PRONÓSTICO DE ACCIDENTES DE TRÁNSITO POR TIPO DE GRAVEDAD EN MEDELLÍN 2019-2021

Edwin Villarraga Ossa
Jorge Eliécer Montoya Gallo
Juan David Arango Buitrago

Universidad Nacional de Colombia
Sede Medellín

Facultad de Minas
Especialización en analítica
Primer semestre
2019

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 toda 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 analítica 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 los datos utilizados; la quinta parte presenta la metodología utilizada; la sexta parte presenta los resultados y por último se ofrecen algunas conclusiones.

2. 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 cifras por medio de interacciones sobre representaciones gráficas de la accidentalidad observada entre el 2014 y 2018 y su 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 o disminución de gravedad de los accidentes. Finalmente, por parte de las compañías de seguros para poder contrastar o realizar estimaciones sobre la siniestralidad de los productos del seguro obligatorio SOAT.

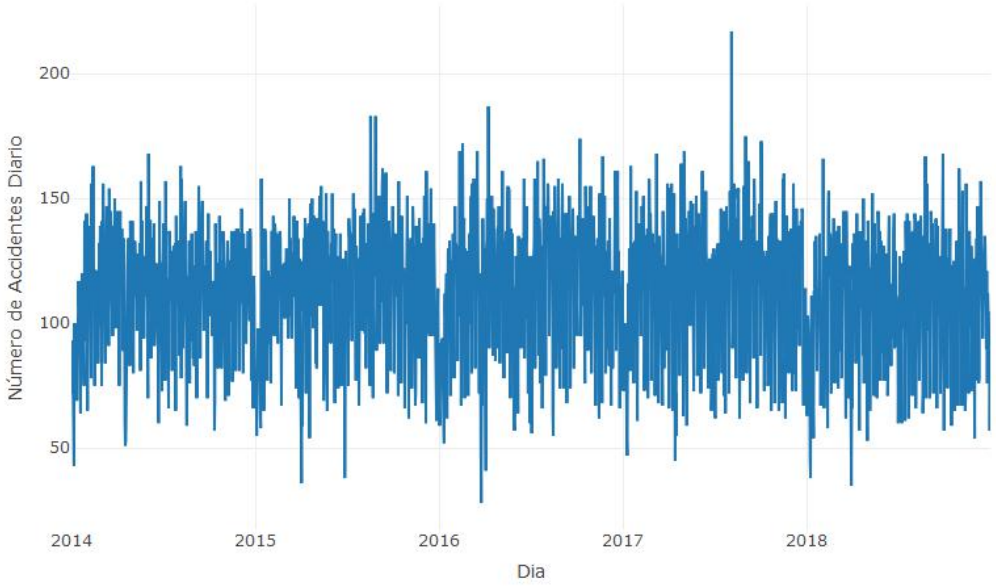
3. Datos

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. En 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).

4. Análisis descriptivo

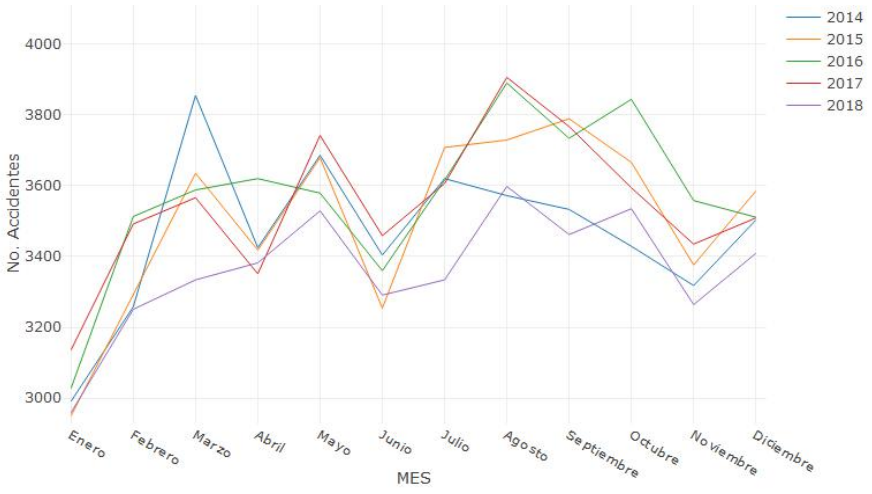
En el gráfico 1 se observa el comportamiento del número de accidentes de tránsito en Medellín en el periodo de análisis. En esta primera aproximación, no se logra observar algún patrón importante salvo las caídas en el número de accidentes a inicio de cada año.

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



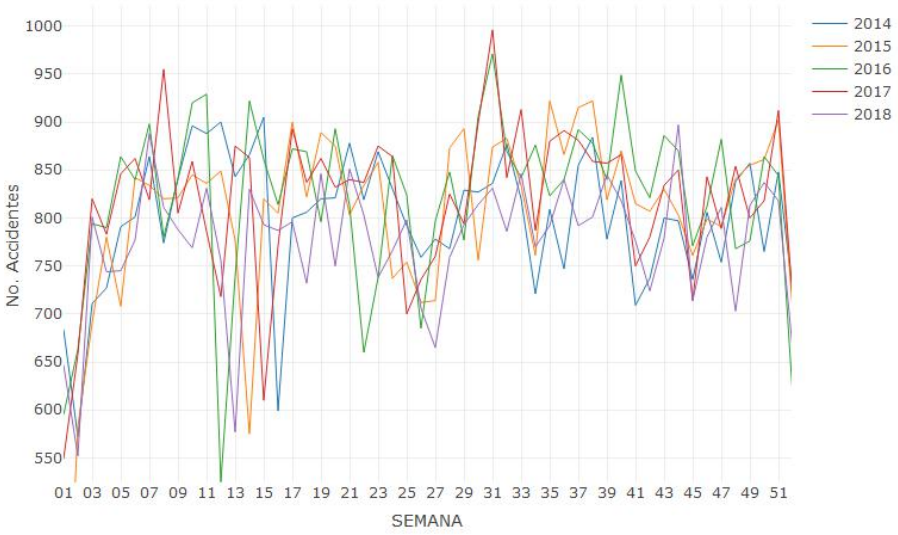
Realizando una presentación en el gráfico 2 del comportamiento mensual y por año del número de accidentes, este nos permite identificar un patrón de caída recurrente en el número de accidentes para los meses de enero, abril, junio y noviembre. Siendo más pronunciado para el mes de de enero. Este último,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

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



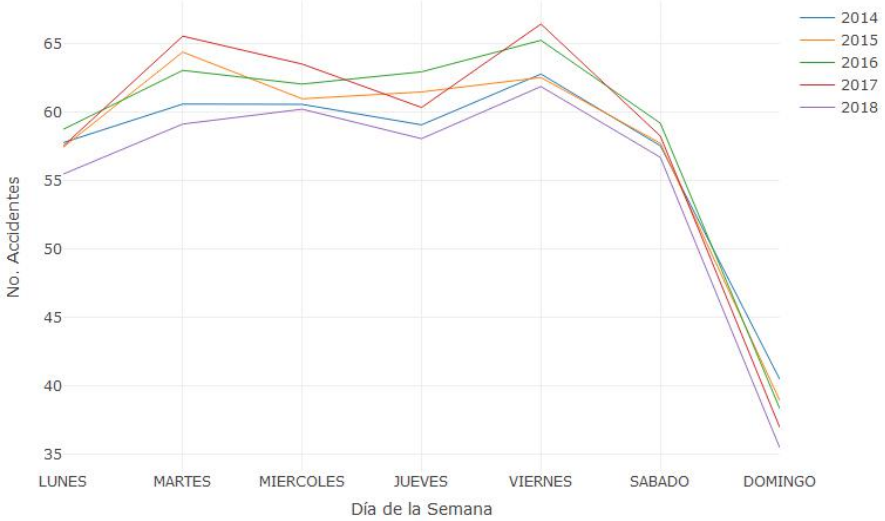
En una representación semanal de los accidentes de tránsito, en el gráfico 3 se observan ciertas semanas con recurrencia en la disminución de la accidentalidad. Las cuales van en línea con los hallazgos en la visualización mensual. Al indagar por la particularidad de estos patrones se identifica la importancia 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.



Continuando con el análisis en otras unidades de medida, se explora por día de la semana donde se observa un patrón marcado para los domingos con una disminución importante en la accidentalidad.

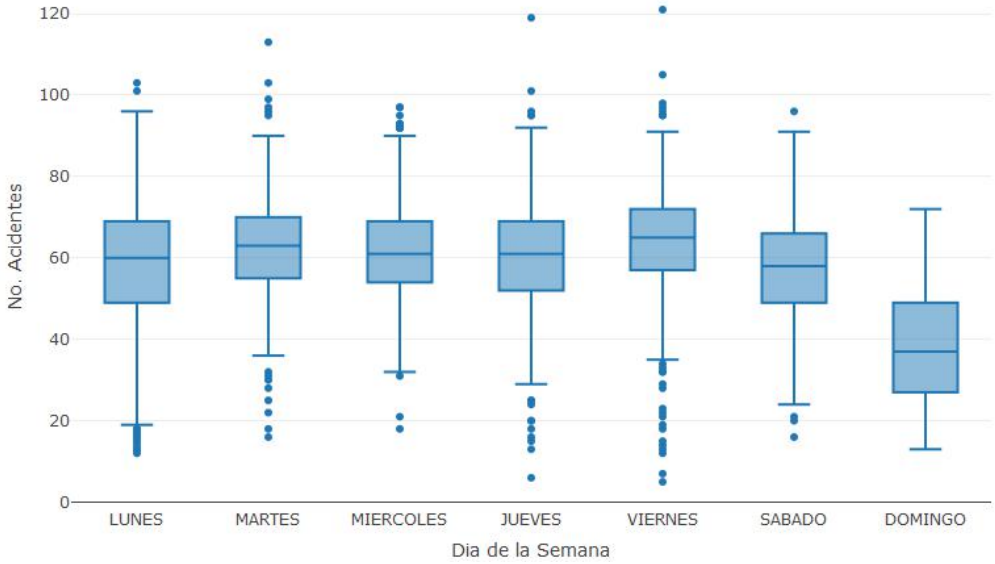
G**ráfico 4.** Promedio de accidentes por día de la semanal en Medellín 2014-2018.



Teniendo en cuenta la importancia de los días festivos en la accidentalidad, se representa por medio del diagrama de caja de bigotes la distribución de los accidentes de tránsito por día de la semana para identificar posibles desviaciones sobre la media y mediana de los registros. Para el día lunes, se identifica un agrupamiento importante por debajo del bigote inferior, dando fortaleza a la

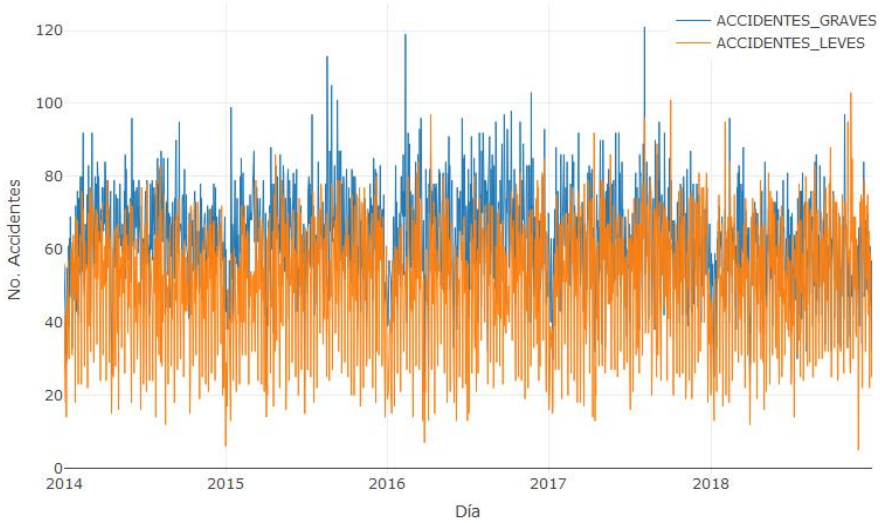
hipótesis sobre la importancia de los días festivos con una reducción importante en la accidentalidad. Para los días viernes, se identifica mayor volatilidad y agrupamientos más importantes que en los otros días. En los demás días se observan registros por encima y por debajo de los bigotes, con un agrupamiento importante en el número de accidentes sobre la mediana de los registros.

Gráfico 5. Distribución de los accidentes de tránsito por día de la semana en Medellín 2014-2018.



A continuación se presentará un análisis descriptivo de la accidentalidad en Medellín por la gravedad, donde se agruparon los accidentes de *solo choques* a "ACCIDENTES_LEVES" y los accidentes de *lesiones* o *muertes* en "ACCIDENTES_GRAVES". En la representación gráfica se observa un patrón común con una accidentalidad grave superior a la leve.

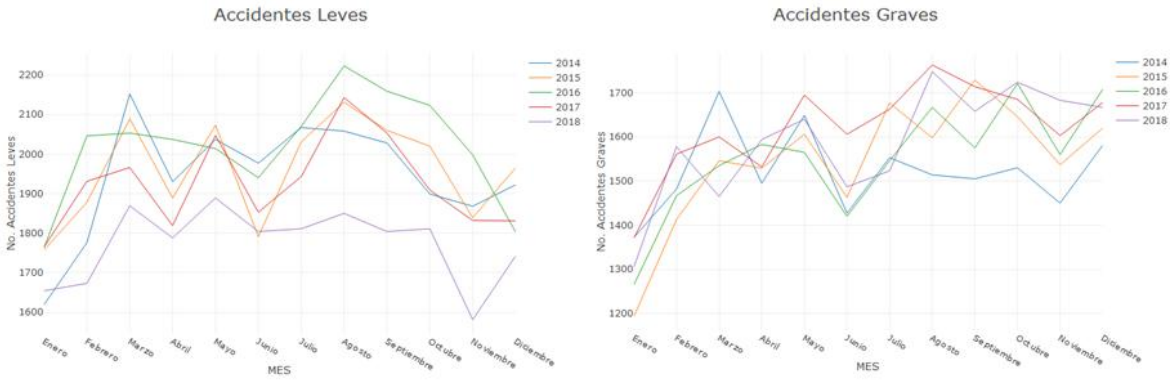
Gráfico 6. Distribución de los accidentes de tránsito por día de la semana en Medellín 2014-2018.



Para la accidentalidad mensual agrupada por gravedad, se sigue presentando comportamientos recurrentes para los meses de enero, abril, junio y noviembre.

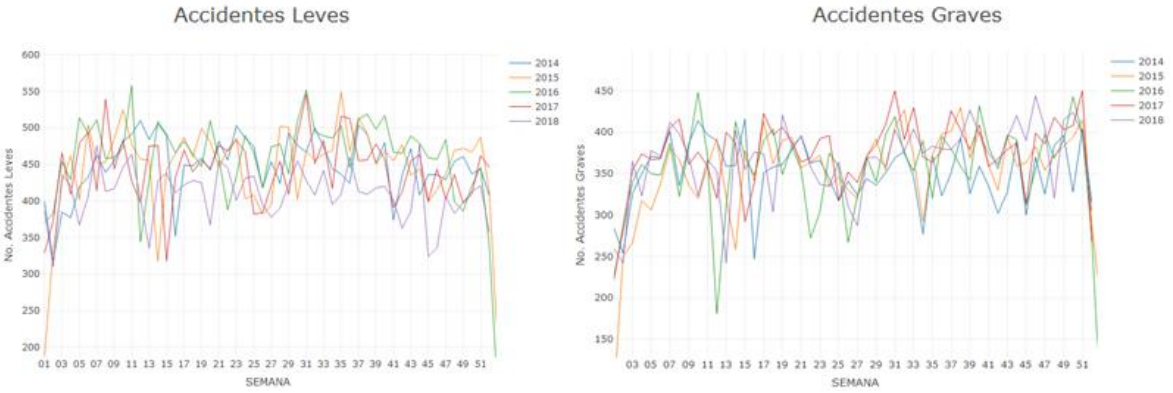
Sin embargo, el efecto no es tan marcado respecto a los accidentes totales y es menos notorio en los accidentes leves.

Gráfico 6. Número de accidentes por gravedad mensual en Medellín 2014-2018



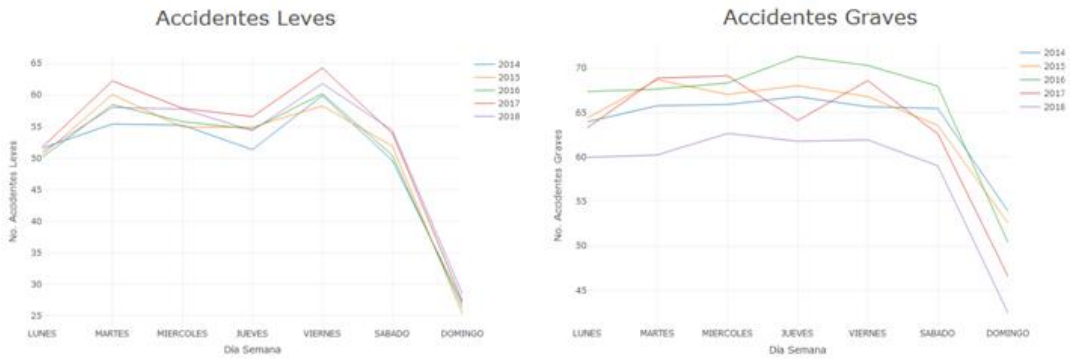
Realizando una representación semanal de los accidentes de tránsito por gravedad, se sigue identificando ciertas semanas con recurrencia en la disminución de la accidentalidad. Siguiendo los hallazgos para los accidentes totales, el gráfico semanal nos permite visualizar la particularidad de ciertas semanas las cuales se asocian a semana santa y semanas con días festivos.

Gráfico 7. Número de accidentes por gravedad semanal en Medellín 2014-2018.



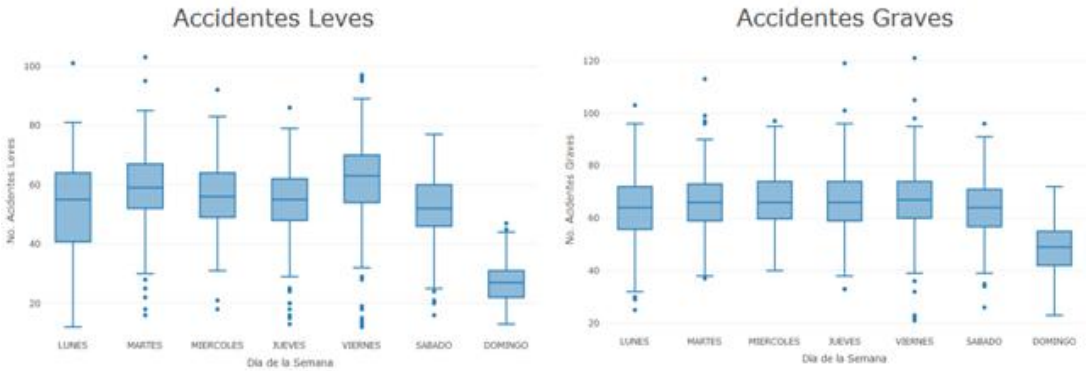
Después del análisis semanal, se explora por día de la semana donde se identifica el mismo patrón marcado para los domingos con una disminución en la accidentalidad tanto para los accidentes graves como los accidentes leves.

Gráfico 8. Promedio de accidentes por gravedad y día de la semanal en Medellín 2014-2018.



Teniendo en cuenta la importancia de los días festivos en la accidentalidad, se representa por medio del diagrama de caja de bigotes su distribución por día y por gravedad del accidente. El agrupamiento en la parte inferior de los lunes ya no es tan marcado. Sin embargo, se identifican agrupamientos en la parte inferior de la caja de bigotes de los accidentes leves.

G**ráfico 9.** Distribución de los accidentes de tránsito por gravedad y día de la semana 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 días específicos o grupo de días con particularidades. A continuación, se presenta una descripción sobre estas variables y sus posibles relaciones en la estimación del mejor modelo**:**

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 semana.

DÍA: Día de ocurrencia de accidentes. Patrón estacional por día de la semana.

Feriado: Identificador de día 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 día 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 día 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: Día de pago de la prima. Cambio de nivel el día del pago de la prima.

Mujer: Día de celebración del día de la mujer. Cambio de nivel el día de celebración.

Padre: Día de celebración del padre. Cambio de nivel el día de celebración.

Madre: Día de celebración del día de la madre. Cambio de nivel el día de celebración.

AmoryAmistad: Día de celebración del día de amor y amistad. Cambio de nivel el día de celebración.

Quincena: Día 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.

5. Metodología

5.1. 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. De acuerdo a lo anterior, se agrupan los datos de los diferentes años y se estandariza la información de la variable gravedad; la cual presenta 3 categorías: “Solo daños, heridos y muertos”. Al revisar la cantidad de registros de la categoría de muertos se encuentran 1.253; un valor no tan significativo frente al total registros. Para lo cual, se toma la decisión de agrupar la información y lograr una mejor manipulación de los datos:

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

Los accidentes leves solo involucran datos materiales y los accidentes graves implican lesiones personales y muertos. 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 unas 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 patrones de incidencia en los accidentes presentados en la ciudad durante esas fechas.

Después de unificada la información en un solo base de datos, se procede a organizar las variables y a crear nuevas bases de datos por cada uno de los periodos de la predicción de los accidentes (diario, semanal y mensual) y por tipo de gravedad del accidente.

5.2. Análisis descriptivo

Seguido de tener la información organizada, se realiza un análisis descriptivo con el objetivo de observar relaciones entre variables, describir las tendencias claves en los datos existentes y determinar las situaciones que conduzcan a nuevos hechos.

Con este análisis se pretende identificar comportamiento y tendencia de los accidentes de tránsito de acuerdo a la particularidad de la fecha. En esa exploración se identificó algunos días de la semana y fechas especiales que disminuyen y/o aumento la accidentalidad. Para identificar estos patrones, se utilizaron los gráficos de dispersión lineales y los diagramas de caja y bigotes, los cuales hacen más comprensible las interacciones y relaciones de estas variables

y se transforman en un insumo valioso en la construcción de los modelos predictivos.

5.3. Análisis descriptivo

Para la estimación de los modelos se procedió a dividir los diferentes dataframes en datos de entrenamiento que comprende los años del 2014 al 2017 y de validación con el año 2018 con el fin de evaluar la efectividad 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 por el método, para evaluar el p-valor de cada una y determinar las variables definitivas de los modelos, quedando de la siguiente manera:

Tabla 1. Variables seleccionadas para cada periodicidad

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 ajuste:

- Regresión lineal
- Knn
- Regresión lineal generalizado
- Árboles de regresión
- 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 (diario, semanal, mensual) de estas categorías.

Luego 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 el 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 crea tablas con las predicciones por cada una de las periodicidades.

6. Aplicación WEB

El objetivo de la aplicación web desarrollada es que el usuario pueda ver gráficamente los pronósticos diarios, semanales y mensuales para los años 2019, 2020 y 2021. También, que los usuarios pueda observar los datos predichos tablas que contienen el detalle de los mismos.

También se incluye ver los datos históricos del año 2014 al 2018 con su representación gráfica y sus respectivas tablas. Toda la información se presenta en diferentes pestañas para que el usuario interactúe de una mejor manera con la información de la accidentalidad de Medellín y los pronósticos que se realizaron sobre la misma.

7. Resultados

En la tabla 2 se muestra el resumen y el comparativo del RMSE que es el criterio de selección y validación de los modelos. De la tabla se concluye que los mejor ajustes para los accidentes totales y accidentes leves son la regresión lineal y para los accidentes graves los árboles de regresión. Un factor a favor de los resultados, es la poca variación entre los RMSE de los modelos con los datos de entrenamiento y los datos de validación donde la variación no superó el 15%.

Tabla 2. Comparación del RMSE de los modelos para la periodicidad Diaria.

Modelo	Total accidentes			Accidentes Graves			Accidentes Leves		
	RMSE Entrenamiento	RMSE Validación	% Variación	RMSE Entrenamiento	RMSE Validación	% Variación	RMSE Entrenamiento	RMSE Validación	% Variación
Regresión lineal	14,701	16,026	9,01%	9,816	11,337	15,50%	9,237	10,412	12,72%
Knn	17,492	20,910	19,54%	9,814	13,416	36,70%	11,152	12,573	12,74%
Regresión lineal Generalizado	14,493	33,566	131,60%	9,769	14,717	50,65%	9,082	21,546	137,24%
Árboles de regresión	16,143	16,320	1,10%	10,389	10,918	5,09%	10,021	11,141	11,18%
Bosques aletorios	11,053	16,887	52,78%	10,808	12,037	11,37%	6,963	11,194	60,76%

Modelos elegidos periodicidad diaria:

- **Total Accidentes:** Regresión lineal
- **Accidentes Graves:** Árboles de regresión
- **Accidentes Leves:** Regresión lineal

En la tabla 3 se muestra el resumen y el comparativo del RMSE que es el criterio de selección y validación de los modelos. De la tabla se concluye que los mejor ajustes para los accidentes totales y accidentes leves en su presentación semanal son la regresión lineal y para los accidentes graves son los bosques aleatorios. A pesar de no contar con bajos niveles de variación entre los RMSE de entrenamiento y validación (Factor que se puede ver influenciado por el nivel de los datos que presenta suma de los accidentes por semana) se logran ajustes significativos.

Tabla 3. Comparación del RMSE de los modelos para la periodicidad Semanal.

Modelo	Total accidentes			Accidentes Graves			Accidentes Leves		
	RMSE Entrenamiento	RMSE Validación	% Variación	RMSE Entrenamiento	RMSE Validación	% Variación	RMSE Entrenamiento	RMSE Validación	% Variación
Regresión lineal	44,036	63,668	44,58%	30,0	55,998	86,66%	25,765	33,319	29,32%
Knn	91,818	58,523	-36,26%	49,560	44,351	-10,51%	47,295	40,473	-14,42%
Regresión lineal Generalizado	44,056	85,567	94,22%	29,953	47,756	59,44%	25,878	47,341	82,94%
Árboles de regresión	92,584	62,529	-32,46%	51,982	49,940	-3,93%	49,718	42,026	-15,47%
Bosques aletorios	76,689	58,164	-24,16%	29,286	50,832	73,57%	41,795	37,559	-10,14%

Modelos elegidos periodicidad diaria:

- **Total Accidentes:** Regresión lineal
- **Accidentes Graves:** Bosques aleatorios
- **Accidentes Leves:** Regresión lineal

En la tabla 4 se muestra el resumen y el comparativo del RMSE que es el criterio de selección y validación de los modelos. De la tabla se concluye que los mejor ajustes para los accidentes totales y accidentes graves en su presentación mensual son la regresión lineal generalizada y para los accidentes leves es la regresión lineal. A pesar de no contar con bajos niveles de variación entre los

RMSE de entrenamiento y validación (Factor que se puede ver influenciado por el nivel de los datos que presenta suma de los accidentes por mes) se logran ajustes significativos.

Tabla 4. Comparación del RMSE de los modelos para la periodicidad Mensual

Modelo	Total accidentes			Accidentes Graves			Accidentes Leves		
	RMSE Entrenamiento	RMSE Validación	% Variación	RMSE Entrenamiento	RMSE Validación	% Variación	RMSE Entrenamiento	RMSE Validación	% Variación
Regresión lineal	83,919	205,348	144,70%	66,0	201,341	205,23%	59,583	66,603	11,78%
Knn	216,306	177,615	-17,89%	128,533	164,804	28,22%	114,519	125,714	9,78%
Regresión lineal Generalizado	83,803	110,882	32,31%	65,243	72,043	10,42%	59,309	74,353	25,37%
Árboles de regresión	204,392	247,030	20,86%	122,569	207,067	68,94%	110,910	121,382	9,44%
Bosques aletorios	91,790	178,597	94,57%	63,843	203,002	217,97%	63,930	66,223	3,59%

Modelos elegidos periodicidad diaria:

- **Total Accidentes:** Regresión lineal generalizado
- **Accidentes Graves:** Regresión lineal generalizado
- **Accidentes Leves:** Regresión lineal

8. Conclusión

En este trabajo se hizo un ejercicio aplicado de analítica predictiva con el uso de datos públicos ofrecidos por las entidades gubernamentales con el ánimo de identificar variable que puedan incidir en la accidentalidad y entregar las proyecciones de la accidentalidad para los años 2019, 2020 y 2021. A pesar de contar con información limitada de la accidentalidad de Medellín, a partir de esta se pudo identificar patrones con alta significancia que pueden dar indicios a la comunidad en general para entender factores que influyen en la accidentalidad.