**Pronóstico de accidentes de tránsito por tipo de gravedad en Medellín 2019-2021**

1. **Introducción**

Día tras días la recolección de información en distintos entornos ha permitido explorar y tratar de entender problemas de todo índole con el objetivo de mejorar el bienestar de un grupo de interés. Es así, como en este trabajo se aborda una problemática que ha tomado gran relevancia con el aumento del parque automotor como es la accidentalidad en el municipio de Medellín (Colombia). Con este trabajo se pretende ofrecer el mejor pronóstico sobre los accidentes de tránsito por tipo de gravedad (Accidentes Graves y Leves) para los años 2019-2021 con uso y evaluación de herramientas de análitica predictiva. Para tal fin, el presente trabajo estará dividido en X secciones: la primera consta de una introducción; la siguiente presenta el público objetivo y enumera posibles usos del trabajo; posteriormente, se detalla la los datos utilizados; la quinta parte presenta la metodología utilizada; la sexta parte presenta los resultados y por último se ofrecen algunas conclusiones.

1. **Público objetivo**

El trabajo está enfocado para el uso de la comunidad en general con la interacción de una aplicación web, por parte de entidades públicas (Alcaldía de Medellín y Secretaría de Tránsito) y privadas como Compañías de seguro (Seguro Obligatorio de Accidentes de Tránsito-SOAT) que pueden tener el detalle de los modelos ajustados con la identificación de patrones que pueden intensificar o reducir la accidentalidad.

La información estimada y publicada puede ser usada por los ciudadanos en general para familiarizarse y entender de forma amigable las cifra por medio de interacciones sobre representaciones gráficas de la accidentalidad observada entre el 2014 y 2018 y el pronóstico del 2019 al 2021. También puede ser utilizado por las entidades públicas como insumo para la toma de decisiones e incluso para la evaluación o validación en las políticas enfocadas en la reducción disminución de gravedad de los accidente y finalmente por parte de las compañías de seguros para poder contrastar o realizar estimaciones sobre la siniestralidad de los productos del seguro obligatorio SOAT.

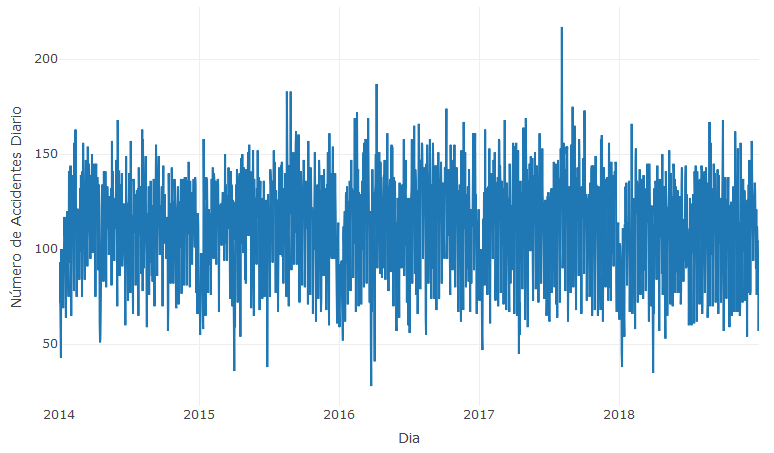
1. **Datos**

La alcaldía de Medellín por medio del portal MEData ofrece datos abiertos a la comunidad en general para el uso y aprovechamiento de la misma. De este portal, obtuvimos la información de accidentes de tránsitos entre el 2014 y 2018 tipificada por variables como: Tipo de accidente, gravedad, ubicación, Fecha y hora. Para el presente trabajo, se utilizaron los datos de accidentes de tránsitos con su respectiva fecha y clasificada por gravedad (Heridos, Muerto y Solo Choque).

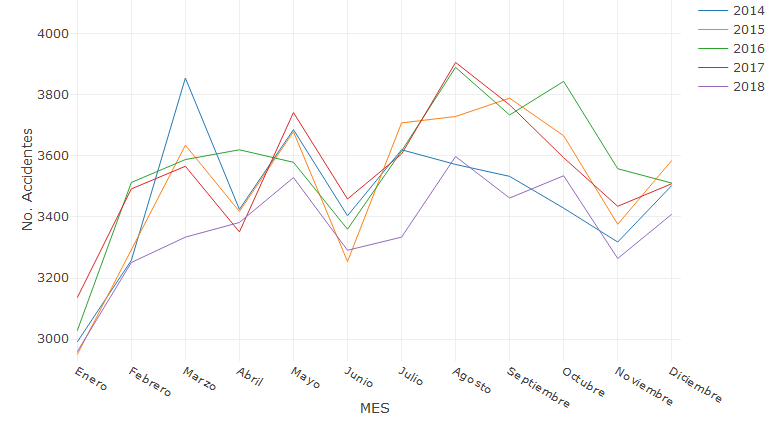
1. **Análisis descriptivo**

En el gráfico 1 se observa el comportamiento del número de accidentes de tránsito en Medellín en el periodo de análisis. En esta primera aproximación, no se logra observar algún patrón importante salvo las caídas en el número de accidentes a inicio de cada año.

**Gráfico 1.** Evolución de los accidentes diarios en Medellín 2014-2018

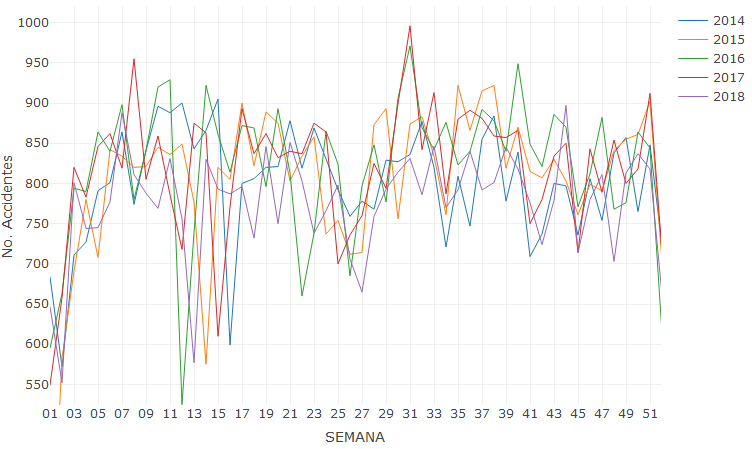
Realizando una presentación en el gráfico 2 del comportamiento mensual y por año del número de accidentes, este nos permite identificar un patrón de caída recurrentes en el número de accidentes para los meses de enero, abril, junio y noviembre. Siendo más pronunciado para el mes de de enero. Este último, podría estar asociado a la temporada de vacaciones**.**

**Gráfico 2.** Número de accidentes mensual en Medellín 2014-2018.

****

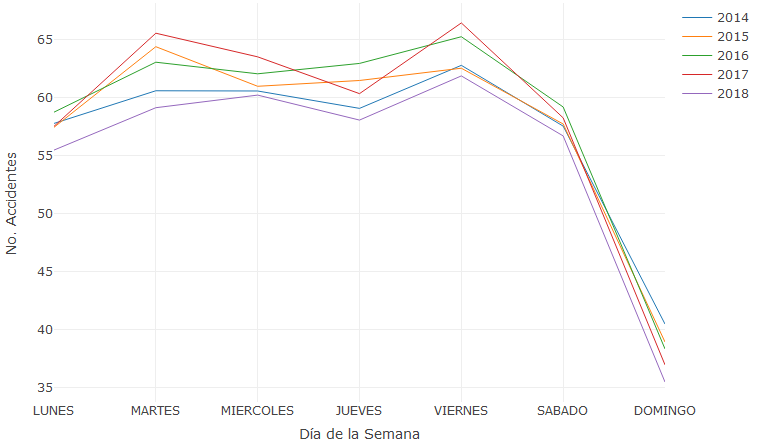
En una presentación semanal de los accidentes de tránsito, en el gráfico 3 se observan ciertas semanas con recurrencia en la disminución de la accidentalidad. Las cuales van en línea con los hallazgos en la visualización mensual.Al indagar por la particularidad de estos patrones se identifica la importancia de la Semana Santa y la existencia de días festivos en cada una de las semanas.

**Gráfico 3.** Número de accidentes semanal en Medellín 2014-2018.

****

Despues del analisis semanal, se explora por dia de la semana donde se identifica un patrón marcado para los domingos con una disminución en la accidentalidad.

**Gráfico 4.** Promedio de accidentes por día de la semanal en Medellín 2014-2018.

****

Teniendo en cuenta todo el análisis descriptivo de la accidentalidad en Medellín, se construyó un conjunto de variables dicotómicas para evaluar la importancia dias especificos o grupo de dias con particularidades. A continuación, se presenta una descripción sobre dichas variables que se van a evaluar en la estimación del mejor modelo y su posible impacto en la estimación del número de accidentes:

**AÑO:** Año de ocurrencia del accidente. Tendencias o cambios de nivel de los accidentes a través de los años.

**MES:** Mes de Ocurrencia del accidente. Patrones estacionales por mes.

**SEMANA:** Semana de ocurrencia del accidente. Patrón estacional por semanal.

**DIA:** Dia de ocurrencia del accidentes. Patrón estacional por día de la semana.

**Feriado:** Identificador de dia feriado en el calendario colombiano excluyendo el jueves y viernes santo. Cambios de nivel en días feriados.

**Feriado\_v1:** Identificador de día feriado en el calendario colombiano incluyendo el jueves y viernes santo. Cambios de nivel en días feriados.

**Feriado\_Lunes:** Identificador de los días feriado lunes en el calendario colombiano. Cambios de nivel en días lunes feriados.

**Feriado\_Otro:** Identificador de díia feriado distinto a lunes en el calendario colombiano. Cambios de nivel en días distintos a lunes y feriados.

**Previo\_feriado:** Identificación de los tres días previos al lunes feriado. Cambios de nivel en días previos a lunes feriados.

**Viernes\_Antes\_Puente:** Identificación de dia viernes antes de lunes feriado. Cambios de nivel para viernes previos a lunes feriados.

**Semana\_Santa:** Identificación de días de semana santa desde el sábado antes del domingo de ramos hasta el domingo de resurrección. Cambios de nivel para los días de semana santa.

**Semana\_Santa\_v1:** Identificación de días de semana santa desde lunes de semana santa hasta el domingo de resurrección. Cambios de nivel para los días de semana santa.

**Prima:** Dia de pago de la prima. Cambio de nivel el dia del pago de la prima.

**Mujer:** Dia de celebracion del dia de la mujer. Cambio de nivel el dia de celebracion.

**Padre:** Dia de celebracion del padre. Cambio de nivel el dia de celebracion.

**Madre:** Dia de celebracion del dia de la madre. Cambio de nivel el dia de celebracion.

**AmoryAmistad:** Dia de celebracion del dia de amor y amistad. Cambio de nivel el dia de celebracion.

**Quincena:** Dia de pago de la quincena asociado al 15 o último día del mes. Cambio de nivel el día de pago.

**Viernes\_Desp\_Quincena:** Viernes más próximo a la quincena. Cambio de nivel.

**Viernes\_Desp\_Quincena\_v1:** Viernes más próximo a la quincena en la misma semana. Cambio de nivel.

**Feria\_Flores:** Días de celebración de feria de flores. Cambio de nivel durante la celebración.

**Metodología**

1. **Procesamiento de los datos:**

Se hizo una revisión de la información que se presenta desde el año 2014 hasta el año 2018 con el fin de conocer las variables, número de registros y la calidad del dato. Con base en la información se identifican las variables fecha y gravedad como aquellas que son de interés para el desarrollo del modelo de predicción de accidente de las cuales se partirá para agregar nuevas variables. Posterior a lo anterior se agrupan los datos de los diferentes años y se estandariza la información de la variable gravedad, esta presenta 3 categorías: “Solo daños, heridos y muertos”, pero se revisa que la cantidad de registros de muertos (1253) no es significativa frente al total registros, para lo cual se toma la decisión de realizar las siguientes agrupaciones que permitan una mejor manipulación de los datos:

* Accidentes Graves: Heridos y Muertos
* Accidentes Leves: Solo Daño
* Total Accidentes: Accidentes Graves y Leves

Por cada una de las agrupaciones se calcula la frecuencia respectiva por la fecha, se crean las variables, mes,año, semana y día.

Adicional a lo anterior se cuenta con una variables que caracterizan fechas especiales tales como día del padre, día de madre, amor y amistad, feria de flores, semana santa, etc. con las cuales se busca identificar si presentan alguna incidencia en los accidentes presentados en la ciudad durante el periodo de ocurrencia.

Después de unificada la información en un solo dataframe, se procede a organizar las variables y a crear un dataframe por cada uno de los periodos de la predicción de accidentes, los cuales son: diaria, semanal y mensual.

1. **Estimación de los modelos:**

Para la estimación de los modelos se procedió a dividir los diferentes dataframes en datos de entrenamiento desde el año 2014 hasta el 2017 y de validación el año 2018 con el fin de ver el comportamiento de los modelos.

Para determinar las mejores variables predictivas para el modelo, se utiliza el método *FORWARD SELECTION* teniendo como criterio aquella selección de variables que presente mejor R^2 ajustado. Sin embargo, para cada una de las periodicidades se realizó un modelo de regresión lineal con las variables seleccionadas anteriormente, con esta información se revisa el p-valor de cada una para determinar las variables definitivas de los modelos, quedando las siguientes:

|  |  |
| --- | --- |
| **Periodicidad** | **Variables** |
| Diaria | Ano\_Base+DIA+SEMANA, Feriado\_Lunes, Feriado\_Otro, Madre, Semana\_Santa, Viernes\_Desp\_Quincena\_v2, Feria\_Flores |
| Semanal | Ano\_Base, SEMANA, Feria\_Flores\_Semana, Semana\_Santa\_Semana, Feriados\_Lunes, Feriados\_Otros |
| Mensual | Ano\_Base, MES, Feriados |

Después de la selección de variables se utilizan los siguientes métodos de modelamiento:

1. Regresión lineal
2. Knn
3. Regresión lineal generalizado
4. Árboles de regresión
5. Bosques aleatorios

Con cada uno de los métodos se construyen los modelos predictivos con los datos de entrenamiento y se validan con el año 2018; esto se lleva a cabo por accidentes totales, accidentes graves y accidentes leves y por cada una de las periodicidades (diaria, semanal, mensual) de estas categorías.

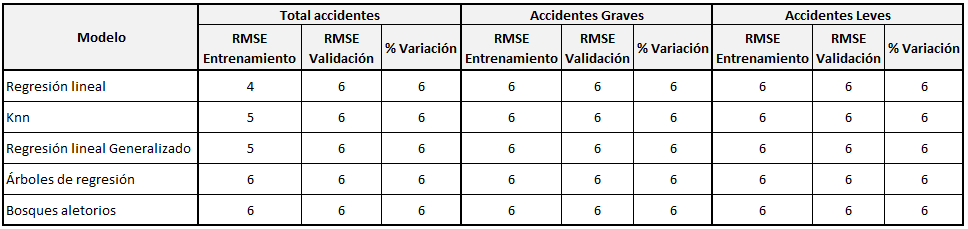
Después de realizar los modelos, se hizo una gráfica comparativa, de acuerdo al tipo de accidente y periodicidad entre los datos reales y de entrenamiento como también los de validación para observar mejor el ajuste.

El factor clave de decisión de los modelos son aquellos que presenten el mínimo error cuadrático medio de la predicción, para lo cual se hacen unos cuadros comparativos que permitieran visualizar esta información y tomar la decisión de los mejores modelos.

Por último, con los modelos elegidos se realizan los diferentes pronósticos de accidentes para los años 2019, 2020 y 2021 y se crean tabla por cada una de las periodicidades.

1. **Aplicación WEB**

**Resultados**

****

**Conclusión**