se identifican los siguientes indicadores: a) la cobertura y número de matriculados, b) el número de instituciones y programas, c) el nivel formativo de los profesores y d) la tasa de deserción.

## 5.2 NÚMERO DE MATRICULADOS Y TASA DE COBERTURA

Uno de los indicadores más importantes para medir el acceso de la población a la Educación Superior es la tasa de cobertura bruta, indicador que mide la participación de los jóvenes y adultos que se encuentran efectivamente cursando un programa de educación superior. Así mismo, se encuentran los datos estadísticos de la Educación Superior en Colombia, específicamente aquellos matriculados para técnico profesional. Para Tecnólogo o universitario, se muestra el incremento en la demanda de la educación.

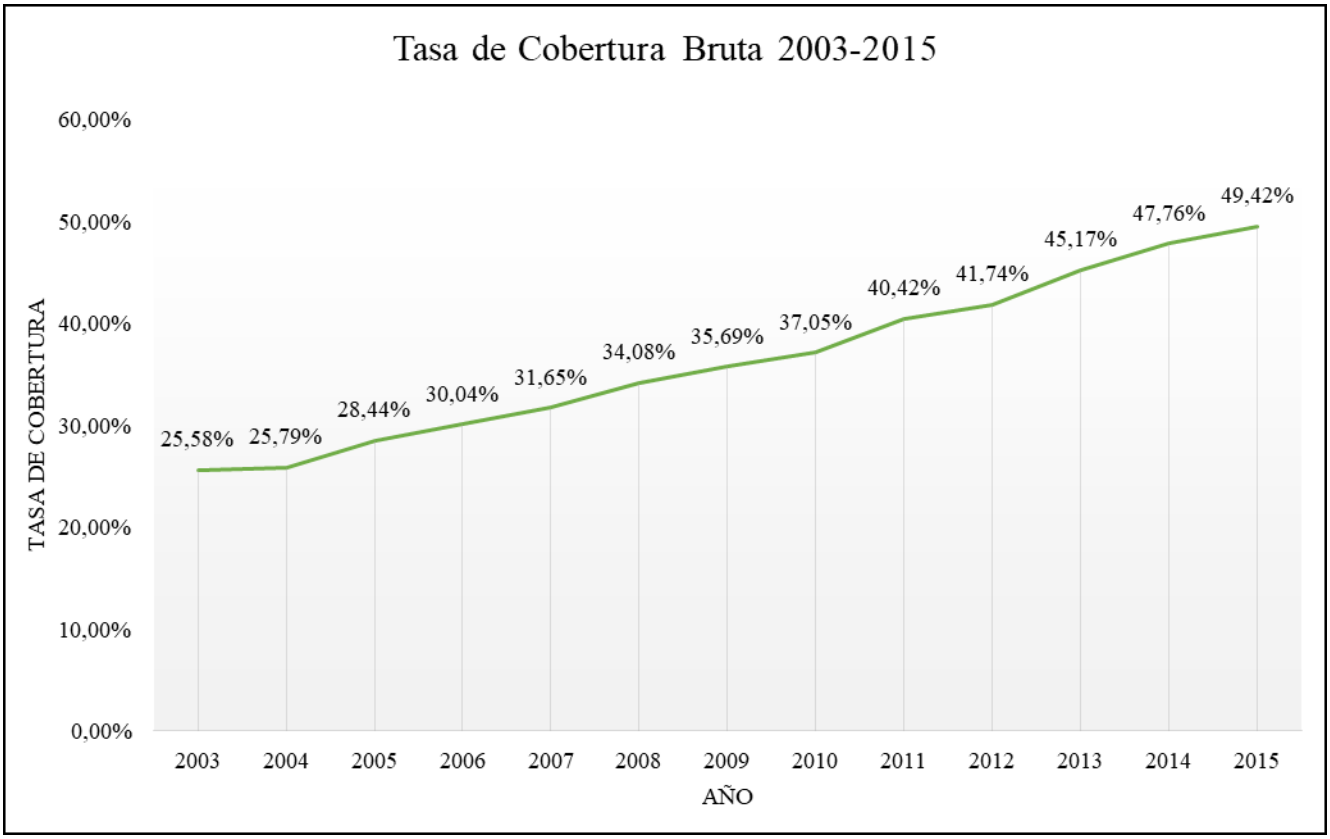
En el periodo comprendido entre los años 2004 y 2008 los matriculados para carreras a nivel técnico y profesional fueron aumentando respecto del total de matriculados. Luego del año 2009,

se da una tendencia de disminución teniendo una participación de 4.4% respecto del total, mientras que los tecnólogos desde 2003 han aumentado considerablemente su participación, incrementando del 13.1% a 29% en los últimos 12 años. Por su parte, las carreras universitarias acumulan la mayor cantidad de gente matriculada, pasando de 781.595 matriculados en el 2003 a 1'431.983 en el 2015.

| | Técnico Profesional | Part % | Tecnólogo | Part % | Universitario | Part % | Total |
|---|---|---|---|---|---|---|---|
| **2003** | 84.674 | 8,5% | 130.419 | 13,1% | 781.595 | 78,4% | 996.688 |
| **2004** | 84.648 | 8,3% | 133.121 | 13,1% | 799.979 | 78,6% | 1.017.748 |
| **2005** | 136.533 | 12,0% | 159.112 | 14,0% | 842.127 | 74,0% | 1.137.772 |
| **2006** | 171.386 | 14,0% | 175.862 | 14,4% | 872.720 | 71,5% | 1.219.968 |
| **2007** | 207.188 | 15,9% | 188.249 | 14,4% | 910.228 | 69,7% | 1.305.665 |
| **2008** | 224.026 | 15,7% | 239.954 | 16,8% | 963.167 | 67,5% | 1.427.147 |
| **2009** | 150.641 | 9,9% | 347.741 | 23,0% | 1.015.608 | 67,1% | 1.513.990 |
| **2010** | 92.941 | 5,9% | 449.686 | 28,3% | 1.045.133 | 65,8% | 1.587.760 |
| **2011** | 82.358 | 4,7% | 504.113 | 28,9% | 1.159.512 | 66,4% | 1.745.983 |
| **2012** | 78.555 | 4,3% | 515.129 | 28,4% | 1.218.816 | 67,2% | 1.812.500 |
| **2013** | 83.016 | 4,2% | 587.914 | 29,9% | 1.296.123 | 65,9% | 1.967.053 |
| **2014** | 96.466 | 4,6% | 614.825 | 29,6% | 1.369.149 | 65,8% | 2.080.440 |
| **2015** | 93.970 | 4,4% | 623.551 | 29,0% | 1.431.983 | 66,6% | 2.149.504 |

*Tabla 2 Número de matriculados por nivel de formación 2003-2015. Datos tomados del "Compendio estadístico de la Educación Superior Colombiana" de Ministerio de Educación, 2016, p.112, Bogotá, Colombia.*

La cobertura bruta a nivel nacional ha ido en constante crecimiento desde el año 2003 donde presentó una tasa del 25,58% al año 2015 con el 49.42%.

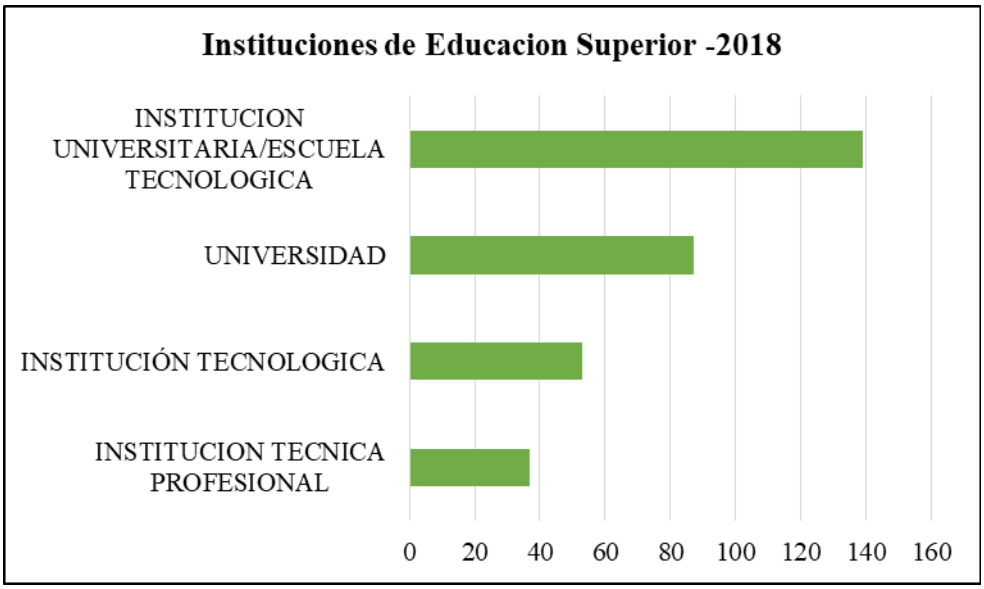



*Grafica 6 Tasa de cobertura Bruta 2003-2015 .Datos tomados del "Compendio estadístico de la Educación Superior Colombiana" de Ministerio de Educación, 2016, p.112, Bogotá, Colombia.*

## 5.3 NÚMERO DE INSTITUCIONES Y PROGRAMAS

Según el Ministerio de Educación "las Instituciones de Educación Superior (IES) son las entidades que cuentan, con arreglo de las normales legales, con el reconocimiento oficial como prestadoras del servicio público de la educación superior en el territorio colombiano". Estas instituciones de Educación Superior se dividen en: a) Instituciones Técnicas Profesionales, b) Instituciones Tecnológicas, c) Instituciones Universitarias o Escuelas Tecnológicas y d) Universidades. (Ministerio de Educación, 2015).

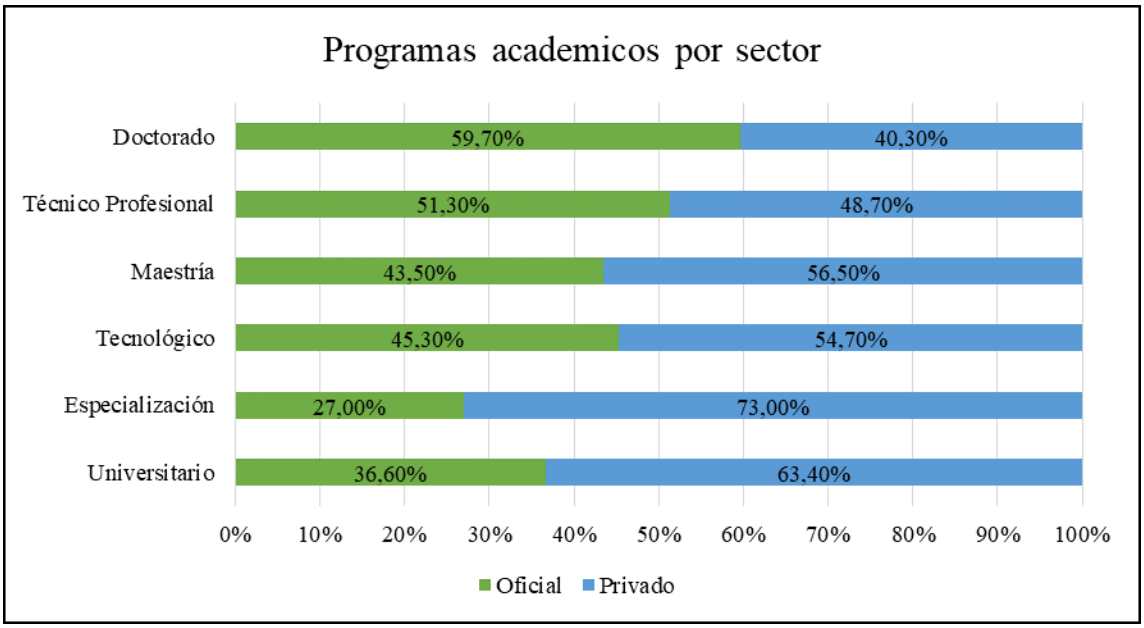
"El carácter de Universidad lo pueden alcanzar por el mandato legal (Artículo 20 de la Ley 30) las instituciones que, teniendo el carácter académico de instituciones universitarias o escuelas tecnológicas, cumplan con los requisitos indicados en el Artículo 20 de la Ley 30 de 1992, los cuales están desarrollados en el Decreto 1212 de 1993". (Ministerio de Educación, 2015). Un reporte para el año 2015 del SNIES (Sistema Nacional de Información de la Educación Superior) mostraba que las instituciones para brindar educación superior se dividían en: 119 instituciones universitarias, 83 universidades, 51 instituciones tecnológicas y 37 Instituciones técnicas. Al año 2018, se han incrementado en 10 las instituciones universitarias, en 4 las universidades y en 2 las instituciones tecnológicas. Las instituciones técnicas profesionales no han incrementado en los 3 años.




*Grafica 7 Instituciones de educación superior 2018 .Datos tomados del "Compendio estadístico de la Educación Superior Colombiana" de Ministerio de Educación, 2016, p.112, Bogotá, Colombia*

Para el año 2015, se ofrecieron un total de 11.213 programas académico dividido en 4.246 para el sector oficial y 6.967 para el sector privado. Así mismo, para el sector oficial se ofrecieron

1.376 programas universitarios, 856 especializaciones, 688 tecnólogos, 638 maestrías, 547 técnicos profesionales y 141 doctorados. Por otro lado, para el sector privado se ofrecieron 2.380 programas universitarios, 2.315 especializaciones, 831 tecnólogos, 827 maestrías, 519 técnicos profesionales y 95 doctorados.




*Grafica 8 Programas académicos por sector. Datos tomados del "Compendio estadístico de la Educación Superior Colombiana" de Ministerio de Educación, 2016, p.51, Bogotá, Colombia.*

## 5.4 NIVEL FORMATIVO DE LOS PROFESORES

En el reporte del Ministerio de Educación para el año 2015, Colombia contaba con 149.280 docentes de los cuales 71.937 (41,8%) son docentes de cátedra, 32.496 (21,8%) son de medio tiempo o tiempo parcial y 44.847 (30%) son de tiempo completo. A nivel académico, de un total de 144.390 docentes, el 0.73% tienen título de técnico profesional, el 1.48% título de tecnólogo, el 30,3% son profesionales, el 30,82% tienen una especialización, el 30,37% tienen una maestría y el 6.56% tienen un doctorado. Ahora bien, para el periodo del 2007 al 2015, la cantidad de docentes aumentó en un 47.18% Sin embargo, el promedio de docentes que tienen una maestría o doctorado (28.027 con maestría y 6.448 con doctorado) resulta ser bajo.

| Máximo nivel de formación | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|
| **Técnico Profesional** | 449 | 509 | 710 | 569 | 1.135 | 1.078 | 965 | 1.054 | 1.055 |
| **Tecnólogo** | 506 | 615 | 711 | 820 | 933 | 1.095 | 969 | 1.650 | 2.131 |
| **Profesional** | 25.651 | 30.582 | 31.317 | 37.125 | 45.934 | 44.101 | 43.239 | 43.285 | 43.366 |
| **Especialización** | 26.446 | 31.268 | 30.159 | 35.045 | 34.789 | 36.962 | 36.885 | 43.468 | 44.505 |
| **Maestría** | 18.780 | 21.705 | 21.598 | 23.519 | 25.419 | 27.546 | 27.944 | 41.875 | 43.856 |
| **Doctorado** | 4.526 | 4.994 | 5.370 | 5.649 | 5.961 | 6.358 | 6.808 | 8.893 | 9.477 |

*Tabla 3 Cantidad de docentes por máxima nivel de formación. Datos tomados del "Compendio estadístico de la Educación Superior Colombiana" de Ministerio de Educación, 2016, p.112, Bogotá, Colombia.*

## 5.5 TASA DE DESERCIÓN

El concepto de deserción se trata por primera vez en el año 1897 con una analogía a la Teoría del Suicidio de Durkheim, donde se presenta una similitud entre la sociedad y la institución educativa. En esta teoría se llega a la conclusión de que al igual que en una sociedad, es razonable que los bajos niveles de integración social aumenten la posibilidad de desertar.

Se puede pensar que el concepto de deserción puede conllevar un razonamiento sencillo, pero existe un gran análisis y complejidad que abarca este tema por lo que varios autores, a través de los años, han aportado distintas teorías e ideas que demuestran una vez más la complejidad que lo rodea. Uno de los autores más importantes que ha desarrollado este tema, es Vincent Tinto quien identifica la dificultad de analizar el tema de la deserción puesto que "No hemos sido capaces de convenir que tipos de comportamientos merecen esta denominación de deserción. Existe confusión y contradicción en lo que se refiere al carácter y las causas del abandono en la educación superior". (Vincent Tinto, 1982)

Se define entonces deserción como "una situación en que se enfrenta un estudiante cuando aspira y no logra concluir un proyecto educativo ,considerándose como desertor a aquel individuo que siendo estudiante de una institución de educación superior no presenta actividad académica durante dos semestres académicos consecutivos, lo cual equivale a un año de inactividad académica" (Tinto & Giovagnoli, 1982) .También se define deserción como "El resultado de ausencia de interés que de incapacidad para satisfacer los requisitos del trabajo académico." (Hackman & Dysinger, 1970)

Teniendo en cuenta el momento en que el estudiante deserta de su programa académico, se puede clasificar en tres etapas:

- **Deserción precoz:** cuando un individuo habiendo sido admitido por la Institución de Educación Superior no se matricula.
- **Deserción temprana:** cuando un individuo abandona sus estudios en los primeros semestres del programa.
- **Deserción tardía:** cuando un individuo abandona sus estudios en los últimos semestres.

Para Tinto, existen dos periodos críticos en donde es más probable que un estudiante deserte. El primer momento es cuando un estudiante tiene el primer contacto con la institución, momento en el cual percibe las primeras impresiones sobre las características particulares de la institución, por lo que se está ante la etapa del proceso de admisión. En esta etapa puede pasar que, por ejemplo, por falta de la información adecuada y veraz del programa académico y de la misma institución educativa, el estudiante deserte de manera precoz.

El segundo se presenta durante los primeros semestres del programa académico, en el que el estudiante inicia su proceso de adaptación social y académica mediante el contacto directo con el ambiente universitario. En este momento algunos estudiantes no logran adaptarse de la mejor manera  o simplemente deciden retirarse por razones diferentes a aquellas en las que la institución puede intervenir. En este sentido, puede presentarse la deserción temprana. En este segundo periodo la formación de expectativas equivocadas sobre las condiciones de vida académica y estudiantil en el medio universitario, la ausencia de compatibilidad entre sus intereses o preferencias y las exigencias de la vida académica, pueden conducir al estudiante, a decepciones que tienen como consecuencia la deserción temprana. (Ministerio de Educación, 2015)

En los últimos años de la carrera la deserción tiende a disminuir, puesto que abandonar los estudios en ese momento significa un costo muy alto en términos de la cantidad de dinero que se ha invertido hasta el momento. Por otro lado, algunos estudiantes solo entran a la universidad con la necesidad de cumplir con cierta cantidad determinada de créditos para certificarse o lograr ascensos en el trabajo y para los estudiantes que tienen un trabajo y estudian al mismo tiempo, asistir a la universidad implica adquirir un conjunto de habilidades especificadas en las tareas que desempeñan laboralmente.

Es así como, Tinto clasifica la deserción en dos grandes grupos donde se han ido construyendo conceptos nuevos de nuevos autores:

### 5.5.1 LA DESERCIÓN DESDE EL PUNTO DE VISTA INDIVIDUAL

Desde el punto de vista de un profesor, el abandono se puede definir como un fracaso en completar un programa académico, mientras que, desde el punto de vista del estudiante, se puede interpretar como un paso positivo hacia la consecución de una meta individual. “Rotular comportamientos como abandono con la connotación de fracaso significa en realidad desconocer la importancia de la maduración intelectual y del efecto deseado que se supone tiene la universidad en el proceso de desarrollo intelectual” (Vincent Tinto, 1982) y esto se afirma debido a que solo algunos abandonos de la educación superior son producto de bajo desempeño académico, mientras que la mayor parte de las deserciones son voluntarias.

Desde el punto de vista del estudiante, “un número sorprendentemente grande de estudiantes que ingresan a la institución de educación superior tienen una idea poco clara acerca de las razones por las cuales están allí y no han reflexionado seriamente sobre la elección de la institución. Clarificar la meta hace entonces que los estudiantes abandonen sus estudios o cambien de programa académico” (Vincent Tinto, 1982)

### 5.5.2 LA DESERCIÓN DESDE EL PUNTO DE VISTA INSTITUCIONAL

Desde el punto de vista institucional, todos los sujetos que abandonan una Institución de Educación Superior pueden ser clasificados como desertores. Que un estudiante abandone la universidad genera grandes problemas financieros en la medida en la que este estudiante genera un vacío que pudo ser ocupado por otro estudiante que seguiría en el programa académico

inicial. Esto se ve de manera más significativa en el sector privado debido a que la matrícula constituye una gran parte de sus ingresos institucionales. “El concepto de deserción es complejo en la medida que no está claro que tipos de abandono requieren la misma atención o exigen similares formas de intervención por parte de la institución. Lo difícil para las instituciones consiste en identificar qué tipos de abandono deben ser clasificados como deserciones en el sentido estricto y cuáles como un resultado normal del funcionamiento institucional.” (Vincent Tinto, 1982)

Este modelo que desarrolla Tinto sobre la deserción tiene como base que los estudiantes poseen ciertas características predeterminadas y “Dichas características influencian la delimitación de los objetivos individuales. La experiencia institucional y algunos factores externos interactúan con los objetivos de los estudiantes y determinan el éxito o fracaso de su proceso de adaptación académica y social” (Vincent Tinto, 1982)

A partir del modelo planteado por Tinto varios autores empezaron a realizar modificaciones y aportes nuevos a este tema.  Entre los aportes más importantes y destacados, se encuentran:

- La deserción es el resultado de ausencia de interés que de incapacidad para satisfacer los requisitos del trabajo académico. (Pascarella & Terenzini, 1977)
- La deserción global a nivel del sistema de educación superior se estimó sobre la base de la eficiencia de titulación o eficiencia académica, entendida como la proporción de estudiantes que se titula en un año en comparación con los que ingresan en el año correspondiente a la duración de las carreras. (Rivera, 2005)

- Las etapas más importantes de un estudiante que no completa sus estudios: a) Previa al ingreso: Escasa madurez, percepción poco adecuada del centro de educación superior e inadecuada orientación académica; b) Durante la permanencia en el centro: Situación de shock por el cambio experimentado en el cambio de compañeros y la no adaptación al nuevo sistema educativo, sentimiento de alienación, conductas defensivas, inhibición de la acción, resignación al abandono; c) Posterior al abandono: Etapa moratoria de reflexión , exploración y reorganización, etapa de auto actualización retornando o no a los estudios. (Ministerio de Educación, 2015)

Toda la información acerca de los factores más importantes que influyen en los estudiantes al desertar fue agrupada de la siguiente manera:

| DETERMINANTE | AUTORES | RAZONES |
|---|---|---|
| INDIVIDUALES | Spady(1970)<br>Bean (1980)<br>Nora y Matonak(1990)<br>Brunsden (2000) | Edad, género y estado civil<br>Posición dentro de los hermanos<br>Entorno familiar<br>Calamidad y problemas de saluda<br>Integración social<br>Incompatibilidad horaria con las actividades extra académicas<br>Expectativas no satisfechas<br>Embarazo |
| ACADÉMICOS | Spady(1970)<br>Tinto (1975) | Orientación profesional<br>Tipo de colegio<br>Rendimiento académico<br>Calidad del programa<br>Métodos de estudio<br>Resultado de los exámenes de ingreso<br>Insatisfacción con el programa<br>Numero de materias |
| INSTITUCIONALES | Ade Iman (1999) | Normalidad académica<br>Becas y formatos de financiamiento<br>Recursos universitarios<br>Orden publico<br>Entorno político<br>Nivel de interacción personal con los programa<br>Apoyo académico<br>Apoyo psicológico |
| SOCIOECONOMICOS | Tinto(1975)<br>Gaviria (2002) | Estrato<br>Situación laboral<br>Situación laboral de los padres e ingresos<br>Dependencia económica<br>Personas a cargo<br>Nivel educativo de los padres<br>Entorno macroeconómico del país |

*Tabla 4 Determinantes de deserción según estudios. Datos tomados de "Deserción estudiantil en la educación superior Colombiana" del Ministerio de Educación Nacional, 2009, pp. 27, Bogotá, Colombia.*

Aunque el modelo de Tinto sirvió de base para el análisis de la deserción realizada por varios autores, otros afirmaron que los enfoques tradicionales describen por qué el estudiante decide abandonar sus estudios, pero no permite explicar el proceso de su deserción. Autores como Desjardins, Ahlburg y McCall (1999) le dieron un enfoque denominado historia de eventos, donde se obtiene una descripción y explicación del proceso de deserción, puesto que este análisis permite seguir la variable dependiente hasta que ocurra el evento de interés.

"En estas investigaciones se incluye la dimensión dinámica del proceso de deserción y se compara la probabilidad de abandonar los estudios en cada periodo concluyendo que la posibilidad de desertar o de graduarse no es constante a lo largo del tiempo. La estructura conceptual del proceso comprende cuatro posibles resultados de interés en cada periodo observado: Suspender los estudios por un tiempo y luego regresar, desertar, graduarse o seguir estudiando. Cada uno de estos resultados son afectados por variables exógenas tanto tiempos variantes como estáticas y aunque los valores de estas últimas son constantes en el tiempo, el efecto que tienen en la decisión de Abandono cambia, por lo que son necesarias "(Ministerio de Educación, 2015)

### 5.5.3 DESERCIÓN EN MODELOS DE EDUCACIÓN A DISTANCIA

Aunque el modelo de Tinto es el más utilizado para estudiar la deserción estudiantil, "este modelo no es aplicable para sistemas de educación no tradicionales como la educación a distancia lo que incluye la idea de separación geográfica entre profesores y estudiantes "(David Kember, 1989)

El estudio de la deserción a distancia ha sido importante en países como Estados Unidos, Inglaterra y Australia. Los estudiantes que por lo general reciben educación a distancia son alumnos adultos, estudiantes en tiempo parcial, trabajadores de tiempo completo, con responsabilidades familiares y que viven en zonas rurales y alejadas. Los factores más importantes para culminar un programa de pregrado a distancia se refieren más a las características personales, tipos de programa y el soporte y apoyo que la institución le brinda al estudiante.

Los factores más importantes para explicar la deserción de estos estudiantes son: i) carencia de tiempo, ii) escasa tutoría, iii) poca información sobre el proceso de enseñanza-aprendizaje, iv) falta de soporte y v) dificultad de comunicación con las instituciones.

Para analizar este problema, Jemer plantea “un modelo longitudinal para estudiar dichos factores. En dicho modelo se incluyen aspectos de motivación y componentes de integración académica y social. A diferencia del modelo propuesto por Tinto, la integración social se refiere a la capacidad del estudiante para alternar el estudio con la familia, el trabajo y la sociedad” (Ministerio de Educación, 2015)

## 5. 6 ASPECTOS GENERALES DE PYTHON

Python es un lenguaje de programación multiparadigma que soporta orientación a objetivos, programación imperativa y en menor medida, programación funcional. Es administrado por la Python Software Foundation y posee licencia de código abierto, denominada Python Software Foundation License.

Python posee ciertos paquetes que ayudan al manejo de los datos y creación de códigos. Uno de los más importantes es Pandas Basics, una herramienta de manipulación de datos de alto nivel desarrollada por Wes McKinney. Es construido con otro paquete importante, NumPy y su estructura de datos claves es llamada DataFrame.

Además de estos dos, para este trabajo se utilizaran las siguientes herramientas en Python:

1. La biblioteca Scikit Learn: utilizada para el aprendizaje automático de software libre. Entre sus principales paquetes, incluye varios algoritmos de clasificación, regresión y análisis de grupos. Está diseñada para operar de la mano de las bibliotecas numéricas y científicas Numpy y Scipy.
2. La biblioteca Matplotlib: utilizada para la generación de gráficos a partir de datos en Python en la extensión matemática de NumPy.

### 5.6.1 ANÁLISIS DE CORRELACIÓN DE VARIABLES

El análisis de correlación de variables consiste en un procedimiento estadístico para determinar si dos o más variables están relacionadas o no. El resultado de este análisis es el coeficiente de correlación que puede tomar valores entre -1 y +1. Un signo positivo indica que hay una relación positiva entre las variables, es decir que mientras una variable incrementa la otra también.

Más allá de que R es una buena herramienta de análisis, tiene algunas desventajas: Los coeficientes de correlación más utilizados solo miden una relación lineal. Por lo tanto, es posible que, si bien existe una fuerte relación no linean entre las variables, r podría estar cerca de 0 o igual a 0. En tal caso, un diagrama de dispersión puede indicar aproximadamente la existencia o no de una relación lineal.

### 5.6.2 MÉTODOS DE CLASIFICACIÓN

Para los métodos de clasificación, cabe conocer los tres tipos de algoritmos de aprendizaje automático que existen: a) supervisado, B) no supervisado y c) por refuerzo

- **Aprendizaje supervisado:** la máquina se enseña con el ejemplo. De esta manera, el operador proporciona al algoritmo de aprendizaje automático un conjunto de datos conocidos que incluye las entradas y salidas deseadas, y el algoritmo debe encontrar un método para determinar cómo llegar a esas entradas y salidas.

  Mientras que el operador conoce las respuestas correctas del problema, el algoritmo identifica los patrones en los datos, aprende de las observaciones y hace predicciones. El algoritmo realiza predicciones y es corregido por el operador hasta que alcanza un alto nivel de precisión y rendimiento.

- **Aprendizaje sin supervisión: el** algoritmo de aprendizaje automático estudia los datos para identificar patrones. No hay una clave de respuesta o un operador humano para proporcionar la instrucción sino que la maquina determina las correlaciones y las relaciones mediante el análisis de datos disponible.
- **Aprendizaje por refuerzo:** se centra en los procesos de aprendizajes reglamentados en las que se proporcionan algoritmos de aprendizaje automáticos con un conjunto de acciones, parámetros y valores finales. Al definir las reglas, el algoritmo de aprendizaje automático intenta explorar las diferentes opciones y posibilidades, monitorizando y evaluando cada resultado para determinar cuál es el optimo

### 5.6.3 CONJUNTO DE ENTRENAMIENTO Y PRUEBA

Para un buen ejercicio de clasificación contando con la suficiente información, se divide el conjunto de datos en dos subconjuntos de Entrenamiento y Prueba.

- **Conjunto de entrenamiento:** es el subconjunto que se utiliza para entrenar el modelo
- **Conjunto de prueba:** es el subconjunto que se utiliza para probar el modelo entrenado.

Es importante asegurarse de que el conjunto de prueba sea lo suficientemente grande para generar resultados significativos y que sea representativo para todos los conjuntos de datos. No se debe elegir un conjunto de prueba con características diferentes al conjunto de entrenamiento.

### 5.6.4 MUESTREO ESTRATIFICADO

El muestreo estratificado es un procedimiento de muestreo en el que el objetivo de la población se separa en segmentos exclusivos, homogéneos (estratos) y luego una muestra aleatoria simple

se selecciona de cada segmento. Las muestras seleccionadas de los diversos estratos se combinan en una sola muestra.

### 5.6.5 SEMILLA ALEATORIA

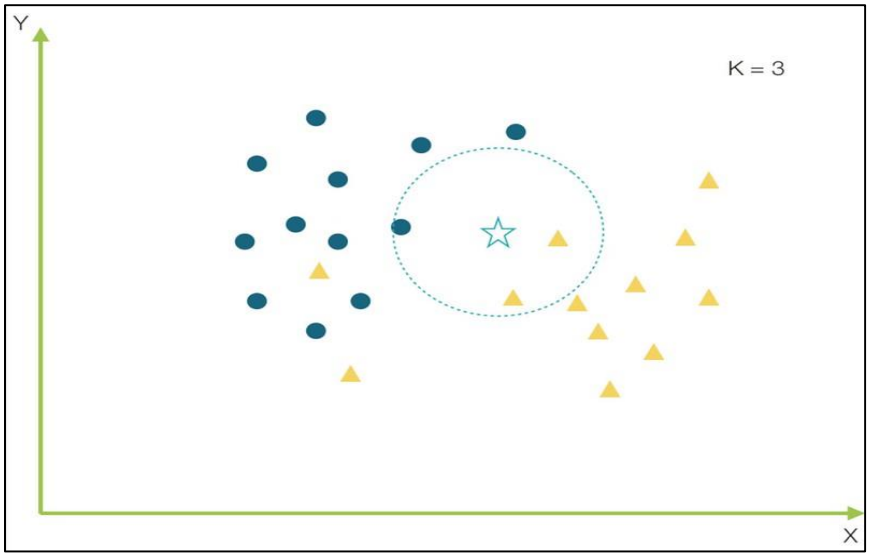
Una semilla aleatoria es un número utilizado para inicializar un generador de números pseudoaleatorios

### 5.6.7 K VECINOS MÁS CERCANOS

K Vecinos más cercanos es uno de los algoritmos de clasificación más básicos en Machine Learning. Pertenece al dominio del aprendizaje supervisado y encuentra una aplicación intensa en el reconocimiento de patrones, la minería de datos y la detección de intrusos.

Entre sus características más importantes se encuentran:

- No hace suposiciones explicitas sobre la forma funcional de los datos, evitando modelar mal la distribución subyacente de los datos.
- Es un modelo con aprendizaje basado en la instancia, lo que significa que nuestro algoritmo no aprende explícitamente un modelo. Lo que hace este modelo es memorizar las instancias de formación que posteriormente se utilizan como “conocimiento” para la fase de predicción.




*Grafica 9 Representación de K Vecinos más cercanos.*

Suponiendo que la estrella es el punto el cual se quiere predecir. Primero, se encuentra el punto más cercano a la estrella y luego se clasifican los puntos para el voto mayoritario de sus vecinos K. Cada objeto vota por su clase y la clase con más votos se toma como la predicción. Para encontrar los puntos similares más cercanos, se encuentra la distancia entre puntos utilizando medidas de distancias. Para las medidas de distancia se tienen:

- Distancia Euclidiana: $\sqrt{\sum_{i=1}^{k}(X_i - Y_i)^2}$

- Distancia Manhattan: $\sum_{i=1}^{k}|X_i - Y_i|$
- Distancia Minkowsky : $\left[\sum_{i=1}^{k}(|X_i - Y_i|)\right]^4$

El número de vecinos (K) es un hiperparámetro que se debe elegir en el momento de la construcción del modelo. Se puede pensar en K como una variable de control para el modelo de predicción.

El algoritmo para KNN tiene los siguientes pasos:

1. Calcular la distancia
2. Encontrar los vecinos más cercanos
3. Votar por las etiquetas

### 5.6.8 BOSQUES ALEATORIOS

Los bosques aleatorios comprenden otro algoritmo de aprendizaje supervisado y puede utilizarse tanto para clasificación como para la regresión.

Un bosque, como su nombre lo indica está compuesto de árboles. Los bosques aleatorios crean arboles de decisión a partir de una muestra de datos seleccionadas al azar, obteniendo predicciones de cada árbol y selecciona la mejor solución mediante votación.

Técnicamente es un método de conjunto basado en el enfoque de "dividir y conquistar" de árboles de decisión generados en un conjunto de datos dividido al azar. Los arboles de decisión individuales se generan utilizando un indicador de selección de atributos tales como la ganancia de información, la relación de ganancia y el índice Gini.

Este algoritmo funciona mediante los siguientes pasos:

1. Construir un árbol de decisión para cada muestra y obtener un resultado de predicción de cada árbol de decisión
2. Realizar una votación por cada resultado previsto
3. Seleccionar el resultado de la predicción con más votos como la predicción final.

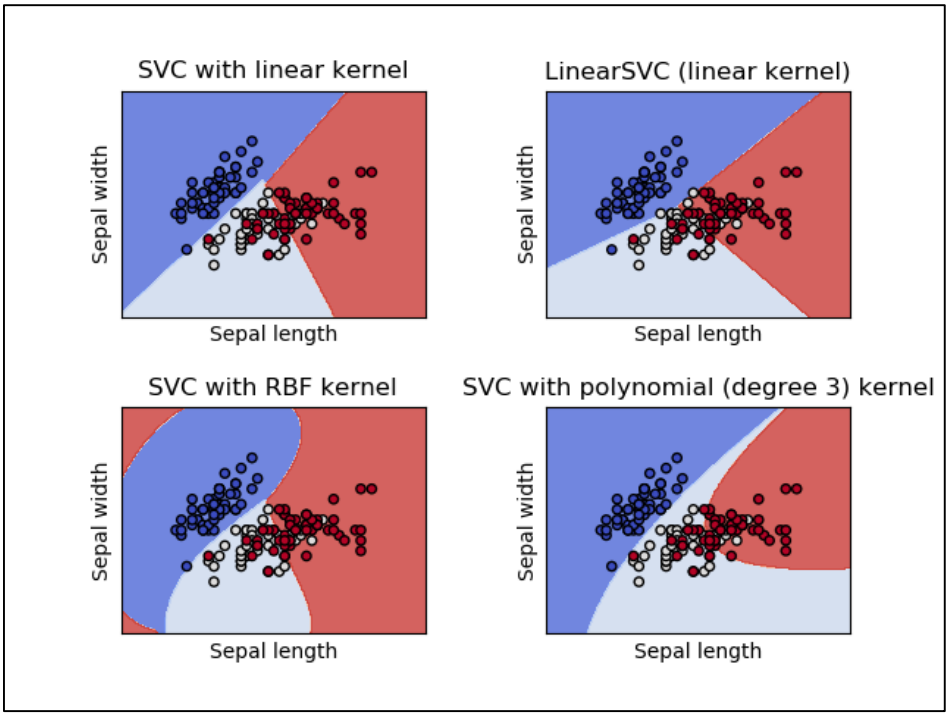
Algunas de las ventajas de los Bosques Aleatorios son:

- Se consideran un método muy preciso y robusto debido al número de árboles de decisión que participan en el proceso.
- No sufre el problema de sobreajuste dado que toma el promedio de todas las predicciones, lo que anula los sesgos.
- El bosque aleatorio también puede manejar valores faltantes.

### 5.6.9 MÁQUINA DE SOPORTE VECTORIAL

Una máquina de soporte vectorial aprende la superficie de decisión de dos clases distintas de los puntos de entrada. Como un clasificador de una clase, la descripción dada por los datos de los vectores de soporte es capaz de formar una frontera de decisión alrededor del dominio de los datos de aprendizaje con poco conocimiento de los datos de esta frontera.

Los datos son mapeados por medio de un Kernel Gaussiano u otro tipo de Kernel a un espacio de características, en un espacio dimensional más alto, donde se busca la máxima separación de los datos.




*Grafica 10 Representación de Maquinas de Soporte Vectorial para distinto tipo de Kernel.*

### 5.6.10 MATRIZ DE CONFUSIÓN

La matriz de confusión es una herramienta que permite identificar la precisión de un modelo teniendo en cuenta sus predicciones luego de ser entrenado. La matriz tiene la siguiente estructura:

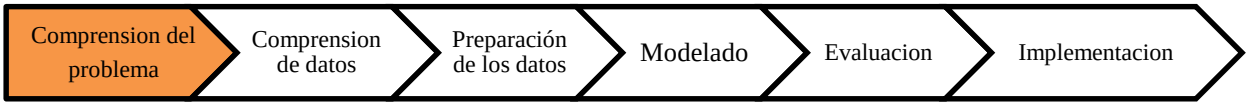
| | | **Predicción** | |
|---|---|---|---|
| | | **Positivos** | **Negativos** |
| **Observación** | **Positivos** | Verdaderos Positivos (VP) | Falsos Negativos (FN) |
| | **Negativos** | Falsos Positivos (FP) | Verdaderos Negativos (VN) |

*Grafica 11 Representación teórica de una matriz de confusión*

- **VP** es la cantidad de positivos que fueron clasificados correctamente como positivos en el modelo.
- **VN** es la cantidad de negativos que fueron clasificados correctamente como negativos en el modelo.
- **FN** es la cantidad de positivos que fueron clasificados incorrectamente como negativos.
- **FP** es la cantidad de negativos que fueron clasificados incorrectamente como positivos.

## 6. METODOLOGIA DE LA INVESTIGACION

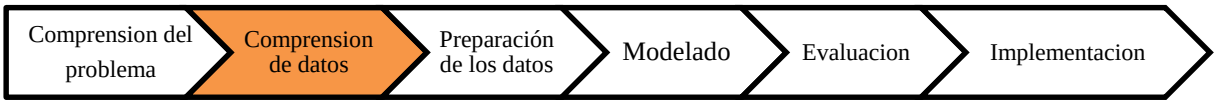
El trabajo se basa en identificar cómo es el comportamiento del fenómeno de la deserción en los estudiantes de la Escuela Colombiana de Ingeniería mediante técnicas de análisis de datos. Utilizando la metodología CRISP-DM, el proceso se puede organizar de la siguiente manera:




### 6.1 FASE DE COMPRENSIÓN DEL PROBLEMA

En esta primera etapa se busca comprender la situación actual del fenómeno de la deserción en la universidad, con el fin de determinar los objetivos y el plan del proyecto. Las principales tareas a desarrollar en esta etapa son las siguientes:

a) **Determinar los objetivos del problema:** plantear los objetivos del proyecto teniendo en cuenta la necesidad de utilizar análisis de datos y definir los criterios de éxito.
b) **Evaluación de la situación:** plantear una situación actual del fenómeno de la deserción, teniendo en cuenta datos estadísticos y trabajos previos que se hayan realizado. Determinar si se tiene o no buen conocimiento del problema, si existen los datos necesarios para llevarse a cabo y que beneficio hay de realizar el trabajo de investigación.
c) **Determinación de los objetivos de Análisis de Datos**: interpretar los objetivos del proyecto en término de Análisis de datos.
d) **Producción de un plan del proyecto:** identificar las técnicas a utilizar y los pasos a seguir para poder llevar a cabo el trabajo de investigación.




## 6. 2 FASE DE COMPRENSIÓN DE DATOS

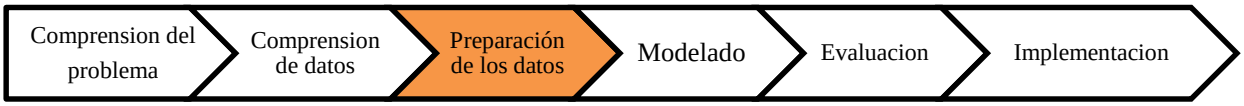
Comprende la recolección inicial de los datos para tener un primer conocimiento del tema para tener un primer punto de vista. Las principales tareas a desarrollar en esta etapa son las siguientes:

a) **Recolección de datos iniciales:** recolectar toda la información posible sobre la situación académica de los estudiantes teniendo en cuenta los parámetros que más se relacionan con la deserción.

b) **Descripción de los datos:** descripción de los datos iniciales. Se establece entonces el volumen de los datos (Campos y número de registros), identificación, significado de cada campo y descripción del formato.

c) **Exploración de los datos:** el fin de esta tarea es encontrar una estructura general de los datos iniciales. Para esto resulta apropiado, utilizar las tablas de frecuencias y distribución.

d) **Verificación de la calidad de los datos:** se busca eliminar el ruido de los datos iniciales, identificando los valores nulos o valores incoherentes.




## 6.3 FASE DE PREPARACIÓN DE LOS DATOS

En esta etapa se busca preparar los datos y adaptarlos según convenga. Esta preparación de los datos incluye tareas generales de selección de datos, limpieza de datos, generación de nuevos campos y cambios de formato. Las tareas involucradas en esta fase son las siguientes:

a) **Selección de datos:** se selecciona entonces un grupo de datos teniendo en cuenta la verificación de la calidad de los datos.

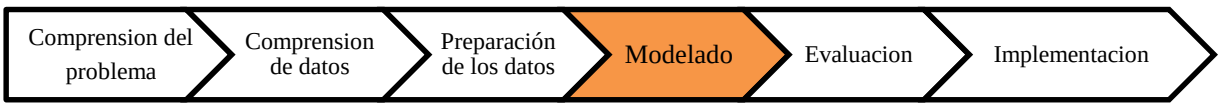
b) **Limpieza de los datos:** para esta tarea se requiere optimizar la calidad de los datos para prepararlos para la siguiente fase de modelado. Entre estas técnicas están: la

normalización de los datos, manejo de los campos numéricos, tratamiento de valores nulos, reducción de volumen de datos, entre otros.

c) **Estructuración de los datos:** incluyendo operación de preparación de datos tales como integración de nuestros registros, transformación de valores para atributos existentes, entre otros.

d) **Integración de los datos:** involucra la creación de nuevas estructuras a partir de los datos seleccionados. Entre estas están la creación de nuevos registros fusión de tablas, tablas resumen, entre otros.

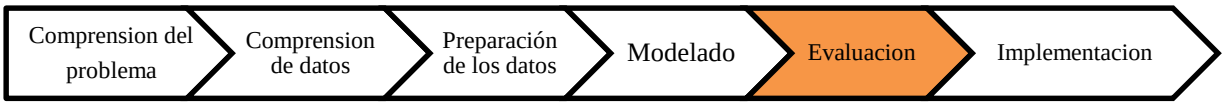
e) **Formateo de los datos:** consiste en la realización de trasformaciones sin modificar el significado con el fin de emplear la Técnica de Análisis de Datos escogida en particular. Entre estas, se encuentran: la reordenación de campos, ajuste de valores de un campo en específico, eliminación de comas y caracteres especiales.




## 6.4 FASE DE MODELADO

En esta etapa se busca la técnica más apropiada para el análisis de datos. Este tiene que estar en función de los siguientes criterios: ser apropiada al problema, disponer de los datos adecuados, cumplir con los requisitos del problema, tiempo adecuado para obtener un modelo y conocimiento de la técnica. Las tareas involucradas en esta fase son las siguientes:

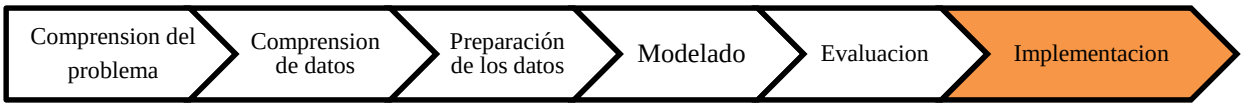
a) **Selección de la técnica de modelado:** teniendo en cuenta los objetivos del trabajo de investigación, se selecciona la Técnica de Análisis de Datos apropiada para el problema a resolver.
b) **Generación del plan de prueba:** seleccionada la técnica apropiada de análisis de datos se debe generar un procedimiento para probar la validez de este.
c) **Construcción del modelo:** ejecución del modelo sobre los datos previamente seleccionados.
d) **Evaluación del modelo:** se determina si el modelo cumple con los objetivos planteados anteriormente.




## 6.5 FASE DE EVALUACIÓN

En esta fase se evalúa el modelo teniendo en cuenta los objetivos y criterios de éxito del problema. Las tareas involucradas en esta fase son las siguientes:

a) **Evaluación de resultados:** evaluación del modelo en relación con los objetivos y buscar si hay maneras de mejorar o identificar sus restricciones.
b) **Proceso de revisión:** revisión de todo el proceso de análisis de datos para identificar que objetos pueden ser mejorados.
c) **Determinación de futuras fases:** determinar si se han cumplido los objetivos. Si no es así, identificar fallos y cambiar los parámetros del modelo.




## 6.6 FASE DE IMPLEMENTACIÓN

Una vez que el modelo se haya validado, se analiza y se arrojan los resultados respectivos en un informe final.

# 7. PROCESO INVESTIGATIVO

## 7.1 ANÁLISIS DEL DATAFRAME

El dataframe a analizar es el compilado de los registros históricos en cuanto a matriculas de estudiantes de La Escuela Colombiana de Ingeniería – Julio Garavito durante el periodo comprendido entre 2018-1 a 2020-1.

De forma general, el dataframe es un archivo en formato Excel que contiene un total de 9 columnas nombradas así: ID, Periodo, Estado, Sexo, Ciudad, Edad y Estado final. Así como también, un total de 5.109 filas con 1.465 datos únicos para la columna principal ID.

**ID** se refiere a la identificación de cada uno de los estudiantes

**PERIODO** se refiere al año en que este estudiante cambia su estado en las oficinas de registro de La Escuela. Los periodos son: 2018-1, 2018-2, 2019-1, 2019-2 ,2020-1.

**ESTADO** se refiere a la categoría que tiene un estudiante en La Escuela teniendo en cuenta su condición académica. Los estados pueden ser:

a. Anulado
b. Convenio Act Col
c. Convenio Act  Univ
d. Convenio Act Col
e. Transferencia
f. Readmitido
g. Terminó estudios
h. Reintegro
i. Reintegro Prueba
j. Reintegro 2 Prueba
k. Prueba
l. 2da Prueba Cons
m. Seguimiento Académico
n. Admit
o. Nuevo
p. Registro
q. Matric
r. Nuevo
s. Cancelado

Las tres últimas categorías de la variable ESTADO son las más importantes y en las que se hará mayor énfasis a lo largo del desarrollo del presente trabajo.

**SEXO** se refiere al género de cada estudiante: Masculino (M) y femenino (F).

**CIUDAD** se refiere a su ciudad de origen: 96 ciudades de Colombia.

**EDAD** se refiere a la edad del estudiante para el presente año 2020.

**ESTADO FINAL** se refiere a un compilado final de la variable estado teniendo en cuenta dos categorías principales: a. cancelado y b) registro.

## 7.2 TRATAMIENTO PREMILIMINAR DE LOS DATOS EN EXCEL

La fase preliminar de tratamiento consiste en la formulación de la variable "ESTADO FINAL "mediante fórmulas en Excel que tienen como fin, determinar para el presente año 2020, en qué condición se encuentra el estudiante en La Escuela. Para esto se definen dos categorías:

- "**CANCELADO":** el estudiante abandonó sus estudios en alguno de los 5 periodos transcurridos.
- **"REGISTRO":** el estudiante se ha registrado en La Escuela en el último periodo 2020-1

Con esta nueva columna en nuestro Dataframe se procede al desarrollo del código en Python.

## 7.2 CARGA Y CONVERSIÓN DE VARIABLES EN EL DATAFRAME

Para comenzar, se realiza la carga del archivo de Excel en el NoteBook de Yupiter, una aplicación en línea que permite el manejo de información en forma de códigos para Python.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt


datos = pd.read_excel('./Desktop/TRABAJO FINAL/Proyecto_datos.xlsx')
datos
```

El archivo se carga en forma de tabla y se muestra inicialmente de la siguiente manera, identificando la cantidad de filas y columnas en total:

| | ID | Periodo | Estado | Sexo | Ciudad | Edad | ESTADO FINAL |
|---|---|---|---|---|---|---|---|
| **0** | 8963611 | 2018-1 | Reintegro | F | BOGOTA | 34 | CANCELADO |
| **1** | 8963611 | 2019-2 | Reintegro-Prueba | F | BOGOTA | 34 | CANCELADO |
| **2** | 8980017 | 2018-1 | Registro | M | ZIPAQUIRA | 31 | CANCELADO |
| **3** | 8982244 | 2018-1 | Seg.Académico | F | BOGOTA | 31 | CANCELADO |
| **4** | 9003525 | 2018-1 | Seg.Académico | F | BOGOTA | 29 | CANCELADO |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **5104** | 13409463 | 2018-1 | Registro | F | BOGOTA | 26 | CANCELADO |
| **5105** | 13409463 | 2018-2 | Registro | F | BOGOTA | 26 | CANCELADO |
| **5106** | 13409463 | 2019-1 | Registro | F | BOGOTA | 26 | CANCELADO |
| **5107** | 13436958 | 2019-2 | Registro | M | BOGOTA | 26 | REGISTRO |
| **5108** | 13436958 | 2020-1 | Reintegro | M | BOGOTA | 26 | REGISTRO |

5109 rows × 7 columns

*Tabla 5 Dataframe original.*

Para facilitar el manejo de los datos, se hace un cambio en el nombre de la columna “ESTADO FINAL”, puesto que el espacio entre palabras podría afectar la formulación. Para esto hacemos uso de la función rename que modifica el nombre una columna en específico y la remplaza en el Data Frame original, quedando de la siguiente manera:

```
datos.rename(columns= {'ESTADO FINAL': 'Estado_final'}, inplace = True )

datos.head ()
```

| | ID | Periodo | Estado | Sexo | Ciudad | Edad | Estado_final |
|---|---|---|---|---|---|---|---|
| **0** | 8963611 | 2018-1 | Reintegro | F | BOGOTA | 34 | CANCELADO |
| **1** | 8963611 | 2019-2 | Reintegro-Prueba | F | BOGOTA | 34 | CANCELADO |
| **2** | 8980017 | 2018-1 | Registro | M | ZIPAQUIRA | 31 | CANCELADO |
| **3** | 8982244 | 2018-1 | Seg.Académico | F | BOGOTA | 31 | CANCELADO |
| **4** | 9003525 | 2018-1 | Seg.Académico | F | BOGOTA | 29 | CANCELADO |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **5104** | 13409463 | 2018-1 | Registro | F | BOGOTA | 26 | CANCELADO |
| **5105** | 13409463 | 2018-2 | Registro | F | BOGOTA | 26 | CANCELADO |
| **5106** | 13409463 | 2019-1 | Registro | F | BOGOTA | 26 | CANCELADO |
| **5107** | 13436958 | 2019-2 | Registro | M | BOGOTA | 26 | REGISTRO |
| **5108** | 13436958 | 2020-1 | Reintegro | M | BOGOTA | 26 | REGISTRO |

5109 rows × 7 columns

*Tabla 6 Dataframe con cambio de variable ESTADO FINAL.*

.

Teniendo el DataFrame listo, inicia la transformación de las variables. Al revisar el tipo de variables que contiene el Dataset encontramos que las variables que tenemos que pasar a Categóricas Numéricas son: el Periodo, Estado, Sexo, Ciudad. El tipo de variable para cada una es:

```
ID              int64
Periodo        object
Estado         object
Sexo           object
Ciudad         object
Edad            int64
Estado_final   object
dtype: object
```

*Tabla 7 Tipos de variable en el Dataframe.*

Para el cambio de variables a categóricas numéricas. Usamos el código cat.codes que asigna un valor numérico a cada una de las categóricas de las variables. Para identificar qué código es asignado a cada una de las categóricas hacemos lo siguiente:

```
datos.Periodo = datos.Periodo.astype ('category')

d = dict (enumerate (datos.Periodo.cat.categories))

print (d)

{0: '2018-1', 1: '2018-2', 2: '2019-1', 3: '2019-2', 4: '2020-1'}


datos.Estado_final = datos.Estado_final.astype ('category')

d = dict (enumerate (datos.Estado_final.cat.categories))

print (d)

{0: 'CANCELADO', 1: 'REGISTRO'}


datos.Estado = datos.Estado.astype ('category')

d = dict (enumerate (datos.Estado.cat.categories))

print (d)

{0: '2da Prueba Cons', 1: 'Admit. Nuevo', 2: 'Anulado', 3: 'Cancelado', 4: 'Convenio Act Col', 5: 'Convenio Act Univ', 6: 'Matric. Nuevo', 7: 'Prueba', 8: 'Readmitido', 9: 'Registro', 10: 'Reintegro', 11: 'Reintegro 2da Prueba', 12: 'Reintegro-Prueba', 13: 'Seg.Académico', 14: 'Termino estudios', 15: 'Transferencia'}


datos.Sexo = datos.Sexo.astype ('category')

d = dict (enumerate (datos.Sexo.cat.categories))

print (d)

{0: 'F', 1: 'M'}


datos.Ciudad = datos.Ciudad.astype ('category')

d = dict (enumerate (datos.Ciudad.cat.categories))

print (d)

{0: 'ACACIAS', 1: 'AGUAZUL', 2: 'ANAPOIMA', 3: 'ARAUQUITA', 4: 'ARMENIA', 5: 'BARRANCABERMEJA', 6: 'BARRANQUILLA', 7: 'BOGOTA', 8: 'BOJACA', 9: 'BOYACA', 10: 'BUCARAMANGA', 11: 'CAJICA', 12: 'CALDAS', 13: 'CALI', 14: 'CAQUEZA', 15: 'CARTAGENA', 16:
```

'CARTAGO', 17: 'CHIA', 18: 'CHIQUINQUIRA', 19: 'CHOACHI', 20: 'CHOCONTA', 21: 'COGUA', 22: 'CUCUTA', 23: 'CUMARAL', 24: 'DUITAMA', 25: 'EL COLEGIO', 26: 'EL ROSAL', 27: 'FACATATIVA', 28: 'FLANDES', 29: 'FLORENCIA', 30: 'FLORIDABLANCA', 31: 'FOMEQUE', 32: 'FUNZA', 33: 'FUSAGASUGA', 34: 'GACHANCIPA', 35: 'GACHETA', 36: 'GARAGOA', 37: 'GARZON', 38: 'GIRARDOT', 39: 'GRANADA', 40: 'GUADUAS', 41: 'GUASCA', 42: 'GUATEQUE', 43: 'IBAGUE', 44: 'IPIALES', 45: 'LA CALERA', 46: 'LA DORADA', 47: 'LA MESA', 48: 'LA PALMA', 49: 'LA VEGA', 50: 'LENGUAZAQUE', 51: 'MADRID', 52: 'MAICAO', 53: 'MANIZALES', 54: 'MEDELLIN', 55: 'MEDINA', 56: 'MELGAR', 57: 'MIRAFLORES', 58: 'MOCOA', 59: 'MONIQUIRA', 60: 'MONTERIA', 61: 'NEIVA', 62: 'NEMOCON', 63: 'OCAÑA', 64: 'OTRA', 65: 'PACHO', 66: 'PALMIRA', 67: 'PASTO', 68: 'PEREIRA', 69: 'PITALITO', 70: 'QUIBDO', 71: 'RAMIRIQUI', 72: 'RIOHACHA', 73: 'SAN JUAN DE RIOSECO', 74: 'SANTA MARIA', 75: 'SANTA MARTA', 76: 'SIMIJACA', 77: 'SINCELEJO', 78: 'SOGAMOSO', 79: 'SOPO', 80: 'SUESCA', 81: 'SUTATAUSA', 82: 'TABIO', 83: 'TENJO', 84: 'TOCAIMA', 85: 'TOCANCIPA', 86: 'TUNJA', 87: 'UBATE', 88: 'VALLEDUPAR', 89: 'VELEZ', 90: 'VILLANUEVA', 91: 'VILLAPINZON', 92: 'VILLAVICENCIO', 93: 'VILLETA', 94: 'YOPAL', 95: 'ZIPAQUIRA'}

Posteriormente, se convierte cada variable en numérica con la generación de sus respectivos códigos números de la siguiente manera:

```
datos.Periodo = datos.Periodo.astype ('category').cat.codes

datos.Estado_final = datos.Estado_final.astype ('category').cat.codes

datos.Estado = datos.Estado.astype ('category').cat.codes

datos.Sexo = datos.Sexo.astype ('category').cat.codes

datos.Ciudad = datos.Ciudad.astype ('category').cat.codes
```

Con este código, a cada una de las categorías se le asigna un valor numérico transformado el DataFrame original y cada una de las variables en:

| | |
|---|---|
| ID | int64 |
| Periodo | int8 |
| Estado | int8 |
| Sexo | int8 |
| Ciudad | int8 |
| Edad | int64 |
| Estado_final | int8 |
| dtype: object | |

*Tabla 8 Tipos de variable después del cambio a numéricas.*

| | ID | Periodo | Estado | Sexo | Ciudad | Edad | Estado_final |
|---|---|---|---|---|---|---|---|
| **0** | 8963611 | 0 | 10 | 0 | 7 | 34 | 0 |
| **1** | 8963611 | 3 | 12 | 0 | 7 | 34 | 0 |
| **2** | 8980017 | 0 | 9 | 1 | 95 | 31 | 0 |
| **3** | 8982244 | 0 | 13 | 0 | 7 | 31 | 0 |
| **4** | 9003525 | 0 | 13 | 0 | 7 | 29 | 0 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **5104** | 13409463 | 0 | 9 | 0 | 7 | 26 | 0 |
| **5105** | 13409463 | 1 | 9 | 0 | 7 | 26 | 0 |
| **5106** | 13409463 | 2 | 9 | 0 | 7 | 26 | 0 |
| **5107** | 13436958 | 3 | 9 | 1 | 7 | 26 | 1 |
| **5108** | 13436958 | 4 | 10 | 1 | 7 | 26 | 1 |

5109 rows × 7 columns

*Tabla 9 Dataframe transformado con variables numericas.*

## 7.3 ANÁLISIS DE CORRELACIÓN ENTRE VARIABLES

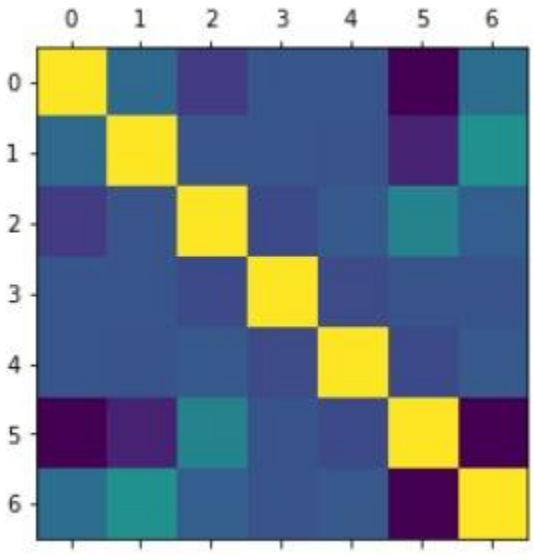
Con cada una de las columnas en su respectivo formato, se puede realizar un análisis de correlaciones. Este análisis muestra en qué proporción una variable está relacionada con otra de manera matemática.

La manera correcta de mostrar esta relación es por medio de una matriz de correlación que se puede ilustrar e interpretar según un rango de colores.

```
datos.corr ()

plt.matshow (datos.corr ())
```

| | ID | Periodo | Estado | Sexo | Ciudad | Edad | Estado_final |
|---|---|---|---|---|---|---|---|
| **ID** | 1.000000 | 0.103443 | -0.122778 | 0.004299 | 0.005752 | -0.356435 | 0.130512 |
| **Periodo** | 0.103443 | 1.000000 | 0.000665 | 0.003751 | -0.008294 | -0.226013 | 0.330958 |
| **Estado** | -0.122778 | 0.000665 | 1.000000 | -0.052183 | 0.025797 | 0.250081 | 0.047078 |
| **Sexo** | 0.004299 | 0.003751 | -0.052183 | 1.000000 | -0.044257 | -0.007633 | -0.006488 |
| **Ciudad** | 0.005752 | -0.008294 | 0.025797 | -0.044257 | 1.000000 | -0.052720 | 0.021008 |
| **Edad** | -0.356435 | -0.226013 | 0.250081 | -0.007633 | -0.052720 | 1.000000 | -0.358907 |
| **Estado_final** | 0.130512 | 0.330958 | 0.047078 | -0.006488 | 0.021008 | -0.358907 | 1.000000 |

*Tabla 10 Matriz de correlación de Variables.*




*Grafica 12 Representación gráfica de la matriz de correlaciones*

Empezando desde la columna 0 “ID”, la matriz de correlación es una matriz simétrica que muestra la relación entre dos variables en general. Con el DataFrame y los resultados mostrados, se puede evidenciar que no existe una relación entre ninguna de las variables, siendo la relaciones máximas 33.0958 % (Variables “Estado” y “Estado_final”) y 25.0081% (Variables Estado y Edad).

La primera relación no es significativa en el sentido que la variable “Estado_final” es una recopilación de la variable “Estado”, por lo que la relación de cierta manera es implícita. La

relación entre la variable Edad y Estado puede interpretarse como una pequeña probabilidad de que para cierta edad un estudiante este en una condición especifica en la universidad. De igual forma, ninguna de estas relaciones da la información suficiente para sacar una conclusión sólida de los datos analizados.

Realizado el análisis preliminar de los datos, se procede con la implementación de los métodos de clasificación, en donde se utilizarán: K Vecinos más cercanos (KN), Máquinas de Soporte Vectorial (SVM) y Bosques aleatorios (RF)

## 7.4 MÉTODOS DE CLASIFICACIÓN

Para el desarrollo de los métodos de clasificación, se deben definir dos parámetros principales: a) las variables de entrada o variables predictoras y b) la variable de salida u objetivo a predecir. Se trabajarán dos casos dependiendo de las variables que se van a utilizar del Dataframe.

**CASO 1:** Las variables de entrada serán la Edad y el Estado, específicamente, aquellos estudiantes que se hayan matriculado por primera vez durante los 5 periodos académicos y la variable de salida será el Estado_final.

**CASO 2:** Las variables de entrada serán la Edad, Género y Ciudad y la variable de salida será el Estado_Final, cubriendo la información de todos los estudiantes.

### 7.4.1 CLASIFICACIÓN- CASO 1

Para empezar, se debe hacer el filtrado de los datos con el fin de manejar solo los estudiantes que se hayan matriculado en alguno de los 5 periodos presentes en el Dataframe. Para eso se filtran aquellos estudiantes en cuyo historial se haya categorizado como “Matric. Nuevo”.

```
datos1=datos [datos ['Estado'] ==6]

datos1
```

Lo que indica que el filtro se debe hacer si en la columna Estado es “Matric. Nuevo” teniendo la nuevos únicos registros

| | ID | Periodo | Estado | Sexo | Ciudad | Edad | Estado_final |
|---|---|---|---|---|---|---|---|
| **4153** | 9346235 | 0 | 6 | 0 | 7 | 21 | 0 |
| **4252** | 9352380 | 0 | 6 | 0 | 43 | 22 | 1 |
| **4270** | 9353671 | 0 | 6 | 1 | 7 | 20 | 1 |
| **4275** | 9353814 | 0 | 6 | 1 | 54 | 20 | 1 |
| **4280** | 9354183 | 0 | 6 | 1 | 7 | 21 | 1 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **5092** | 9417133 | 4 | 6 | 0 | 7 | 17 | 1 |
| **5093** | 9417328 | 4 | 6 | 1 | 7 | 20 | 1 |
| **5097** | 9417623 | 4 | 6 | 0 | 7 | 18 | 1 |
| **5098** | 9417666 | 4 | 6 | 0 | 7 | 18 | 1 |
| **5099** | 9417783 | 4 | 6 | 1 | 7 | 19 | 1 |

306 rows × 7 columns

*Tabla 11 Dataframe tras filtrar por estado – Caso 1.*

La biblioteca Scikit Learn posee una amplia variedad de paquetes de clasificación por lo que es necesario importar aquellos que se vayan a utilizar para los tres métodos.

```
from sklearn.model_selection import train_test_split

from sklearn.neighbors import KNeighborsClassifier

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import accuracy_score, confusion_matrix

from sklearn.tree import export_graphviz
```

```
from sklearn import svm

from matplotlib import style
```

Con los paquetes ya importados es necesario definir cuáles son las columnas que contienen las variables de entrada y salida para la clasificación, la partición en entrenamiento- prueba y la semilla para la fijación de los números aleatorios.

Definimos las variables como X y Y de tal manera que:

```
X= datos1.iloc[:, [1,5]]

Y = datos1.Estado_final
```

Fijamos la semilla para números aleatorios, con el fin de evitar que para cada corrida del código se generen nuevos números que afecten los resultados. Para eso utilizamos el paquete de NumPy.

```
np.random.seed (2020)
```

Para la separación de la muestra de entrenamiento y prueba, utilizamos el paquete de Skicit Learn llamado train_test_split que se encarga de revolver el Dataframe y escoger aleatoriamente los datos para cada uno de los grupos teniendo en cuenta una proporción especifica. Para este proyecto se utilizará 70% para entrenamiento y 30% de prueba.

Otro de los parámetros que uno puede manipular es la manera en que se desarrolla el muestreo. Se implementa muestreo estratificado para que los datos totales se puedan dividir en varios grupos que sean homogéneos, aumentando la eficiencia del proceso.

```
X_train, X_test, Y_train, Y_test =train_test_split(X, Y, test_size = 0.3, stratify = Y)
```

Estos pasos sirven para el análisis del caso 1 y caso 2, dado que se usa la misma semilla y la misma distribución para el muestreo. Lo único que cambia es la definición de las variables de entrada y de salida X y Y.

BOSQUES ALEATORIOS – Caso 1

Para utilizar los arboles de decisión utilizamos el parámetro RandomForestClassifier cuyos parámetros son: n_estimators, para determinar la cantidad de árboles a construir; criterion, la función que se va a utilizar para determinar la calidad de cada nivel (en este caso utilizamos ganancia de la información con la entropía); min_samples_split, la mínima cantidad de muestras para crear un nodo del árbol.

```
Rfc=RandomForestClassifier (n_estimators=100, criterion= 'entropy', min_samples_split= 20)
```

Con el modelo ya creado, lo corremos sobre nuestras variables X y.

```
rfc.fit (X_train, Y_train)
```

Utilizamos el modelo ya entrenado con los datos de entrenamiento para predecir las variables de salida para los datos de prueba.

```
y_pred= rfc.predict (X_test)
```

Podemos ver la precisión del modelo con los parámetros accuracy_score y confusión_matrix. De manera general, lo que muestran este indicador es que tan eficiente es el modelo creado para predecir la variable de salida.

```
accuracy_score (Y_test, y_pred)

confusion_matrix (Y_test, y_pred)
```

Los resultados fueron los siguientes:

**Accuracy score: 0.7065217391304348**

**Matriz de confusión:** $\begin{bmatrix} 2 & 21 \\ 6 & 63 \end{bmatrix}$

La capacidad del modelo para predecir es de un 70.65%. De la matriz de confusión podemos ver que el modelo predijo:

- Predijo correctamente 2 en estado “Cancelado” y 63 en estado “Registro”
- Predijo incorrectamente 21 “Cancelados” categorizándolos en “Registro” y 6 en “Registro” categorizándolos como “Cancelado.

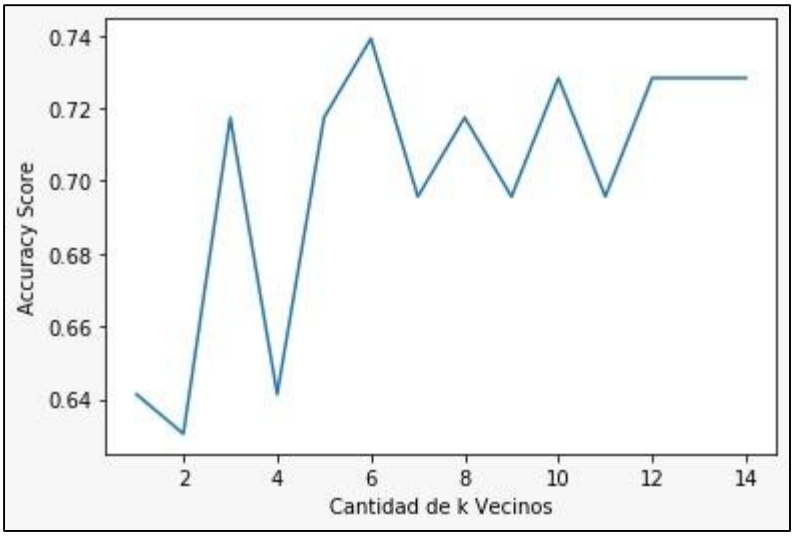
K NEAREST NEIGHBORS – CASO 1

Para utilizar K Vecinos más cercanos utilizamos el parámetro KNeighborsClassifier al cual no se le agrega ningún parámetro. Dado el funcionamiento de la técnica de clasificación, podemos hacer una serie de iteraciones para determinar en qué numero k (Número de vecinos) presenta mayor precisión el modelo.

```
n = np.arange (1,15)

entrenamiento_r = []

for i, j in enumerate (n):

        kn = KNeighborsClassifier (n_neighbors = j)

       kn.fit (X_train, Y_train)

        entrenamiento_r.append (kn.score (X_train, Y_train))
```

Así, estamos creando un modelo cambiando el valor del número de vecinos de 1 a 15 y buscando cuál de estos tiene mayor precisión. El resultado es el siguiente:

```
entrenamiento_r

plt.plot(n, entrenamiento_r)

[0.6413043478260869,0.6304347826086957,0.717391304347826,

0.6413043478260869,0.717391304347826,0.7391304347826086,

0.6956521739130435,0.717391304347826,0.6956521739130435,

0.7282608695652174,0.6956521739130435,0.7282608695652174, 0.7282608695652174,

0.7282608695652174]
```




*Grafica 13 Variación del Accuracy Score para distinta cantidad de vecinos K – Caso 1*

Por lo tanto, la cantidad de vecinos a utilizar para la predicción será 5. Realizamos la predicción y determinamos el accuracy_score y la matriz de confusión para este método

```
kn = KNeighborsClassifier (n_neighbors = 5)
kn.fit (X_train, Y_train)
y_predk = kn.predict (X_test)
accuracy_score (Y_test, y_predk)
confusion_matrix (Y_test, y_predk)
```

Los resultados fueron los siguientes:

**Accuracy score: 0,717391304347826**

**Matriz de confusión:** $\begin{bmatrix} 7 & 16 \\ 10 & 59 \end{bmatrix}$

La capacidad del modelo para predecir es de un 71.73%. De la matriz de confusión podemos ver que el modelo predijo:

- Predijo correctamente 7 en estado “Cancelado” y 59 en estado “Registro”

- Predijo incorrectamente 16 “Cancelados” categorizándolos en “Registro” y 10 en “Registro” categorizándolos como “Cancelado”.

MÁQUINAS DE SOPORTE VECTORIAL (SVM) – CASO 1

Para hacer la clasificación mediante el uso de SVM utilizamos el paquete svm.SVC que trae la biblioteca de Scikit Learn. Los parámetros más importantes son: el Kernel que indica el método del hiperplano que va a separar los datos, para este caso utilizaremos Linear, Poly, Rbf y Sigmoid; y C como un parámetro de tolerancia del modelo a la hora de separar los datos, en este caso se utiliza el estándar 1.

A continuación se hacen los modelos para cada uno de las opciones.

Tipo de Kernel: Linear

```
msv1 =svm.SVC (kernel='linear', C = 1.0)
msv1.fit (X_train, Y_train)
y_predsvm1 = msv1.predict (X_test)
print (confusion_matrix (Y_test, y_predsvm), accuracy_score (Y_test, y_predsvm))
```

**Accuracy Score:** 0.7391304347826086

**Matriz de confusion:** $\begin{bmatrix} 0 & 23 \\ 1 & 68 \end{bmatrix}$

La capacidad del modelo utilizando Kernel Linear para predecir es de un 73.91%. De la matriz de confusión podemos ver que el modelo predijo:

- Predijo correctamente 0 en estado “Cancelado” y 68 en estado “Registro”
- Predijo incorrectamente 23 “Cancelados” categorizándolos en “Registro” y 1 en “Registro” categorizándolos como “Cancelado.

Tipo de Kernel: Poly

```
msv2 =svm.SVC (kernel='poly', C = 1.0)
msv2.fit (X_train, Y_train)
y_predsvm2 = msv2.predict (X_test)
print (confusion_matrix (Y_test, y_predsvm2), accuracy_score (Y_test, y_predsvm2))
```

**Accuracy Score:** 0.7391304347826086

**Matriz de confusión:** $\begin{bmatrix} 0 & 23 \\ 1 & 68 \end{bmatrix}$

La capacidad del modelo utilizando Kernel Poly para predecir es de un 73.91%. De la matriz de confusión podemos ver que el modelo predijo:

- Predijo correctamente 0 en estado “Cancelado” y 68 en estado “Registro”
- Predijo incorrectamente 23 “Cancelados” categorizándolos en “Registro” y 1 en “Registro” categorizándolos como “Cancelado.

Tipo de Kernel: rbf

```
msv3 =svm.SVC (kernel='rbf', C = 1.0)
msv3.fit (X_train, Y_train)
y_predsvm3 = msv3.predict (X_test)
print (confusion_matrix (Y_test, y_predsvm3), accuracy_score (Y_test, y_predsvm3))
```

**Accuracy Score:** 0.75

**Matriz de confusión:** $\begin{bmatrix} 0 & 23 \\ 0 & 69 \end{bmatrix}$

La capacidad del modelo utilizando Kernel Rbf para predecir es de un 75% . De la matriz de confusión podemos ver que el modelo predijo:

- Predijo correctamente 0 en estado “Cancelado” y 69 en estado “Registro”
- Predijo incorrectamente 23 “Cancelados” categorizándolos en “Registro” y 0 en “Registro” categorizándolos como “Cancelado.

Tipo de Kernel: Sigmoid

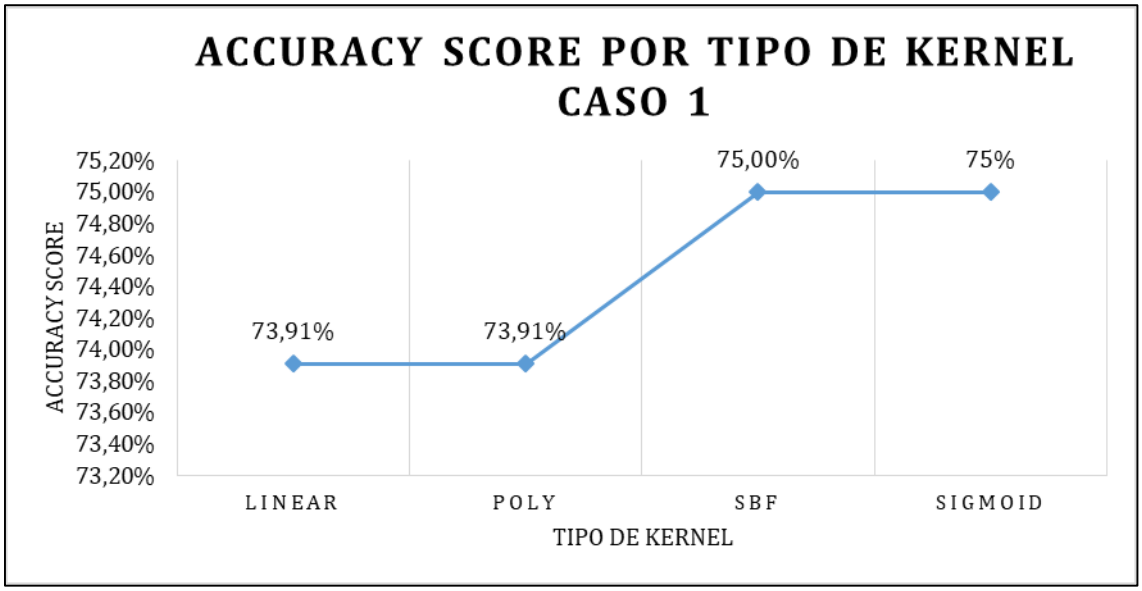
```
msv4 =svm.SVC (kernel='sigmoid', C = 1.0)
msv4.fit (X_train, Y_train)
y_predsvm4 = msv4.predict (X_test)
print (confusion_matrix (Y_test, y_predsvm4), accuracy_score (Y_test, y_predsvm4))
```

**Accuracy Score:** 0.75

**Matriz de confusión:** $\begin{bmatrix} 0 & 23 \\ 0 & 69 \end{bmatrix}$

La capacidad del modelo utilizando Kernel Rbf para predecir es de un 75%. De la matriz de confusión podemos ver que el modelo predijo:

- Predijo correctamente 0 en estado “Cancelado” y 69 en estado “Registro”
- Predijo incorrectamente 23 “Cancelados” categorizándolos en “Registro” y 0 en “Registro” categorizándolos como “Cancelado.




*GRAFICA 6: Variación del Accuracy Score para los distintos tipos de Kernel- Caso 1*

En resumen, los tipos de Kernel a utilizar que dan mejor exactitud son el “Sigmoid “ y “Rbf”.

### 7.4.2 CLASIFICACION - CASO 2

Para el caso 2 , como lo hicimos para el caso 1, empezamos fijando la semilla para números aleatorios, definiendo las variables X2,Y2 y agrupando los datos en datos para entrenamiento y datos para prueba.

```
np.random.seed(2020)
X2 = datos.iloc[:, [1,3,4,5]]
Y2 = datos.Estado_final
X2_train , X2_test, Y2_train , Y2_test =train_test_split(X2,Y2,test_size = 0.3 , stratify = Y2)
```

Con esto ya definido, se realiza el mismo procedimiento que el caso 1 incluyendo las nuevas variables. A continuación se muestran los resultados para cada uno de los métodos de clasificación.

RANDOM FOREST – CASO 2

**Accuracy Score:** 0.7912589693411611

**Matriz de confusión:** $\begin{bmatrix} 155 & 234 \\ 86 & 1058 \end{bmatrix}$

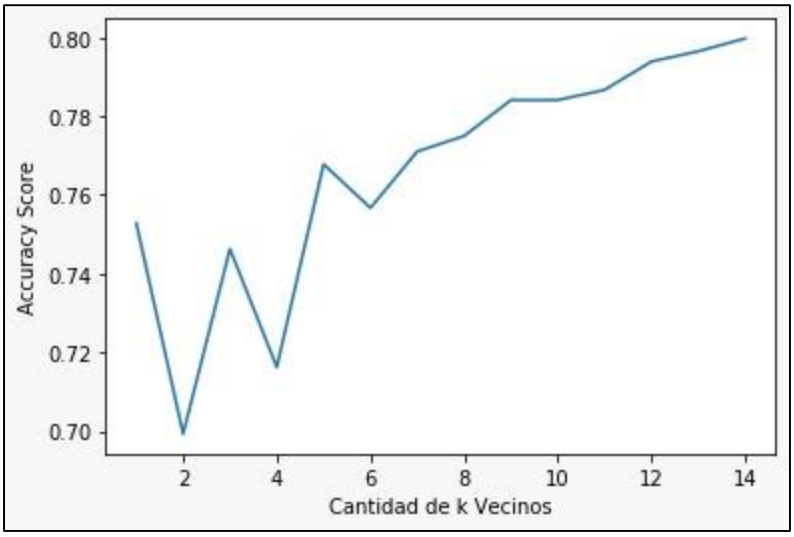
La capacidad del modelo para predecir es de un 79.1258%. De la matriz de confusión podemos ver que el modelo predijo:

- Predijo correctamente 155 en estado “Cancelado” y 1058 en estado “Registro”
- Predijo incorrectamente 234 “Cancelados” categorizándolos en “Registro” y 86 en “Registro” categorizándolos como “Cancelado.

K –NEAREST NEIGHBORS – CASO 2

El cálculo de la precisión del modelo para las iteraciones en el caso 2 fue:

[0.7527723418134377, 0.6992824527071102,0.746249184605349,
0.7162426614481409,0.7677756033920418,0.7566862361382909,
0.7710371819960861,0.7749510763209393,0.7840834964122635,
0.7840834964122635,0.786692759295499,0.7938682322243966, 0.7964774951076321,
0.7997390737116764]




*Grafica 14 Variación del Accuracy Score para distinta cantidad de vecinos K – Caso 2*

Calculando resultados para un número de vecinos K =5 , se tiene que:

***Accuracy score:*** 0.7677756033920418

***Matriz de confusión:*** $\begin{bmatrix} 147 & 242 \\ 114 & 1030 \end{bmatrix}$

La capacidad del modelo para predecir es de un 76.77%. De la matriz de confusión podemos ver que el modelo predijo:

- Predijo correctamente 147 en estado “Cancelado” y 1030 en estado “Registro”
- Predijo incorrectamente 242 “Cancelados” categorizándolos en “Registro” y 114 en “Registro” categorizándolos como “Cancelado.

## MÁQUINAS DE SOPORTE VECTORIAL (SVM) – CASO 2

Los resultados para cada uno de los modelos cambiando el tipo de Kernel, al igual que el caso 1 son las siguientes:

Tipo de Kernel : Linear

***Accuracy Score:*** 0.7849834964122635

***Matriz de confusion:*** $\begin{bmatrix} 94 & 295 \\ 36 & 1108 \end{bmatrix}$

La capacidad del modelo utilizando Kernel Lineal para predecir es de un 78.498%. De la matriz de confusión podemos ver que el modelo predijo:

- Predijo correctamente 94 en estado “Cancelado” y 1108 en estado “Registro”
- Predijo incorrectamente 295 “Cancelados” categorizándolos en “Registro” y 36 en “Registro” categorizándolos como “Cancelado.

Tipo de Kernel : Poly

***Accuracy Score:*** 0.76777756033920418

***Matriz de confusión:*** $\begin{bmatrix} 46 & 343 \\ 13 & 1131 \end{bmatrix}$

La capacidad del modelo utilizando Kernel Poly para predecir es de un 76.77%. De la matriz de confusión podemos ver que el modelo predijo:

- Predijo correctamente 46 en estado “Cancelado” y 1131 en estado “Registro”
- Predijo incorrectamente 343 “Cancelados” categorizándolos en “Registro” y 13 en “Registro” categorizándolos como “Cancelado.

Tipo de Kernel : Rbf

***Accuracy Score:*** 0.761252446183953

***Matriz de confusion:*** $\begin{bmatrix} 26 & 363 \\ 3 & 1141 \end{bmatrix}$

La capacidad del modelo utilizando Kernel Rbf para predecir es de un 76.125%. De la matriz de confusión podemos ver que el modelo predijo:

- Predijo correctamente 26 en estado “Cancelado” y 1141 en estado “Registro”
- Predijo incorrectamente 363 “Cancelados” categorizándolos en “Registro” y 3 en “Registro” categorizándolos como “Cancelado.

Tipo de Kernel : Sigmoid

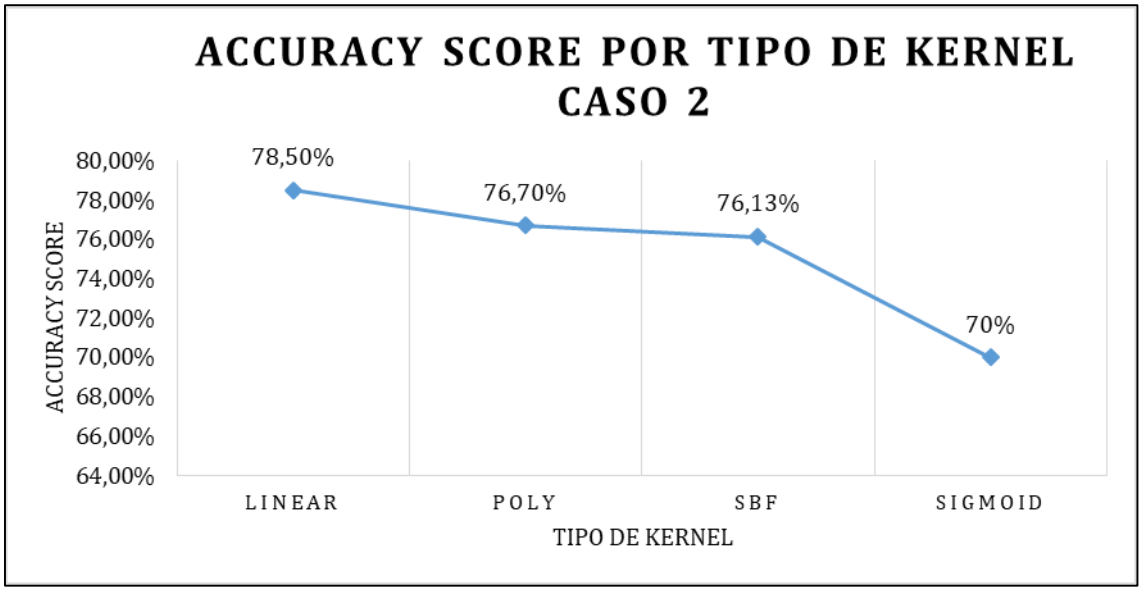
***Accuracy Score:*** 0.700587084148728

***Matriz de confusión:*** $\begin{bmatrix} 153 & 236 \\ 223 & 921 \end{bmatrix}$

La capacidad del modelo utilizando Kernel Lineal para predecir es de un 70 %. De la matriz de confusión podemos ver que el modelo predijo:

- Predijo correctamente 153en estado “Cancelado” y 921 en estado “Registro”
- Predijo incorrectamente 236 “Cancelados” categorizándolos en “Registro” y 223 en “Registro” categorizándolos como “Cancelado.

En resumen, el tipo de Kernel que da mas exactitud al modelo es “Linear”




*Grafica 15Variación del Accuracy Score para los distintos tipos de Kernel- Caso 2*

## 8. CONCLUSIONES

- El análisis de estos datos busca encontrar una relación entre datos personales de los estudiantes de la Escuela y su estado en la misma. A pesar de que los modelos creados tienen una buena capacidad de predecir, los resultados no son una base sólida para una toma de decisiones debido a su manipulación previa.

  Estos modelos desconocen la causa de porqué el estudiante está matriculado o no en la escuela para este año 2020. Si no está matriculado, solo toma las variables a las cual se

tuvo acceso que son los datos personales, buscando a la fuerza una relación, cuando estas no están directamente relacionadas con este fenómeno.

- Cada uno de los modelos da un accuracy_score para los dos casos:

**CASO 1**

| **ACCURACY SCORE PARA CADA UNO DE LOS METODOS UTILIZADOS** | |
|---|---|
| **METODO** | **VALOR** |
| Bosques Aleatorios | 70,65% |
| K vecinos | 71,73% |
| Máquina de soporte Vectorial : Linear | 73,91% |
| Máquina de soporte Vectorial : Poly | 73,91% |
| Máquina de soporte Vectorial : Rbf | 75% |
| Máquina de soporte Vectorial :Sigmoid | 75% |

*Tabla 12 Resumen de resultados – Caso 1.*

**CASO 2**

| **ACCURACY SCORE PARA CADA UNO DE LOS METODOS UTILIZADOS** | |
|---|---|
| **METODO** | **VALOR** |
| K Vecinos | 79,12% |
| Bosques aleatorios | 76,7% |
| Máquina de soporte Vectorial : Linear | 78,50% |
| Máquina de soporte Vectorial : Poly | 76,70% |

| | |
|---|---|
| Máquina de soporte Vectorial : Rbf | 76,13% |
| Máquina de soporte Vectorial :Sigmoid | 70% |

*Tabla 13 Resumen de resultados – Caso 2.*

Cada uno de los modelos tiene una precisión entre el 70 al 80% por lo que corrobora la elaboración del modelo para los dos casos.

- Los datos suministrados por La Escuela no alcanzan para hacer un buen modelo de análisis de datos dado que las variables utilizadas no tienen ninguna relación con su estado final en la institución. No hay fundamentos para probar que los datos personales de una persona puedan estar relacionadas con su estado universitario.

  Este análisis hubiera podido tener mejores resultados si se hubieran tenido variables directamente relacionadas al estado final. Entre estas variables podrían estar: La cantidad de asignaturas vistas por semestre, las notas del semestre, los resultados en las pruebas de admisión, entre otros.

## 10. ANEXOS

- Código en Python.
- Datos suministrados por la universidad formato Excel

## 10. BIBLIOGRAFÍA

- ¿Cómo saber las etiquetadas asignadas por tipo de categoria cat.codes ? , (29 de Junio 2018), Recuperado de https://www.it-swarm.dev/es/python/como-saber-las-etiquetas-asignadas-por-tipo-categoria.-cat.codes/806322417/

- Betancourt, Gustavo A. (2005) Universidad Tecnologica de Pereira , Las maquinas de soporte vectorial(SVMs). recuperado de https://revistas.utp.edu.co/index.php/revistaciencia/article/view/6895

- Zelada , Carlos (2017) RPubs by RStudio : Evaluacion de modelos de clasificacion(Matriz de confusion), recuperado dehttps://rpubs.com/chzelada/275494

- Redaccion APD (2019) APD: ¿Cuales son los tipos de algoritmos del Machine Learning? , recuperado de https://www.apd.es/algoritmos-del-machine-learning/

- Gonzalez , Ligdi (2019), Maquinas Vectores de Soporte Clasificacion y Teoria , recuperado de https://ligdigonzalez.com/maquinas-vectores-de-soporte-clasificacion-teoria/

- Gonzalez , Ligdi (2019) , Bosques aleatorios Clasificacion y Teoria, recuperado de https://ligdigonzalez.com/bosques-aleatorios-clasificacion-teoria-machine-learning/

- Gonzalez, Ligdi (2019), K vecinos más cercanos Teoria, recuperado de https://ligdigonzalez.com/k-vecinos-mas-cercanos-teoria-machine-

learning/#:~:text=K%20vecinos%20m%C3%A1s%20cercanos%20es,y%20la%20detecci%C3%B3n%20de%20intrusos.